Containerisation Of Java Microservices

Are Low-Code Platforms Helpful For Professional Developers?

₹120
ISSN-2456-4885

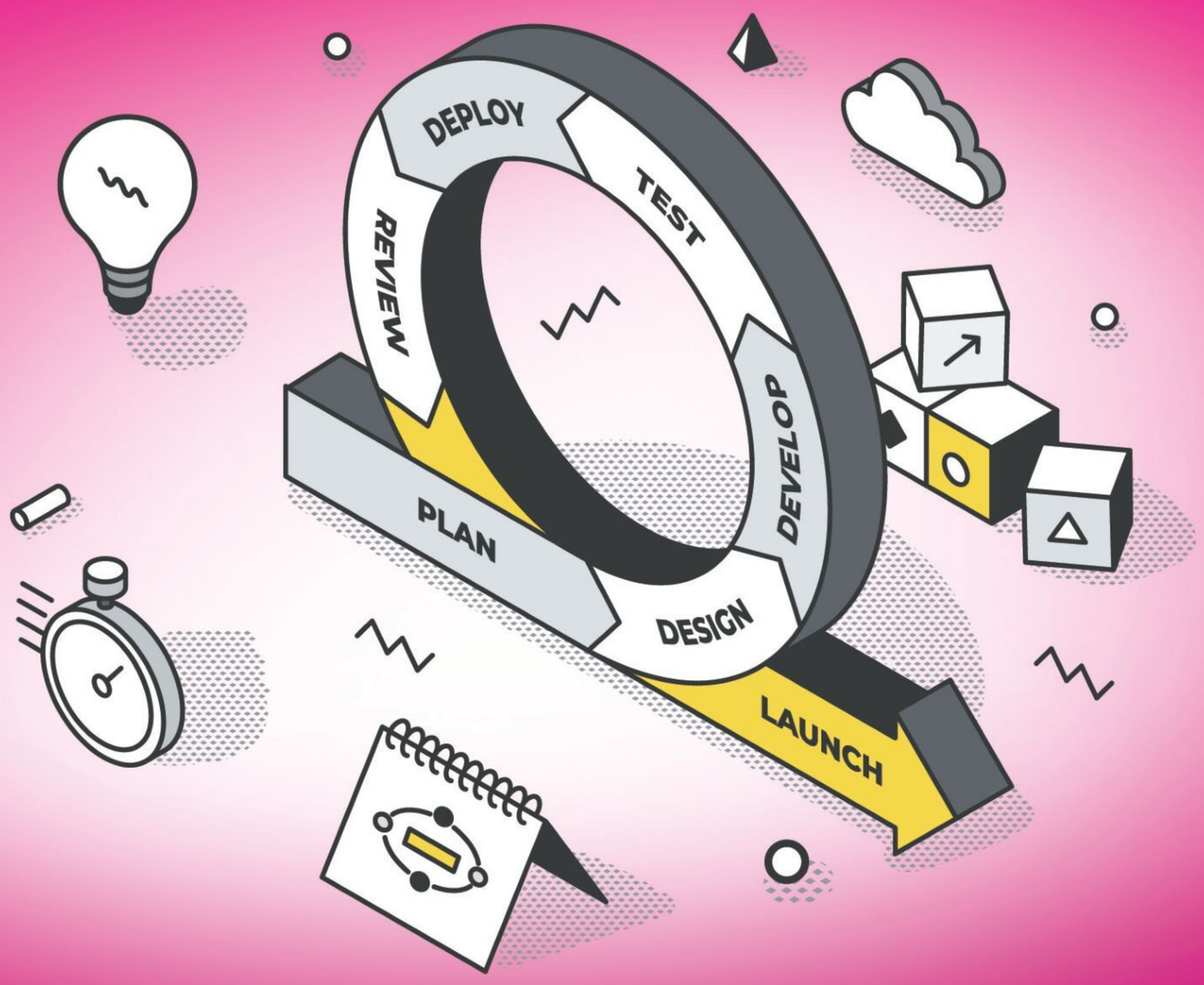
OpenSource For You

Volume: 10 | Issue: 07 | Pages: 100 | May 2022

THE COMPLETE MAGAZINE ON OPEN SOURCE

EFY GROUP Publication

Understanding the DevOps Culture



The Role Of DevOps In Digital Transformation

DevSecOps: A Philosophy That Puts Security First

DevOps And Agile Computing: Enabling A True Cultural Shift



"It's time to contribute to open source"

Nilesh Vaghela, Founder, ElectroMech Corporation

Why You Should Deploy Immutable Infrastructure On The Cloud

www.opensourceforu.com

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.

Does Your Antivirus Solution Provide You With Complete Protection?

Don't RELY on 20 years old technology to FIGHT current date viruses



TOP 10 things that your Antivirus solution should provide to tackle today's threat...

- Ransomware File Protection
- ATP- Advance Threat Protection
- EDR- Endpoint Detection And Response
- Application Blocking
- Deep Learning Malware Analysis
- Exploit Prevention
- URL Blocking
- Disk and Boot Record Protection
- Peripheral Control
- Respond Investigate Remove (Root Cause Analysis)

If it doesn't, contact us for a solution which can

Contact: Santosh on 9971696319 or Email at santosh.gupta@itsipl.com

ITS
A Total Solutions Company

I. T. Solutions India Private Limited
D-88/5, Okhla Industrial Area, Okhla Phase I, New Delhi -110020
Ph: 011-47695000 • Email: sales@ itsipl.com • www.itsipl.com

- Mumbai
- Jaipur
- Chandigarh

SOPHOS 2020 BEST NEXTGEN PARTNER | Delhi/ NCR
SOPHOS 2019 BEST NEXTGEN PARTNER | Delhi/ NCR
SOPHOS 2018 BEST SI PARTNER | Delhi

CONTENTS

MAY 2022 | ISSN-2456-4885

FOR U & ME

23 AI Accelerators: A Long Way to Go

FOCUS

39 Are Low-Code Platforms Helpful for Professional Developers?

44 Why You Should Deploy Immutable Infrastructure on the Cloud

50 Demystifying Observability for Modern Applications

55 Open Source Tools for MLOps: An Overview

58 DevOps and Agile Computing: Enabling a True Cultural Shift

60 The Role of DevOps in Digital Transformation

DEVELOPERS

62 Hamming Codes: The Error-Correcting Codes

71 Recommender Systems: An Overview of the Mathematics

78 Data Visualisation in R: Graphs

ADMIN

89 Manual Renewal of SSL Certificates: A Simple Guide

91 Analyse Logs with Elasticsearch

OPEN GURUS

95 Supervised Machine Learning for Car Evaluation Using KNIME



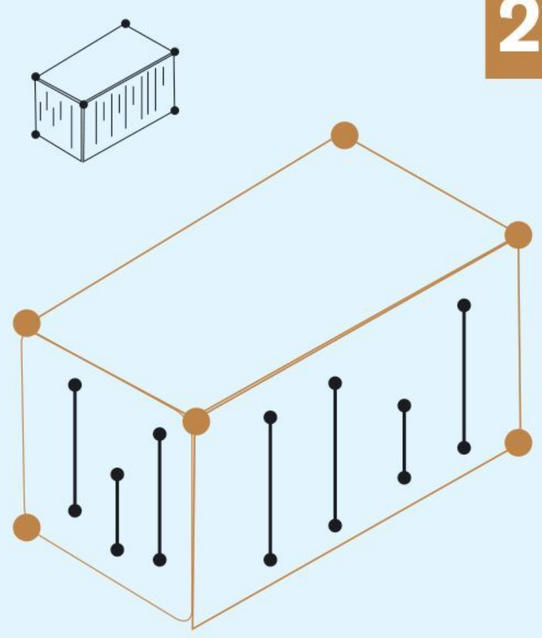
"It's time to contribute to open source"

— **Nilesh Vaghela,** AWS community hero and founder, ElectroMech Corporation

The Design Odyssey

20

Containerisation of Java Microservices





Enter the Metaverse for Better or for Worse

REGULAR FEATURES

07 FossBytes

www.opensourceforu.com

Job Guaranteed.

You don't need a training institute. Just do it yourself.

WORK WITH OPEN SOURCE.

Demand for Open Source is sky rocketing. Be it for managing IT infrastructure or development of software —Open Source experts are what organisations are seeking. Develop expertise in an Open Source stack, and watch your career graph zoom.

You don't need a training institute for this. You can do it yourself. 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.



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: efyenq@efy.in

MUMBAI
E-mail: rmowest@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 © 2021. 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.

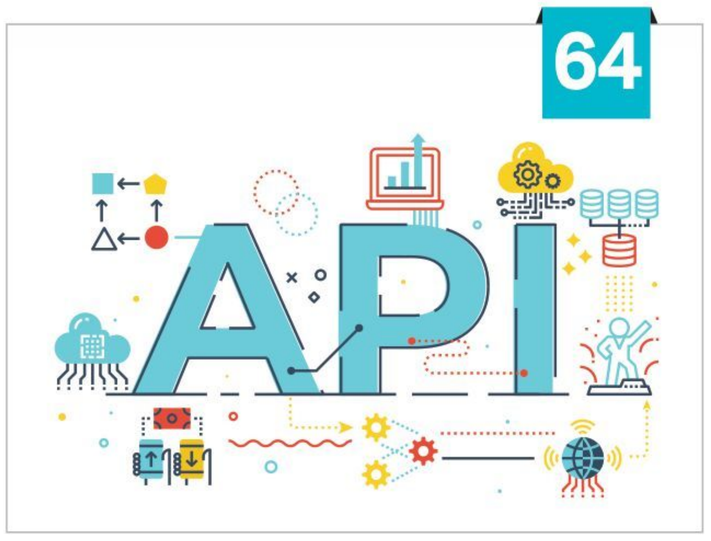
CONTENTS



DevSecOps: A Philosophy that Puts Security First



Planning the DevOps Journey for an Enterprise



Generating APIs Using Django Rest Framework with Insomnia



An Introduction to Teradata

JiHu GitLab gets Series A round of financing



JiHu GitLab, an open source startup, has completed the Series A round of funding for hundreds of millions of yuan. This round of funding will be utilised for product development, market expansion, open source ecological construction, and independent intellectual property research.

This batch of funding was split into two parts. Temasek led the first stage, with Alpha Prime, GGV Capital, Shanghai Artificial Intelligence Industry Fund, and Nokia Growth Partners (NGP) Fund following. Taikang Insurance Group led the second stage, followed by Cheers Fund, Lenovo Capital, and Incubator Group. Sequoia Broadband Cross-border Digital Industry Investment Partnership and Gaocheng Capital, among others, increased their investments in the company during this round.

JiHu GitLab is a DevOps firm founded by GitLab Inc., a US-based open source platform, and supported with Chinese funds. It began operations in March 2021, with its US parent providing technology and branding. The company is run by local staff that is separate from the rest of the operations. GitLab Inc.'s shareholding ratio will be decreased to less than 50 per cent following this round of funding, and China's own independent operating system will be implemented.

The firm independently investigated, created, and launched the 'GitLab Integrated Security DevOps Platform', 'GitLab SaaS', and 'GitNative DevOps Integrated Machine', which are more tailored to the demands of Chinese developers; the company's infrastructure and data are stored in China.

Open Robotics to provide open source simulator for maritime contest

Open Robotics will provide the open source simulator for the simulation phase of the MBZIRC Maritime Grand Challenge, according to ASPIRE, the technology programme management arm of Abu Dhabi's Advanced Technology Research Council (ATRC).

GitHub's push protection capability makes security proactive

Security breaches continue to affect all of us. To date, GitHub has detected more than 700,000 secrets across thousands of private repositories using secret scanning for GitHub Advanced Security. GitHub also scans for its partner patterns across all public repositories (for free). It has now added an option in GitHub Advanced Security to help prevent leaks from happening altogether by scanning for secrets before a Git push is accepted.

By scanning for highly identifiable secrets before they are committed, security can turn proactive instead of being reactive, preventing secrets from leaking altogether.

GitHub's new push protection capability embeds secret scanning in the developer workflow. To make this possible without disrupting development productivity, push protection only supports token types that can be detected accurately. The new capability has been launched with support for 69 high confidence patterns, each of which has a signal-to-noise ratio that developers can trust.

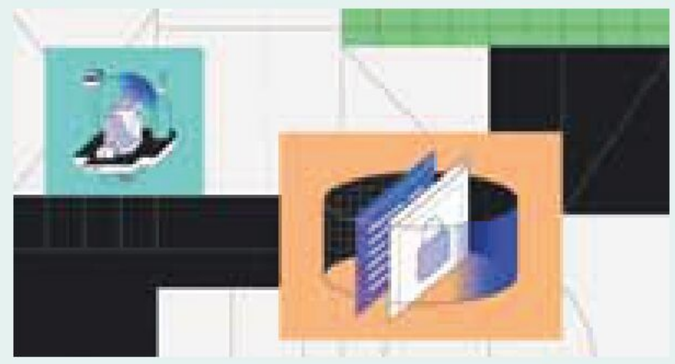


Image Source: https://github.blog/

With push protection, GitHub will check for high-confidence secrets as developers push code and block the push if a secret is identified. High-confidence secrets have a low positive rate, so security teams can protect their organisations without compromising developer experience.

NetApp acquires Instaclustr

NetApp Inc., a data storage company, has acquired the open source database startup Instaclustr Pty Ltd for an undisclosed sum.

Enterprises who don't want to handle their own software can use Instaclustr to get fully managed versions of the most popular open source databases. Its 'database as a service' platform hosts and supports a variety of open source technologies, including Apache Cassandra, Apache Kafka, Apache Spark, Apache Lucene, and ScyllaDB, among others.

NetApp, which provides companies with hybrid cloud data services and data management, said it will merge Instaclustr into its CloudOps offering after the deal closes, subject to customary closing conditions.

In a blog post, NetApp executive



vice president and general manager Anthony Lye stated that the transaction is all about improving client experiences by making open-source software installation easier. "The acquisition of Instaclustr is a strategic next step for NetApp -- to do more for the customer, simplify the experience, continuously optimize and secure the platform so customers don't have to do it themselves," he wrote.

NetApp has been on a buying spree of late, acquiring startups like Spot, CloudCheckr Inc., and, most recently, Fylamynt, in order to expand its reach into areas like cybersecurity, cloud cost control, and cloud automation. The purchases are all part of NetApp's strategy to transition from a legacy storage supplier to one that is more relevant to organisations that manage the majority of their workloads in the cloud.

For the upcoming part of the competition, semi-finalists will conduct inspection and intervention tasks in simulation, and the technology will combine previous weather data from the Swiss firm meteoblue. Competitors will be able to access historical Abu Dhabi weather data from the past 20 years to replicate realistic weather conditions.



The simulation platform will be built on Gazebo and Robot Operating System by Open Robotics, a global leader in software for robotics simulation and application development. Gazebo is an open source 3D robotics simulator that simulates robots for a variety of applications, including marine robotics.

Dr Ray O. Johnson, ASPIRE's acting chief executive, stated, "With Open Robotics and meteoblue, we have selected two organisations that are both at the pinnacle of their respective fields. The MBZIRC Maritime Grand Challenge will push the boundaries in maritime robotics. We are providing the competitors with a world-class simulation environment with Open Robotics and meteoblue."

"We are making all the software open source," Dr Johnson continued. "Our mission is to drive the creation of transformative technologies and encourage global collaboration to help stimulate innovation across the field of maritime robotics. Sharing the technology to help make innovation easier was our goal."

Airbyte acquires Grouparoo

Airbyte, an open source data integration platform, has acquired Grouparoo, an open source startup that helps businesses sync data between data warehouses and cloud-based applications. This compliments Airbyte's products in many respects, as Airbyte concentrates on loading data into data warehouses, whereas Grouparoo focuses on operationalising that data.

"It's an open source reverse ETL [extract, transform, load] company," Airbyte co-founder and CEO Michel Tricot said. "They focus exactly on the other direction [from Airbyte]. They have a very strong technical team and they've already built a part of the product; it's going to be about how we can leverage everything that they've done and get them into the team to expand the Airbyte product for reverse ETL."



Airbyte isn't so much buying the product as it is the team's knowledge, according to Tricot. He stated, "We are not integrating their technology. It's more about their knowledge and experience." Tricot also mentioned that the Grouparoo team has always focused on making its service accessible to a non-technical audience, which is something Airbyte has also been focused on for a long time (though it also offers a command-line tool for technical users that want more flexibility, too).

Brian Leonard, Grouparoo's CEO and co-founder, wrote: "Thank you to our users and investors for your continued support. Grouparoo certainly had a set of early believers and users that saw what we were trying to accomplish.

They deployed Grouparoo in their infrastructure or on our cloud, and some even built their own plugins to extend the platform. When we really took a hard look at it, though, we were not on the right path to have the impact that we wanted to have in the world."

He felt Airbyte was on the right track, because in order to operationalise data, practically every organisation today must extract and load it into their warehouses.

ForAllSecure contributes US$ 2 million to improve the security of open source software

ForAllSecure has contributed US$ 2 million to improve the security of open source software. The firm also announced that a free version of its flagship game, Mayhem, will be available. Mayhem, the winner of the DARPA Cyber Grand Challenge and a Smithsonian Institution AI display, will now be available for free to anyone incorporating the product into any GitHub project.



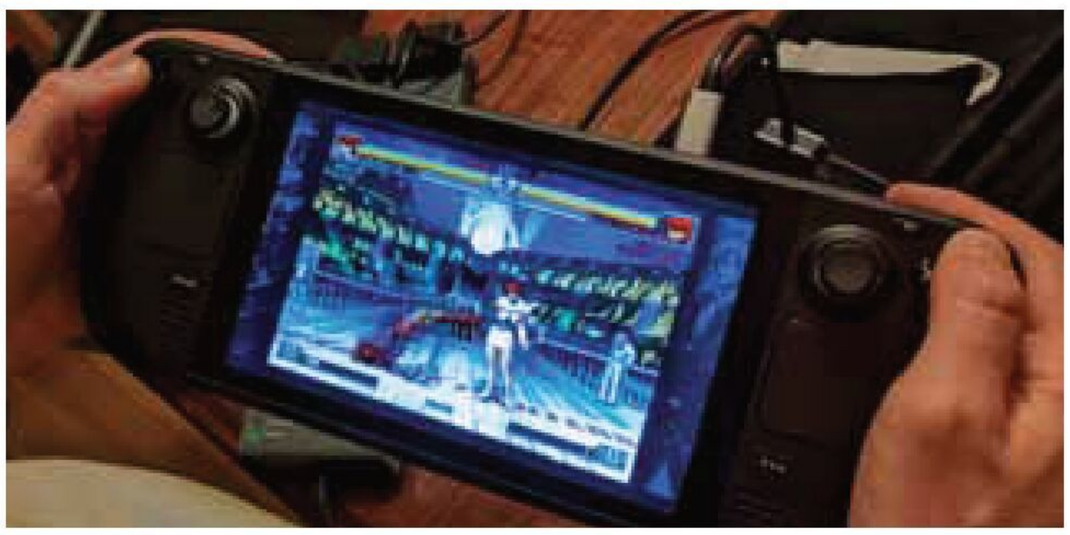
Although open source software is mission-critical, it is chronically under-tested in terms of security. OSS developers, according to the Linux Foundation, desire free security assessments and the incorporation of security to their continuous integration pipeline. With its Mayhem Heroes program, ForAllSecure is donating up to US$ 2 million to help satisfy these demands. Anyone who successfully integrates Mayhem into a qualifying OSS GitHub project will be awarded US$ 1,000.

"We're on a mission to automatically find and fix the world's exploitable bugs before attackers can succeed. OSS developers need help, and don't have access to the tools they need to quickly and easily find vulnerabilities," said David Brumley, chief executive officer and co-founder of ForAllSecure. "Our Mayhem Heroes program democratizes software security testing; it will make tens of thousands of OSS projects safer, and ultimately impact the security of systems used by everyone around the world."

Mayhem for Code and Mayhem for API, both versions of the Mayhem security tools, are now available for personal use for free. Mayhem's unique algorithms were developed at Carnegie Mellon University and were rated groundbreaking in the Cyber Grand Challenge by the US Defense Advanced Research Projects Agency.

Start gaming platforms directly from Steam Deck with this games manager

The Lutris games manager's recent upgrade, which comes with an open source solution, allows you to start a variety of top gaming platforms directly from your Steam Deck, all from a single, unified UI.



According to Computer Base, the software uses proprietary scripts

EleutherAI releases GPT-NeoX-20B, a 20-billion-parameter NLP AI model

GPT-NeoX-20B, a 20-billion parameter natural language processing (NLP) AI model similar to GPT-3, has been publicly sourced by EleutherAI researchers. The model was trained on 825GB of publicly available text data and performs comparably to GPT-3 models of similar size.

The EleutherAI blog announced the release. Using NVIDIA A100-SXM4-40GB GPUs, GPT-NeoX-20B was trained on EleutherAI's open source Pile data set. It obtained accuracy similar to a linear interpolation between OpenAI's Curie and DaVinci models when tested on various typical NLP benchmark tasks, and its one-shot performance on the MATH test data set outperformed GPT-3 175B. EleutherAI claims that GPT-NeoX-20B is the largest open source pre-trained autoregressive language model available.

EleutherAI's own software (also known as GPT-NeoX) was used to train GPT-NeoX-20B, which is built on Megatron and DeepSpeed and implemented in PyTorch. The team used model parallelism as well as data parallelism during training because the model was too huge to fit on a single GPU. Furthermore, because the team's compute budget limits made hyperparameter search "intractable," they elected to re-use the GPT-3 paper's hyperparameters.

Airbyte makes its cloud service generally available

Airbyte, an open source data integration platform for developing ELT data pipelines, has made its cloud service generally available. The company says this will make it simple for enterprises to develop durable ELT (extract, load, and transform) pipelines by allowing them to use the company's pre-built connectors or design their own.

The fast-growing company, which recently raised US$ 150 million in Series B funding, launched its cloud service in a limited beta last October. It has since improved a number of user experience aspects, including the onboarding process. It has also taken the time to learn how its cloud users' needs differ from those of its open source users.

Airbyte will compensate developers who design, maintain, and fix connectors for their platform, which is an innovative concept.

"We've always had contributions — and a growing number of contributors every week," Airbyte co-founder and COO John Lafleur said. "One way to make sure that part of the community is focused on what has the most impact for the user community is to incentivise them. We've been doing a beta test for that over the last few weeks and it's about giving some kind of bounty, some rewards, to address some issues in priorities. This means getting a connector from alpha to beta and from beta to production, and address what is needed to get to GA certification — or bug fixing."



to get runtime environments and compatibility layers like Wine to function via a subsystem. The Epic Games Store, Humble Bundle, Dolphins emulator, GOG launcher, and, of course, Steam, are all supported by these programs, commonly known as runners. It's compatible with both Steam for Windows and Steam for Linux.

Lutris also allows you to integrate local installations from a variety of sources into your game collection. Origin, Ubisoft Connect, and the Bethesda launcher should all operate well as long as the client is running in the background. Not that installing the latter is really useful right now, given Bethesda's decision to discontinue its launcher in May 2022.

Lockheed Martin releases open source spacecraft docking standard

Lockheed Martin, the aerospace company, has released an open source, non-proprietary mission augmentation port (MAP) interface standard for on-orbit spacecraft docking, which may be downloaded from its website. The company thinks that the standard will hasten the development of technology that allows docking satellites to communicate with one another.

The Lockheed Martin standard includes data that can be used by designers to develop their own MAP-compliant docking adapters, such as the dimensions of plates and petals required for a compliant physical mate of docking port halves; suggestions for electrical interfaces and docking profiles; and data that can be used by designers to develop their own MAP-compliant docking adapters.

Lockheed Martin is working on its own Augmentation System Port Interface (ASPIN), which will be MAP-compliant. Between a host spacecraft and a satellite augmentation vehicle, the ASPIN adapter provides an electrical and data link (SAV).



"Just like USB was designed to standardise computer connections, these documents are designed to standardise how spacecraft connect to each other on orbit," said Paul Pelley, senior director of Advanced Programs at Lockheed Martin Space. "We believe it's in the best interest of the nation for the industry to have common interface standards to provide mission agility and enterprise interoperability."

Trezor is being targeted for phishing attacks

Phishing assaults are particularly prevalent in the crypto and NFT industries. Trezor, a Bitcoin wallet service, is now being targeted by hackers. Trezor's email list was used to target customers and fool them into downloading a fraudulent version of the program aimed at stealing their crypto assets, according to *Bleeping Computer.*

The original Trezor software is open source. When the software is installed, it prompts the user for a recovery phrase, which they should have entered when

they first set up their wallet. This recovery phrase serves as a key to regaining access to the wallet if it is misplaced. It's game over once the user enters the key. The scammers will receive the recovery phrase and will be able to claim all your crypto assets.



When it comes to phishing scams, distinguishing a reputable site or program from a phoney can be quite difficult. Due to the usage of unusual characters, even the websites related with the download for this fraud appeared authentic. It's usually a good idea to double-check anything that requests a security key or password; there are definitely dragons lurking about.

Trezor believes that this particular dragon was co-opted for malevolent motives and targeted one of its newsletters hosted on the automated email provider Mailchimp. Mailchimp admitted that an insider had targeted crypto firms, according to Trezor, but the company has yet to publish a statement on the matter.

MHKiT gets updated for modelling marine energy

The Marine and Hydrokinetic Toolkit (MHKiT) is a massive, searchable, open source knowledge hub that provides developers with the code needed to analyse how well their technology might perform in various ocean and river sites. The toolbox has been upgraded to include data on tidal and river energy resources as well as high waves, elements that affect how equipment works underwater (such as turbulence and sediment), and more.

Developed in 2019 with funding from the US Department of Energy's Water Power Technologies Office, MHKiT is accessible in Python and MATLAB. For each platform, the MHKiT team is releasing increasingly robust versions.

MHKiT now has data about tidal and river resources on its Python platform, including how much energy is available at which areas, how fast and turbulent the waves are at certain sites, and what underwater barriers, such as sand and stones, might obstruct a device's performance.

In addition, wave energy producers can now simulate the extreme waves that their machines may encounter at their chosen deployment location. Unprepared devices may not survive the waves; with MHKiT, engineers can improve their chances of survival.

RoninX is building an open source protocol for decentralised content streaming

RoninX is building and developing an open source glass-to-glass (creation-to-viewing or end-to-end) protocol to enable decentralised 8K quality streaming, according to Sandeep Casi, the company's president. He also stated that the foundation is transferring large amounts of data to decentralised storage.

The RoninX Foundation, a non-profit organisation that aims to bring together pioneers in camera technology, streaming content, and blockchain to change the media and entertainment sector, has been unveiled by three renowned organisations.

Streamonix, IoTeX, and Videogram have joined forces as founding stewards of the world's first endeavour to encourage the education, advancement, and adoption of decentralised content management and streaming. In the next few months, the RoninX Foundation will look to add a few more founding stewards, such as camera manufacturers, content studios, and streaming service providers.

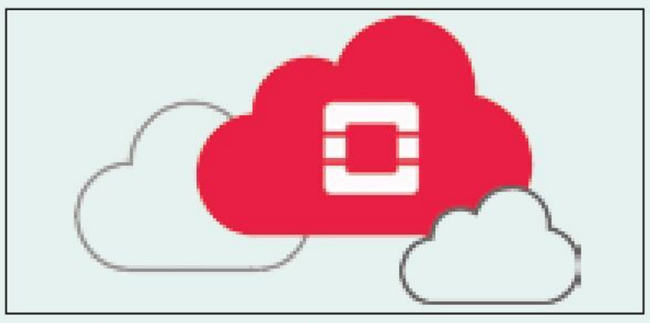
"We are a non-profit educational foundation that operates with full transparency and believes in open source," explained the chair, Anoop Nannra. "We also have a permissive licensing model, which is why we work with Apache and MIT licensing, which essentially allows developers to do whatever they need to or want to with our software."



OpenStack releases Yoga, a new IaaS cloud

The OpenStack community has released Yoga, the 25th version of open source cloud infrastructure software. Yoga highlights include support for advanced hardware features such as SmartNIC DPUs, improved integration with cloud-native software such as Kubernetes and Prometheus, and reduction of technical debt to maintain a stable and reliable OpenStack core.

OpenStack, the open infrastructure-as-a-service standard, is the one infrastructure platform for deployment of diverse architectures—bare metal, virtual machines (VMs), graphics processing units (GPUs) and containers. It is one of the most active open source projects in the world, supported by a vibrant and engaged community of developers globally. Over the span of just 25 weeks, almost 13,500 changes authored by over 680 contributors from more than 125 organisations and 44 countries were merged into the Yoga release.



The feature advancements in the new release include hardware enablement extended for SmartNIC DPUs. Local IP is added to Neutron, focused on high efficiency and performance of the networking data plane for very large scale clouds, or clouds with high network throughput demands.

Some other basic changes include that Ironic, OpenStack's bare metal provisioning program default deployment boot mode, now supports UEFI instead of legacy BIOS. Meanwhile, Kolla is depreciating binary images. In the next release, all binary image support will be removed. Users must migrate to source-based images going forward.

The MHKiT team uses the MATLAB platform to import a large amount of data from a fleet of buoys, such as wave heights, current speeds, and ocean temperatures. The team has also included historical wave data to help predict the weather and climates that marine energy technology may encounter offshore. This so-called hindcast data set will eventually encompass the whole of United States.

MHKiT is a popular tool that has been downloaded over 4,000 times in just four years (although some of those downloads are repeat customers coming back for more).

Google and Linux Foundation release Nephio for telecom market

The Linux Foundation has extended its push to bring open source projects into the telecom market by partnering with Google Cloud to develop the Kubernetes based Nephio project. This project aims to make it easier for telecom operators to deploy and manage multi-vendor cloud infrastructure and network operations across large-scale edge deployments by providing Kubernetes based cloud-native intent automation and automation templates.

Nephio sits on top of a Kubernetes substrate, either directly handled by an operator or via a hyperscaler-based platform such as Google Config Connector, Amazon Web Services (AWS) Kubernetes Controllers, and Azure Service Operator.

Kubernetes' cloud-native orchestration capability is used to enable operators to roll out and manage new services in their 5G and edge deployments. This will allow for a speedier on-boarding of network functions into a production environment, similar to how hyperscalers employ the DevOps process.

"We believe that true cloud nativeness lies in the fact that Kubernetes can actually automate all the way down from describing a network function or a service to the infrastructure including all the underlying networking components," stated Gabriele Di Piazza, senior director of product management for telecom at Google Cloud.

The project is open about the fact that it does not provide end-to-end service orchestration, cloud infrastructure, or network operations.

The project's largest hurdle appears to be providing a "carrier-grade" option for telecom operators as they build up their 5G networks. This has long been an issue since telecom operators compete in a highly regulated market with government and public safety regulations that often exceed those of traditional cloud based operators.

Oracle makes Solaris 11.4 CBE available

Oracle has started making a new version of Solaris. Oracle Solaris 11.4 CBE was announced in early March this year, and is the 'Common Build Environment' for open source developers and strictly non-production personal use, which is required if you want Solaris for new installations after 2022.

The new Solaris 11.4 CBE is effectively a rolling release, and Oracle hopes that by doing so, it will make it easier to integrate the open source software that Solaris relies on, rather than being locked onto the dated 11.4.0 GA release.

An Oracle account is required to download the new Solaris 11.4 CBE. The CBE builds are sometimes defined as "pre-release builds of a particular SRU, analogous to a beta." The Oracle Technology Network Early Adopter License Agreement for Oracle Solaris grants the non-production usage licence.

Under Oracle support contracts, users will be able to upgrade from these free CBE releases to paid SRU releases. The Oracle Solaris blog has more information for people interested in Oracle Solaris 11.4 CBE.



Cocos raises US$ 50 million to develop its game engine further

Cocos Technology, the Beijing based digital interactive content development platform founded in 2010, has announced a US$ 50 million funding round to support the development of its open source Cocos Creator game engine. Cocos is a platform for creating 2D games for mobile and other devices that has been around for a long time. With a new version of the Cocos game engine, which traces its roots back a decade to Cocos2d-x, it also included its first all-in-one 3D engine and editor last year.



To-date, Cocos has been used by over 1.4 million developers in over 100,000 games available in app stores, and by over 1.6 billion users.

CCB Trust, GGV Capital, Agora, and others are among its investors. Cocos will use this round of funding to continue to expand the engine's fundamental technology and push the engine's greater integration into numerous industries, including gaming, autos, education, XR, home design, architectural engineering design, and other future ventures.

Baidu released its product 'Xi Rang' toward the end of 2021 to speed up the formation of the metaverse. Cocos has recently joined Baidu as a partner, offering a technical basis for the metaverse. Over time, the two companies will work together to create a metaverse universe.

Directus releases a patch to fix an XSS vulnerability

According to a recent advisory from the Synopsys Cybersecurity Research Centre (CyRC), a stored cross-site scripting (XSS) vulnerability in the widely used Directus content management system (CMS) might lead to account compromise in the service's admin application if not swiftly remedied.

CVE-2022-24814 is an open source, Web based framework used to manage SQL-based databases and connect their contents via an application programming interface (API) into multiple clients or websites. It was discovered and identified by CyRC researcher David Johansson.

CVE-2022-24814 is identical to two previously reported vulnerabilities – CVE-2022-22116 and CVE-2022-22117 – and bypasses a previous mitigation provided in Directus 9.4.2 for these bugs. It has a CVSS base score of 5.4, indicating that it has a medium influence.

It allows an authenticated user with Directus access to craft a stored XSS attack that executes automatically when other users view collections or files in Directus by abusing its file upload functionality.

The vulnerability was first reported on January 28, 2022, and confirmed on March 7, 2022. Directus released version 3.7.0 on March 18th, which includes a remedy for CVE-2022-24814. Users who haven't updated to this version yet should do so immediately. Directus, according to Synopsys, had moved quickly and responsively throughout the investigation.

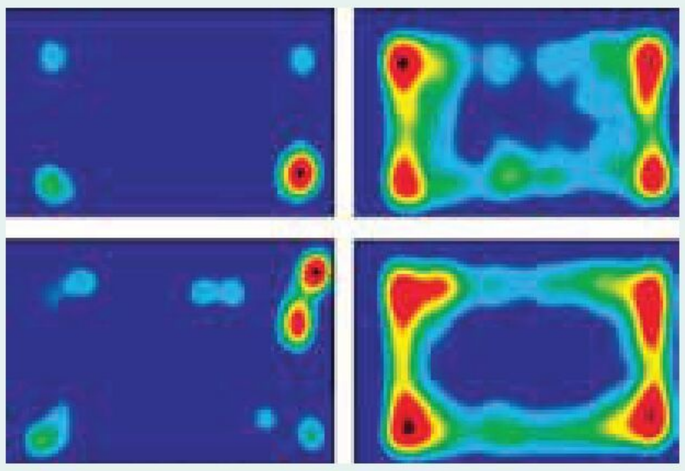


Open source algorithm can improve the reliability of autistic mice study

A new open source program has the potential to increase the reliability and depth of data acquired from a marble-burying assay, which is routinely used in autistic mice models to infer anxious or repetitive behaviours. The program tracks a mouse's movement, counts how many marbles it buries, and keeps track of how long it takes to do so.

Researchers generally photograph mice in a cage with many marbles in order to do this test. They take another photograph after a specified period of time has passed and compare the two. Mice with anxiety-like or repetitive habits are known to bury more marbles than usual.

