

How To Use Genetic Algorithms To Build Efficient Machine Learning Models



Bun: Ushering In A New Era Of JavaScript Runtime

₹120
ISSN-2456-4885

OpenSource

Volume: 12 | Issue: 02 | Pages: 100 | December 2023

THE COMPLETE MAGAZINE ON OPEN SOURCE

ForYou

An **EFY** GROUP Publication

OPTIMISING DATA MANAGEMENT

Large Language Models:
Helping Manage Data

Managing Data On
The AWS Platform

Why Data Management And Security
Can Spell Success For Organisations

How To Use
RSS For Privacy
And Profit

From 'Mostly Harmless' Distros On A Pen
Drive, To Liberated Hardware And More
– a chat with **Abhas Abhinav**, entrepreneur and free software activist



Wanna Be Your Own Boss?

DO OPEN SOURCE. ←



Demand for Open Source is sky rocketing. Be it for managing IT infrastructure or development of software--Open Source solutions are what customers are seeking.

All you need to do is develop expertise in an Open Source stack, and then build a team around it!

And, Open Source For You can be your friend and a guide through this journey.

TO READ OUR PRINT EDITION Visit: <https://subscribe.efyindia.com>

TO READ OUR EZINE EDITION Visit: <https://ezine.lfymag.com>

WORLD'S LEADING PUBLICATION ON OPEN SOURCE

Looking for marketing solutions to engage with cutting edge techies?
Contact us at growmybiz@efy.in OR call us at +91-9811155335.



India's #1 Event For R&D and New Product Development Of Electronics

9TH EDITION
INDIA
ELECTRONICS
WEEK 2024

31st Jan -
2nd Feb 2024

KTPO Expo Center |
Whitefield | Bengaluru | India



For more information on sponsoring, exhibiting or attending, please call +91-9811155335 or growmybiz@efy.in

CONTENTS

DECEMBER 2023 | ISSN-2456-4885

FOR U & ME

- 20 Responsible AI: A Call by the US Government
- 23 Explainable AI: Enhancing Transparency and Trust
- 26 How to Use RSS for Privacy and Profit
- 30 Web Development: Explore Types and Career Paths

FOCUS

- 35 Role of Open Source in Designing SAN and NAS Systems
- 55 Large Language Models: Helping Manage Data
- 62 Synthetic Data: Illusions, Lies, Deceptions and the Reality

DEVELOPERS

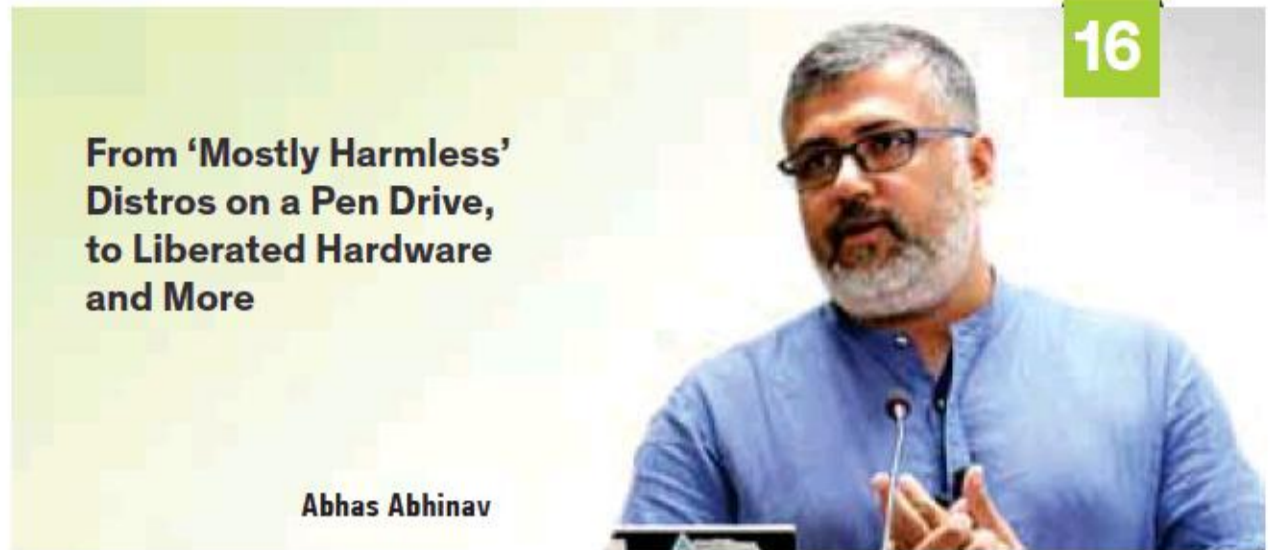
- 67 Putting G, P, and T Together
- 71 Blockchain Basics
- 75 How to Use Genetic Algorithms to Build Efficient Machine Learning Models
- 82 Bun: Ushering in a New Era of JavaScript Runtime
- 85 Real-Time Linux Applicability for Hard Real-Time Systems

ADMIN

- 93 Reducing the Size of Your Docker Image

OPEN GURUS

- 95 EdgeX Foundry: Transforming IoT with Edge Computing



REGULAR FEATURES

- 07 FossBytes

Wanna Support Open Source For You? Subscribe to the Magazine— so that we can keep promoting Open Source.



AMAZING OFFER

**Pay for
6 Issues
Get 12!**

(Buy One Get One Free!)

WORLD'S LEADING PUBLICATION ON OPEN SOURCE

ORDER FORM

Please
Tick (✓)
Your Choice

**Pay for 12 Issues
Get 24 Issues**
and save 50%
(Buy 12 Issues Get 12 Issues Free!)
₹1440

**Pay for 24 Issues
Get 48 Issues**
and save 50%
(Buy 24 Issues Get 24 Issue Free!)
₹2880

**Pay for 36 Issues
Get 72 Issues**
and save 50%
(Buy 36 Issues Get 36 Issue Free!)
₹4320

To subscribe online, visit
<https://tinyurl.com/y5kuv4la>

OR

SCAN
THIS
CODE



Name _____ Organisation _____ Mailing Address _____

Pin Code _____ State _____ Phone No. _____ City _____

Subscription No. (for existing subscribers only) _____ Email _____

I would like to subscribe to the above (✓)marked Open Source For You magazine starting with the next issue. Please

find enclosed a sum of Rs _____ by DD/MO/crossed cheque bearing the No. _____ dt. _____ in favour of EFY Enterprises Pvt Ltd, payable at Delhi. (Please add Rs 50 on non-metro cheque)

Please mark one (nearest) relating to your subscription: Indian Company MNC R&D organisation Engineering institute College/School Any other (specify): _____

Send this filled-in form or its photocopy to : EFY Enterprises Pvt Ltd, D-87/1 Okhla Industrial Area, Phase 1, New Delhi 110 020 | Ph: 011-40596600 | e-mail: support@efy.in

Terms:- # These rates are applicable for new subscribers as well as renewal by existing subscribers. # Can access ezine till your subscription is active # The rates are valid for subscribers within India only. # Please allow 4-6 weeks for processing of your subscription. # The subscription copies will be dispatched through ordinary post only # Subscription Agents will not get agency commission against this scheme # Disputes, if any, are subject to exclusive jurisdiction of competent courts and forums in Delhi/New Delhi only. * Replacement will be made if intimation of damaged / non-receipt of copies is received within 30 days of its publication ** After three months, if you are not satisfied with the magazine, your balance amount will be returned (Not applicable for gift offer)

EDITOR
RAHUL CHOPRA

EDITORIAL, SUBSCRIPTIONS & ADVERTISING
Delhi (HQ)
D-87/1, Okhla Industrial Area, Phase I, New Delhi 110020
Phone: +91-9811155335
E-mail: info@efy.in

MISSING ISSUES
Phone: +91-9811155335
E-mail: support@efy.in

BACK ISSUES
Phone: +91-9811155335
E-mail: support@efy.in

NEWSSTAND DISTRIBUTION
Phone: +91-9811155335
E-mail: efycirc@efy.in

ADVERTISEMENTS
NEW DELHI (HEAD OFFICE)
Phone: +91-9811155335
E-mail: efenq@efy.in

MUMBAI
E-mail: rmwest@efy.in

BENGALURU
E-mail: rmosouth@efy.in

CHINA
Worldwide Focus Media
E-mail: china@efy.in

GERMANY
pms Plantenberg Media Service GmbH
E-mail: germany@efy.in

JAPAN
Tandem Inc.
E-mail: japan@efy.in

TAIWAN
J.K. Media
E-mail: taiwan@efy.in

UNITED KINGDOM
ASA Media
E-mail: uk@efy.in

UNITED STATES
E & Tech Media
E-mail: usa@efy.in

Printed, published and owned by Ramesh Chopra. Printed at Tara Art Printers Pvt Ltd, A-46,47, Sec-5, Noida, on 28th of the previous month, and published from D-87/1, Okhla Industrial Area, Phase I, New Delhi 110020. Copyright © 2023. All articles in this issue, except for interviews, verbatim quotes, or unless otherwise explicitly mentioned, will be released under Creative Commons Attribution-NonCommercial 3.0 Unported License a month after the date of publication. Refer to <http://creativecommons.org/licenses/by-nc/3.0/> for a copy of the licence. Although every effort is made to ensure accuracy, no responsibility whatsoever is taken for any loss due to publishing errors. Articles that cannot be used are returned to the authors if accompanied by a self-addressed and sufficiently stamped envelope. But no responsibility is taken for any loss or delay in returning the material. Disputes, if any, will be settled in a New Delhi court only.

SUBSCRIPTION RATES			
Year	Newstand Price (₹)	You Pay (₹)	Overseas
Five	7200	4320	—
Three	4320	3030	—
One	1440	1150	US\$ 120

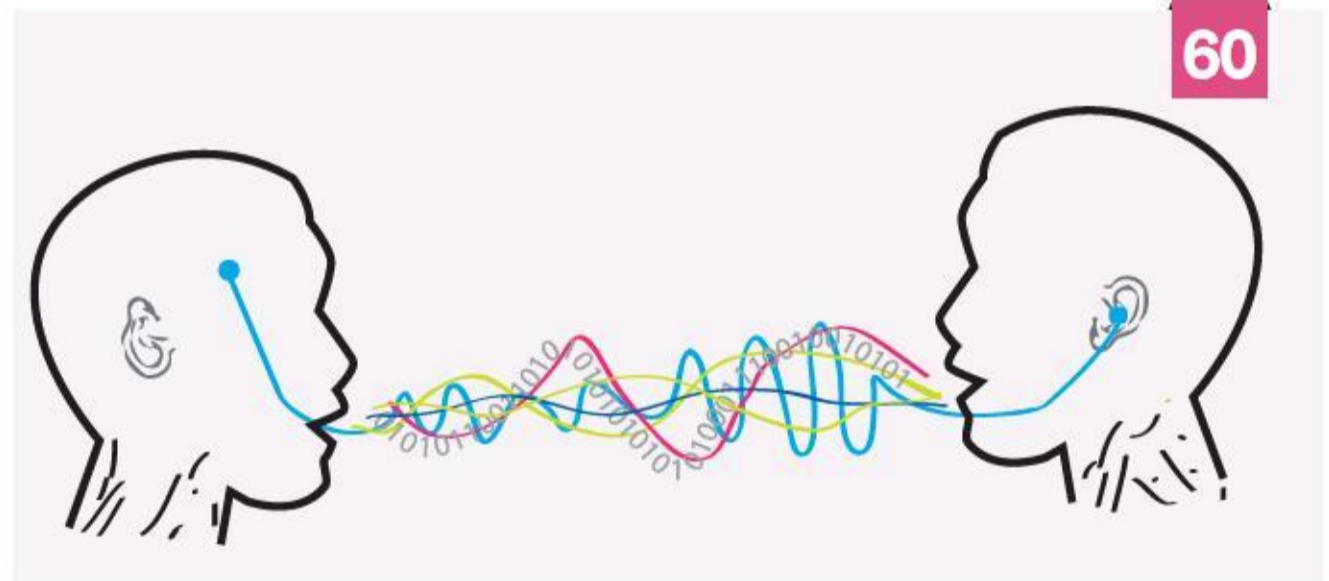
Kindly add ₹ 50/- for outside Delhi cheques.
Please send payments only in favour of EFY Enterprises Pvt Ltd.
Non-receipt of copies may be reported to support@efy.in—do mention your subscription number.

CONTENTS



50

Why Data Management and Security can Spell Success for Organisations



60

Encoding and Decoding in Python for Managing Data



64

Large Language Models: Enhancing IoT Security



78

Python: How it Helps Run IoT Devices

Elon Musk envisions X as an open source news platform



In a significant development, Elon Musk, the visionary behind SpaceX and Tesla, is positioning X, formerly Twitter, as an open source news outlet. Musk is causing a stir by advocating for a community-driven approach to news sharing on this platform.

Responding to concerns about misinformation on X, Musk highlighted the role of Community Notes, a tool that allows users to annotate posts and provide context. He made a clear distinction between X and mainstream media, emphasising that X operates on an open source philosophy, providing an alternative to tightly regulated narratives found in traditional news outlets.

Musk firmly believes that all the crucial information from legacy media is already available on X, asserting the platform's position as a hub for relevant news and discussions. However, some users have expressed doubts about its effectiveness in combating misinformation. Concerns include delays in the appearance of annotations, which may allow misleading information to spread, as well as accusations that X may inadvertently promote misinformation targeting American institutions.

X has recently changed its news-sharing format, abandoning headlines in favour of showing only images and linked web domains. This shift raises questions about how news is perceived and shared on the platform.

Additionally, Musk has invited journalists to publish their work directly on X, with the promise of higher earnings, aiming to entice professionals in the journalism industry. Musk's push for citizen journalism on the platform, encouraging live video reporting from ordinary citizens, is another significant development.

OpenText unveils an integrated partner network

OpenText has unveiled its integrated OpenText Partner Network. This consolidated network merges the partner ecosystems of both OpenText and the recently incorporated Micro Focus. This integration paves the way for enhanced support and expanded opportunities for a robust network of over 30,000 partners. Partners can tap into the company's extensive knowledge and wide-ranging

Tektronix unveils Python driver package for test & measurement

Tektronix has introduced an open source Python-native driver package for test and measurement (T&M) instrumentation. The package, named `tm_devices`, offers a native Python user experience, empowering users with auto-complete, precise type hinting, built-in help, real-time syntax checking, and debugging capabilities.

Rick Kuhlman, GM, Portfolio Software, said, "We aim to enhance productivity with Pythonic development tools and code reuse, encouraging collaboration and customisation."

The `tm_devices` package seamlessly integrates across Tektronix and Keithley devices, ensuring users in the test and measurement sector have access to the latest features and improvements. It facilitates ongoing development and updates, creating a dynamic environment.

"Automation typically involves dated string-based command APIs with shortcomings," explained Kuhlman. The `tm_devices` package, however, providing a single-line, seconds-long installation for Python auto-completion, type hinting, and real-time syntax checking – all free and open source.

Designed for simplicity, the driver package is now an integral part of Tektronix's development pipeline, adding an extra layer of quality and scrutiny. "With a quick install, no prerequisites, or unnecessary dependencies, it's extremely easy to get up and running," said Kuhlman.

The `tm_devices` open source Python driver package is globally accessible, marking a significant stride in democratising advanced T&M instrumentation tools.

Giskard's testing framework focuses on AI regulation

French startup Giskard's open source testing framework, designed for large language models (LLMs), addresses biases, security gaps, and content risks, and is positioned as a critical tool for AI regulation.

Giskard's open source Python library integrates seamlessly into LLM projects, notably retrieval-augmented generation (RAG) projects. Compatible with tools like Hugging Face, MLFlow, and TensorFlow, it caters to the practical needs of developers.

Post-setup, Giskard assists in generating a comprehensive test suite covering performance, hallucinations, misinformation, biases, and harmful content generation. CEO Alex Combessie said, "You'll have the performance aspect, which will be the first thing on a data scientist's mind. But more and more, you have the ethical aspect, both from a brand image point of view and now from a regulatory point of view."



Developers integrate tests into the continuous integration and continuous delivery (CI/CD) pipeline for regular scrutiny. Any anomalies trigger a detailed scan report directly in the GitHub repository.

Giskard's second product, the AI Quality Hub, serves as a debugging tool, gaining traction with companies like Banque de France and L'Oréal. It aims to evolve into a comprehensive documentation generator, facilitating compliance with regulations. The third product, LLMon, acts as a real-time monitoring tool for LLM answers.

With 20 dedicated team members, Giskard is gearing up to be the leading LLM antivirus in the market.

information management capabilities by coming under a unified programme structure. This ensures they can provide top-tier solutions that bring value to enterprise clients.

The company is channelling resources into R&D to forge cutting-edge technologies that the partners can leverage. The aim is to aid global customers in enhancing and safeguarding their foundational data and information channels to harness artificial intelligence's power.

Its programme for enterprise partners offers globally uniform procedures and benefits, all aimed at fostering future growth. The cloud acceleration programme



is tailored for mid-market partners, and emphasises generating new business and expansion avenues via the company's public cloud solutions.

Then there is the aviator thrust for partners, a dynamic new proposition that enables partners to design bespoke solutions for clients, leveraging the company's

cloud API services. This is especially beneficial for clients embarking on their AI journey or crafting sector-specific solutions.

The SolEx programme allows partners to function as an extension of OpenText products, presenting solutions that cater to distinct market demands.

The OpenText Partner Network operates in 150 nations worldwide. The revamped network is now accessible to all qualifying partners.

Tidelift's open source insights strengthen cybersecurity and risk management

In today's evolving cybersecurity landscape, reacting to late-stage risks is no longer adequate. Timely and accurate information is essential for early action and cost-saving. Tidelift provides access to a meticulously curated feed of open source package insights, allowing users to generate reports in line with government cybersecurity regulations.

The company collaborates with open source maintainers to ensure compliance with new security standards. It pays maintainers for their efforts, offering organisations better risk management and software assurance.

Tidelift's subscription includes exclusive insights from open source maintainers,



consolidated and structured data from various package manager ecosystems, and enhanced human-researched data to provide more contextualised insights.

Moreover, Tidelift aligns with US government standards, confirming adherence to the NIST Secure Software Development Framework

(SSDF), which is crucial for government contracts. The subscription provides a standardised attestation report and an automated solution to actively monitor and maintain attestation records for open source components.

To assist organisations in managing and enforcing open source policies, Tidelift offers a software bill of materials within the subscription, creating a centralised inventory of open source components. This simplifies the identification of compromised packages when addressing vulnerabilities, ensuring the use of approved, secure components across all development teams.

iXsystems names Patrick Keane as vice president of engineering

iXsystems, the driving force behind TrueNAS open storage, has announced the appointment of Patrick Keane as its vice president of engineering, effective October 17, 2023. This move is a significant step forward for TrueNAS technology and the company's continued expansion in the market.



Keane's track record over two decades of tech leadership spans various industries, including direct marketing, telecommunications, broadcast media, medicine, data visualisation, and high-performance computing. He has a history of delivering enterprise storage solutions, having previously led engineering

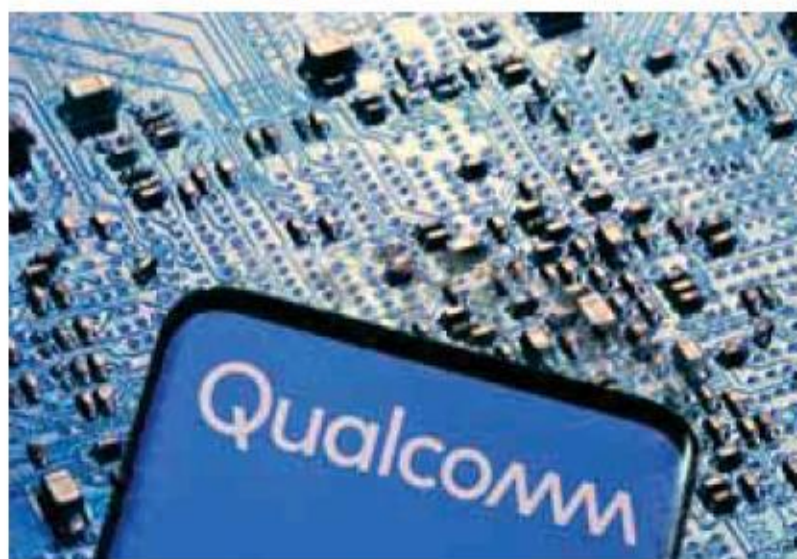
teams at renowned organisations such as NetApp and, most recently, as vice president of engineering at Panasas, Inc.

Kris Moore, senior vice president of engineering at iXsystems, expressed excitement about this development: "Patrick's addition to our engineering leadership team brings valuable enterprise knowledge and experience to our company at an exciting time in the history of companies. More enterprise customers are adopting TrueNAS, and he will be instrumental in advancing the technology to serve in an expanding set of use cases in enterprise environments."

Keane's primary mission will be to drive innovation in iXsystems' storage solutions. He said: "I have been a TrueNAS enthusiast for years, both in my personal and professional life. TrueNAS is a stand-out product in the storage industry for its rich feature set, ease of use, and rock-solid reliability for enterprise customers and home users. The opportunity to collaborate with the open source community and the amazing team at iXsystems to drive TrueNAS innovation is a dream job."

Qualcomm to partner with Google for RISC-V based wearables

Semiconductor chipmaker Qualcomm has announced an extension of its long standing partnership with Alphabet's Google to make wearable devices using RISC-V technology. The wearables will have Google's Wear OS, which is an



Android operating system designed for smart watches and other wearables.

RISC-V is an open source instruction set architecture (ISA) that competes with ARM's proprietary ISA, which is a dominant choice for embedded systems and mobile devices.

Qualcomm announced its intention to launch the RISC-V-based wearables solution worldwide, including in the United States.

Fleet boosts Mac enterprise security

Fleet, the open source endpoint security company, has shaken up Mac enterprise security by integrating osquery for script execution on managed devices. This move addresses the needs of admins constrained by conventional mobile device management platforms.

Fleet, known for its open architecture and GitOps embrace, now enables IT departments to approve and execute script changes on demand. This development offers a flexible solution for businesses managing hybrid IT operations.

Co-founder Zach Wasserman, also the co-creator of osquery, emphasised Fleet's open source nature, providing business owners with a versatile approach to managing and protecting their IT infrastructure. He noted, "Fleet allows you to escalate from a read-only system to a script execution system, allowing for not just monitoring but remediation as well."



Wasserman underscored the advantage of an open source approach, steering away from the need to continuously add software to tackle problems. The aim is to provide a streamlined architecture that can easily adapt to new challenges.

In contrast to one-size-fits-all solutions, Wasserman believes Fleet offers the flexibility and freedom to adapt and grow without being locked into proprietary solutions, especially beneficial for businesses with budget constraints.

Dell infrastructure adopts Hugging Face's GenAI models

Dell Technologies and Hugging Face, the generative AI language model supplier, have joined forces to integrate custom GenAI models onto Dell server and storage products. The partnership extends Hugging Face's influence beyond its renowned open source large language models, including Distilbert-sst-2-int8.

AMD and CloudFlare have previously aligned with Hugging Face. Now, Dell enters the fray, facilitating on-premise deployment of customised large language models through a dedicated portal on the Hugging Face platform.

Jeff Boudreau, Dell's chief AI officer, emphasised the partnership's benefits: "Dell Technologies' GenAI capabilities, combined with Hugging Face's collection of data sets and libraries, offer users the freedom of open source GenAI models with the peace of mind of on-premise data security and reliability."

Jeff Boudier, head of product and growth at Hugging Face, concurred: "This collaboration will empower companies to build their own AI systems leveraging the incredible innovations of the open source community, while benefiting from the security, compliance and performance of Dell systems."

The partnership ushers in the Hugging Face Dell portal, a platform enabling users to deploy open source models seamlessly on Dell servers and storage. Dell, solidifying its commitment to AI, recently unveiled the ObjectScale XF960 all-flash storage appliance and Dell Validated Designs for AI.

Hugging Face lists over 300,000 large language models in its catalogue. The platform provides pre-trained models, libraries, tools, and resources, simplifying AI utilisation.

Recently, both companies teamed up with other industry leaders to initiate the RISC-V software ecosystem (RISE). Qualcomm Technologies also disclosed its investment in a firm to further RISC-V hardware development.

Dino Bekis, vice president and general manager, wearable and mixed signal solutions, Qualcomm Technologies, said that they are excited to leverage RISC-V and expand their Snapdragon Wear platform as a leading silicon provider for Wear OS.

General manager of Wear OS by Google, Bjorn Kilburn, said that Qualcomm has been a pillar of the Wear OS ecosystem, providing high performance, low power systems for many of their OEM partners.

VNC releases VNClagoon 2.0, the collaboration software for enterprises

VNC has unveiled VNClagoon 2.0, an iteration of its modular software stack tailored for enhanced communication and collaboration. This highly scalable



application significantly reduces rollout and deployment times by leveraging Kubernetes-driven microservices and micro-modules. The backend optimisation ensures seamless operation in complex setups.

VNClagoon 2.0 has the native integration of all functional modules in the

front end. This innovation allows businesses to create customised workspaces for individual employees, teams, departments, or entire organisations, harnessing the power of the fluid principle. Modules and dashlets can be combined and rearranged as needed, giving users great flexibility. Users can choose to deploy the suite on-premise, in the cloud, or hybrid architectures.

Recognising the importance of partnerships, VNC has introduced tailored partner packages. These packages are designed to serve the diverse needs of partners in the indirect sales channel. They range from VNCbasic (P1), ideal for start-ups, small and medium-sized enterprises (SMEs), municipalities, and local authorities, to VNCprofessional (P5), which caters to large enterprises as well as federal and state authorities. VNCcharity (P0) offers a cost-effective solution for non-profit organisations. Partners can also leverage the white-labelling feature to customise VNClagoon 2.0 to their branding or clients' corporate identity.

Andrea Wörrlein, managing director of VNC in Berlin and a board member of VNC AG in Zug, emphasised, "VNClagoon 2.0 is suitable for virtually all sizes of companies and public authorities, from start-ups to large corporations, thanks to its modular design, infinite combination options, easy deployment, flexible provisioning, and end-to-end interoperability of modules and dashlets."

OpenAI's ChatGPT Plus unveils a host of new features

OpenAI is making waves in the world of artificial intelligence once again with the introduction of new features for its ChatGPT Plus, a generative AI chatbot. Among the notable additions is the 'advanced data analysis' feature, currently in the testing phase and available to select users on paid tiers.

One of the standout capabilities of this feature is its ability to allow users to seamlessly upload files within a conversation and work with them, all while being assisted by the AI-based chatbot. This functionality extends to the analysis and summarisation of data sheets that users upload. ChatGPT goes beyond



simply responding with a brief textual summary; it can also generate visual representations based on the provided data. Additionally, the chatbot can analyse textual data to answer questions posed by the user. Notably, the version currently under testing exhibits a

remarkable level of autonomy, as it doesn't necessitate users to select a specific model for data analysis manually. Instead, it can intelligently discern the appropriate model for the given data type, streamlining the user experience.

However, ChatGPT's talents are not confined solely to text files. Users can instruct the chatbot to generate iterations of uploaded image files as well. To accomplish this, ChatGPT leverages the power of the DALL-E3 model, which excels in generating images within the context of a conversation.

xAI unveils witty AI chatbot Grok

Elon Musk's new AI venture, xAI, has introduced Grok, an open source artificial intelligence (AI) chatbot. Inspired by 'The Hitchhiker's Guide to the Galaxy', Grok is designed to possess a witty personality and to tackle questions other AI chatbots may avoid.

With access to data from X, Grok is expected to have a competitive edge, although xAI acknowledges that, like any large language model (LLM), Grok is



still capable of producing inaccurate or conflicting information.

Currently in early beta and two months into its training, Grok is open to a select group of users, with a wider release planned for the future. Interested individuals

can join a waitlist to try the bot, which will eventually become a part of the X Premium+ subscription, priced at \$16 per month.

Musk positions xAI, backed by a team from OpenAI and DeepMind, as a competitor to existing AI firms. In preliminary tests, Grok has outperformed other models in its compute class on certain tasks, although it is still outmatched by bots with access to larger data sets (like ChatGPT 4.0).

The term 'Grok' originates from Robert A. Heinlein's novel 'Stranger in a Strange Land', signifying a profound, intuitive understanding. xAI, which was set up in July, aims to create inclusive AI that caters to diverse perspectives, countering the trend of 'politically correct' systems. Grok is an experiment to apply xAI's ethos in a real-world scenario.

WordPress 6.4.1 maintenance release fixes bugs in version 6.4

WordPress users can breathe a sigh of relief as a maintenance release, version 6.4.1, has been swiftly rolled out to address critical issues that emerged shortly after the release of WordPress 6.4, code named Shirley, on November 7.

This new maintenance release comes as a lifesaver for users who encountered several bugs, with two of them having significant implications for the functionality of certain plugins, potentially wreaking havoc on affected websites.

Atlassian acquires AirTrack

Atlassian Corporation has successfully acquired Melbourne-based AirTrack, bolstering its commitment to enhancing enterprise asset and configuration management. This strategic move will empower organisations to manage vital assets more effectively, reduce operational risks and expenses, and enhance cybersecurity.

The integration of AirTrack's expertise with Jira Service Management will enable enterprises to oversee their critical assets through several key benefits such as access to 30+ pre-built data connectors for streamlined data management, swift identification of discrepancies across record-keeping systems, and enhanced visibility into the infrastructure that supports critical services and engineering components.

This acquisition extends Atlassian's capability to assist enterprises in reimagining asset management beyond traditional IT applications, addressing a broader array of security and operational challenges.



Additionally, Atlassian has introduced advancements to Jira Service Management, including the availability of virtual agent capability, with further AI-driven features on the horizon.

The acquisition of AirTrack reinforces Jira Service Management's asset management capabilities, enhancing its overall functionality. The Premium and Enterprise Editions of Jira Service Management now offer accessible virtual agent capabilities empowered by artificial intelligence. These virtual agents assess user intent, sentiment, context, and profile data for personalised interactions, continuously improving through natural language processing.

Canonical launches MicroCloud

Canonical has introduced MicroCloud, a game-changer in cloud computing designed for scalable clusters and edge deployments. MicroCloud prioritises simplicity, security, and automation, allowing businesses to deploy a fully functional cloud solution with a single command, minimising deployment time.

Security takes centre stage as MicroCloud ensures strict confinement and over-the-air transactional updates for data integrity. The solution offers flexibility, supporting deployment on-premise or at the edge, providing a consistent experience regardless of location.

Cédric Gégout, VP of product at Canonical, highlighted MicroCloud's response to the evolving cloud landscape: "As data becomes more distributed, the infrastructure has to follow. MicroCloud is our answer to that, delivering a simpler, more prescriptive cloud experience that makes zero-ops a reality."

MicroCloud is versatile, running on both commodity and high-end hardware, and is compatible with Ubuntu Desktop, Server, and Core. Workloads can be executed using Kubernetes or system containers, ensuring efficiency with reduced resource consumption.



Offered within Canonical's Ubuntu Pro subscription, MicroCloud provides users with an enterprise-grade, open source solution supported by different tiers. It adopts a transparent per-node pricing model, eliminating hidden fees.

The most pressing issue addressed by WordPress 6.4.1 was a critical bug that caused users to encounter download failures, displaying a disheartening error message - "cURL error 28: Operation timed out." The root of this problem can be traced back to a major change in WordPress 6.4, specifically an update to the Requests library version. This update included a breaking change for users on hosts with cURL version 7.29 or older, which was, unfortunately, more common than expected. The latest cURL version is 8.4.0, highlighting the need to stay up to date.

Another critical issue stemmed from the removal of code that was no longer used by the WordPress core, an action taken in an attempt to streamline and optimise the platform. However, the removal of this code inadvertently disrupted various plugins that still relied on it. The consequences were far-reaching, causing issues for users and their websites. The solution provided by WordPress 6.4.1 was to reintegrate the removed code, ensuring that backward compatibility was maintained and that plugins could function without hiccups.

The final issue tackled by this maintenance release was relatively minor in comparison but still noticeable. A typographical error affected the appearance of a nag screen in the administrator panel, causing it to stretch awkwardly across the top of the page. This typo was rectified with the release of WordPress 6.4.1.

India's developer community surges; eyes 2027 crown

India's developer community is booming, with a remarkable addition of 3.5 million developers in 2023, bringing the total to an impressive 13.2 million. GitHub's 2023 State of the Octoverse report predicts that by 2027, India could overtake the United



States, becoming the world's largest developer hub. This significant growth is not just numerical but is substantiated by India's contributions to open source projects.

The report highlights a 36% increase in India's

developer community in 2023, demonstrating its consistent upward trajectory. Moreover, India plays a pivotal role in collaborating with the UN-backed Digital Public Goods Alliance, contributing to the development of a digital public infrastructure encompassing software and AI models.

The report underscores the global surge in individual contributors to generative AI projects, showcasing 148 percent year-over-year growth. Besides India, the United States and Germany stand out as leaders in developer communities, with Japan, Hong Kong, the United Kingdom, and France following closely behind.

MikroElektronika unveils ClickID

MikroElektronika (MIKROE) has introduced ClickID, a hardware/software open source solution that aims to transform the landscape of embedded development. Based on a one-wire device, it enables effortless hot-plugging of Click or any mikroBUS-compatible add-on peripheral boards into development boards running embedded Linux or similar operating systems.

ClickID eliminates the need for manual configurations at boot-time. According to Nebojsa Matic, MIKROE's CEO, "ClickID brings a new level of automation and self-diagnosis to embedded tools. We have eliminated the challenge of configuring devices at boot-time with device trees, which is time-consuming and error-prone."

Facilitating complex operations, ClickID allows software to identify add-on board peripheral configurations seamlessly. A demo library is available, demonstrating the reading, writing, and processing of data from the ClickID. The full application code and ready-to-use projects are easily accessible through NECTO Studio Package Manager, LibStock, or MIKROE's GitHub account.

For Linux-based systems like BeaglePlay, ClickID simplifies the addition and operation of peripheral add-on boards. Leveraging the 1-Wire protocol, it communicates with the operating system kernel, enabling automatic recognition of silicon ICs on add-on boards. This results in the delivery of manifest files to the operating system, streamlining the loading of appropriate drivers for interfaces such as SPI, I2C, UART, PWM, and ADC.

Christine Long, CEO of BeagleBoard.org Foundation, commented, "This is a big step enabling users of boards like our BeagleBoard.org BeaglePlay to simply use the well-tested code that is already there without needing to search and integrate; just connect and play!"

Silo AI launches Poro, the open source LLM for EU languages

Silo AI, the Helsinki-based AI startup, has unveiled Poro, an open source large language model (LLM) designed to enhance multilingual AI capabilities for all 24 official European Union languages. Poro is the inaugural model in a series, developed by SiloGen and the University of Turku's TurkuNLP research group.

"It is a digital sovereignty question where you want to ensure that there are models that are capturing the value base, the culture, the languages," said Peter Sarlin, CEO of Silo AI. "Ultimately, it's about value creation, ensuring that not



only European but any company out there can create value that stays within Europe and stays within that organisation."

Named after the Finnish word

for 'reindeer', Poro features a 34.2 billion parameter model utilising a BLOOM transformer architecture with ALiBi embeddings. Training on a 21 trillion token multilingual data set, including English, Finnish, Python, and Java, it addresses the challenge of creating effective natural language models for lower-resourced European languages.

Poro is trained on LUMI, Europe's fastest supercomputer in Kajaani, Finland, equipped with 512 AMD Instinct MI250X GPUs. The Poro Research Checkpoints program ensures transparency in documenting the model's training progress. Sarlin noted, "There aren't initiatives that have given such transparency to model training."

Teaming up with the University of Turku, Poro represents collaboration between applied AI expertise and academic leadership in multilingual language modelling research. "If Poro delivers on its promise, it could democratise access to performant multilingual models, giving Europe a homegrown alternative to systems from US tech companies," Sarlin noted.

Microsoft and Google collaborate on OneTable

Microsoft and Google have teamed up with Onehouse to back the OneTable project, reshaping the cloud data lake scene. This collaboration tackles the challenges organisations face in choosing data lake table formats, aiming to prevent vendor lock-in and compatibility hurdles for data analytics and AI workloads.

OneTable, initiated by Onehouse and announced in February 2023 alongside a \$25 million funding raise, is positioned to become a crucial layer above existing data lake table formats—Apache Iceberg, Apache Hudi, and Databricks' Delta Lake. This layer allows seamless conversions and access across these formats, unlocking new possibilities for interoperability.

Microsoft's involvement is aligned with its data lake approach called Fabric, supporting the Delta Lake table format. Raghu Ramakrishnan, CTO for data at Microsoft, stressed the importance of enabling openness, stating, "We want a pathway where people can buy into our ecosystem without feeling blocked." This move is part of Microsoft's strategy to create a



single, open framework.

Gerrit Kazmaier, VP data and analytics at Google Cloud,

explained that organisations often struggle to choose a primary table format, leading to potential long-term consequences. OneTable, he noted, is not replacing different data lake table formats but alleviates the burden of choosing one that might result in lock-in.

For more news, visit www.opensourceforu.com





Asia's #1 Open Source Conference

Total registrations
15,372

Total attendees
4743

Speakers
80



20th Edition

OPEN

SOURCE INDIA

Thank You

VISITORS, for making the event a huge success

SPEAKERS, for contributing your valuable thoughts

PARTNERS, for your support



For more details, call on 98111 55335 or email us at info@opensourceindia.in

From 'Mostly Harmless' Distros on a Pen Drive, to Liberated Hardware and More

Abhas Abhinav



Abhas Abhinav's FOSS journey began in the late nineties. His love for open source has only grown stronger with the years, though he is not too happy about the way free and proprietary software are being mixed at will in today's world.

Have you heard of the Mostly Harmless USB disk? What does it consist of? How does it work?

To learn more, meet Abhas Abhinav. He is a long-time community member of Free and Open Source Software (FOSS) networks in India, particularly (but not restricted to) Bengaluru. He tells the story of this disk: "It started as an experiment to offer a convenient and always accessible tool to showcase and use a free OS anywhere. I have made multiple variants of it over the past two years. One was for the LibrePlanet conference in Boston, where we gave it out to some of the participants gratis — or at a price of their choosing."

There are ten OSs in the Mostly Harmless USB disk — Debian, Ubuntu, Fedora, and Alpine Linux are live and installable. Tails is a portable operating system that protects against surveillance and censorship. Tails bootable allows you to browse the internet over TOR. Short for The Onion Router, TOR is free and open source software for enabling

anonymous communication. As Wikipedia explains, it directs internet traffic via a free, worldwide, volunteer overlay network that consists of more than seven thousand relays. The price of the disk is ₹650.

As a part of the DebConf'23 conference (held in Kochi in 2023) swag, a unit of the MH-USB was given out to all the participants. This meant a total of around 300 units, in all, got distributed. This USB contained multiple images of various Debian GNU/Linux operating systems and derivatives.

Abhas recalls: "If you remember, there was a time when [Indian computer magazines] *PC Quest*, *Linux For You* (now known as *Open Source For You*), *DeveloperIQ*, and *CHIP* distributed a GNU/Linux CD. This helped precipitate GNU/Linux adoption in the late 1990s. Then, in the early 2000s, Ubuntu would send you a nice colourful CD if you just filled out a form on their website [and requested it]."

Abhas says his own endeavour is to try and replicate these instances in a

different way. He hopes that having a bootable multi-OS USB handy on their key rings all the time "might enable people to do something that they found challenging in the past."

But, he admits, it's not a financially viable product. Abhas actually gives out more units than he sells. "But that's fine with me since I look at it as a way to increase access to a bootable GNU/Linux OS."

Many people appreciate it, and he finds that as "nice feedback." Even a lack of feedback "does not really affect my motivation to continue spending money on fabricating and distributing MH-USB units. I think I will keep building and updating it as long as I can."

The MH-USB partition of the USB drive is formatted with exFAT, allowing universal access to the data on any operating system that supports exFAT.

Today, he feels, the dominant way of using software is to just use a browser. "Then, the software is not on your computer or a computer that you control," he adds.

Liberated hardware products

So what are his new initiatives?

“Several,” he says.

‘Liberated hardware’ is hardware that can run free software but isn’t necessarily designed to do so. At the LibreTech.Shop, over the past year, Abhas has listed quite a few ‘liberated hardware’ products that offer free software on a device by default.

These include:

- Mobile phones running LineageOS, the open source Android distribution available for several devices (or Ubuntu Touch, designed primarily for mobile devices, or PostmarketOS, tiny and under 10MB in size).
- Wireless routers running OpenWRT, meant for embedded devices.
- ‘Smartwatches’ that run a free OS called AsteroidOS, the FOSS operating system for smart watches.
- Liberated computers (laptops and desktops) that offer a perfect GNU/Linux experience out-of-the-box and are powered by Coreboot (the free software BIOS replacement). Coreboot describes itself as “an extended firmware platform that delivers a lightning fast and secure boot experience on modern computers and embedded systems. As an open source project it provides auditability and maximum control over technology.”
- Home automation solutions, including wireless switches and Wi-Fi tubelights to start with. Wireless switches enable you to switch on/off a mains outlet over the network. A Wi-Fi tubelight lets you switch the tubelight over the network as well as change its colour and brightness, all with 100% free software and without requiring internet access or a proprietary mobile app.

Abhas says: “What I am attempting to do is remove the last bit of hesitation that someone might have with using free software



in more places. If I can sell the same products that people might purchase off-the-shelf with free software pre-loaded, that might be the convenience they’re searching for. The ‘sales’ are not really encouraging right now, but I am hoping that setting up a stall at various community events might help change that.”

“I’m trying to tell somebody about the beauty of GNU/Linux, and I don’t have a USB disk at hand to show it to them live on their computer. Or to even install a thing that they can test and see first-hand. We also need to fix computers sometimes, or to install something. We need to have something all the time,” he explains.

Since he got started in the late 1990s, how has his work changed, and how has the discourse changed?

“My vision was to run a business where I could charge people for running free software for them. It remains the same even today. Ironically, it’s getting more difficult to do that sometimes, as compared to the early 2000s. At that time, people were used to the idea of installing software on their own computers,” Abhas adds.

Tech... and politics

Does he agree with the view that the Indian techie often doesn’t understand

— or care for — the politics involved in technology? Says Abhas: “It depends on who you talk to and which circles you’re in. But a lot of people do not. That is true.”

Thankfully, as he puts it, journalists and activists are more aware of this aspect today. “They understand the dangers of proprietary tracking systems.”

He argues: “The respect for the freedom we have only grows when we lose some of it or all of it. But many have not thought about losing freedom.”

Server space

Two decades back, when we first met, Abhas was very much into the server space. Does his passion for the server continue?

“Yes. I still work with servers. It’s important to do self-hosting at the individual or family level,” he says.

Many people want to use free software. One option is to host it yourself. You need a small server that you would pay for. When you install your own software, you run your email, activity-powered social presence, microblogging, and videos. Use it to sync pics from your phone, or run a chat server. All of that is possible, he points out.

But it doesn’t come easy. The problems one encounters can be twofold — most people don’t know

how to get started. They often think there are many challenges to get a server up and running.

“But by opting out of that decision making, they are also opting for proprietary software. We are trying to show them various ways to do it differently. Do it on Raspberry Pi, a cheap server, or even on a wireless router. We show the examples, and people can choose the place where they want to host it,” he explains.

“If others control all this software, then they control the software which runs on it, and the data which resides on it. They control the way it goes out to others, and to whom. That is not what the internet was meant to be. It was a place where we all run our own space, and run it in a truly federated manner,” he notes.

The FOSS journey

As a student, in 1997-98, Abhas recalls constantly seeking new things to do with his computer. This led to experimentation with other operating systems (OSs) such as OS2 and Windows NT. During this time, he bought a book to learn about Linux, which everyone was talking about. He taught himself a “little bit of LaTeX” too. “Then, there was no looking back,” he remembers.