"The marble-burying assay, the way it's currently analysed, does show reproducible results," stated study investigator Lucas Wahl, a graduate student in Aleksandra Badura's lab at Erasmus University in Rotterdam, the Netherlands.



However, according to Badura, associate professor of neurology and the study's lead investigator, labs do the assay differently. "Behavioural studies are a little bit of a messy ground," he said. Machine learning is used to standardise the procedure in the new tool, which was detailed in *eNeuro* in March and is freely available online.

Intel takes over the Open Networking Foundation's development team

Intel has purchased the Open Networking Foundation's (ONF) development team as well as Ananki, an independent venture-backed company that was recently spun out of the ONF to supply open source based software-defined private 5G as a commercial service, according to a release. The financial details have been kept under wraps.



Guru Parulkar, the ONF executive director, has joined Intel as vice president of the Network and Edge Group, reporting to Nick McKeown, the founding ONF board member and senior vice president and general manager of Intel's Network and Edge Group. In addition, the majority of ONF's internal development team will join Intel's workforce.

As it transitions away from code development by an internal team, the Open Networking Foundation (ONF), which was founded in 2011 with the mission of transforming networking by championing software-defined networking (SDN), disaggregation, and open source, has released all of its projects to open source with permissive software licences. ONF has made its full portfolio of production-ready platforms open sourced, including solutions for public 5G (SD-Core, SD-RAN), private 5G networks (Aether), software-defined broadband (SEBA/VOLTHA), and P4 programmable networks (SD-Fabric, PINS).

"ONF has made important contributions to the industry with its software-defined development projects, and I am confident that its move to a community-sourced model as part of its next stage of growth will fuel further contributions as the industry and its needs advance," said Nick McKeown. "As a result of ONF's shift in direction, we are pleased to welcome ONF's development team to Intel."

For more news, visit *www.opensourceforu.com*

EFYGROUP
YOURS SINCE 1969

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

TEAR OUT — **ORDER FORM** — TEAR OUT

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_____

_____City_____

Pin Code_____State_____Phone No._____ Email_____

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

"It's time to contribute to open source"

Nilesh Vaghela is an ***AWS community hero*** and ***founder, ElectroMech Corporation,*** a cloud and open source company. According to him, contributing to open source is a rewarding act in itself. However, it needs commitment and there are many steps involved in the process, right from selecting a project to ensuring your contribution is noticed. In a conversation with ***Abbinaya Kuzhanthaivel*** of ***OSFY,*** he shares a few tips on how developers can help to improve India's contributions to open source.



Nilesh Vaghela,
AWS community hero and
founder, ElectroMech Corporation

Q. Can you tell us a bit about your current role and contributions to open source.

A. I am currently an architect working on automation. I lead multiple teams and also contribute majorly to Invinsense, an open source security service platform. I started open source groups in early 1998 and had around 1500 members even then. A group *(https://groups.google.com/g/vglug)* I am handling now has been very active since 2014-15.

Q. How did you start working with open source projects?

A. I am a mechanical engineer by qualification, and was dealing with modems and UPS systems at my firm ElectroMech Corporation. I slowly got dragged into handling PCs, networking and Linux. I started experimenting in 1996 after getting inspired by seeing over 150 computer servers running on Linux at a Nuclear Science Centre. That's when I converted my company entirely, focusing on open source for training and support.

I can proudly say that I was one of the first and early adopters of open source — helping customers to understand what is open source, what are its benefits, what's available for free, security or code issues, and so on. We got at least four or five customers in Vadodara supporting us, and we eventually promoted ourselves through advertisements in the Yellow Pages. We partnered with Red Hat and the journey continues till now.

Q. How do you think open source has evolved since then?

A. I can say that, earlier, open source was a passion that fascinated and attracted people to participate. I was particularly impressed hearing user-friendly stories across the world when some Siberian contributors were working to improve water scarcity. It was more like a corporate social responsibility (CSR) activity. A board would be created by people and experts to govern and take projects forward. People would come in for the love of technology without expectations.

I did not believe then that open source could get commercial, but it is the driving force for most of the innovation and technology today, and many more enterprises are adopting it. We are looking forward to a great balance in contribution to and use of open source, as we have people, communities and big companies coming into play. This is the real future and power of open source.

Q. Could you share some of your challenges?

A. Initially, I walked alone but people joined once they knew my intentions were good. I created a lot of communities without any expectation, but did get paid indirectly in terms of reputation or fame; some people understood that I was a technical expert and gave projects in the long term. As it was very early days, people started joining the community and contributing without much effort. The goal wasn't to get business and hence I can say I didn't really face any hurdles.

Q. What are your leadership mantras and lessons from being a community leader?

A. First, if you want to start a community, then be neutral and don't harbour a biased opinion. It may look like you are running a community as a leader, but remember those joining you are contributing equally. And never demotivate anyone. Be polite while making comments and addressing queries. Whatever the question, if you don't want to give an answer, then choose to be quiet. But don't stop anyone from asking. Help them build expertise.

Second, don't involve the community in business. Do not mix and match the goals of your business and community. Have a clear differentiation.

Always try to encourage people to participate instead of delivering instructions as an expert. If you find people are interested to lead and take initiatives, then give them the stage. Invite and engage them in the community. That will help you to make more community leaders. Also, keep your community simple and don't involve sponsors in the initial stages.

Q. Who do you look up to for inspiration?

A. Richard Stallman, the father of the open source movement, is my inspiration and I have always admired his projects.

Apart from him, I have an interesting incident to share that inspired me to work on open source. At the time when I started working on open source, most of the software for the Nuclear Science Centre was based on the Windows OS. However, many scientists wanted to work with Linux based software. And within two or three months, they actually created Linux drivers. This is what fascinated me – that the user can create these drivers which may not be possible in the case of proprietary software. I really liked the fact that open source empowered the user.

Q. Your thoughts on the open source landscape in India and the scope for improvement.

A. India is the largest consumer of open source and we are focusing on becoming a contributor. With so many developers around, we still do not have software giants in India. What we have mostly are service providers and not innovators. More people should become contributors to open source and develop something with international labels.

The very thought of contributing to open source should begin from the level of schools and colleges. Fortunately, the Gujarat government has already introduced lessons based on Linux from Class 8 to Class 10. It is important to educate and make youngsters aware of open source models.

Second, we have to develop good mentors. When people start contributing, it is important to find an open source mentor working in that particular project. The mentor gives a small assignment, tries the code and then commits it. If everything goes fine, the contribution is increased gradually. Unfortunately, we have very few mentors available in India. We need to prepare a lot of mentors or maybe connect to those across the world.

Third, we need to encourage those who come forward to contribute. Once you are a recognised developer or a person contributing to open source development, you also progress in your career and business.

India can be a major contributor to open source by following such simple methods.

Q. What do you think about the coding requirements to contribute to open source?

A. From my experience, if you know the internal parts, how to develop the application, what code standard you should maintain, and how to manage the team and other best practices, you may not have to actually worry about coding expertise.

There are other roles too with respect to designing, security maintenance and integration. See what works for you. Develop and strengthen your skill in what you like



The team at ElectroMech Corporation

to do. If you feel coding still interests you, then take the support of fellow developers to learn it.

Q. How do you shortlist a project you would like to contribute to?

A. You need to understand your top few interest areas and then do your research on the projects happening around them. You need to figure out the area of requirements or openings for contributors and more volunteers. You can start small to practice and then build expertise.

Avoid going by the trendy topics; what's important is your individual interest. For instance, DevOps is in high demand now and you may tend to go for a DevOps project. Do not make this mistake.

You can find open source projects on Cloud Native Computing Foundation (CNCF), Apache, Fedora, Red Hat, and so on. This way you can also find mentors who are already working on projects and can get proper guidance.

Q. Projects have their own purpose and target audience. Sometimes they even misalign with open source goals. So, what does one check before contributing?

A. I agree it becomes challenging when somebody starts an open source project and then commercialises it. But this is always a risk, and should not discourage you.

First try to check out the group — how popular are the contributors working in the group, how long have they been contributing, and how reputed are they. And once you join, observing everyone and everything is the key. Try to learn at least for three to six months, and understand how everything works. You can always leave the project if you find the intention is wrong. But if you feel it's all right, then go ahead and contribute.

There are certain licence checks that you can do, say, like GPL version 3. You can also look at unmodified licence versions like the Apache open source licence.

Q. Do you think big companies will accept contributions from freshers?

A. Yes, of course. Companies also like mentoring. They usually don't allow you to contribute directly, but may give you a small assignment initially. A mentor will first try to understand what skill you have and how good you are at it. Once they recognise you have the kind of skill that is needed, they will continue to guide you or assign you to some other mentor based on your skill. The initial stages are very crucial. Many companies do some sort of screening, and you may be allowed to contribute only after you have proved your ability.

Q. What are the initial challenges that contributors have to overcome when picking up projects?

A. First, you should be very serious about your contribution. There are no written commitments and contributors may tend to take the work lightly. That's totally wrong. Try to dedicate 8-10 hours or whatever is feasible each day. If you are skipping this commitment because you feel there are no immediate returns, then you will not be a good contributor.

Always adhere to a mentor's guidance strictly at the initial stages. This is very crucial for a healthy contribution. Sometimes it may happen that you believe you are good at something, while your mentor may not assign a project based on that skill. Simply approach your mentor in such scenarios and ask him what you should do, what is your role, and how you can contribute.

Q. Many developers do not get replies after submitting the contribution to a project. How does one make a submission noticeable?

A. Write a small blog on the project you are planning to contribute to, covering aspects like what you like in it, what you don't like, and what can be improved. Such a positive and promotional approach can greatly help you.

Be a part of the group and be involved in activities related to that particular project. Instead of contributing, first try to engage with the group, and this will increase the chances for you to get adopted as a contributor.

Once you have a better understanding of the project, not only will your work be accepted but you will be able to better align yourself with that project.

Q. How do you overcome a situation where your contribution is not accepted?

A. Just understand that this can happen for many reasons — maybe you are not in the right project or you have not contributed correctly. If the project is country driven, your request may not be accepted. Hence, remember to have a checklist as stated earlier. Don't worry if your contribution is not accepted because either you are not fit for the project or the project is not fit for you.

All I would recommend is try to identify four or five projects, and at least one among those projects you work on will probably be accepted.

Q. What is your message for our readers?

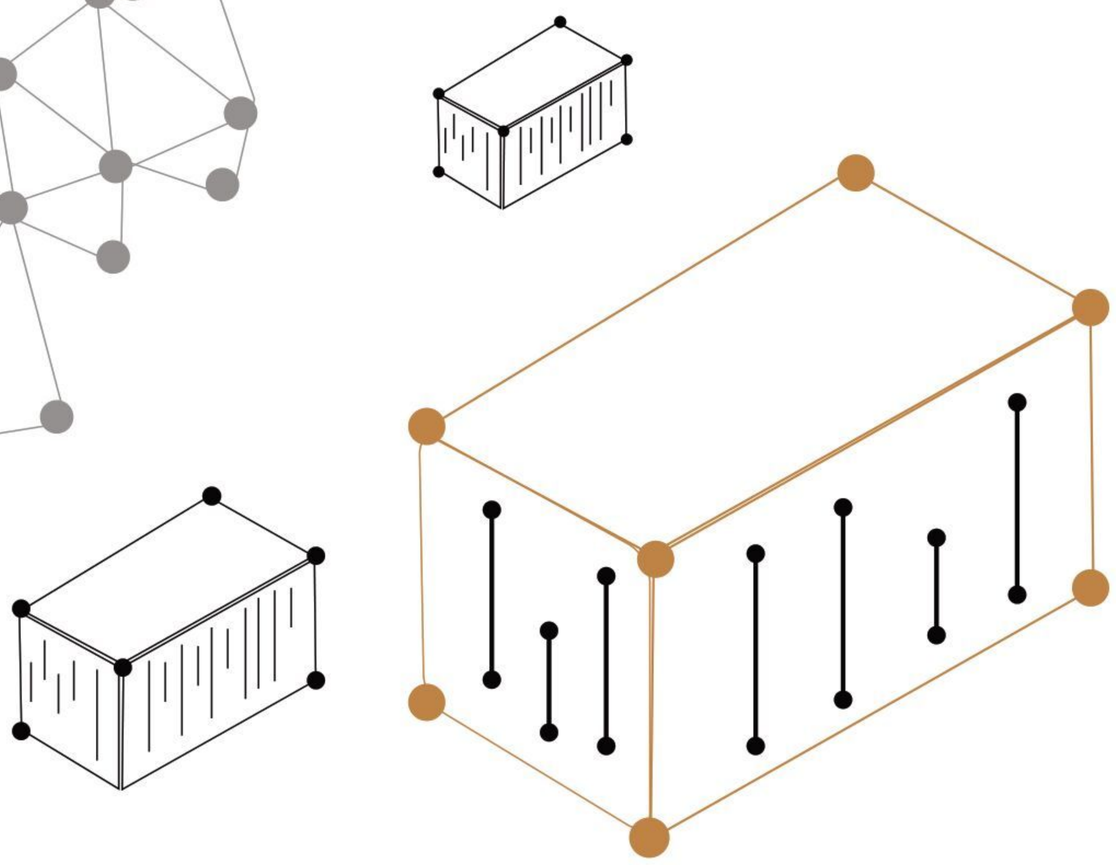
A. Open source is the driving force behind most of the innovation happening today. Instead of just using open source, let us try to contribute according to our capacity and skills. Contributions can be in terms of code, documentation, testing, blogs, money, etc. It's time to contribute.

Q. Any hiring plans for ElectroMech Corporation — what are the roles and skill expectations?

We have requirements in cloud DevOps, and are hiring cloud architects, Python developers, Linux architects and security professionals. END

OSFY Magazine Attractions During 2022-23	
Month	**Theme**
March 2022	Security, Network Management and Monitoring
April 2022	Open Source Programming (Languages and tools)
May 2022	DevOps Special
June 2022	AI, Deep learning and Machine Learning
July 2022	Database management and Optimisation
August 2022	Cloud Special: Everything from management to implementation
September 2022	Mobile/Web App Development, Optimisation and Security
October 2022	Blockchain and Open Source
November 2022	Open Source and IoT and Edge
December 2022	Data science, Data security, Data Management,Storage and Backup
January 2023	Docker and containers
February 2023	Open Source on Windows and Best in the world of Open Source (Tools and Services)

Containerisation of Java Microservices



Frequent roll outs of microservices demand self-contained units that ensure safe and successful deployments. Containerisation offers an elegant solution. In this part 7 of the Design Odyssey, we will build and deploy the AddService of UMS as a Docker container.

In the previous part of this series, it was observed that a packaged Spring Boot application carries an embedded Tomcat Web server along with class files and dependent libraries. However, the JRE was left out of the application package. The deployment engineer is still required to find and install a suitable JRE before deploying the application. How about including the JRE also into the package and distributing it so that the deployment can be automated fully?

That's exactly what containerisation is!

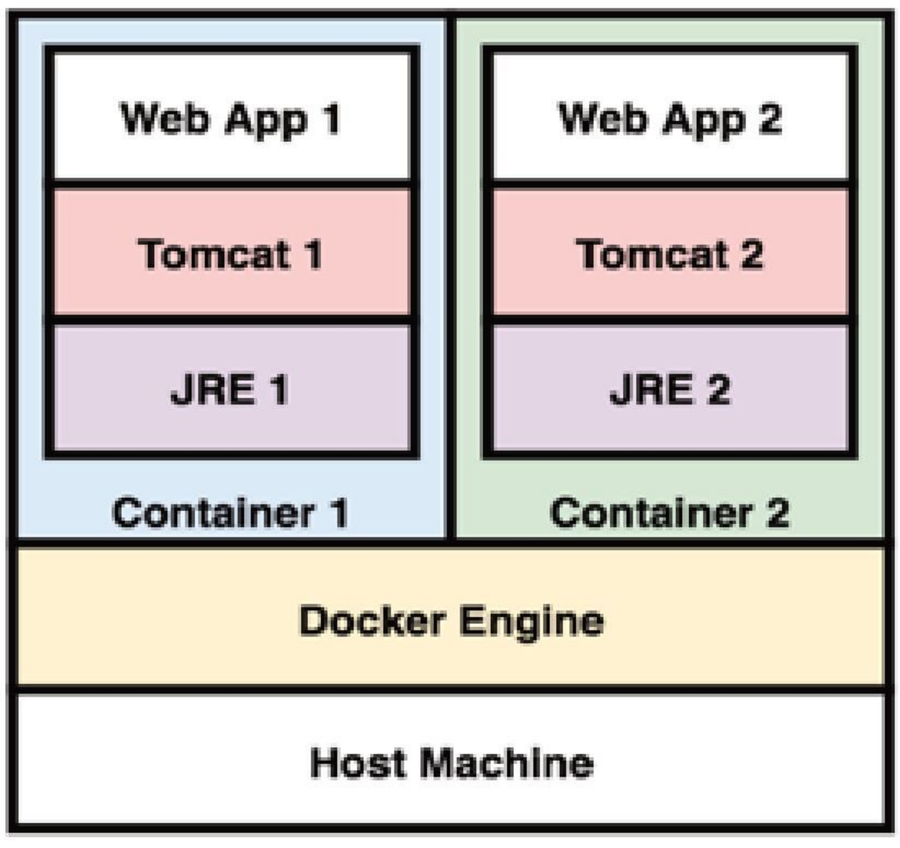


Figure 1: Deployment unit with containerisation

As depicted in Figure 1, a container is a deployment unit that consists of not only the application code but also the required operating environment, which includes artifacts like JRE, Python runtime, Node runtime, etc, as required.

Docker containers

In theory, the concept of containerisation is not very new. For example, the Linux operating system has already been offering the necessary bits and pieces in that direction. However, credit goes to the Docker containers for the popularity of this approach in the last decade.

The architecture of Docker containers can be summarised as shown in Figure 2. It consists of: (1) a Docker Hub, which is a repository of distributable images of Docker containers,

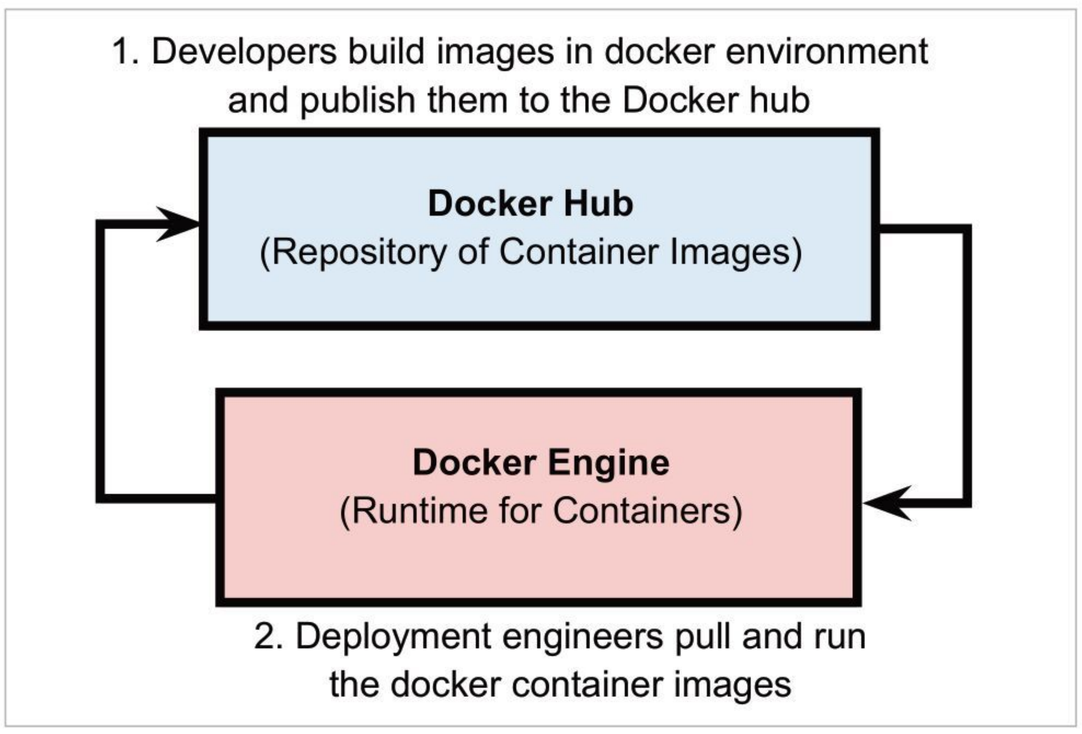


Figure 2: The Docker architecture

and (2) a Docker Engine, which runs the Docker containers.

The architecture is well supported by the other ecosystem players like Git, Jenkin, Maven, etc.

As part of the development cycle, the engineers write and push the code to Git repositories. As part of the build cycle, build tools like Jenkins pull the code from Git repositories, build the Docker image and push it to the Docker Hub.

As part of the deployment cycle, the Docker runtime pulls the image from Docker Hub and creates Docker containers that bring the applications to life. Any number of images can be pulled and launched on a single instance of Docker runtime. Where is the magic?

Docker containers are language-agnostic. They are capable of running applications on any platform. For example, they can run an Express Web service on Node or a Flask Web service on Python, or a Spring Boot Web service on JRE. The answer is simple. The container just offers a sandbox with a Linux environment. The container image must package everything else that is required for the application. For example, the image of a Flask service must include Python runtime into the container image and similarly, the image of a Spring Boot service must include JRE.

Let's explore the last option a bit more.

Spring Boot and Docker container

The Maven build tool packages a Java application. The Docker build tool packages a container. Just like the *pom.xml* acts as the manifest file for the Maven build tool, *Dockerfile* acts as the manifest file for the Docker build tool.

A *Dockerfile* specifies the root container as the prerequisite, and presents directives to be followed in order to create the image. For instance, the following is a manifest for building a Docker image of a Spring Boot Web application.

```
FROM maven:3.5-jdk-8
COPY src /usr/glarimy/src
COPY pom.xml /usr/glarimy
RUN mvn -f /usr/glarimy/pom.xml clean package
EXPOSE 8080
ENTRYPOINT ["java","-jar","/usr/glarimy/target/ums-add-service.jar"]
```

Let's decipher it line by line.

```
FROM maven:3.5-jdk-8
```

The above directive specifies *maven:3.5-jdk-8* as the root container. This makes sure that a Docker container is started with a JDK of version 8 and the Maven of version 3.5 is available for the rest of the packaging process. This container is used only for building the image. Let's call this a build-container.

```
COPY src /usr/glarimy/src
COPY pom.xml /usr/glarimy
```

The above directives create */usr/glarimy* folder on the build-container and copy the code from a local *src* folder to a */usr/glarimy/src* folder on the build-container. They also copy the local *pom.xml* file to */usr/glarimy* on the build-container.

```
RUN mvn -f /usr/glarimy/pom.xml clean package
```

This directive kicks off the Maven build process to package the Java code as per the referred *pom.xml*. Note that all this happens on the build-container. Once the process is completed, the container image for the Spring Boot application is ready. This image can be run as a container, which we refer to as deployment-container. However, we still need to specify the directives on how to launch this on the deployment-container.

```
EXPOSE 8080
```

The above directive specifies that port 8080 be opened up for traffic on the deployment-container when this image runs.

```
ENTRYPOINT ["java","-jar","/usr/glarimy/target/ums-add-service.jar"]
```

This directive specifies the command to launch the application on the deployment-container.

There are several other directives possible, but the approach is more or less similar. Once the *Dockerfile* is ready, and the Docker is installed on a given machine, the following command on a Docker runtime is sufficient to build the container image.

```
docker build -t glarimy/ums-add-service .
```

Observe that the last token of the above command points to the location of *Dockerfile*, which is the current folder in this case.

The following command lists all available images on the local Docker Engine:

```
docker images
```

Figure 3 presents a sample output of the above command. This list includes the images that are pulled from the Docker Hub as well as built locally.

It's not useful if the created Docker image of the application just lies on the local Docker Engine. In order

```
$ docker images
REPOSITORY                      TAG          IMAGE ID       CREATED              SIZE
glarimy/ums-add-service         latest       7cd1a919486b   44 minutes ago       764MB
<none>                          <none>       64288c003081   47 minutes ago       635MB
<none>                          <none>       cb7ec9732570   About an hour ago    764MB
maven                           3.5-jdk-8    985f3637ded4   3 years ago          635MB
```

Figure 3: Listing the Docker images

```
$ docker container ps
CONTAINER ID   IMAGE                          COMMAND                  CREATED          STATUS
PORTS                    NAMES
d3849b745533   glarimy/ums-add-service   "java -jar /usr/glar…"   13 seconds ago   Up 13 seconds
0.0.0.0:8080->8080/tcp   ums-add-service
```

Figure 4: Listing the containers

for it to be deployed on the target infrastructure, the image needs to be distributed. The following command publishes the local image on the Docker Hub so that anyone can pull it for deployment. Of course, an account needs to be created on *https://hub.docker.com* before pushing the image to the hub.

```
docker push glarimy/ums-add-service
```

The images on the Docker Hub can be searched online at *https://hub.docker.com/*. Individuals and organisations can create private spaces to ensure the confidentiality of the images. In fact, automated builds can also be set up on Docker Hub in such a way that it listens to the code commits on Git and kicks off building a new version of the image immediately. Organisations may use several other tools like Jenkins and design a CI pipeline.

One way or another, once an image is available in the hub, it can be pulled to any Docker machine with the following command:

```
docker pull glarimy/ums-add-service
```

And the following command launches a new Docker container from the image and exposes it to the clients:

```
docker container run -d --name ums-add-service -p 8080:8080
glarimy/ums-add-service
```

This command may look a bit complicated, but it's not really so once it is understood.

- The *docker container run* is the command that runs the container.
- The *-d* switch makes it a daemon process so that the prompt is back.
- The *-name* switch names the container as expected. Without this switch, the Docker Engine generates a name that may not be intuitive.
- The *-p* switch maps the port 8080 of the container to the port 8080 of the host machine so that the clients outside of the container can also access the service.
- And the last part is the actual name of the image that needs to be run, which is *glarimy/ums-add-service* in this case.
- The list of containers running currently can be found using the following command:

```
docker container ps
```

Now the service can be accessed from the host machine using the appropriate CURL command or using any REST client, just like it was done in the previous parts of this series.

Next steps

Though the Docker containers solve the problem of building ready-to-run self-contained images that can be deployed onto any Docker machine, this is certainly not sufficient. What if microservices running in two different containers want to access each other? What if a microservice wants to access a database server? What if N number of instances of a given service must always be run to handle the load?

The answer lies in networking and orchestrating the containers, which will be covered in the next part of this series. END

By: Krishna Mohan Koyya

The author is the founder and the principal technology consultant at Glarimy Technology Services, Bengaluru. He has delivered more than 500 programs to corporate clients on design and architecture in the last 12 years. Prior to that, he held various positions in the industry and academia for a decade.

AI Accelerators: A Long Way to Go

We are witnessing the AI revolution. In order to make smart decisions, the devices equipped with AI need lots and lots of data. And making sense of that data needs power and speed. This is where AI accelerators come into the picture.

As IoT systems are getting more and more efficient, data acquisition is getting easier. This has increased the demand for AI and ML applications in all smart systems. AI based tasks are data-intensive, power-hungry, and call for higher speeds. Therefore, dedicated hardware systems called AI accelerators are used to process AI workloads in a faster and more efficient manner.

Co-processors like graphics processing units (GPUs) and digital signal processors (DSPs) are common in computing systems. Even Intel's 8086 microprocessor can be interfaced with a math co-processor, 8087. These are task-driven additions that were introduced because CPUs alone cannot perform their functions. Similarly, CPUs alone cannot efficiently take care of deep learning and artificial intelligence workloads. Hence, AI accelerators are adopted in such applications. Their designs revolve around multi-core processing, enabling parallel processing that is much faster than the traditional computing systems.

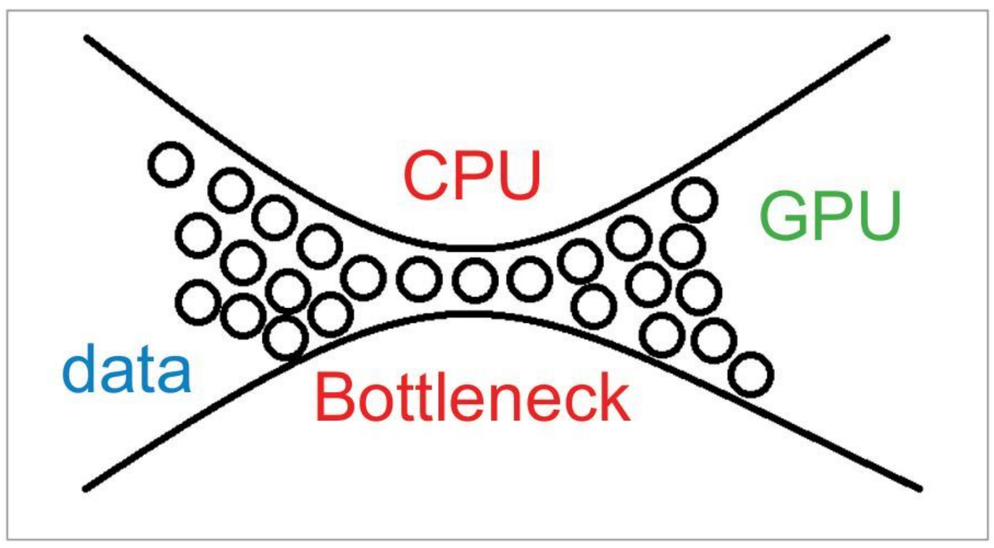


Figure 1: CPU bottleneck

At the heart of ML is the multiply-accumulate (MAC) operation. Deep learning is primarily composed of a large number of such operations. And these need to happen parallelly. AI accelerators have the capacity to significantly reduce the time it takes to perform these MAC operations as well as to train and execute AI models. In fact, Intel's head of architecture, Raja Koduri, noted that one day every chip will be a neural net processor. This year, Intel plans to release its Ponte Vecchio HPC graphics card (accelerator), which, Intel claims, is going to be a game-changer in this arena.

AI in smartphones

- Apple's iPhone 13 series comes with A15 bionic chips that contain dedicated neural network hardware
- OnePlus 9 Pro and other phones featuring Qualcomm's Snapdragon 888 processor have AI capabilities in their cameras thanks to Qualcomm artificial intelligence (AI) engine
- Samsung Galaxy S20 series includes an AI-powered camera
- Google's Pixel 6's processor has features powered by AI and ML

User-to-hardware expressiveness

'User-to-hardware expressiveness' is a term coined by Adi Fuchs, an AI acceleration architect in a world-leading AI platforms startup. Previously, he has worked at Apple, Mellanox (now NVIDIA), and Philips. According to Fuchs, we are still unable to automatically get the best out of our hardware for a brand new AI model without any manual tweaking of the compiler or software stack. This means we have not yet effectively reached a reasonable user-to-hardware expressiveness.

It is surprising how even slight familiarity with processor architecture and the hardware counterpart of AI can help in improving the performance of our training models. It can help us understand the various bottlenecks that can cause the performance of our models to decrease. By understanding processors, and AI accelerators in particular, we can bridge the gap between writing code and having it implemented on the hardware.

Bottleneck in AI processing

As we add multiple cores and accelerators to our computing engines to boost performance, we must keep in mind that each of them has a different speed. When all

these processors work together, the slowest one creates a bottleneck. If you use your PC for gaming, you would probably have experienced this problem. It does not matter how fast your GPU is if your CPU is slow.

Similarly, it would not matter how fast hardware accelerators are if the read/write data transfer is slow between RAM storage and the processor. Hence, it is crucial for designers to select the right hardware for all components to be in sync.

Popular AI accelerator architectures

Here are some popular AI architectures.

Graphics processing unit (GPU): Well, GPUs were not originally designed for ML and AI. When AI and deep learning were gaining popularity, GPUs were already in the market and were used as specialised processors for computer graphics. Now we have programmable ones too, called general-purpose GPU (GPGPU). Their ability to handle computer graphics and image processing makes them a good choice for usage as an AI accelerator. In fact, GPU manufacturers are now modifying their architectures for use in AI or ML.

In 2012, Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton from the University of Toronto presented a paper on AlexNet, a deep neural network. It was trained on readily available, programmable consumer GPUs by NVIDIA (NVIDIA GTX 580 3GB GPU). It is basically a GPU implementation of a CNN (convolutional neural network). AlexNet won the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC).



Figure 2: AlexNet's most probable labels on ImageNet images, with five labels considered most probable for each. The probability assigned to each label is also shown by the bars (Credit: Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton)



Figure 3: Xilinx' Alveo U50 data centre accelerator card that comes with the Zebra software (Credit: Xilinx)

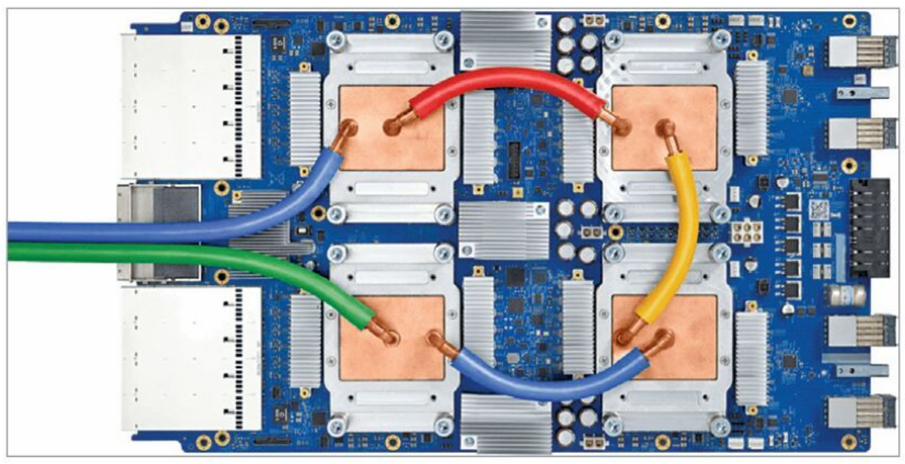


Figure 4: Cloud TPU v3 (Credit: Google)

FPGAs and ASICs: FPGAs are semiconductor devices that, as the name suggests, are field programmable. Their internal circuitry is not mapped when you buy them; you have to program them right on the field using hardware description languages (HDLs) like VHDL and Verilog. This means that your code is converted into a circuit that is specific to your application. The fact that they can be customised gives them a natural advantage over GPUs.

Xilinx and an AI startup called Mipsology are working together to enable FPGAs to replace GPUs in AI accelerator applications, just using a single command. Mipsology's software Zebra converts GPU code to run on Mipsology's AI compute engine on an FPGA, without any code changes.

FPGAs are cheaper, reprogrammable, and use less power to accomplish the same work. However, this comes at the cost of speed. ASICs, on the other hand, can achieve higher speeds and consume low power. Moreover, if ASICs are to be manufactured in bulk, the costs are not that high. But they cannot be reprogrammed on the field.

Tensor processing unit (TPU): Tensor processing unit is an AI accelerator built by Google specifically

for neural network machine learning. Initially, it was used by Google only, but since 2018, Google has made it available for third-party use. It supports TensorFlow code and allows you to run your own programs on the TPU on Google Cloud. The documentation provided by Google is comprehensive and includes several user guides. The TPU is specifically designed for vector-by-matrix multiplication, an operation that happens multiple times in any ML application.