FOSS, he believes, is a serious option for youngsters to generate employment and acquire valuable skills. He points to the bigger picture: “When you feel that pressure is building up in the place you’re in right now, you will make a different choice. Free software is one of the best ways to do that. Using free software can often be the best way to solve a problem.”

Biggest challenges

What does Abhas see as the biggest challenges before FOSS (Free and Open Source Software) or FLOSS (Libre, too) software in India today?

He says, “I feel that, in general, GNU/Linux usage is decreasing, and that is a concerning trend. For a lot of

Some of the sites worth visiting, related to Abhas’ work:


1. Hardware solutions that don’t compromise your privacy, freedom, opportunity to learn, do things on your own: <https://mostlyharmless.io>
2. Ethical and hackable products: <https://libretech.shop>
3. Online support; discourse forum: <https://ask.libre.support>
4. What is a free software business: where you combine respect for your users’ freedom with your choice of building a business: <https://freesoftware.business>
5. Commemorating 20 years of DeepRoot GNU/Linux: <https://20.deepproot.in>
6. ActivityPub: <https://sovrان.me>
7. Hacks.Wiki (an experiment to organise quick hacks, notes, bookmarks and tools into an easy-to-build-and-maintain ‘digital garden’): <https://hacks.wiki/About/>
8. Sovran Video powered by PeerTube, an ActivityPub-federated video streaming platform using P2P directly in your web browser: <https://sovrان.video>
9. Sovran Photos: <https://sovrان.photos/@abhas>
10. Keypad: <https://sovrان.dev/mh/vcky66> / <https://libretech.shop/product/kids-klack> (Abhas built this to help his daughter attend online school classes during Covid)
11. Writings: <https://abhas.io>

younger people, it is entirely normal to run free software on proprietary operating systems or alongside a mix of proprietary software.”

Apart from desktop operating systems, there are other systems involved too that are proprietary. A lot of free software is developed around GitHub, a proprietary version control system and community operated by Microsoft. There are many more such online proprietary ecosystems. “For some reason, we no longer feel uncomfortable about mixing proprietary and free software in this way,” he argues.

If you ask a group of students how they would build an online group for collaboration, the most popular responses are: WhatsApp, Telegram, Discord, etc — all proprietary systems. “The first tools of choice are not email, mailing lists, matrix or something else

that is free software and can be self-hosted and maintained by ourselves. This blurring of the line between free and proprietary software is what is worrisome,” he adds.

Several things have disappeared from the common ‘FLOSS discourse’ in India. No one really talks about localisation (Indic language solutions in FOSS) anymore. Or the freedom to compute freely and privately. Or even having fun, Abhas says, with a tinge of regret. As he puts it: “Lest we forget, Torvalds wrote the Linux kernel ‘just for fun’ (as goes the title of his autobiography). He wanted to understand how the 80386 CPU from Intel worked, and he decided to write an operating system kernel for that. The rest is history... but it all started with the need to have a certain sort of fun. And I feel that this too is missing.” **END** 

 By: Frederick Noronha

The author is a Goa-based freelance journalist, who has been tracking FOSS, Creative Commons and the Wikipedia issues for a long time.

Stay Connected. Stay Informed. Stay Ahead.



SUBSCRIBE AND SAVE

ORDER FORM

PRINT MAGAZINE	1 YEAR (12 copies each)	3 YEARS (36 copies each)	5 YEARS (60 copies each)
Electronics For You (Rs 100/copy)	WITHIN INDIA (IN RUPEES)		
	840 <input type="checkbox"/>	2150 <input type="checkbox"/>	3000 <input type="checkbox"/>
	SAARC COUNTRIES (IN US\$ BY AIR)		
	50 <input type="checkbox"/>	135 <input type="checkbox"/>	NA
OTHER COUNTRIES (IN US\$ BY AIR)			
100 <input type="checkbox"/>	270 <input type="checkbox"/>	NA	

PRINT MAGAZINE SUBSCRIBERS GET:

- Free e-magazine every month
- Free delivery of print magazine by post
- And much more (check: subscribe@efy.in)
- For delivery by courier, please add Rs 50 for each copy

To subscribe online, visit
<https://payment.efyindia.com>

OR

SCAN
THIS
CODE



e-magazine subscriptions within India are available at half the rates mentioned here.
Overseas rates for each e-magazine in US\$: 1 year: \$12; 3 years: \$33; 5 years: \$50 only

Name _____ Organisation _____ Mailing Address _____

City _____

Pin Code _____ State _____ Phone No. _____ Email _____

Subscription No. (for existing subscribers only) _____ I would like to subscribe to the above (✓)marked magazine(s) starting with the next issue. Please find enclosed a sum of Rs. _____ by DD/MO/crossed cheque bearing the No. _____ dt. _____ in favour of EFY Enterprises Pvt Ltd, payable at Delhi.

Please mark one (nearest) relating to your subscription: Indian Company MNC R&D organisation Engineering institute College/School Any other (specify): _____

Send this filled-in form or copy to : EFY Enterprises Pvt Ltd, D-87/1 Okhla Industrial Area, Phase 1, New Delhi 110 020 | Ph: 011-40596600 | e-mail: support@efy.in

Terms:- # These rates are applicable for new subscribers as well as renewal by existing subscribers # Please allow 4-6 weeks for processing of your subscription.
Please include your pincode for prompt delivery of your copy.

Responsible AI: A Call by the US Government

A new Biden executive order calls for responsible AI principles that ensure innovation and growth with safety and privacy, without promoting biases of any kind. It's something all of us need to follow.

As I write this, the US administration has issued an executive order on ensuring the safe, secure, and trustworthy development and use of AI. This government involvement underscores the rapid adoption of AI, necessitating the need for an executive order — a first of its kind on technology adoption. While the order focuses on the US and its interests in addressing security, the economy, and society, it also provides a framework for industries and organisations to adopt in their AI journey, with a deliberate focus on its core principles. The article attempts to elaborate on the guiding principles and to outline key actions that will shape the industry in the coming years.

As per the executive order, responsible AI adoption with secure and reliable data is to be governed by eight guiding principles and priorities. Let's look at each one of them to understand what they entail and the expectations they set.

Ensuring AI safety and security

The first guiding principle of the executive order emphasises the importance of ensuring that AI is developed and used in a safe and secure manner. This principle recognises that AI has the potential to cause harm if it is not properly designed, tested, and monitored. To achieve this goal, the order calls for the development of standards and best

practices for AI safety and security, as well as mechanisms for identifying and mitigating risks associated with AI. The order also encourages the use of privacy-enhancing technologies and other measures to protect sensitive data and prevent unauthorised access to AI systems.

Promoting responsible innovation, competition, and collaboration

The second guiding principle of the executive order focuses on promoting responsible innovation, competition, and collaboration in the development and use of AI. This principle recognises that AI has the potential to drive economic growth and innovation while also acknowledging the need to ensure that these benefits are shared fairly and equitably. To achieve this goal, the order calls for the promotion of public-private partnerships, the development of AI standards and best practices, and the establishment of mechanisms for sharing data and other resources. The order also encourages the use of open data and open source software to promote collaboration and innovation.

Supporting American workers in responsible AI development

The third guiding principle of the executive order emphasises the importance of supporting American workers in the development and use of AI. This principle recognises that AI has the potential to displace workers and



Image Source: <https://jacksonvillefreepress.com>

exacerbate existing inequalities. It calls for measures to ensure that workers are trained and equipped to succeed in the AI-driven economy. To achieve this goal, the order calls for the development of workforce training programs, the promotion of apprenticeships and other forms of on-the-job training, and the establishment of mechanisms for worker participation in the development and deployment of AI systems. The order also encourages the use of AI to improve working conditions and increase productivity.

Advancing equity and civil rights in AI

The fourth guiding principle of the executive order emphasises the importance of advancing equity and civil rights in the development and use of AI. This principle recognises that AI has the potential to perpetuate and exacerbate existing biases and inequalities, calling for measures to ensure that AI is developed and used in a fair and equitable manner. To achieve this goal, the order calls for the development of AI standards and best practices that promote fairness, transparency, and accountability, as well as mechanisms for identifying and mitigating bias in AI systems. It also encourages the use of AI to promote diversity and inclusion in the workplace and in society at large.

Protecting user interests in AI and AI-enabled products

The fifth guiding principle of the executive order emphasises the importance of protecting the interests of Americans who use, interact with, or purchase AI and AI-enabled products in their daily lives. This principle recognises that AI has the potential to impact consumers in a variety of ways. It calls for measures to ensure that consumers are protected from harm and have access to accurate and transparent information about AI systems. To achieve this goal, the order

calls for the development of consumer protection policies and regulations, as well as mechanisms for monitoring and enforcing these policies. It also encourages the use of AI to improve consumer experiences and increase access to information.

Protecting privacy and civil liberties in AI

The sixth guiding principle of the executive order emphasises the importance of protecting Americans' privacy and civil liberties as AI continues to advance. This principle recognises that AI can collect and analyse vast amounts of personal data and calls for measures to ensure that this data is collected, stored, and used in a manner that respects individuals' privacy and civil liberties. To achieve this goal, the order calls for the development of privacy-enhancing technologies and other measures to protect sensitive data, as well as mechanisms for monitoring and enforcing privacy policies. It also encourages the use of AI to enhance privacy and security.

Regulating and governing AI risks

The seventh guiding principle of the executive order emphasises the importance of managing the risks associated with the federal government's use of AI and increasing its internal capacity to regulate, govern, and support responsible use of AI. This principle recognises that the federal government is a major

user and developer of AI, calling for measures to ensure that AI is developed and used in a manner consistent with public interest. To achieve this goal, the order calls for the development of AI policies and regulations that promote transparency, accountability, and public participation, as well as mechanisms for monitoring and enforcing these policies. The order also encourages the use of AI to improve government services and increase efficiency.

Promoting innovation and competition with leadership

The eighth guiding principle of the executive order emphasises the importance of promoting innovation and competition in the development and use of AI. This principle recognises that AI has the potential to drive economic growth and innovation, and calls for measures to ensure that these benefits are shared fairly and equitably. To achieve this goal, the order calls for the promotion of public-private partnerships, the development of AI standards and best practices, and mechanisms for sharing data and other resources. It also encourages the use of open data and open source software to promote collaboration and innovation.

These principles aim to ensure that AI is developed and used in a responsible and safe manner, while also promoting innovation, competition, and collaboration. They also prioritise supporting American workers, advancing equity and civil rights, protecting consumers,



Figure 1: AI guiding principles in the US executive order



Figure 2: US government calls for action on development and use of AI

and safeguarding privacy and civil liberties. The executive order emphasises that these principles should be adhered to by executive departments and agencies while considering the views of other agencies, industry, members of academia, civil society, labour unions, international allies and partners, and other relevant organisations.

These calls for action aim to promote safe, responsible, fair, privacy-protecting, and trustworthy AI systems; support small businesses in innovating and commercialising AI, advance responsible global technical standards for AI development and use outside of military and intelligence areas, and address cross-border and global AI risks to critical infrastructure.

‘Red Teaming’ is a common practice where organisations hire ethical hackers to emulate real attacks against enterprise systems to ensure appropriate security measures are in place. This practice is evolving to

ensure the adoption of responsible AI. OpenAI has created a community for the open source world to collaborate and communicate in the development of a framework with a network of industry experts. Although private organisations are going to invest in similar efforts, it’s encouraging to see rapid developments in the open source community for establishing a platform for individuals to join and contribute to the positive impacts of AI.

In conclusion, AI is a transformative force that unlocks the creative potential within all of us. By harnessing the power of AI responsibly in the data age, we can unleash our imagination, redefine creativity, and embark on a journey of boundless innovation. It’s the dawn of a new era as technology advances and industries adopt responsible AI, making it an integral part of our daily lives. **END** 🐧

References

- The White House executive order on development and use of AI: <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/?tsrc=366>
- OpenAI Red Teaming Network: <https://openai.com/blog/red-teaming-network>

By: Bala Kalavala

The author is a technical architect, evangelist, thought leader, and sought-after keynote speaker. He currently works as digital transformation business development executive in a global technology consulting firm.

Disclaimer: This article expresses the views of the author and not of the organisation he works in.

Explainable AI:

Enhancing Transparency and Trust

AI is giving us answers to a million questions, but we don't really know how it arrives at them. Explainable AI (XAI) helps understand what went into making a decision, enhancing trust in the system, especially with respect to ethical practices.



Artificial intelligence (AI) is transforming industries with sophisticated decision-making and automation technologies. However, as AI systems become more complicated, the need for openness and interpretability to understand and trust AI-driven choices grows. The field of explainable AI (XAI) has emerged just to address this.

Explainable AI and model interpretability are critical not just for comprehending AI decisions, but also for fostering trust and assuring ethical AI practices. As AI continues to enter numerous industries, the demand for transparent and interpretable AI models grows. Striking the right balance between transparency and performance,

using relevant XAI methodologies, and adhering to legislation and ethical principles are crucial for realising AI's full potential in a responsible and accountable manner.

As AI plays an increasingly important role in our lives, it is critical that AI systems are transparent, accountable, and adhere to ethical norms. Adopting explainable AI not only assures compliance with present and new rules, but also upholds ethical norms, allowing AI to benefit society while respecting individual rights and values.

Table 1 gives a quick overview of how explainable AI fits into the paradigm of artificial intelligence research activities.

The need for explainable AI

The field of AI has experienced exponential growth, driven by factors such as the increasing volume of data, post-COVID expectations in automation, and the pursuit of operational excellence in digital experiences. There are multiple factors that have contributed to the importance of explainable AI.

Trust and accountability:

AI decisions in vital fields like healthcare, banking, and autonomous cars can have life-changing effects. Understanding how and why AI arrives at certain conclusions is critical for building trust and responsibility.

Type of AI	Key features	Use cases
Explainable AI	<ul style="list-style-type: none"> Provides transparent explanations for AI decisions Helps users understand how AI models work Enhances trust and accountability 	Healthcare diagnostics, Fintech - financial risk assessment, legal decisions, regulatory compliance
AI fairness	<ul style="list-style-type: none"> Mitigates biases in AI algorithms Ensures equitable treatment of different groups Prevents discrimination in AI-driven decisions 	Hiring processes, lending and credit decisions, criminal justice system
Emotion AI	<ul style="list-style-type: none"> Detects and interprets human emotions Analyses facial expressions, voice tone, and text sentiment Enhances human-computer interaction 	Customer service, market research, mental health monitoring
Responsive AI	<ul style="list-style-type: none"> Adapts and responds to user inputs and context Improves user experience through natural interactions Personalised recommendations and services 	Chatbots, virtual assistants, recommendation systems, e-commerce
Generative AI	<ul style="list-style-type: none"> Creates new content, text, images, or music Can produce creative and original outputs Used in content generation and creative tasks 	Content generation, art creation, language translation, chatbots

Table 1: What different types of AI do

Fairness and bias: AI models might inherit biases from their training data, resulting in discriminating conclusions. Explainability assists in identifying and correcting these biases, assuring fairness and regulatory compliance.

Regulations and compliance: Many businesses are governed by stringent regulations that demand transparency and responsibility. Explainable AI assists organisations in adhering to legal and ethical requirements.

Human-AI collaboration: To effectively collaborate with AI, humans must grasp AI’s reasoning. This is especially true in medical and legal applications.

Techniques for enhancing model interpretability

Model interpretability in the context of explainable AI (XAI) refers to the extent to which humans can comprehend and explain an AI model’s predictions and decisions. It involves making complex and frequently inscrutable machine learning models more transparent and understandable, so that users, stakeholders, and experts can gain insight into how the model arrives at

particular outputs or decisions.

Model interpretability seeks to increase transparency in artificial intelligence models. This requires revealing how the model processes input data and produces output predictions. Transparent models are simpler to comprehend and trust. Here are some methods that improve model interpretability.

SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations): These methods assign relevance scores to model elements, making it easier to comprehend their impact on predictions.

Conclusion trees and rule-based models: These present a clear and interpretable collection of the factors that led to a specific conclusion.

Grad-CAM (Gradient-weighted Class Activation Mapping): These techniques highlight the most relevant regions in images that influenced the model’s choice.

Local explanations vs global explanations: Depending on the context and requirements, AI models can provide explanations for individual

predictions (local) or for the entire model (global).

Balancing transparency and performance in AI models

Balancing transparency and performance in AI models is crucial for responsible AI development and deployment. Transparency refers to the ability to understand and explain the decision-making process of an AI model, while performance refers to the accuracy and effectiveness of the model in achieving its intended purpose.

Trust and accountability: When users can understand how AI models make decisions, they are more likely to trust and accept those decisions. This is particularly important in high-stakes applications, such as healthcare or criminal justice.

Fairness and bias detection: Transparency allows for the detection and mitigation of biases in AI models. By understanding how models arrive at their conclusions, it becomes possible to identify and address potential sources of bias, ensuring that AI systems operate fairly and impartially.

Debugging and improvement:

Transparency facilitates the debugging and improvement of AI models. By examining the inner workings of a model, developers can identify areas where the model is making errors or suboptimal decisions, enabling them to refine and enhance its performance.

High-performance AI models are essential for solving real-world problems. If an AI model consistently makes inaccurate or ineffective decisions, it will not be adopted or trusted by users. However, challenges do arise in balancing transparency and performance.

Challenges in balancing transparency and performance

There is often a trade-off between transparency and performance in AI models. Highly transparent models, such as rule-based systems, may lack the complexity and flexibility to achieve high levels of accuracy. Conversely, highly complex models, such as deep learning models, may be difficult to understand and explain.

Several strategies can be employed to balance transparency and performance in AI models.

Explainable AI (XAI) techniques: XAI techniques aim to make AI models more transparent without compromising their performance. This can involve using techniques such as feature importance analysis, partial dependence plots, and SHapley values to explain how models make decisions.

Model architectures: Choosing appropriate model architectures can also help balance transparency and performance. For instance, decision trees are inherently more transparent than deep neural networks, while still being able to achieve high accuracy in certain domains.

Ensemble models: Combining multiple models into an ensemble can improve overall performance while maintaining a degree of transparency. By understanding the contributions of

individual models to the ensemble's predictions, users can gain insights into the decision-making process.

Human-AI collaboration: In certain scenarios, human-AI collaboration can be a viable solution. Humans can provide oversight and guidance to AI models, ensuring that decisions are made in a transparent and accountable manner.

Use cases of explainable AI in different sectors

XAI has found practical applications in quite a few sectors.

Healthcare: AI that can be explained is essential in medical diagnosis and therapy recommendations. Interpretable models can assist clinicians in understanding AI predictions and making educated judgements.

Finance (banking, financial services, securities, capital markets, insurance): XAI is used in banking and finance for risk assessment, fraud detection, and credit scoring. Models that can be interpreted provide transparency and aid in regulatory compliance.

Legal: XAI supports legal professionals with contract analysis, legal research, and case prediction. Clear explanations boost trust and productivity.

Autonomous vehicles: It is critical to ensure road safety. Explainable AI in self-driving cars facilitates comprehension of how the AI makes decisions in difficult traffic circumstances.

Manufacturing: Artificial intelligence (AI) is employed in Industry 4.0 for quality control, predictive maintenance, and process optimisation. Explainability is required for problem solving and process improvement.

Explainable AI (XAI) has gained attention due to its potential to address the opacity of many AI models, often referred to as 'black boxes'. By making AI systems more transparent and understandable, XAI aims to

enhance trust, accountability, fairness, and responsible AI development. However, the regulatory and ethical implications of XAI also raise important considerations.

Regulatory implications of XAI

As AI systems become more pervasive and impactful, regulatory bodies worldwide are increasingly addressing the need for transparency and explainability. Several jurisdictions have implemented or are considering regulations that mandate or encourage XAI for certain AI applications, particularly those with significant societal impact, such as healthcare, finance, and criminal justice. Examples of regulatory initiatives include:

The General Data Protection Regulation (GDPR) in the European Union requires that AI systems that make decisions about individuals be explained to those individuals upon request.

The National Artificial Intelligence Strategy of France emphasises the importance of XAI and calls for the development of guidelines and standards for explainable AI systems.

The Algorithmic Accountability Act proposed in the United States aims to promote fairness and transparency in AI systems, including requirements for explainability in certain applications.

These regulatory developments underscore the growing recognition of the need for XAI to ensure responsible AI practices and protect individuals from the potential harm caused by opaque AI systems.

Ethical implications of XAI

Ethical considerations also play a crucial role in the development and deployment of XAI. The ability to explain AI models raises questions about the potential misuse of this information and the need for safeguards to protect individuals' privacy and autonomy. Key ethical concerns include:

Continued to page...29



How to Use RSS for Privacy and Profit

With RSS you can reclaim your online freedom. Say no to tracking, censorship, and information overload. Experience the internet on your terms!

Social media may tempt website owners by bringing in a swarm of visitors once in a while, but that is essentially a bait-and-switch trap. Ultimately, the social media companies will want to be paid. There are other less onerous ways to grow the audience for a website. Email is one of them, but not everyone wants to consume news, opinion or entertainment through email. How about RSS? It is a time-tested strategy to build a loyal audience without exposing them to the intrusive tracking done by tech giants.

Even if you do not leave the mobile data or the wireless connection of your phone on all the time, there will be continuous tracking. Smartphones are designed to continue to report your movements even in airplane mode or when the phone is turned off. You agree to this incessant tracking when you mindlessly click the 'Next' button when you first set up the phone (<https://www.opensourceforu.com/2016/05/how-to-run-android-without-google/>). The only way to avoid this tracking is to use a feature/button phone or no phone at all. Unfortunately, this is not an easy option for everyone.

When you are using a desktop or a laptop, do not be logged into Google. Then, how will you check your email? Do not check your mail in the browser. Use dedicated software for it, known as an email client like Thunderbird or Seamonkey to separate your email from your browsing (<https://www.opensourceforu.com/2018/02/tricks-try-thunderbird-seamonkey/>).

Similarly, do not log into Facebook or Twitter to read your news. It gives them the ability to censor news and track you. Go directly to the source, anonymously. But there are potentially dozens or hundreds of news sources. Who has the time to visit all of them? There is a simple solution — RSS.

An RSS feed

RSS stands for 'really simple syndication'. It was created by Aaron Swartz, who was also the creator of Markdown, Creative Commons and Reddit.

RSS is an XML-based technology for notifying consumers of fresh content published online. An RSS file

published by a website is known as an RSS feed or a news feed or a web feed. Like a web page, an RSS feed will be accessible from a World Wide Web address.

<http://www.vsubhash.in/rss.xml>
<http://rss.cnn.com/rss/edition.rss>

A website can have multiple news feeds.

<http://www.vsubhash.in/blogs/blog/rss.xml>
<http://www.vsubhash.in/blogs/tweets/rss.xml>

Some websites provide topic-specific RSS feeds.

<https://news.google.com/rss/search?q=ukraine>
<https://stackoverflow.com/feeds/tag?tagnames=java>

A website may provide reader comments in a thread-specific RSS feed.

<https://www.opensourceforu.com/2016/05/how-to-run-android-without-google/feed/>
<https://stackoverflow.com/feeds/question/77349789>

Real-time data can be published as RSS feeds.

<https://rssfeed.azure.status.microsoft/en-us/status/feed/>
https://w1.weather.gov/xml/current_obs/PAMR.rss

RSS clients

End users who wish to read or be notified of the fresh content using these RSS links will have to use RSS client applications. RSS clients are variously known as RSS readers, news readers, or feed readers. RSS readers are available as desktop and mobile applications. Browsers and email clients may natively support RSS feed or with add-ons. There are some online RSS reader services but they defeat the idea of using RSS for anonymity.

When you encounter a site that you like, find out its RSS link. It is usually behind an orange icon as shown in Figure 1. Some news websites may have a dedicated page where they list all their RSS links.



Figure 1: RSS feeds are advertised using this easily identifiable icon

An undeclared war on RSS?

Facebook and Twitter used to have RSS feeds when they were growing their audience. To monetise their apps, they removed RSS support and became walled gardens. There

was also an undeclared war on RSS (<https://www.smh.com.au/technology/apple-joins-the-war-on-rss-20120731-23d8y.html>). Web browsers used to indicate the availability of an RSS feed in the address bar, but they quietly removed the feature. If Google, Facebook and Twitter can become the gateways of online content, it becomes easier to censor news, opinion and information.

Does YouTube have RSS feeds?

Strangely, YouTube video channels continue to support RSS. They do not display the links anywhere. You have to manually construct it in this format:

https://www.youtube.com/feeds/videos.xml?channel_id=_____

The blank part must be replaced with the internal channel ID of whichever video you are watching. Now that YouTube has assigned Twitter-like user names to channels, the channel IDs are not easy to copy. However, you can open the web console (*Tools » Web Developer Tools » Console*) and type 'ytInitialPlayerResponse.videoDetails.channelId'; you will get the channel ID in quotation marks. For the *Electronics For You* video channel, I got the ID as "UCJbnMYV_yigckub3RjtXD1Q". With this ID, the channel RSS link can be constructed as:

https://www.youtube.com/feeds/videos.xml?channel_id=UCJbnMYV_yigckub3RjtXD1Q

To automate this task for other channels, create a bookmarklet in your browser with this code:

```
window.alert("https://www.youtube.com/feeds/videos.xml?channel_id=" + ytInitialPlayerResponse.videoDetails.channelId);
```

What is a bookmarklet? It is a bookmark that uses JavaScript code instead of an ordinary WWW address. The code has to be specially formatted for use in a bookmark. There are a lot of bookmarklet formatter website services that you can use for this purpose.

Use RSS for profit

A major problem with many web design and marketing professionals is that they know all about social media and nothing about RSS. When they are asked to improve the reach of an existing website, the first thing they do is remove support for RSS feeds. Their reasoning is that RSS readers allow people to stay inside their RSS readers and never visit the site publishing the content. This is a misconception. A website needs to publish only the article headlines in its RSS feed. It does not have to publish the entire article content in

the feed. Persons interested in reading an article will click the link and visit the site.

Paradoxically, the same web marketing individuals will add social media plugins to every page in the website. Will that not cause the reader to leave the website? Apparently, that is okay. Don't you want your content to become viral? Unfortunately, this traffic is very fickle and will not even return. Easy come, easy go. Another important thing that web marketers ignore is that not every social media mention is seen by all followers. To limit information overload and to monetise traffic leaving to external sites, social media companies deliberately limit the visibility of users' posts. In other words, subtle censorship. To gain visibility, they want you to advertise. In other words, subtle extortion. The sites that advertise get more visibility. You have to pay money to one guy to bring in traffic and then make money from another who displays ads. The gamble may not pay off.

In contrast, the anonymous RSS costs almost nothing extra. Growth of this organic traffic may be linear and slow. Over time however, the publisher can reap the benefits of an ever-growing audience that no middleman can censor. Publishers must foster this kind of loyal traffic that is dependable. Social media traffic can dry up suddenly one day when they inexplicably change their algorithm (as they have done many times in the past) or if payments cannot be made.

Instead of diverting traffic to social media sites, publishers should ask visitors to subscribe to their RSS feeds. When a feed indicates that there is valuable content on a site, the subscribers will click and visit the site. When they visit the site, the publisher can then monetise the traffic by:

- Serving unobtrusive ads to the visitors, that is, no annoying overlays or pop-ups
- Provide easy links to subscribe to:
 - An ad-free paywall
 - A print edition

RSS is not the only game in town. Online publishers can use other mediums to monetise their content. Email newsletters are one of them. Some people prefer email newsletters as their email clients enable them to perform offline searches of news articles. The same headline, link, snippet, and intro image strategy used in RSS feeds can be used in email newsletters too. However, email is subject to filters and deliverability issues. You need to proceed with great care as irresponsible use of email newsletters can result in blacklisting of your domain. You can read more about this subject in my free ebook 'Email Newsletter Strategies For Profit'.

Anatomy of an RSS feed

An RSS feed is so simple that it can be hand-coded. Most website content management systems (CMS) and blogging software automatically generate RSS feeds.

The initial part of the XML is metadata about the feed. In this metadata, the RSS feed identifies itself and also provides a link back to itself. What follows next are individual news items or article metadata wrapped inside item tags. Each item tag includes a title (or headline) and a link for the news item, status update, article, or blog post. A GUID tag is required to provide a globally unique ID to make it easy for an RSS reader to uniquely identify the item in its database. It is common practice to use the URL of the news item as a GUID if it happens to be unique.

```
<?xml version="1.0"?>
<rss version="2.0">
  <channel>
    <title>Weather forecast for Palakkad</title>
    <description>Daily weather forecast provided by ISRO</description>
    <link>https://www.isro.gov/weather-forecast/palakkad.xml</link>

    <item>
      <title>Warm sunny weather</title>
      <description>Clear sky.</description>
      <link>https://www.isro.gov/weather-forecast/palakkad-2023-10-20.html</link>
      <guid>https://www.isro.gov/weather-forecast/palakkad-2023-10-20.html</guid>
      <pubDate>Sun, 20 Oct 2023 00:00:00 +0530</pubDate>
    </item>

    <item>
      <title>Possibility of rain</title>
      <description>It may or may not rain.</description>
      <link>https://www.isro.gov/weather-forecast/palakkad-2023-10-19.html</link>
      <guid>https://www.isro.gov/weather-forecast/palakkad-2023-10-19.html</guid>
      <pubDate>Sat, 19 Oct 2023 00:00:00 +0530</pubDate>
    </item>

    <item>
      <title>Stormy weather and lightning</title>
      <description>Fisherman are advised not to venture into the sea.</description>
      <link>https://www.isro.gov/weather-forecast/palakkad-2023-10-18.html</link>
      <guid>https://www.isro.gov/weather-forecast/palakkad-2023-10-18.html</guid>
      <pubDate>Fri, 18 Oct 2023 00:00:00 +0530</pubDate>
    </item>
  </channel>
</rss>
```



Figure 2: RSS reader with display of news

RSS readers display news items in chronological order, just like an email client. They can display content from multiple sites.

You can place rich-text content including images inside the description tags using HTML (with CDATA formatting for XML). In this respect, RSS feeds are better than email newsletters. Email clients will usually block images, but RSS clients display images in all their glory. Just add a good intro image and an interesting text snippet to each news item, and there will be no stopping the traffic.

Podcasts

Podcasts are also published as RSS feeds. They have an extra enclosure tag providing the location of the MP3 file.

```
<enclosure url="https://www.allindiaradio.gov.in/news-in-english/2023-10-20-nie-en.mp3" length="2324324324" type="audio/mpeg" />
```

RSS readers and podcast apps consume the RSS feeds and enable you to download and play the MP3 file.

Most RSS readers have built-in support for podcasts.

A lot of people have no time to read articles. Instead, they listen to audiobooks and podcasts when they are commuting.

A news organisation can convert a select few of its news articles as podcasts and deliver them to these subscribers using the same RSS feeds. Podcasts can also be monetised by adding advertisements or other promotional material. **END** 🐧

By: V. Subhash
The author has written 30 books including 'Quick Start Guide to FFmpeg (Apress)', 'Linux Command-Line Tips & Tricks', 'How To Install Solar' and 'Cool Electronic Projects'.

Continued from page...25

Data privacy and discrimination: XAI techniques may reveal sensitive personal information about individuals, raising concerns about data privacy and the potential for discrimination based on this information.

Explainability for whom?: XAI explanations may be tailored for technical experts, making it difficult for non-experts to understand or interpret them. This could exacerbate existing power imbalances and hinder public trust in AI systems.

Explainability vs bias: XAI techniques may inadvertently reinforce existing biases in AI models if they fail to account for the underlying factors that contribute to biased outcomes.

Explainability for what purpose?: XAI explanations should not be used to justify or legitimise decisions made by AI systems, especially in ethically

sensitive domains. Instead, they should serve as a tool for critical evaluation and accountability.

Navigating these ethical challenges requires a thoughtful approach to

XAI development and deployment. It is essential to consider the potential impact of XAI explanations on individuals, communities, and the society as a whole. **END** 🐧

- References**
- <https://cloud.google.com/explainable-ai>
 - <https://www.ibm.com/topics/explainable-ai>
 - <https://insights.sei.cmu.edu/blog/what-is-explainable-ai/>
 - <https://www.darpa.mil/program/explainable-artificial-intelligence>
 - <https://www2.deloitte.com/nl/nl/pages/innovatie/artikelen/a-call-for-transparency-and-responsibility-in-artificial-intelligence.html>
 - <https://www.mckinsey.com/capabilities/quantumblack/our-insights/why-businesses-need-explainable-ai-and-how-to-deliver-it>
 - <https://www.finantrix.com/explainable-ai-and-regulatory-reporting/>

By: Dr Magesh Kasthuri and Dr Anand Nayyar
Dr Magesh Kasthuri is a senior distinguished member of the technical staff and principal consultant at Wipro Ltd. This article expresses his views and not that of Wipro.
Dr Anand Nayyar is a PhD in wireless sensor networks and swarm intelligence. He works at Duy Tan University, Vietnam, and loves to explore open source technologies, IoT, cloud computing, deep learning and cyber security.

Web Development: Explore Types and Career Paths

Embarking on a web development journey, whether you're drawn to front-end, back-end, or full-stack, promises a rewarding and exciting experience in this ever-evolving field. Just take the plunge!



In today's digital age, web development plays a crucial role in shaping our online experiences. From the interactive websites we browse to the complex web applications we rely on, skilled web developers bring these virtual worlds to life. But what exactly are the different types of web development? And where do the job opportunities lie? In this article, we will delve into the realms of front-end web development, back-end web development, and full-stack web development, exploring their unique characteristics, job prospects, and market trends.

Front-end web development

Front-end web development focuses on crafting visually appealing and interactive user interfaces. Front-end developers are responsible for translating designs into functional

websites. They utilise HTML, CSS, and JavaScript to structure and style web content, ensuring seamless interaction with users. They bring designs to life, implementing responsive layouts and interactive elements. With the help of frameworks like React, Angular, and Vue.js, front-end developers create dynamic and engaging websites and web applications.

Technologies and tools: At the core of front-end development are HTML, CSS, and JavaScript—the foundational trio of web technologies. They provide the building blocks for structuring web content, styling elements, and adding interactivity. Additionally, front-end developers leverage frameworks like React, Angular, and Vue.js to enhance efficiency and streamline development processes.

Use cases and examples: Front-end developers excel at creating captivating user interfaces, optimising website performance, and designing user experiences. From implementing responsive designs to building dynamic web applications, their skills shine in bringing ideas to life. For instance, a front-end developer might work on an e-commerce platform, designing intuitive product pages and enabling seamless user interactions.

Challenges: When it comes to front-end development challenges, ensuring seamless connectivity across different devices and internet speeds can be a real head-scratcher. Here is the list of challenges that are most critical to front-end development.

- **Cross-browser compatibility:** Ensuring that websites and web applications function consistently across different web browsers can be a challenge due to variations in rendering engines.
- **Responsive design:** Designing and developing interfaces that adapt seamlessly to different devices and screen sizes requires knowledge of responsive design techniques.
- **Performance optimisation:** Balancing aesthetics and functionality while optimising website performance can be challenging, especially when dealing with large amounts of data or complex animations.

Job opportunities and market trends: The demand for front-end developers remains high across industries. Companies prioritise user experience, driving the need for skilled front-end

PIONEERS IN TECHNOLOGY. PARTNERS IN SUCCESS.

EFYGROUP
YOURS SINCE 1969

**"It has helped me
reach this position."**

**-Mukesh Ajwani
MD, Electronic Components
Trading Firm**



Mr Ajwani is not the only one who owes his success to *Electronics For You*. The technology magazine that started off in 1969 is today, South Asia's most popular one. Covering the latest on emerging technologies, changing industry trends, evolving trade practices...

Not surprisingly, top decision makers and those high up on the corporate ladder in the electronics fraternity treat EFY as their Bible.

- Over half a million readers
- India's largest selling technology magazine
- Caters to the entire electronics fraternity

professionals. According to industry reports, the average salary for front-end developers ranges from ₹ 800,000 to ₹ 1.5 million per annum, depending on experience and location. Progressive web apps (PWAs) have emerged as a notable trend, offering enhanced offline functionality and app-like experiences.

Learning curve: A key part of the learning curve for front-end development is mastering techniques to optimise website connectivity for smooth user experiences.

- Learning HTML, CSS, and JavaScript fundamentals is essential.
- Staying updated with evolving web technologies, frameworks, and best practices is crucial.
- Gaining design sensibilities to create visually appealing and user-friendly interfaces is an asset.

Back-end web development

While front-end development focuses on the client-side experience, back-end web development handles the server-side logic and data management. Back-end developers work with programming languages like Python, PHP, Ruby, or Java to handle data processing, integrate with databases, and implement complex business logic. They build the functional backbone of web applications, ensuring smooth operations and secure data storage. Back-end developers work behind the scenes, making sure that user actions are processed accurately and efficiently.

Technologies and tools: Back-end developers work with server-side programming languages like Python, PHP, Ruby, and Java. These languages facilitate handling data, integrating with databases, and implementing complex business logic. Popular databases such as MySQL, MongoDB, and PostgreSQL are utilised to store and manage application data. Frameworks like Node.js, Django, and Laravel aid in building robust web applications and provide APIs for seamless integration.

Use cases and examples: Back-end developers excel at building scalable systems, handling data processing, and implementing security measures. They work behind the scenes to ensure that user actions are processed correctly, and data is securely stored. For example, a back-end developer might build the backend for a financial application, handling complex calculations, and ensuring data security.

Challenges: Back-end development comes with its fair share of challenges.

- **Scalability and performance:** Building robust and scalable back-end systems that can handle high traffic and large data sets can be challenging.
- **Security:** Implementing strong security measures to protect sensitive data and prevent unauthorised access requires knowledge of secure coding practices.
- **Integrations and APIs:** Working with third-party services, databases, and APIs can present challenges, especially when dealing with different data formats and authentication mechanisms.

Job opportunities and market trends: The tech industry continues to seek talented back-end developers to support the ever-growing number of web applications. The average salary for back-end developers ranges from Rs 1 million to Rs 1.8 million per annum. Notable trends in back-end development include serverless architecture and the adoption of microservices to build scalable and modular systems.

Learning curve: Here are a few most crucial factors that form the learning curve for back-end development.

- Learning server-side programming languages such as Python, PHP, or Ruby.
- Understanding database management systems and query languages.
- Acquiring knowledge of web frameworks and libraries for efficient development.

Full-stack web development

Full-stack web development combines both front-end and back-end skills, allowing developers to work across the entire web development stack. Full-stack developers possess a well-rounded understanding of web technologies, enabling them to handle end-to-end development tasks. They are proficient in both client-side and server-side technologies, helping them to design user interfaces, implement business logic, and connect with databases. Their versatile skillset allows them to handle end-to-end development tasks, making them valuable assets in building comprehensive web applications.

Technologies and tools: Full-stack developers are proficient in front-end technologies like HTML, CSS, JavaScript, as well as back-end technologies such as server-side programming languages, databases, and web frameworks. They often work with MEAN (MongoDB, Express.js, Angular, Node.js), MERN (MongoDB, Express.js, React, Node.js), or LAMP (Linux, Apache, MySQL, PHP) stacks to build dynamic and versatile web applications.

Use cases and examples: Full-stack developers excel in managing the entire project life cycle, from front-end design and development to back-end implementation and database integration. They are well-equipped to handle both client-side and server-side aspects of web development. For example, a full-stack developer might create a social media platform, where they build the user interface, handle user authentication, and manage data storage and retrieval.

Challenges: The main challenge for full-stack development is to develop a deep understanding of data flow and communication. Here is the list of the most critical challenges for the process.

- **Balancing multiple technologies:** Full-stack developers need to master both front-end and back-end technologies, which can be demanding and require constant learning and adaptation.

A comparison of important metrics and data

Metrics	Front-end development	Back-end development	Full-stack development
Focus	User interface design	Server-side logic	End-to-end development
Technologies	HTML, CSS, JavaScript	Server-side languages (e.g., Python, PHP, Ruby, Java)	Combination of front-end and back-end technologies
Key skills	UI/UX design, HTML, CSS, JavaScript, Front-end frameworks (e.g., React, Angular, Vue.js)	Server-side programming languages, Database management, API integration	Front-end and back-end technologies, Problem-solving, Project management
Challenges	Cross-browser compatibility, Responsive design, Performance optimisation	Scalability, Security, Integrations and APIs	Balancing multiple technologies, Project management, Debugging and troubleshooting
Learning curve	HTML, CSS, JavaScript fundamentals, Design principles, Front-end frameworks	Server-side programming languages, Database management systems, Web frameworks	Comprehensive knowledge of front-end and back-end technologies, Problem-solving skills, Project management
Job opportunities	High demand for UI-focused roles, Front-end development agencies, Tech companies	Growing demand for back-end developers, Software development firms, E-commerce platforms	Versatile job opportunities, Startups, Full-stack development agencies
Popular frameworks	React, Angular, Vue.js	Node.js, Django, Laravel	MEAN, MERN, LAMP stacks
Market trends	Progressive web apps (PWAs), Responsive design, User experience enhancements	Serverless architecture, Microservices, Data security	Mobile web development, UI/UX development, Specialisations in frameworks/technologies

- **Project management:** Handling end-to-end development tasks, collaborating with teams, and managing project timelines can be challenging, requiring effective communication and organisational skills.
- **Debugging and troubleshooting:** Identifying and resolving issues that can arise in both front-end and back-end components of a web application can be complex.

Job opportunities and market trends: The demand for full-stack developers is on the rise as companies seek professionals who can handle end-to-end development tasks. Full-stack developers command competitive salaries, with the average ranging from Rs 1.2 million to Rs 2 million per annum. With the rise of new technologies and frameworks, full-stack developers must stay updated to remain relevant in the rapidly evolving web development landscape.

Learning curve: Here is what developers must try and master when they dive into full-stack web development.

- Acquiring comprehensive knowledge of front-end and back-end technologies and frameworks
- Understanding the interplay between client-side and server-side processes
- Developing problem-solving skills to tackle challenges across the entire development stack

The future outlook

As businesses increasingly rely on web applications, the demand for skilled web developers will continue to grow. By understanding the different types of web development and the job opportunities they offer, aspiring web developers can make informed decisions about their career paths.

Front-end, back-end, and full-stack web development each have their own unique characteristics and job opportunities. Whether you're passionate about creating captivating user interfaces, handling server-side logic, or bridging the gap between front-end and back-end, there is a niche within web development that suits your skills and interests. As the digital landscape continues to expand, web development remains a promising and rewarding career choice.

Note: The salary and market trends mentioned are based on general industry insights and may vary depending on factors such as location, experience, and market conditions.



By: Meghraj Singh Beniwal

The author is the founder of Schooling for Next Generation, a tech enthusiast and Android hobbyist. He is a test automation analyst working in Virginia in the US.

FOCUS

Data Management

Managing Data on the AWS Platform



AWS data management services offer scalability, flexibility, and strong security measures, making them an excellent solution for organisations wishing to maximise their data's value. You need to right-size them for peak efficiency and avoiding unnecessary costs from overprovisioned resources.

Read more on page.....40

Blockchain-Based Storage Platforms that You Could Use



Blockchain's immutability and cryptographic signatures stand guard against hacking, system failures, and data breaches, fortifying data integrity. These features are invaluable when it comes to data storage. It's no surprise, therefore, that blockchain-based storage platforms are gaining in popularity. *Read more on page.....46*

Role of Open Source in Designing SAN and NAS Systems

Open source SAN and NAS solutions can be an excellent option for companies needing to store vast amounts of data. Although the setup and maintenance may require more technical skill, the resulting cost-efficiency, flexibility, control, and community support can all make it worth the effort.

Read more on page.....35

Why Data Management and Security can Spell Success for Organisations

In an increasingly data-centric world, where information is not just power but also a valuable commodity, data management and security stand as the twin pillars upon which the modern business landscape rests. From multinational corporations to small startups, the ability to effectively manage and secure data has the potential to be a game changer, shaping the future and success of organisations across industries. *Read more on page.....50*

Large Language Models: Helping Manage Data

Generative AI and large language models (LLMs) are the future, and promise a revolution in data management. However, development of LLMs is still very costly and inaccessible to smaller organisations. This will change as the years go by, and AI becomes more commonplace.

Read more on page.....55

Role of Open Source in Designing SAN and NAS Systems

Open source SAN and NAS solutions can be an excellent option for companies needing to store vast amounts of data. Although the setup and maintenance may require more technical skill, the resulting cost-efficiency, flexibility, control, and community support can all make it worth the effort.



possible to deploy them in hybrid environments, storing data both on-premise and in the cloud.

What are SAN and NAS?

SAN and NAS are two types of data storage systems used in a network environment.

A SAN is a high-speed network that connects multiple servers and storage devices over a dedicated network. It allows multiple servers to access the same storage devices simultaneously and provides block-level access to data. SAN is ideal for large-scale enterprises that require high performance, reliability, and scalability. Data is typically stored on arrays of high-speed magnetic disks and accessed via Fibre Channel or iSCSI protocols.

On the other hand, NAS is a file-level storage device that allows data to be stored and accessed over a network. NAS devices are essentially dedicated file servers providing data access to multiple users or applications. They use an IP-based file sharing protocol like NFS or SMB/CIFS to enable client systems to access data as if it were local. NAS systems are typically used for departmental or office-level applications, and are designed for user-level access to files such as documents, multimedia files, and databases.

In summary, SAN is a network designed for block-level access to storage, typically used in large-scale enterprises, while NAS is a network device that allows file-level access to

Open source technologies play a critical role in providing cost-effective and flexible solutions to meet the storage needs of organisations. Open source software enables companies to customise and optimise their storage infrastructure to fit specific requirements and budgets. Its storage solutions offer scalability, high availability, performance, and security.

Storage area networks (SANs) and network attached storage (NAS) systems can be easily integrated with open source software such as Ceph, GlusterFS, and OpenFiler, which offer a wide range of features that are ideal for enterprise storage. Cloud computing and virtualisation have further enabled the adoption of open source storage solutions, making it

storage, typically used in smaller-scale departments or offices.

Overview of open source software in SAN design

Open source software has become an increasingly popular option for designing SAN systems in recent years, as it enables individuals and organisations to take advantage of the community of developers who contribute to various software tools and frameworks.

Open source SAN design offers several benefits, including cost savings, flexibility, and the ability to customise tools and platforms to meet specific needs. Additionally, open source software often comes with strong reliability and security features, as developers are usually committed to ensuring the products they build are trustworthy and dependable.

There are many established and emerging open source software options for SAN design. Some of the most widely used platforms include ZFS, Ceph, DRBD, and GlusterFS. Each of these offers unique benefits, ranging from optimal storage utilisation to efficient data protection and replication.

Overall, open source software has become an integral part of the SAN design landscape, offering users a range of powerful features and resources for building robust and reliable storage networks. As the open source community continues to grow and evolve, we can expect even more exciting developments and innovations in the world of SAN design.

Advantages of open source software in SAN design

Cost-effectiveness: One of the most significant advantages of open source software in SAN design is its cost-effectiveness. The use of open source software is free and does not require any licensing fees, making it a great option for businesses with limited budgets.

Flexibility: Open source software is highly customisable, allowing it to be tailored to meet the specific needs of a business and SAN design. This flexibility permits businesses to add or modify features as necessary without worrying about the limitations of proprietary software.

Community support: The open source software community is vast and active. This means that businesses can benefit from peer feedback and support for issues that may arise.

Continuous development: Open source software is continually updated and improved upon, ensuring that businesses can benefit from continued development and bug fixes.

Disadvantages of open source software in SAN design

Security: The use of open source software requires the sharing of source code, raising potential security concerns as businesses may be hesitant to share their proprietary information.

Lack of vendor support: While the open source community can be a great resource, businesses may face a lack of dedicated vendor support when using open source software.

Limited compatibility with proprietary software: Open source software may not be fully compatible with proprietary software used by a business, which can lead to compatibility issues.

Potential for instability: Open source software may not be as reliable or stable as proprietary software, which can lead to downtime and increased risks.

A few open source software for SAN design

OpenSAN: A software-defined storage platform that allows users to create and manage their own storage infrastructure.

Ceph: A distributed storage platform providing unified object, block, and file storage. It is designed to scale to thousands of nodes and supports multiple storage protocols.

Openfiler: An operating system that provides networked storage capabilities, including iSCSI, NFS, and SMB/CIFS. It supports hardware RAID, encryption, and snapshotting.

Open-E DSS V7: A network storage software platform that supports NAS and iSCSI protocols. It provides high availability, failover, and snapshots for data protection.

GlusterFS: An open source, distributed storage system allowing users to create virtual disk resources that can be accessed over a network.

NexentaStor: A software-defined storage solution supporting NAS, iSCSI, and Fibre Channel protocols. It provides replication, snapshots, compression, and deduplication for data protection and efficiency.

LIO Target: A software package offering a framework for building iSCSI targets on Linux. It supports multiple initiators, authentication, and security features.

Overview of open source software in NAS design

Open source software has become increasingly popular in NAS design. NAS systems are used to store and share data between multiple devices over a network, and open source software provides an economical and versatile option for creating and maintaining these systems.

The collaborative nature of open source development also helps. A multitude of users and developers contribute to the development and improvement of the software, enhancing its quality and reliability. Consequently, NAS devices featuring open source software are often highly stable and secure.

Advantages of open source software in NAS design

Cost-effectiveness: Open source software is generally free of cost or has minimal licensing fees, making it more affordable for budget-conscious organisations.

Customisation: Users can tailor the software to meet specific business needs.

Flexibility: Open source software provides users with greater flexibility in terms of incorporating new features and making modifications.

Wide range of available applications: The open source community has developed a wide variety of applications, which can be easily integrated into a NAS system.

Security: Open source software is subject to scrutiny from a global community, leading to a more secure system.

Disadvantages of open source software in NAS design

Lack of support: Open source software often lacks dedicated technical support, requiring users to rely on forums and online communities for technical assistance.

Complexity: It can be more complex and harder to install and configure than proprietary software.

Compatibility issues: It may not be compatible with certain hardware or software, hampering system performance.

Quality control: Open source software is not always checked for bugs or issues before its release, potentially leading to system instability.

Reliability: It may not have the same level of reliability as proprietary software, potentially affecting business operations.

A few open source software for NAS design

There are several open source software options for NAS design.

OpenMediaVault: OpenMediaVault is a simple and user-friendly NAS operating system designed for users seeking a complete and ready-to-install NAS solution. It comes with several features such as support for RAID,

remote access, and user management, and is built on Debian.

FreeNAS: FreeNAS is also popular open source software for NAS design, known for its flexibility, scalability, and reliability. It includes advanced functions such as support for ZFS, virtualisation, and plugins for cloud backup and multimedia streaming.

Openfiler: Openfiler is a Linux-based NAS/SAN platform that offers several file systems and protocols, including SMB/CIFS, NFS, HTTP/HTTPS, and FTP. It also provides features for multiple disks, LUNs, and RAID.

NAS4Free: NAS4Free is lightweight NAS software that is designed for small businesses and home users. It offers support for various file systems such as ZFS, UFS, and EXT.

Rockstor: Rockstor is open source NAS software that comes with several features such as support for Docker plugins, Btrfs file system, and snapshots. It can be installed on any hardware platform and offers smooth integration with cloud services.

Trends in the industry

Here are some general trends in and characteristics of SAN and NAS design.

SAN trends

Convergence: The convergence of SANs with Ethernet-based networks has been happening for some time now. It allows for greater flexibility and cost savings.

Flash storage: Flash storage's speed and efficiency make it ideal for SANs. As the cost of flash storage continues to decline, it is becoming more popular for use in SANs.

Software-defined storage: Software-defined storage (SDS) is gaining popularity for SANs. SDS allows for greater flexibility and scalability by separating the control plane from the data plane.

NAS trends

Hybrid cloud: Many organisations are adopting a hybrid cloud approach to their NAS design. This approach allows them to leverage both on-premise and cloud-based storage for greater flexibility.

Increased security: With the growing volume of data stored on NAS devices, security becomes a greater concern. Many NAS devices now include advanced security features such as encryption and access controls.

Integration with backup and disaster recovery: NAS devices are often used for backup and disaster recovery purposes. Many NAS devices now include built-in backup and disaster recovery tools, making these tasks easier and more efficient.

Forecasting the future

Forecasting the future of SAN and NAS design involves predicting how these technologies will evolve and adapt to meet changing storage demands and emerging technologies. Some potential trends that could impact SAN and NAS design include:

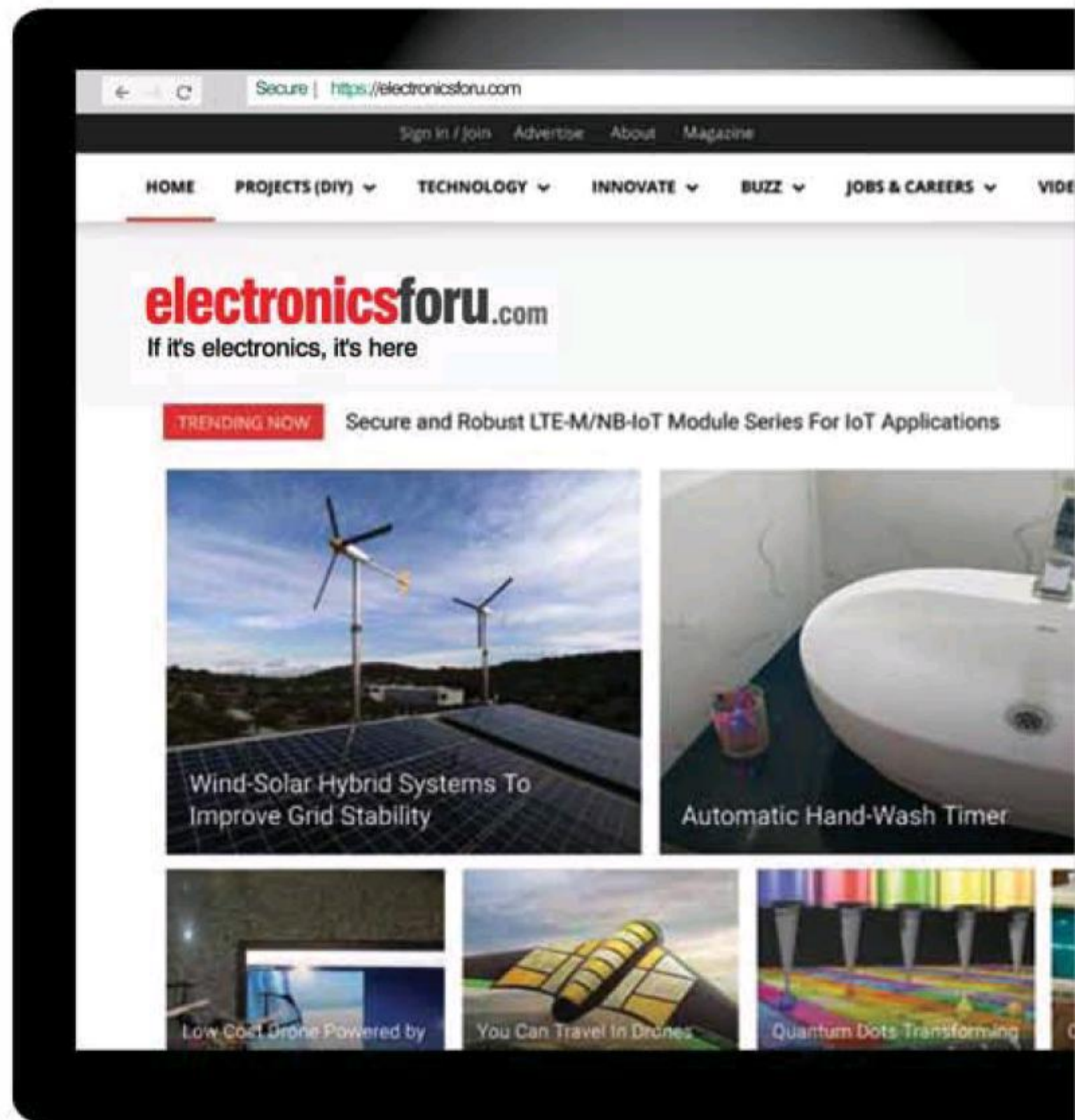
Data growth and the need for scalable storage solutions: As more data is generated by businesses and consumers, SAN and NAS storage solutions will need to be scalable to accommodate the increasing storage demands.

Rise of cloud-based storage: Cloud-based storage solutions offer a flexible and cost-effective option for storing data, which may impact the demand for on-premise SAN and NAS storage solutions.

Proliferation of IoT (Internet of Things) devices: As more devices become connected to the internet, SAN and NAS solutions may need to adapt to efficiently manage the large volumes of data generated by these devices.

Continued to page... 45

Your favourite website has



electronicsforu.com

THANKS TO YOU—OUR ONLINE NETWORK IS

FACTS & FIGURES

- 4 websites (two more coming soon)
- Five major Facebook communities
- Seven major LinkedIn groups & pages
- Million-plus active users (monthly)
- Million-plus reach through Facebook
- Fifty-thousand-plus industry connections through LinkedIn

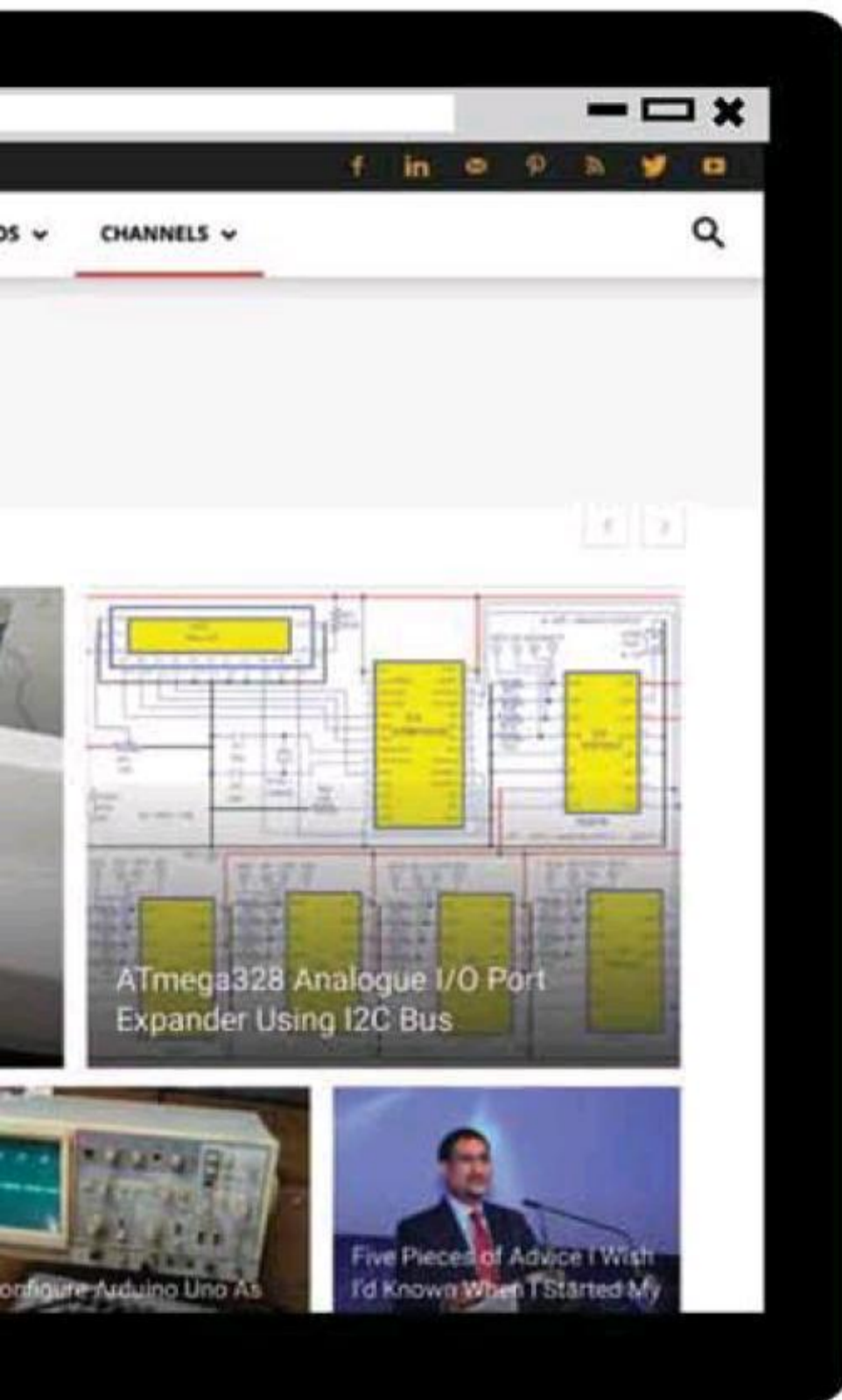
READERS

- You can access all content for FREE
- You can subscribe to newsletters for FREE --on most websites
- Register on our websites to get free invites to technical webinars and seminars

EXPERTS

- Experts who want to share their knowledge through articles, DIY Projects, etc are welcome
- We also welcome experts who want to share their knowledge through webinars or seminars
- You can contact us at editop@efy.in

fast growing peers now...



Amazing DIY Projects. Latest Tech trends.
The hang-out for electronics enthusiasts.



The Latest in IOT.
A platform for enablers, creators and providers of IOT solutions.



India. Electronics. Directory.
Enabling commerce between buyers & sellers of electronics in India.



Business. Electronics. India.
Everything you want to know about India's electronics industry.

AMONGST THE WORLD'S TOP 5 AND GROWING!

INDUSTRY

- You can advertise for as little as US\$ 100 per month
- Special combo offers for advertisers in our print publications
- We've now enabled flexible CPM-based advertising
- You can advertise on the platform of your choice (based on your target audience)
- We invite press releases at efy-edit-team@efy.in
- Press releases are published free of cost, subject to discretion of the editorial team

RESPONSE GUARANTEED SOLUTIONS

We now also act as marketing partners for our clients and drive entire marketing for them, where we charge them on basis of results and not efforts!

CONTACT US: Shrikant Rao • growmybiz@efy.in • +91-98111 55335

Managing Data on the AWS Platform

AWS data management services offer scalability, flexibility, and strong security measures, making them an excellent solution for organisations wishing to maximise their data's value. You need to right-size them for peak efficiency and avoiding unnecessary costs from overprovisioned resources.



Data is the lifeblood of the modern business world, and good data management is crucial for organisations looking to obtain insights, make educated decisions, and keep a competitive advantage. Amazon Web Services (AWS) provides a diverse set of data management solutions across various use cases, including data storage, processing, analytics, and security.

In the development of cloud native architecture, backend storage plays a critical role and employing a cloud native database has many advantages

like cost efficiency, agility and seamless migration from legacy platforms.

To harness the power of their data, organisations must carefully choose and efficiently combine AWS data management services. AWS's scalability, flexibility, and strong security measures make it an excellent solution for organisations of all sizes and industries wishing to maximise their data's value.

Data migration to the AWS platform

A significant factor for achieving better total cost of ownership (TCO) in

cloud adoption involves transitioning from commercial OLTP databases to native offerings like RDS (Relational Database Service) or Aurora in AWS. This transition offers cost benefits and facilitates integration with native services.

Data migration from one source type to another is a complex task, contingent upon factors such as database complexity, data volume, compatibility between source and target (data types, data objects), and more.

AWS provides native migration tools like AWS Schema Conversion

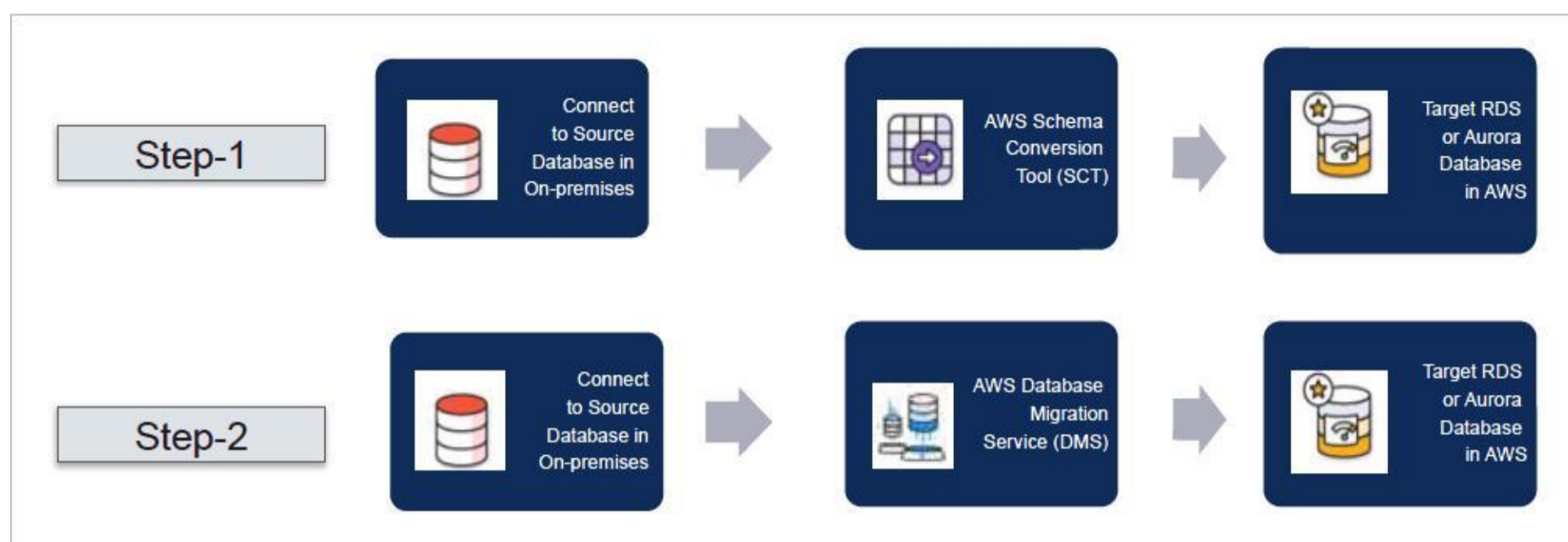


Figure 1: Steps in AWS native database migration

Tool (SCT) for migrating data objects and structures from source to target, encompassing tables, indexes, data constraints, functions, and procedures, among others. Once completed, the target database in AWS will have the like-to-like structure of the source, enabling the use of AWS data migration service (DMS) tools. Figure 1 shows these steps in order.

DMS ensures high reliability and recovery of data migration, supporting scheduled migrations to migrate data during weekends, minimising live data traffic in databases from running applications. Also, using S3 bucket for storing buffered data expedites migration to AWS database, and DirectConnect facilitates connection and transfer of data from on-premise to the cloud. For higher data security, DMS encrypts data in transit during migration ensuring data integrity.

Architectural decisions on choosing the right database solution for the target architecture hinge on various factors, including functional and non-functional requirements (NFR). Table 1 lists various AWS database services and their potential use cases aiding in decision-making for selecting the right database solutions.

Right-sizing database solutions in AWS

Right-sizing is the process of

matching the database resources with your actual workload requirements. This involves analysing your database usage patterns, such as CPU, memory, and I/O, and then selecting the instance type and size that offers the best performance for your workload at the lowest possible cost.

Right-sizing your database solutions in AWS is an essential step in optimising your cloud costs and ensuring that your databases are at peak performance. By selecting the right instance type and size for your workload, you can avoid both overprovisioning and under-provisioning, mitigating unnecessary costs and potential performance bottlenecks.

Right-sizing is important for several reasons.

Cost optimisation: By right-sizing your databases, you avoid paying for unused resources, resulting in significant cost savings on your AWS bill.

Performance optimisation: Under-provisioned databases may struggle to keep up with workloads, leading to performance bottlenecks. Conversely, overprovisioned databases incur costs for unused resources.

Compliance: Certain regulatory requirements mandate that your databases be right-sized. For example, the HIPAA regulations require that healthcare organisations right-size their databases to protect patient data.

How to right-size your databases in AWS

There are several steps you can take to right-size your databases in AWS.

Analyse your database usage patterns: Utilise AWS CloudWatch or other monitoring tools to analyse CPU, memory, I/O, and other key metrics.

Identify your workload requirements: Once you have a good understanding of your database usage patterns, you need to identify your workload requirements. This includes understanding the number of concurrent users, the type of queries that are being run, and the amount of data being stored.

Select the right instance type and size: AWS offers various instance types and sizes. The best instance type and size for your workload will depend on your specific requirements. Use the AWS RDS instance advisor tool to guide your selection of the right instance type and size.

Monitor and adjust: Continuously monitor database performance after right-sizing it. This will help you to identify performance bottlenecks and make necessary adjustments to instance type and size.

Creating testable solutions with AWS Data Lab

Throughout the cloud adoption journey, challenges arise in database

Database service	Features/Design requirements	Potential use cases
Amazon RDS	<ul style="list-style-type: none"> • Managed relational databases and support for various database engines (MySQL, PostgreSQL, etc) • Automated backups and scaling • High availability • Security and compliance 	Web applications, e-commerce, content management
Amazon DynamoDB for RDBMS	<ul style="list-style-type: none"> • Managed NoSQL database • Seamless scalability • High performance • Low latency access • Automatic sharding • Encryption at rest and in transit • Global tables for multi region replication 	Mobile and web applications, gaming, IoT
Amazon Redshift for data warehouse	<ul style="list-style-type: none"> • Data warehousing • High performance analytics • Columnar storage • Massively parallel processing (MPP) • Integration with data visualisation tools • Security and compliance 	Data analytics, business intelligence
AWS Kinesis data streams for Big Data processing	<ul style="list-style-type: none"> • Scalable Big Data solutions • Suitable for data lake architecture 	Data analytics and data predictions
Amazon Aurora for large scale OLTP applications	<ul style="list-style-type: none"> • High performance, MySQL, and PostgreSQL compatible relational database • Automated backups and scaling • Low latency read replicas • Data replication across multiple availability zones • Security and compliance 	High performance applications, e-commerce, SaaS
Amazon Neptune for graphical data store	<ul style="list-style-type: none"> • Managed graph database • Fully managed and highly available • Supports property graph and RDF graph models • SPARQL and Gremlin query languages • Encryption at rest and in transit • Security and compliance 	Social networks, recommendation systems, knowledge graphs
Amazon DocumentDB for document/content managed store (DMS/CMS)	<ul style="list-style-type: none"> • Managed NoSQL database compatible with MongoDB • Fully managed and scalable • High availability with automatic failover • Encryption at rest and in transit • Security and compliance 	Content management systems, catalogues, user profiles
Amazon Timestream (time series data store)	<ul style="list-style-type: none"> • Managed time series database • Scalable and serverless • Storage and analytics for time series data • Data retention policies • Data life cycle management • Security and compliance 	IoT applications, industrial monitoring, log data
Amazon Quantum Ledger DB (QLDB)	<ul style="list-style-type: none"> • Managed ledger database • Fully managed and scalable • Immutable, transparent, and cryptographically verifiable journal • Support for SQL-like queries • Security and compliance 	Supply chain, financial services, audit trails

Table 1: AWS database services and their potential use cases

migration, data analytics, data lake solution development, and creating real-time data pipelines with machine learning solutions. Addressing these challenges requires expertise and highly skilled resources to manage risks and mitigate them in real-time.

AWS provides engineering support through AWS Data Lab, facilitating accelerated data engineering activities through skilled AWS technical resources and lab facilities. AWS Data Lab is like a black box setup for data engineering activity to develop solutions; it comprises two categories — Build Lab and Design Lab.

Design Lab offers 1.5 to 2 days of engagement with the AWS engineering and technical team to build a solution design for data engineering projects for customers. Build Lab offers 2-5 days of intense workshop engagement with the AWS technical team to develop a working trailblazer for the full-blown architecture/solution design. There are five AWS Data Lab hubs, located in Seattle, New York, Herndon, London and Bengaluru. You can engage with these labs online or through in-person visits.

Using AWS Data Lab, you can develop solutions using AWS Quick Sight, build machine learning solutions for data analytics using SageMaker, develop data lake formation templates, and carry out real-time streaming using AWS EMR for Apache Spark. AWS Data Lab works on three principles:

- Think big to develop a robust data engineering solution
- Start small to create a prototype solution
- Scale fast to ensure all the best practices advised by the AWS technical team are incorporated in the solution design

In-memory data storage services

For Big Data processing like data aggregation or data analytics, we may need high-speed data operations like querying, updating and storing data to physical data stores. With disk-based services, this cannot be achieved to the high speed that we need, as they are physical IO operations. Modern techniques introduced later with first level and second level caching, and then data storage with solid-state devices (SSD) gave better transaction rates and fault-tolerant read-write operations.

However, these could not solve the problem and hence modern cloud architecture introduced in-memory data stores like Memcached, which can store entire databases in memory and does not need any physical IO operation. The ultra-speed design includes an in-memory data store with persistent state using a framework like Redis to handle main memory data for backend synchronisation.

Amazon MemoryDB for Redis is such an in-memory datastore with persistent facility, and is now generally available (GA) across multiple regions. It doesn't use any disk IO for non-change (read-only) operations, and uses a backend synchronised data store for data change operations like update/add/delete. This design ensures reliable, real-time synchronisation.

AWS Data Pipeline

AWS Data Pipeline is a fully managed service designed to automate the movement and transformation of data between AWS compute and storage services, as well as on-premise data sources. It helps you build reliable, managed data pipelines to collect, transform, and move your data at scale.

These are the key characteristics of this service.

Orchestrated data movement: AWS Data Pipeline allows the orchestration of data movement between different data stores.

Data transformation: It allows you to transform data using a variety of pre-built AWS services, such as Amazon EMR and Amazon Redshift.

Dependency management: It also allows you to define dependencies between tasks, so that tasks can only run after their dependencies have completed successfully.

Error handling: AWS Data Pipeline provides several features for handling errors, such as retries and notifications.

Monitoring and alerting: It provides monitoring and alerting capabilities that allow you to track the status of your pipelines and receive notifications when there are problems.

AWS Data Pipeline works by orchestrating the movement and

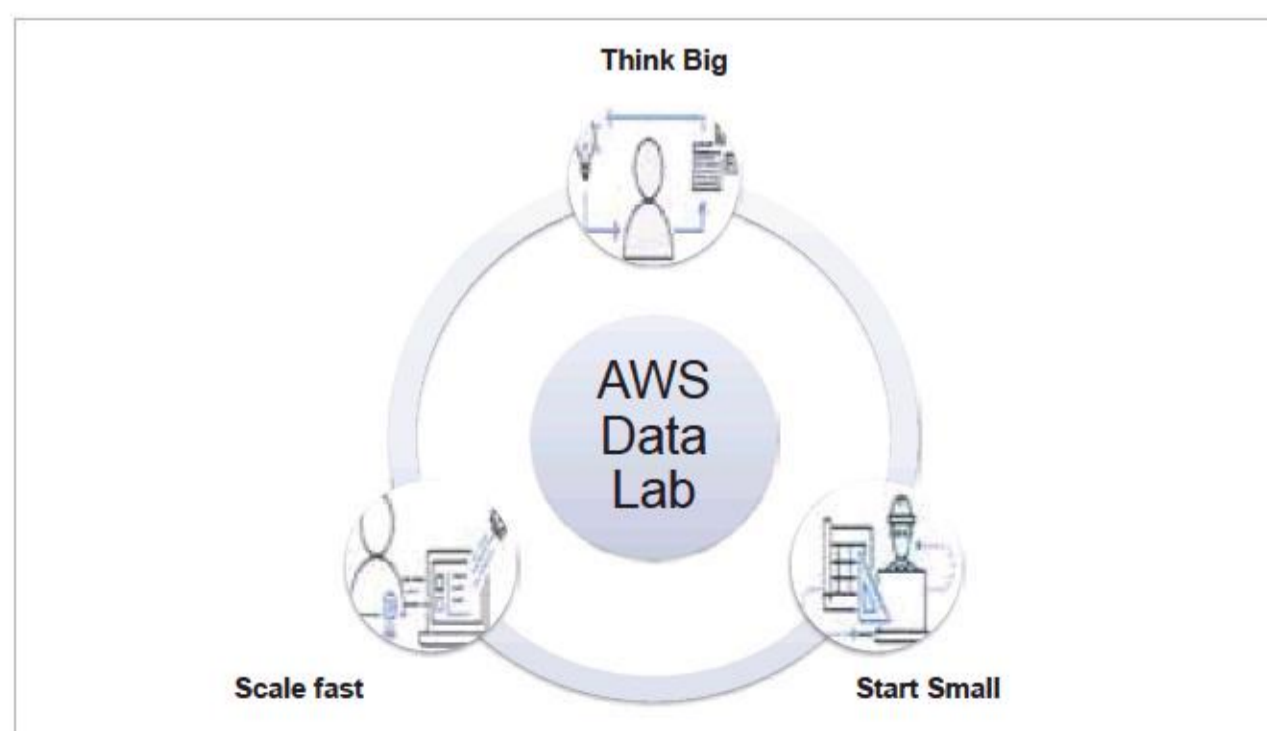


Figure 2: AWS Data Lab

transformation of data between different AWS compute and storage services, as well as on-premise data sources. It does this by defining a pipeline, which is a collection of tasks that are scheduled to run at specific times or when certain conditions are met.

Here is a step-by-step overview of how AWS Data Pipeline works.

Create a pipeline: A pipeline definition is created by specifying the data sources, data destinations, and transformation tasks that you want to run. You can use the AWS Data Pipeline console, the SDK, or the CLI to create a pipeline.

Schedule the pipeline: Schedule the pipeline to run on a regular basis or when certain conditions are met. You can specify the frequency of the pipeline execution, as well as the start time and end time of each execution.

Data movement: AWS Data Pipeline transfers data between the data sources and destinations using the appropriate AWS services. For example, it can use Amazon S3 to transfer data between data lakes or Amazon EMR to transform data in batch.

Data transformation: It can transform data using a variety of pre-built AWS services, such as Amazon EMR and Amazon Redshift. You can use these services to clean, filter, and aggregate data before loading it into the destination.

Error handling: AWS Data Pipeline automatically retries failed tasks and sends notifications when there are errors. This ensures that your pipelines are resilient to failures and that data processing is not interrupted.

Monitoring and alerting: It provides monitoring and alerting capabilities that allow you to track the status of your pipelines and receive notifications when there are problems. This helps to identify and troubleshoot issues quickly.

The key components of the AWS Data Pipeline are:

Agent: The agent is a small piece of software that runs on each worker node in the pipeline. It is responsible for executing tasks and communicating with the data pipeline service.

Catalogue: The catalogue is a database that stores information about data sources, data destinations, and transformation tasks. This information is used by the data pipeline service to schedule and execute pipelines.

Jobs: A job is a collection of tasks that are scheduled to run together. Each job has a start time, an end time, and a list of tasks that need to be executed.

Tasks: A task is a unit of work that is executed by the data pipeline service. Each task has a type, such as Copy, Transform, or HiveJar, and a set of parameters.

There are many benefits of using the AWS Data Pipeline.

- **Reduced operational overhead:** AWS Data Pipeline automates many of the tasks involved in managing data pipelines, such as scheduling, error handling, and monitoring. This can free up your team to focus on more strategic projects.
- **Increased reliability:** This service is designed to be fault-tolerant and can handle transient failures. This means that your data pipelines will continue to run even if there are occasional problems with the underlying infrastructure.
- **Scalability:** AWS Data Pipeline can handle large volumes of data and can scale up or down to meet changing needs. This means that you can easily accommodate spikes in data volume without having to worry about infrastructure capacity.
- **Security:** AWS Data Pipeline uses encryption and access control to protect your data, securing it from unauthorised access and tampering.
- **Cost-effectiveness:** The pipeline is a pay-as-you-go service, so you only pay for the resources that you use. This can help save money on data processing costs.

Disaster recovery (DR) architecture for AWS database services

Disaster recovery (DR) architecture for AWS database services refers to the strategies and infrastructure that are implemented to protect and restore AWS database services in the event of a disaster. A comprehensive DR plan ensures business continuity and minimises data loss by enabling rapid recovery from disruptive events such as natural disasters, hardware failures, or cyberattacks.

The key components of DR architecture for AWS database services are:

Primary database: The primary database resides in the production environment and handles all read-write operations.

Recovery region: A separate AWS region serves as the recovery site, where a replica or standby database is maintained.

Replication technology: Data replication techniques like Amazon RDS replication or Aurora Global Database are used to synchronise data between the primary database and the recovery site.

Failover mechanism: A failover mechanism is established to automatically switch to the recovery database in case of a disaster.

Monitoring and alerting: Monitoring tools are implemented to track the health and performance of the DR infrastructure, and alerts are configured to notify administrators of potential issues.

Common DR strategies for AWS databases are:

- **Backup and restore:** Regularly create backups of the primary database and store them in a remote location, such as an S3 bucket or Glacier vault in a different AWS region. In case of a disaster, restore the most recent backup to a new database instance in the recovery region.


- **Pilot light:** Maintain a read-only replica of the primary database in the recovery region. The replica is kept up-to-date using replication technology, and it can be quickly promoted to a read-write instance if the primary database becomes unavailable.
- **Warm standby:** Maintain a warm standby database in the recovery region. The standby database is kept in a near-ready state, with data refreshed periodically or on-demand. It can be activated as the primary database with minimal downtime in case of a disaster.
- **Multi-site active/active:** Operate active-active production databases in multiple AWS regions. Data is replicated between the regions in real-time, allowing failover to either region with minimal disruption.

The choice of DR strategy depends on factors such as recovery time objective (RTO), recovery point objective (RPO), cost, and complexity.

RTO: This is the time acceptable for downtime following a disaster. Backup and restore strategies may have longer RTOs compared to pilot light or warm standby.

RPO: This is the acceptable amount of data loss following a disaster. Warm standby and active/active strategies minimise data loss compared to backup and restore.

Cost: Backup and restore is the most cost-effective DR strategy, while active/active is the most expensive. Pilot light and warm standby offer a balance between cost and performance.

Complexity: Backup and restore is the simplest, while active/active is the most complex to implement and manage. **END** 

References

- <https://aws.amazon.com/memorydb/>
- <https://aws.amazon.com/big-data/featured-partner-solutions-data-governance-compliance/>
- <https://aws.amazon.com/data-exchange/providers/>
- <https://aws.amazon.com/aws-cost-management/aws-cost-optimization/right-sizing/>
- <https://aws.amazon.com/datapipeline/>
- <https://aws.amazon.com/athena/>
- <https://aws.amazon.com/blogs/architecture/disaster-recovery-dr-architecture-on-aws-part-i-strategies-for-recovery-in-the-cloud/>

By: Dr Magesh Kasthuri and Dr Anand Nayyar

Dr Magesh Kasthuri is a senior distinguished member of the technical staff and principal consultant at Wipro Ltd. This article expresses his views and not that of Wipro.

Dr Anand Nayyar is a PhD in wireless sensor networks and swarm intelligence. He works at Duy Tan University, Vietnam, and loves to explore open source technologies, IoT, cloud computing, deep learning and cyber security.

Continued from page...37


Advances in flash storage: The continued development of flash storage technology is likely to impact SAN and NAS design, as flash storage can offer significantly faster data access speeds compared to traditional storage solutions.

Overall, forecasting the future of SAN and NAS design involves considering factors such as data growth, emerging technologies, and shifting business needs to predict how storage solutions will need to evolve to remain effective.

Final thoughts

Here are some final thoughts and recommendations for open source

SAN and NAS design.

- Open source SAN and NAS solutions provide businesses with cost-effective and flexible storage options. They offer a significant cost advantage over proprietary solutions and can be customised to meet unique business needs.
- These solutions are fully transparent, giving users complete control over configurations, updates, and upgrades.
- Community support is essential for open source SAN and NAS solutions. A strong community can help refine the design and functionality of these solutions, which ultimately benefits everyone.
- Security is of utmost importance in SAN and NAS design. User and data safety must be maintained, and open source solutions must keep pace with existing security threats. **END** 

By: Dr S. Balakrishnan

The author heads the Department of Computer Science and Business Systems at Sri Krishna College of Engineering and Technology, Coimbatore, Tamil Nadu.

Blockchain-Based Storage Platforms that You Could Use

Blockchain's immutability and cryptographic signatures stand guard against hacking, system failures, and data breaches, fortifying data integrity. These features are invaluable when it comes to data storage. It's no surprise, therefore, that blockchain-based storage platforms are gaining in popularity.

Data is a crucial asset today, and ensuring its security and accessibility is of paramount importance. Traditional centralised storage solutions have long been the norm, but they come with vulnerabilities such as data breaches, single points of failure, and susceptibility to hacking.

The classical centralised storage systems, often governed by a single authority or entity, hold massive repositories of data within specific servers or data centres. Despite their efficiency and ease of management, these systems present a significant risk – a single point of failure. If breached, manipulated, or disrupted, the entire system could crumble, leading to catastrophic data loss or unauthorised access. This has brought about a quest for alternative solutions that can address these vulnerabilities effectively.

Blockchain technology was initially popularised as the technological backbone of cryptocurrencies like Bitcoin. Its inherent characteristics – decentralisation, immutability, and cryptographic security – make it a promising candidate for revolutionising data storage and backup practices.

The decentralised nature of blockchain means that data isn't stored in a single location or controlled by a sole authority. Instead, it is distributed across a vast network of nodes, creating a decentralised system that eliminates the risk associated with a central point of failure. Each piece of data is cryptographically secure, time-stamped, and linked together, forming an immutable chain of blocks. This ensures that the stored information remains tamper-proof and highly resistant to unauthorised changes or breaches.

The integration of blockchain technology into decentralised storage solutions offers a promising alternative to centralised repositories, promising enhanced security, resilience, and accessibility. It signifies a paradigmatic change



Image Source: <https://www.freepik.com>

in how data can be secured and managed. By spreading data across a network of nodes, utilising advanced encryption methods, and employing consensus protocols, blockchain offers a compelling solution to the inherent vulnerabilities of traditional centralised storage.

Blockchain technology and use cases

The hallmark of blockchain lies in its ability to record and store data in a secure and immutable manner. This is achieved through the chaining of individual blocks of data in a chronological sequence, where each block contains a cryptographic hash of the previous one. This linking structure ensures that any attempt to alter past data would necessitate changing the entire subsequent chain, making it practically impossible to tamper with the information without consensus across the network.

The feature that distinguishes blockchain from traditional databases is its tamper-proof nature. The distributed nature of the ledger ensures that no single entity can unilaterally manipulate the information stored within it. Consensus mechanisms, such as proof of work (PoW) or proof of stake (PoS), validate and agree upon the accuracy of transactions and data, further enhancing the system's security and reliability.