On-edge computing and AI

On-edge computing gives an engine the power to compute or process locally with low latency. This leads to faster decision-making due to faster response time. However, the biggest advantage would be the ability to send just the processed data to the cloud. Hence, with edge computing, we will no longer be as dependent on the cloud as we are today. This means that lesser cloud storage space is required, and hence lower energy usage and lower costs.

The power consumed by AI and ML applications is no joke. So, deploying AI or machine learning on edge has disadvantages in terms of performance and energy that may outweigh the benefits. With the help of on-edge AI accelerators, developers can leverage the flexibility of edge computing, mitigate privacy concerns, and deploy their AI applications on edge.

Boards that use AI accelerator

As more and more companies make a mark in the field of AI accelerators, we will slowly witness a seamless integration of AI into IoT on edge. Companies like NVIDIA are, in fact, known for their GPU accelerators. Given below are a few examples of boards that feature AI accelerators or were created specifically for AI applications.

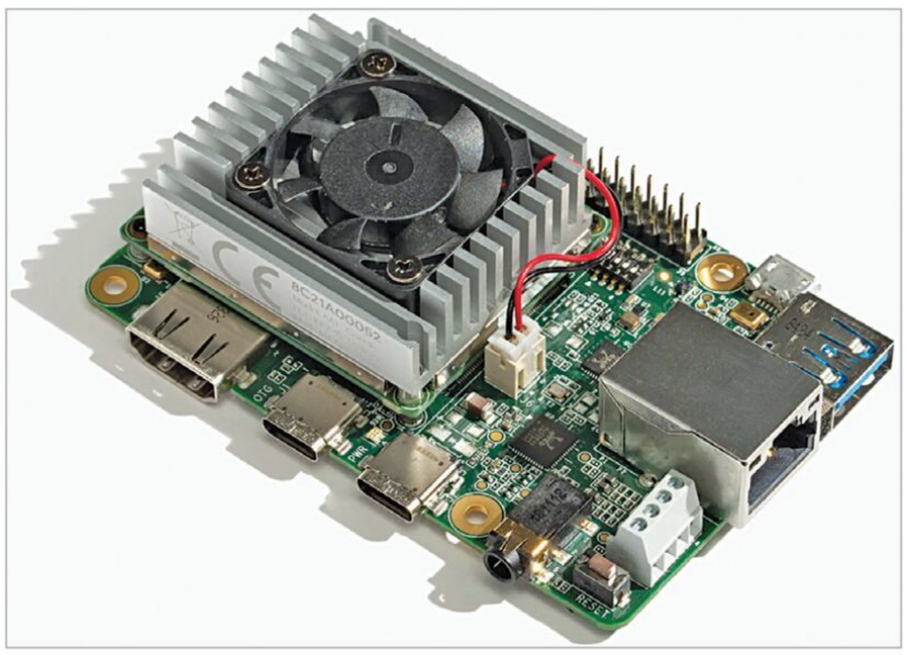


Figure 5: Coral Dev Board (Credit: Coral)

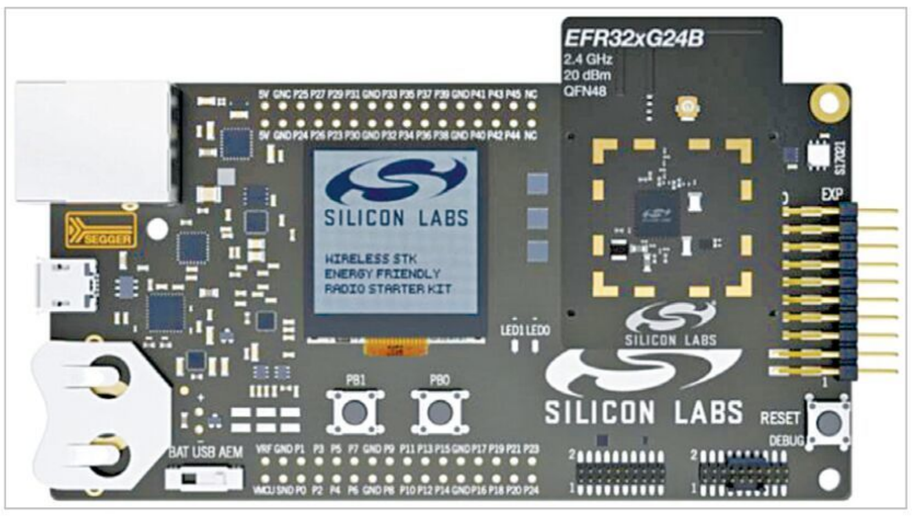


Figure 6: The complete Pro Kit for the new BG24 and MG24 SoCs with all the necessary hardware and software for developing high-volume, scalable 2.4GHz wireless IoT solutions (Credit: Silicon Labs)

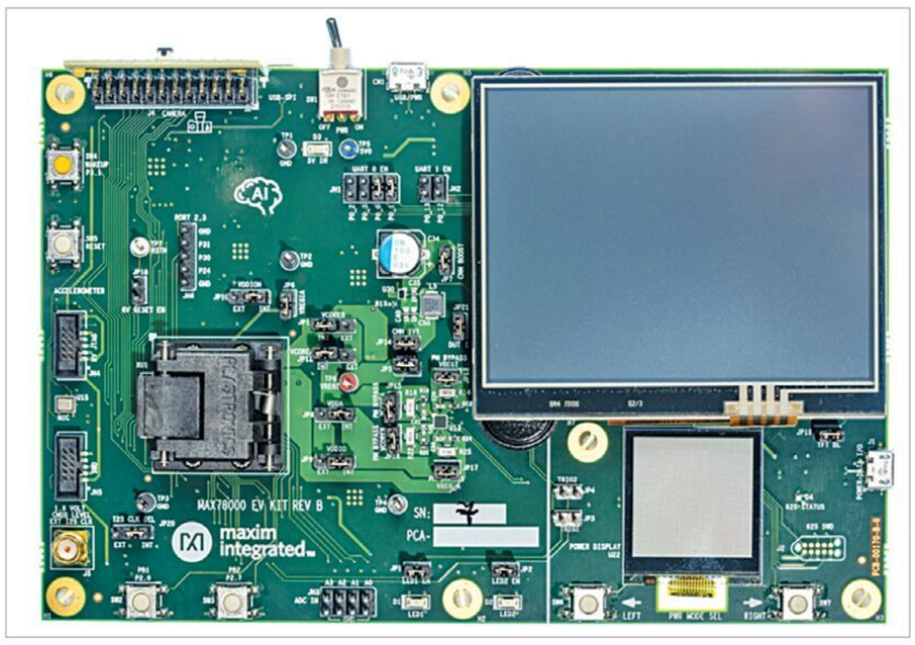


Figure 7: Evaluation Kit for the MAX78000 (Credit: Maxim Integrated)

Google's Coral Dev Board: Google's Coral Dev Board is a single-board computer (SBC) featuring an edge TPU. As mentioned before, TPUs are a type of AI accelerator developed by Google. The edge TPU in the Coral Dev Board is responsible for providing high-performance ML inferencing with a low power cost.

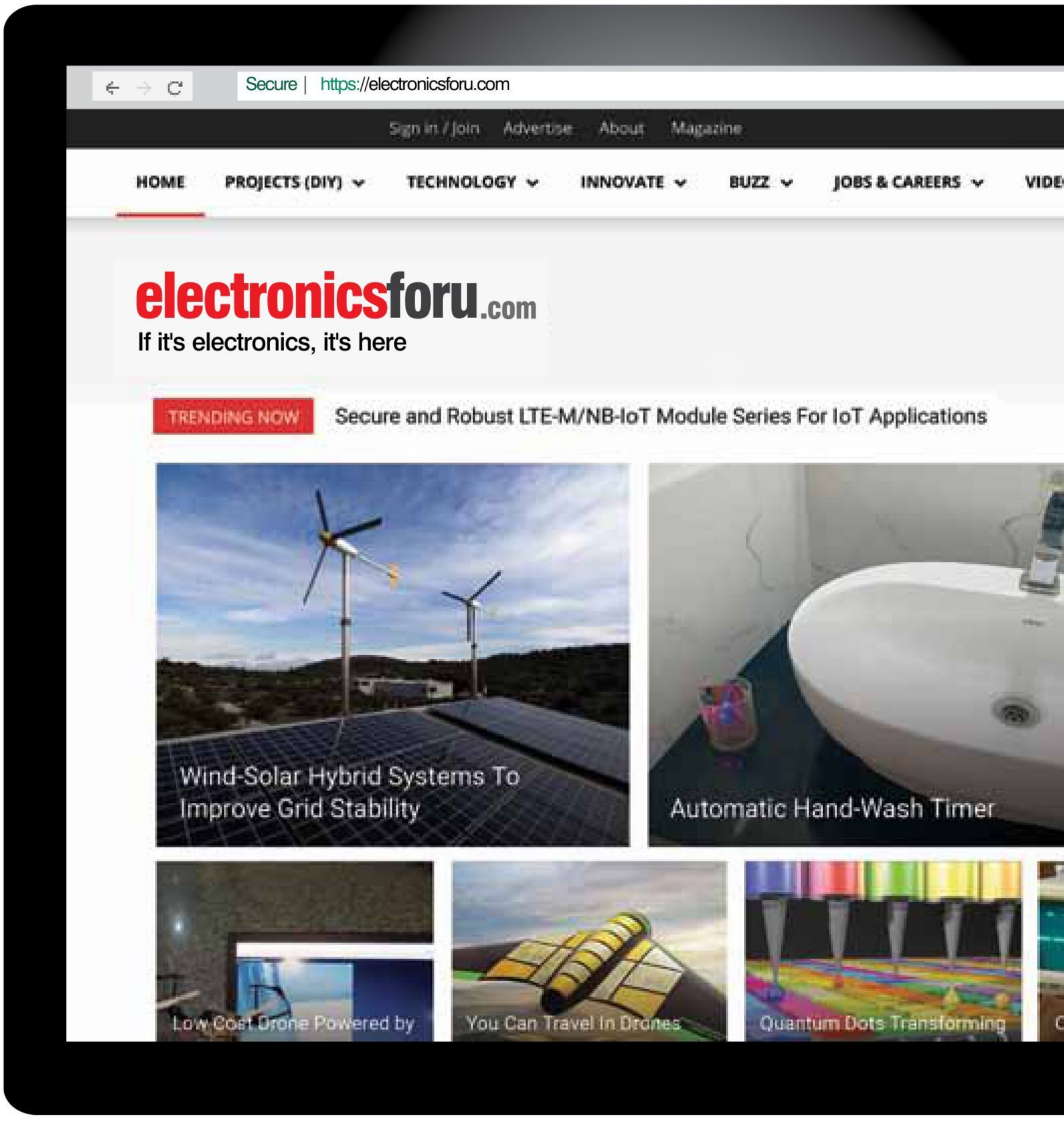
The Coral Dev Board supports TensorFlow Lite and AutoML Vision Edge. It is suitable for prototyping IoT applications that require ML on edge. After successful prototype development, you can even scale it to production level using the onboard Coral system-on-module (SoM) combined with your custom PCB. Coral Dev Board Mini is Coral Dev Board's successor with a smaller form factor and lower price.

BG24/MG24 SoCs from Silicon Labs: In January 2022, Silicon Labs announced the BG24 and MG24 families of 2.4GHz wireless SoCs with built-in AI accelerators and a new software toolkit. This new co-optimised hardware and software platform will help bring AI/ML applications and wireless high performance to battery-powered edge devices. These devices have a dedicated security core called Secure Vault, which makes them suitable for data-sensitive IoT applications.

The accelerator is designed to handle complex calculations quickly and efficiently. And because the ML calculations are happening on the local device rather than in the cloud, network latency is eliminated. This also means that the CPU need not do that kind of processing, which, in turn, saves power. Its software toolkit supports some of the most popular tool suites like TensorFlow.

"The BG24 and MG24 wireless SoCs represent an awesome combination of industry capabilities, including broad wireless multiprotocol support, battery life, machine learning, and security for

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

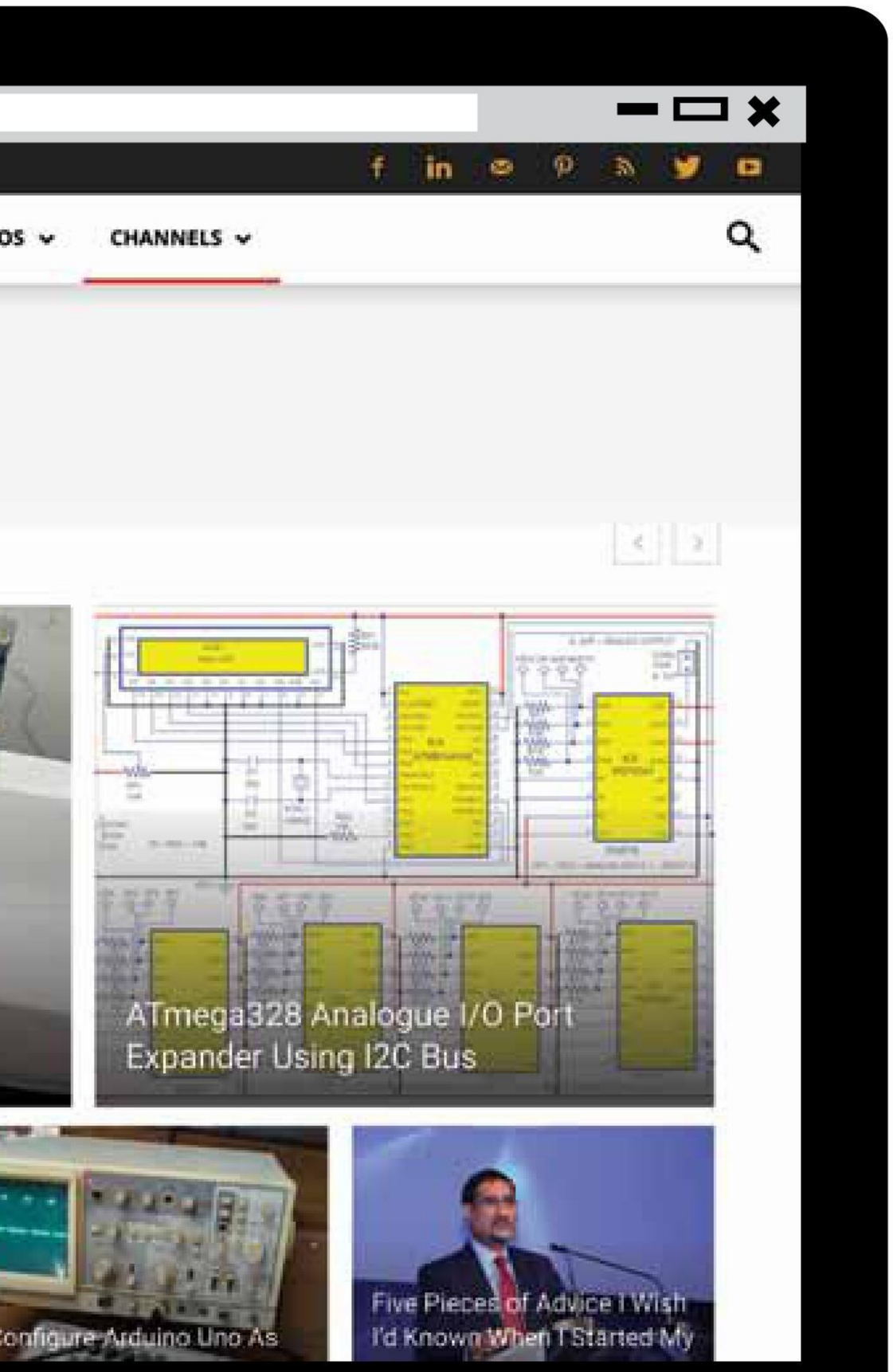
READERS

- 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...





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.

Amazing DIY Projects. Latest Tech trends.

The hang-out for electronics enthusiasts.

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

IoT edge applications," says Matt Johnson, CEO of Silicon Labs. These SoCs will be available for purchase in the second half of 2022.

MAX78000 Development Board by Maxim Integrated: The MAX78000 is an AI microcontroller built to enable neural networks to execute at ultra-low power. It has a hardware based convolutional neural network (CNN) accelerator, which enables battery-powered applications to execute AI inferences.

Its CNN engine has a weight storage memory of 442kB and can support 1-, 2-, 4-, and 8-bit weights. Being SRAM based, the CNN memory enables AI network updates to happen on the fly. The CNN architecture is very flexible, allowing networks to be trained in conventional toolsets like PyTorch and TensorFlow.

Kria KV260 Vision AI Starter Kit by Xilinx: This is a development platform for Xilinx's K26 system on module (SOM), which specifically targets vision AI applications in smart cities and smart factories. These SOMs are tailored to enable rapid deployment in edge based applications.

Because it is based on an FPGA, the programmable logic allows users to implement custom accelerators for vision and ML functions. "With Kria, our initial focus was Vision AI in smart cities and, to some extent, in medical applications. One of the things that we are focused on now and moving forward is expanding into robotics and other factory applications," says Chetan Khona, director of Industrial, Vision, Healthcare & Sciences markets at Xilinx.

Gluon AI co-processor by AlphaICs: The Gluon AI accelerator is optimised for vision applications and provides maximum throughput with minimum latency and low power. It comes with an SDK that ensures easy deployment of neural networks.



Figure 8: Kria KV260 Vision AI Starter Kit (Credit: Xilinx)

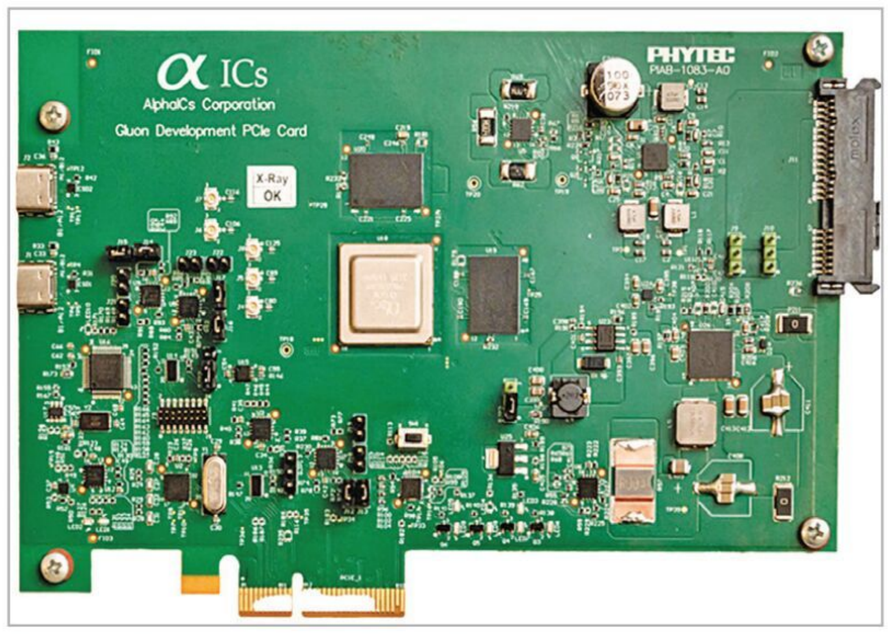


Figure 9: Gluon Evaluation Board (Credit: AlphaICs)



Figure 10: Intel NCS2 (Credit: Intel)

AlphaICs is currently sampling this accelerator for early customers. It is engineered for OEMs and solution providers targeting vision market segments, such as surveillance, industrial, retail, industrial IoT, and edge gateway manufacturers. The company also offers an evaluation board that can be used to prototype and develop AI hardware.

Intel's Neural Compute Stick 2 (Intel NCS2): Intel's NCS2 looks like a USB pen drive but actually brings AI and computer vision to edge devices in a very easy manner. It contains a dedicated hardware accelerator for deep neural network interfaces and is built on the Movidius Myriad X Vision processing unit (VPU).

Apart from using the NCS2 for their PC applications, designers can even use it with edge devices like Raspberry Pi 3. This will make prototyping very easy and will enhance applications like drones, smart cameras, and robots. Setting up and configuring its software environment is not complex, and detailed tutorials and instructions on various projects are provided by Intel.

A long way to go

The field of AI accelerators is still niche, and there is still a lot of room for innovation. "Many great ideas were implemented in the past five years, but even those are a fraction of the staggering number of AI accelerator designs and ideas in academic papers. There are still many ideas that can trickle-down from the academic world to industry practitioners," says Fuchs. END

By: Aaryaa Padhyegurjar
The author is an Industry 4.0 enthusiast with a keen interest in innovation and research.

The article was originally published in the April 2022 issue of Electronics For You.

Enter the Metaverse for Better or for Worse

If we told you that another universe is evolving in the digital space, where people can live just as in ours, would you believe it or write it off as fan fiction? It is time to take a serious look at the metaverse, which coexists—and even interoperates—with our real world, because it is poised to get bigger.



HoloLens enables collaboration among surgeons worldwide (Courtesy: Microsoft)

You cannot escape the term Metaverse. It stares back at you from the newspapers, television, social media, conferences, investor meets—practically every place except maybe your grandma's living room. (Then again, you might be visiting your grandma in the metaverse!)

Facebook changed the name of the parent company to Meta, while Microsoft is betting big on collaboration using Mesh. Boeing and Airbus want to build their next planes on the metaverse. Barbados is building an embassy in Decentraland. Sotheby's has a store there too, by the way, and Christie's recently sold one of the most expensive pieces of digital art in the metaverse for nearly US$70 million. Adidas opened a virtual showroom in The Sandbox, and Ariane Grande performed a grand concert in Fortnite.

People are buying land and building malls and mansions in the metaverse, while youngsters in Philippines are giving up their real-world jobs as they earn more money by winning battles on Axie Infinity. Closer home, people are depositing their cryptocurrency in fixed deposits on Zionverse, while two Potter-heads in Tamil Nadu hosted their Harry Potter themed wedding reception on TardiWorld. The metaverse is all over the place—but what really is it?

The metaverse, 29 years after its conceptualisation

A lot of the advanced technology we see in reality today was once the realm of fiction. Neal Stephenson coined the term metaverse in his 1992 book *Snow Crash*. He described it as a parallel digital world where humans interact with each other in *avatar* form. Almost three decades later, venture capitalist Matthew Ball describes the metaverse as a massively-scaled and interoperable network of real-time rendered three-dimensional (3D) virtual worlds, which can be experienced synchronously and persistently by an effectively unlimited number of users.

That answers all of you who are asking how the metaverse is different from the massively multi-player online (MMO) games, which are so popular today. The multiverse is not one virtual world—rather, it is a universe that brings together multiple virtual worlds. Over time, the virtual worlds will be

connected and people will be able to go from world to world, carrying with them their things and money. Basically, their activities will not be tied to one world, but be spread across multiple platforms in the metaverse. That is the broad picture—although it is going to take time to get there.

Not just that, the metaverse will also be a realm that blends real and digital worlds. For example, you will be mingling with real people, and making real money on the metaverse. It is not an entirely fictional world or a game. Satya Nadella, CEO of Microsoft, nailed it when he said the idea is to embed computing into the real world and to embed the real world into computing, bringing real presence to any digital space.

Dissecting the definitions

As we cut through the jargon, you will see that the metaverse is actually made up of several components—virtual worlds, virtual reality gear, digital currency, etc. This will also help you understand the spate of acquisitions going on in this space, as the tech majors like Meta and Microsoft try to bring together the different cogs that make the metaverse run.

"The metaverse is a concept of virtual collaboration where people join shared three-dimensional (3D) environments as *avatars*, where they can talk, teach, learn, work, play and can also transact using blockchain tokens. Overall, it is a combination of 3D collaboration platforms, blockchain, non-fungible tokens (NFT), crypto-currency and gamification," explains Pushpak Kypuram, founder-director, NextMeet, a Hyderabad based startup that has developed an immersive metaverse platform for meetups.

For the love of the game

Filipinos loved to play Axie Infinity. So, they earned Smooth Love Potions (SLP) and Axie Infinity Shards (AXS) by winning battles against zombies! They converted these crypto-tokens into ETH and subsequently to US$ and found that this was more than what some of them were earning in real-world jobs, especially during the Covid lockdowns. So, many of them gave up their real jobs to earn money playing their favourite game!

The metaverse is made up of virtual worlds

The metaverse, as we said, is made up of several virtual worlds, many of which are evolved gaming platforms. Second Life, for example, was launched in 2003. Although it seems like a game from the outside, its 'residents' take it quite seriously—it is their chance at a second life, where the first one has not met their expectations. It feels like being in a real world, where the residents explore, meet other residents, socialise, build, create, shop, and trade virtual property and services with one another.

Way back in 2008, a couple even married each other on Second Life. They later met each other in the real world, and, thankfully hit it off. But it can also be the other way round; with couples getting divorced due to affairs in the virtual world.

Not all virtual worlds are, like, complete worlds. Some of them focus on some activities alone. For example, in World of Warcraft, Fortnite, Call of Duty, or Axie Infinity, you battle with villains to save the world. In Farmville, you focus on developing an awesome and profitable farm. In The Sandbox, you play God's assistant, exploring and bringing together resources like water, soil, humans, wildlife, and machines to create stuff.

If you are tired of superhuman environments, then maybe you can go for the LokaLocal, a virtual world based on 3D maps of real-world cities and locations, where players can chill out in 3D versions of their favourite hangouts like Connaught Place in Delhi, Marine Drive in Mumbai, or MG Road in Bengaluru.

There are also virtual spaces for more serious work, like Microsoft Mesh and Meta's Horizon Workplaces for collaborative meetings and design sessions, or NextMeet for meetups, and VRXtream for organising events.

Opportunities for all

The metaverse means opportunities galore, for contributors from various fields. If there is another universe evolving, it means more demand for goods and services, so more jobs are created too.
Here is a small sample:

Chipmakers: Huge amounts of data and heavy processing obviously means huge demand for semiconductors. Likewise, the heavy use of ultrasonic and infrared sensors, force and strain sensing, and motion sensing in VR headsets creates a demand for micro-electro-mechanical systems (MEMS).

Developers: A few months ago, Mark Zuckerberg said Meta would create 10,000 jobs in Europe over five years to make a metaverse. Developers can use the software development kits (SDKs) available for various platforms to build immersive applications for the metaverse.

Designers: When we talk about creating virtual spaces, it obviously requires designers to do the job. In fact, people buying plots on virtual worlds and building houses in them are sometimes engaging real-life architects to do the job. And there is a lot of scope for creativity here, because what you build can defy all the laws of physics!

Creators: Creators of anything digital can sell their wares in the virtual worlds and make real money out of it. The biggest example is Beeple, a digital artist who had erstwhile not sold a single print for over US$10, being able to sell an NFT for nearly US$70 million!

Digital justice

As the lines between real and virtual worlds blur, our avatars need to be as alert as us, because the dangers of the real world are bound to occur in the virtual one too. We need to be aware of cyber frauds and privacy infringement. Cases of groping on MR platforms have also been reported. Soon after Meta launched Horizon Worlds, a beta tester reported on their website that she had been groped by a stranger on Horizon Worlds. Meta reviewed the incident internally, and suggested that this could have been averted if the tester had used a tool called Safe Zone, which is a protective bubble users can activate when feeling threatened. Within it, no one can touch them, talk to them, or interact in any way until they signal that they would like the Safe Zone lifted. While some platforms have such features in place, many do not. And even where such features are available, most users do not know about it. So, more awareness has to be created.

Tech futurist and author Dr Ayesha Khanna suggested in one of her interviews that all virtual worlds must have some systems in place for digital justice and good governance, to prevent digital harm and cyber-crime, and to punish such occurrences if they occur despite best efforts.

Meta hopes to also bring you closer to your family and friends with Horizon Worlds and Horizon Homes, where, rather than fighting to save the world, you will be building escape rooms, social hangout spaces and party games to foster relationships.

So, now we know that there are multiple virtual worlds in this metaverse, each with different themes and purposes. If you look at it very objectively, these worlds are nothing but interoperable software platforms that you log in to for work or play. What makes them worthy of being called 'worlds' is their immersive nature, and ability to host thousands of people in a single instance of the server.

These worlds are immersive, and you enter them as *avatars*

Take Zoom or Google Meet—these are also ways of interacting with people. But everyone, including you, is just a thumbnail in it. You can see and talk to each other, but you are not really in the same space. The metaverse, on the other hand, is immersive and experiential. You enter this 3D digital space and become part of it, along with others. So, if you are playing with friends in Horizon Worlds, your *avatars* will be together in the same park or beach. If you are attending a math session, you will be sitting in a virtual classroom next to your friend.

So, to experience the metaverse, you first need an *avatar,* and then you need the tech gear to immerse yourself in the virtual world. Your *avatar* is, simply put, a 3D character that represents you in the digital world. You can create your own *avatar,* or customise available ones, most of which look like animated characters. However, the more serious and real-world applications emerging today will feel better if your *avatar* looks something like you. So, Epic Games owned company Unreal Engine has developed a tool called MetaHuman Creator, which would help you develop digital humans that look and behave pretty much like you.

Once you have an *avatar,* you enter the virtual world. "Metaverse can be experienced using personal computers (PCs), laptops, mobiles, etc. Virtual reality (VR) and augmented reality (AR) are hardware technologies used to experience the metaverse. With VR, you are in a 360-degree immersive environment, whereas with AR you augment the 3D elements with the real world," explains Kypuram. Going a step forward, we now have mixed reality (MR)—a form of AR where the real world and digital objects are not just visually interleaved but also interact with each other.

As the Rs (VR, AR, MR) keep increasing, we see them being brought under one umbrella term called extended reality (XR). According to VRdirect, the term simply represents any sort of technology that alters reality for the user by blurring the line between real and digital worlds. So, it combines VR, AR, and MR, as well as any new technologies that will come up in the future.

VR/MR headsets need no introduction to gaming enthusiasts—Oculus from Meta, HoloLens from Microsoft, and Playstation VR headsets from Sony, to name a few. VR headsets, most commonly used by serious gamers, will totally immerse you in a simulated environment using 3D vision, spatial sound, and motion trackers that sense movement of your eyeballs, and sometimes even jaw movements. The headsets usually come with handheld gaming controllers too. Some advanced sets might also include cameras to detect full body movement, and haptic feedback vests to give you a sense of touch—so you can feel it when your opponent punches you!

MR headsets go a step further to combine real-world and digital elements, usually using holographic display tech. Not an easy task, and right now they cost a whopping lot, but MR headsets might be the actual solution to balancing real and digital lives. Betting big on the metaverse, most of the industry majors are working on developing advanced MR headsets. Last year, Meta revealed two such projects—Project Nazare and Project Cambria. Apple is also developing an MR headset. Microsoft HoloLens 2 is a preferred MR headset for professional and industrial uses—but its lowest-end model costs more than US$3,000.

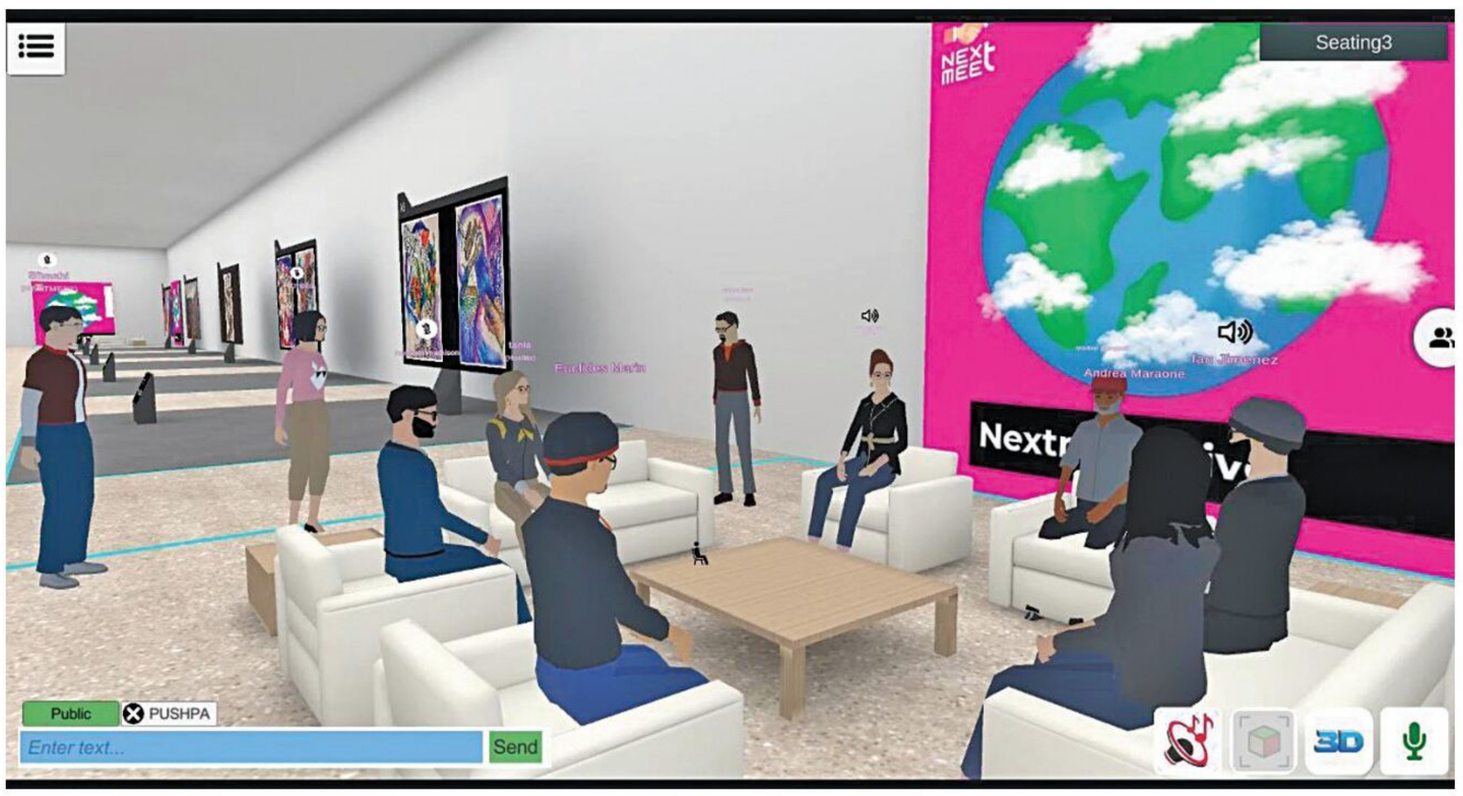
The problem with most headsets today is that they are not comfortable to wear for long durations, and do look a little odd when worn in public. Or perhaps it is a saving grace, as it naturally pulls out someone who tends to spend too much time in the virtual world.

You can manage without these Rs too

Leave alone an expensive MR set, the average Internet user does not even have a VR headset. So, if you are conducting a university course on the metaverse, or a company town hall, how can all the employees participate?

"The metaverse can be experienced via PCs and mobiles, but a VR or MR headset would enable users to experience real immersion. Right now, a decent VR headset would cost around US$300 and an MR headset would be around US$2,000. VR and MR technology is in motion, and it has gained good traction in some niche markets like industrial training and simulation. Once more affordable and comfortable VR/MR headsets are in the market, it will become a consumer product. Right now, people who want to attend an event in the metaverse—say, a townhall conducted by their company—can come either on PC, mobile, or VR depending on the convenience and availability," says Kypuram of NextMeet.