The application of blockchain extends well beyond its initial use in the realm of financial transactions. Various industries are actively exploring and implementing its

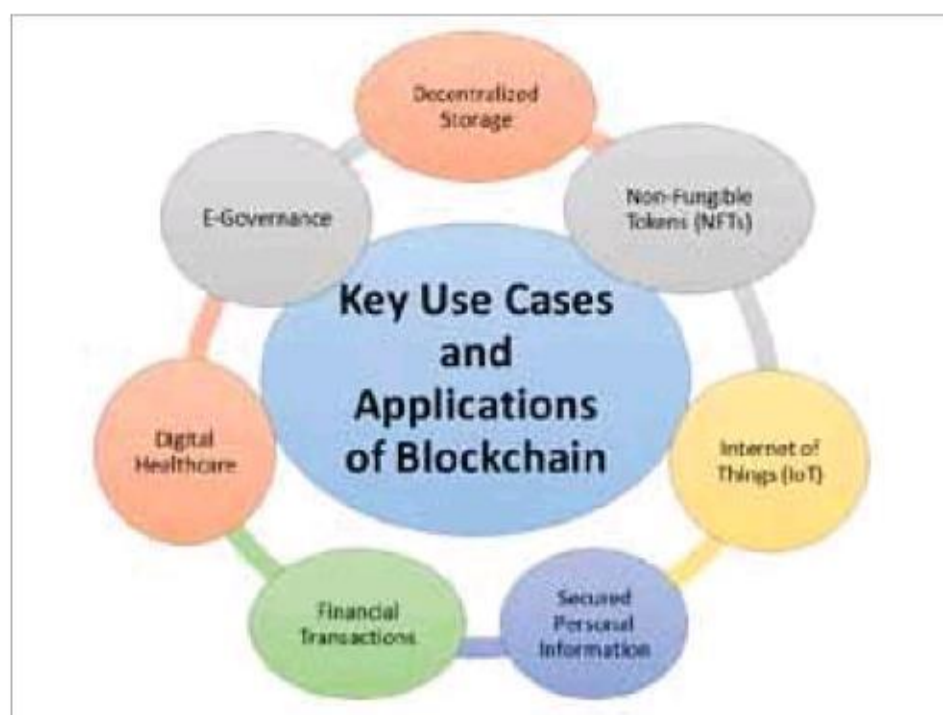


Figure 1: Key use cases and applications of blockchain

potential in different sectors. In supply chain management, for instance, blockchain's transparency and immutability have been harnessed to track and verify the origins and movements of products, thereby enhancing trust and authenticity within supply chains.

Similarly, in healthcare, blockchain technology has been considered for securely managing patient data, ensuring privacy and integrity while enabling authorised access by healthcare providers. In the real estate sector, blockchain's

ability to provide transparent and immutable records of property ownership and transactions could revolutionise the industry, reducing fraud and streamlining the cumbersome processes involved in property transactions.

Crucially, the sphere where blockchain showcases immense promise is in data storage and backup. As industries increasingly grapple with the vulnerabilities of centralised data storage, blockchain's decentralised nature and robust security features offer a compelling solution. The distributed and immutable storage model of blockchain reduces the risk of data breaches, offers enhanced resilience against cyber threats, and provides a more secure and reliable framework for data backup.

Blockchain for decentralised storage

Decentralised storage, powered by blockchain technology, introduces a transformative approach to handling data. This involves breaking down data into smaller, encrypted segments, a process known as sharding. These encrypted fragments are then distributed across a network of nodes in a decentralised manner. This fragmentation and distribution not only fortify the security of the stored data but also augment its accessibility and reliability.

Also, the redundancy created by distributing data across multiple nodes in a decentralised blockchain network ensures that even if one or several nodes fail or are compromised, the data remains accessible and intact due to its availability across other nodes.

Figure 2: IPFS decentralised platform in blockchain

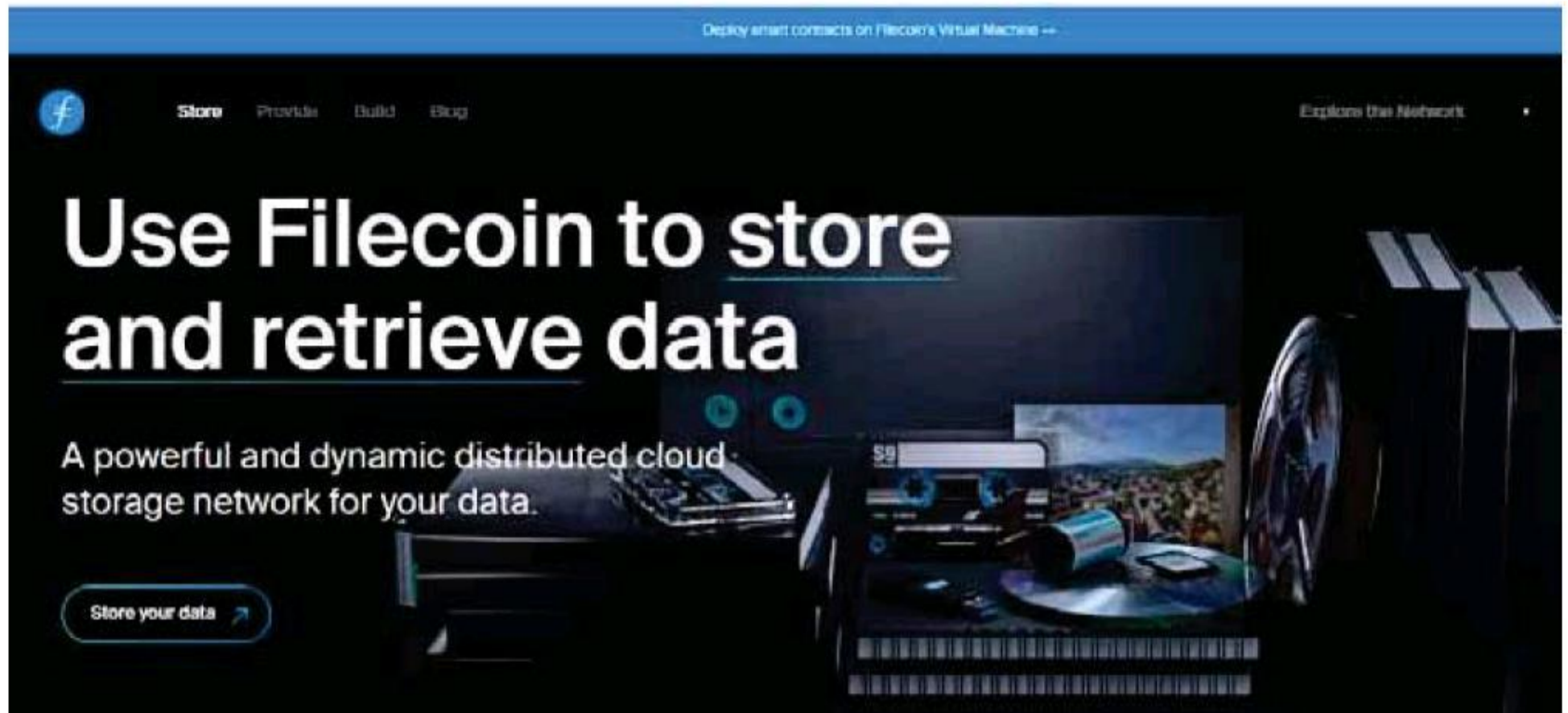


Figure 3: Filecoin platform in blockchain

One of the core pillars of blockchain technology is its immutability – the quality of being unchanging over time. Blockchain maintains the integrity and security of stored data through its immutable nature. Each block within the chain contains a cryptographic hash, a unique digital signature derived from the data in the block, including a reference to the previous block's hash. This chaining mechanism creates a chronological sequence, ensuring that any attempt to alter or tamper with data within a block would necessitate altering subsequent blocks in the chain, thereby alerting the entire network to the unauthorised change. This feature makes the blockchain extremely resilient to unauthorised modifications, making data tampering exceedingly difficult.

The immutability of the blockchain ensures that once data is stored and confirmed within the chain, it remains secure and unaltered. Any modifications or tampering attempts become instantly visible across the network, making the system robust against fraudulent or unauthorised activities.

Working with blockchain-based storage platforms

Several blockchain-based storage platforms now provide innovative, decentralised and secured backup solutions. For instance, Filecoin allows users to rent out spare storage space in exchange for Filecoin tokens. This incentivises individuals and organisations to contribute their unused storage and create a decentralised storage network. Another example is Storj, which operates by breaking down files into smaller encrypted pieces and distributing them across its network of nodes. Users can rent out their unused storage space and earn cryptocurrency in return, creating a decentralised and efficient storage system.

These platforms utilise blockchain technology to create a trustless and secure environment for data storage, ensuring privacy, integrity, and accessibility.

IPFS (InterPlanetary File System) (<https://ipfs.io/>): IPFS is a peer-to-peer hypermedia protocol designed to create a distributed and decentralised method for storing and accessing files. It's not exactly a blockchain but is often used in conjunction with blockchain technology. IPFS is engineered to establish a distributed and decentralised system for file storage and retrieval.

IPFS's core concept is content-addressing, which means files are identified and retrieved based on their unique content, rather than using location-based URLs. At its core, IPFS functions by breaking down files into smaller chunks, and each chunk is assigned a distinct cryptographic hash. This hash acts as a unique identifier for that particular piece of data. Rather than traditional hierarchical file systems, IPFS utilises a Merkle DAG (directed acyclic graph) data structure to link these chunks together, allowing for efficient tracking and retrieval of content.

A key advantage of IPFS lies in its decentralised nature. Instead of relying on a central server, files on IPFS are stored across a network of nodes. These nodes collaborate to store and distribute the chunks of files. When a user requests a specific file, IPFS searches the network for the nodes storing the relevant chunks and retrieves the data, reassembling it on the user's end. One of the critical advantages of IPFS is its resilience against data loss and censorship. Due to the distributed nature of the system, files are replicated across multiple nodes, reducing the risk of losing data if a single node fails. Furthermore, IPFS fosters an environment where information remains available even if certain parts of the network go offline or if there are disruptions.

IPFS can be used for a wide array of applications, including website hosting, distributing software updates, sharing large data sets, and more. It offers a more efficient and potentially faster method of content delivery by utilising a peer-to-peer network, enabling users to access and share information in a decentralised, secure, and reliable manner.

Although not a blockchain itself, IPFS often synergises with blockchain technology to complement and enhance decentralised applications, including data storage, authentication, and decentralised web services. It represents a significant step forward in reshaping how data is stored, accessed, and shared on the internet, offering a decentralised and secure alternative to traditional web protocols.

Filecoin (<https://filecoin.io/>): Filecoin operates on the principles of blockchain and incentivises users to contribute their excess storage space to create a large-scale, distributed storage system. Users can earn Filecoin tokens by contributing storage space, creating a decentralised and efficient storage network. It was developed to address the growing need for secure, decentralised, and efficient storage solutions.

The network enables users to both store and retrieve data in a decentralised manner through a marketplace where unused storage space can be rented out in exchange for Filecoin tokens. These tokens serve as a form of payment within the ecosystem and can also be traded on various cryptocurrency exchanges. This creates a peer-to-peer economy where individuals or organisations with spare storage capacity can offer it to those in need of storing their data.

The storage system divides files into smaller pieces, which are encrypted, duplicated, and distributed across a network of storage providers. This process enhances data security and resilience, ensuring that the information remains available and intact even if some nodes in the network fail or go offline.

Filecoin implements a proof system called 'Proof of Replication' to ensure that the storage providers are storing the data as promised. This means that storage providers must prove that they are storing the exact copy of the data they were hired to store. This mechanism helps to maintain data integrity and trust within the network. Moreover, the Filecoin network is designed to be self-sustaining and decentralised. As more users contribute storage space, the network's capacity and resilience increase, making it more attractive for those seeking reliable and distributed storage solutions.

Storj (<https://storj.io/>): Storj is a decentralised cloud storage platform that employs blockchain to secure and distribute files across a network. Users can rent out their unused storage space and earn cryptocurrency in return.

Sia (<https://sia.tech/>): Sia is a decentralised cloud storage platform that utilises blockchain to secure data through a peer-to-peer network. It aims to provide secure and private storage at a lower cost.

Arweave (<https://www.arweave.org/>): Arweave is a permanent and low-cost data storage solution using a blockchain-like structure called the 'blockweave'. It aims to store data indefinitely, ensuring it remains accessible over time.

Bluzelle (<https://bluzelle.com/>): Bluzelle is a decentralised database service that uses blockchain to provide scalable and reliable data storage. It offers secure and cost-effective data management.

Swarm (<https://ethersphere.github.io/swarm-home/>): Swarm is a decentralised storage and communication system that operates as a base layer for Ethereum's web3 stack. It provides decentralised and redundant storage for various dApps.

NKN (New Kind of Network) (<https://www.nkn.org/>): NKN is a blockchain-powered decentralised data transmission network. It incentivises users to share network connections and spare bandwidth to create a more extensive and secure network.

SpaceChain (<https://spacechain.com/>): SpaceChain aims to combine blockchain technology with satellite infrastructure to create a decentralised and secure data storage network accessible from anywhere in the world.

The potential for blockchain-based tools in decentralised and secured backup is immense, with ample room for further research and development. Advancements in this field could focus on enhancing scalability, improving user experience, and increasing interoperability between different blockchain networks. Additionally, the integration of smart contracts into decentralised storage systems could automate and ensure secure backup processes. Innovations in encryption methods and data retrieval techniques are also areas that researchers are actively exploring to bolster the security and efficiency of blockchain-based storage.

Blockchain technology holds promise for secure, decentralised data backup. As research and development in this field continue to evolve, the potential for safer, more accessible, and efficient storage solutions remains high. With a focus on innovation and overcoming existing challenges, blockchain-based tools for data backup are poised to revolutionise the way we secure and store our valuable information. **END** 🐧

 **By: Dr Gaurav Kumar**

The author is associated with various academic and research institutes for delivering expert lectures and conducting technical workshops on the latest technologies and tools.

Why Data Management and Security can Spell Success for Organisations

In an increasingly data-centric world, where information is not just power but also a valuable commodity, data management and security stand as the twin pillars upon which the modern business landscape rests. From multinational corporations to small startups, the ability to effectively manage and secure data has the potential to be a game changer, shaping the future and success of organisations across industries.



Data management involves acquiring, organising, storing, and maintaining data in a way that makes it accessible, reliable, and usable. It encompasses the strategic planning, administration, and implementation of policies, procedures, and technologies to manage data throughout its lifecycle. Effective data management ensures data quality, integrity, and availability, making it a valuable asset for decision-making and business operations.

Data security is the practice of protecting data from unauthorised access, disclosure, alteration, or destruction. It involves implementing various security measures and technologies to safeguard sensitive information and maintain its confidentiality, integrity, and availability. Data security encompasses encryption, access controls, authentication, and security policies to prevent data breaches, cyberattacks, and data loss, ensuring that data

remains protected from potential threats and vulnerabilities.

Importance of data in the digital age

Data has transcended its status as mere information; it has become the lifeblood of modern business and society. Data fuels innovation, drives personalisation, and empowers organisations to make informed decisions and stay competitive. With the ability to anticipate market trends,

optimise processes, and enhance customer experiences, data is the cornerstone of success in today's interconnected world. Data is the new currency of the digital age, essential for success in this data-driven era.

Data management ensures that organisations have the right data at the right time, facilitating swift and informed decisions. As Lou Gerstner, former IBM CEO, aptly put it, "Data is not just the new oil; it's the new soil for growing businesses." Effective data management fosters a fertile ground for growth and innovation.

On the other hand, data security is the guardian of an organisation's trust and reputation. In the words of the American cryptographer Bruce Schneier, "Data is a liability." It's an asset only when protected. Data security is the shield that safeguards an organisation from reputational damage, financial losses, and legal entanglements.

Together, data management and security enhance operational efficiency, boost customer trust, and ensure compliance with regulatory mandates. They're not just components of a strategy; they're the bedrock upon which organisations build their digital success stories.

Challenges in data management and security

Although valuable, managing and securing data can be a complex endeavour. In this section, we'll examine the key challenges organisations face in data management and security, along with potential solutions.

Data growth and complexity

Challenge: The exponential growth of data, often referred to as Big Data, can overwhelm organisations. Managing, storing, and making sense of this sheer volume and variety of data is a significant challenge.

Solution: Implement data management and analytics tools that

can handle large data sets efficiently. Consider data warehousing solutions, cloud storage, and data governance policies to streamline data organisation. Employ machine learning and AI for data analysis to derive valuable insights from vast data pools.

Data privacy and compliance regulations

Challenge: Data privacy regulations, such as GDPR (General Data Protection Regulation) and HIPAA (Health Insurance Portability and Accountability Act), impose strict requirements on how data is handled, making compliance a significant challenge.

Solution: Create a robust data governance framework that includes policies and procedures for data protection, access control, and encryption. Regularly audit data practices to ensure compliance. Invest in secure data storage solutions and educate employees on data privacy best practices.

Cybersecurity threats and breaches

Challenge: The ever-evolving landscape of cybersecurity threats, including malware, phishing, and ransomware attacks, poses a constant risk to data security.

Solution: Employ a multi-layered cybersecurity strategy that includes firewalls, intrusion detection systems, and regular software patching. Educate employees on recognising and mitigating security threats through cybersecurity training. Implement data encryption and regular security audits to detect vulnerabilities.

Data silos and fragmentation

Challenge: Data silos and fragmentation occur when data is isolated within different departments or systems, making it difficult to access and utilise effectively.

Solution: Establish an integrated data management system that promotes data sharing and collaboration across departments. Implement data

warehouses, data lakes, or master data management solutions to consolidate and centralise data. Develop standardised data formats and metadata to improve data accessibility and consistency.

The game-changing benefits of effective data management and security

Effective data management and security practices safeguard the organisation's most critical asset and offer numerous transformative benefits. Let's delve into these advantages.

Improved decision-making and competitive advantage

The ability to access and analyse data with precision empowers organisations to make data-informed decisions. With comprehensive data at their disposal, they gain insights into customer behaviour, market trends, and operational efficiency. This data-driven decision-making offers a distinct competitive advantage. For example, companies like Google have harnessed data analytics to refine their search algorithms continually, ensuring that users receive the most relevant results. By doing so, they maintain their edge in the highly competitive search engine market.

Improve customer trust and loyalty

Customer trust is invaluable in today's data-driven landscape. Effective data security and privacy practices not only protect sensitive customer information but also foster a sense of trust and reliability. Customers are more likely to engage with businesses that respect their privacy and protect their data. Amazon, as a prime example, has gained customer loyalty by personalising shopping experiences. By leveraging data while ensuring security, businesses enhance the customer's sense of trust and satisfaction.

Compliance and risk mitigation

In an environment where data privacy regulations are continually evolving, adherence to compliance mandates is crucial. Effective data management and security ensure organisations remain compliant with data protection laws. By maintaining compliance, organisations not only mitigate legal risks but also build a reputation for trustworthiness. Healthcare providers, for instance, adhere to stringent regulations like HIPAA to safeguard patient data. Compliance not only avoids legal consequences but also ensures the security and confidentiality of sensitive medical information.

Cost savings and operational efficiency

Efficient data management practices contribute to cost savings by eliminating data redundancy, ensuring data accuracy, and streamlining processes. Data security measures prevent costly data breaches and their associated financial and reputational costs. For instance, companies like IBM have employed data management and security practices to enhance operational efficiency. By reducing data duplication and improving data quality, they save on storage costs and enhance productivity.

Key components of data management and security

Effective data management and security encompass several key components that are essential for safeguarding and optimising an organisation's data. The coding examples given below demonstrate the practical application of key components in data management and security, emphasising the importance of proper implementation for data protection and effective utilisation.

Data collection and storage

Data collection and storage are fundamental to data management. Efficient organisation and storage

ensure accessibility and reliability. For example, consider this Python code for storing data in a text file:

```
# Data Collection
data = "This is sample data to be stored
in a text file."

# Data Storage
with open("data.txt", "w") as file:
    file.write(data)
```

Data quality and accuracy

Maintaining data quality and accuracy is crucial. Python offers libraries like pandas to clean and preprocess data:

```
import pandas as pd

# Load data
data = pd.read_csv("data.csv")

# Clean and filter data
cleaned_data = data.dropna()
```

Data access and sharing

Controlling data access and sharing is vital. Use access control lists (ACLs) in a web application:

```
// JavaScript example
const userAccess = {
  "user1": ["read", "write"],
  "user2": ["read"]
};

// Verify user permissions
function checkPermission(user, action) {
  if (userAccess[user] &&
  userAccess[user].includes(action)) {
    return true;
  }
  return false;
}
```

Data encryption and protection

Data encryption and protection are essential for security. Use the 'cryptography' library in Python to encrypt data:

```
from cryptography.fernet import Fernet
```

```
# Generate a key
key = Fernet.generate_key()
cipher_suite = Fernet(key)
```

```
# Encrypt and decrypt data
data = b"Sensitive data to be encrypted"
encrypted_data = cipher_suite.
encrypt(data)
decrypted_data = cipher_suite.
decrypt(encrypted_data)
```

Data backup and disaster recovery

Creating data backups and disaster recovery plans are critical. Use a simple shell script to automate data backups:

```
# Shell script for data backup
#!/bin/bash

source_dir="/path/to/important/data"
backup_dir="/path/to/backup/location"
timestamp=$(date +"%Y%m%d%H%M%S")

tar -czvf "$backup_dir/
backup_${timestamp}.tar.gz" "$source_dir"
```

Open source tools are revolutionising the way organisations approach data management and security. These versatile solutions offer a cost-effective and flexible alternative to proprietary software, empowering businesses to strengthen their data management capabilities while bolstering their security measures.

Open source tools: Empowering data management and security

Open source tools and technologies empower organisations to efficiently manage and secure data, ensuring better data governance, compliance, and peace of mind. Their active communities contribute to ongoing development, ensuring that they stay up-to-date with evolving security threats and data management requirements. Here are some of the key open source tools and technologies that are making a significant impact in this space.

Table 1: How open source tools enhance data management and security

Cost-effective data management	<p>Open source tools provide cost advantage by being freely available, unlike proprietary software with high licensing fees.</p> <p>They allow organisations to allocate their budgets more efficiently and provide an opportunity for smaller companies and startups to access powerful data management solutions without breaking the bank.</p>
Customisability	<p>Open source tools offer a high degree of customisability. Organisations can tailor the software to their specific needs, creating solutions that fit unique data requirements.</p> <p>With open source tools, you're not bound by the limitations of proprietary software, and you can adapt the tools to your evolving data needs.</p>
Collaboration and community support	<p>Open source tools benefit from extensive user communities that contribute to their development and support.</p> <p>When you use open source tools, you have access to a community of users who can provide advice, solutions to issues, and share best practices for data management and security.</p>
Security through transparency	<p>Open source tools are considered more secure due to their transparency. The source code is open for review, making it difficult for malicious actors to insert hidden vulnerabilities or backdoors.</p> <p>Security experts worldwide scrutinise the code, and any vulnerabilities or weaknesses are quickly identified and patched. This open, transparent development process contributes to the overall security of open source solutions.</p>
Rapid updates and fixes	<p>Open source tools typically have a shorter response time to security vulnerabilities. When a flaw or bug is discovered, the open source community can respond quickly by releasing patches or updates to address the issue.</p> <p>This rapid response is crucial in an era where cyber threats are constantly evolving, and data breaches can have severe consequences.</p>
Integration with other tools	<p>Open source tools are designed to be highly integrative, facilitating their use in conjunction with other software and systems.</p> <p>Their ability to connect and collaborate with various tools is particularly important in data management, as it enables organisations to build comprehensive data ecosystems that cater to their unique requirements.</p>

Apache Hadoop

- **Data management:** Hadoop is a powerful framework for storing and processing large data sets. It provides distributed storage and processing capabilities that help organisations effectively manage their Big Data, ensuring data integrity and accessibility.
- **Security:** Hadoop offers several security features, such as authentication, authorisation, and encryption, to protect sensitive data stored within the platform.

Elasticsearch

- **Data management:** Elasticsearch is a distributed, RESTful search and analytics engine that excels at indexing and searching large volumes of data. It's widely used for log and event data management, making it easier to uncover valuable insights from various data sources.
- **Security:** Elasticsearch has robust security features, including role-based access control and data encryption, to ensure that sensitive information is safe from unauthorised access.

OpenStack

- **Data management:** OpenStack is an open source cloud computing platform that enables organisations to build and manage private and public clouds. It offers storage solutions like Cinder (block storage) and Swift (object storage) for efficient data management.
- **Security:** OpenStack includes security features like identity and access management, firewall as a service, and encryption, providing a secure infrastructure for data storage and processing.

Snort

- **Data management:** Snort is an open source network intrusion detection system and intrusion prevention system. It helps organisations monitor network traffic for suspicious activity and protect against data breaches.
- **Security:** Snort uses real-time traffic analysis and packet logging to identify and mitigate security threats, making it a valuable asset in the data security landscape.

KeePass

- **Data management:** KeePass is an open source password manager that allows users to store and manage their passwords securely. It helps individuals and organisations keep their login credentials safe and easily accessible when needed.
- **Security:** KeePass employs strong encryption and provides features like two-factor authentication, making it a reliable solution for safeguarding critical access credentials.

Best practices for data management and security

A robust data management and security framework safeguards sensitive information, ensures compliance, and fosters a culture of security awareness among employees. Here are some essential tips for achieving robust data management and security.

Establishing data governance policies

- **Define clear data governance policies:** Begin by establishing comprehensive data governance policies that outline data ownership, classification, and usage guidelines. These policies should also address compliance requirements, data retention, and disposal procedures.
- **Assign data stewards:** Designate individuals within your organisation as data stewards responsible for overseeing data governance and compliance. These stewards should

ensure that data is used in a manner consistent with established policies.

- **Data classification:** Implement a data classification system that categorises data based on its sensitivity. This system helps determine the appropriate level of security measures for different types of data.
- **Regular policy review:** Periodically review and update data governance policies to reflect changing regulations and evolving business needs. Ensure that employees are aware of and adhere to these policies.

Training employees on data security

- **Security training programs:** Develop and provide comprehensive training programs on data security for all employees. Training should cover best practices for handling sensitive data, recognising security threats, and following data security protocols.
- **Phishing awareness:** Educate employees on recognising and avoiding phishing attacks, which are common entry points for cyber threats. Conduct simulated phishing exercises to reinforce training.
- **Incident response training:** Train employees on how to respond to security incidents, including data breaches. Establish clear reporting procedures and response protocols to minimise the impact of security breaches.

Implementing encryption and access controls

- **Data encryption:** Implement data encryption solutions to protect data both in transit and at rest. Use encryption protocols like HTTPS

for data in transit and encryption tools for data at rest, such as file-level or database-level encryption.

- **Access controls:** Enforce strong access controls by limiting data access to authorised personnel. Implement role-based access control (RBAC) to ensure that employees only have access to data necessary for their roles.
- **Two-factor authentication (2FA):** Put in place 2FA for accessing sensitive systems and data. This additional layer of security helps prevent unauthorised access, even if login credentials are compromised.

Regular data audits and assessments

- **Scheduled audits:** Conduct regular data audits to identify vulnerabilities and data security gaps. These audits should assess data handling processes, user access, and compliance with data governance policies.
- **Vulnerability scanning:** Use automated vulnerability scanning tools to identify potential weaknesses in your IT infrastructure. Regular scans help detect and address security issues proactively.
- **Penetration testing:** Periodically perform penetration testing to simulate real-world cyberattacks. This testing helps assess the resilience of your security measures and identify areas for improvement. To reiterate, data is the new currency in the digital age. Those who harness its potential through effective data management and security practices will thrive in the competitive and data-centric business landscape. It's time to recognise data as the true game changer, shaping the future of successful organisations. **END** 🐧

By: Dhaval Gajjar

The author is the CTO of Textdrip, and the CEO of Pranshtech Solutions and WeTechnolabs Solutions. With a passion for technology and innovation, he explores the ever-evolving world of digital solutions, sharing insights and expertise to drive progress in the tech industry.

Large Language Models: Helping Manage Data

Generative AI and large language models (LLMs) are the future, and promise a revolution in data management. However, development of LLMs is still very costly and inaccessible to smaller organisations. This will change as the years go by, and AI becomes more commonplace.



AI technology is changing the way the world does business. Generative artificial intelligence (generative AI) refers to the use of large language models (LLMs) to create new content, like text, images, music, audio, and videos.

LLMs are generative AI models that use deep learning techniques known as transformers. These models excel at natural language processing (NLP) tasks, including language translation, text classification, sentiment analysis, text generation, and question-answering. LLMs are trained with vast data sets from various sources, sometimes boasting hundreds of billions of parameters. They could

fundamentally transform how we handle, interact with and master data.

Prominent examples of large language models include OpenAI's GPT-3, Google's BERT, and XLNet, based on a whopping 175 billion parameters.

Industry adoption of large language models

Generative AI is primed to make an increasingly strong impact on enterprises over the next five years.

- The generative AI-based LLMs market is poised for remarkable growth, with estimations pointing towards a staggering valuation of \$188.62 billion by the year 2032. - *Brainy Insights*

- The world's total stock of usable text data is between 4.6 trillion and 17.2 trillion tokens. This includes all the world's books, all scientific papers, all news articles, all of Wikipedia, all publicly available code, and much of the rest of the internet, filtered for quality (e.g., web pages, blogs, social media). Recent estimates place the total figure at 3.2 trillion tokens. One of today's leading LLMs was trained on 1.4 trillion tokens. - *Forbes*
- By 2025, 30% of enterprises will implement an AI-augmented development and testing strategy, a substantial increase from 5% in 2021. - *Gartner*

- By 2026, generative design AI will automate 60% of the design effort for new websites and mobile apps - *Gartner*
- By 2027, nearly 15% of new applications will be automatically generated by AI without human intervention. This is not happening at all today. – *Gartner*
- According to Accenture’s 2023 Technology Vision report, 97% of global executives agree that foundation models will enable connections across data types, revolutionising where and how AI is used.

The key characteristics of LLM technology are:

- LLMs can automate data cataloguing, enhancing speed and efficiency.
- They can continuously monitor and improve data quality and detect anomalies.
- These models can generate insights from both structured and unstructured data.
- LLMs enhance data quality control by detecting inconsistencies and anomalies. They can automate the process of checking data against predefined quality standards, making it faster and more efficient.
- They can complete a given text coherently, translate text between languages, and summarise concisely.
- LLMs perform various NLP tasks, understanding and processing human language, and allowing users to ask questions in a conversational manner. They can generate insights in natural language, making them accessible to non-technical stakeholders.

LLM data preparation

Figure 1 shows the steps involved in LLM data preparation.

Data requirements: Establish the AI strategy. Identify areas where LLMs add value and the types of data that are needed for training LLMs, and select the right tools and technology for GenAI and data management. Identify the

various data source systems covering enterprise smarter applications, machine learning tools and real-time analytics. Categorise data into structured, semi-structured, and unstructured types. Structured data refers to databases of the enterprise. Unstructured data comprises videos, images, text messages, etc.

Data collection: Data sources provide the insight required to solve business problems. The various data sources are web, social media, text documents, etc. Collected data is used for training LLMs.

For example, to perform sentiment analysis by the trained LLM, the collected data should include a large number of reviews, comments, and social media posts.

Web scraping is the automated method of extracting data from various websites. Crawlee and Apify Universal scrapers are examples of web scraping tools.

Data organisation: Preprocess data by cleansing, normalising, and tokenising it.

- Data cleansing involves the identification and removal of inaccurate, incomplete, or irrelevant data. Duplicate data will be removed and incorrect data values will be fixed.
- Data normalisation transforms

data to standard format for easy comparison and analysis.

- Data tokenisation breaks the text into individual words or phrases using natural language processing (NLP). Tokenisation happens at word level, character level or sub-word level.

Data encoding: Feature engineering is crucial here. It involves creating features based on pre-processed data. Features are numerical representations of the text that the LLM can understand. The data is split, augmented, and encoded during this stage.

- Splitting is the process of dividing the data into training, validation, and testing sets. Training set is leveraged to train LLMs.
- Augmenting helps in synthesising new data and transforming the existing data.
- Encoding embeds data into tokens or vectors.

Data storage: This step primarily involves the Model hub, blob storage, and databases. The Model hub consists of trained and approved models that can be provisioned on demand, acting as a repository for model checkpoints, weights, and parameters. Comprehensive data architecture covering both structured and unstructured data sources is defined

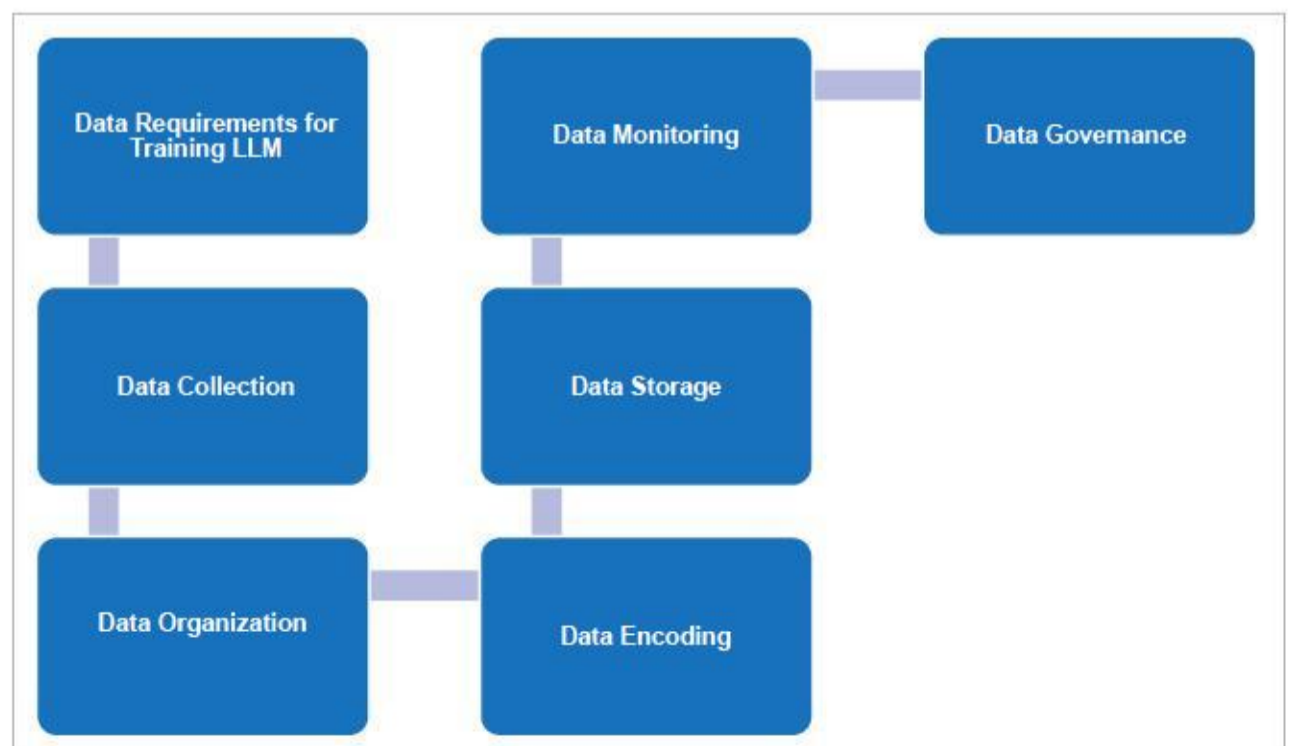


Figure 1: Data preparation for LLM

as part of the repository. The data is categorised and organised so that it can be used by generative AI models.

Data monitoring: This step involves monitoring the quality and relevance of the data, and updating it to improve the performance of the LLM.

Data governance: This step covers the automation of policies, guidelines, principles, and standards. Data governance must align with AI governance at the enterprise level to realise business objectives.

LLM models can be built using open source or proprietary models. Open source models are off-the-shelf and customisable, while proprietary models are offered as LLMs-as-a-service. Figure 2 shows a few LLM tools.

Let's now take a look at the use cases of LLMs across industries.

Healthcare and pharma

Generative AI-based applications help healthcare professionals be more productive, identifying potential issues upfront, and providing insights to deliver interconnected health. This helps in:

Better customer experience: Automating administrative tasks, such as processing claims, scheduling appointments, and managing medical records.

Patient health summary: Providing healthcare decision support by generating personalised patient health summaries, thus improving patient response times and experience.

Faster analysis of publications: Reducing the time it takes to create research publications on specific drugs by analysing vast amounts of data from multiple sources faster than ever. This helps in accelerating the speed and quality of care, and can also improve drug adherence.

Personalised medicine: Creating individual treatment plans based on a patient's genetic makeup, medical history, life style, etc.

Healthcare virtual assistant: Providing end users with conversational and engaging access to the most relevant and accurate healthcare services and information.

Manufacturing

Generative AI enables manufacturers to create more with their data, leading to advancements in predictive maintenance and demand forecasting. It also helps in simulating manufacturing quality, improving production speed, and efficiency in the use of materials.

Predictive maintenance: Helps in estimating the life of machines and their components, proactively providing information to technicians about repairs and replacement of parts and machines. This helps in reducing the downtime.

Performance efficiency: Anticipates production problems in real-time, covering the risks of production disruptions, bottlenecks, and safety.

Other usages of generative AI in the manufacturing industry include:

- Yield, energy, and throughput optimisation
- Digital simulations
- Sales and demand forecasting
- Logistics network optimisation

Input	Generative AI LLM Tools	Results
Text Generation & Chatbots Tools	Jasper Copy.ai OpenAI AYLIEN LinkDelta Curated AI Crystal Ryters Block Automated Insights ChatGPT Notion AI Artificial Fiesta agolo alltius Contlo InsurStaq.ai LONGSHOT Scalenut Social Beat senseforth.ai SYNCSENSE ZPOD	Blog posts, Articles, Product descriptions, Creative writing, Language Translation
Image Generation & 3D Modelling Tools	OpenAI Canva breeze.ai Wix.com FIREDROP Poly PRISMA Photo Room Booth.ai Adobe Sensei GRAPHIG BlendAI HEXO Nuro Pixel.AI Postify AI Spyne Uns Zapero Predis.ai Zocket Magic studio Pebblely	Image Generation, Image Captioning
Audio & Video Generation Tools	Fliki abridge amper Voicemaker Jukedeck audio bot Speechfy DeepMind Lyrebird Visla AI Topaz AI Synthesia HeyGen WIBBITZ 03 runway Scalenut SOLOSHOT WSC Sports WordAI Magesto Syllaby beatoven.ai dubpro.ai Listnr kroop AI MURF AI Predis.ai Rephrase.ai SYNCSENSE trureach.ai Unscript.ai vidyo.ai Vitra.ai	Text to Speech, Instruction Following, Chatbot builder Video Generation, Visually captivating videos, Text to Video, Streamline video creation process
Code & Data Generation Tools	GitHub copilot Replit Ghostwriter Codemate eizen FluidAI hype ai kombai observe-AI pixis workhack portkey Maya DhiWise CodeGPT	Generate code snippets, Respond to queries, Assisting developers with writing code, debugging existing code bases

Figure 2: LLM tools

Retail

Generative AI helps in personalising offerings and brand management, thus optimising marketing and sales activities. It enables retailers to tailor their offerings more precisely to customer demand, and supports dynamic pricing and planning.

Personalised offerings:

Generative AI enables retailers to deliver customised experiences, offerings, pricing, and planning. It also modernises the online and physical buying experience.

Dynamic pricing and planning:

It helps predict demand for different products, providing greater confidence for pricing and stocking decisions.

Other usages of generative AI in the retail industry include:

- Campaign management
- Content management
- Augmented customer support
- Search engine optimisation

Banking

Generative AI applications deliver personalised banking experiences to customers. They improve the financial simulations, developing risk analytics.

Risk mitigation and portfolio optimisation: Generative AI helps banks to build data foundations for developing risk models, identify events that are impacting the bank, mitigate that risk, and optimise portfolios.

Customer pattern analysis: It can analyse patterns in historic banking data at scale, helping relationship managers and customer representatives to identify customer preferences, anticipate needs, and create personalised banking experiences.

Customer financial planning: Generative AI can be used to automate customer service, identify trends in customer behaviour, and predict customer needs and preferences. This helps to understand the customer better and provide personalised advice.

Other usages of generative AI in the banking industry include:

- Anti-money laundering regulations
- Compliance
- Financial simulations
- Applicant simulations
- Next best action
- Risk analytics
- Fraud prevention

Insurance

The capability of analysing and processing large amounts of data by generative AI helps in accurate risk assessments and effective claims processes.

Customer support: Generative AI can provide multilingual customer service by translating customer queries and responding to them in the preferred language.

Policy management: It analyses large amounts of unstructured data related to customer policies, various policy documents, customer feedback, and social media literature for better policy management.

Claims management: Generative AI helps in analysing various claims artefacts to enhance the overall efficiency and effectiveness of claims management.

Other uses of generative AI in the insurance industry are:

- Customer communications
- Coverage explanations
- Cross-sell and up-sell products
- Accelerate the product development life cycle
- Innovation of products

Education

Generative AI offers real-time collaboration between teachers, administrators, and technology innovators.

Student enablement: Generative AI helps students who speak different languages with real-time lesson translations, and helps the visually impaired with classroom accessibility.

Student success: It offers deep analytic insights into student success, and helps teachers to make informed

decisions on how to improve results.

Telecommunications

Generative AI adoption improves operational efficiency and network performance in the telecom industry. It can be used to:

- Analyse customer purchasing patterns
- Personalise recommendations of services
- Enhance sales
- Manage customer loyalty
- Give insights into customer preferences
- Provide better data and network security, enhancing fraud detection

Public sector

The goal of most governments is to establish a digital government and provide better citizen services. Generative AI enables smart cities and optimised service operations.

Smart cities: Generative AI helps in toll management, traffic optimisation, and sustainability.

Better citizen services: It offers easier access to connected government services through tracking, search, and conversational bots.

Other services that are enabled using generative AI are:

- Service operations optimisation
- Contact centre automation

Limitations of current LLMs

Enterprises face several challenges in implementing LLMs as a part of generative AI solutions.

Data preparation: Identifying data sources for LLM, labelling data for algorithms, data policies, data security, data storage and data governance are the big challenges.

High volume data: LLMs reference huge amounts of data to generate meaningful output. This training and deploying of LLMs require significant computational resources, making them inaccessible to smaller enterprises or researchers with limited budgets.

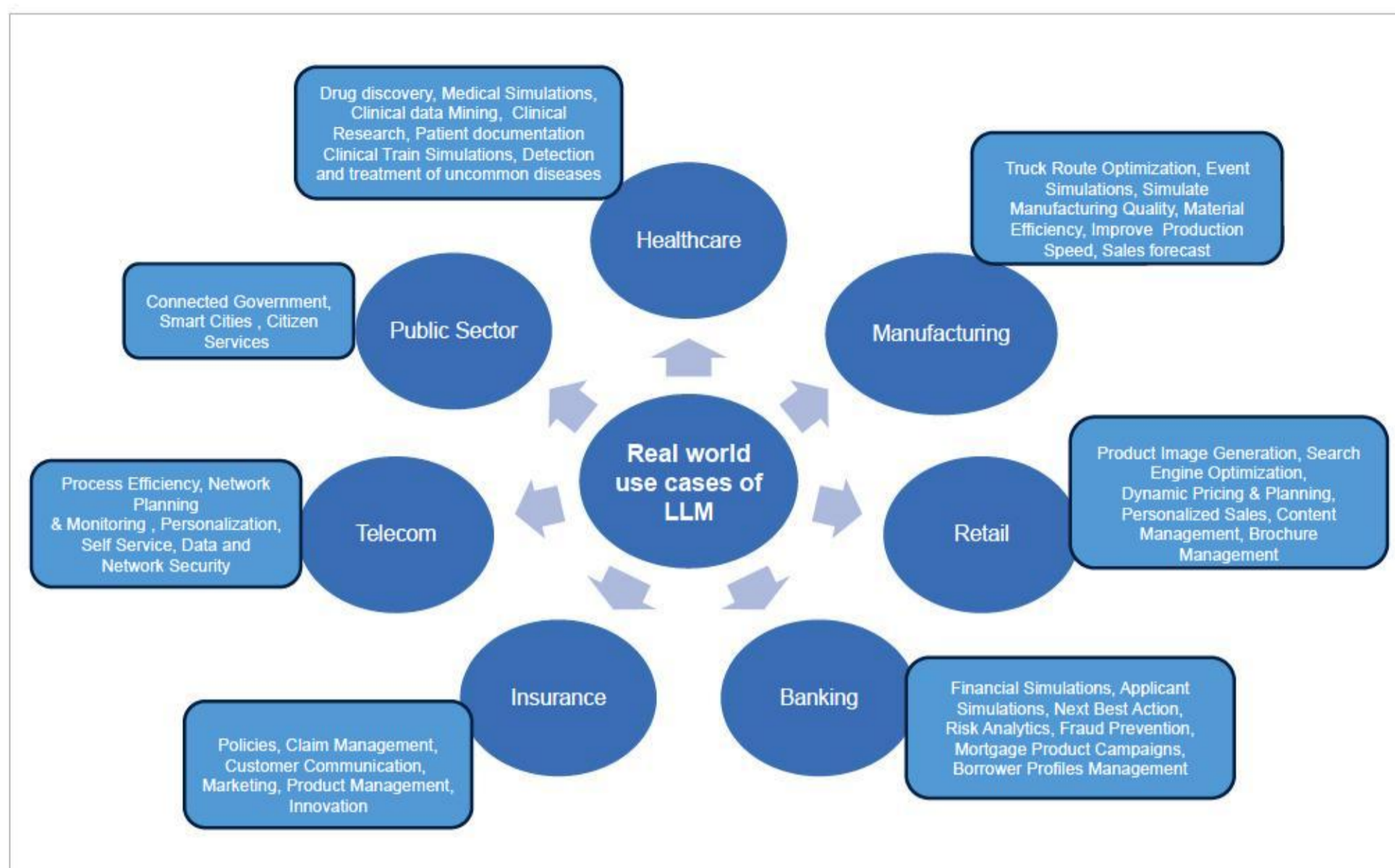


Figure 3: LLM use cases

High cost: Training costs of LLMs range from a couple of million dollars to ten million dollars, making them financially inaccessible for some organisations.

Data quality: Most LLMs have been trained in data based on web scraping. This may lead to plausible but factually incorrect or nonsensical responses, leading to poor data quality.

Data security: Sometimes, LLMs accept publicly available data as input. This can expose enterprise secrets, posing data security risks.

Reliability: Trained models are often 'black boxes', leading to false, harmful, and unsafe results.

Skills gap: Generative AI initiatives requires expertise in machine learning, deep learning, prompt engineering, and large language models, which many enterprises lack in-house.

In summary, LLMs are revolutionising the technology that helps in bridging the gap between

human communications and machine understanding.

The major steps involved in building custom LLM models for an enterprise are:

- Data gathering and preparation
- Training the model
- Evaluating the model
- Integrating with applications
- Deploying the model
- Continuous retraining

The future of LLMs is promising,

powering virtual assistants, improving machine translation quality and enabling innovative applications across various industries. Organisations such as Google, Microsoft, Amazon, Facebook, IBM and OpenAI leverage LLMs for applications like NLP, chatbots, content generation, sentimental analysis, etc. However, LLM technology is very new and not well understood, and many applications remain exploratory. **END** 🐧

Acknowledgements

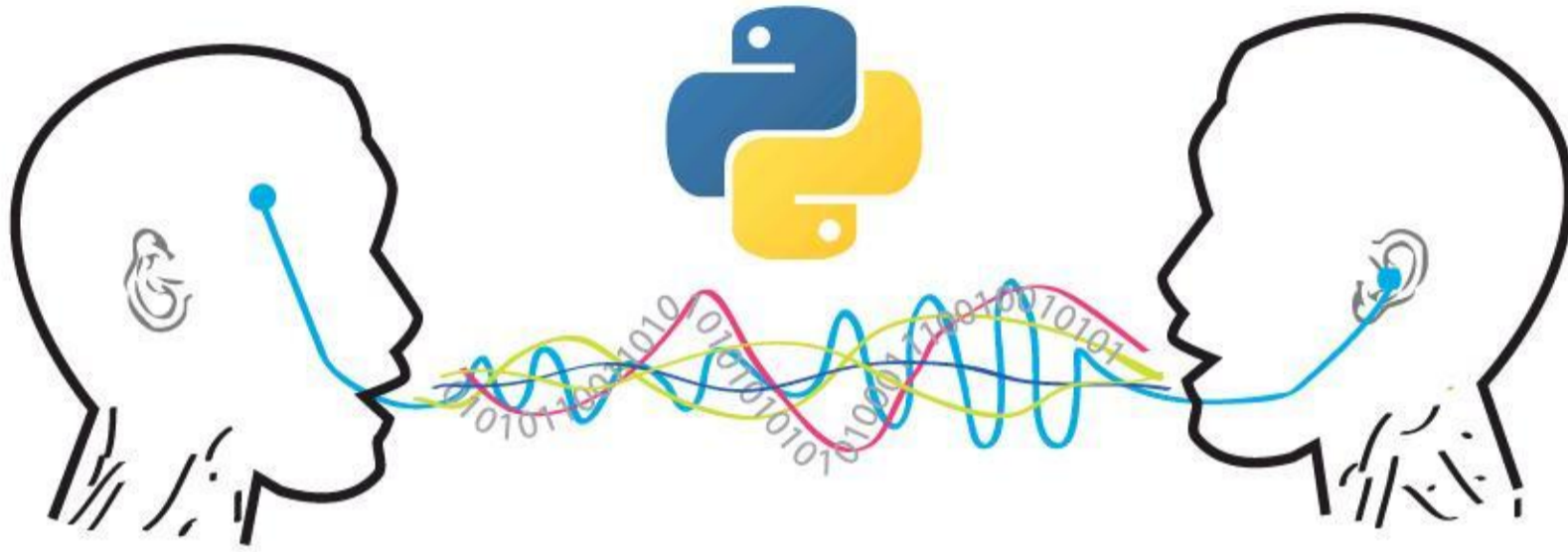
The author would like to thank Santosh Shinde of BTIS, Enterprise Architecture division of HCL Technologies Ltd, for the support in bringing this article to fruition as part of architecture practice efforts.

By: Dr Gopala Krishna Behara

The author is an enterprise architect in the BTIS Enterprise Architecture division of HCL Technologies Ltd. He has a total of 28 years of experience in the IT industry.

Disclaimer: The views expressed in this article are that of the authors and HCL does not subscribe to the substance, veracity, or truthfulness of the said opinion.

Encoding and Decoding in Python for Managing Data



Encoding and decoding help transform images, videos, graphics and text into the binary format that computers understand. Python has built-in functions and modules to perform encoding and decoding operations on string data.

Data is the core of any digital application, but how does it communicate with the computers that process it? Computers only understand the binary format, which is a system of zeros and ones, but data can come in various forms, such as images, videos, graphics, or text. How do we convert these data forms into binary format so that computers can understand and manipulate them? This is where encoding and decoding techniques come in handy. Let's explore how to encode and decode data in Python, one of the most popular and powerful programming languages for data science.

Encoding and decoding are processes of converting string data from one format to another, depending on the context and the goal. For example, you may want to encode string data that contains special characters or symbols into a format that can be transmitted over the internet, such as URL encoding. Or, you may want to decode a string of data that is encrypted

or compressed into a format that can be read and understood, such as Base64 decoding. Now this is all confusing, right? Let's delve deeper to clearly understand when to use which method. But first let's list different types of encoding and decoding methods, such as ASCII, Unicode, Base64, or URL encoding. Each method has its own rules and standards for how to represent string data in different formats.

Python is a programming language that has built-in functions and modules to perform encoding and decoding operations on string data. For example, you can use the `encode()` and `decode()` methods to convert string data between different encodings, such as UTF-8, Latin-1, or ASCII. You can also use the 'binascii' module to convert string data between binary and ASCII formats, such as Base64, hexadecimal, or binary. And you can use the 'urllib' module to convert string data between URL and plain text formats, such as URL encoding, URL decoding, or URL parsing.

However, encoding and decoding string data in Python can also

encounter errors and exceptions, such as `UnicodeError`, `ValueError`, or `TypeError`. These errors and exceptions may occur when the string data is not valid, compatible, or supported by the encoding or decoding method. For example, you may get a `UnicodeError` if you try to encode a string data that contains characters that are not in the ASCII range. Or, you may get a `ValueError` if you try to decode a string data that is not a valid Base64 format. And you may get a `TypeError` if you try to encode or decode a data type that is not a string, such as an integer or a list.

Python uses UTF-8 encoding as the default method for the `encode()` function. In the example given in Figure 1, we try to encode the string A using different

```
>>> "A".encode()
b'A'
>>> "A".encode('latin-1')
b'A'
>>> "A".encode('utf-8')
b'A'
>>> "A".encode('ASCII')
b'A'
```

Figure 1: Encoding using different forms

```

>>> "\xF1".encode('ASCII')
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
UnicodeEncodeError: 'ascii' codec can't encode character '\xf1' in position 0: ordinal not in range(128)
>>> "\xF1".encode('utf-8')
b'\xc3\xb1'
>>> "\xF1".encode('latin-1')
b'\xf1'
>>>

```

Figure 2: Difference in UTF-8 and Latin-1 encoding

encoding methods in Python. We use the `encode()` function with the name of the encoding method as the argument. As you can see, there is no difference in the output for the string A, as it is a common character that can be represented by the same byte in all the methods. However, this may not be the case for other characters, such as symbols, emojis, or foreign languages.

To understand encoding method differences, let's examine how they represent characters using bytes. A byte is a unit of data that consists of 8 bits, each of which can have a value of 0 or 1. Different encoding methods use different combinations of bits to map characters to bytes, depending on the range and the standard of the method. ASCII is one of the oldest and simplest encoding methods, which uses 7 bits to represent 128 characters, such as letters, numbers, punctuation marks, and control codes. ASCII is a signed method, which means that the first bit is reserved for indicating the sign of the byte, either positive or negative. Therefore, ASCII can only use the remaining 7 bits to encode characters, which limits its range to $2^7 = 128$ characters. ASCII can't encode characters beyond this range, such as symbols, emojis, or foreign languages.

UTF-8, an advanced universal encoding method, uses variable-length bytes to represent over a million characters from different languages and scripts. It is an unsigned method, which means that it does not use the first bit for the sign of the byte, but for indicating the length of the byte. Therefore, UTF-8 can use all 8 bits

to encode characters, which increases its range to $2^8 = 256$ characters for a single byte. However, UTF-8 can also use more than one byte to encode characters that require more bits, such as two bytes for $2^{16} = 65,536$ characters, three bytes for $2^{24} = 16,777,216$ characters, or four bytes for $2^{32} = 4,294,967,296$ characters.


Latin-1, also known as ISO-8859-1, is another encoding method. It uses 8 bits to represent 256 characters, mostly from Western European languages, such as French, German, or Spanish. Latin-1 is also an unsigned method, which means that it can use all 8 bits to encode characters, which gives it the same range as UTF-8 for a single byte. However, Latin-1 does not use variable-length bytes, which means that it cannot encode characters that require more than 8 bits, such as Chinese, Arabic, or Hindi. Latin-1 is also not compatible with UTF-8. This may cause

errors or corruption when converting between the two methods.

In Figure 2, we can see some characters differing in their representations in UTF-8 and Latin-1, with the latter having a much greater scope. Hence, not all encodings give the same bytes.

While Latin-1 consists of most of the characters, another encoding form called Base64 allows you to encode images, files and other media into bytes. Base64 decoding works by reversing the encoding process. It converts the encoded text string into groups of 4 characters, each of which corresponds to one of the 64 values in the Base64 alphabet. These values are then converted back to 6 bits, and the resulting bits are concatenated to form the original binary data.

In Python, Base64 is used to convert encoded string into Base64 bytes. This means that characters

unable to be represented by encoding methods like ASCII, Latin-1 and UTF-8 get shown as '??' but get a valid meaning once encoded with Base64. In this way, as shown in Figure 3, emojis, images, files and string are converted into the binary for the computer to understand. **END** 

```

>>> import base64
>>> "😊".encode()
b'??'
>>> "😊".encode('ASCII')
b'??'
>>> "😊".encode('UTF-8')
b'??'
>>> "😊".encode('latin-1')
b'??'
>>> base64.b64encode("😊".encode())
b'Pz8='
>>>

```

Figure 3: Explanation of Base64 encoding of non-ASCII characters

By: Anisha Ghosh

The author is a cybersecurity researcher and an active contributor to open source communities and repositories. She is interested in developing novel, scalable and secure systems.

Synthetic Data: Illusions, Lies, Deceptions and the Reality

It will be supremely tough to differentiate between reality and deception in an AI dominated world that uses synthetic data.

Data is synonymous with the word ‘information’ and in the 21st century, it is the most valuable asset. With every passing moment, more and more data is pumped into society. A common mistake that people make is to link the word ‘data’ with something related to computers or digital devices. It isn’t. Every book written and every picture painted forms a part of the ‘data’ circulating through society. The first form of data to be ever recorded can be traced back to cavemen’s paintings, stone sculptures, and so on.

Human beings have the innate tendency to try to replicate whatever they see or perceive. For example, the cavemen drew the animals they hunted, the storms that occurred, and so on. In the digital age, it is much easier to capture pictures of what we see, and record videos of what we experience. Apart from just what we perceive, what we do also forms a major part of the data in circulation -- for example, every bank transaction we carry out, every location we visit, and so on. And this data is not mere information; it is a valuable part and parcel of everyone’s life, and defines who they are. A computer system is able to perfectly predict one’s behavioural patterns with this data. This is nothing new. All of us have witnessed random ads on websites about the exact thing we are thinking about. There are even jokes around this, saying a computer is able to read minds, but in reality, it is able to perfectly predict your behaviour.

Having used Google as my default search engine since I first touched a

computer, and an Android device for over 7 years, my entire digital history can be summed up on a single web page within my Google account. This is not true just for me; it’s true for millions or even billions around the globe. You could try this out; explore what a single organisation ‘Google’ knows about your life. Just visit myactivity.google.com. And if one organisation has so much information about you, imagine how much more is out there if we count the information with multiple organisations including banks, travel agencies, e-commerce websites, and so on.

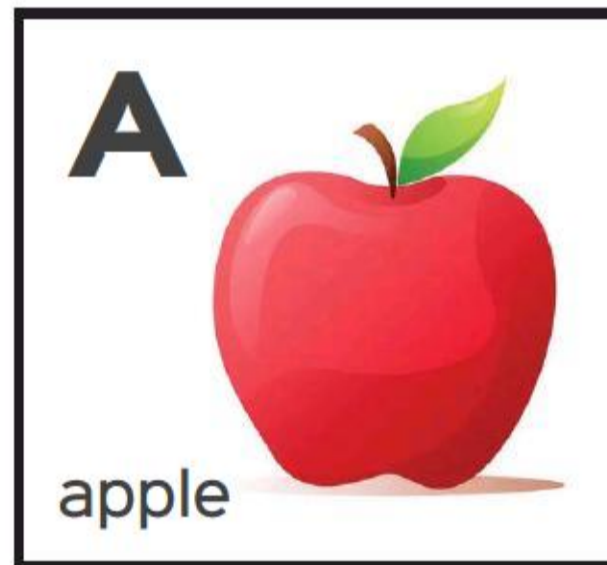


Figure 1: Children’s book showing letter ‘A’

Well, this was about the information of what we do and what we see. But a significant amount of data is about what does not exist, is imaginary and an illusion. As discussed before, even this illusionary data is not limited to the computer realm. Every fairy tale ever written, every lie ever told, and every image drawn from one’s imagination forms a part of the data about things that do not actually exist.

However, it is to be noted that even if data is not true, it doesn’t mean that it is not useful. Every child picks up their first book of alphabets, and the first word on the first page is usually ‘A for Apple’. But the picture beside that is generally a hand-drawn picture of an apple coloured uniformly red. As adults, we would agree that a real apple does not look like that picture. But that picture will help a child identify an apple when he or she finds one in real life. However, if a child is presented with a green apple, he or she may mistake it for a guava or a pear. This is because the pictures of apples are usually red, and the pictures of guavas or pears are usually green. So what can be done to avoid this situation? The solution is to introduce more pictures of apples to the child — both red and green. And the child’s mind then starts to identify the fruit by its shape and not the colour.

This is similar to what is done when building a machine learning model. A model needs to be introduced to or trained with a vast amount of data to be able to identify objects properly. And these ‘vast amounts of data’ for general items are available as data sets for model training. One of the most widely used data set for object recognition is the COCO (common objects in context) data set containing over 330,000 images.

But what is to be done when we are trying to identify some rare objects, about which so much of data does not exist? Let us go back to the analogy of a child. How would a child ever identify a dragon or a unicorn if he or she ever comes across one? There

is obviously no original picture of a dragon or a unicorn the child has ever possibly seen. But he or she has gone through drawings and artist impressions, which is a form of representation from imagination. It is data about something that does not exist. The data is entirely made up to help the child identify the dragon even if it never existed. Similarly, even for computers, in such cases, the data has to be made up. Made-up data is also known as ‘synthetic data’ and plays a very important part in the training of machine learning models.

But this is just about object recognition. Machine learning models can be used for a lot more purposes than just object recognition. For example, a model can be made to identify fraudulent transactions from bank statements. However, there may not be so many records of fraudulent transactions available to train the model. In such scenarios, synthetic data has to be prepared.

A very different type of artificial neural network called generative adversarial neural (GAN) network is used to generate synthetic data. But what are the moral obligations to the generation and use of synthetic data? That boils down to the fact that synthetic data is not real data and does not have a physical significance. For example, let’s look at this question — how safe is lying? Well, telling a child that a dragon lives in the mountains may be safe, but stating someone was killed by a dragon is not legal and an insurance fraud. It is the same lie, but the situation it is used in shows whether it is safe and legal.

As we approach the end of 2023, recent months have witnessed a substantial increase in the capabilities of AI. Generative AI now crafts stories, draws pictures, and performs various tasks with heightened proficiency. But GANs have been around for a long time. It’s just that generative AI has made them more accessible to everyone. This has increased cybercrimes and frauds a



Figure 2: Picture created by DALL-E-3 with the prompt ‘Realistic image of a teenager fighting a dragon in mountains, cloudy sky, teenager is using a sword, dragon breathing fire.’

lot. For example, the frauds over voice calls sound more realistic now because the scammer can actually use AI to modulate his or her voice to the voice of someone the victim knows. For the past few months, after Photoshop added the Generative AI fill feature, it has been a joke that people could fake being sick by photoshopping being admitted into a hospital. This shows how easy it is to tamper with evidence these days.

A prime example is deepfakes, which can accurately swap a person’s face with another with a high level of accuracy, often escaping detection by the naked eye. A recent example might be the deepfake of a popular south Indian actress doing the rounds on social

media, which was made from a video of a social media influencer. Deepfakes have been used by many criminals and in many cyber frauds. There has been a rising case of deepfakes on indecent pictures or videos to harm someone’s reputation.

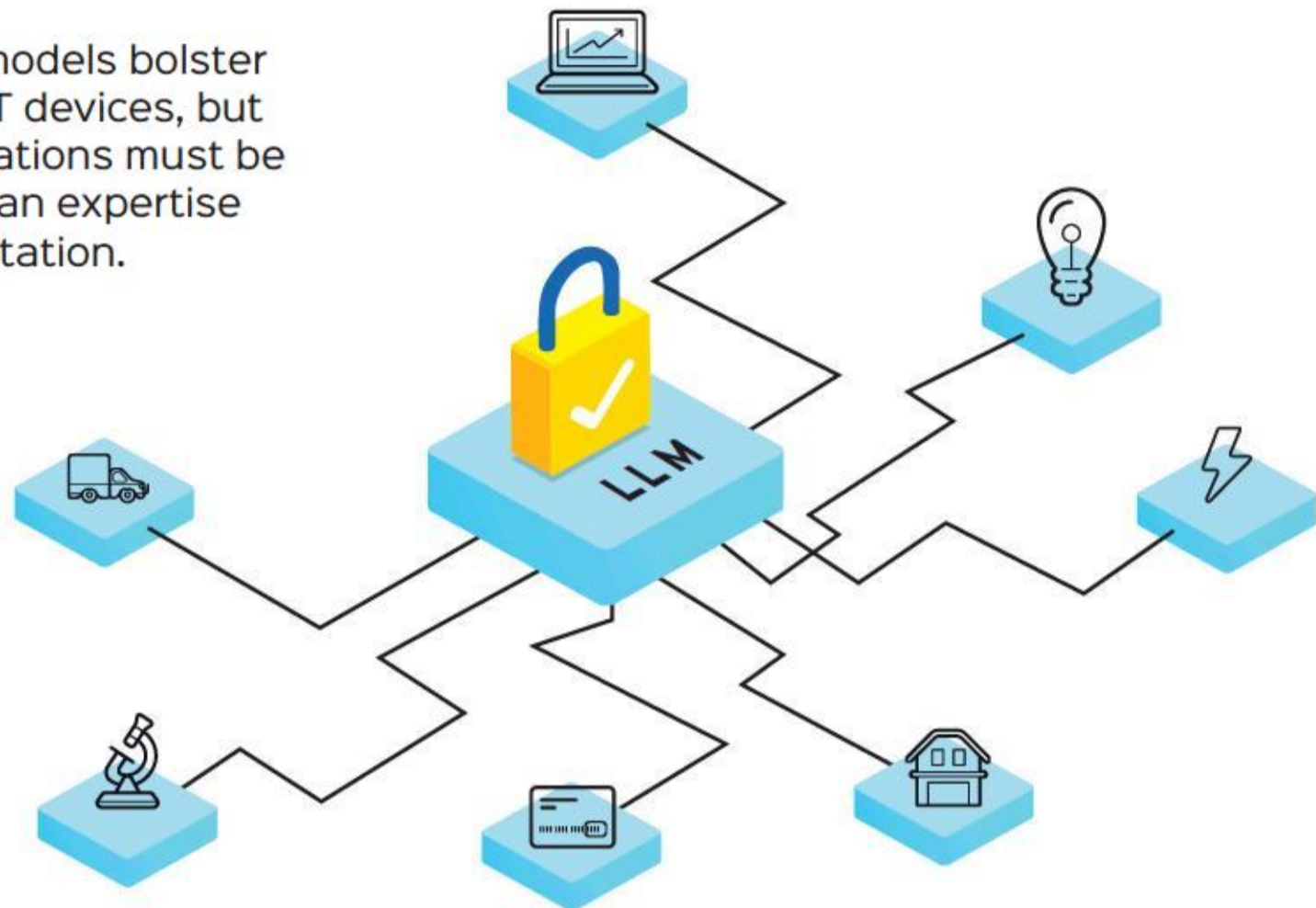
With the amount of synthetic data being pumped into society every single day, it is quite difficult to differentiate between reality, the illusion, and the deception and lies. As children, we trusted what we were told, and believed in dragons, unicorns, and fairies. But we soon outgrew that. As adults we may now need to enhance our discerning capabilities substantially to figure out what’s real and what’s not. **END** 🐧

By: Aditya Mitra

The author is a B. Tech, M. Tech, and PhD in CSE. He works as an assistant professor in the School of Computer Science Engineering and Technology at Bennett University, Greater Noida. He is experienced in cloud computing and open source technology.

Large Language Models: Enhancing IoT Security

Large language models bolster the security of IoT devices, but their recommendations must be validated by human expertise prior to implementation.



The Internet of Things (IoT) is a network of physical objects that can connect and exchange data with other devices and systems over the internet. These can range from ordinary household objects to sophisticated industrial tools. IoT finds applications in various industries, including smart homes, wearable devices, smart cities, and Industrial IoT.

IoT is a rapidly growing field with the potential to revolutionise many industries and aspects of our lives. IoT devices can enhance our lives by increasing the efficiency and security of our homes and businesses, making our cities more livable, and facilitating proactive healthcare systems.

Why IoT security?

IoT devices play a critical role in essential infrastructure, such as energy grids, transportation systems,

manufacturing plants and healthcare. Unauthorised access to these systems can result in severe repercussions, including power failures, transportation disruptions, and potential loss of life. Hence, protecting data, privacy, and the functionality of IoT devices is crucial.

In addition, IoT devices are often integrated into enterprise networks, creating a pathway for attackers to infiltrate and compromise corporate networks, leading to data breaches, and intellectual property theft.

Hence, as IoT devices proliferate, addressing IoT security challenges becomes increasingly important.

How is IoT security different from traditional security?

IoT security differs from traditional security in several ways due to the unique characteristics and challenges presented by IoT devices and

ecosystems. Here are some of the primary differences between IoT security and traditional security.

- **Attack surface:** IoT devices are often connected to the internet, creating a much larger attack surface compared to traditional devices.
- **Resource constraints:** IoT devices frequently have limited computing, memory, network, and energy resources, which restrict the ability to implement complex security measures.
- **Heterogeneity:** IoT devices come from various manufacturers, featuring different operating systems and security features, making it challenging to uniformly implement traditional security measures.
- **Physical environments:** IoT devices are often deployed in diverse and difficult-to-access remote locations, making them susceptible to physical

tampering and environmental factors that traditional IT infrastructure may not encounter.

- **Scale and quantity:** IoT networks can consist of thousands or even millions of devices, far exceeding the scale of traditional IT devices and networks. Managing and securing this large number of endpoints poses a significant challenge.

Additionally, there are other challenges like life cycle management, supply chain vulnerabilities, and the lack of proper user interfaces.

Traditional security solutions are designed to protect IT systems such as enterprise networks, servers, and workstations and are not effective at safeguarding IoT devices due to these unique challenges.

How to ensure IoT security

Securing IoT requires a tailored approach, including measures such as device authentication, encryption, secure device management and regular updates.

- **Device security:** IoT devices need security at the hardware and software levels. We need to implement lightweight embedded agents that can keep the device's operating system and software up-to-date. In addition, they should offer runtime protection to monitor state, and respond when there are anomalies detected to identify and combat zero-day attacks.
- **Network security:** We need to implement a zero-trust strategy and segment the network to isolate

IoT devices from other devices, minimising the attack surface.

- **Firmware security:** In-depth examination of an IoT device's firmware should be conducted to detect potential vulnerabilities within its code.

IoT security solutions should also include auto device discovery and identification, secure firmware updates, and real-time monitoring and anomaly detection, using security information and event management systems to collect and analyse log data from IoT devices.

IoT security with LLMs

LLMs (large language models) are AI models based on artificial neural networks, trained using self-supervised learning and semi-supervised learning. As autoregressive language models, they work by taking an input text and repeatedly predicting the next token or word. They have the potential to significantly impact IoT security in various ways.

- **Automated threat detection:** LLMs can be used to develop natural language processing models to analyse logs, reports, and security alerts from IoT devices and networks. This can help in automated threat detection and response by identifying patterns and anomalies in textual data.
- **Automated patch management:** These models can automate security tasks like vulnerability scanning and improved patch management of IoT devices.

- **Security awareness and education:** LLMs can generate educational content, articles, and documentation about IoT security to raise awareness and understanding of IoT security best practices.
- **Comprehensive security policy development:** They can assist in creating and refining security policies and guidelines for IoT deployments.
- **Secure configuration templates:** LLMs can help generate secure templates for IoT devices and provide guidance on how to configure devices for better vulnerability management. LLMs also have the potential to be used to develop new and more sophisticated IoT attacks that can have an adversarial impact on IoT security.
- **Adversarial attacks:** LLMs can be manipulated to generate deceptive content, including phishing emails, fake security advisories and malicious code that can lead to social engineering attacks against IoT users.
- **Automated exploitation:** LLMs can be used to automate malicious tasks, such as scanning for vulnerable IoT devices and exploiting them.
- **Privacy concerns:** Conversations between users and AI systems can disclose sensitive information about IoT devices and their vulnerabilities. LLMs are proving to be highly effective at solving general-purpose tasks since they are trained on large data sets based on a broad generalist knowledge base. However, their ability to access specialised, private, and up-to-date knowledge is limited. To enhance LLM knowledge in a specific domain, fine-tuning them on domain-specific data sets is crucial.

It is important to use LLMs judiciously in IoT security to maximise their positive impact and mitigate the challenges. Human expertise is essential to validate the recommendations and findings generated by AI models. This includes implementing safeguards to

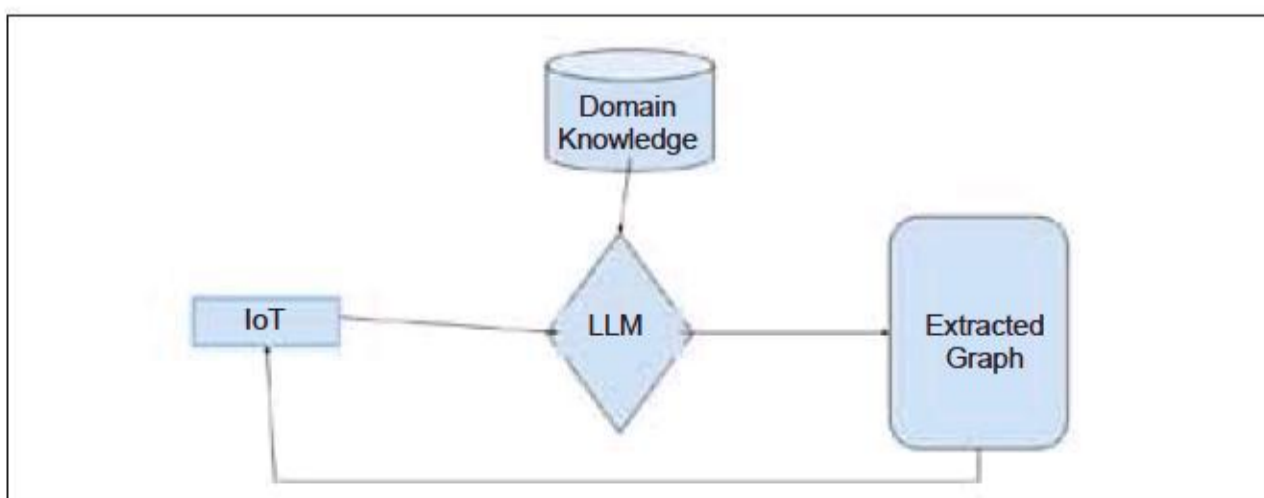


Figure 1: IoT security with LLMs

prevent adversarial attacks, curating AI-generated content, and ensuring human oversight in critical security decision-making processes.

IoT security is expected to evolve in response to the growing number of IoT devices and the increasing complexity of IoT ecosystems. Following are some key trends and developments that will guide the future of IoT security.

- Increased adoption of zero trust security with a focus on continuous verification and strict access controls for both devices and users
- Implementation of enhanced authentication methods like multi-factor authentication, biometrics, and hardware-based authentication
- Blockchain technology using tamper-proof ledgers for device identity, transactions, and data integrity to help secure device updates and firmware integrity
- AI and machine learning algorithms to help identify abnormal behaviour

and potential threats in real-time

- Edge computing security solutions to protect data and devices at the source
- Comprehensive device life cycle management, including secure onboarding, patching, and end-of-life decommissioning
- Introduction of stricter IoT security regulations and standards by governments and regulatory bodies to safeguard data and privacy
- Secure coding practices and rigorous security testing embedded into the development process
- Post-quantum cryptography

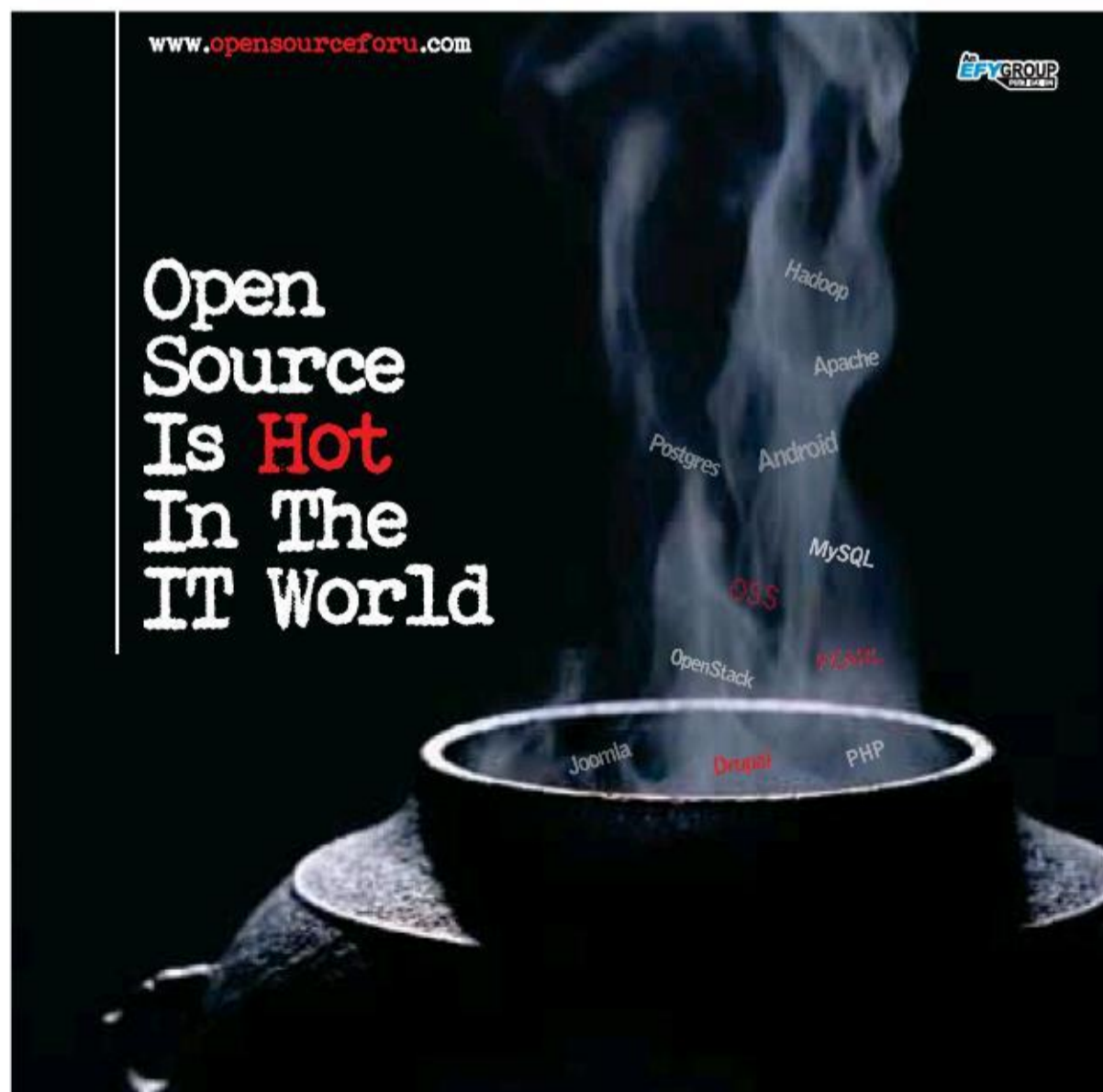
to secure IoT data against quantum attacks

Overall, IoT security requires a proactive and holistic approach to address the unique challenges posed by IoT ecosystems. As LLMs continue to develop, they will play an innovative role in protecting IoT devices and networks. Incorporating LLMs into IoT security operations requires careful planning, integration with existing security infrastructure, and ongoing monitoring to ensure that AI-generated insights align with the unique security requirements of IoT ecosystems. **END** 🐧

By: Radhakrishna Singuru

The author has more than 25 years of industrial experience in product and system software development spanning multiple industry domains. His areas of interest include cloud and virtualisation technologies, security engineering, blockchain, confidential computing, quantum computing, AI/ML, scalable platforms, network switching software, etc. He is currently working as senior principal software engineer, Palo Alto Networks.

Disclaimer: The views expressed in this article are that of the author and Palo Alto Networks does not subscribe to their substance, veracity or truthfulness.



THE COMPLETE MAGAZINE ON OPEN SOURCE



- >> To find dealers to buy copy from news-stand, visit: www.ezine.fymag.com
- >> To subscribe, visit: pay.efyindia.com
- >> To buy an ezine edition, visit: www.magzter.com choose Open Source For You

Putting G, P, and T Together

Get to know the Generative Pre-Trained Transformer (GPT) better, before you consider putting it to any serious use.



ChatGPT is writing essays, answering questions on the Bhagavad Gita, suggesting menu options for lunch, coding, and debugging code, writing new episodes of Star Wars, proposing itineraries for vacation travel, and even completing assignments for students. It garnered a million users faster than any well-known app in recent times. Perhaps because it is filling in for the much-missing human interaction in people's lives today. Almost everyone is talking about it and enjoying the conversation too—till it gets technical.

Artificial intelligence (AI) of this scale is not easy to understand—or explain. And the profuse use of jargon makes things worse. On the other hand, it is mandatory to know more about it, especially if we

plan to make any professional use of the tool. So, in this article we attempt to cut through the jargon with the help of some experts on the subject. This will help us understand the technology behind this chatbot, how it is different from others, what it is capable or incapable of, how to use it, whether there are other such worthy AI models, what does the future hold for it, and so on.

For some time, let us cut the chat and focus on the GPT!

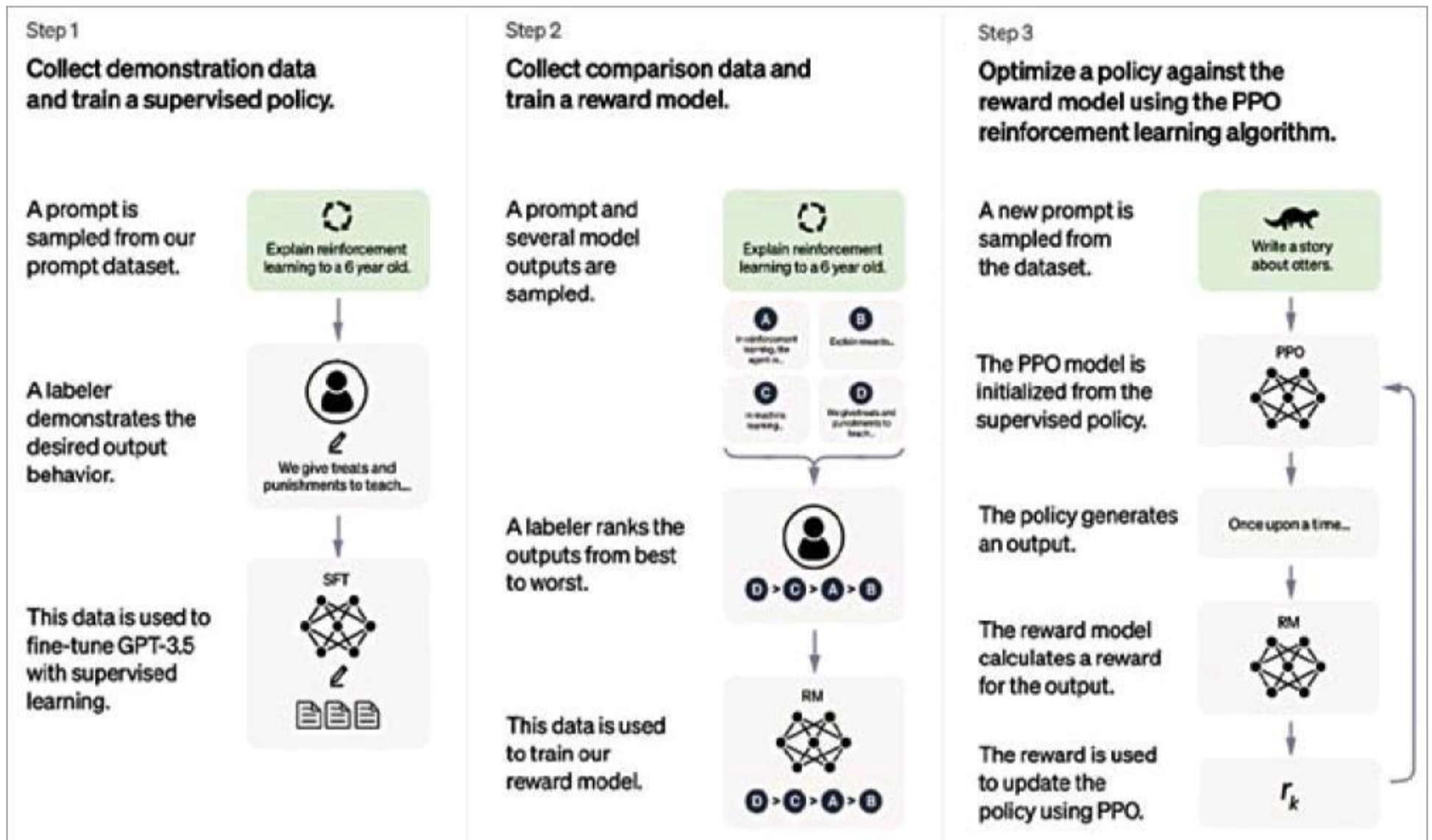
AI is everywhere, how is ChatGPT different?

AI is all around us. From Netflix recommendations and browser ads to a game of chess, and those decision-influencing insights that a banking

platform's dashboard provides, it is all artificial intelligence. By now, we are all used to specific intelligence—being smart at one or few things (which applies to most of us also!). Yet, the kind of general intelligence and conversational skills displayed by OpenAI's ChatGPT makes one's nerves tingle—in anticipation, and in fear.

“There are two main differences between ChatGPT and traditional conversational AI models: the technology and user experience,” remarks Anurag Sahay, CTO and managing director - AI and Data Science, Nagarro.

In terms of technology, ChatGPT distinguishes itself by being based on the most advanced generative



Reinforcement Learning from Human Feedback (RLHF) used to train ChatGPT (Courtesy: OpenAI)

pre-trained transformer (GPT) language model available today, which uses a transformer architecture with a large number of parameters that allows it to learn and generate human-like text. This model exhibits greater capability in comprehending context, drawing generalisations from examples, and furnishing responses that are more precise and pertinent when contrasted with conventional conversational AI models.

The free version of ChatGPT is currently based on GPT-3.5, while the paid ChatGPT Plus is based on the GPT-4 platform, launched in March 2023. GPT-4 is multi-modal. It can handle text and image inputs. You can scan graphs, books of accounts or question papers and get answers from it.

OpenAI says GPT-4 can solve difficult problems with greater accuracy thanks to its broader general knowledge and problem-solving abilities. It is apparently more intelligent, because it has scored higher in benchmark

exams like the graduate record examinations (GRE) and the scholastic assessment test (SAT).

With respect to user experience, ChatGPT conversations are more engaging and interactive compared to traditional conversational AI models. Sahay feels that, “This is because ChatGPT has a clearer understanding of context, which allows it to maintain a more coherent conversation and respond specifically to user queries. As a result, users feel like they are talking to someone who is intelligent, knowledgeable, and empathetic, which makes the interaction more natural and human-like. On the other hand, when users talk to a bot built with traditional conversational AI models, they may not feel that the agent on the other end truly understands them or is empathetic. Overall, the combination of ChatGPT’s advanced technology and more natural user experience sets it apart from traditional conversational AI models.”

Sanjeev Azad, vice president—technology, GlobalLogic, also feels that one of the key differentiators of ChatGPT is its adaptability—“The way it quickly adapts to the user’s communication style based on the questions asked, and personalises responses as per the user’s need.”

He further explains that the power and effectiveness of large language models can be evaluated in various ways, but the two most considered factors are the number of parameters and the extent of language coverage. GPT-3 has more than 175 billion parameters. Although the number of parameters of the multi-modal GPT-4 is not officially available as of the date this story is filed, it is rumoured to be more than a trillion.

ChatGPT can converse in any language as long as it has been trained on enough data in that language. According to a report by seo.ai, ChatGPT knew 95 natural languages, as of February 2023. It has also been trained on several programming

languages, including Python, Java, C++, JavaScript, and Ruby.

OpenAI's GPT is not without contenders in this respect. WuDao 2.0, a multi-modal AI developed by the Beijing Academy of Artificial Intelligence, has around 1.75 trillion parameters and can handle conversations in English and Chinese. The Megatron-Turing Natural Language Generation Model, a collaboration between Microsoft and Nvidia, uses novel parallelism to achieve 530 billion parameters.

BigScience's open-access AI, Bloom, can supposedly handle 46 natural languages and 13 programming languages, with its 176 billion parameters. Google's Language Model for Dialogue Applications (LaMDA), which powers Bard, has 137 billion parameters.

"Yes, there are other powerful transformers and large language models besides GPT-3, but GPT-3 has gained the maximum popularity to date," remarks Azad.