Vignesh Selvaraj, CEO, TardiVerse, shares similar sentiments. TardiVerse developed a Harry Potter themed space in their blockchain powered TardiWorld for Dinesh and Janaganandhini to celebrate their wedding reception with friends and family, without any restrictions on the number of guests! Given the cost of headsets, it would not have been possible for all the attendees to login with one. So, they decided to keep it simple. While they had special equipment for the wedding couple to be present at Hogwarts' Great Hall, others just had to click on a web link, create a login for themselves, choose and customise an *avatar,* and enter. They could participate in events and games using simple mouse or touch screen controls. This is what is practically possible currently, but over time as headsets become more prevalent, people can immerse themselves fully—just like video gaming enthusiasts do already!



Art event at Broward College, Florida, USA, powered by NextMeet (Courtesy: NextMeet)

The metaverse supports economic activity

Virtual reality platforms are not new—we have been watching this space for a long time. The real twist in the metaverse story probably came when these make-believe worlds suddenly became hotbeds of commercial activity, and people started minting money. This is a subject in its own right, but for the sake of completeness we will attempt to give you a short overview of commerce in the metaverse.

When you think of trade, the two basic elements that come to mind are assets and currency. In a virtual world, these are powered by blockchain technology.

Assets are represented by non-fungible tokens or NFTs. A non-fungible token is a non-interchangeable unit of data stored on a blockchain. Think of it as a unique digital token that represents a digital asset. So, buying an NFT gives you ownership of the associated digital asset. Once an NFT has been encoded using blockchain technology, no one can alter its ownership data or metadata. Owners can prove its validity using blockchain verification. NFTs can be moved across different ecosystems owing to the open nature of blockchain standards. So, you can trade NFTs outside the world in which you bought them! Ethereum is the most popular blockchain platform used for encoding NFTs, but one can also use others like Tezos, Algorand, or Polkadot.

Just as NFTs represent assets, crypto-currencies serve the equivalent of money. These are fungible tokens, that is, they can be interchanged. They are identical to each other and traded, hence an ideal medium for commercial transactions. Ether (ETH) is the native crypto-currency of the Ethereum platform. Every virtual world can have its own native currency, which can be converted into other crypto-currencies, or even real-world money based on prevailing conversion rates. For example, MANA is the native crypto-currency in Decentraland. Approximately 1123 MANA equals 1 ETH, and 1 ETH equals US$3,137.5 (you heard it right!), so 1 MANA is roughly US$2.8.

With the basic building blocks—assets and currencies—in place, marketplaces began to emerge in the virtual world. Some of the well-known ones are Decentraland to buy, develop, and sell virtual plots of land; OpenSea to trade in rare digital items and crypto-collectibles; SoRare for digital player



NFTs priced at US$ 10 were sold on BeyondLife.club to celebrate Dinesh and Janaganandhini's metaverse wedding reception (Courtesy: BeyondLife.club)

EFY GROUP
YOURS SINCE 1969

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 ☐	2150 ☐	3000 ☐
		SAARC COUNTRIES (IN US$ BY AIR)	
	50 ☐	135 ☐	NA
		OTHER COUNTRIES (IN US$ BY AIR)	
	100 ☐	270 ☐	NA

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

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**

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.



Daler Mehndi hosted a Republic Day concert on the PartyNite metaverse from Gamitronics, becoming the first Indian to perform on the metaverse (Courtesy: Daler Mehndi on Instagram)

cards and virtual teams; SuperRare and MakersPlace for authentic digital art; and so on. Almost every virtual world, in fact, has a marketplace that sells related ware.

Last year, Sotheby's bought a plot of land in Decentraland and developed a metaverse store, which is an exact replica of its London store. In March 2021, Mike Winkelmann, a digital artist popularly known as Beeple, sold the NFT of one of his digital works for US$69 million at Christie's. In December 2021, Adidas released its first range of NFTs (that sell for approximately 0.2ETH) at its showroom in The Sandbox. Nike has filed several trademark applications in the past couple of months for upcoming virtual rollouts. Last year, in February, the US based National Basketball Association's online platform for NFT Basketball cards registered a sale of about US$250 million.

Gucci, Christian Dior, Ralph Lauren, all the fashion biggies have stores in 3D virtual worlds! And on the other end of the spectrum, Pizza Hut, Pringles, and Taco Bell are also selling themed NFTs.

One of the reasons for the popularity of these meta marketplaces is the ease of trading. You just set up a digital wallet, buy crypto-currency, connect your wallet to an NFT marketplace and start selling or bidding for an NFT—be it a plot of land, a collectible football card, or a rare piece of digital art.

Putting it to real use

A technology can garner attention but not sustain itself unless it proves to be really useful. Entertainment is a given, but will the meta worlds also improve the quality of our lives in the real world?

"People have been playing 3D games since the last three decades. Now, since remote working and learning is the norm, gamification will become necessary in these spaces too. Just like businesses have been developing websites on the Internet for nearly three decades, now every brand will start having 3D experiential stores. Like mobile devices replacing desktop computers, eventually holographic projections will replace TV. Holographic eyewear are being manufactured by giants like Microsoft and Apple. It is not like we will be living inside 3D, but 3D will be a superimposed element in everyone's life. Just like we have moved from button phones to touch phones, in the future we will operate gadgets using gestures as the user interfaces and user experiences shift to the 3D paradigm," says Kypuram.

Experts say they are unable to visualise even the tip of the iceberg fully yet, because the possibilities are endless. Jog your mind, and you can list at least a dozen within a minute! Here is a small sample of what is possible.

Learning: The pandemic has taught us the importance of remote working. With the metaverse, learning could be so much more interactive and fun at school and university levels. You can get the feel of being in a real classroom with your friends, and it would be easier for the teacher too to explain stuff with holograms and 3D working models.

Work: Effective collaboration is expected to be one of the biggest benefits of the metaverse—and one that companies like Meta and Microsoft are betting big on. Colleagues can interact and brainstorm as if they were in the same room. Plus, the standards of collaboration—especially in design—would go up significantly with MR platforms.

In December 2021, a Reuter report spoke of Boeing's plans to build its next plane in the metaverse. Its future factory will bring together engineers around the world using Microsoft HoloLens headsets, robots that speak to each other, and immersive 3D designs using holograms. I know that sounds like Tony Stark building something in the Marvel universe, but well it is bound to happen.

Events and workshops: Perhaps what people miss the most in remote working are the occasional events and workshops that liven things up and create opportunities to interact across the spectrum. Now, close-to-real events are possible in the virtual world.

NextMeet, for instance, has worked with Indian and international universities to organise virtual events and conferences. Their immersive 3D environment comes replete with *avatars* that can move around the space, 3D spatial voice, expo halls with banner hosting stands, presentation screens in the conference hall, help desks and networking lounges! VRXtream is another metaverse for events, where you can even buy spaces (stalls) and rent them out, or vice versa.

Healthcare: From training medical students to remote consultation and counselling, the metaverse is expected to benefit the medical world notably—and we can already see fragments of it in use. In a blog post that Microsoft put up last year showcasing its immersive

MR platform, Mesh, the company imagined *avatars* of medical students learning about human anatomy, gathered around a holographic model and peeling back muscles to see what's underneath. Hospitals are using MR platforms to help paramedics give first aid to patients, the World Health Organisation (WHO) used AR and smart phones to train Covid-19 response teams in remote areas, and psychiatrists are using VR to treat post-traumatic stress among combat soldiers.

Fitness and training: Maybe, exercising in the metaverse would be more fun! Zippy is a virtual world for runners, where people can run across major marathon cities like Boston, London, Mumbai, and Tokyo, or through scenic environments like jungle trails and beach runs. The virtual environment syncs with the real-world kinematics of the runners using their fitness wearables or treadmill sensors.

If you love fiction, then OliveX could have just the solution to add a spot of fun to your run! In association with award-winning novelist Naomi Alderman, the company has developed an ultra-immersive running game called *Zombies, Run!*, where every run is a mission to save the world. You run faster when you are chased by the zombies, and slow down to pick up supplies. The accompanying narration and music set the tempo for your run. What a fun way to exercise!

Or, if it is serious athletic practice that you are thinking about, Ghost Pacer might be of help. A solution developed by runners, for runners, these AR glasses project a live holographic opponent into your field of view, for you to compete with and train against.

Retail: Everyone loves to shop. But the pandemic made us realise that physical store shopping is not something we can take for granted. So, for those times when you want to shop but cannot go out, or if you are too busy to shop hop, then you can turn to the showrooms that are slowly mushrooming in the metaverse. From real estate to automobiles, sellers are building unique experiences to woo shoppers. Real estate agents could give you virtual tours of their properties, you could walk around a virtual clothing store and experience the latest in fashion, or test drive a car virtually!

Social media. This is one space Meta is betting big on. Imagine what happens when your social media apps become immersive? Your friends from different parts of the world could manifest in your living room as holograms! Quite some time for that level of interaction to become mainstream, but, as of now, we commonly see friends hanging out together in shared virtual spaces.

Honestly, there is too much to dream about in the metaverse. *Avatars* might be living in huge mansions in plush downtown neighbourhoods, driving luxury cars, going on cruises and watching their favourite movies on wall-sized screens. But, in real life, the same people might be living in hostels, having no television, not even a bicycle, and maybe no real friends to call their own.

When one snaps out of the dream, it might be very difficult to come to terms with reality. The huge disparity between the real and virtual lives could even lead to psychological disorders. So, we need to know where to draw the line, as in the case of any technological advancement. END

Block and tackle

"We're moving from a mobile and cloud era to an era of ubiquitous computing and ambient intelligence, an era that will experience more digitisation over the next 10 years than the last 40. We need that next level of real-time hyper-connectivity between businesses and between consumers and businesses, where data and intelligence flow freely to tackle the challenges of supply and demand." ***— Satya Nadella, CEO, Microsoft***

"I don't know if I necessarily buy into this metaverse stuff, although people talk to me a lot about it. You know, when I grew up it was like, 'Don't sit too close to the TV, it's gonna ruin your eyesight,' and now TV is like literally right here (puts his hand very close to his face), I'm like, umm, what? Is that good for you? I don't see someone strapping a friggin' screen to their face all day." ***— Elon Musk, CEO of SpaceX and Tesla***

"The next platform will be even more immersive—an embodied Internet where you're in the experience, not just looking at it. We call this the metaverse, and it will touch every product we build. The defining quality of the metaverse will be a feeling of presence—like you are right there with another person or in another place. Feeling truly present with another person is the ultimate dream of social technology. That is why we are focused on building this." ***— Mark Zuckerberg, founder, chairman and CEO of Meta***

"With digital risks already being high, especially for vulnerable groups, safety risks could be more prevalent in the metaverse. There are numerous ways such risks could be exacerbated in the metaverse. Unwanted contact could become more intrusive. The rise of virtual currencies is another challenge. Finding ways to incentivise better behaviours and reward positive interactions may need to become a bigger part of a safer digital future." ***— Cathy Li and Farah Lalani, World Economic Forum***

By: Janani Gopalakrishnan 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 March 2022 issue of Electronics For You.

An EFY GROUP EVENT

5243 ATTENDEES

70+ SESSIONS HOSTED

100+ SPEAKERS

18 COUNTRIES REPRESENTED

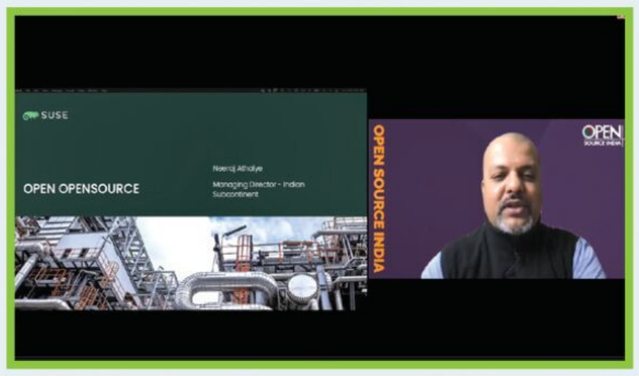


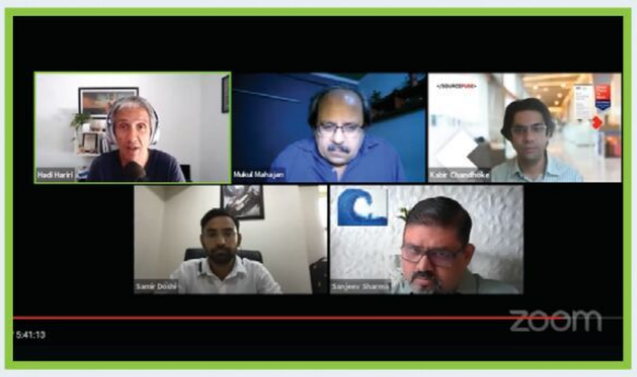
















18th Edition

OPEN SOURCE INDIA

VIRTUAL EVENT

Thank You

VISITORS, for making the event a huge success

SPEAKERS, for contributing your valuable thoughts

PARTNERS, for your support

Microsoft Azure | soda foundation | Google | aws

MySQL | HUAWEI | iTEXT | GitHub

SUSE | ENTREPRENEUR FIRST JOINEF.COM | TRUEFOUNDRY BY ENSEMBLE LABS | Application Library Engineering Group

openLooKeng | EDGE GALLERY | soda foundation meetupindia | BANGALORE CNCF MEETUP

www.OpenSourceIndia.in

For more details, call on +91-9667743666 or email us at info@opensourceindia.in

FOCUS

Latest in DevOps

Are Low-Code Platforms Helpful for Professional Developers?



Over the years, low-code platforms have matured immensely, and are being used in varied domains of software development for better productivity and quality. This article explores the possibility of leveraging low-code platforms for complex product development by professional developers. *Read more on page.....39*

DevSecOps: A Philosophy that Puts Security First



DevOps has helped remove barriers between development and operations, to improve productivity as well as the reliability of software. However, it has fallen short on some aspects such as security, leading to the emergence of the DevSecOps philosophy. Find out more about it in this article. *Read more on page.....47*

Why You Should Deploy Immutable Infrastructure on the Cloud

Even as DevOps and cloud adoption have increased significantly in recent years, the use of immutable infrastructure has also become pretty popular. This article outlines the core benefits of deploying immutable infrastructure on the cloud.

Read more on page.....44

Demystifying Observability for Modern Applications

To come up with an effective solution to a problem, you need to have the relevant insights into it. Distributed systems are prone to unexpected failures and malfunctions, and observability helps us find out the reasons for these in order to arrive at an effective solution.

Read more on page.....50

Planning the DevOps Journey for an Enterprise

Large enterprises are looking to implement efficient site reliability engineering and a DevOps culture that is adaptable, agile and intelligent. But they need to strategise for that.

Read more on page.....53

Are Low-Code Platforms Helpful for Professional Developers?

Over the years, low-code platforms have matured immensely, and are being used in varied domains of software development for better productivity and quality. This article explores the possibility of leveraging low-code platforms for complex product development by professional developers.



In the last several years, companies have invested a lot of time and energy in innovating and improving the overall process of software product development. Agile working methods have helped in making the process of product development a lot smoother. However, developers still face the challenge of meeting ever expanding customer requirements quickly and easily.

In order to meet these requirements in totality, developers need tools and platforms that can enable quick delivery of software by reducing the coding timelines and without compromising on the quality aspects.

No-code platforms are one set of tools that enable creation of application or product software with zero coding. These use a 'lego building blocks' approach that eliminates the need for hand coding and focuses on just configuration of functions based on graphical modelling experiences. These tools are more relevant to a class of business users called citizen developers, who can use them to optimise a specific process or a function by developing their own applications.

Contrary to no-code platforms, low-code platforms do not try to eliminate the need for coding. Instead, they aim to make development of software easier than the traditional method of hard coding each line of a program or software. This approach minimises hard coding with prepackaged templates, graphic design techniques, and drag-and-drop tools to make software.

The focus of this article is on general-purpose low-code platforms

that can be used by professional developers to build enterprise grade applications or products.

Different types of low-code platforms

Low-code platforms cater to different use cases. Depending on the intended usage or purpose, these can be classified as follows.

- **General-purpose:** These platforms can create virtually any type of product or application. With a general-purpose platform, users can build software that serves a wide variety of needs and can be deployed on cloud or on-premise.
- **Process:** These platforms focus specifically on software that runs business processes such as forms, workflows or integrations with other systems.
- **Request handling:** Request handling low-code platforms are similar to process-based low code, but are less capable. They can only handle processing requests for fixed processes.
- **Database:** Database low-code platforms are useful if users have large amounts of data that needs to feed into a system, without spending a lot of time on the task.
- **Mobile application development platform (MADP):** These platforms help developers code, test, and launch mobile applications for smartphones and tablets.

Key features of generic low-code platforms

General-purpose enterprise low-code development platforms support rapid software development, deployment, execution and management using declarative, high-level programming abstractions such as model-driven and metadata-based programming languages, and one-click deployments.

The key features supported by general-purpose low code platforms are as follows.

- *Visual modelling:* These platforms have a comprehensive visual modelling capability, including business processes, integration workflows, UIs, business logic, data models, Web services, and APIs.
- *Databases:* Support for a visual editor, Excel sheet import or use of existing data models from a different database.
- *Pre-built templates:* Support a variety of pre-built templates that can serve as a starting point to get an application up and running quickly. Using a well designed and tested template not only increases productivity but also helps in building a more reliable and secure application.
- *Integration:* Provide easy integration with external enterprise systems, databases, and custom apps.
- *Security and scalability:* The right low-code platform makes it easy to create enterprise grade software that is secure and scalable.
- *Metrics:* Support for gathering metrics and monitoring the software.
- *Life cycle management:* Support for version management and ticket management, as well as Agile or scrum tools.
- *Reusability:* The code generated is reusable and can integrate easily with general-purpose IDEs, with multi-platform support for testing and staging.
- *Deployment options:* Ability to deploy on public or private clouds with support for container images. Also support preview functionality before publishing to the cloud.
- *Licensing:* Flexible licensing models with no vendor lock-in.
- *Others:* The platforms include other services, such as project management and analytics.

Generic low-code platforms for professional developers

The main objective of a generic low-code platform is to help reduce the time spent by a developer working on a product as compared to traditional hand coding. Using visual interfaces, drag-and-drop modules, and more, these low-code platforms reduce the manual effort of coding largely, but will need some coding requirements for completely building a product.

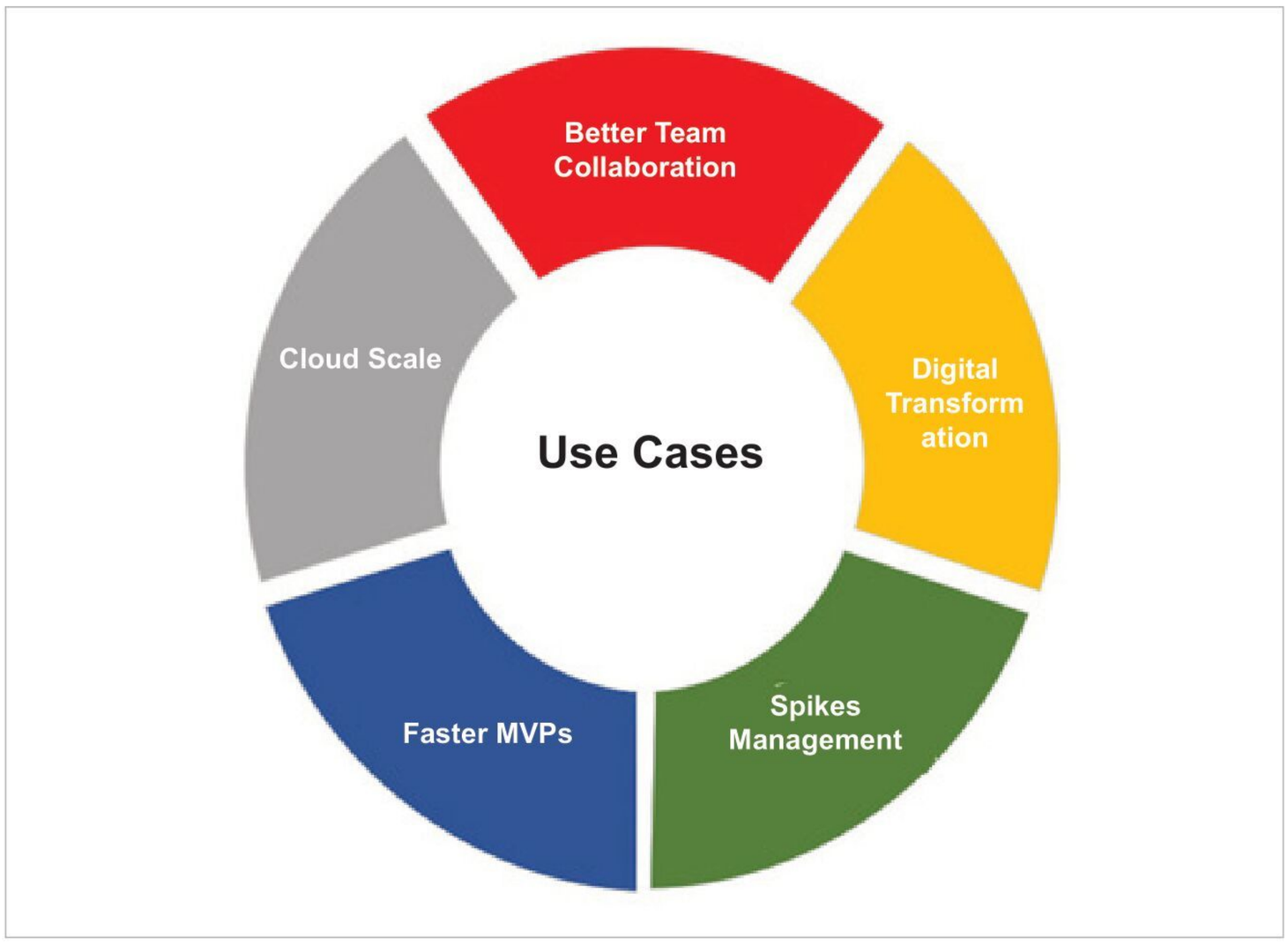


Figure 1: Important use cases of low code

Generic low-code platforms cannot be directly used to build low-level products like in-memory grids, Big Data processing algorithms, image recognition, etc. However, over the years they have evolved a lot to cover the widest range of capabilities for enterprise-grade development and full life cycle management, including business process management (BPM), integration workflows, UIs, business logic, data models, Web services, and APIs. They enable high-productivity development and a faster time to market for all types of developers. They also help create applications of any complexity including enterprise applications with generic architectures and complex backends, using microservices and service bus.

Figure 1 illustrates some of the key use cases that have been influenced positively in terms of productivity, scale and quality by leveraging low-code platforms.

Enable cross-functional team collaboration: Any product development needs a combination of business or functional experts as well as professional developers. Low-code platforms provide features that are relevant to a business user as well as professional developer. Cross-functional teams can leverage these platforms to turn great business ideas into readily deployable products much faster.

Rapid digital transformation of existing product suites: By leveraging client and server-side APIs of the platform, developers will be able to build, package and distribute new functionalities, such as connectors to external services like machine learning and AI. Low-code platforms enable developers to push beyond the boundaries of the core platform to build better solutions faster by extending the native features of the platform with some code.

Help meet sudden spikes in demand: Automated code generation combined with the one-click

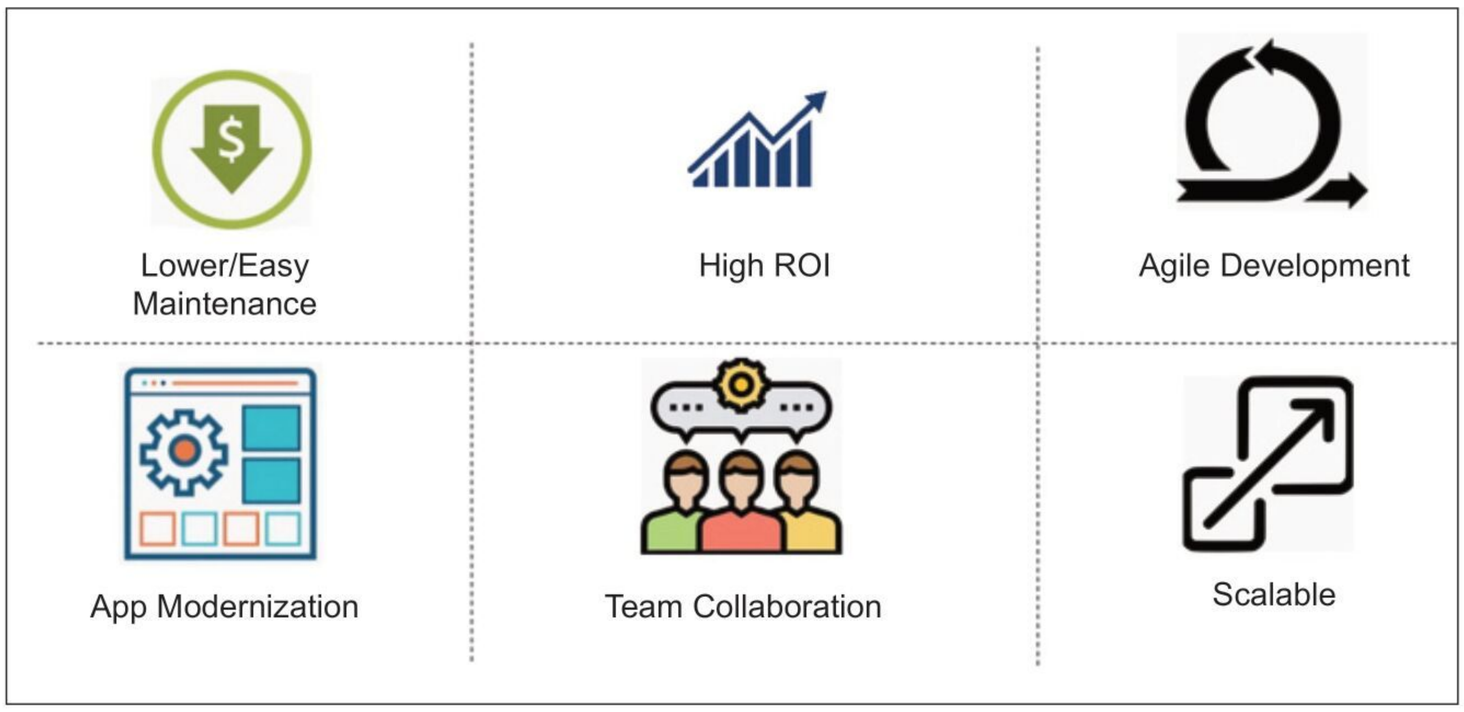


Figure 2: Benefits of low-code platforms

deployment option of low-code platforms helps in quicker product customisations, as well as building product variants and burning backlog of features faster.

Help build MVPs at a rapid rate: The end-to-end development and operational tools in low-code platforms provide a cohesive ecosystem that allows an enterprise to rapidly bring products to life and manage the entire SDLC process. They are used to build quick MVPs or PoCs for technology, frameworks, and architecture or for feature evaluation.

Support cloud scale architecture and design: Low-code platforms provide flexible microservices

Table 1

Evaluation criteria	Description
Functional features	Productivity, UI flexibility, ease of use
Cloud and containerisation	Capability to utilise popular cloud providers services like serverless, AI/ML, blockchain, etc
CI/CD integration	Out-of-the-box support for automation and CI/CD toolchain
Integration capabilities	REST and cloud app support, ability to connect to different SQL and NoSQL databases
Performance	Parallel and batch execution with support for elastic scalability
Security	Enable security by design: security tools, development methods, and governance tooling
Language support	Support for popular languages like Java, .NET, Python, etc
Development methodologies	Support standard Agile development methodologies like Scrum, XP, Kanban, etc
Extensibility	Ability to extend features of existing applications
Others	Platform support, learning curve, documentation, etc

Table 2

Low-code platform	Description	Type
Mendix	This is designed to accelerate enterprise app delivery across the entire application development life cycle, from ideation to development, deployment, and maintenance on cloud or on-premise.	Proprietary
OutSystems	This platform provides tools to develop, deploy and manage omni-channel enterprise applications. It addresses the full spectrum of enterprise use cases for mobile, Web and core systems.	Proprietary
Appian	An enterprise grade platform that combines the key capabilities needed to get work done faster using AI, RPA, decision rules, and workflow on a single low-code platform.	Proprietary
Budibase	Helps in building business applications. Supports multiple external data sources, and comes with prebuilt layouts, user auth, and a data provider component. Supports JavaScript for integrations.	Open source
WordPress	Powers more than 41 per cent of the Web -- from simple blogs to enterprise websites. With over 54,000 plugins, the level of customisation without writing code is incredible.	Open source
Node-RED	Helps to build event-driven IoT applications. A programming tool for wiring together hardware devices, APIs, and online services.	Open source

integration options (custom, data, UI, infra) to support building next-gen low-code products with multi-cloud deployment options. They are able to scale and handle thousands of users and millions of data sets.

Pros and cons of low-code platforms

The primary advantage of low-code software development is speed (months to weeks). On an average, there is six to ten times productivity improvement using a low-code platform over traditional hand coding approaches to software development.

Figure 2 lists some of the key benefits of low-code platforms.

- *Better software, lower maintenance:* Low-code platforms standardise the development approach and reduce the complexity as well as the error rate of the source code.
- *Cross-team collaboration:* Team members with different skills and capabilities can collaborate to realise the final product faster.
- *Enable Agile software development:* They offer the team members a consistent product that begins as a single screen, and grows from sprint to sprint as the full product takes shape.
- *Faster legacy app modernisation:* Enable faster UI generation based on the existing data model, reuse of the logic from legacy databases and smart transfer of existing screens/ persistence layers.
- *Low risk and high RoI:* Due to shorter development cycles, the risk of undertaking a new project is low, and there is a good chance of getting high returns.
- *Scaling through multiple components:* Allow use of a common platform to develop multiple services.
- *Easy maintenance:* The software can be easily updated, fixed and changed according to customer requirements.

In the last few years, low-code platforms have evolved a lot in terms of functionality and applicability. However, they still have some limitations when being used fully for any generic product development, some of which are listed below.

- *Low-level product development:* Cannot be used for building products like in-memory grids, Big Data processing algorithms, image recognition, etc.
- *Custom architectures and services:* Have limited use in enterprise software with unique architecture, microservices, custom back-ends, unique enterprise service bus, etc.
- *Source code control:* These platforms totally lose control over the code base; it's difficult to debug and handle edge conditions where the tool does not do the right thing automatically.
- *Limited integration:* Custom integration with legacy systems needs some significant coding.
- *Security and reliability:* These platforms are vulnerable to security breaches, because if the low-code platform gets hacked, it can immediately make the product built on it also vulnerable.
- *Customisation:* Custom CSS, custom integration, advanced client-side functionality, etc, will require a good amount of coding.

Popular generic low-code platforms

While there are many low-code platforms available in the market, one should leverage the evaluation criteria given in Table 1 before choosing the right one relevant to the business.

Table 2 lists the highly popular generic low-code platforms, both proprietary and open source.

Gartner predicts that by 2024, 65 per cent of application development projects will rely on low-code development.

Will low-code development replace traditional engineering completely? This is unlikely in the near future, but it will definitely help to significantly improve developers' productivity and bring quality products to market faster. It can bridge the talent gap in the mid-term by arming non-technical people with tools to build, whilst alleviating some of the product backlog for maxed out professional engineering teams

Low-code platforms can address critical issues like the high demand for new enterprise software, the need to modernise aging legacy systems and the shortage of full-stack engineers. The choice of when and how much to use a low-code platform depends on the range of applicability, development speed, manageability and flexibility for performance limitations.

However, if there is a need to develop a product that is high quality and unique while being specific to business requirements, custom product development may be a better option. END

By: Narasimha Sekhar Kakaraparthi and Radhakrishna Singuru

Narasimha Sekhar Kakaraparthi has 27 years of experience in product development for infrastructure management, cloud, as well as telecommunication, manufacturing and technical delivery. He is currently working as principal architect for virtual desktops and end user services at Wipro Technologies. He is a senior member of Wipro's eminent architects group called Distinguished Member of Technical Staff (DMTS).

Radhakrishna Singuru 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 chief architect, Cloud Platform Products and Security Engineering Practice, ER&D at Wipro Technologies. He is also a senior member of DMTS.

Disclaimer: *This article expresses the views of the authors and not that of Wipro.*

FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK

We Value Your Feedback

We love to hear from you as **OSFY** consistently strives to make its content **informative, interesting** and **engaging.**



Please share your feedback/ thoughts/ views via email at osfyedit@efyindia.com.

We welcome your comments/ suggestions and encourage you to send them to:
The Editor, D-87/1, Okhla Industrial Area, Phase 1, New Delhi-110020.

OpenSource For You
THE COMPLETE MAGAZINE ON OPEN SOURCE

FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK FEEDBACK

Why You Should Deploy Immutable Infrastructure on the Cloud

Even as DevOps and cloud adoption have increased significantly in recent years, the use of immutable infrastructure has also become pretty popular. This article outlines the core benefits of deploying immutable infrastructure on the cloud.



Immutable infrastructure refers to a state where the infra components like VMs, servers and containers are never modified once deployed. In case any changes are required, the existing components are destroyed and replaced by new ones.

In the real world, we often make changes in security patching and while deploying a new application version, as well as OS-specific changes. All these changes are made 'in-place' or with the help of configuration management tools like Chef, Ansible, Puppet, or SaltStack. That makes the whole infrastructure unstable and prone to configuration drift. This is where the concept of immutable infrastructure comes in.

Under this approach, a golden (master) image of a server is created that has all the required software stacks preconfigured. The same image is used to provision new servers. If any changes are required in the configuration of the servers, a new image is baked, tested, and rolled out across environments.

Some of the benefits of this approach are listed below.

- Creates consistency across environments (dev/stage/prod), as all the servers are provisioned with the help of a particular golden image.
- Makes infrastructure more stable because everything is immutable, as 'in-place' changes are not permitted. Additionally, eliminates the risk of configuration drift.
- Simplifies the deployment process because all the components are preconfigured on the server. Also makes rollback safer since the same process used to deploy the new version is used to roll back to older versions.
- Makes debugging much faster during production issues since all the servers, provisioned using golden images, are properly versioned and well documented.
- Enhances security as the images are baked while keeping compliances in mind and validated after thorough testing.
- Finally, embraces Agile software deployment in practice by implementing DevOps principles.

Terraform, Vagrant, Docker and Packer are some of the open source tools available for creating and managing immutable infrastructure on the cloud.

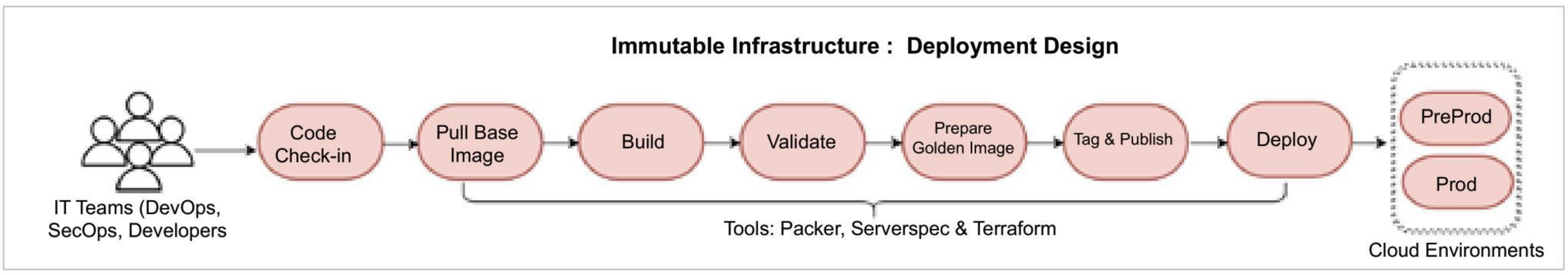


Figure 1: Deployment design

The challenge of using immutable infrastructure

One of the core practices of DevOps is 'continuous delivery', wherein every single change is promoted from development to the production environment automatically. This sounds pretty fascinating and rather suitable for releasing a new version of an application, but in the real world, bringing it into practice is quite challenging, primarily when the enterprises fall under one of the following criteria.