Behind the scenes

Azad spends some time explaining the working of ChatGPT to us. "ChatGPT uses a large neural network to generate responses to user inputs. The neural network is trained on massive amounts of text data and uses this knowledge to generate responses similar in style and content to human-generated text."

He explains the working of ChatGPT in five steps.

Step 1 - Pre-processing: This is the first step to ensure we have the relevant data to feed into the neural network, as neural networks require a lot of computing power to process the information. This includes steps such as tokenisation, where the input is split into individual words, and encoding, where each word is converted into a numerical representation.

Step 2 - Neural network: At its core, a neural network uses a 'transformer architecture' technique to process the input and generate a response. This architecture allows ChatGPT to model complex relationships between words

Demystification zone

Let us first cut through some of the jargon you are likely to encounter in this and other articles on ChatGPT...

Foundation models: We could say that foundation models are, well, foundational to today's AI revolution! Foundation models induce an element of reusability in AI systems, rather than having to create and train task-specific AI models for every single requirement that arises. Foundation models are trained on large, unlabelled datasets and then fine-tuned or trained for specific applications.

"Typically, foundation models are trained using some self-supervised tasks, such as learning, to predict the next word in a sequence of words. Using a foundation model, very little supervised data is required to develop a model for any specific task. A very popular deep learning architecture that is used nowadays for learning language based foundation models is the transformer architecture," says Sachindra Joshi of IBM Research India.

Transformer models: Transformer models track relationships in sequential data and use that to understand context and meaning. Transformers use a technique called self-attention to focus on relevant parts of an input and make more accurate predictions.

"Transformer models are a specific kind of deep learning architecture, which learn the representation of a token (word) based on all the other token representations in a text. This model is very effective in learning good token representations, and the computations required to learn these representations can be easily parallelised using GPUs. This makes it possible to use an enormous amount of textual data to learn very effective token representations," explains Joshi.

(The paid subscriptions to ChatGPT and other AI tools measure your usage based on tokens—now, we have demystified that term as well! According to OpenAI, 1000 tokens would roughly measure up to 750 words.)

Large language models: Large language models (LLMs) are considered to be one of the most successful applications of the transformer model. LLMs are pre-trained using massive volumes of unlabelled data. The LLM learns high-level features at this stage. It can then be fine-tuned for specific tasks.

If you take the example of ChatGPT, it has been pre-trained using troves of textual data from the internet, from which it has learnt to understand context and meaning well enough to hold mostly meaningful conversations with users. As a next step, users can utilise the ChatGPT API to make their own apps, fine-tuned and trained for specific purposes, say, to answer questions on Indian history, the Bhagavad Gita, or customer queries about a company's products.

LLMs can be used for more than natural language processing. They can also be used to study and predict patterns in weather, protein structures, financial data, cyber frauds, and more!

Generative AI: Generative AI is a term that is used to refer to any model that can generate some content—could be text, images, or audio—in response to a prompt. ChatGPT and Dall-E are popular examples in today's context.

Multimodal AI: It can understand and respond in multiple modes, such as text, images, and audio. You could, for example, input an image or a graph as part of your prompt. The latest version from OpenAI, GPT-4, released in March 2023, is multimodal.

and generate grammatically correct and semantically meaningful responses.

Step 3 - Language modelling: The neural network uses a language model to predict the most likely words to follow the user's input. The language model is trained on vast text data and uses this knowledge to make predictions.

Step 4 - Beam search: ChatGPT uses a technique called 'beam search' to ensure that the generated response is grammatically correct and semantically

meaningful. This involves generating multiple possible responses and selecting the one that is most likely to be correct based on a set of criteria, such as the likelihood of each word in the response.

Step 5 - Post-processing: Once the neural network has generated a response, it is post-processed to ensure that it is in a format that is easy for the user to read and understand. This includes steps such as converting

numerical representations back into words and ensuring that the response is grammatically correct.

“These steps necessitate a significant amount of computing power, and ChatGPT relies on high-performance hardware to ensure that it can generate responses quickly and accurately. It generally requires powerful graphics processing units (GPUs) or tensor processing units (TPUs) to handle the massive amounts of data involved in natural language processing,” he explains.

Platforms like Microsoft Azure and IBM Cloud come to the rescue, providing a seamless fabric of compute power required to train AI models. According to reports, ChatGPT initially used a Microsoft Azure supercomputing infrastructure, powered by Nvidia GPUs, which Microsoft built specifically for OpenAI. This year, after Microsoft announced its multi-year investment in OpenAI, it also phenomenally upgraded OpenAI’s infrastructure.

Sachindra Joshi, IBM distinguished engineer, Conversational Platforms, IBM Research India, remarks that, “Building and deploying foundation models requires proper infrastructure. Training models like ChatGPT requires thousands of GPUs, making it extremely difficult for many entities to train these kinds of models from scratch. Deploying these models for inference is also very resource consuming and a lot of research is happening around better learning models with fewer parameters, and making inference costs cheaper.”

With great power comes great responsibility

All that impressive technology working in the background gives you not just a lovable chatbot that you can chat and while away your time with, but also a very capable API that can be used to enhance applications and business platforms.

Is ChatGPT open source

Despite the company being called OpenAI, the code of ChatGPT is not open sourced. Since the company was not available for comment, we asked ChatGPT itself about it, and here is what it had to say:

“As an AI language model, my code is not open sourced. However, the underlying technology that powers me is based on research and development by OpenAI, and some of the tools and libraries used to create me are open source. OpenAI has also released several pre-trained models for use in natural language processing tasks, which are available for free download and use. If you’re interested in exploring these resources, you can find more information on the OpenAI website.”

This is something that has miffed many in the open source world, because ChatGPT uses quite a lot of open source components and open knowledge as well. (Practically most of its training has happened on information garnered from the Web, contributed by people from around the world.) Perhaps it is not giving back as much as it takes?


ChatGPT can converse on varied subjects, draft emails, make lists, translate text into different languages, summarise long texts, assist in various domains like education, healthcare, finance, and customer service, and do much, much more. It can adapt to different communication channels and be integrated with external data sources and application programming interfaces (APIs) to provide more customised and comprehensive responses.

Hand-in-hand with all this power comes a lot of risk as well. From privacy and plagiarism to cybersecurity issues, and the very serious possibility of it dumbing down the human brain, there are several issues that we must be aware of, so we can make responsible use of generative AI.

Indeed, ChatGPT has taken trendwatchers by surprise, by racing to the peak of the hype cycle within a phenomenally short span of time. Unlike the metaverse, which people are still struggling to understand, let alone adopt, generative AI (especially ChatGPT) has already proved its worth to people and seems to be here to stay.

“ChatGPT has revolutionised the accessibility of AI to the masses, providing people with the opportunity to interact with it in ways that were once exclusive. The pace at which enterprises are investing in AI is remarkable, having surged by 2.5 times over the last five years. Yet, the number of AI products available today is several multiples higher than it was in 2017. The impact of this disruption is truly remarkable,” says Sahay.

He explains that, in the near future, these generative AI models will become better, less expensive, and more accurate. Because these models are becoming multimodal, a single model will soon be able to synthesise information from all four modalities—text, image, video, as well as speech.

Sahay signs off saying, “This is a critical future element that everyone is anticipating, and will be more closely monitored and moderated for ethical reasons. There is no doubt that everyone is on board to ensure that both privacy and accountability are addressed, and that more regulation and governance in this area are clearly needed.” 

By: Janani G. Vikram

The author is a freelance writer based in Chennai, who loves to write on emerging technologies and Indian culture. She believes in relishing every moment of life, as happy memories are the best savings for the future.

The article was originally published in the May 2023 issue of Electronics For You.

Blockchain Basics

Blockchain is a way of storing data that makes it difficult or impossible for the system to be altered, hacked, or otherwise manipulated. It is a type of distributed ledger that distributes and copies transactions among the network of computers involved. Let's take a quick look at all that it entails.



Blockchain technology is a system for storing public transactional records, also known as blocks or 'chains', across multiple databases that are linked by peer-to-peer nodes. Termed as a 'digital ledger' this storage method ensures transaction authenticity through the owner's digital signature, offering protection against fraud. Consequently, the data in the digital ledger is extremely secure. In simpler terms, the digital ledger can be likened to a Google spreadsheet shared among many computers connected to a network, where transactional records are kept in accordance with actual purchases. The data can be seen by anyone, but nobody can alter it, which is an intriguing aspect.

Blockchain 1.0 (cryptocurrency)

Spearheaded by Hall Finley, Blockchain Version 1.0 emerged in 2005 as the inaugural implementation of distributed ledger technology (DLT) within the realm of cryptocurrency.

This development facilitated financial transactions rooted in blockchain or DLT, with Bitcoin as its enabling vehicle. Notably, Version 1.0 is characterised by its permissionless nature, allowing any participant to

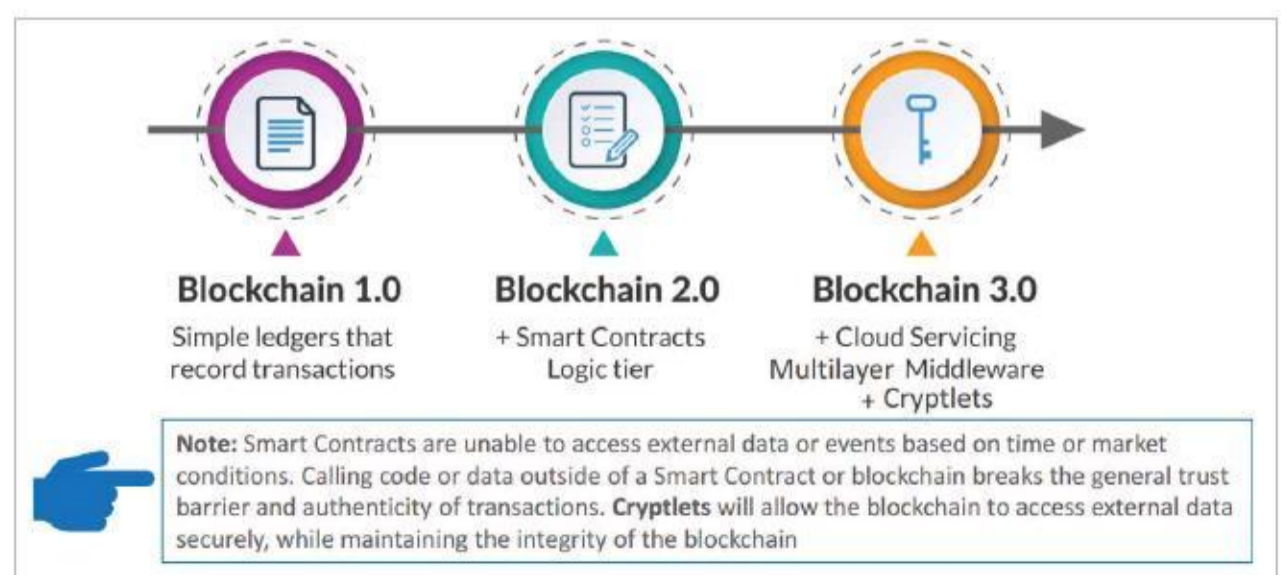


Figure 1: Evolution of blockchain

engage in valid Bitcoin transactions. This iteration finds its primary utility in the domains of currency and payments.

The overarching goal of Blockchain 1.0, also referred to as Blockchain Version 1.0, was to usher in a transparent, publicly accessible, fully decentralised, immutable ledger, and a distributed transaction system within the global financial landscape. It takes its foundational cues from the concept and structure of Bitcoin, with a primary emphasis on the genesis of novel cryptocurrencies. Blockchain 1.0 is often described as a digital, decentralised, distributed ledger that meticulously records transactions within a universally shared database. These records are updated by blockchain miners, and collectively overseen and maintained by all participants, without any single entity claiming ownership.

Blockchain 2.0 (smart contracts)

The introduction of Blockchain Version 2.0 stemmed from inherent issues found in its predecessor, Version 1.0. Notably, Version 1.0 faced challenges related to the resource-intensive nature of Bitcoin mining and a notable lack of network scalability. Version 2.0 was conceived as a means to rectify these shortcomings.

In this upgraded iteration, blockchain extends its functionality beyond merely handling cryptocurrencies to encompass smart contracts. These smart contracts can be likened to compact computer programs residing within the interconnected blocks of the chain. They autonomously execute predefined conditions, such as facilitation, verification, or enforcement, thus streamlining processes and enhancing transaction cost efficiency.

One significant shift in Blockchain 2.0 involved replacing Bitcoin with Ethereum as its foundational cryptocurrency. This transition proved highly successful, enabling Blockchain 2.0 to efficiently process a substantial volume of transactions on the public network with remarkable speed.

Blockchain 3.0 (DApps)

Version 3.0, that followed Version 2.0, features the inclusion of DApps, commonly known as decentralised applications. These DApps resemble traditional applications in that they can have a frontend coded in any language, facilitating communication with their backend. However, what sets them apart is that their backend operations are executed within a decentralised peer-to-peer network. They harness decentralised storage and communication mechanisms, such as Ethereum Swarm, to function.

DApps are inherently decentralised, devoid of a single owner or central authority, ensuring several key advantages, including transparency, heightened security, universal data accessibility, absence of censorship, and a flexible development environment. The adoption of DApps brings forth numerous benefits, including uninterrupted operation (zero downtime), robust privacy protection, data integrity assurance, and trustworthy, yet secure, communication for various purposes such as business and transactions.

Why is blockchain important?

Transparency and trust: Blockchain ensures transparency and an immutable record of transactions, making it difficult to alter or manipulate data. This transparency and immutability build trust among participants and eliminate the need for intermediaries, reducing costs and increasing efficiency.

Security: Blockchain uses cryptographic algorithms to secure transactions and data. The decentralised nature of blockchain networks makes them resilient to hacking and fraud attempts. This enhanced security is crucial for applications like financial transactions, supply chain management, and identity verification.

Decentralisation: Blockchain operates on a distributed network

of computers or nodes, eliminating the need for a central authority or intermediary. This decentralisation ensures that no single entity has control over the network and reduces the risk of single points of failure or censorship.

Innovation and disruption:

Blockchain technology has the potential to disrupt traditional industries and business models by introducing new ways of transacting, storing data, and executing contracts. It enables the development of decentralised applications, smart contracts, and new financial instruments, fostering innovation and opening up new business opportunities.

How does blockchain work?

The developments in blockchain are relatively nascent, yet they harbour the potential to bring about revolutionary changes in the future. So, let's start the process of analysing this technology as shown in Figure 2.

Blockchain comprises a synergy of three key technologies:

1. Cryptographic keys
2. A peer-to-peer network housing a shared ledger
3. A computational framework for storing the network's transactions and records

Cryptography keys, comprising the private and public keys, are foundational to blockchain. These keys play a crucial role in facilitating secure and prosperous transactions between two parties. Every individual possesses both these keys, deploying them to establish a robust and unassailable digital identity reference, an element of paramount significance within the blockchain framework.

In the realm of cryptocurrency, this digital identity takes the guise of a 'digital signature', a term widely employed to authorise and oversee transactions. The digital signature seamlessly integrates into the peer-to-peer network, where a diverse

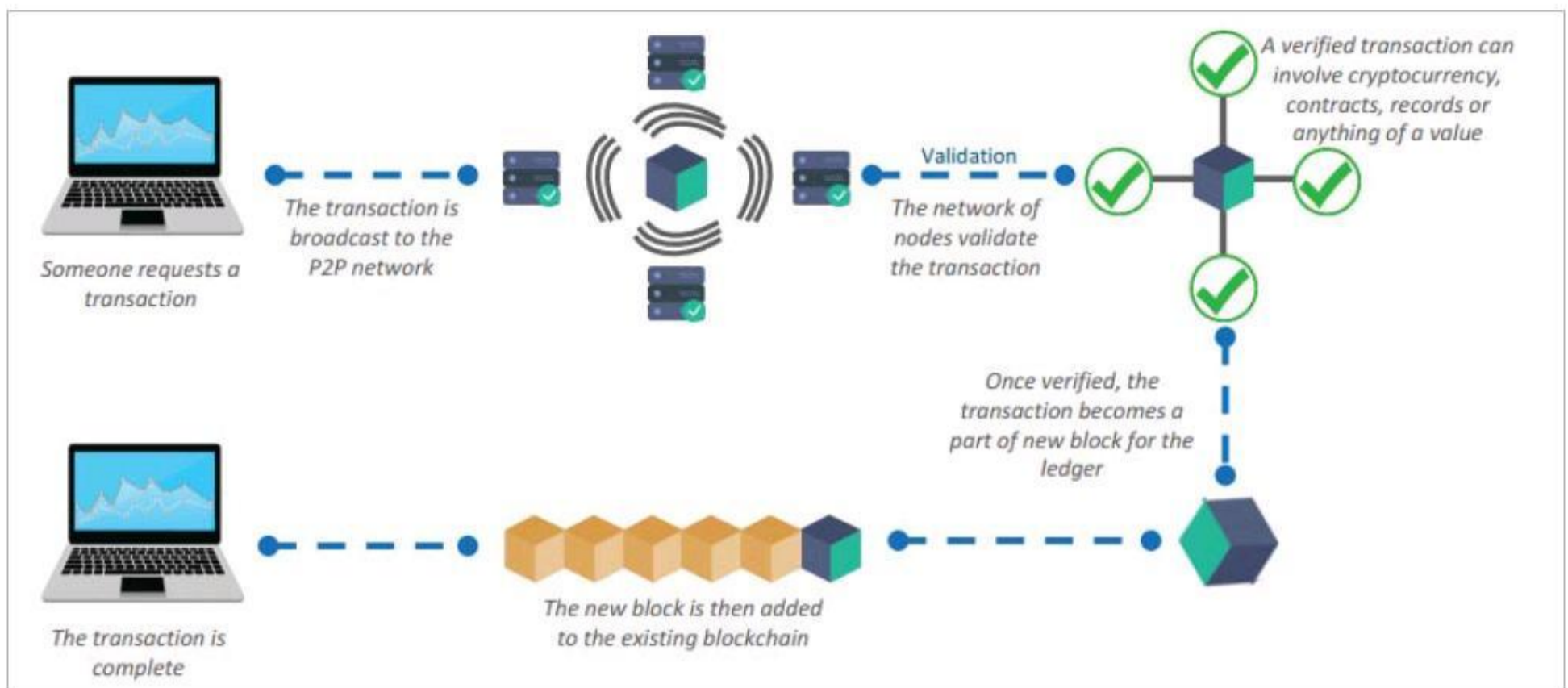


Figure 2: How the blockchain works

community of individuals collectively act as governing authorities. These individuals employ digital signatures to achieve consensus on various facets.

Upon receiving authorisation, a transaction undergoes meticulous mathematical verification, culminating in a successful and secure transaction between the interconnected parties within the network. In essence, blockchain users wield cryptography keys as they engage in a myriad of digital interactions across the peer-to-peer network.

Here's a use case that illustrates how blockchain works.

Hash encryption: The SHA256 algorithm secures data through hashing and encryption in blockchain technology. It transmits information about the sender's address (public key), the recipient's address, the transaction, and his or her private key. After verification, the hashed data is sent across the globe and added to the blockchain in encrypted form. The sender and receiver's authentication is made easier by the SHA256 algorithm's virtually impenetrable hash encryption.

Proof of work: Each block in a blockchain has four main headers.

- **Previous hash:** The previous block can be found at this hash address.
- **Transaction details:** Information about each transaction that must take place.
- **Nonce:** In cryptography, an arbitrary number is used to distinguish the block's hash address.
- **Hash address of the block:** A hashing algorithm transmits the preceding hash, transaction information, and nonce. The output value with a length of 64 characters and 256 bits is referred to as the distinct 'hash address' or block hash.

Numerous people around the world try to figure out the right hash value to meet a predetermined condition using computational algorithms. The transaction completes when the predetermined condition is met. In simpler terms, blockchain miners solve a mathematical puzzle, known as a proof of work problem. Whoever solves it first gets a reward.

Mining: The act of appending transactional data to the existing digital public ledger is commonly known as 'mining'. Although this term is closely linked to Bitcoin, it extends its application to other

blockchain technologies. The mining process entails the creation of a block transaction's hash, a highly secure and tamper-resistant mechanism. This process safeguards the entire blockchain network, eliminating the requirement for a centralised system.

Types of blockchains

There are primarily three types of blockchains.

Public blockchain: Public blockchains are open and permissionless, meaning anyone can participate in the network, validate transactions, and create new blocks. They are decentralised and rely on a consensus mechanism, usually proof of work (PoW) or proof of stake (PoS), to validate and add new blocks. Bitcoin and Ethereum are examples of public blockchains. Public blockchains provide transparency and security but may have scalability limitations due to their openness.

Private blockchain: Private blockchains, also known as permissioned blockchains, are restricted to a specific group of participants. These participants are known and trusted, and they are given permission to access and

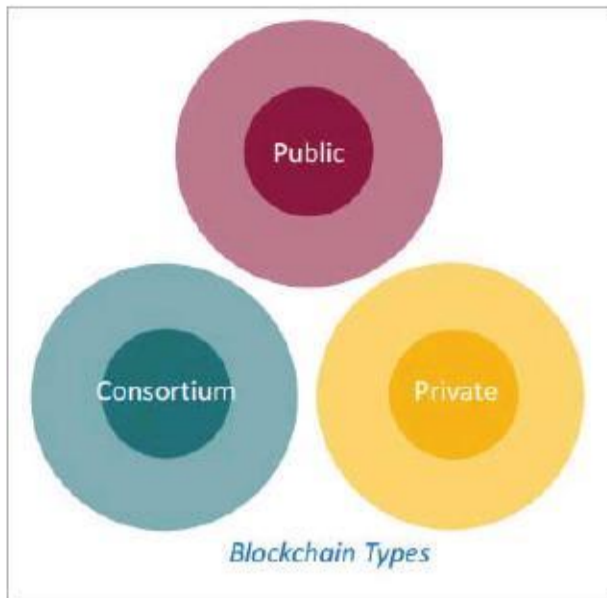


Figure 3: Types of blockchains

validate transactions on the network. Private blockchains are often used by businesses or organisations for internal purposes, such as supply chain management or record-keeping. They provide greater control and privacy compared to public blockchains but sacrifice some decentralisation.

Consortium blockchain: Consortium blockchains are a hybrid between public and private blockchains. They are managed and operated by a group of organisations or entities working together in a consortium. Consortium blockchains offer a more decentralised approach compared to private blockchains while maintaining a level of control among the consortium members. They are often used in industries where multiple stakeholders need to collaborate, such as banking consortia or supply chain networks.

Where is blockchain used?

Blockchain technology has a wide range of uses across various industries.

- **Cryptocurrencies:** Blockchain technology is most commonly associated with cryptocurrencies like Bitcoin and Ethereum. It enables secure and transparent peer-to-peer transactions without the need for intermediaries like banks.
- **Supply chain management:** Blockchain can be used to track and trace the movement of goods

across the supply chain. It provides real-time visibility, enhances transparency, and ensures the integrity of data, reducing fraud, counterfeiting, and improving efficiency.

- **Healthcare:** Blockchain technology can securely store and share medical records, ensuring patient data privacy, interoperability, and accessibility. It can also streamline the process of clinical trials, drug traceability, and supply chain management in the healthcare industry.
- **Finance and banking:** Blockchain has the potential to revolutionise the financial sector by enabling faster, more secure, and cost-effective cross-border payments, remittances, and settlements. It can also facilitate identity verification, fraud prevention, and decentralised lending and borrowing platforms.
- **Voting and governance:** Blockchain can enhance the transparency, security, and integrity of voting systems. It enables verifiable and tamper-proof voting records, reducing the risk of fraud and ensuring fair and trustworthy elections.
- **Intellectual property:** Blockchain technology can be used to establish proof of ownership and protect intellectual property rights. It enables creators to securely register and manage their copyrights, patents, and trademarks, reducing infringement and facilitating licensing and royalty payments.

Pros and cons of blockchain


Even with all of its complexity, blockchain has virtually limitless potential as a decentralised method of record-keeping. Blockchain technology


may very well have uses beyond those mentioned above, ranging from increased user privacy and heightened security to lower processing costs and fewer mistakes. There are a few downsides, though.

Pros

- **Enhanced accuracy:** Blockchain eliminates the need for human involvement in verification processes, leading to improved accuracy.
- **Cost reductions:** By eliminating the necessity for third-party verification, blockchain reduces operational costs.
- **Decentralisation:** Its decentralised nature makes tampering with blockchain data more challenging.
- **Secure and efficient transactions:** Blockchain ensures secure, private, and efficient transactions.
- **Financial inclusion:** It serves as a banking alternative and provides a means to secure personal information, particularly beneficial in countries with unstable or underdeveloped governments.

Cons

- **Technology costs:** Some blockchains incur significant technology expenses.
- **Low transaction throughput:** Blockchains often exhibit low transactions per second.
- **History of illicit use:** Blockchain has been associated with illicit activities, including those on the dark web.
- **Regulatory variations:** Regulation of blockchain varies by jurisdiction and remains uncertain.
- **Data storage constraints:** Blockchains may encounter limitations in data storage capacity. 

 **By: Rajnesh Devi**

The author is an assistant professor at the Yamuna Group of Institutions.

How to Use Genetic Algorithms to Build Efficient Machine Learning Models

Genetic algorithms (GAs) are optimising various stages of a machine learning pipeline, focusing on data preparation and model tuning. By employing GAs, we can automate labour-intensive steps, including handling missing data, feature engineering, and hyperparameter optimisation. This step-by-step guide offers an end-to-end blueprint for building more robust and efficient machine learning models to maximise the value extracted from data.



In the age of Big Data, the amount of information we can collect and analyse is unprecedented. While this provides incredible opportunities for learning and growth, it also presents a challenge: How do we make the most out of this vast sea of data? Merely collecting data isn't enough; what makes the difference is how efficiently we can process and analyse it. This is where genetic algorithms (GAs) come into play.

Genetic algorithms are optimisation heuristics based on the principles of natural selection. They offer a way to find good solutions to complex problems, and in the context of machine learning, they can help us to fine-tune models for better performance and more effective data utilisation.

Further on, we'll explore how genetic algorithms can be employed to make your data work harder for you. From data preparation to model selection, we'll look at how GAs can enhance each step of the machine learning pipeline.

Data-centric approach in machine learning

We live in times of data overload. That means we have to take a data-centric approach in machine learning. While algorithms and models often take the spotlight, the quality and efficiency of the data being fed into these models are just as critical—if not more so. Optimising the algorithms alone won't yield the desired results if the data itself isn't optimised. It's akin to trying to make a delicious meal; even the best chefs can't produce a culinary masterpiece with subpar ingredients.

So, what does it mean for data to be 'efficient'? Efficiency in this context refers to maximising the useful information that can be extracted from a given set of data. This could involve eliminating redundant features, fine-tuning hyperparameters to better suit the specific data set, or even selecting a machine learning model that's particularly well-suited for the data you have.

Here is where genetic algorithms can add value. By helping us automate the process of feature selection, hyperparameter tuning, and even model selection, GAs can play an instrumental role in making your data more effective.

Understanding genetic algorithms

Before we delve into the application of genetic algorithms in data optimisation, it's essential to have a fundamental grasp of what they are and how they work. Originating from the natural processes of biological evolution, genetic algorithms work on the principles of selection, crossover (or recombination), and mutation.

Selection: This is the process of choosing the fittest individuals from a population to act as parents for the next generation. In machine learning, this could mean selecting the models that produce the best results on a given data set.

```
def select_parents(population, fitness):
    # Select two parents based on their fitness scores
    return sorted(zip(population, fitness), key=lambda x:
x[1])[-2:]
```

Crossover: Once the parents are selected, the next step is to combine their traits to create offspring. In the context of machine learning, this could involve mixing the hyperparameters of two well-performing models.

```
def crossover(parent1, parent2):
    # Perform crossover between two parents
    crossover_point = len(parent1) // 2
    child = parent1[:crossover_point] + parent2[crossover_
point:]
    return child
```

Mutation: This introduces small changes in the offspring, adding some level of randomness and diversity. In machine learning, a mutation might be a slight change in a hyperparameter value or a feature's weight.

```
import random

def mutate(child):
    # Apply mutation to a child
    mutation_point = random.randint(0, len(child) - 1)
    child[mutation_point] = random.uniform(0, 1)
    return child
```

The power of genetic algorithms lies in their ability to optimise complex functions efficiently, making them a valuable tool for enhancing data utility in machine learning models.

Data preparation

Before feeding data into a machine learning model, it's crucial to ensure that it's well-prepared and clean. Data preparation involves multiple steps, such as handling missing values, normalisation, and feature engineering. These steps aim to improve the model's performance by enhancing the data's quality.

Genetic algorithms can offer an automated way to tackle these data preparation challenges. Instead of manually picking features or trying various normalisation techniques, GAs can be programmed to explore a range of options to find the most efficient data preparation strategy.

Handling missing values

```
from sklearn.impute import SimpleImputer
import numpy as np
```

```
def handle_missing_values(data, strategy='mean'):
    imputer = SimpleImputer(strategy=strategy)
    return imputer.fit_transform(data)
```

Feature engineering

```
def feature_engineering(data, selected_features):
    return data[:, selected_features]
```

Normalisation

```
from sklearn.preprocessing import MinMaxScaler
```

```
def normalise(data):
    scaler = MinMaxScaler()
    return scaler.fit_transform(data)
```

By employing genetic algorithms in these preparatory steps, you can optimise your data set for the most effective machine learning outcomes.

Applying genetic algorithms to model tuning

Once the data is prepared, the next critical step is model selection and tuning. Machine learning offers a plethora of algorithms to choose from, each with its own set of hyperparameters. The number of possible combinations can be overwhelming, but genetic algorithms can help narrow down the choices to the most effective ones.

Model selection

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
```

```
def select_model(model_type):
    if model_type == 'RandomForest':
        return RandomForestClassifier()
    elif model_type == 'SVM':
        return SVC()
```

Hyperparameter tuning

```
def tune_hyperparameters(model, hyperparameters):
    model.set_params(**hyperparameters)
    return model
```

Fitness function

```
from sklearn.metrics import accuracy_score

def fitness_function(model, X_train, y_train, X_test, y_test):
    model.fit(X_train, y_train)
    predictions = model.predict(X_test)
    return accuracy_score(y_test, predictions)
```

Genetic algorithms can automate the selection and tuning process by exploring the model and hyperparameter space efficiently. The GA will evaluate the performance of each candidate solution (combination of model and hyperparameters) using a fitness function — in this case, the model's accuracy score.

The end-to-end pipeline

The ultimate goal is to bring all these individual pieces into a coherent whole—an end-to-end pipeline that takes raw data and outputs an optimised machine learning model. In this pipeline, genetic algorithms play a pivotal role in automating multiple steps, from data preparation to model tuning.

End-to-end pipeline

```
from sklearn.model_selection import train_test_split

def end_to_end_pipeline(raw_data, target, model_
type='RandomForest'):

    # Step 1: Data preparation

    clean_data = handle_missing_values(raw_data)
    normalized_data = normalise(clean_data)

    # Step 2: Feature selection

    X_train, X_test, y_train, y_test = train_test_
split(normalized_data, target, test_size=0.2)
    selected_features = [i for i in range(len(X_train[0]))]
```

```
# Placeholder, would be determined by GA

    # Step 3: Model selection and tuning

    model = select_model(model_type)
    hyperparameters = {} # Placeholder, would be determined
by GA
    tuned_model = tune_hyperparameters(model,
hyperparameters)

    # Step 4: Evaluate fitness

    fitness = fitness_function(tuned_model, X_train[:, selected_
features], y_train, X_test[:, selected_features], y_test)

    return fitness

# Example usage

raw_data = np.random.rand(100, 10) # 100 samples, 10
features
target = np.random.randint(0, 2, 100) # Binary target
variable

fitness = end_to_end_pipeline(raw_data, target)
```

This is a simplified example, but it gives you a blueprint for constructing an end-to-end pipeline that employs genetic algorithms at every key stage. This ensures that you're extracting the most value from your data at each step of the machine learning process.

From automating the tedious process of data preparation to fine-tuning machine learning models, genetic algorithms provide an efficient, automated approach to optimise the entire data pipeline. By leveraging these algorithms, we're not just simplifying the model development process but also ensuring that the highest quality insights are gleaned from our data. Whether dealing with large-scale data sets, multi-dimensional features, or diverse machine learning models, genetic algorithms equip you with the versatility to handle a broad array of data challenges. As a result, they become an indispensable asset in any data scientist's toolkit for crafting robust and effective solutions. **END** 🐧

By: Mir H.S. Quadri

The author is a research analyst with a specialisation in artificial intelligence and machine learning. He is the founder of Arkinfo, which focuses on the research and development of tech products using new age technologies. He shares a deep love for the analysis of technological trends and understanding their implications. Being a FOSS enthusiast, he has contributed to several open source projects.

Python: How it Helps Run IoT Devices

Python's strength lies in its English-like syntax, the nimble MicroPython, and an extensive library of modules. It's a great tool for prototyping, developing, and operating various IoT devices.



IoT, or the Internet of Things, is a technology we encounter daily. It encompasses various applications, including automatic lights controls based on presence, fan speed and AC temperature adjustments according to outdoor climate conditions, self-driving vehicles, surveillance and security systems, smart electricity meters that send monthly bills directly to your email, production lines, real-time health monitoring, and more.

IoT is defined as a network of electronic devices equipped with sensors, actuators, or accessories that possess processing capabilities through software. These devices can connect, control and exchange data with other systems over the internet or other communication networks in real-time, enhancing decision making processes.

Why Python

Python is widely used in IoT due to its simplicity and versatility. Here are some key reasons for its popularity.

- **Easy to learn and use:** Python's syntax is simple, like the English language. Also, there is a lighter version

of Python called MicroPython making it a great fit for small computing resources. MicroPython is a lean and efficient Python 3 programming language implementation that includes a small part of the Python standard library, and is optimised to work on microcontrollers and in limited contexts.

- **Cross-platform compatibility:** Python can run on multiple operating systems such as Linux, Windows, MacOS, or Raspberry Pi ensuring compatibility. In most cases, Python is preinstalled.
- **Existing large community:** Python boasts a substantial user community actively contributing to the development of tools and support for IoT advancements.
- **Interoperability:** Python seamlessly integrates with other languages and protocols like MQTT, HTTP, BLE, and more.
- **Scalability:** Python's modular architecture suits both large and small-scale IoT applications.
- **Extensive library support:** Python offers a vast array of tools for machine learning, deep learning, data intensive applications, data analysis, data inference, visualisation, and more. Its huge set of development libraries, tools and frameworks makes development faster and compatible with IoT.
- **Open source:** Python is an open source framework, readily available for download at no cost.

Using Python for IoT

Python offers various modules for IoT programming, facilitating development at various stages. Let us explore around 30 modules that assist in IoT development.

- **IoT devices**
Arduino and Raspberry Pi are the two most popular controller boards for hardware projects.
 - **Raspberry Pi**
 - MicroPython is a full Python 3 implementation that works directly on embedded hardware such as the Raspberry Pi Pico.
 - It provides an interactive prompt (the REPL) and a built-in file system, as well as the ability to run commands directly through USB serial.

- MicroPython's Pico port offers modules for interacting with low-level chip-specific hardware.
- Detailed documentation can be found at <https://www.raspberrypi.com/documentation/microcontrollers/micropython.html>
- **Arduino**
 - Arduino is an open source platform comprising hardware and software, allowing quick production of interactive electronics projects.
 - Arduino employs its own programming language, akin to C++; however, for applications that require integration with sensors and other physical devices, Arduino can be used with Python.
 - The Firmata protocol is the standard method to control Arduino from PC. It enables serial communication with microcontrollers from software on a host computer, allowing it to receive digital and analogue inputs and transmit data to digital and analogue outputs.
 - To get started with Arduino and Python, visit <https://realpython.com/arduino-python/>
- **IoT system and sensors interaction**
 - **PySensors:** Use 'pip install PySensors' to install it.
 - lm-sensors (Linux monitoring sensors) is a free and open source application that provides temperature, voltage, and fan monitoring tools and drivers.
 - The PySensors Python module is used for ctypes bindings for *libsensors.so* from the lm-sensors project.
 - Reference: <https://pypi.org/project/PySensors/>
 - **esptool:** Install using 'pip install esptool'.
 - esptool is a Python-based, open source, platform-independent software for communicating with the Espressif SoC's ROM bootloader.
 - The toolbox for working with Espressif chips includes *esptool.py*, *espefuse.py*, and *espsecure.py*.
 - They can, for example, accomplish these tasks:
 - Binary data saved on flash can be read, written, erased, and verified.
 - Can read chip characteristics and other data relating to the chip, such as the MAC address or flash chip ID.
 - The one-time-programmable efuses can be read and written.
 - Make binary executable images that are ready for flashing.
 - Binary images can be analysed, assembled, and merged.
 - Reference: <https://pypi.org/project/esptool/>
 - **pyusb:** Install with 'pip install pyusb'.
 - pyusb simplifies communication with USB devices in Python.
 - It should function without any additional code in any Python >= 3.6 environment with ctypes and a pre-built USB backend library.
- **pyserial:** Use 'pip install pyserial' to install it.
 - pyserial is a Python library responsible for serial port access.
 - It provides the same class-based interface for Python running on Windows, OSX, Linux, BSD (potentially any POSIX-compliant system), and IronPython on all supported platforms.
 - This allows it to communicate serially with devices like voltmeters, oscilloscopes, strain gauges, flow meters, actuators, and lights.
 - Reference: <https://pypi.org/project/pyserial/>
- **pybluez:** Install with 'pip install pybluez2'.
 - This module is responsible for serial port access.
 - It provides the same class-based interface for Python running on Windows, OSX, Linux, BSD (potentially any POSIX-compliant system), and IronPython on all supported platforms.
 - Reference: <https://pypi.org/project/pybluez2/>
- **gpiozero:** Install via 'pip install gpiozero'.
 - gpiozero provides a straightforward interface to GPIO devices on the Raspberry Pi.
 - Component interfaces are provided to make getting started with physical computing as simple as possible.
 - One may rapidly connect components together with very little code.
 - Reference: <https://pypi.org/project/gpiozero/>
- **mraa**
 - mraa is a Python-compatible structural GPIO library.
 - Eclipse Mraa (Libmraa) is a C/C++ library with Java, Python, and JavaScript bindings for interacting with I/O pins and buses on various IoT and edge systems, providing an organised and reasonable API, where port names/numbering match the board that you are on.
 - The use of Libmraa does not bind you to any particular hardware. Since board detection is done at runtime, you can write portable code that works on all supported platforms. The most appealing aspect of this is that there is only one library for each individual device.
 - The goal is to make it easy for developers and sensor manufacturers to map their sensors and actuators on top of compatible hardware, and to allow high level languages and structures to govern low level communication protocols.
 - Reference: <https://github.com/eclipse/mraa/tree/master/examples/python>
- **IoT communication protocols**
 - **sockets:** Install using 'pip install sockets'.
 - TCP/IP and UDP are communication transport layer protocols.

- sockets package facilitates TCP/IP and UDP communication, which are used to enable networking in IoT devices.
- Reference: <https://pypi.org/project/sockets/>
- *paho-mqtt*: Install using 'pip install paho-mqtt'.
 - The MQTT or Message Queue Telemetry Transport protocol is a publish/subscribe messaging transport that is intended for machine-to-machine (M2M)/Internet of Things connectivity, enabling high speed data exchange with low payload.
 - When a minimal code footprint is necessary or bandwidth on the network is at a premium, it is beneficial for communications with faraway locations.
 - Applications can connect to a MQTT broker to publish messages, subscribe to topics, and receive published messages thanks to the client class provided by the Paho library code.
 - Additionally, it offers certain assistance functions to make it simple to publish one-time messages to MQTT servers.
 - Reference: <https://pypi.org/project/paho-mqtt/>
- *smtplib*
 - This protocol handles the sending and routing of emails between mail servers.
 - It is part of Python's standard library.
 - Reference: <https://docs.python.org/3/library/smtplib.html>
- *asyncio*: Install with 'pip install asyncio'.
 - The asyncio module provides architecture for building single-threaded concurrent programming with co-routines, multiplexing I/O access through sockets and other resources, running network clients and servers, and other primitives.
 - Reference: <https://pypi.org/project/asyncio/>
- **IoT database**
 - *MySQLdb*: Install with 'pip install MySQL-python'.
 - MySQLdb is a popular go-to relational database that aids in the building of remote storage for IoT systems.
 - Reference: <https://pypi.org/project/MySQL-python/>
 - *SQLite*
 - SQLite is a C library that provides a lightweight disk-based database that does not require a separate server process and may be accessed using a non-standard variation of SQL.
 - SQLite can be used by IoT to store internal data.
 - The sqlite3 module does not require installation. Since Python 2.5, it has been included in the standard library.
 - Reference: <https://docs.python.org/3/library/sqlite3.html>
- **IoT graphical user interface**
 - *Tkinter*
 - Python provides numerous GUI (graphical user interface) development choices. Tkinter is the most commonly used of all the GUI techniques.
 - It is a standard Python interface to the Python-supplied Tk GUI toolkit.
 - Python with Tkinter is the quickest and most straightforward approach to construct GUI apps.
 - Reference: <https://docs.python.org/3/library/tkinter.html>
- **IoT backend**
 - *Flask*: Install via 'pip install Flask'.
 - Flask is a simple micro framework for building web applications, to build http requests.
 - IoT applications require a web-based interface in order to provide monitoring and controlling devices.
 - Reference: <https://pypi.org/project/Flask/>
 - *websockets*: Install using 'pip install websockets'.
 - websockets is a Python package for creating WebSocket servers and clients. It prioritises consistency, simplicity, resilience, and performance.
 - The WebSocket API is a cutting-edge technology that allows a two-way interactive communication session to be established between a user's browser and a server. This API sends messages to a server and receives event-driven responses without polling the server for a response.
 - Reference: <https://pypi.org/project/websockets/>
 - *requests*: Install with 'pip install requests'.
 - The requests module allows sending HTTP requests; the HTTP request generates a 'response object' containing response data (content, encoding, status, and so on).
 - Supports multipart file uploads, streaming downloads, and connection pooling.
 - Enables automatic decoding and decompression of content.
 - Reference: <https://pypi.org/project/requests/>
 - *aiohttp*: Install with 'pip install aiohttp'.
 - aiohttp provides an asynchronous HTTP client or server framework, including websockets support.
 - Reference: <https://pypi.org/project/aiohttp/>
 - *pushsafer*: Install with 'pip install python-pushsafer'.
 - pushsafer can send and receive push notifications in real-time to iOS, Android, and Windows devices (mobile and desktop), as well as browsers such as Chrome, Firefox, Opera, and others.
 - Reference: <https://pypi.org/project/python-pushsafer/>
- **IoT data analysis and visualisation**
 - *Numpy*: Install via 'pip install numpy'.

- Numpy package helps to power N-dimensional array computing.
- It's often used in IoT to read sensor bulk data from the system's database inherent functions.
- Reference: <https://pypi.org/project/numpy/>
- **Pandas:** Install with 'pip install pandas'.
 - Pandas is a Python module that provides quick, versatile, and expressive data structures that are intended to make working with 'relational' or 'labelled' data simple and natural.
 - It aspires to be the basic high-level building block for doing realistic, real-world data analysis in Python.
 - Reference: <https://pypi.org/project/pandas/>
- **Matplotlib:** Install using 'pip install matplotlib'.
 - Matplotlib is a comprehensive library for creating static, animated, and interactive visualisations in Python.
 - Reference: <https://pypi.org/project/matplotlib/>
- **JSON**
 - JSON is a data storage and exchange syntax.
 - It has been written in JavaScript object notation.
 - Python includes a library named json that may be used to work with JSON data.
 - Reference: <https://docs.python.org/3/library/json.html>
- **IoT machine learning and deep learning (AI)**
 - **Tensorflow:** Install with 'pip install tensorflow'.
 - TensorFlow is a high-performance numerical computing open source software package.
 - Its adaptable architecture enables simple computing deployment over a wide range of platforms (CPUs, GPUs, TPUs), from PCs and clusters of servers to mobile and edge devices for handling non-linear data sets.
 - Reference: <https://pypi.org/project/tensorflow/>
 - **Opencv:** Install via 'pip install opencv-python'.
 - OpenCV is a large open source library for computer vision, machine learning, and image processing, and it currently plays an essential part in real-time operations.
 - Reference: <https://pypi.org/project/opencv-python/>
 - **Pycaret:** Install using 'pip install pycaret'.
 - PyCaret is an open source, low-code machine learning library in Python.
 - When compared to other open source machine learning libraries, PyCaret is a low-code library that may be used to replace hundreds of lines of code with only a few lines. Experiments become significantly faster and more efficient as a result, making it apt to use for IoT apps.
 - Reference: <https://pypi.org/project/pycaret/>
 - **LightGBM:** Install using 'pip install lightgbm'.
 - LightGBM is an abbreviation for Light Gradient Boosting Machine. It provides plenty of memory-

efficient yet quick computational power and is capable of handling massive volumes of data.

- Reference: <https://pypi.org/project/lightgbm/>
- **IoT cloud integration**
 - **Azure-CLI:** Install using 'pip install azure-cli'.
 - Azure-CLI provides Python tools for Azure.
 - Reference: <https://pypi.org/project/azure-cli/>
 - **AWS:**
 - Client devices can interface with AWS IoT and AWS IoT Greengrass core devices using the Python programming language by using the AWS IoT Device SDK for Python.
 - Reference: <https://docs.aws.amazon.com/greengrass/v1/developerguide/IoT-SDK.html>

Challenges of using Python for IoT

Python has many advantages, but no programming language is flawless.

- Python is slower than other programming languages because it is dynamically typed and run line by line.
- Python's key limitations are its runtime speed, memory management, and concurrency and parallelism support. When compared to other languages, the runtime speed is poor. Due to the amount of memory Python needs, projects with many objects active in RAM may encounter challenges when using Python.
- While Python is a good server-side language, it is rarely used on the client-side as working with mobile applications and apps is slow and inconvenient.

IoT is evolving, and the choice of tools to meet its demands is crucial. Several programming languages have proven effective in IoT development. Python has shown itself to be a valuable tool for prototyping, developing, and operating various IoT devices and systems. Its rapid development speed, minimal learning curve, and extensive library set make it indispensable for IoT. **END** 🐧

References

- <https://www.datamation.com/wp-content/uploads/2021/04/IOT-4.png>
- <https://www.deepeatdev.com/en/blog/iot-with-python-reasons-to-use-it/>
- <https://edu.varistor.in/python-in-iot/>
- <https://www.oodlestechnologies.com/blogs/role-of-python-in-iot-development/>
- <https://svitla.com/blog/internet-of-things-with-python>

By: Thangaselvi Arichandrapandian

The author works in a leading bank as AVP. She is a perennial learner and loves the quote: "The more I learn, the more I realise how much I don't know." -Albert Einstein

Bun:

Ushering in a New Era of JavaScript Runtime



JavaScript development demands speed, and Bun delivers with lightning-fast performance, native TypeScript support, and streamlined workflows. But it is still work in progress.

In the fast-paced and ever-evolving landscape of JavaScript development, a new and promising player has recently made its debut: Bun. This all-in-one toolkit is poised to transform the way JavaScript and TypeScript applications are built, tested, and deployed. It goes beyond being just a runtime environment – it combines the roles of a package manager, a bundler, and a test runner into one cohesive package.

Bun's primary objective is to simplify the development process, ushering in an era of smoother and more efficient workflows for developers. Its creators aim to provide a solution that is not only faster and leaner but also aligns more closely with modern development practices.

In doing so, Bun aspires to serve as a compelling alternative to Node.js, which has long been the go-to choice for developers seeking top-notch performance, versatility, and robust community support.

As we explore Bun further, we'll delve into its key features, its potential impact on the JavaScript ecosystem, and how it aims to meet the evolving demands of today's web and application development.

Why Bun matters: A new chapter in JavaScript runtime history

The world of JavaScript development is ever-advancing, and even in the presence of well-established tools like Node.js, there is always room for innovation and progress. Bun has

generated quite a buzz in the developer community. But what sets it apart from its predecessor and why was it developed in the first place? Here are the claims made by Bun developers.

- **Performance that packs a punch:** One of the key motivations behind creating Bun was to achieve superior performance. By leveraging the JavaScript core engine, optimised for faster startup times, Bun has been known to outperform Node.js. In specific scenarios, it has demonstrated a remarkable performance improvement, being up to ten times faster than its predecessor. This is no small feat and underscores Bun's commitment to delivering a lightning-fast runtime experience.

- **A comprehensive toolkit for streamlined development:** Bun is not merely a runtime; it's a comprehensive toolkit that covers all aspects of the development process. It offers an all-in-one approach, encompassing a package manager, bundler, and test runner. This integrated solution simplifies and streamlines development, eliminating the need to juggle multiple tools and processes. Bun's developers have recognised that an efficient workflow is crucial for modern web and application development.
- **Native TypeScript support:** One distinctive feature that sets Bun apart from Node.js is its native support for TypeScript (TS). Unlike Node.js, which may require third-party transpilers or complex configurations, Bun embraces TypeScript natively. This means that developers working with TypeScript can do so seamlessly and efficiently, without additional layers of complexity.
- **Expanding possibilities with .jsx and .tsx files:** Another remarkable aspect of Bun is its support for .jsx and .tsx files. This feature enables Bun to natively convert HTML and markup into JavaScript, which is a functionality not readily available in Node.js.
- **Efficiency meets memory management:** While Node.js is known for its minimal memory footprint, Bun has taken memory efficiency a step further. It's designed to be even more memory-efficient, addressing the needs of developers who demand optimal resource utilisation.
- **Interoperability and compatibility:** Bun aims for seamless compatibility with Node.js, making the transition from one to the other as smooth as possible. This allows developers to experiment with Bun without disrupting their existing workflows or investments in Node.js-based projects.

'Baking' Bun for your JS project

Getting started with Bun is as simple as running the 'bun init' command. This command scaffolds an empty project, providing developers with a ready-to-go environment. However, before that, we need to install Bun.

To install Bun on Linux/macOS/WSL you can use curl:

```
curl -fsSL https://bun.sh/install | bash
```

Bun also has a Docker image; to use it, simply run:

```
docker pull oven/bun
docker run --rm --init --ulimit
memlock=-1:-1 oven/bun
```

Interestingly, you can even install it using its competitor npm:

```
npm install -g bun
```

If you're a Windows user, note that at the time of writing this article, Bun only has an experimental build for Windows, which contains the Bun

runtime only. However, this is expected to change in the near future.

You can check the installation using the command:

```
bun --version
```

Writing a simple HTTP web server in Bun

Let's understand Bun by writing a simple HTTP web server in it. To create a Bun project, we use the *init* command as follows. You can also refer to Figure 1 for clarity.

```
mkdir OSFY-bun
cd OSFY-bun
bun init
```

Here, I created a directory named 'OSFY-bun' and executed the *init* command there. This is essentially a setup command that asks for multiple questions related to our project. When naming the entry point (which is the main file we are going to execute), Bun accepts both .ts and .js for TypeScript and JavaScript, respectively. In the

```
sreyas@de4bian: ~/Desktop/OSFY-bun
sreyas@de4bian:~/Desktop/OSFY-bun$ bun init
bun init helps you get started with a minimal project and tries to guess sensible defaults. Press
^C anytime to quit

package name (osfy-bun): osfy-bun
entry point (index.ts): index.js

Done! A package.json file was saved in the current directory.
+ index.js
+ .gitignore
+ jsconfig.json (for editor auto-complete)
+ README.md

To get started, run:
  bun run index.js
sreyas@de4bian:~/Desktop/OSFY-bun$ bun run index.js
Hello via Bun!
sreyas@de4bian:~/Desktop/OSFY-bun$
```

Figure 1: Creating a Bun project

```
index.js - OSFY-bun - Visual Studio Code
File Edit Selection View Go Run Terminal Help
EXPLORER
OSFY-BUN
  node_modules
  .gitignore
  bun.lockb
  JS index.js
  jsconfig.json
  package.json
  README.md
JS index.js
1 console.log("Hello via Bun!");
2
```

Figure 2: Simple 'hello world' program

example, I used `index.js`, making my project a JavaScript project. Opening the code in `index.js` we can see a simple 'hello world' program written in JS (Figure 2).

Writing an HTTP server is straightforward using Bun's native functions. We can write a simple one like this (Figure 3):

```
-- console.log("Hello via Bun!");
1+ Bun.serve({
2+     fetch(request) {
3+         return new
Response("Hello OSFY readers!");
4+     },
5+ })
```

To execute this, simply type:

```
bun run index.js
```

And go to `http://localhost:3000` to view the result as shown in Figure 4.

Instead of re-running the `run` command, Bun also provides an option for hot reloading, so that any changes in the file are updated instantaneously. This is achieved by:

```
bun --hot index.js
```

Benchmarking Bun and Node

Bun's selling point is speed, so I had to test it for myself. I used a simple 'hello world' program in JS as the test subject. Also, I used a hyperfine command-line tool to benchmark Bun's performance against Node's (Figure 5).

Not surprisingly, Bun ran almost five times faster than Node. Additionally, Bun's package manager installs 30 times faster than Node's npm.

So should I go for Bun?

While Bun shows considerable promise, it is not without its share of limitations and considerations. One of the primary concerns regarding Bun is its stability. As an emerging technology, it is still maturing and may exhibit occasional instability. This may deter

```
JS index.js > ...
1  - console.log("Hello via Bun!");
1+ Bun.serve({
2+     fetch(request) {
3+         return new Response("Hello OSFY readers!");
4+     },
5+ })
2  6
```

Figure 3: HTTP server



Figure 4: Server response in the browser

```
sreyas@de4bian: ~/osfy-bun-test
sreyas@de4bian:~/osfy-bun-test$ hyperfine "bun index.js" "node index.js"
Benchmark 1: bun index.js
Time (mean ± σ): 21.5 ms ± 4.9 ms [User: 11.6 ms, System: 4.6 ms]
Range (min ... max): 14.8 ms ... 30.9 ms 111 runs

Benchmark 2: node index.js
Time (mean ± σ): 98.9 ms ± 11.6 ms [User: 53.0 ms, System: 26.8 ms]
Range (min ... max): 85.5 ms ... 128.5 ms 28 runs

Summary
'bun index.js' ran
4.60 ± 1.18 times faster than 'node index.js'
sreyas@de4bian:~/osfy-bun-test$ S
```

Figure 5: Benchmarking Bun

some developers from adopting it on a widespread basis, as they seek robust and reliable runtime environments for their projects. Also, Bun's compatibility with existing Node.js modules is not seamless. It faces hurdles in achieving full compatibility with the extensive array of modules in the Node.js ecosystem. This presents an impediment for developers who wish to transition to Bun seamlessly without encountering compatibility issues. Another evolving aspect of Bun is its support for Windows-based systems. Currently, it may not provide the same level of compatibility and performance for Windows users as it does for other

platforms. This is a consideration for developers who work primarily within the Windows environment.

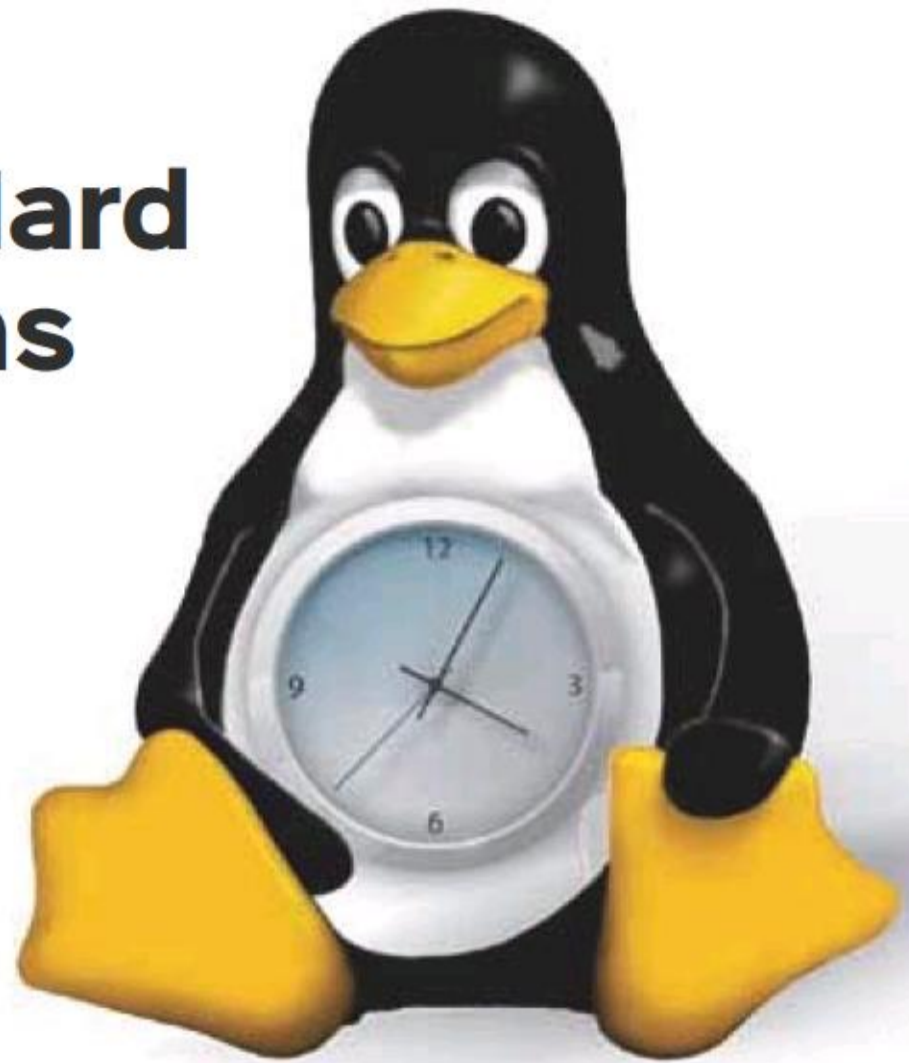
Given its recent introduction to the JavaScript scene, Bun has yet to gain widespread adoption compared to more established counterparts, such as Node.js and Deno. Consequently, it may not be the first choice for applications that demand proven reliability and broad industry acceptance. However, its potential is promising. Bun is open source and you can check out its development on its GitHub repository at <https://github.com/oven-sh/bun>. **END** 🐧

By: S. Sreyas

The author is a skilled programmer who has gained expertise in the field through self-learning. He has made significant contributions to various open source projects, with a particular focus on reverse engineering, compiler design, and system level programming. His contributions are available on GitHub.

Real-Time Linux Applicability for Hard Real-Time Systems

Let's get an in-depth understanding of real-time systems, analysing the real-time characteristics of two prominent operating systems: Linux with PREEMPT_RT patch and QNX. By examining their real-time characteristics, determinism, performance, and features, developers can get an idea of how the real-time behaviours of Linux with PREEMPT_RT patch compare with QNX.



Real-time operating systems (RTOS) play a critical role in the development of hard real-time systems and meeting the stringent timing requirements of real-time applications. QNX (Quantum Neutrino eXecutive) stands out as a preferred choice for hard real-time systems, such as medical systems, due to its real-time capabilities and features.

Linux, in its standard form, is not considered a hard real-time operating system. However, efforts have been made to make Linux more suitable for real-time applications. Linux with PREEMPT_RT patch attempts to enhance the real-time capabilities of the standard Linux kernel by adding specific real-time scheduling policies, reducing interrupt latency, and providing mechanisms for prioritising time-critical tasks. These improvements have made Linux with PREEMPT_RT patch more predictable and responsive.

Business challenges

A real-time operating system is specialised, designed to meet the stringent timing requirements of hard

real-time systems. These systems often have critical applications in industries such as aerospace, automotive, industrial automation, and medical devices. Nowadays, medical industries are migrating from (proprietary) RTOS such as QNX to Linux (open source) drawn by licence, community support, security, and cost considerations. Linux with strong real-time capabilities is required for this scenario, and for understanding how PREEMPT_RT patch improves real-time capabilities.

Figure 1 lists some business challenges associated with using RTOS.

Problem statement

RTOS like QNX are used and deployed in real-time systems in

critical industries, such as medical devices. Industries considering migrating from RTOS to Linux need a detailed understanding of Linux with the PREEMPT_RT patch -- its real-time capabilities, behaviour, and features. Additionally, a comparison between Linux with PREEMPT_RT patch and QNX is necessary. The problem is that this information and comparison results are not available. This article will address this issue comprehensively. We also do not know how to use PREEMPT_RT patch and the real-time behaviours, features and latency reduction offered by it. This issue will be addressed here with PREEMPT_RT patch guidelines and case studies.

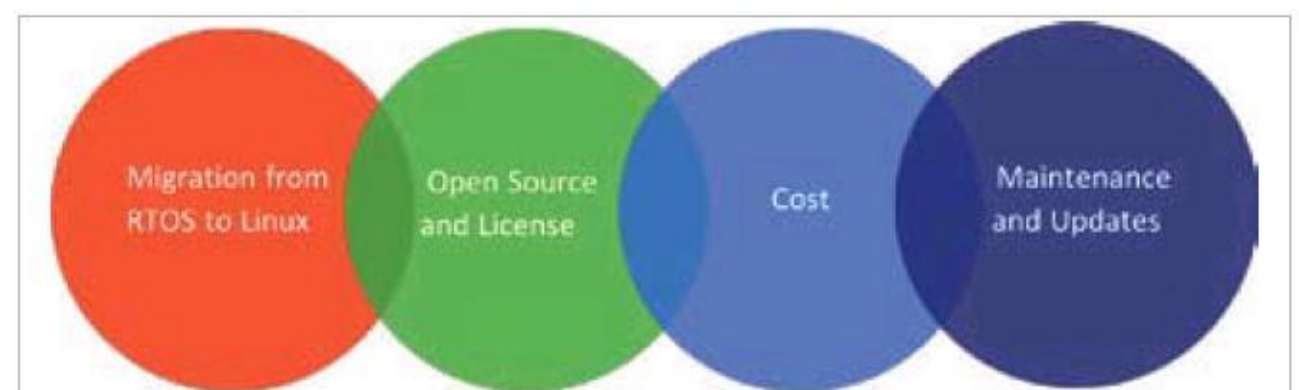


Figure 1: Business challenges

Real-time systems

Real-time systems are computer systems designed to respond to input and produce output within a specified timeframe, known as a deadline.

The primary characteristic that distinguishes real-time systems from other computer systems is their ability to guarantee timely and predictable responses to events or tasks. These systems are widely used in various domains, including industrial control, automotive systems, aerospace, medical devices, multimedia, and telecommunications.

There are two main types of real-time systems.

Hard real-time systems: Meeting deadlines is crucial in hard real-time systems. Failure to complete a task within its specified deadline can lead to dangerous consequences, system failure, or compromise safety. Examples include flight control systems in aeroplanes, medical devices for life support, and airbag deployment in vehicles.

Soft real-time systems: In soft real-time systems, meeting deadlines is important, but occasional misses may be tolerable without dangerous consequences. While timely execution is desirable, the system can still provide useful results even if some deadlines are missed. Examples of soft real-time systems include streaming multimedia applications, online video conferencing, and online gaming. The user experience may degrade if deadlines are missed, but it doesn't lead to system failure or severe safety issues.

Apart from these main types, there is a middle ground known as 'firm real-time systems', a hybrid between hard and soft real-time systems. Firm real-time systems prioritise meeting most deadlines, but occasional misses are allowed if they don't significantly impact the overall system behaviour. These systems are often used in multimedia streaming and telecommunications applications.

QNX

QNX is a real-time operating system known for its strong real-time characteristics, reliability, and low and deterministic latencies. Originally developed by QNX Software Systems, it is now a part of BlackBerry Limited and is widely used across various industries and applications that require real-time performance, determinism, and high reliability. Its characteristics are:

Deterministic behaviour: QNX is designed to be highly deterministic, guaranteeing that critical tasks and processes meet their deadlines consistently. This attribute is particularly crucial in safety-critical applications, such as automotive systems, industrial automation, and medical devices.

Real-time scheduling: QNX employs a priority based preemptive scheduler, allowing tasks with higher priorities to preempt lower-priority tasks. This scheduling mechanism ensures that time-critical tasks receive immediate attention without interference from lower-priority tasks.

Fault tolerance and reliability: QNX is designed with a strong focus on fault tolerance and reliability. Its microkernel architecture isolates critical components, reducing the impact of failures and enhancing system resilience.

Resource management: QNX provides efficient resource management, allowing developers to allocate CPU time, memory, and other system resources to tasks in a controlled and predictable manner. This capability is essential for meeting the timing requirements of real-time applications.

Preemption: QNX supports preemption, enabling a higher-priority task to interrupt the execution of a lower-priority task promptly.

Fast interrupt handling: QNX provides efficient interrupt handling mechanisms. When an interrupt occurs, the kernel quickly determines the appropriate handler for the interrupt and dispatches it. This minimises the time spent in the interrupt handling routine.

Latencies: QNX is known for its low and deterministic latencies, making it suitable for real-time applications with strict timing requirements. The latencies in a real-time operating system like QNX refer to the time it takes for the system to respond to an event or an interrupt and start executing a specific task. There are several factors that can influence latencies in QNX:

- Context switching
- Task scheduling
- Kernel design
- Interrupt handling

It's important to note that while QNX offers low and deterministic latencies, the actual latencies experienced in a specific system can depend on various factors, such as the hardware platform, system workload, and application design. The above techniques are used to optimise systems for low latencies and predictable behaviour.

Linux

Linux also possesses some real-time capabilities. Examples include EDF and PREEMPT_RT patch.

EDF: EDF (earliest deadline first) is a scheduling algorithm commonly used in real-time operating systems and some real-time extensions for general-purpose operating systems, including Linux. It prioritises tasks based on their deadlines, giving preference to tasks with the earliest deadlines for execution.

In standard Linux, the default scheduling algorithm is the completely fair scheduler (CFS), which is designed for fairness and responsiveness in non-real-time workloads. However, there are ways to implement EDF in Linux for real-time applications. One way to enable EDF scheduling in Linux is by using the PREEMPT_RT (real-time) patch.

PREEMPT_RT patch: The PREEMPT_RT patch is a set of patches for the Linux kernel that aim to turn Linux into a soft real-time operating system. The goal of the PREEMPT_RT patch is to improve the responsiveness and predictability of the

Linux kernel for real-time applications while retaining the general-purpose capabilities of Linux.

The main features introduced by the PREEMPT_RT patch are listed in Table 1.

When the PREEMPT_RT patch is applied to the Linux kernel and booted, the real-time features become available. Real-time tasks can be created, and use the real-time scheduling classes to achieve deterministic and timely execution.

To use real-time capabilities, you need to configure the applications to run with elevated privileges (often as root) since real-time tasks have higher scheduling priorities and require special permissions.

PREEMPT_RT patch significantly improves real-time performance; it may not be suitable for certain ultra-high-performance real-time applications where absolute timing guarantees are critical. This patch is an external addition to the Linux kernel and may not always be in sync with the latest official kernel releases. Therefore, it is crucial to carefully select the kernel version and the corresponding PREEMPT_RT patch version that are compatible and well-maintained for your application needs. Additionally, the kernel configuration options and tuning for real-time performance should be done with care and understanding of specific requirements.

Cyclictest

We take a look at a test conducted in the PREEMPT_RT patch, using the Cyclictest command. Cyclictest is a command-line utility in Linux, often used for measuring real-time performance and latency of the system. It is particularly helpful for testing and analysing the Linux kernel's ability to handle periodic tasks with strict timing requirements.

To use Cyclictest, you need to install it on your Linux system. It is often available in most distribution repositories. If it is not installed already, use the package manager specific to your distribution. For example, on Ubuntu and Debian-based systems, the following command can be used to

Feature	Description
Preemption	<ul style="list-style-type: none"> Enhances the kernel's ability to be preempted at more points during its execution This allows higher-priority tasks to interrupt lower-priority ones more quickly, reducing latencies and improving responsiveness
Kernel preemption	<ul style="list-style-type: none"> Introduces kernel preemption, which allows high-priority kernel threads to preempt lower-priority ones Improves the responsiveness of the kernel to real-time events
Priority inheritance	<ul style="list-style-type: none"> Implements priority inheritance protocols to prevent priority inversion, a situation where a higher-priority task is blocked by a lower-priority task holding a shared resource Priority inheritance ensures that the resource-holding task temporarily inherits the priority of the blocked task, preventing priority inversions
High-resolution timers	<ul style="list-style-type: none"> Provides support for high-resolution timers, which allow applications to set timers with finer granularity, making it easier to meet tight timing constraints
Real-time scheduling	<ul style="list-style-type: none"> Enhances the Linux scheduler to support real-time scheduling processes, such as SCHED_FIFO and SCHED_RR in addition to the standard CFS scheduler These policies allow developers to create real-time tasks with strict timing requirements
Throttle mechanism	<ul style="list-style-type: none"> Introduces throttle mechanism to limit the execution time of low-priority tasks, preventing them from excessively delaying high-priority real-time tasks
Interrupt handling	<ul style="list-style-type: none"> The patch optimises interrupt handling to reduce the latency associated with servicing interrupts It reduces the time spent in interrupt context, making the system more responsive to real-time events
Soft and hard real-time support	<ul style="list-style-type: none"> PREEMPT_RT patch offers both soft and hard real-time capabilities Soft real-time systems aim to meet timing constraints most of the time but allow occasional deadline misses. Hard real-time systems must meet timing constraints strictly, without any misses
Kernel locking and synchronisation	<ul style="list-style-type: none"> To ensure data consistency and prevent race conditions, the patch addresses kernel locking and synchronisation mechanisms to handle concurrent access to shared data

Table 1: PREEMPT_RT patch features

install Cyclictest:

```
sudo apt-get install rt-tests
```

Once installed, run Cyclictest with distinct options. The basic syntax is as follows:

```
cyclictest <options>
```

Here are some common options used with Cyclictest:

- t: Specifies the number of threads to use (default is 1)
- n: Specifies the number of iterations (default is 10000)
- i: Specifies the interval (in microseconds) between iterations (default is 1000µs)
- l: Specifies the maximum thread lock latency (in microseconds) before considering a result as an outlier (default is 10000µs)
- m: Specifies the CPU mask, which restricts the test to specific CPUs (default is all CPUs)
- p: Prints the raw latency results
- H: Prints the latency histogram
- S: Prints the latency statistics (mean, min, max, etc)
- q: Quiet mode; suppresses the summary and histogram outputs.

Here is a Cyclictest sample command:

```
cyclictest -p 90 -t80 -m -n 10000 -l 100000 -priospread
```

This command uses the values given below with options.

- The option -p is used for thread priority level and -p as 90.
- The option -t is used for number of threads and -t as 80.
- The option -m is used to enable *lockall*, which locks all memory used by the 'cyclictest' process into RAM.
- Number of iterations or number of test cycles -n is 10000.
- The option -l sets the length of each test cycle in microseconds and -l as 100000.

- The option -priospread is used to enable 'priority spreading'. Priority spreading is a feature in the Linux kernel that helps mitigate priority inversion issues.

- Kernel version: 6.1.35
- Hardware: Raspberry Pi 4 model B (64-bit)

Kernel configuration options for the server should be 'No Forced Preemption' as shown in Figure 2.

These changes are reflected in the .config file as shown below:

```
CONFIG_PREEMPT_NONE_BUILD=y
CONFIG_PREEMPT_NONE=y
CONFIG_PREEMPT_VOLUNTARY is not set.
CONFIG_PREEMPT is not set.
CONFIG_PREEMPT_DYNAMIC is not set.
CONFIG_SCHED_CORE is not set.
```

Kernel configurations

The kernel configurations given below enable the PREEMPT_RT patch. The latencies are calculated with all the kernel configurations as well as with the enabled PREEMPT_RT patch. The testing is done with 'cyclictest' Linux command-line utility with some required options. Hackbench load is also given with the Cyclictest command.

The kernel config options (including PREEMPT_RT Patch) are:

Kernel configuration options for the desktop should be 'Voluntary Kernel

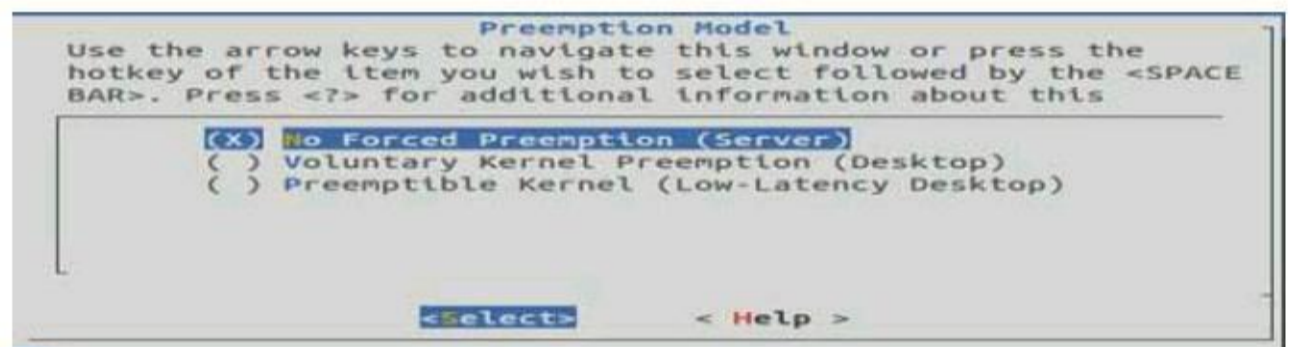


Figure 2: Kernel config options for server – 'No Forced Preemption'

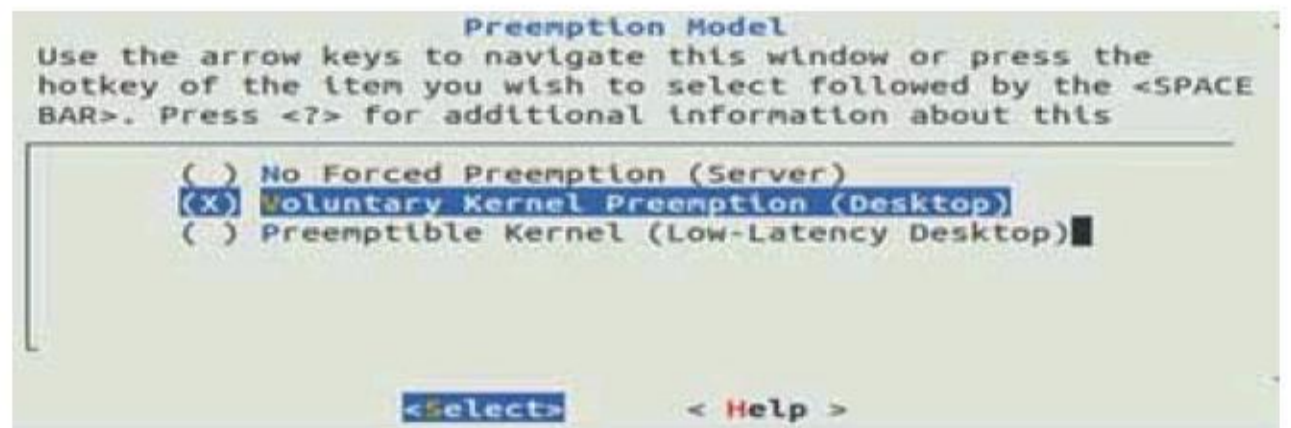


Figure 3: Kernel config options for the desktop – 'Voluntary Kernel Preemption'

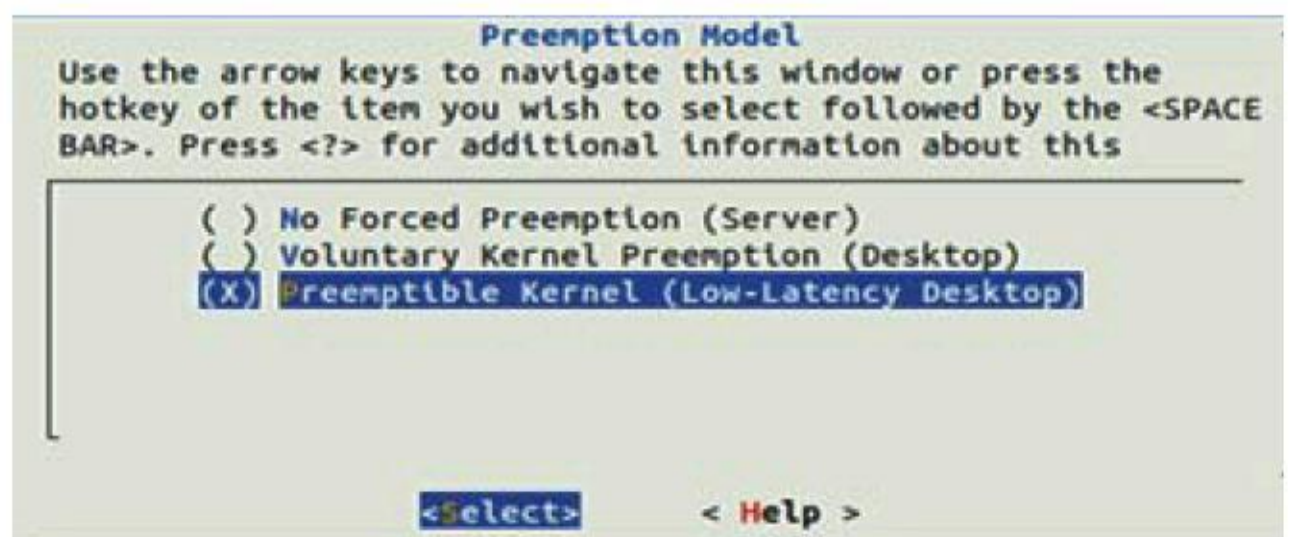


Figure 4: Kernel config options – 'Preemptible Kernel'

Preemption' as shown in Figure 3. The changes are reflected in the .config file, as shown below:

```
CONFIG_PREEMPT_VOLUNTARY_BUILD=y
CONFIG_PREEMPT_NONE is not set.
CONFIG_PREEMPT_VOLUNTARY=y
CONFIG_PREEMPT is not set.
CONFIG_PREEMPT_DYNAMIC is not set.
CONFIG_SCHED_CORE is not set.
```

Figure 4 shows kernel configurations for 'Preemptible Kernel' (low latency desktop).

These changes are reflected in .config as shown below:

```
CONFIG_PREEMPT_BUILD=y
CONFIG_PREEMPT_NONE is not set.
```

```
CONFIG_PREEMPT_VOLUNTARY is not set.
CONFIG_PREEMPT=y
CONFIG_PREEMPT_COUNT=y
CONFIG_PREEMPTION=y
CONFIG_PREEMPT_DYNAMIC is not set.
CONFIG_SCHED_CORE is not set.
```

Figure 5 shows kernel configurations for 'Fully Preemptible Kernel' (real-time).

These changes are reflected in .config as shown below:

```
CONFIG_HAVE_PREEMPT_LAZY=y
CONFIG_PREEMPT_LAZY=y
CONFIG_PREEMPT_NONE is not set.
CONFIG_PREEMPT_VOLUNTARY is not set.
CONFIG_PREEMPT is not set.
CONFIG_PREEMPT_RT=y
```