- **Heterogeneous environments:** Here, applications are deployed on multiple environments (i.e., dev, stage, preprod or prod) or on more than one cloud (i.e., AWS, GCP or Azure).
- **Infrastructure auditing:** Lots of security tickets are received from SecOps routinely, which require immediate attention to fix the requested vulnerabilities.
- **Monolith applications:** The applications deployed on the cloud still follow the monolith design pattern, and decoupling them is nearly impossible at the moment.
- **Complex tooling:** The toolchain designed to promote changes from development to production environments is quite complex and requires manual intervention while making changes.

Since there are no clearly defined design practices to automate the deployment of immutable infrastructure, enterprises still prefer to manage the infrastructure in traditional ways, with the help of configuration management tools like Chef, Ansible, Puppet or SaltStack.

Tools	Description
Packer	Packer is a great tool to create identical machine images from a single source configuration for multiple platforms. For example, the pipeline depicted in Figure 1 uses Packer for baking the golden images from a base image.
Terraform	Terraform is an Infrastructure as Code tool that enables infrastructure provisioning on cloud and on-premise environments. Likewise, we use it to deploy cloud servers using the golden images.
Serverspec	Serverspec is an integration test framework that checks the servers' actual state and ensures they are provisioned correctly. It is used in the pipeline in Figure 1 for validating the infrastructure state.
Jenkins	Jenkins is an open source automation server that enables lots of integrations with the help of some great plugins. Jenkins is used in designing the end-to-end deployment pipeline described in Figure 1. Additionally, it allows a flawless integration of Packer, Serverspec and Terraform.

Solution: Automate immutable infrastructure deployments

Now that we have covered the core concepts of immutable infrastructure and the barriers to bringing it into practice, let us try and understand the patterns, processes, and practices needed to deploy it in any cloud environment.

Figure 1 illustrates the logical representation of the entire workflow, where the different phases of a pipeline are mentioned. The pipeline starts its execution by fetching the vanilla (base) machine image from the respective cloud repository. Then it pulls the latest application code commits, builds, and compiles. Upon successful validation, it creates a golden image from a running server, and publishes it to cloud repositories. And, finally, the golden image is deployed to the respective cloud environments.

The table above lists some excellent open source tools that can be used to implement the design workflow illustrated in Figure 1.

Deploying a demo application on AWS cloud

Now, let's implement the deployment design illustrated in Figure 1 and use it for launching an application instance on AWS.

The first step is to set up a Jenkins server and configure AWS credentials on it. Then, install Packer and Terraform binaries on the Jenkins server to launch the application instance in the AWS environment. Next, create a Jenkins job using the declarative pipeline given below:

```
pipeline {
    stages {
        // Pull latest code
        stage('Code Check-in') {
            steps {
                checkout([$class:
'GitSCM', branches: [[name: '*/
main']], userRemoteConfigs: [[url:
'https://github.com/vijayg92/immutable-
infrastructure']]])
                sh('ls -ltr')
            }
```

```
        }
        // Validate Packer template
        stage('Template Validation') {
            steps {
                withCredentials([str
ing(credentialsId: 'AWS_ACCESS_KEY_
ID', variable: 'AWS_ACCESS_KEY_ID'),
string(credentialsId: 'AWS_SECRET_
ACCESS_KEY', variable: 'AWS_SECRET_
ACCESS_KEY')]) {
                    sh script: 'cd
packer/webserver/ && /opt/tools/packer
validate -var-file=vars.json app.json',
returnStatus: true
                }
            }
        }
        // Build Golden Image
        stage('Build Golden Image') {
            steps {
                withCredentials([str
ing(credentialsId: 'AWS_ACCESS_KEY_
ID', variable: 'AWS_ACCESS_KEY_ID'),
string(credentialsId: 'AWS_SECRET_
ACCESS_KEY', variable: 'AWS_SECRET_
ACCESS_KEY')]) {
                    sh script: 'cd
packer/webserver/ && /opt/tools/packer
build -var-file=vars.json app.json',
returnStatus: true
                }
            }
        }
        // Validate Terraform Templates
        stage('Validate') {
            steps {
                withCredentials([str
ing(credentialsId: 'AWS_ACCESS_KEY_
ID', variable: 'AWS_ACCESS_KEY_ID'),
string(credentialsId: 'AWS_SECRET_
ACCESS_KEY', variable: 'AWS_SECRET_
ACCESS_KEY')]) {
                    sh script: 'cd
terraform/webserver/ && /opt/tools/
terraform plan -auto-approve',
returnStatus: true
                }
            }
        }
        // Deploy golden ami to AWS
cloud
        stage('Deploy') {
            steps {
                withCredentials([str
ing(credentialsId: 'AWS_ACCESS_KEY_
ID', variable: 'AWS_ACCESS_KEY_ID'),
string(credentialsId: 'AWS_SECRET_
ACCESS_KEY', variable: 'AWS_SECRET_
ACCESS_KEY')]) {
                    sh script: 'cd
terraform/webserver/ && /opt/tools/
terraform apply -auto-approve,
returnStatus: true
                }
            }
        }
    }
}
```

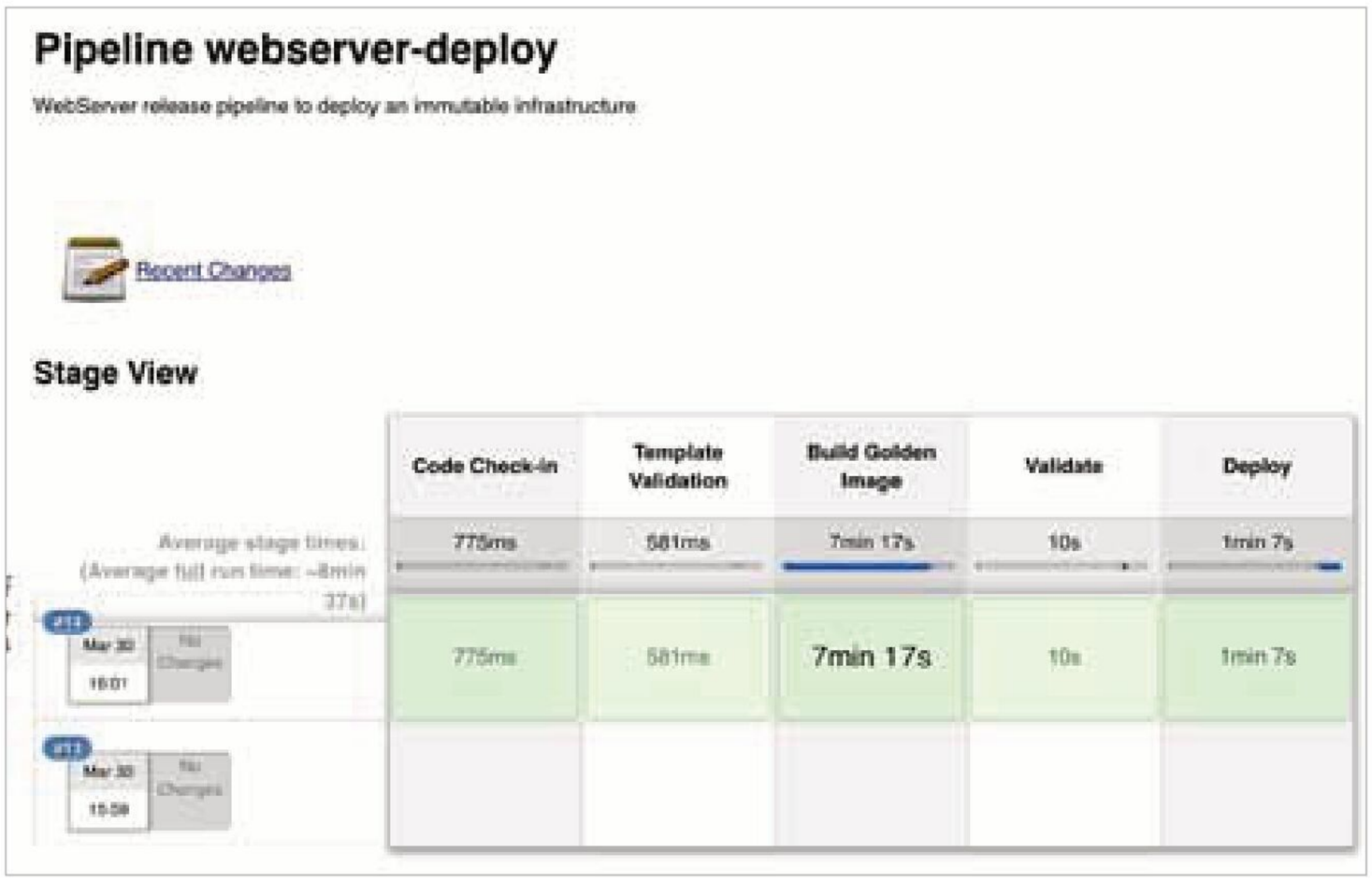


Figure 2: Jenkins pipeline



Figure 3: Final output

Finally, trigger the pipeline job, and wait for it to finish. If everything goes well, as shown in Figure 2, the application instance should be up and running in the AWS cloud (Figure 3).

The benefits of immutable infrastructure are not just restricted to faster deployments, safer rollbacks, enhanced security and scaling, but allow service owners like SysOps, SecOps, SREs, developers, etc, to move away from older, time-consuming ways of managing servers. END

References

[1] *https://github.com/vijayg92/immutable-infrastructure*
[2] *https://martinfowler.com/bliki/ImmutableServer.html*
[3] *https://www.hashicorp.com/resources/what-is-mutable-vs-immutable-infrastructure*
[4] *https://eplexity.com/blog/benefits-of-immutable-infrastructure/*

By: Vijay Gosai

The author is an open source enthusiast and is working as a senior site reliability engineer. His interests include distributed systems, cloud computing, infrastructure automation, software development, design patterns, DevOps and emerging technologies.

DevSecOps: A Philosophy that Puts Security First

DevOps has helped remove barriers between development and operations, to improve productivity as well as the reliability of software. However, it has fallen short on some aspects such as security, leading to the emergence of the DevSecOps philosophy. Find out more about it in this article.



Though a lot of people understand the security concerns that are inherent in DevOps, most do not know how to overcome the problem. The new security-focused spin-off DevSecOps could solve these problems if implemented properly.

In order to understand and appreciate DevSecOps, we first need to understand the benefits and the shortcomings of DevOps.

Pros of DevOps

Reduced time-to-market: Not too long ago, it used to take anywhere close to a year or two for a simple software release, with big companies claiming that they would need to estimate time, resources, and the manpower required to complete a task, and so on. Today there are companies that are shipping out software on a daily basis. This has been possible because of the adoption of the DevOps philosophy.

Global presence: DevOps has without a doubt helped companies manage a global presence. Rewinding back a few years, only the major players in the field of software development were able to ship their software across the globe. Today, that has almost entirely changed as a company situated in India can assist a client from any corner of the world. Similarly, Indian customers can get the services of any company located anywhere in the world.

Software democratisation: One of the reasons why companies demanded a lot of time to build software was because they had to literally develop everything from scratch. Every library and all the necessary tools had to be built from the ground up, which was a huge task. But today that has changed completely, as all the required libraries that can perform basic tasks can be used by anyone, and a developer

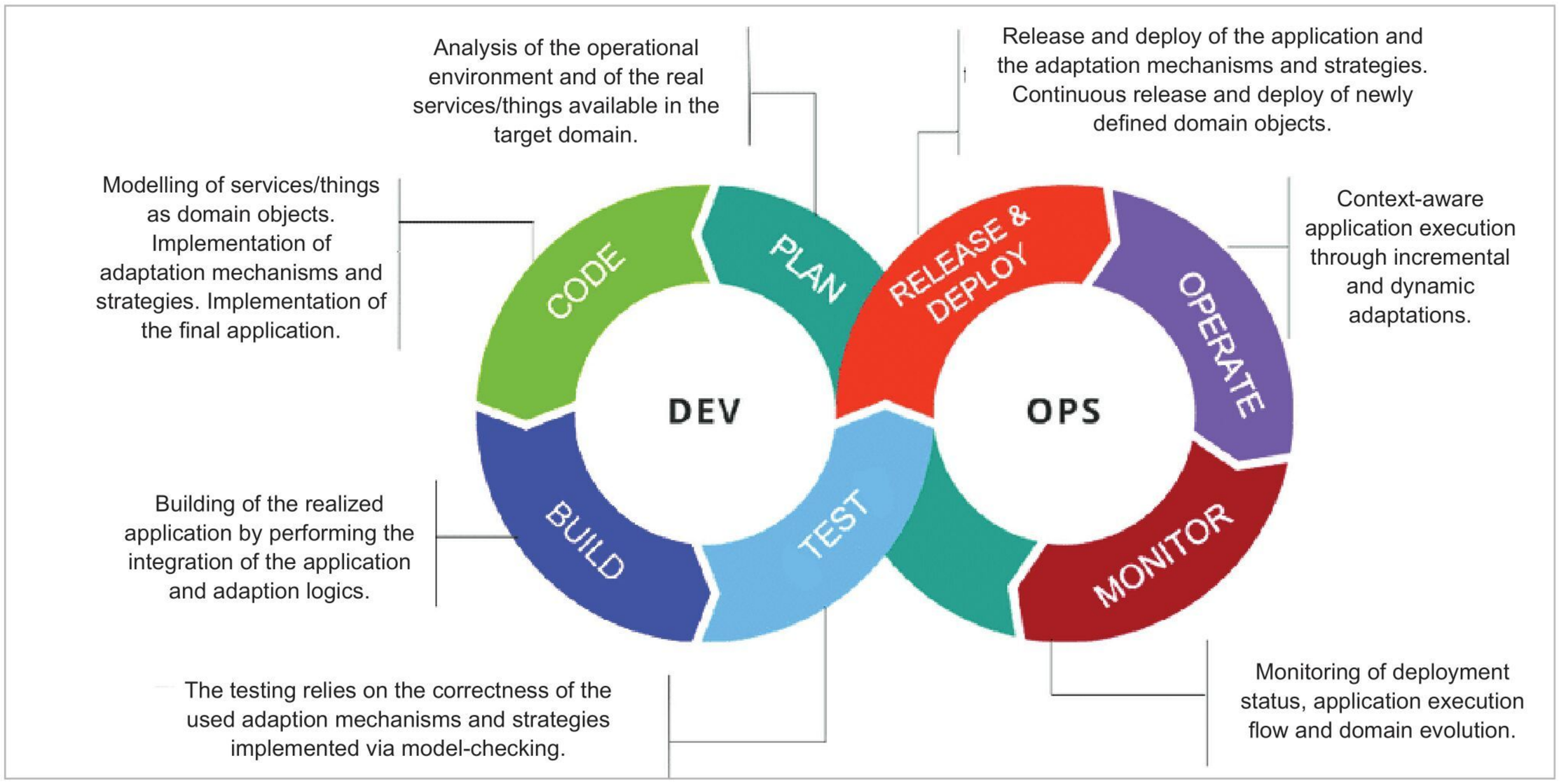


Figure 1: Detailed DevOps workflow

can just start to build on top of these like a matrix. This is a convenience, especially for those who rely heavily on these libraries. A good example of this is what React and Angular have been doing for Web development.

The con: Lagging security practices

However, software security still lags behind in spite of all the progress that is being made. We get to hear about major data breaches occurring around the world and often in large companies even to this day. One of the biggest data breaches happened at Equifax in 2017 on 13th May and was not discovered until 29th July! Once the breach was discovered, effective action was taken within 24 hours. However, a lot of sensitive data such as credit card details, social security numbers, and other important documents were accessed over the two-month period. According to some estimates, almost half of the American population was impacted.

Damages incurred due to poor security

In India, awareness about such security concerns is low, so the severity of this might be hard to grasp for us. These types of data breaches tend to occur from time to time in our country as well. Here are some of the major data breaches that have taken place in India just over the last year.

There was a data breach at Bigbasket in October 2021, which sacrificed the information of over 20 million people. Then there was a breach at Juspay, a payment gateway, in which over 35 million people's primary information such as names and addresses, as well as financial details such as card numbers and banking credentials were revealed. In a security breach at Dominos, the data of over one million people was put at risk. Another major breach occurred very recently at Air India, where over 4.5 million people were affected.

According to research by IBM, it has been found that the average cost of a data breach is over 5 million US dollars. But what is interesting is that it is about 30 times cheaper to fix the defects during the development stage of software, compared to once it goes out into production. Though this is still a considerable amount of money, at least it saves companies from incurring a bigger loss.

So the people behind the development of DevOps are now fostering a new paradigm or a philosophy, which they term as 'Dev + Sec + Ops' (philosophy of integrated security practices within the DevOps process). This basically integrates security practices with the existing DevOps process and philosophy to make sure that security is the most important aspect in every stage of software development, and not an afterthought.

What exactly is DevSecOps?

DevSecOps is still in its early phases and is continuing to mature. There are already a whole bunch of tools available, but there is still plenty of work to be done. The first important aspect that it covers is that it helps in identifying the security issues, unlike in DevOps where these are generally just an afterthought. The security team goes through a higher level of overview and makes sure that security is on point, before giving the final nod in the development process.

DevSecOps also gives speed and agility to the security team so they can

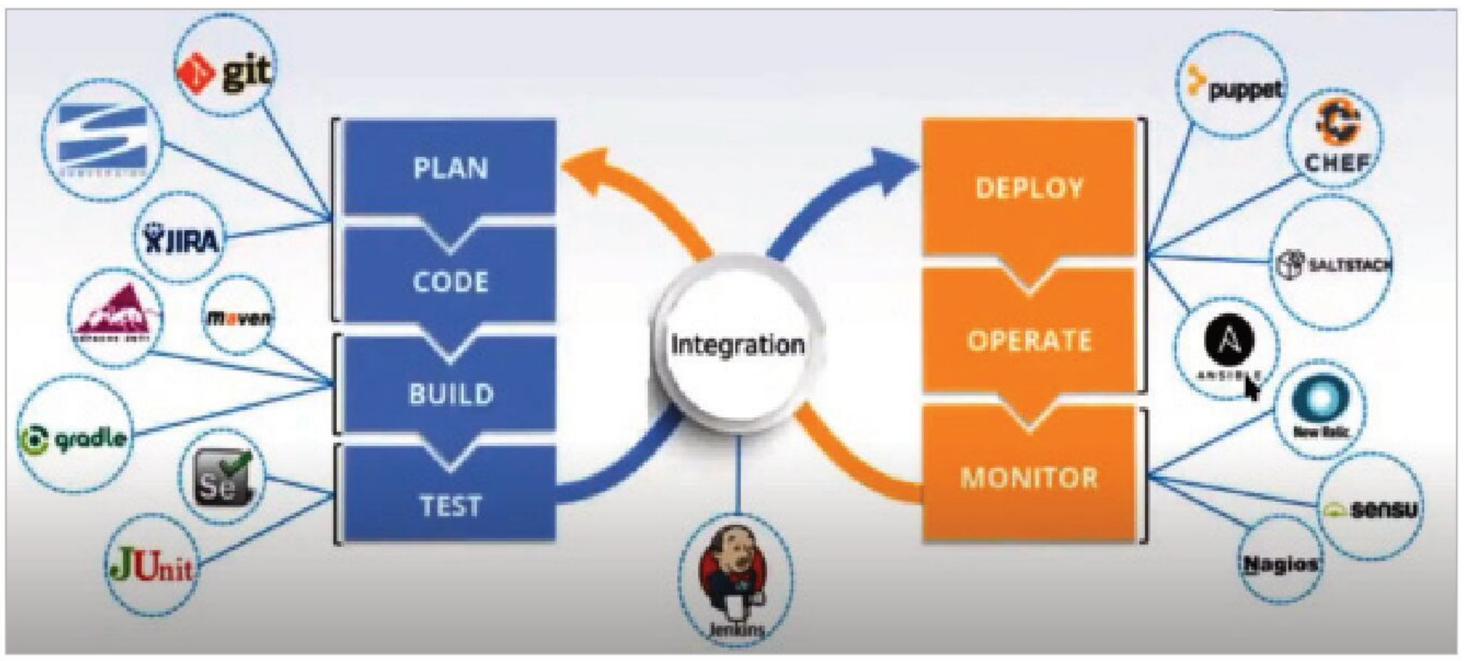


Figure 2: Integration of various DevOps tools with DevSecOps tools

Tools available at your disposal

Some of the tools that are already being put to use in the DevOps toolchain can be used to manage DevSecOps too. No new tools are needed for DevSecOps. Tools such as Jira, Git, Gradle, and Selenium are a few that can be used for planning, coding, building and testing software. They are used in DevOps but are compliant with DevSecOps as well — it is just a matter of how they are implemented. Tools such as Puppet, Chef, and Sensu can be used for deployment, operation and monitoring (refer to Figure 2).

figure out the issues right from the word go rather than finding them at the end. It allows them to respond to the changes rapidly when there are a lot of new laws that have to be abided by such as the General Data Protection Regulation (GDPR). When the security team is involved from the beginning, it helps them to deal with issues more swiftly. More importantly, it helps to create more automated builds, and more QA testing can be done.

Automating the entire process in DevOps is a very crucial aspect, and is one of the core pillars on which it stands today. The more things can be structured and automated the better the software is, as it tends to be more scalable that way. Automation helps in maintaining better communication between the security and development teams, which is something DevOps has always advocated.

Figure 1 depicts the workflow that DevOps usually advocates, on top of which the layers of security are supposed to be added at the planning stage. This is the stage at which the security team is already thinking about what kind of threat model policies they want to build in, and the issues that they could possibly encounter when deploying the final product.

Once the coding pipeline where code is being generated is reached, the static analysis and code reviews are to be done dynamically. When it gets into testing, the security team must make sure that the code is hardened against the known vulnerabilities. This will significantly lower the risk of attacks. Some kind of pen testing can be done while making sure that the security planning happens at the planning stage itself and not at the test stage.

Once the code is ready to be deployed, it is essential to continuously monitor and assess it for any kind of threats and vulnerabilities so that the application is not impacted. It must also meet all necessary compliances.

The benefits of using security as code

As much as the automation of the entire workflow is important, there is an extent to which it can be done using tools. Managing a single server manually is not a big deal, but to manage tens of thousands of servers together every day can be super challenging. A code can be put in place in the form of an application, which can very easily manage these servers on its own. It can be instructed to do so according to the required security protocol. Let us look at some of the benefits of using security as code.

Collaboration: A code is an unambiguous common language that can be understood by literally everyone, even if a person doesn't have any prior understanding of it. Anyone can learn to code and it will perform the same way wherever it is implemented. This makes it significantly easier to work with people from all walks of life.

Scalability: It is highly scalable. Irrespective of the number of servers or nodes being managed, a piece of code is going to perform the same way wherever it has been deployed. This leaves almost no room for error and minimises the risk of malfunctions.

Visibility: Since the code is always visible, it can be edited anytime it is necessary to do so. Whenever there is a bug that might cause any kind of hindrance, it can be rectified very easily without any significant damage. It can be altered at will at any given point of time.

It's a no-brainer that security shall always be the utmost priority for anything that you do on the Web, and anything that comprises system hardware and software. Weak security can lead to irreversible losses. Hence, it becomes imperative for developers to implement foolproof security using evolving philosophies like DevSecOps. END

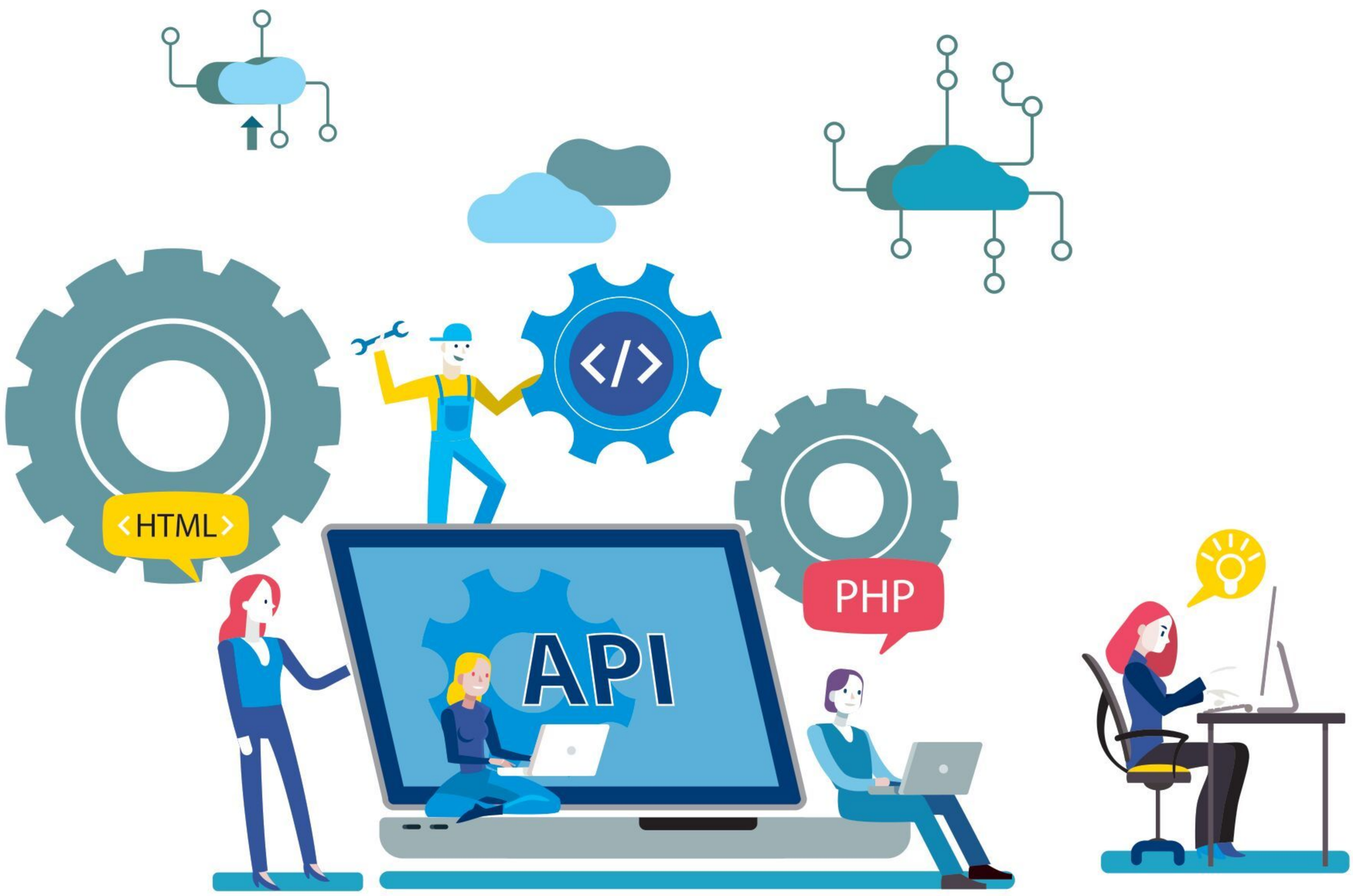
By: Jonathan Pereira

The author is DevRel at Polygon. This article is based on a talk he gave at OSI 2021, a virtual event.

Transcribed and curated by Laveesh Kocher

Demystifying Observability for Modern Applications

To come up with an effective solution to a problem, you need to have the relevant insights into it. Distributed systems are prone to unexpected failures and malfunctions, and observability helps us find out the reasons for these in order to arrive at an effective solution.



Operating a distributed system is hard, not just because of its complexity, but also because of the unpredictability of failure modes. With quick software deliveries, continuous build deployments and modern cloud architectures, the possible failure modes are increasing. Unfortunately, traditional monitoring tools cannot help us to deal with these challenges anymore.

What caused these problems to begin with?

Tech giants like Amazon and Microsoft started building their applications with monolithic architecture as it was more convenient back in the day. They all had to go through the same challenges and hence chose microservices as a better solution, as these can be developed, scaled, and deployed independently. This significantly increased the delivery speed and scalability of applications, but on the flip side, they now have to deal with a different kind of operational complexity related to managing these microservices deployments. One of the benefits of working with older technologies was that only a limited set of failures occurred.

Using application programming interfaces (APIs) to expose core business functionality and facilitating service-to-service communication gives several control points, and makes it easier to build these complex modern applications. However, API-driven applications come with their own set of issues such as design complexity, lack of end-to-end visibility, and security concerns. We are going to discuss that in detail.

A peek into modern problems

Every company or organisation that is adapting to these microservices and

API-based architectures is addressing these four basic questions:

- Are the services/APIs providing the functionality they were designed to deliver?
- Are the services/APIs secure?
- As a business do you understand the usage of APIs by users?
- Are the services/APIs providing the optimum performance to the user?

Visibility into dependencies

While addressing these questions, there are several problems that come up. Even as there are different solutions to these problems, there are simply too many gaps in them. There are several approaches to deal with and answer these questions. We can look into their dependencies. Microservices architecture, in general, is inherently very complex compared to the current distributed architecture that is being used. There is no end-to-end visibility into dependent services.

It is extremely important to understand the failure point. This is, in fact, the very first step of the debugging process because whenever something breaks we can only fix it if we have a good understanding of what is broken. It is difficult to decipher the connected metrics between services. So, in general, the visibility aspect is essential to this modern and complex architecture.

In order to find the problem areas, you need to correlate the multiple dashboards that are being used. Most developers are often trying to figure out this correlation between different services and metrics manually, which also increases the cognitive load on the development.

The role of logs

The metrics will tell you what broke, and the logs will tell you why it broke in the first place. The first logical step after looking at your metrics is to go to logs and understand why the failure happened that led everything to behave abnormally. Query logs by requestID result in a huge number of logs in the microservices world. This is because if you have one call that is touching 1000 different services, then you have to deal with a huge pile of the log. And sequential navigation through logs to find out the root cause is hard. It is also hard to correlate due to the sheer amount of logs present.

Debugging slow queries

Sometimes your customer data does not load properly and your customers are unable to log in because the database having their data is slow. To deal with this, the traditional way is to manually log slow queries into the logging system. This also leads to an increased dependency on DBAs to get slow query logs.

Why we need observability

"Monitoring tells you whether a system is working; observability lets you ask why it isn't working"
— *Baron Schwartz*

All questions about visibility, service interactions and slow queries can be answered by observability. In control theory, observability is defined as a measure of how well the internal states of a system can be inferred from the knowledge of that system's external outputs. Simply put, observability is how well you can understand your complex system.

Need for observability

- The monoliths to microservices transition is happening and has happened in the past as well.
- Microservices create complex interactions.
- Failure modes are unpredictable these days.
- There is a rise in the API ecosystem.
- Monitoring can no longer help us.

Ability to answer abstract questions

If we are placing an order on Amazon and something goes wrong, then what characteristics did the queries that timed out at 500ms common — service versions, browser plugins?

Instrumentation produces data and querying the data answers this question. You may be wondering what this instrumentation data is. Well, there are three data sources that we are talking about here — metrics, logs and distributed tracing.

Metrics are basically the aggregated summary statistics. These are valuable because your operating system and applications generate metrics that are going to contain a name or a time stamp and a field that represents some value. They are the first obvious place to start with when it comes to monitoring. One issue that metrics present are that they lack context. You can see something weird going on with your system but you will never know what, unless you look into the second pillar — logs.

Logs contain detailed debugging information emitted by processes. They play perhaps the second most important role in your toolbox because they tend to give us more in-depth information about resources than metrics. So, if metrics show us that a resource has died, logs will let us know why it died in the first place. The problem with logs is that they are hard to correlate and bring into context. This brings us to our next pillar — distributed tracing.

Distributed tracing provides insights into the full life cycle, aka traces of requests to a system, allowing you to pinpoint failure and performance issues. It adds critical visibility into the end-to-end health of the application. For instance, whenever a request comes to the front-end, a unique ID is assigned to it and that request gets tracked throughout the system. Along the way, we collect telemetry data, which can be shown in different formats.

What is Hypertrace?

Hypertrace is an open source distributed tracing and observability platform that helps to make sense of the production request. It converts data into relevant insights for everyone. It helps the infrastructure team to identify which services are causing overload, allows the application team to understand which requests put their service objectives at risk, and lets the deployment team find the problems.

Why should you choose Hypertrace?

From extensibility and interoperability to extremely effective logs and alerts, Hypertrace has several features that make it a good choice for observability. Here is a quick look at some of the key features of this platform.

- *Topology:* Hypertrace lets you view the topology of all services and backends in real-time. It is scalable for hundreds of them.
- *Dashboards:* It provides you pre-scanned dashboards for services, APIs, and backends.
- *Extensibility:* You can write your own enrichers to enrich trace data for your business needs and create a view around them.
- *Interoperability:* Hypertrace works out-of-the-box with all open source tracing formats like Zipkin, OTel and Jaeger.
- *GraphQL APIs:* It traces, spans, and exposes entities through GraphQL APIs. It helps build the next creative use case.
- *API and trace analytics*: It helps to slice and dice all data powered by Apache Pinot.
- *Span logs:* Hypertrace speeds up root cause analysis (RCA) and reduces mean time to detect (MTTD) with span logs on the platform.
- *Error analytics:* It lets you find issues quickly with error highlighting and filters.
- *Alerting:* You get an alert on any deviation/anomalies in the out-of-bag (OOB) metrics.

Visibility into dependent services

The old way: Whenever something broke, say, you found out that service X had some latency issues, you had it checked with the service owner to understand dependencies. Then you also looked at metrics that were exported to Prometheus manually.

The new way: In the new way, the dependency graph is already created with real-time metrics based on the available data.

Finding hotspots

The old way: Multiple dashboards were required to correlate.

The new way: Complete request is visualised in a single pane.

Log fatigue

The old way: You had to query logs by requestID and dig through all the logs generated across all the services for this request. The number of log lines also increased based on the number of dependencies.

The new way: You just figure out the problematic span and find correlated logs in the same view. If a particular span is taking more time than was accounted for, then you can click on it and see correlated logs in the context.

Debugging slow queries

The old way: There was a dependency on DBAs to get slow query logs. You had to manually log slow queries into the logging platform and create dashboards.

The new way: 'Automagic' creation of query dashboards can be done today.

Hypertrace deployment

From the infrastructure end of things you can easily deploy Hypertrace on any of the container environments like Kubernetes, EKS, ECS, AKS, VMs, GKE or Docker. It is compatible with Zipkin, OpenTelemetry and Jaeger. Language modules such as Java, Python and Go are also supported (refer to Figure 1).