```
CONFIG_PREEMPT_COUNT=y
CONFIG_PREEMPTION=y
CONFIG_SCHED_CORE is not set.
```

Table 2 lists all kernel config and cyclic test observations.

Figure 6 shows a histogram of No Forced Preemption (server), Figure 7 of Voluntary Kernel Preemption (desktop), Figure 8 of Preemptible Kernel (low latency desktop), and Figure 9 of Fully Preemptible Kernel (real-time).

Figure 10 shows some PREEMPT_RT patch benefits.

Guidelines for using PREEMPT_RT patch in Linux Kernel guidelines

- **Select a compatible kernel version:** Choose a kernel version that is compatible with the PREEMPT_RT patch we intend to apply.
- **Hardware and driver support:** Ensure that hardware and drivers are supported by the selected kernel version.
- **Patch the kernel:** Apply the PREEMPT_RT patch to the chosen kernel version. Follow the patch's documentation and instructions carefully.
- **Configuration options:** Enable necessary real-time options in the kernel configuration.
- **Compile the source code:** Compile the source code after patching as per the instructions.
- **Real-time scheduling policies:** Configure the kernel to support real-time scheduling policies like SCHED_FIFO and SCHED_RR. These are essential for managing real-time tasks.
- **Minimise interrupt handling latency:** Optimise interrupt handling to minimise latency. Make the interrupt service routines as efficient as possible.
- **Community and support:** Engage with the PREEMPT_RT community and mailing lists for assistance, advice, and troubleshooting.

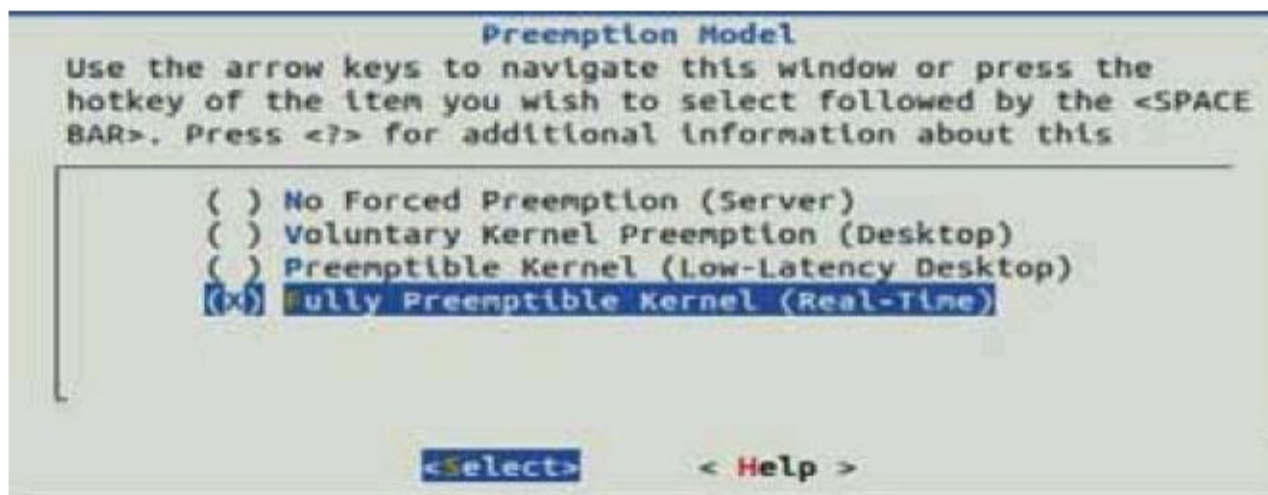


Figure 5: Kernel config options – 'Fully Preemptible Kernel'

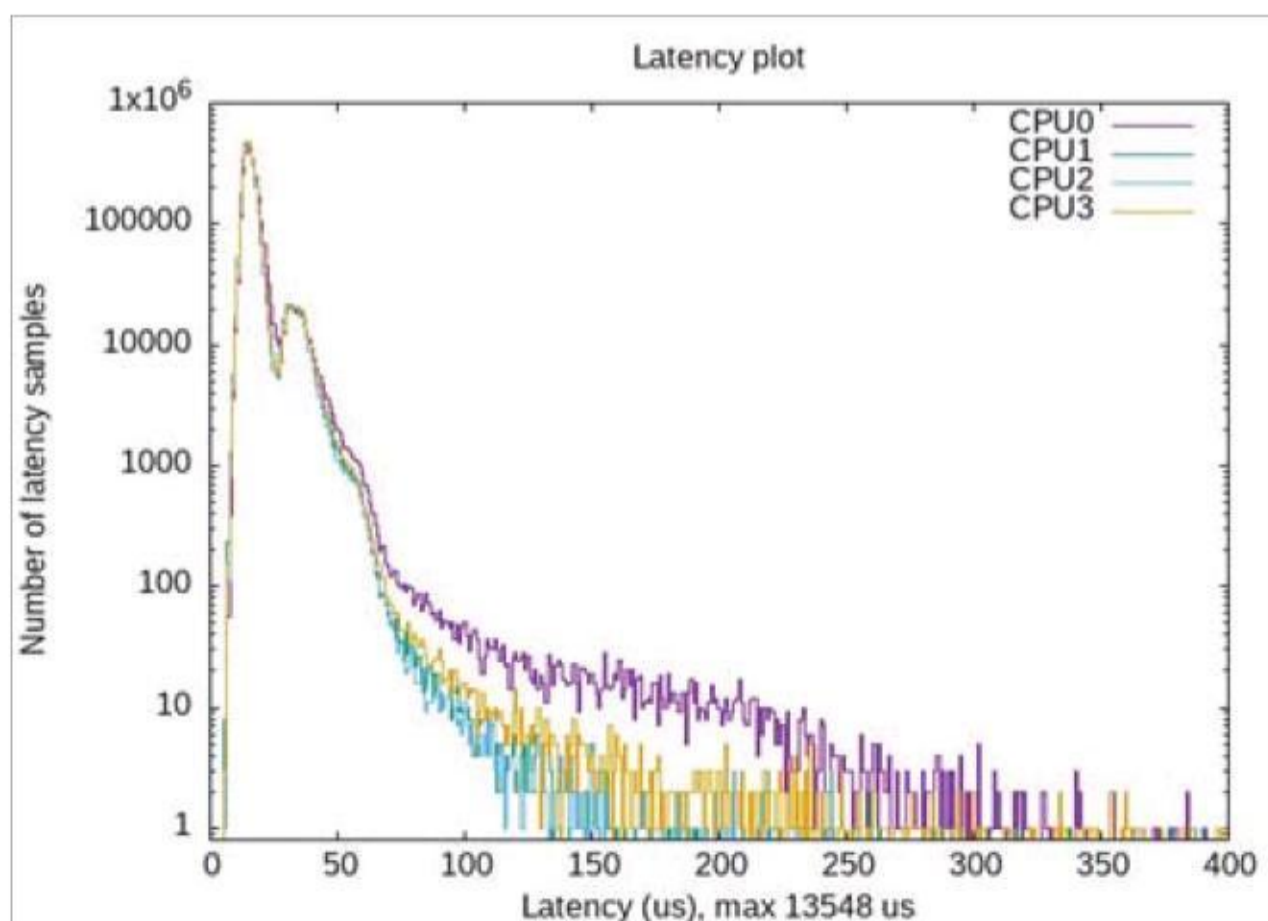


Figure 6: Graph of No Forced Preemption

Name	No Forced Pre-emption (server)	Voluntary Kernel Pre-emption (desktop)	Preemptible Kernel (low latency desktop)	Fully Preemptible Kernel (real-time)
Kernel compilation	Compiled kernel without applied RT-patch	Compiled kernel without applied RT-patch	Compiled kernel without applied RT-patch	Compiled kernel with applied RT-patch
Hackbench load	hackbench -l 100000000	hackbench -l 100000000	hackbench -l 100000000	hackbench -l 100000000
Cyclictest duration	10 min	10 min	10 min	10 min
Latency	13548us	541us	699us	156us

Table 2: Cyclictest observations

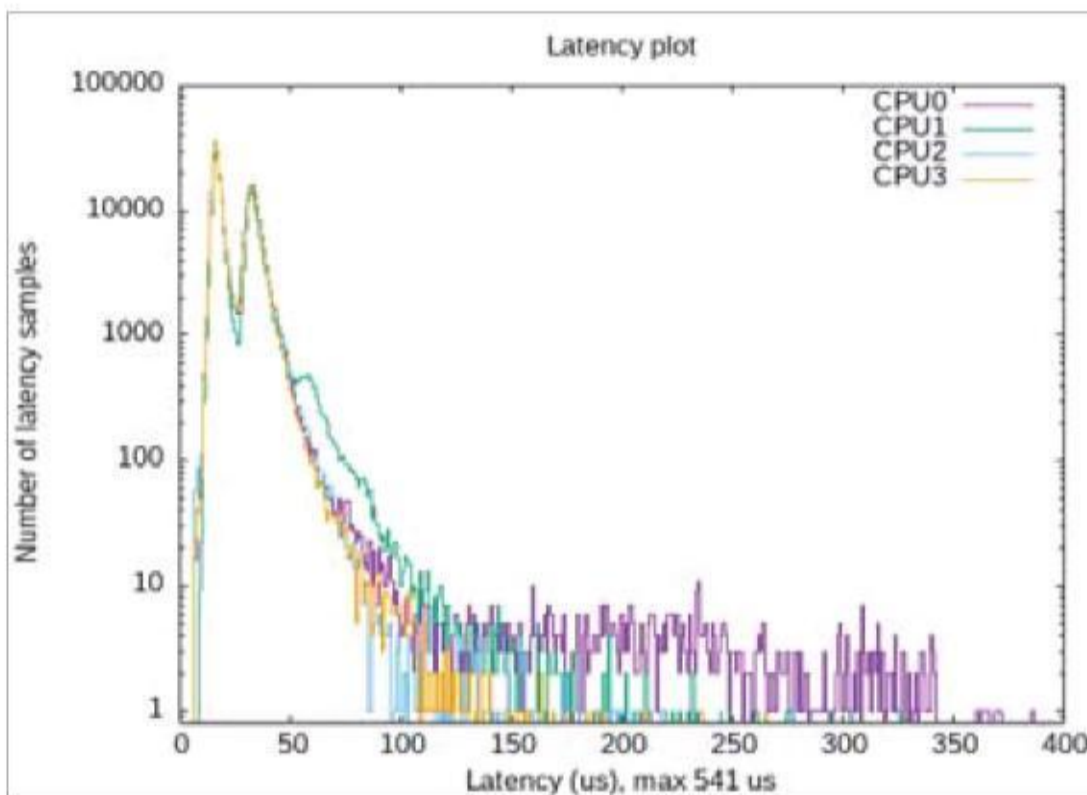


Figure 7: Graph of Voluntary Kernel Preemption

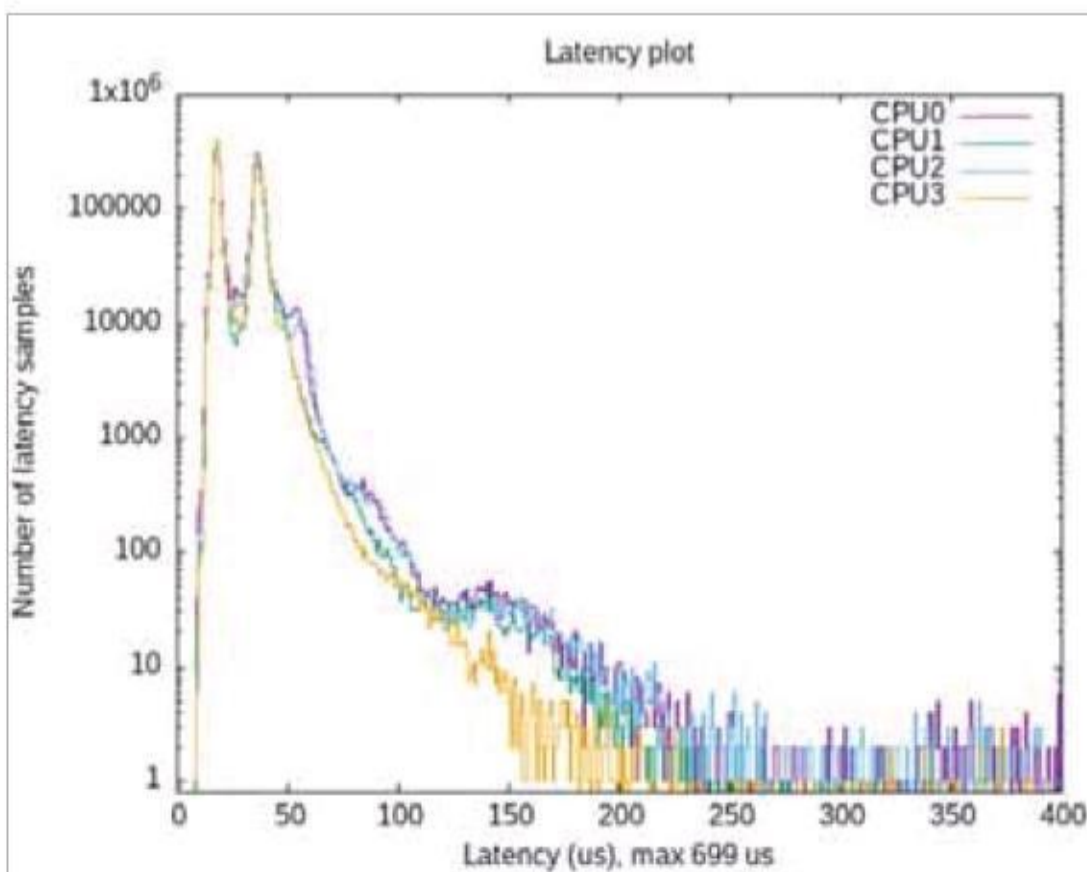


Figure 8: Graph of Preemptible Kernel

Application guidelines

- **Real-time thread priority:** Real-time threads should be assigned the highest priority (SCHED_FIFO or SCHED_RR) to ensure they are not preempted by lower-priority tasks.
- **Use POSIX real-time APIs:** Leverage the POSIX real-time APIs, such as pthreads with SCHED_FIFO, for managing real-time threads. These APIs provide mechanisms for setting thread priorities, managing synchronisation, and avoiding priority inversion.
- **Minimise blocking and I/O:** Real-time threads should minimise blocking operations, especially those that may involve waiting on external events. If blocking is necessary, use non-blocking I/O or asynchronous I/O mechanisms.
- **Avoid busy-waiting:** Avoid busy-waiting (e.g., spinlocks) as it can lead to high CPU usage and increased system latency. Instead, use blocking mechanisms or conditional waiting (e.g., pthread_cond_wait) to efficiently manage synchronisation.
- **Manage memory carefully:** Minimise dynamic memory allocation during real-time operation as it can introduce non-deterministic behaviour. Allocate memory statically or use memory pools to ensure deterministic access.
- **Resource management:** Properly manage resources, such as file descriptors, network connections, and hardware peripherals, to avoid contention and resource exhaustion. Release resources when they are no longer needed.

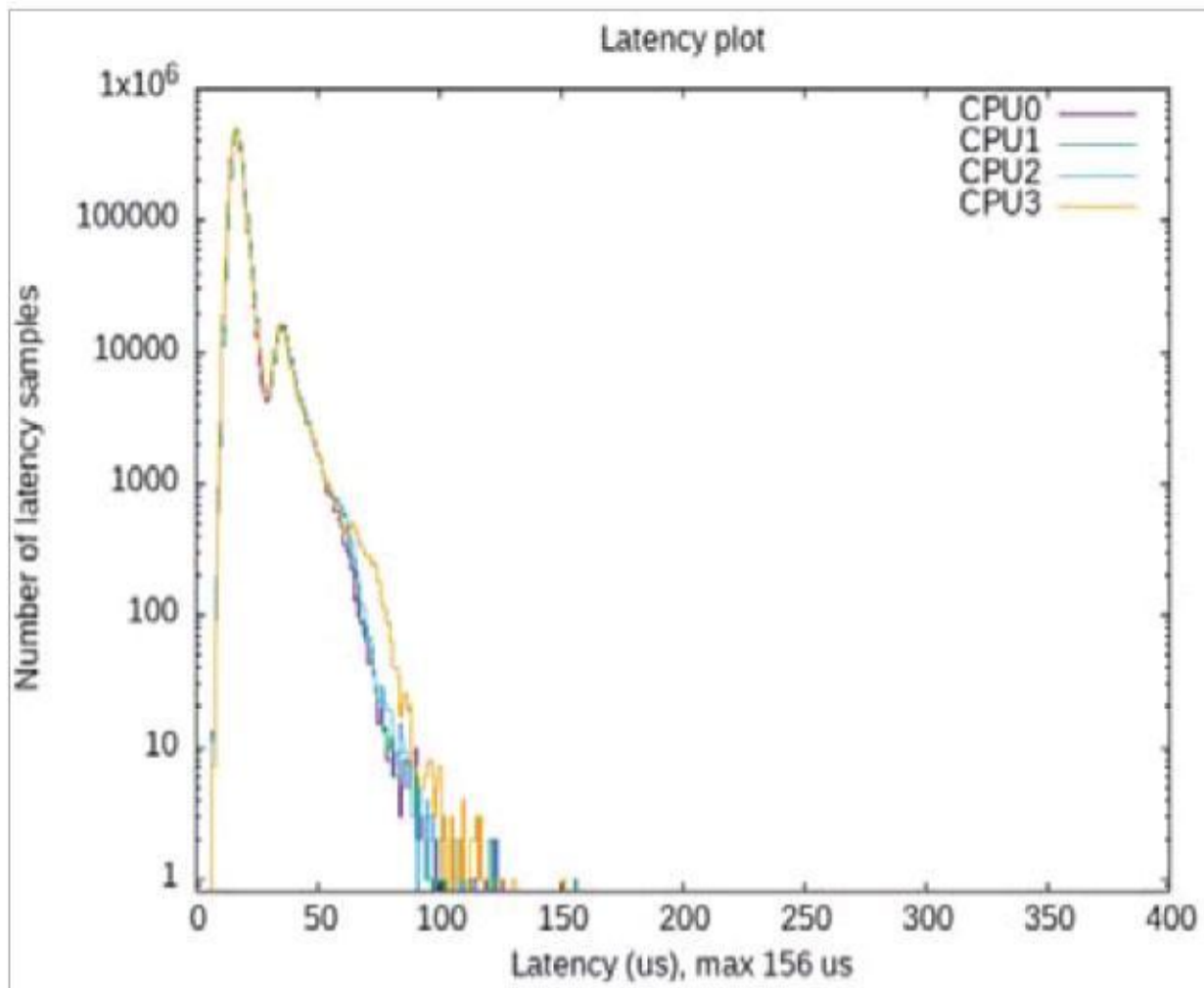


Figure 9: Graph of Preemptible Kernel ((real-time))

- **Error handling:** Implement robust error handling and fault tolerance mechanisms in your real-time application to handle unexpected situations and recover gracefully.

Comparison of QNX and Real-Time Linux

QNX and Real-Time Linux are both popular choices for building real-time and embedded systems, but they have different characteristics, features, and performance considerations. Table 3 gives a comparison of QNX and Real-Time Linux in terms of real-time performance.

In making your decision, consider the following factors.

Timing requirements: PREEMPT-RT patch converts standard Linux to Real-Time Linux. The latter has good determinism to achieve real-time capabilities and reduces scheduling latency. However, QNX may be slightly better compared to Real-Time Linux if your application requires hard real-time capabilities and precise determinism, though the latter can be used based on the requirements.

Reliability and safety: QNX is often used in safety-critical systems including medical devices. It is designed with safety in mind and provides features for safety certification. Real-Time Linux can be used in safety-critical applications, but achieving safety certification may require additional

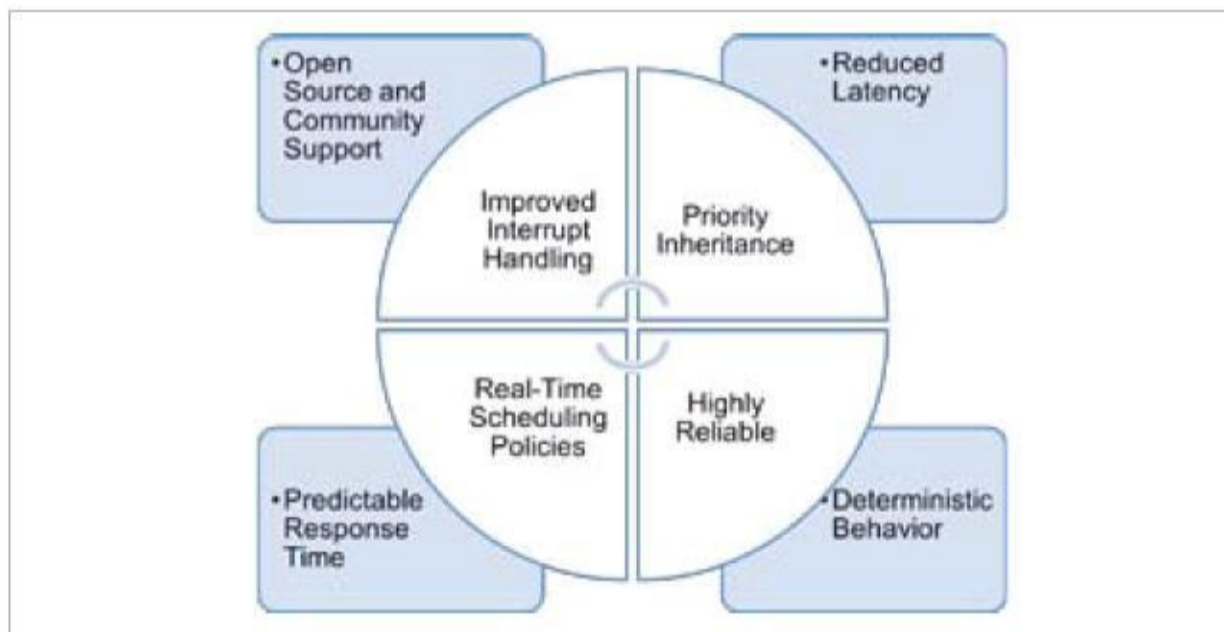


Figure 10: PREEMPT_RT patch benefits

Characteristics	QNX	Real-Time Linux
Determinism	Better determinism and real-time performance out of the box due to its microkernel architecture	Real-Time Linux can achieve good determinism with proper configuration and patches
Latency	Has lower latencies and faster context switching, making it suitable for hard real-time applications	Achieves low latencies with specialised patches but might have slightly higher latencies compared to QNX
Ease of use	QNX is designed specifically for real-time applications and provides a comprehensive environment for embedded systems development	Real-Time Linux may require more configuration and tuning to achieve optimal real-time performance
Ecosystem and compatibility	QNX has a more focused ecosystem tailored to embedded and real-time applications.	Real-Time Linux benefits from the vast Linux ecosystem, including a wide range of device drivers and software libraries



Table 3: Comparison of QNX and Real-Time Linux

effort and adherence to industry-specific safety standards.

Ecosystem and tools: Consider the availability of development tools, libraries, and drivers that match your project's requirements.

Familiarity: If your team is already experienced with Linux development, Real-Time Linux with PREEMPT_RT patch, kernel config options, IRQ handling and thread priorities may be a more seamless choice. SCHED_FIFO, SCHED_RR are often used for Real-Time threads. These factors will improve real-time capabilities in the Linux environment.

Project scope: The size and complexity of the project may influence which platform is more suitable.

In conclusion, both QNX and Real-Time Linux are good options for developing real-time and embedded systems; each comes with its own strengths and considerations. The choice depends on the specific application requirements, the level of real-time performance needed, familiarity with the platforms, the available tools, ecosystem, and the level of determinism needed.  

References

- https://wiki.linuxfoundation.org/realtime/preempt_rt_versions
- <https://mirrors.edge.kernel.org/pub/linux/kernel/projects/rt/>
- <https://bootlin.com/doc/training/preempt-rt/preempt-rt-slides.pdf>
- <https://ubuntu.com/blog/what-is-real-time-linux-part-iii>
- https://elinux.org/images/b/ba/Elc2013_Rostedt.pdf
- https://wiki.archlinux.org/title/Realtime_kernel_patchset
- https://re.public.polimi.it/retrieve/e0c31c12-9844-4599-e053-1705fe0aef77/11311-1076057_Reghenzani.pdf
- <https://static.lwn.net/lwn/images/conf/rtlws11/papers/proc/p19.pdf>
- <https://www.smile.eu/en/white-paperswhite-papers/linux-rt>
- <https://dl.acm.org/doi/abs/10.1145/3297714>
- https://elinux.org/images/d/de/Real_Time_Linux_Scheduling_Performance_Comparison.pdf
- http://www.qnx.com/developers/docs/qnxcar2/index.jsp?topic=%2Fcom.qnx.doc.neutrino.sys_arch%2Ftopic%2Fkernel_Scheduling_latency.html
- http://www.qnx.com/developers/docs/qnxcar2/index.jsp?topic=%2Fcom.qnx.doc.neutrino.sys_arch%2Ftopic%2Fkernel_Interrupt_latency.html
- https://www.mathworks.com/products/connections/product_detail/qnx-neutrino-rtos.html
- <https://blackberry.qnx.com/en/products/foundation-software/qnx-rtos>
- <http://home.iitj.ac.in/~saurabh.heda/Papers/Survey/Report%20on%20%20QNX%20-2007.pdf>

By: Kumar Thangavel, Vijay Annamalaisamy and Sathya Durai

Kumar Thangavel is a senior technical lead with HCL Technologies and has experience in developing embedded software products. He is a part of the Embedded Platform Lab COE and has contributed to various projects.

Vijay Annamalaisamy is a group technical specialist with HCL Technologies. He has extensive experience in developing embedded software for products from multiple domains. He is a part of the Embedded Platform Lab COE and has contributed to various projects.

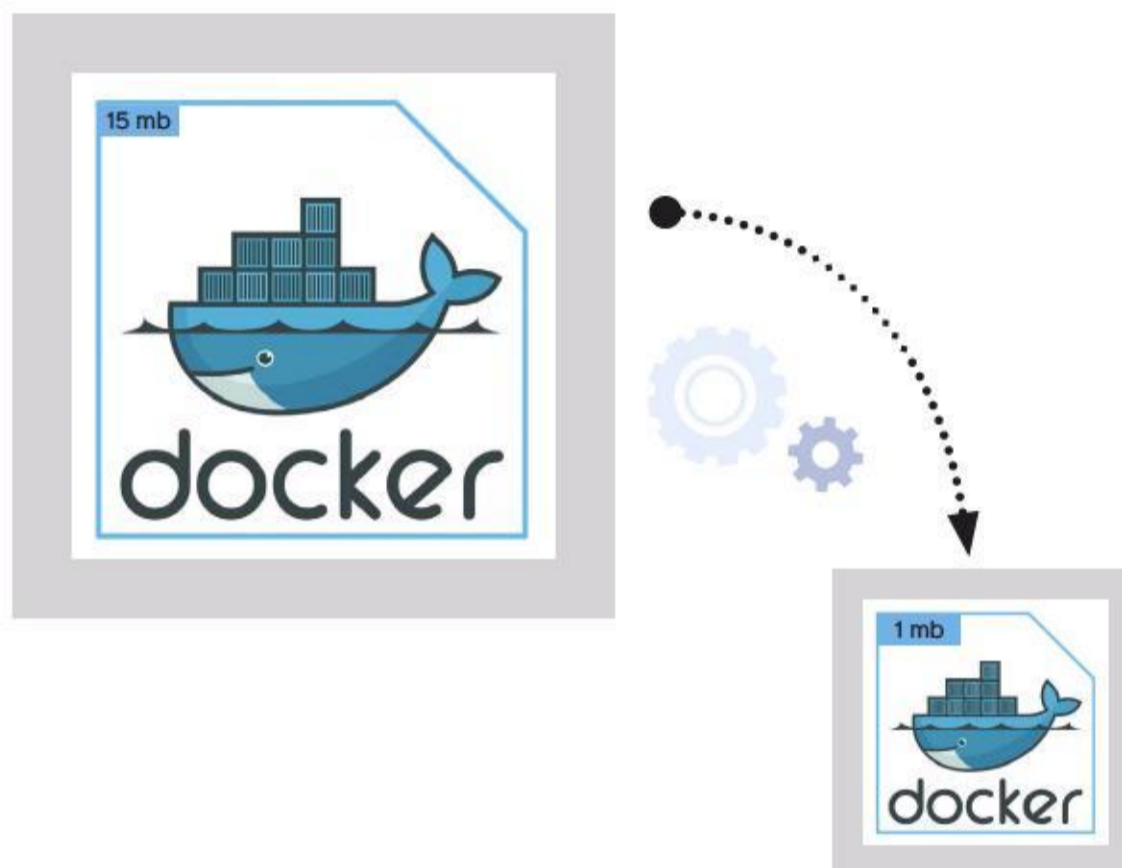
Sathya Durai is a solution director with HCL Technologies. He has extensive experience in playing the architect role for embedded software development for products from multiple domains. He also is a part of the Embedded Platform Lab COE and has contributed to various projects.

OSFY Magazine Attractions During 2023-24

Month	Theme
March 2023	Security, Network Management and Monitoring
April 2023	Open Source Programming (Languages and tools)
May 2023	Cloud Special: Everything from management to implementation
June 2023	AI, Deep learning and Machine Learning
July 2023	Database management and Optimisation
August 2023	Mobile/Web App Development, Optimisation and Security
September 2023	DevOps Special
October 2023	Blockchain and Open Source
November 2023	Open Source and IoT and Edge
December 2023	All About Data Management
January 2024	Containers and Managing Containers
February 2024	Open Source on Windows and Best in the world of Open Source (Tools and Services)

Reducing the Size of Your Docker Image

Why stick with a 6GB Docker image that functions? Downtime's crucial; services require updates and lightweight images mean faster deployments, saving time.



I work on a project called Kube-ez (www.github.com/kitarp29/kube-ez). It is a simple Golang project which has an image of nearly 0.5GB in size. This resulted in excessive time consumption while spinning up any container with this image. This issue not only affected deployments but also slowed down the CI jobs, leading to a delay in my entire software development life cycle.

I was able to fix this issue by reducing the image size by 97%, turning it from ~0.5GB to a mere 16MB. This change enabled me to build more reliable deployments and faster CI jobs.

Let me present the approach I adopted to address this problem, and discuss some alternative solutions.

Understanding the problem

In the world of microservices, most production relies on containers. Images act like the CDs/flash drives used to play a certain song (your software) in your container.

Docker images aim to bring unity among the developers despite the platform and environment. However, achieving this should not come at the cost of memory and speed.

- Bulky images take longer to download.
- You need increased time to spin up a container using them, slowing deployments and CI pipelines.
- They increase the load and cost of the image registry.
- With upgrades, it becomes harder to check for vulnerabilities.

Why solve this problem?

I often hear from developers: “So what if the Docker image is 6GB; it still runs, right? Don’t touch it.”

However, I disagree based on my experience as a site reliability engineer at Juspay (an Indian fintech startup).

Keeping a service up and running is always a significant challenge. But let’s agree that no company has an SLO of 100%. Each second of downtime counts!

Someday, the service is bound to crash or run into errors. Even in the best-case scenario, you will need to update it sometime. On that day, when you need to spin up a new container or an updated container, having a lighter Docker image will help.

A lighter Docker image ensures that containers spin up faster, saving precious seconds for your company. Also, services like AWS ECR charge you nearly \$0.10/GB of images stored each month, making it a cost-effective choice.

Ways to reduce the Docker image size

There are three different approaches to this task. The first option is to use a third-party service such as Docker Slim to handle the heavy lifting. The next one is to delve deep into Docker images and improve them layer by layer. Lastly, use Docker's own service, which has been recently introduced and simplifies the process of developing Docker images according to best practices.

Docker Slim

Docker Slim (<https://slimtoolkit.org/>) is the best and easiest tool to solve this problem.

- Easily reduces the size of your Docker images.
- Connects to DockerHub and gives you a user-friendly platform for managing images on the registry.
- Checks and fixes vulnerabilities.
- Installation and usage are straightforward.

However, it would be great to see APK and other package installation methods in the future from this platform. As you can see from the results (Figure 1), this service can help you reduce the image size by around 37%.

If you wish to reduce the size of a local Docker image, you can refer to the article by the RedHat Developer Group at https://developers.redhat.com/articles/2022/01/17/reduce-size-container-images-dockerslim#using_dockerslim.

Manual approach

This approach requires a good understanding of your project. There are two steps in this approach, and you can stop after the first step if it satisfies your needs. Let's dig deeper.



Figure 1: Results using Docker Slim to shrink image size

Using lighter base images: This is the easiest and the first option to consider. Use a lighter image as the base layer for your container. You can opt for an Alpine image of the OS or the programming language you are using.

Alpine images are basic, barebone OS, missing essential software and drivers like graphics, Wi-Fi, and more.

Your DockerFile would look something like this:

```
FROM golang:alpine
WORKDIR /app
COPY . .
RUN go build -o app
CMD ["/app"]
```

Multistage build: The idea is to use two different stages in the image. One is to build the binary, and the next is to simply use it!

In my case, I applied a multistage build pattern, creating a binary in the build stage and using it to run the container. Now the image size is ~35 MB. Further size reduction was achieved by optimising the GCC dependency, resulting in an image size of around 16MB.

Since Go binaries can run independently, I did not need a Go Alpine image.

Here is my final DockerFile:

```
# Build
FROM golang:alpine AS build
WORKDIR /k8-api
COPY . .
RUN go mod download && \
    G0111MODULE=on CGO_ENABLED=0 go
build -ldflags "-s -w"
# Deploy
FROM alpine
COPY --from=build /k8-api/k8-api .
ENTRYPOINT ["/k8-api"]
EXPOSE 8000
```

Docker init

Docker has just released a feature called 'docker init', which automates the entire process. It not only reduces the image size but also follows best practices for creating a Docker image. I am still trying it with different projects but the preliminary tests seem promising.

You can read about docker init at <https://docs.docker.com/engine/reference/commandline/init/>

We have discussed three different approaches for reducing Docker image size, and provided tips and resources for further optimisation. You can choose the approach that best suits your needs in order to improve production performance and reliability. **END** 🐧

By: Pratik Singh

The author works at NASDAQ as a senior software developer, and has been dabbling with Go and DevOps for the last two years.

EdgeX Foundry: Transforming IoT with Edge Computing

EdgeX Foundry boosts device performance by deploying intelligence at the edge, using CPU and GPU resources effectively. It offers a range of benefits, and together with Ubuntu Core, is transforming the IoT landscape by bringing computing close to the data source.



IoT (Internet of Things) has sparked a revolution in various industries by interconnecting devices and facilitating data-driven decision-making. However, as the volume of IoT data escalates, challenges related to latency and bandwidth become increasingly evident. Edge computing, which processes data closer to its source, thus reducing latency and enhancing efficiency, addresses these issues. At the forefront of this transformative approach stands EdgeX Foundry, an open source framework.

While the concept of edge computing is undoubtedly intriguing, successfully implementing it can be a complex and intimidating endeavour. This is precisely where EdgeX Foundry emerges as a game-changing solution. This open source, vendor-neutral framework aims to streamline and standardise edge computing for IoT applications.

Now, let's explore its notable advantages to understand what sets EdgeX Foundry apart from other solutions. There are several

compelling reasons to consider it as your primary choice for an IoT edge computing platform.

Speed to market

In the fiercely competitive IoT space, speed to market can make or break a product. EdgeX Foundry accelerates development by providing a pre-built foundation with a comprehensive ecosystem of tools and resources, allowing developers to roll out innovative IoT solutions to market swiftly.

An open source powerhouse

EdgeX Foundry embodies the spirit of collaboration and innovation that defines the open source community. It operates under the Apache 2.0 open source licensing model, endorsed by the Linux Foundation’s LF Edge organisation. This open approach guarantees vendor neutrality. It also invites contributions from a global network of developers, ensuring that the platform remains robust and adaptable. It empowers not just developers but technology providers and end users alike to leverage the power of edge computing without any restrictions. The result? An ecosystem of ideas and solutions that benefit everyone involved.

Unmatched flexibility

With a modular architecture and a microservices-based framework, EdgeX Foundry offers unparalleled adaptability. IoT projects vary widely in requirements, and the flexibility and pluggable architecture of EdgeX plus the choice of hardware for the selection

of communication protocols, allows for tailoring IoT solutions to the unique needs of various industries.

Focus on value addition

EdgeX Foundry isn’t just about data; it’s all about adding value. By bringing computing closer to the data source, it allows real-time analytics and decision-making. This shift not only boosts efficiency but also opens new opportunities, letting developers focus on solving real-world problems without worrying about the complexities of edge computing infrastructure.

Cost reduction and risk mitigation

EdgeX Foundry helps realise cost savings in your cloud infrastructure and minimises data transfer expenses. The practice of processing data locally reduces the exposure to risks associated with transmitting sensitive data over networks. EdgeX Foundry not only reduces infrastructure costs but also enhances security and compliance by maintaining sensitive data within a localised environment.

Enabling innovation

Innovation is the heartbeat of EdgeX Foundry. Its open framework nurtures an ecosystem of compatible applications and services, igniting the imaginative sparks of developers. This spirit of innovation knows no bounds, enabling industries and organisations to modernise their operations, optimise processes, and deliver personalised real-time experiences through the power of IoT.

Realising edge computing benefits

Edge computing goes beyond mere tech buzz; it represents a groundbreaking shift that unveils a world of endless possibilities. In this transformative landscape, EdgeX Foundry takes centre stage as the key to unlocking the full potential of edge computing, especially in industries where lightning-fast data processing, ultra-low latency, unwavering reliability, and real-time decision-making are critical. Imagine it as having a cloud-based virtual machine right at the edge, revolutionising the rules of the entire game.

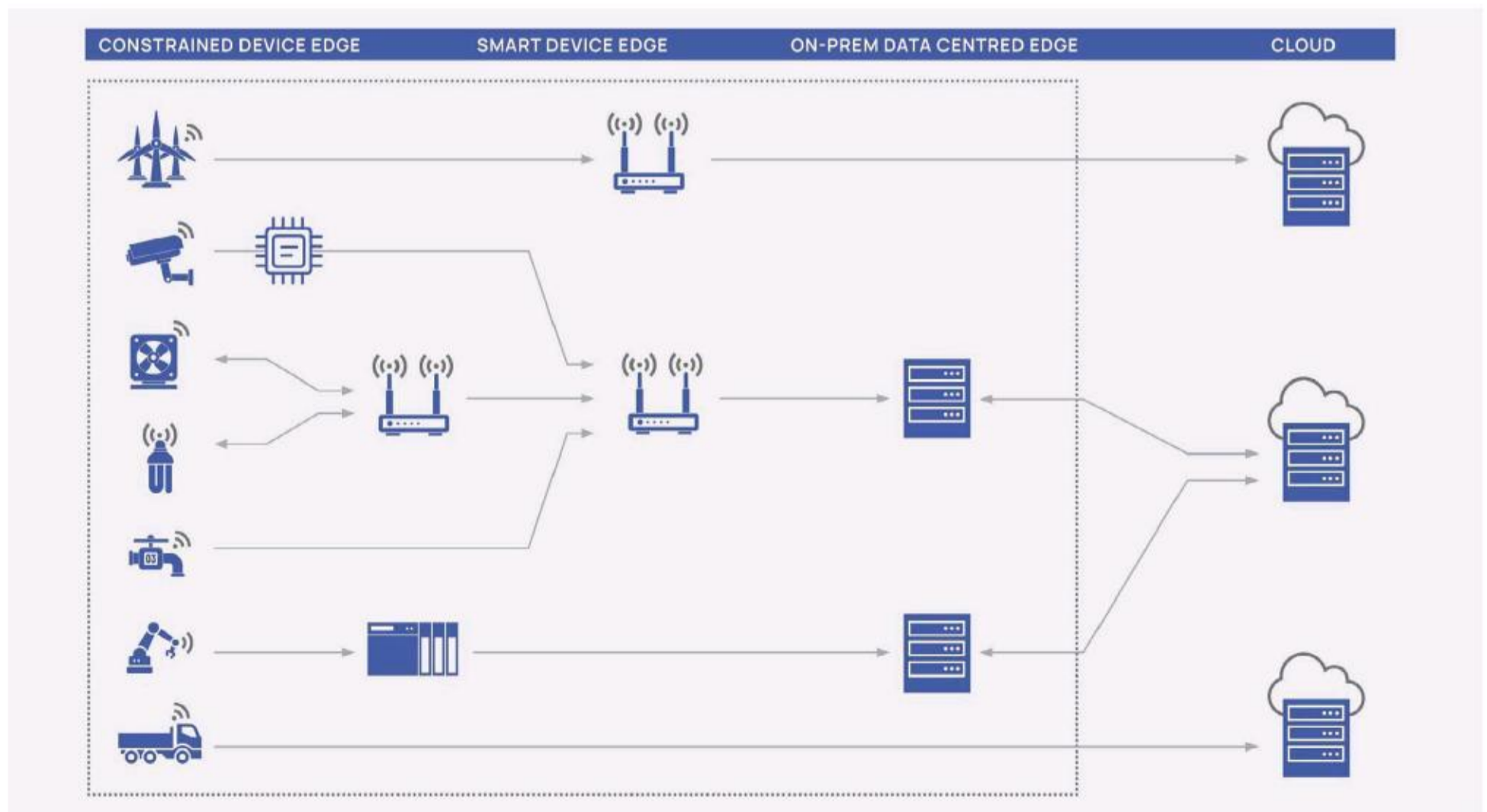


Figure 1: EdgeX platform architecture [Image credits: <https://www.edgexfoundry.org/cmsfiles/image/Platform/edgex-platform-architecture-2.jpg>]

The EdgeX Foundry architecture

The architecture of EdgeX Foundry is the cornerstone of its success and is thoughtfully designed to maximise technological efficiency. It utilises a modular, microservices approach that ensures adaptability and scalability, much like a city with distinct buildings serving unique purposes. This architectural design streamlines data flow, remains platform-agnostic for seamless integration with diverse technologies, and establishes a secure foundation for data exchange and decision-making at the edge. It is a testament to EdgeX Foundry's commitment to delivering a holistic solution for edge computing, akin to a well-planned city that accommodates the dynamic needs of its inhabitants and their varied requirements.

Platform agnosticism

EdgeX Foundry wholeheartedly embraces diversity. It doesn't favour one platform over another, ensuring compatibility with a vast array of hardware from different OEMs and ODMs, a range of operating systems, cloud providers, and applications that span direct interfaces to sandboxed environments and containers. This platform agnosticism guarantees effortless integration into your current infrastructure, enabling businesses to adapt and innovate without the need for extensive overhauls.

Microservices flexibility

At the heart of EdgeX Foundry lies a microservices architecture that serves as a foundation for crafting nimble, scalable, and exceptionally flexible services. This adaptability empowers organisations to tailor their IoT solutions precisely to their needs, avoiding the confinement of rigid, monolithic applications. This approach ensures agility in a market that constantly evolves. It boasts support for containerisation, such as Docker or Snap, promoting application portability and simplifying deployment

and orchestration. It facilitates the development of interoperable, plug-and-play edge software applications and value-added services.

Store and forward capability

EdgeX Foundry offers a crucial feature known as 'store and forward capability'. This capability proves essential in situations where network connectivity is intermittent. It enables data to be temporarily stored at the edge, and when a stable connection is re-established, the data is efficiently forwarded, ensuring seamless operations.

Additionally, EdgeX Foundry optimises data handling by compressing historical data and eliminating redundant information within specific time intervals. This not only reduces the data size but also streamlines the process of collecting data from multiple IoT devices before performing computations for data transmission and storage.

Intelligence at the edge

EdgeX Foundry's strategy of positioning intelligence at the edge serves as a solution to various critical concerns, including actuation latency, bandwidth constraints, storage limitations, and facilitating remote operations. This approach optimises performance and operational efficiency while also acting as a mini-monitor for IoT devices onsite. Furthermore, this intelligent edge placement allows for the effective utilisation of not only higher CPU compute power but also the efficient utilisation of GPU and AI-related CPU resources.

Brown and greenfield deployments

EdgeX Foundry's architectural design inherently resolves critical interoperability challenges that arise at the convergence of various directions in a distributed IoT edge architecture. This ensures seamless integration with legacy systems, enabling the collection of data from existing sensors and controllers. Consequently, EdgeX

Foundry eliminates the need for costly and disruptive infrastructure overhauls in projects such as smart building optimisation. The real-time processing and analysis of data from brownfield devices at the edge result in enhanced energy efficiency and increased occupant comfort.

Moreover, EdgeX Foundry's capability to accommodate both brownfield and greenfield deployments simplifies the gradual expansion and modernisation of IoT systems while honouring existing infrastructure investments. This adaptability proves to be an invaluable asset in real-world applications, where IoT initiatives often comprise a mixture of legacy and cutting-edge technologies.

Security and manageability

EdgeX Foundry places a strong emphasis on security, offering the capability to process and store sensitive data locally, thereby significantly reducing the risk of potential security breaches. This approach empowers companies to enhance the protection of their critical edge data and devices. It enables on-premise data processing while allowing selective exposure of specific data and applications to the cloud, ensuring data security and compliance with industry regulations.

EdgeX Foundry doesn't limit its focus to security alone; it also streamlines the management of widely distributed compute nodes and effectively scales down to accommodate highly-constrained devices. This simplifies operational and system management, making the platform more accessible and versatile across a wide range of use cases.

Support on Ubuntu Core, an IoT-optimised Linux distribution

Ubuntu Core is purpose-built to address the distinctive requirements of IoT deployments. Its lightweight design and minimal resource footprint are well-suited for edge devices with constrained

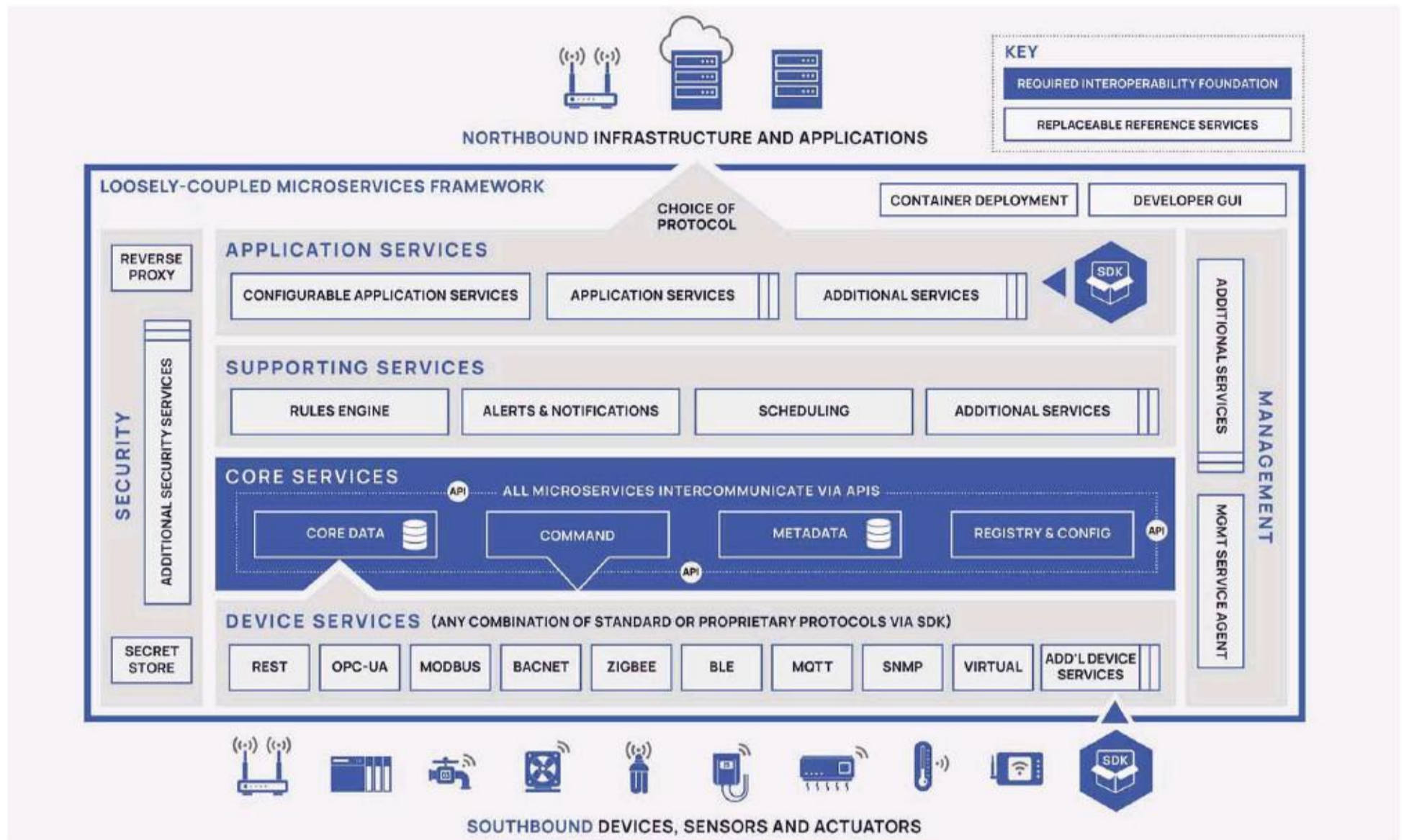


Figure 2: EdgeX focused IoT [Image credits: <https://www.edgexfoundry.org/cmsfiles/image/Platform/edgex-focused-at-the-iot-edge-2.jpg>]

resources, ensuring the efficient utilisation of assets. Ubuntu Core offers a range of features that bolster IoT deployments, including:

- **Transaction-based updates:** This enhances update reliability, crucial for uninterrupted IoT operations.
- **Secure boot and encryption:** These features fortify system security, ensuring data integrity and access control.
- **App store model:** Simplifying software management, it streamlines deployment and offers long-term support for mass-scale production.
- **Strict isolation with snaps:** Snaps, which are containerised, universal Linux packages, bundle applications and dependencies, introducing several benefits:
 - **Isolation:** Each snap operates in its own isolated environment, avoiding conflicts and enhancing security.
 - **Atomic updates:** Snaps can be individually updated, ensuring

the latest security patches and bug fixes.

- **Dependency management:** Snaps encompass all necessary dependencies, simplifying software installation and maintenance.
- **Version rollback:** In case of issues with an update, snaps allow for straightforward rollback to a previous version.
- **Cross-distribution compatibility:** Snaps can be installed on any Linux distribution that supports them, promoting interoperability.
- **Long-term support for mass-scale production:** Organisations can rely on Ubuntu Core for an extended period, ensuring the stability and security of their IoT devices even in

large-scale, long-term production scenarios. This support typically includes regular updates, security patches, and maintenance services to keep the IoT infrastructure running smoothly and securely.

EdgeX Foundry, in combination with Ubuntu Core, is leading the charge in transforming IoT with edge computing. This powerful duo brings computing closer to the data source, enabling real-time processing and decision-making in a secure and efficient manner. As IoT continues to evolve, EdgeX Foundry’s open source, flexible, and adaptable nature makes it the ideal choice for organisations looking to harness the full potential of edge computing in their IoT deployments. **END** 🐧

By: Deepen Dhulla

The author has over 21 years of experience in the Linux and open source realm. He specialises in assisting clients in setting up private clouds using HCI based on ProxmoxVE and Ceph SDS.

IOTSHOW.IN

31ST JAN - 2ND FEB 2024 • KTPO • BENGALURU

PROFIT FROM



INDIA'S ONLY EVENT FOR THE **CREATORS**
AND THE **SOLUTION PROVIDERS** OF IOT

IoTshow.in is India's biggest conference-cum-expo on Internet of Things.

Visitor Registration: www.iotshow.in. Special Offers for first 1000 Registrants!

CO-LOCATED WITH:

INDIA
ELECTRONICS
WEEK

BROUGHT TO YOU BY:



MORE INFO:

Web: www.iotshow.in
Email: support@efy.in
Bulk Registration: iew@efy.in
Tel: +91-9811155335

“

**I started reading it
when I was a student...
...and I am still
reading it, as a student**”

—CEO, Design House



electronics
YOURS SINCE 1969 **FOR YOU**



To Subscribe:
<https://subscribe.efy.in>

OR

Scan This Code



For any query, call: +91-98111-55335 or email: support@efy.in