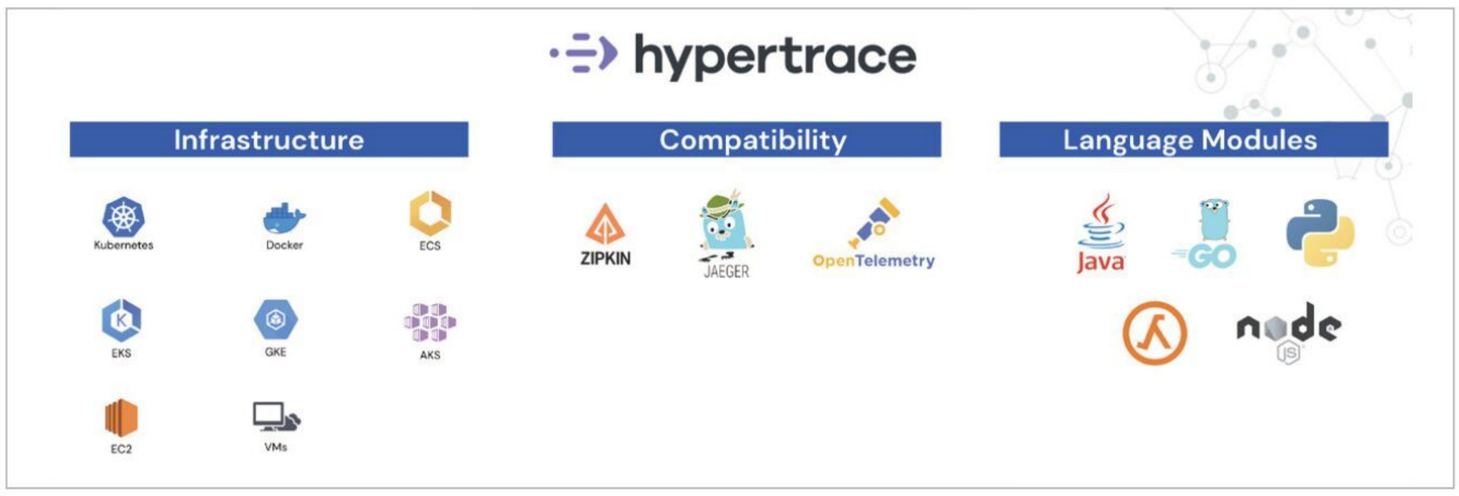


Figure 1: Infrastructure, compatibility and language modules supported by Hypertrace

A host of such desirable features makes Hypertrace a great cloud-native distributed tracing based tool for observability. It is evolving tremendously as people are continuously contributing towards it. Definitely something to check out! END

By: Jayesh Bapu Ahire

The author is founding engineer/product manager at TraceableAI. This article is based on a tech talk he gave at OSI 2021, a virtual event.

Transcribed and curated by Laveesh Kocher

Planning the DevOps Journey for an Enterprise

Large enterprises are looking to implement efficient site reliability engineering and a DevOps culture that is adaptable, agile and intelligent. But they need to strategise for that.



Whether it is an on-premises or a hybrid environment, the architect looks for synergy. This maturity is factored around various processes, including a questionnaire that collects elementary information about an organisation. Indeed, an enterprise must measure its IT, people, tools, and processes, and then create a path based on this snapshot to plan the DevOps journey. Aligning the company and its people, technology and processes to a strategy is more a philosophical and less a technology or operational change. The success, therefore, is human-centric and incremental over time, based on values.

How a company can embrace DevOps in its culture and maturely deliver services is the key aspect we need to explore. An organisational approach enables faster development of applications and easier maintenance of existing deployments by creating stronger bonds between developers, operations and other stakeholders in the company. How a company formulates it, however, is more crucial for a successful DevOps practice. Typically, it should start with a maturity assessment, gap analysis and strategy formulation before actual implementation and iterative roll-out. At the same time, the company must convince the sponsor, architecture and governance, people, key stakeholders, dev and support line about its needs, feedback loop and improvements.

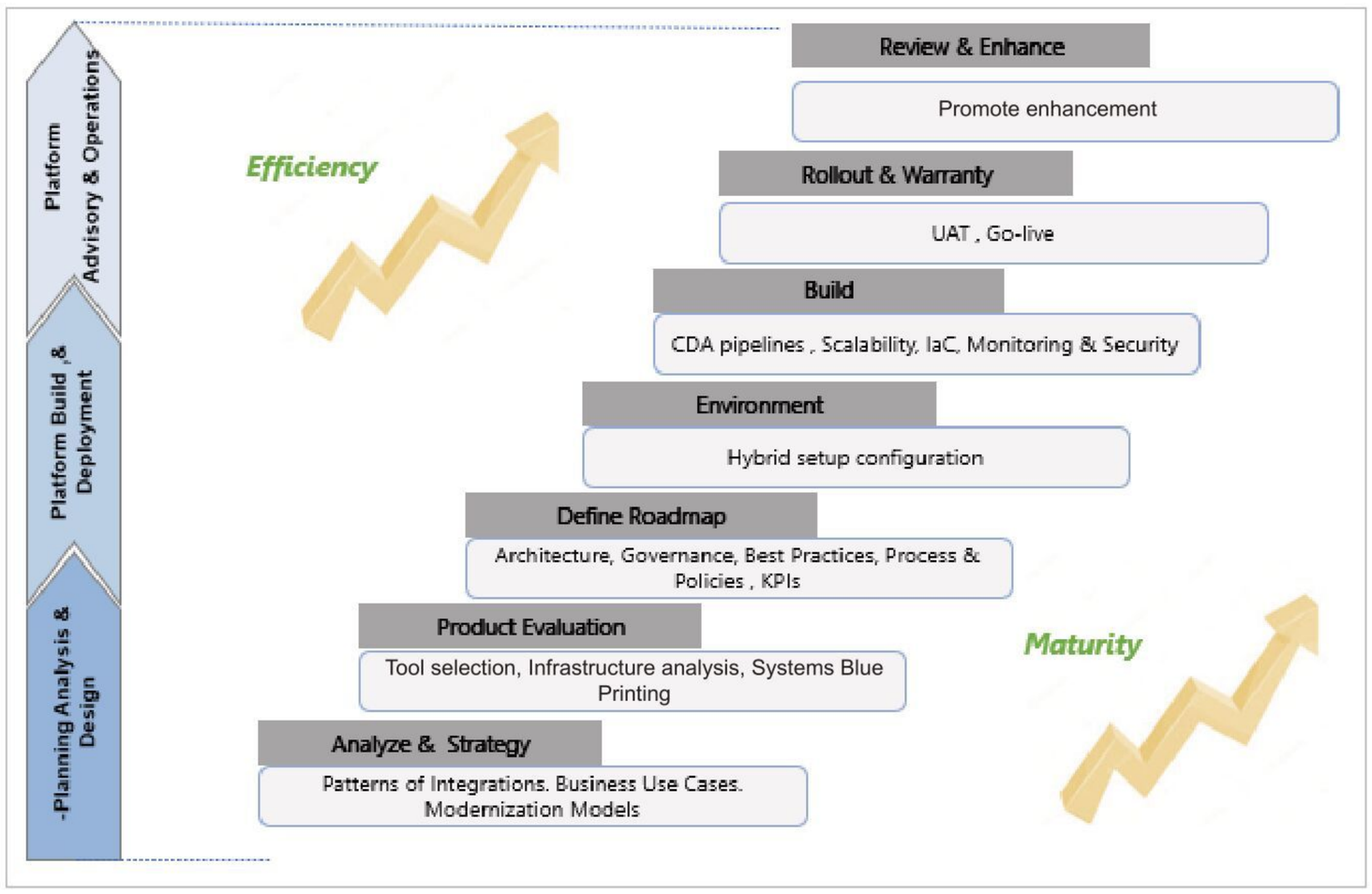


Figure 1: An outline of the approach to typical DevOps adoption

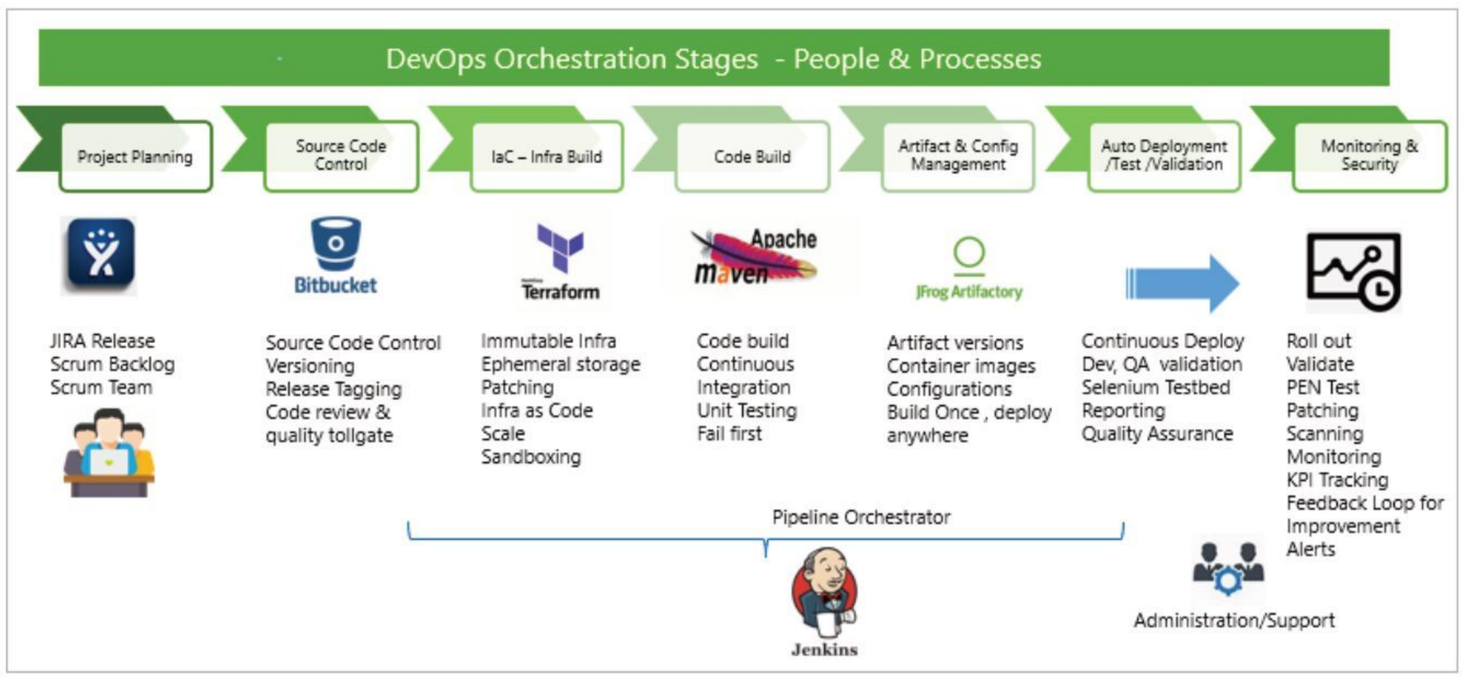


Figure 2: Java app DevOps pipeline

This need with synergy between people and processes. There's also a demand for stringent security reviews, especially in the post-pandemic digital first era where banks, retail outlets, healthcare services and automotive sectors are seeing the biggest boom in digital marketing and sales. Figure 1 gives a brief outline of typical DevOps absorption in an enterprise with iterative improvement and retrospection at every point.

Analysis and assessment are vital to help understand the operations and delivery processes, deployment, monitoring, reliability, traceability of businesses and technical processes in an organisation's landscape. This results in creating a minimalistic model of execution, which tailor-fits an enterprise. Key dimensions to be considered while making an assessment are mentioned below:

- Current people processes and operations
- Retrospection, feedback and improvement
- Technology roadmap – cloud focused?
- Iterative
- Development and deployments
- Operational KPIs and SLA management
- Security and governance
- Continuous integration and automated testing
- Time to market and release cadence

This helps to evolve a consolidated strategic and robust DevOps framework, which is scalable and agile enough to improve with time.

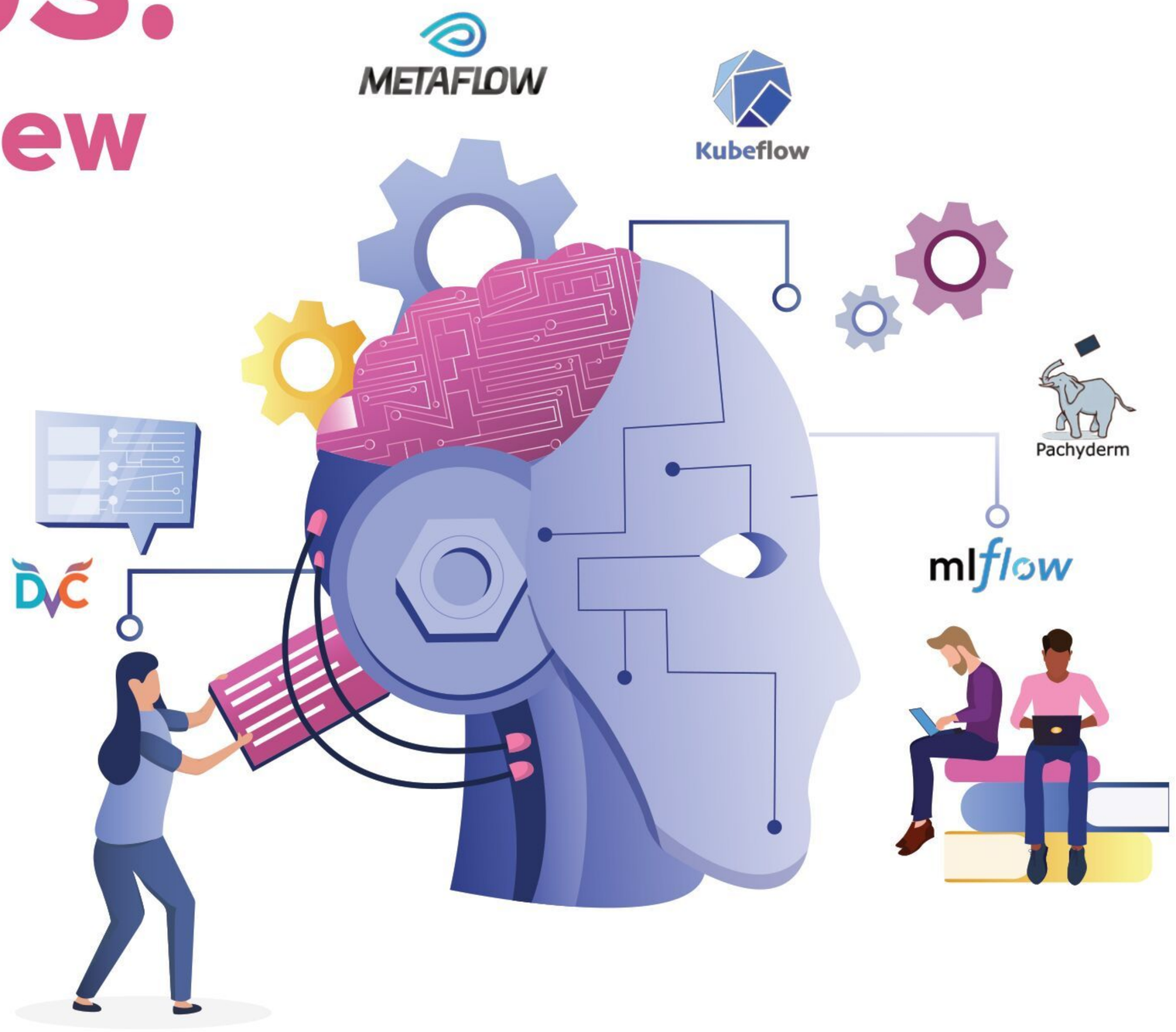
An indicative Java app DevOps pipeline approach in Agile iterations is shown in Figure 2, which is self-explanatory with respect to the various stages and tasks associated. Using the power of open source tools, people and operations with well-orchestrated API components for resiliency is the ultimate goal.

Finally, with the advent of machine learning as well as cognitive tools and services, many from the open source fraternity will be bringing in operational intelligence in the near future. In fact, robotic process automation tools and various service management and APM product lines are in the process of creating an intelligent automated framework for DevOps, with security and controls for their enterprise customers to bridge the gap between culture and synergy. An early security review in a fully automated way while releasing product patches may see the biggest traction in the world of malware and spoofing. By 2026, DevOps and DevSecOps in convergence with SysOps will touch a market of around US$ 15 billion, according to a *Fortune Business Insights* report. END

By: Joydeepta Bhattacharjee

The author is a technical evangelist and cloud platform architect working with an MNC for the last 16+ years. He is very passionate about open source adoption across technologies.

Open Source Tools for MLOps: An Overview



MLOps makes budgeting overall resources and computing easier. It also helps to automate end-to-end machine learning workflows and improve team collaboration. This article will introduce you to a few free and open source tools for MLOps.

The impact of the pandemic on different industries, organisations, project development teams and consumers is still being assessed, and its consequences will be felt for years to come. However, the pandemic years have spurred more innovation and inventions than ever before with the help of developers working remotely in a virtual work environment. Technology has helped people to work, acquire and receive services, and connect with each other for better performance. Developers and DevOps experts suddenly have a plethora of possibilities for rapid improvements and new developments at work.

DevOps is associated with assorted domains for multiple applications. The corporate world is migrating to DevOps based deployment and execution of projects for better performance.

Integration of DevOps with artificial intelligence

Machine learning (ML) operations, abbreviated as MLOps, integrate artificial intelligence (AI) and data science with contemporary DevOps processes. This helps intelligent applications to integrate with DevOps along with data analytics and feature engineering.

This boosts the ML life cycle process by creating a better environment for cooperation and communication amongst teams.

For a higher degree of efficiency, MLOps blends machine learning with pre-existing application development and IT operations. Machine learning also aids in the development of a strong automation system for quality control, pipelining, monitoring, and a range of other processes.

A great open source MLOps solution provides users with a platform that allows them to work with complete operational independence at a cheap cost and with total access to all essential resources. It gives enterprises greater space to cooperate and more freedom on one platform.

MLOps allows IT organisations to enhance efficiency in model deployment and governance while also automating the workflow's repetitive manual tasks. As a result, it is being used in an increasing number of development projects.

A data science project focuses on leveraging data to develop a predictive model, which gives us a near-perfect outcome that mimics a real-world business scenario. When attempting to attain this goal, it's easy to overlook the many other parts of

a data science project, particularly the operational ones. Because ML projects are iterative in nature, keeping track of all the variables, configurations, and outcomes may be a challenging task in itself. Effective team cooperation is becoming increasingly important as data science teams get larger and more dispersed.

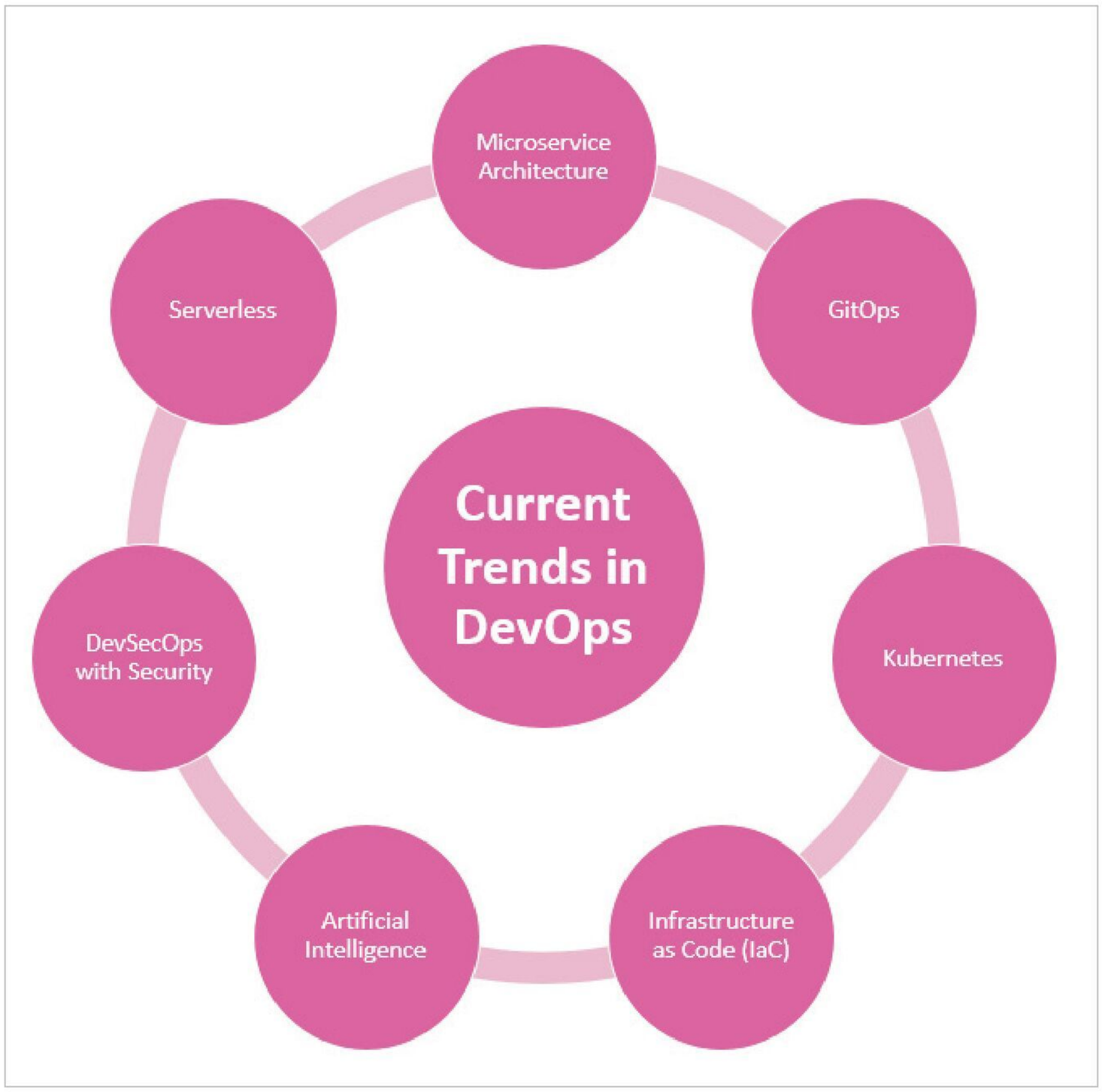


Figure 1: Recent trends in DevOps

Free and open source tools for MLOps

A number of suites and platforms are available for the implementation and deployment of MLOps for different applications. A few key platforms, using which MLOps deployment and overall control on the process becomes easy for project managers and developers, are listed below.

Kubeflow

kubeflow.org

Open source MLOps platform Kubeflow makes orchestrating and deploying machine learning workflows very easy for developers. Machine learning training, pipeline building, and Jupyter notebook management are all supported by Kubeflow's dedicated services and integrations. It's easy to interface with frameworks like Istio, and can even perform TensorFlow training tasks without any issues.

MLflow

mlflow.org

This open source machine learning life cycle management platform provides several components for experiment tracking, project packaging, model deployment, and registry. Machine learning libraries such as TensorFlow and PyTorch may be easily integrated with MLflow to speed up the training, deployment, and maintenance of applications.

Data Version Control (DVC)

dvc.org

An open source tool for data science and machine learning applications, DVC is written in Python programming language. Data sets and machine learning models are managed and versioned in a Git-like fashion. With DVC, machine learning projects may be shared and replicated easily using a command-line interface.

Pachyderm

github.com/pachyderm/pachyderm

Pachyderm is an MLOps data pipeline and versioning leader. You can automate and scale your machine learning life cycle while ensuring reproducibility with its data foundation. It comes in a commercial Pachyderm Enterprise Edition and an open source Pachyderm Community Edition thanks to financing from Benchmark, Microsoft M12, and others. Using Pachyderm, customers can get their ML and AI initiatives off the ground faster, minimise their data processing and storage expenses, and comply with stringent data governance guidelines.

Metaflow

metaflow.org

ML models can be trained, deployed, and managed with ease using Metaflow, a Python based framework that merges machine learning, deep learning, and Big Data. Netflix created Metaflow as an open source MLOps platform that makes it simple to handle large scale enterprise data science initiatives.

Kedro

github.com/quantumblacklabs/kedro

A Python-written open source MLOps framework, Kedro helps data scientists write code that is easy to test, reproduce and maintain. Versioning and modularity are used in machine learning projects as part of the software engineering process. It provides pipeline visualisation, project templates, and the ability to distribute data science projects in a variety of ways.

Seldon Core

seldon.io/solutions/open-source-projects/core

Models can be converted into microservices using Seldon's open source MLOps framework that provides a wide range of logging features and enhanced monitoring. Some of Seldon's high-level capabilities make it easy to containerise ML models, evaluate their usability and security, and make them auditable by interacting with a number of services.

Flyte

flyte.org

Flyte is another open source MLOps tool for managing and maintaining Kubernetes-native machine learning operations, as well as automating them. It ensures that machine learning model execution is repeatable by recording model changes, versioning it, and containerising the model and its dependencies. Flyte is a Python library that supports complicated machine learning workflows written in Python, Java, and Scala.

ZenML

github.com/zenml-io/zenml

ZenML is an open source MLOps platform that incorporates machine learning tools like Jupyter notebooks to deploy ML models in a consistent and intuitive manner. It is used to build machine learning pipelines that can be replicated to produce ML projects.

MLRun

mlrun.org

MLRun is a freely available MLOps framework that assists in managing your machine learning pipeline from development to production deployment. It incorporates model tracking, automation, high efficiency, management and easy scalability into any ML process.

MLReef

mlreef.com

MLReef is an open source MLOps platform that is both robust and simple to use, offering options for newcomers, experienced users, and enterprises. It's a complete machine learning platform designed to make ML development more dependable and efficient.

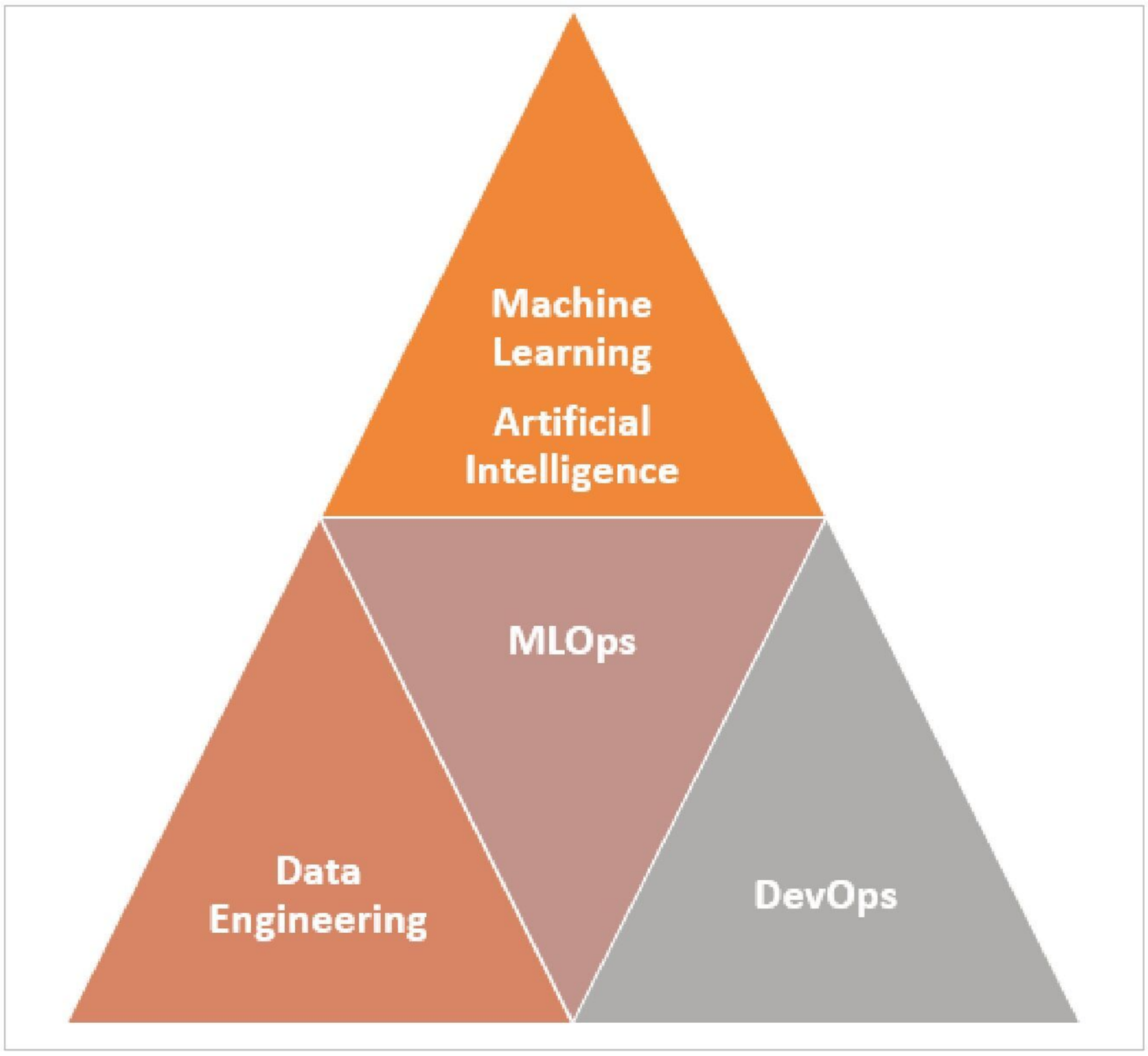


Figure 2: MLOps and its associated perspectives

Bodywork

github.com/bodywork-ml/bodywork-core/

Bodywork is a framework for automating model training and deployment pipelines. It offers automation services for the repetitive operations that most people associate with DevOps. The Bodywork MLOps framework effectively delivers and executes code at the appropriate times. This procedure ensures that the models are well-trained and available at all times. It enables machine learning engineers to use K8s (Kubernetes) containers to deploy model-scoring services. It also aids ML engineers in the delivery of continuous code and deployment automation.

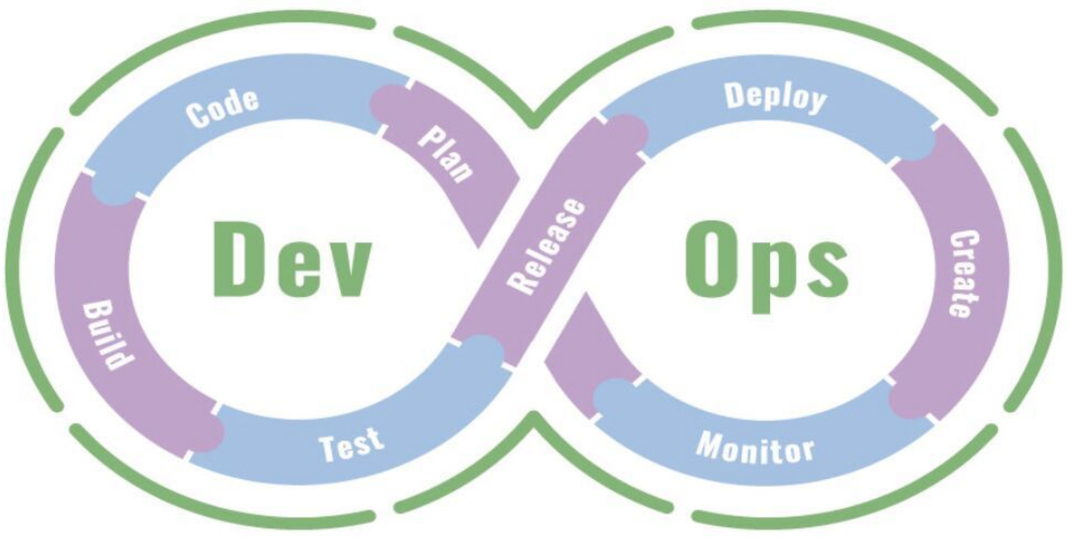
MLOps, with the integration of AI in DevOps, improves the efficiency, reliability, and security of application development and delivery. Businesses and their customers benefit from a more flexible software development life cycle. MLOps is appropriate for teams working collaboratively to design, develop, and deliver secure systems quickly for advanced applications. Through automation, collaboration, rapid feedback, and iterative improvement, MLOps enables AI experts, software engineers and operations teams to speed up the delivery of projects. END

By: Dr Gaurav Kumar

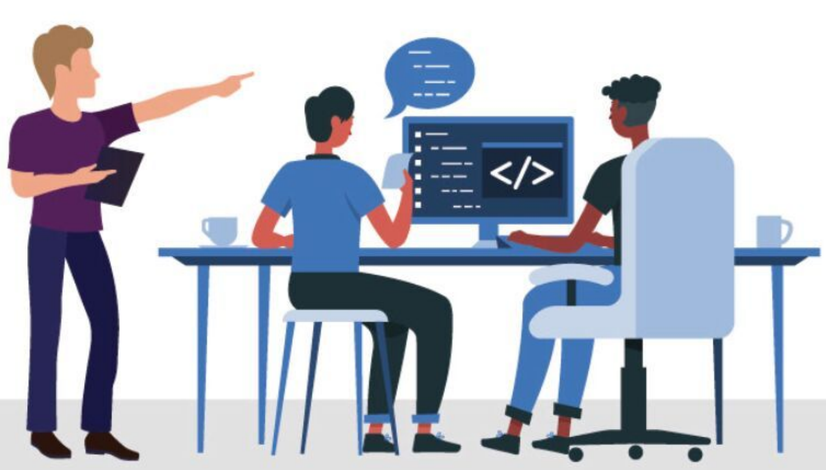
The author is associated with various academic and research institutes, where he delivers expert lectures and conducts technical workshops on the latest technologies and tools.

DevOps and Agile Computing: Enabling a True Cultural Shift

In DevOps, two teams work together to provide better and fast delivery of software. The aim of Agile computing is to provide better collaboration in a small development team to maintain productivity. Let's look at what it takes to implement a true cultural shift in an organisation.







Anyone remotely connected with information technology is aware of the buzzwords like digital transformation, 5G, edge computing, IoT, etc. However, the focus of every technology initiative is the standardisation of software development, delivery and operations. DevOps enables that.

DevOps stands for development and operations, the two key pillars of every organisation — it encompasses all aspects of development and production. At the core are an organisation's philosophies, values, principles, and practices that speed up the development and delivery of applications.

These days Agile computing and DevOps go hand-in-hand, as they have complimenting characteristics that drive the adoption of tools for DevOps and bring about a shift in culture for agility, somewhat like killing two birds with one stone. When the inefficiencies and weaknesses of the organisation are understood and addressed correctly, the software development process improves immensely.

A cultural shift for everyone in the organisation, not just for those responsible for application development and support, is critical in promoting collaboration and transparency. A realisation that silos exist and need to be torn down towards a common goal is critical. In an application life cycle, the roles of non-IT personnel are limited to requesting for a service and using it when it's made available. Development teams develop the software and administrative teams support them.

Figure 1 depicts a traditional organisation versus one that has gone through a cultural shift and has found its true Ikigai or "a reason for being," a Japanese concept that applies well to the adoption of DevOps. A purpose-driven organisation that defines every role and its engagement model

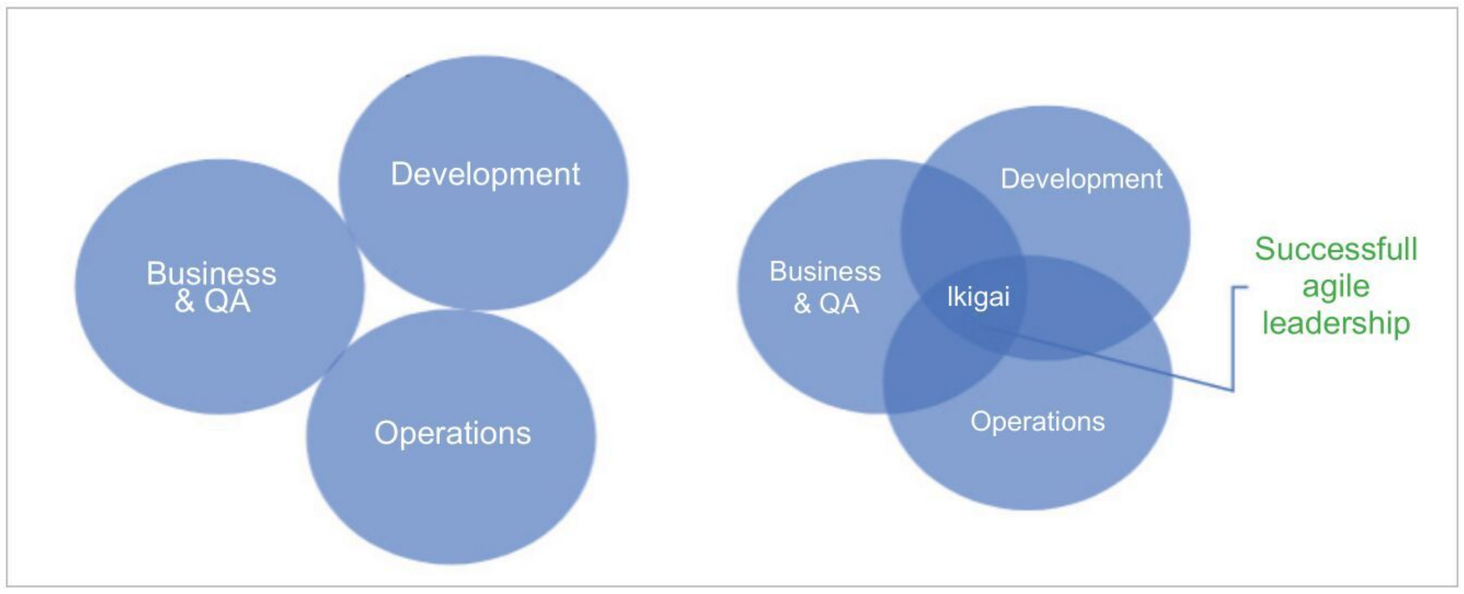


Figure 1: Traditional organisation vs the one that has undergone a cultural shift

is critical for continuous learning and continuous improvement, especially if it empowers teams, enables quick feedback and high empathy, and builds trusted cross-team collaboration.

Successful rollout of a DevOps and Agile culture requires understanding the key roles in a DevOps team and the interdisciplinary best practices to ease the transitions. A sharp focus on user experience remains critical to application and product development. Understanding current ways of working versus the DevOps adoption plan will help to understand the gaps, and develop a successful roadmap for the organisation.

DevOps methodology

The adoption of DevOps is rapidly expanding across the IT industry, with digital transformation pushing the need for speed and businesses focusing on quick product delivery. A DevOps strategy requires a true cultural shift in an organisation. In contrast to how long application and product development have been around, DevOps methodology is relatively new but evolving at breakneck speed. DevOps means different things to different people — there is no set defined concept for it; rather, it's a grouping of concepts that requires an understanding of how to use the operational benefits the methodology can offer successfully.

A focus on continuous integration of development work and continuous deployment of applications (commonly known as CICD) with an eye on automation in provisioning, testing, deployment, etc, and transparent communication will ensure enhanced teamwork. Typical challenges of DevOps adoption are the existing culture, legacy integration, application complexity, and skillsets. Many of these concerns can be addressed with the management mobilising culture shift initiatives, along with an adequate budget and strategy for integrated tools for DevOps.

Tools and terminology

DevSecOps, AIOps, AIMLOps and automation are the buzzwords in the DevOps terminology, but in the end it is all about the user experience. This is overlooked many times because of the various moving parts in the DevOps environment. It is important to remind teams of the importance of the user experience to bring the end goal back into focus. Setting up a knowledge repository that includes the meaning of DevOps, success criteria, methodology, and road map will help an organisation transition to the DevOps culture successfully.

What most organisations have today is a set of tools and procedures, and individuals who make up a DevOps team. This team is only skilled and responsible for putting the tools to good use at best. However, it's the job of everyone in the organisation to embrace the DevOps culture. What is needed is proper planning, and leveraging of automation tools across development and deployment environments. An integrated environment that subjects developed code to automated testing, packages it for deployment, and publishes the code to the production environment removes the frictions typically faced in software adoption. Selecting the appropriate tools that fit an organisation's maturity level and objectives is critical for embracing the DevOps culture towards agility in the development of products and applications.

Creating an Agile culture is a journey that requires more than ticking checklists while providing tools and administrative processes. Just as the overall company culture and values are the drivers of employee satisfaction, an Agile culture can be realised with education and acceptance of the shift in an organisation's use of technology by everyone. Bridging the divide between development, operations, QA and the business is a process and takes time. Understanding how DevOps methodology can bridge this gap means a new way forward towards Agile-IT! END



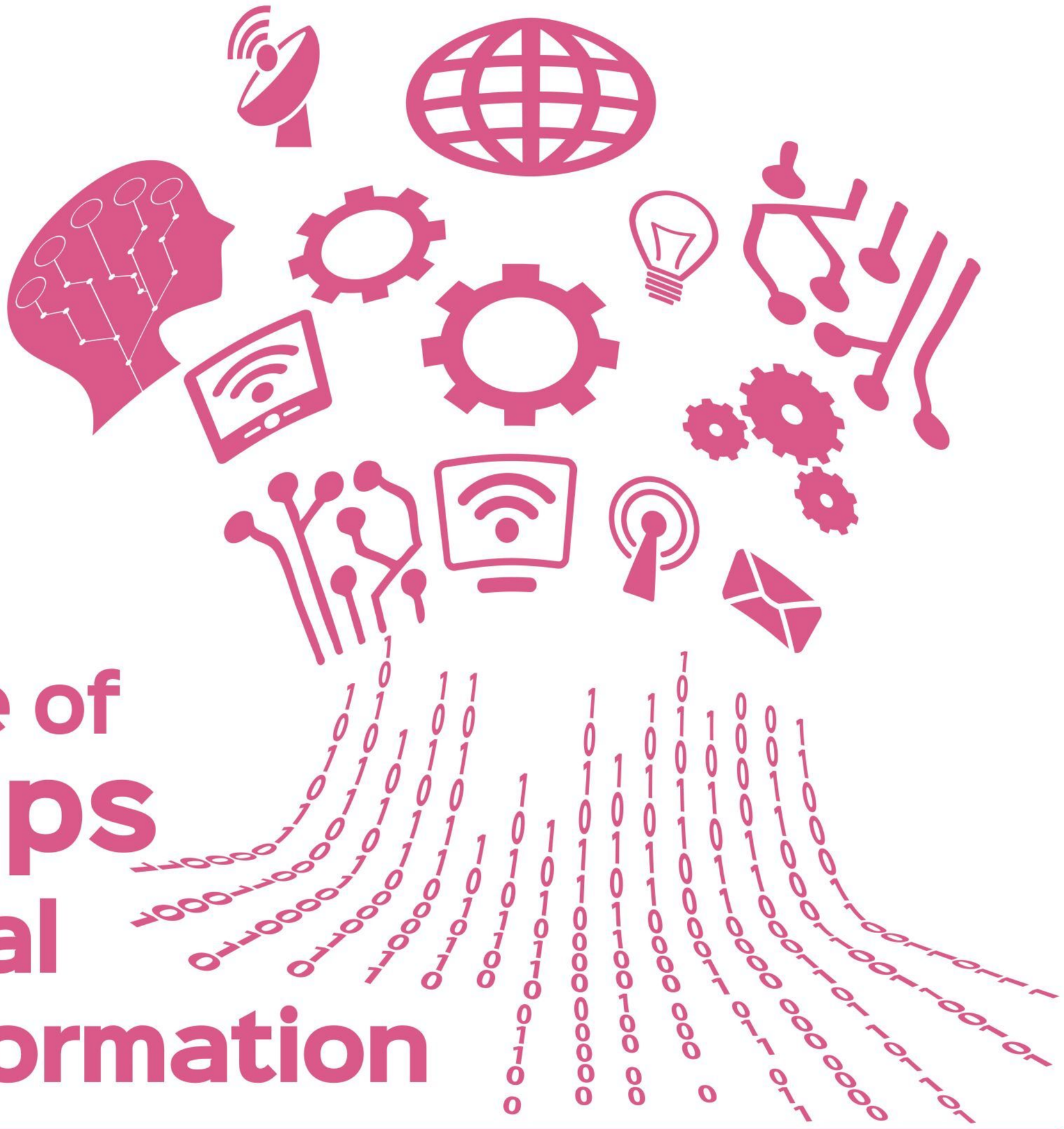
Reference

https://DevOps.com

By: Bala Kalavala

The author is a technical architect, evangelist, thought leader and a sought after keynote speaker. He currently works as a distinguished member of the technical staff and heads the enterprise architecture practice as chief architect in Wipro Ltd.

Disclaimer: *This article expresses the view of the author and not that of Wipro.*



The Role of DevOps in Digital Transformation

DevOps is an end-to-end approach that caters to all aspects of software development and operations. A cultural paradigm in the organisation, it brings development and operations together to achieve the dual objectives of stability and change. The crux of this approach is continuity and automation. This includes continuous integration, continuous deployment and continuous delivery. This article gives an overview of its role in digital transformation.

Before adopting DevOps, the different groups in a company — such as business, IT and operations — worked in silos. In the Waterfall approach to software development, each phase was separated by stage-gates. An example would be the completion of development leading to the starting of the testing phase. In this approach, there were unexpected surprises during the transitioning process. Also, each group had its own tools and systems, leading to a plethora of applications. Virtual walls existed between the different parts of the organisation and there were no proper communication lines, which led to cost and effort overruns.

DevOps tooling for digital transformation

Organisations have a large ecosystem of systems and applications. It's extremely important that the application technology portfolio meets the expectations of the business. Organisations have the option of building their own internal technology stack or leveraging any cloud provider.

Ever since the rapid adoption of cloud, cloud providers like AWS, Azure and GCP have started providing a number of DevOps tools and services. Organisations can choose a provider or build their own tools/accelerators leveraging the capabilities provided by the cloud provider. This brings an enormous amount of flexibility and provides a great opportunity for transforming the business.

DevOps framework for enterprise adoption

An enterprise DevOps framework is a pluggable shared services model that is completely automated. It scales depending on the dynamic needs of the organisation, leading to enterprise-wide adoption. The framework integrates seamlessly with the current toolset of the organisation.

Enterprise DevOps integrates a business with its IT processes, and includes:

- Business planning
- Feature, user-story and task creation
- Agile development
- Version management
- Unit testing (test driven development)

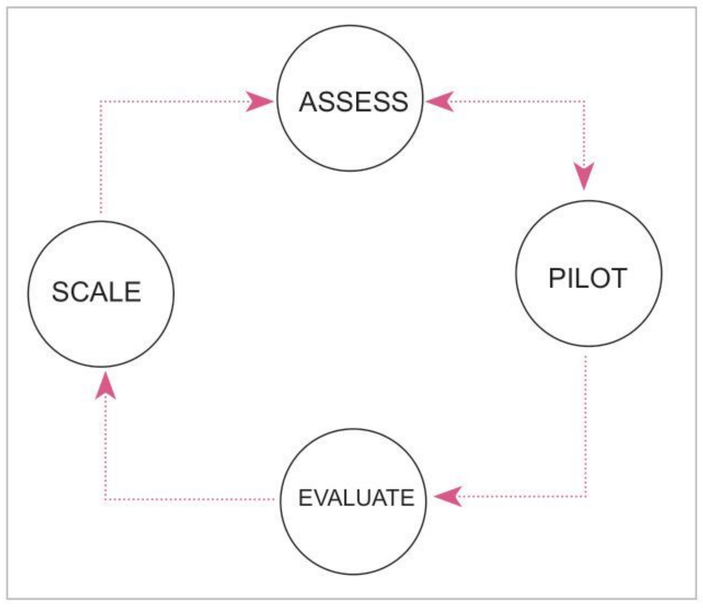


Figure 1: Enterprise DevOps framework

- Continuous integration
- Continuous build
- Test automation
- Security
- Performance testing
- Configuration management
- Release management
- Continuous deployment
- Continuous delivery
- Monitoring

DevOps adoption has a number of benefits like improved customer satisfaction, reduced total cost of ownership, reduced response times, and higher operational efficiency. Predictive techniques that use AIOps can help prevent DevOps issues, and identify patterns and root causes for greater organisational productivity. END

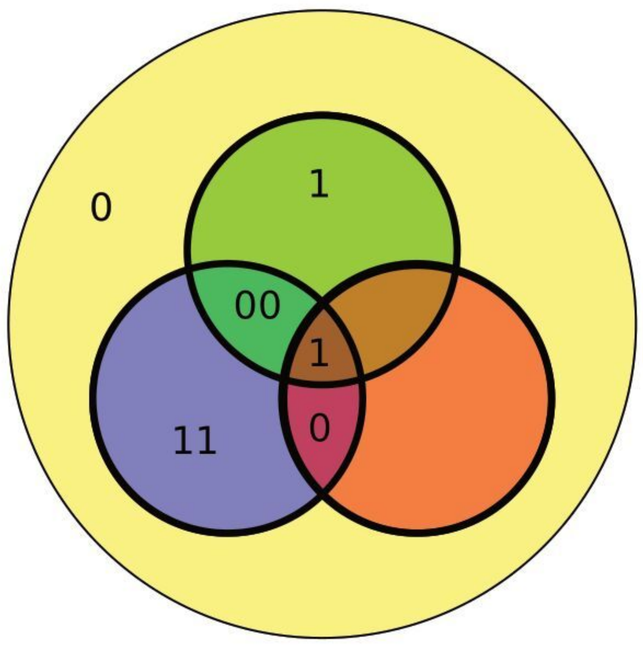
By: Phani Kiran

The author has over 17 years of experience in the IT industry and has worked on Big Data technologies.



Hamming Codes: The Error-Correcting Codes

When bits are transmitted or sent over the computer network, they can easily get corrupted due to interference and network problems, and lead to errors. Hamming codes are linear codes used to detect and correct the errors that can occur when data is moved from the sender to the receiver. Let's take a look at their implementation in Python.



In 1950, R.W. Hamming created an error correction code called the Hamming code. It divides data into fixed length blocks, making it a block based error detection and correction algorithm. Extra or redundant parity bits are generated and embedded into the data message before transmission. After the receiver gets them, these parity bits are recalculated to detect and correct the error.

Up to 2 bit errors can be detected and up to 1 bit errors can be corrected. For example, in case of *even parity*, the number of 1s in specific positions should turn out to be even. This will ensure the integrity or correctness of the message. The procedure can be described in the following steps:

- Calculation of the number of redundant bits
- Finding the position of parity/redundant bits
- Filling the parity bits
- Sending the message

On the receiver side, the steps are:

- Receiving the message
- Finding the position of parity/redundant bits
- Comparing the parity bits to find the error

Algorithm overview

We are going to use *even parity* in this example.

The user has to send a data stream of bits from one endpoint to another. The number of bits in the data stream is represented as 'm'. And the hamming encoding algorithm is going to insert 'r' number of extra parity bits in the message for encoding before sending on the wire.

If 'm' is the number of data bits and 'r' is number of redundant bits, the following condition must be satisfied:

```
2^r ≥ m + r + 1    ....eq1
```

The parity bits will be inserted in specific locations of the encoded data stream. These locations are powers of 2; e.g. (2^0=1, 2^1=2, 2^2=4, 2^3=8, etc).

Let us take an example. Say we have a data stream of 4 bits; so, *m=4*. This is represented as follows:

D1	D2	D3	D4

We now need to calculate the number of parity bits:

```
If r=1 we put in eq1  2^1 < 4+1+1, so not ok
If r=2 we put in eq1  2^2 < 4+2+1, so not ok
If r=3 we put in eq1  2^3 = 4+3+1, this is ok
```

So the number of redundant bits is *r=3*. After the insertion of parity bits, the data stream will look like this:

P1	P2	D1	P3	D2	D3	D4

These parity bits are calculated from XOR of data bits at specific locations. The specific data bit positions are found based on the following table (where the corresponding bit is 1).

bit location	P3	P2	P1
1	0	0	1
2	0	1	0
3	0	1	1
4	1	0	0
5	1	0	1
6	1	1	0
7	1	1	1

So,

P1=B1 **XOR** B3 **XOR** B5 **XOR** B7

P2=B2 **XOR** B3 **XOR** B6 **XOR** B7

P3=B4 **XOR** B5 **XOR** B6 **XOR** B7

—--eq2

This is referred to as (7,4) Hamming code.

Implementation overview (with code walkthrough)

I have written a Python code to implement the algorithm. Here, *even parity* is used.

Data is the input data stream taken as input from the user (example: data='1101').

'm' is the length of data taken as input from user (m=4).

Encoding of data

Step 1: *calculateRedundantBitSize()* function finds the value of 'r'**;** r=3.

Step 2: In *createEncodedDat()*, the encodedList is the list initialised with all 0s.

For (7,4), the Hamming code encodedList will look like this:

0	0	0	0	0	0	0

The next code copies data bits to specific locations:

```
for i in range(1,m+r+1,1):
        if(i==(2**j)):
            j=j+1
        else:
            encodedList[i-1]=data[k]
            k=k+1
```

The *for* loop is iterating over the complete data length of 'm+r'. If 'i' is power of 2 in the encodedList, the i-1 index will hold the data bit. For (7,4), the Hamming code encodedList will look like this:

0	0	1	0	1	0	1

Then the parity bits are calculated as follows:

```
for j in range(0,r,1):
        parityBitLoc = 2**j
        for i in range(1,m+r+1,1):
           if((parityBitLoc & i) != 0):
               p=parityBitLoc-1
               encodedList[p]= str(int(encodedList[i-1]) ^
int(encodedList[p]))
```

'j' is iterating over parity bits and 'i' is iterating over the full data stream. The condition:

```
if((parityBitLoc & i) != 0)
```

…will verify the specific positions for parity bit calculation, as specified in equation 2. 'p' is indicating the position of the parity bit in the list.

For (7,4) Hamming code, the encodedList will look like this:

1	0	1	0	1	0	1

Finally, the encoded bits are stored as a string in encodedData.

Decoding and error correction

We now look at error correction. For error simulation, users can give an input to intentionally alter a bit. *rcvdList* is an erroneous bitString, which is passed on to the decodedData function.

As an example, we assume that bit 4 is corrupted. *rcvdList* will look like this:

1	0	1	1	1	0	1

In the decodedData function, execute the following code:

```
chkList[p] = int(rcvdList[i-1]) ^ int(rcvdList[p]) ^
int(chkList[p])
```

After execution of this code, *chkList* will look like this:

P1	P2		P3			
0	0	0	1	0	0	0

We convert it to a binary:

$P3P2P1 = (100)_b$

…which is 4 in decimals. This matches exactly with the user input. Thus it can also detect and correct the error. You can find the full source code on GitHub at *https://github.com/SupriyoGanguly/python-hamming.*

Hamming code is a cost-effective solution for detecting and correcting single-bit errors in computers. But if multiple bits are erroneous, it can totally spoil the data. Satellites and modem communications use *hamming* code. The simple Python code implementation discussed in this article can be used as a module by programmers for *hamming* code encoding and decoding for any standard data stream length of 4 bits, 7 bits, 13 bits, and so on. END

By: Supriyo Ganguly

The author is an M. Tech in embedded systems, and works as a senior technical officer at the Electronics Corporation of India Limited, Hyderabad.

Generating APIs Using Django Rest Framework with Insomnia

This article gives the fundamentals of generating APIs using Django Rest Framework with Insomnia. It's aimed at absolute beginners.



This article will familiarise you with the basic structural layout and features that come with creating a Django project for API generation. It will help you understand what models, serialisers and views are, how they are interconnected, and how we can use Insomnia to test the functioning of CRUD (create, read, update and delete) operations on our APIs.

Initial setup

You will be required to use the command prompt or carry out the initial steps needed to create a Django project. Check the version of Python you have installed on your system:

```
python --version
```

Pip can be defined as an installation manager for packages/libraries that are written in Python. Let's download a file called *get-pip.py:*

```
curl https://bootstrap.pypa.io/get-pip.py -o get-pip.py
```

To install pip, give the following command:

```
python get-pip.py
```

Note: Figure 1 shows pip as being uninstalled first and then reinstalled, because I already had pip installed on my system.

Now let us understand what 'virtual environments' are before we create a Django project. To state bluntly, virtual environments are environments that can have distinct versions of packages/libraries installed within them, respectively. Let me explain that with an example. We are in April 2022, and we intend to create a project (to be dealt with in VS Code). Let's say the latest version of Python is 3.1 (which we have installed in our system and no virtual environments have been created) having libraries a, b, c, d, e and f. Assume that our project requires libraries a, b and c, all of which are in Python 3.1; hence our project will work without any hindrances.

Now let's say Python's new version 3.2 comes out a few weeks or a few months later which has libraries a, d, e, f, g and h. We install this new

```
Select Command Prompt
Microsoft Windows [Version 10.0.19043.1586]
(c) Microsoft Corporation. All rights reserved.

C:\Users\kamil>python --version
Python 3.10.2

C:\Users\kamil>curl https://bootstrap.pypa.io/get-pip.py -o get-pip.py
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 2596k  100 2596k    0     0  1312k      0  0:00:01  0:00:01 --:--:-- 1314k

C:\Users\kamil>python get-pip.py
Collecting pip
  Using cached pip-22.0.4-py3-none-any.whl (2.1 MB)
Installing collected packages: pip
  Attempting uninstall: pip
    Found existing installation: pip 22.0.4
    Uninstalling pip-22.0.4:
      Successfully uninstalled pip-22.0.4
Successfully installed pip-22.0.4

C:\Users\kamil>
```

Figure 1: The command prompt

version on our system (having created no virtual environments). Our project will not function properly now since it required libraries b and c to do so, which are now not present in 3.2. This is where virtual environments come in. Assume we create a virtual environment ‘venv1’. Within this, we’ll install Python 3.1 and create our project. In another virtual environment ‘venv2’, we can install Python 3.2. Note that the project we created in ‘venv1’ will not work in ‘venv2’ since ‘venv2’ has Python 3.2 installed within it, which does not encompass all the libraries that the project requires. The purpose of this elaborate explanation is to give you an idea of how virtual environments work.

To install a virtual environment wrapper (does not create a virtual environment immediately), type:

```
pip install virtualenvwrapper-win
```

To create a virtual environment ‘venv1’, give the following command:

```
mkvirtualenv venv1
```

Note: The ‘(venv1)’ observed towards the left of the latest command line indicates that we are inside a virtual environment.

Now, install Django:

```
pip install Django
```

Next, create a new directory, access it and then create a Django project within the directory:

```
Syntax for creating a directory: mkdir
(name of directory)
Syntax for accessing a created
directory: cd (name of directory)
Syntax for creating a django project:
django-admin startproject osfyproject1

(venv1) C:\Users\kamil>mkdir osfydir
(venv1) C:\Users\kamil>cd osfydir
(venv1) C:\Users\kamil\osfydir>django-
admin startproject osfyproject1
```

Now that the Django project has been created, let us run the server within our newly created project. The command prompt is:

```
(venv1) C:\Users\kamil\osfydir>cd
osfyproject1
(venv1) C:\Users\kamil\osfydir\
osfyproject1>python manage.py runserver

Watching for file changes with
StatReloader
Performing system checks...

System check identified no issues (0
silenced).

You have 18 unapplied migration(s).
Your project may not work properly
until you apply the migrations for
app(s): admin, auth, contenttypes,
sessions.
Run ‘python manage.py migrate’ to apply
them.
April 12, 2022 - 16:18:22
Django version 4.0.3, using settings
```

```
Command Prompt

C:\Users\kamil>pip install virtualenvwrapper-win
Requirement already satisfied: virtualenvwrapper-win in c:\users\kamil\appdata\local\programs\python\python310\lib\site-packages (1.2.7)
Requirement already satisfied: virtualenv in c:\users\kamil\appdata\local\programs\python\python310\lib\site-packages (from virtualenvwrapper-win) (20.13.0)
Requirement already satisfied: filelock<4,>=3.2 in c:\users\kamil\appdata\local\programs\python\python310\lib\site-packages (from virtualenv->virtualenvwrapper-win) (3.4.2)
Requirement already satisfied: six<2,>=1.9.0 in c:\users\kamil\appdata\local\programs\python\python310\lib\site-packages (from virtualenv->virtualenvwrapper-win) (1.16.0)
Requirement already satisfied: distlib<1,>=0.3.1 in c:\users\kamil\appdata\local\programs\python\python310\lib\site-packages (from virtualenv->virtualenvwrapper-win) (0.3.4)
Requirement already satisfied: platformdirs<3,>=2 in c:\users\kamil\appdata\local\programs\python\python310\lib\site-packages (from virtualenv->virtualenvwrapper-win) (2.4.1)

C:\Users\kamil>mkvirtualenv venv1
created virtual environment CPython3.10.2.final.0-64 in 1157ms
  creator CPython3Windows(dest=C:\Users\kamil\Envs\venv1, clear=False, no_vcs_ignore=False, global=False)
  seeder FromAppData(download=False, pip=bundle, setuptools=bundle, wheel=bundle, via=copy, app_data_dir=C:\Users\kamil\AppData\Local\pypa\virtualenv)
    added seed packages: pip==21.3.1, setuptools==60.5.0, wheel==0.37.1
  activators BashActivator,BatchActivator,FishActivator,NushellActivator,PowerShellActivator,PythonActivator

(venv1) C:\Users\kamil>
```

Figure 2: Command prompt involving creation of virtual environment ‘venv1’

```
(venv1) C:\Users\kamil>pip install django
Collecting django
  Downloading Django-4.0.3-py3-none-any.whl (8.0 MB)
     |████████████████████████████████| 8.0 MB 3.3 MB/s
Collecting tzdata
  Downloading tzdata-2022.1-py2.py3-none-any.whl (339 kB)
     |████████████████████████████████| 339 kB 2.2 MB/s
Collecting asgiref<4,>=3.4.1
  Downloading asgiref-3.5.0-py3-none-any.whl (22 kB)
Collecting sqlparse>=0.2.2
  Using cached sqlparse-0.4.2-py3-none-any.whl (42 kB)
Installing collected packages: tzdata, sqlparse, asgiref, django
Successfully installed asgiref-3.5.0 django-4.0.3 sqlparse-0.4.2 tzdata-2022.1
```

Figure 3: Installation of Django

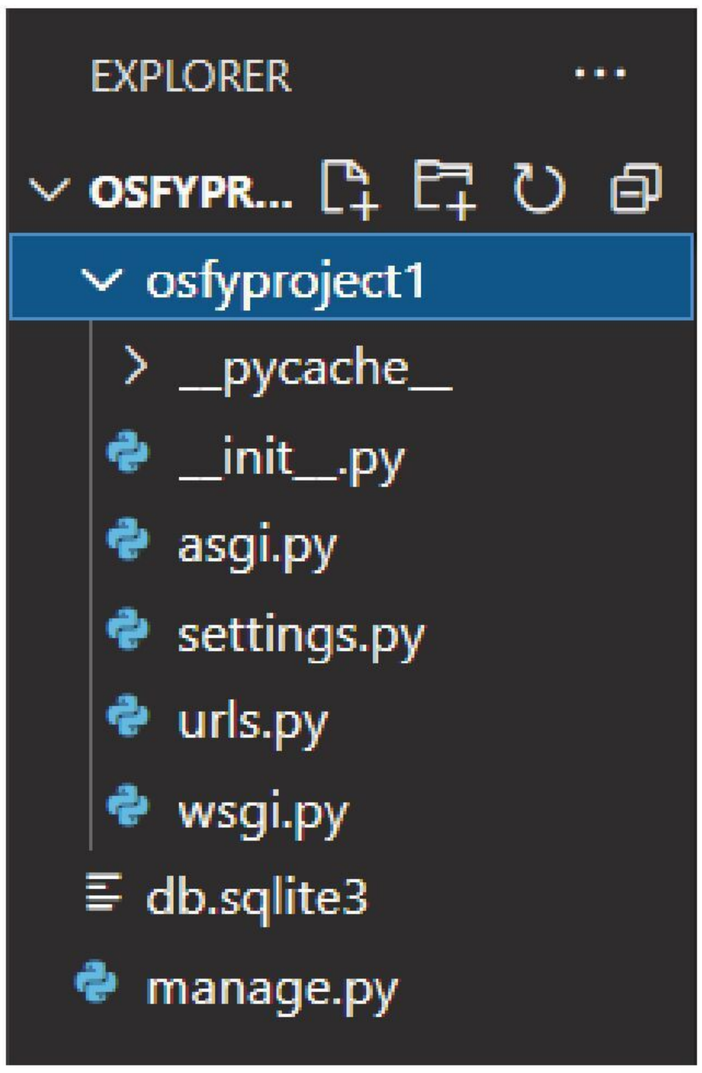


Figure 4: Contents of the project in VS Code

```
'osfyproject1.settings'
Starting development server at
http://127.0.0.1:8000/
Quit the server with CTRL-BREAK.
```

The *http://127.0.0.1:8000/* observed in the second last line of the above code is the base URL of all the operations we wish to perform. We will observe how we generate separate URLs as an extension of this for carrying out CRUD operations, in later sections.

Now let's head to VS Code and open our project by selecting it within the directory, the file path of which is *C:\Users\kamil\osfydir.*

Let us activate our virtual environment within the terminal of VS Code:

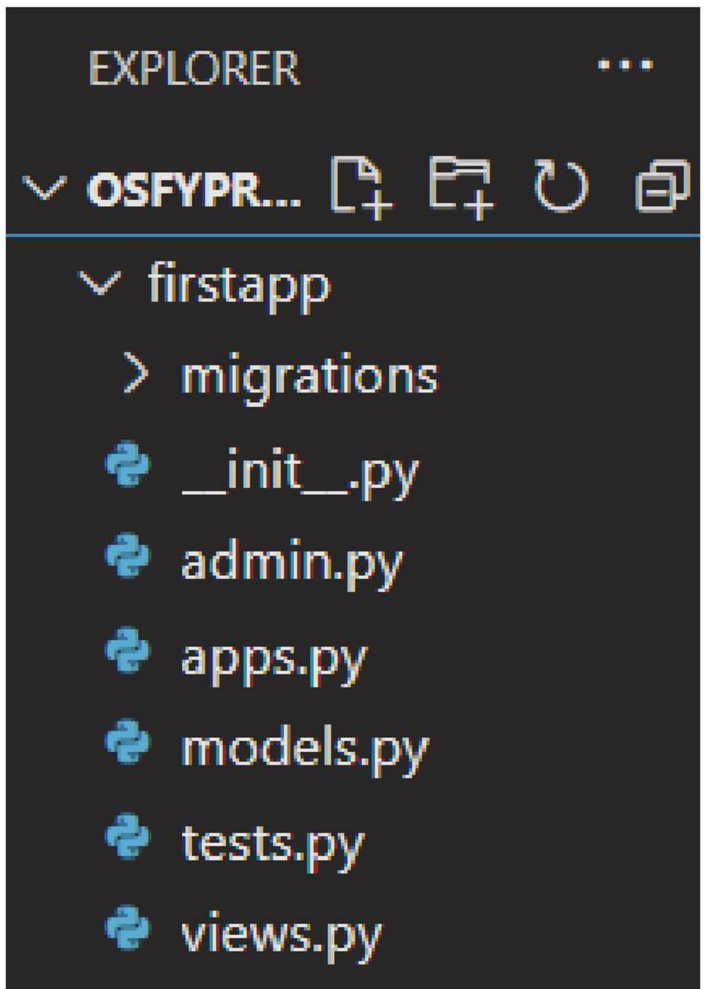


Figure 5: Updated file explorer in VS Code

```
Microsoft Windows [Version
10.0.19043.1586]
(c) Microsoft Corporation. All rights
reserved.

C:\Users\kamil\osfydir\
osfyproject1>workon venv1
(venv1) C:\Users\kamil\osfydir\
osfyproject1>
```

We now create an app within our project. Let's say that your 'project' is a website on the whole and the 'apps' are specific Web pages within the website. Since you wish to write the logics for APIs of different Web pages, you will want to write them in their respective *views.py* files. You will understand as we go further.

```
(venv1) C:\Users\kamil\osfydir\
osfyproject1>python manage.py startapp
firstapp

(venv1) C:\Users\kamil\osfydir\
osfyproject1>
```

Instead of getting an overview of the various purposes the different files within the app and the project serve, which might confuse you in the beginning, let's go about it step-by-step as we move along the path of obtaining the end goal of our task. First, let's create a table in our database with different kinds of fields. But hang on, we first need to establish a connection with the database in order to create a table within it. We will be using pgAdmin (PostgreSQL) for our task.

We first need to establish a connection between our project and the *osfyproject1* database. We need *psycopg2* to do so. Let's install *psycopg2* by entering a command within the prompt in VS Code. The following section of code shows the output of the installation operation.

```
(venv1) C:\Users\kamil\osfydir\
osfyproject1>pip install psycopg2
Collecting psycopg2
 Using cached psycopg2-2.9.3-cp310-
cp310-win_amd64.whl (1.2 MB)
Installing collected packages: psycopg2
Successfully installed psycopg2-2.9.3
WARNING: You are using pip version
```

```
21.3.1; however, version 22.0.4 is
available.
You should consider upgrading via the
'C:\Users\kamil\Envs\venv1\Scripts\
python.exe -m pip install --upgrade
pip' command.
```

Let us now create a table in our database. These 'tables' are referred to, as per the terminology of a Django project, as 'models', and are defined within the 'models.py' file in the app. Within 'models.py' of *firstapp*, let's write the following:

```
from django.db import models

# Create your models here.

class Studentinfo(models.Model):
    First_name=models.CharField(max_len
gth=50,blank=False,null=False)
    Last_name=models.CharField(max_leng
th=50,blank=False,null=False)
    Student_ID=models.IntegerField(blan
k=False,null=False)
    Email_ID=models.EmailField(blank=Fa
lse,null=False)
```

Our idea is to create a simple table/model taking in the information of a student. The fields 'First_name' and 'Last_name' are character fields, which allow a maximum length of 50 characters. The 'Required=True' parameter implies that you cannot discard this field (this will be covered in a case scenario later). The 'null=False' means that these fields will not accept null values. There are many different types of fields and parameters for Django models, which you can find in the official documentation of Django.

We now need to 'send' our defined model/table to our database. For this, let's make a few changes to the 'DATABASES' section within the *settings.py* file of 'osfyproject1' in VS Code. The pre-existing section looks as follows:

```
DATABASES = {
```

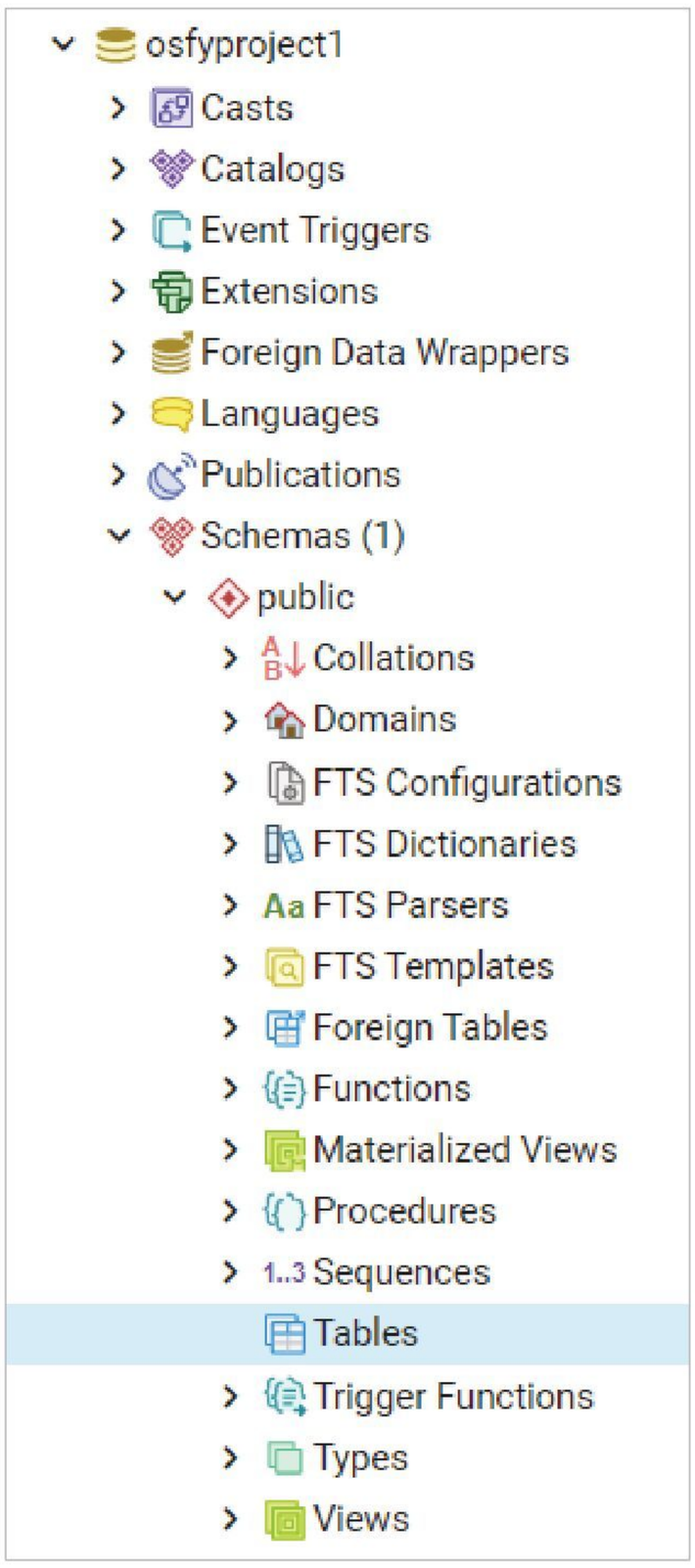


Figure 6: Newly created database in pgAdmin

```
    'default': {
        'ENGINE': 'django.db.backends.
sqlite3',
        'NAME': BASE_DIR / 'db.
sqlite3',
    }
}
```

By default, it considers sqlite3 as its database platform, but in our case, we are using PostgreSQL. The modified section should appear like this:

```
  DATABASES = {
    'default': {
        'ENGINE': 'django.db.backends.
postgresql',
        'NAME': 'osfyproject1',
        'USER':'postgres',
        'PASSWORD':'password',
        'HOST':'localhost',
    }
}
```

The 'PASSWORD' will be the password you set when you installed PostgreSQL on your system (mine was simply 'password' itself).

Another section within 'settings.py' which requires an addition is the 'INSTALLED_APPS' section.

```
INSTALLED_APPS = [
    'firstapp',
    'django.contrib.admin',
    'django.contrib.auth',
    'django.contrib.contenttypes',
    'django.contrib.sessions',
    'django.contrib.messages',
    'django.contrib.staticfiles',
]
```

Kindly add 'firstapp' in the above list.

Now, we must send our table to the database, which in this terminology is 'migrating' our models. Migrating our defined model to our PostgreSQL database involves two simple steps:

```
(venv1) C:\Users\kamil\osfydir\
osfyproject1>python manage.py
makemigrations
Migrations for 'firstapp':
  firstapp\migrations\0001_initial.py
    - Create model Studentinfo
```

The 'makemigrations' step automatically creates a file called '0001_initial.py' in the *migrations* folder within our 'firstapp'. This file has code written in Python, which inculcates the definition of our model in a way that PostgreSQL will understand. Let's take a peek at how this looks:

```
from django.db import migrations,
models

class Migration(migrations.Migration):
    initial = True
    dependencies = [
```

	id [PK] bigint	First_name character varying (50)	Last_name character varying (50)	Student_ID integer	Email_ID character varying (254)

Figure 7: Glimpse of our newly-created model

```
    ]
    operations = [
        migrations.CreateModel(
name='Studentinfo',
 fields=[
                ('id', models.
BigAutoField(auto_created=True,
primary_key=True, serialize=False,
verbose_name='ID')),
                ('First_name', models.
CharField(max_length=50)),
                ('Last_name', models.
CharField(max_length=50)),
                ('Student_ID', models.
IntegerField()),
                ('Email_ID', models.
EmailField(max_length=254)),
            ],
        ),
    ]
```

For the next step, we simply use the keyword 'migrate', which results in a table called 'Studentinfo' being created in our database:

```
(venv1) C:\Users\kamil\osfydir\
osfyproject1>python manage.py migrate
```

The output of the above command shows around 15 items being checked as 'OK'. If you take a look at the 'Tables' tab in the 'osfyproject1' database, you will find a 'firstapp_studentinfo' table, which is shown in Figure 7.

The 'id' field has been assigned as the primary key of the table by default. You can set another field as the primary key by including 'primary_key=True' as one of the parameters of the field within the model definition in 'models.py'. Now let's jump into performing CRUD operations for this model.

Create a new file within 'firstapp' called 'serializers.py' and write the following:

```
from rest_framework import serializers
from .models import Studentinfo

class Studentserializer(serializers.
ModelSerializer):
    class Meta:
        model=Studentinfo
        fields='__all__'
```

A serializer is defined on the Web as follows: "Serializers allow complex data such as query sets and model instances to be converted to native Python datatypes that can then be easily rendered into JSON." Simply put, it converts the object(s) entered into JSON format.

Let's take an example. Say, we have a simple logic for entering new data into the database in our 'views.py', which does not include checks for any errors. Now, if we were to make an entry into our table in the format below:

```
# create a student object
>>> student = Studentinfo(First_
name='Kamil', Last_name='Khan',
Student_ID=085, Email_ID='kamil123@
gmail.com')

# serialize the data
>>> serializer =
Studentserializer(student)

# print serialized data
>>> serializer.data
```

…the output of 'serializer.data' would have been as follows:

```
{'First_name': 'Kamil', 'Last_name':
'Khan', 'Student_ID': 085, 'Email_ID':
'kamilpaigah@gmail.com'}
```

The aim of covering the above code was to provide a practical explanation of the definition found on the Web, but since we will be using Insomnia here onwards, we will not be required to enter data in the format we just observed. We will be entering data in JSON format itself. In simple terms, a serializer creates a serializer class, which maps the fields of the model it is associated with. You may ask, if that's the case, what's the difference between a model and a serializer? From what I have observed, a serializer has the capability to check the validity of the data entered, such as checking if the value entered for field(s) is of the type defined in the model; if not, it will show errors. Now let's move on to *views.py*, wherein we will cover the logic for 'Create' and 'Read'.

```
from firstapp.models import Studentinfo
from .serializers import
Studentserializer
from rest_framework.response import
Response
from rest_framework import status
# Create your views here.
     ---(5)

class Student_post_get_view (APIView):
    def post(self,request):
        serializer=Studentserializer(da
ta=request.data)
        if serializer.is_valid():
            ---(10)
            serializer.save()
            return Response(serializer.
```

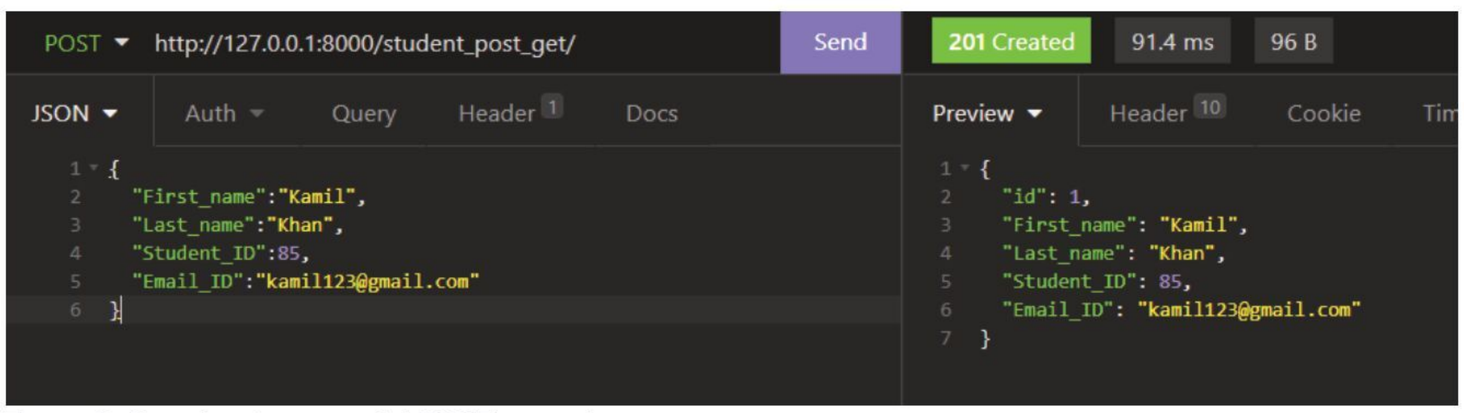


Figure 8: Sample of successful POST operation

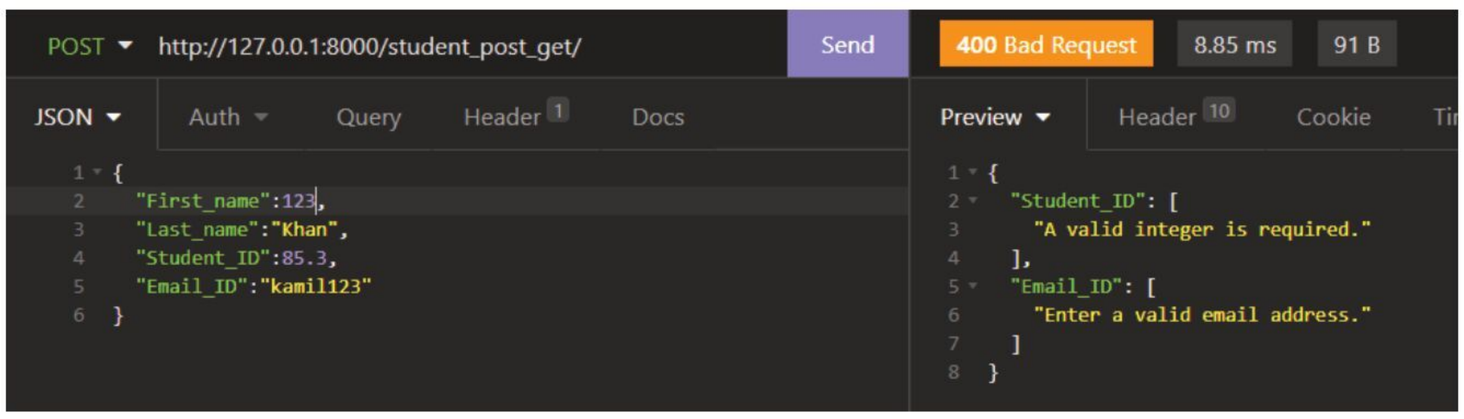


Figure 9: Sample error messages

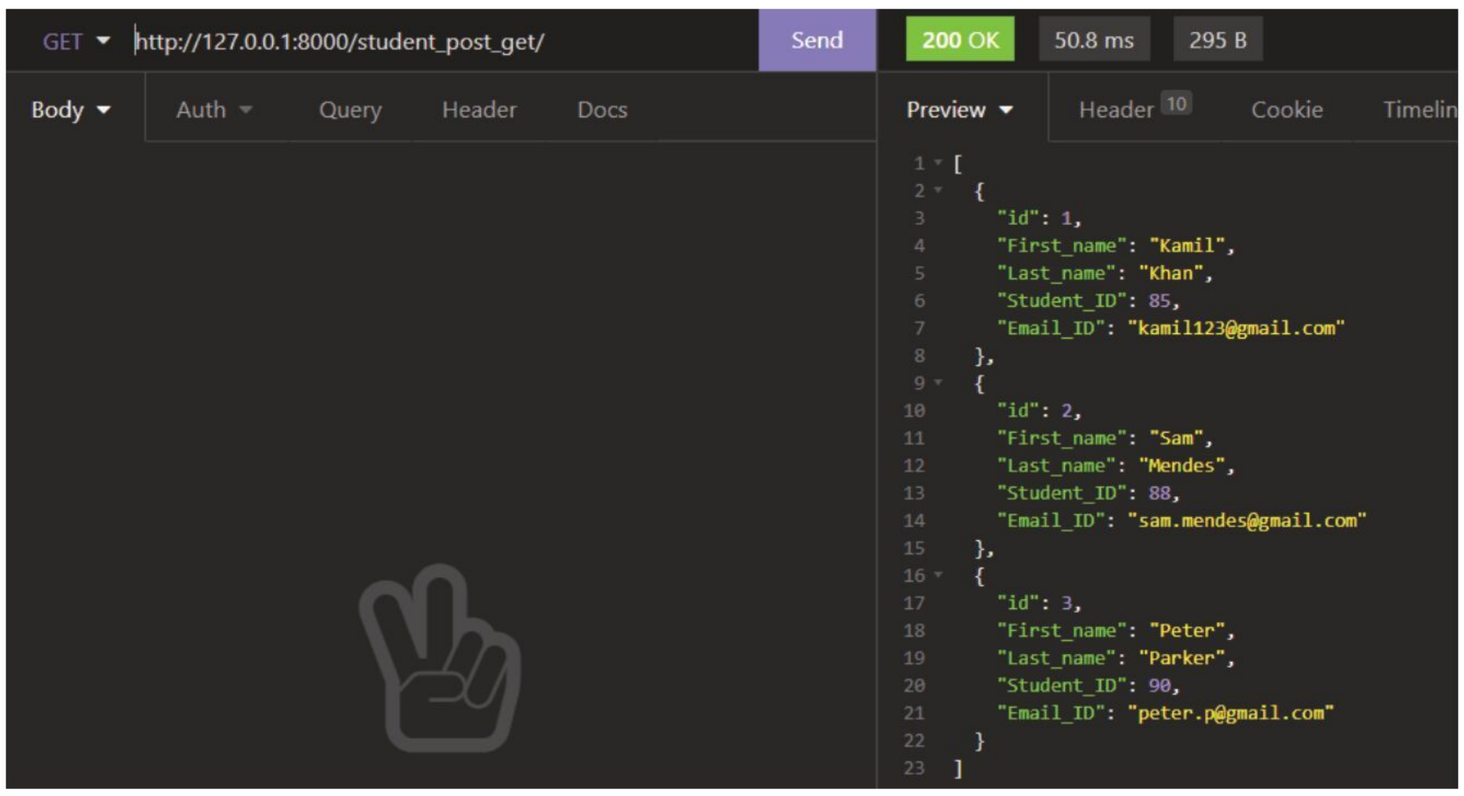


Figure 10: Displaying all information using GET

```
data,status=status.HTTP_201_CREATED)
        return Response(serializer.
errors,status=status.HTTP_400_BAD_
REQUEST)

    def get(self,request):
    ---(15)
        getting_data=Studentinfo.
objects.all()
        serializer=Studentserializer(ge
tting_data,many=True)
        return Response(serializer.
data)
```

The function ‘post’ corresponds to ‘Create’ operation, while the ‘get’ function aligns with the ‘Read’ operation. In line (9), the ‘request.data’ is a representation of the data we ‘POST’ in JSON format. ‘request.data’ being assigned to the ‘data’ parameter within the Studentserializer implies that the entered data is passed to the serializer. Line (10) checks the validation of the data entered for the constituting fields within the model that the ‘Studentserializer’ is associated with, i.e., Studentinfo. Line (12) displays the entered data itself, indicating that there are no validation errors, post the saving mentioned in line (11). Line (13) helps return the errors (if any). As per line (16), all the objects present within the ‘Studentinfo’ model (the existing data) are assigned to ‘getting_data’ variable. These objects are then passed through the serializer to be presented in JSON format as per line (18). Just one step remains before we can initiate inserting values for our table. In ‘urls.py’ within ‘osfyproject1’, insert the following:

```
from django.contrib import admin
from django.urls import path
from firstapp.views import Student_post_
get_view

urlpatterns = [                    ---(5)
    path('admin/', admin.site.urls),
    path('student_post_get/',Student_
post_get_view.as_view()),
]
```

The URL path we entered in line (7) helps us in carrying out the operations present within ‘Student_post_get_view’, i.e., the ‘POST’ and ‘GET’. In the prompt, type:

```
python manage.py runserver
```

You should be able to see the link *http://127.0.0.1:8000/* in the output, implying that the server is running. We can now head to Insomnia to carry out the ‘POST’ and ‘GET’ operations (Figure 8)!

Now let’s try botching these up to see how validation errors are displayed.

As observed in Figure 9, we entered incorrect types of values for ‘First_name’, ‘Student_ID’ and ‘Email_ID’ fields, but why is it that we got error messages for only two of them? This is because a serializer takes the values that you enter for the fields and tries converting them to the field types you defined in the model, and checks for successful conversion. Generally, in Python, a number of any type — be it float, string, etc — can be converted to a string. Since it is possible for the number 123 to be converted to string in Python, no error message is displayed for the same. We will enter valid data for two more students and then perform the ‘GET’

operation to see the information of all three students (Figure 10).

Let's move on to the 'views.py' code for 'PUT' and 'DELETE' operations, which correspond to the 'Update' and 'Delete' operations in CRUD:

```
class StudentDetails(APIView):
    def get_object(self,pk):
        try:
            return Studentinfo.objects.
get(pk=pk)
        except Studentinfo.
DoesNotExist:
            return
HttpResponse(status=status.HTTP_404_
NOT_FOUND)

    def get(self,request,pk):
        article=self.get_object(pk)
        serializer=Studentserializer(a
rticle)
        return Response(serializer.
data)

    def put(self,request,pk):
        article=self.get_object(pk)
        serializer=Studentserializer(ar
ticle,data=request.data)
        if serializer.is_valid():
            serializer.save()
            return Response(serializer.
data)
        return Response(serializer.
errors,status=status.HTTP_400_BAD_
REQUEST)

    def delete(self,request,pk):
        article=self.get_object(pk)
        return Response(status=status.
HTTP_204_NO_CONTENT)
```

The 'get_object' function aims to obtain the object within the existing table having the same primary key value, i.e., 'pk' as the one entered in the URL. It will return an error message if no entry in our model exists with the entered primary key value. The 'get' function simply displays the model entry with the given 'pk'. 'get', 'put' and 'delete' functions make use of the 'get_object' function and treat it as the first step in trying to obtain the required individual value. Add the required URL in 'urls.py':

```
path("student_details/<int:pk>/",Stud
entDetails.as_view(),
    ),
```

In Figure 12, I aimed to change just the 'Last_name' of the student with *pk=1*. The change will be reflected when you 'GET' all student info.

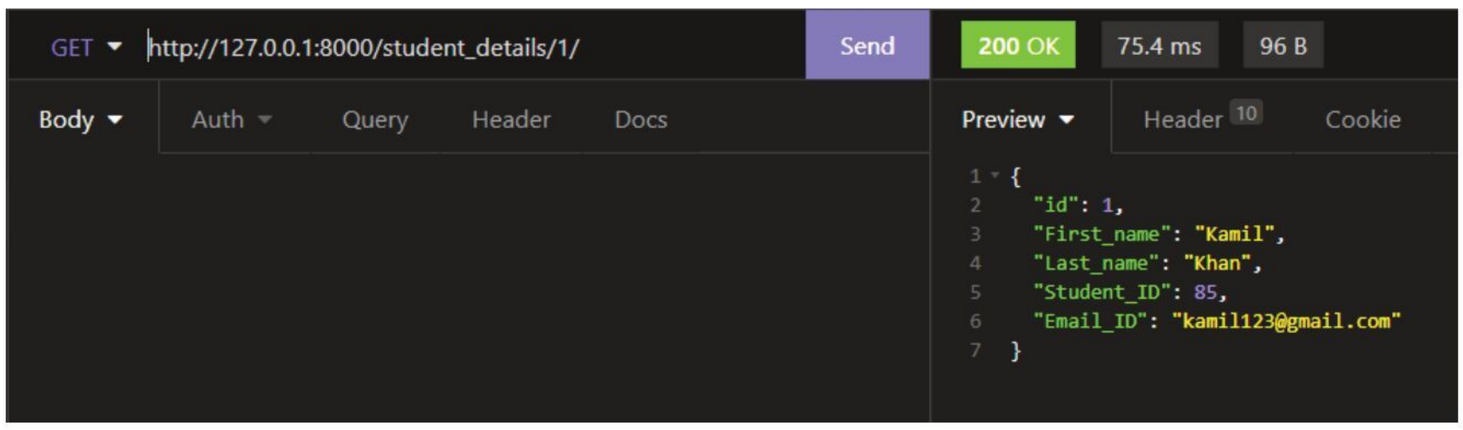


Figure 11: GET-ting individual student info

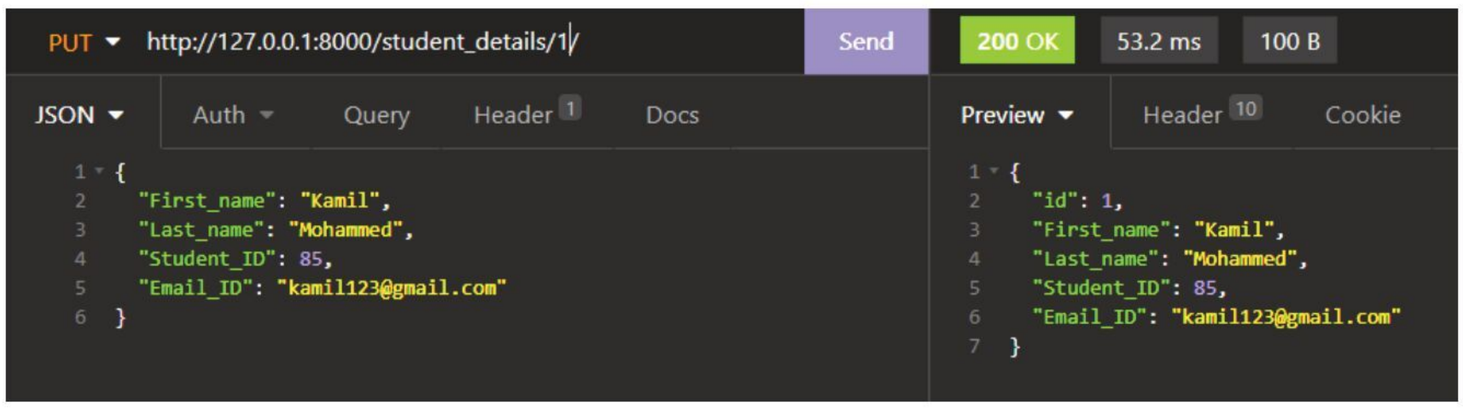


Figure 12: PUT-ting student info

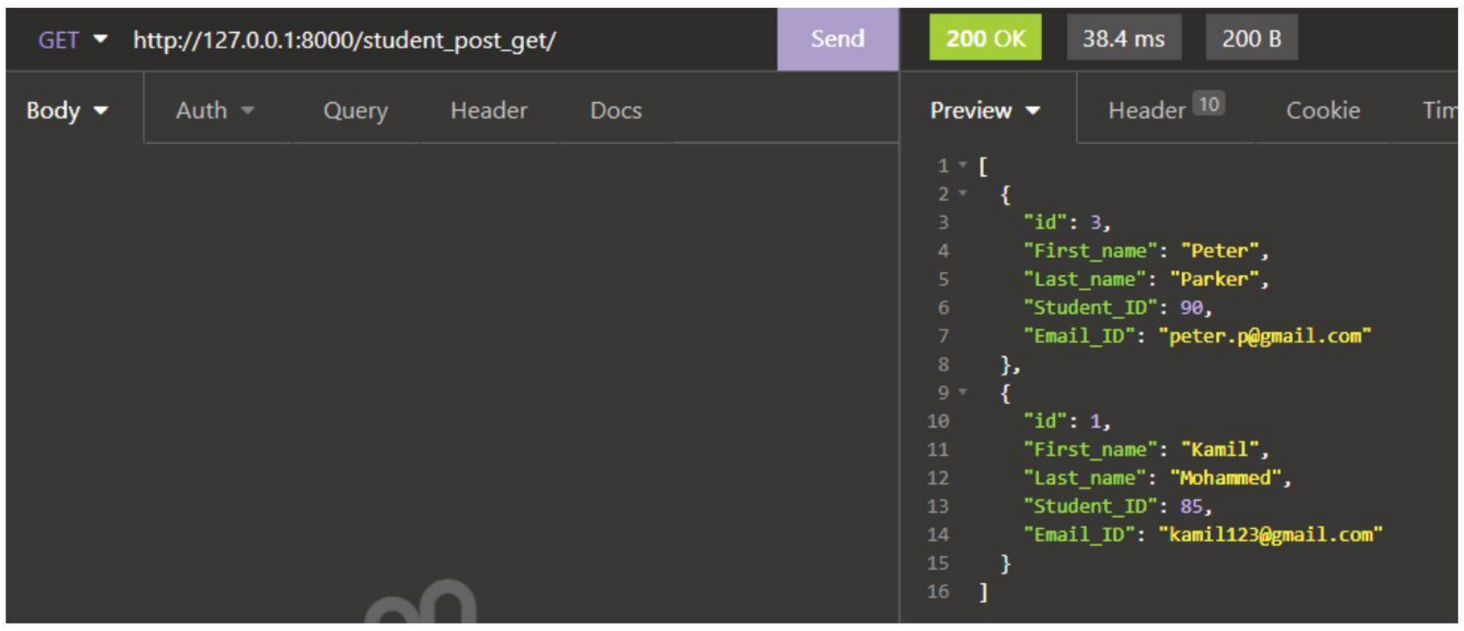


Figure 13: Updated 'GET' output

We will now delete the entry in the model having *pk=2*. All you have to do is enter 2 in the URL as '*http://127.0.0.1:8000/student_details/2/*' and perform the 'DELETE' operation. Let's see if the same is reflected in the model.

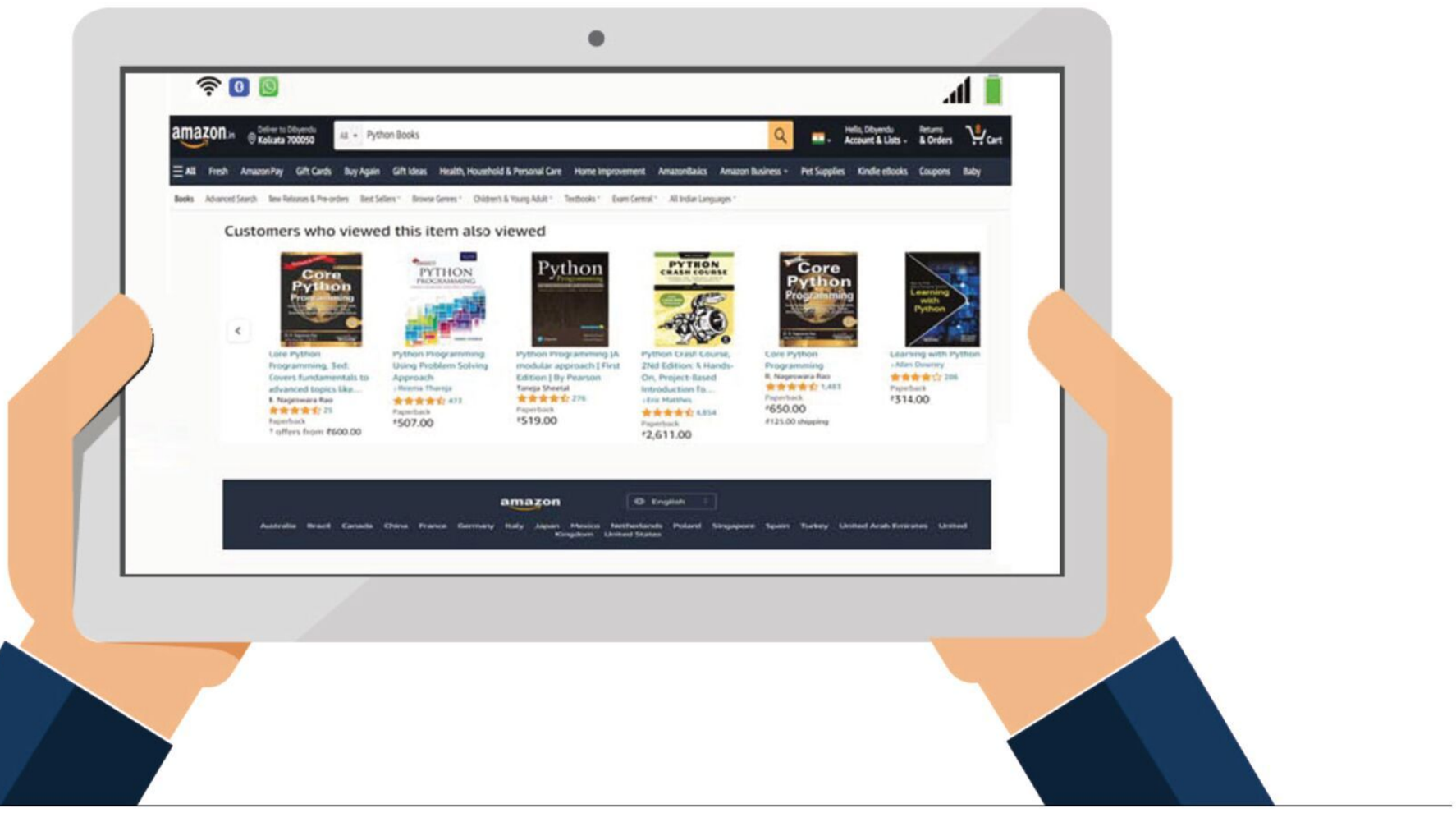
Having gone through the article, you would have grasped an idea of the absolute basics of generating APIs; specifically, how to perform CRUD operations. END

By: Mohammed Kamil Khan

The author is currently working at Vamstar as a data discovery expert.

Recommender Systems: An Overview of the Mathematics

This two-part series of articles explains and demonstrates how to implement a recommender system for an online retail store using Python. In this first part of the series, the focus is on the theory behind such a system. The implementation will be covered in the second part.



Recommender systems are now widely used to personalise your online experience, advising you on what to buy, where to dine, and even who you should be friends with. People's preferences vary, yet they tend to follow a pattern. They have a tendency to like things that are comparable to the other things they enjoy, and they have tastes that are similar to the tastes of the people they know. Recommender systems attempt to capture these trends to forecast what users might enjoy in the future. E-commerce sites, social networks, and online news platforms actively implement their own recommender systems to assist customers in making product choices more efficiently.

You would have observed that Netflix and Amazon Prime recommend content that is similar to what you watched previously. So, how does this happen? It's because of a machine learning based recommendation engine that figures out how similar the videos are to other things you like, and then it predicts your choices.

In 2006, Netflix announced a competition to develop an algorithm that would "significantly increase the accuracy of forecasts about how much someone will enjoy a movie based on their movie tastes." Since then, Netflix has expanded its use of data to include personalised ranking, page generation, search, picture selection, messaging, and marketing, among other things. The Netflix Recommendation Engine (NRE), its most effective algorithm, is made up of algorithms that select content based on each individual user profile. It's so accurate that customised recommendations drive 80 per cent of Netflix's viewer engagement.

Netflix isn't the only business that uses a recommender system. Amazon, LinkedIn, Spotify, Instagram,

YouTube, and a few other websites employ recommendation algorithms to anticipate customer preferences and grow their businesses.

With the growth of YouTube, Amazon, Netflix, and other similar Web services over the last few decades, recommender systems have become increasingly important in our lives. In a broad sense, these systems are algorithms that try to propose relevant items to consumers (these can range from music to perfumes, books, products to buy or anything else, depending on the industries).

Now let's look at the various recommender system paradigms and how they work, outline their mathematical foundations, and discuss the strengths and weaknesses of each of them.

Relationships in the context of developing a recommendation system

User-product relationship: When some users have an affinity or desire for specific items, this is known as the user-product connection. An athlete, for example, may have a taste for athletics-related things; therefore the e-commerce website will create an *Athletics -> Sports* user-product relationship.

Product-product relationship: When items are similar in nature, whether by look or description, they form product-product associations. Books or music in the same genre, and dishes from the same cuisine are all such examples.

User-user relationship: When two or more customers have similar tastes in a product or service, they form user-user relationships. Mutual friends, same backgrounds, comparable ages, and so on, are examples of this.

Providing data for a recommender system, and some challenges

Data can be provided to a recommender system in a variety of methods. Explicit and implicit ratings are two of the most essential methods.

Explicit ratings: The user expresses his or her opinion through explicit ratings. Star ratings, reviews, feedbacks, likes, and following someone are just a few examples. Exact ratings can be difficult to obtain because people do not always rate things.

Implicit ratings: When people interact with an item, they give implicit ratings. These infer a user's activity and are simple to obtain because consumers click implicit ratings subconsciously. Clicks, views, and purchases are all examples of such ratings.

> **Note:** Users spend time and money on what is most important to them; therefore views and purchases are a better entity to promote.

There are some specific technical challenges to building a recommendation engine, some of which are explained below.

- **Lack of data:** To produce effective predictions, recommender systems require a huge amount of data. Large firms like Google, Apple and Amazon can generate better recommendations since they collect data about their customers on a regular basis. There are also recommender systems that are based just on purchases, with no ratings recorded. If clients do not rate a book, Amazon does not know how much they enjoyed it.
- **Cold start:** This situation comes when new customers or new items are added to the system; a new item cannot be recommended to customers when it is first introduced to the recommender system because it lacks ratings or reviews, making it difficult to predict the choice or interest of customers, resulting in less accurate recommendations. A newly released movie, for example, cannot be recommended to the user until it has received some ratings. A new customer item added-based problem is difficult to handle because it is impossible to find similar users without prior knowledge of their interests or preferences.
- **Sparsity:** This occurs when the majority of users do not provide ratings or reviews for the items they purchase, causing the rating model to become very sparse, potentially leading to data sparsity issues. This reduces the chances of finding a group of users with similar ratings or interests.
- **Latency:** Many products are added frequently to the database of recommendation systems; however, only the previously existing products are generally recommended to users because newly added products have not yet been rated.

Use cases for recommender systems

The following are some of the potential use cases of a recommender system that can add value to a business.

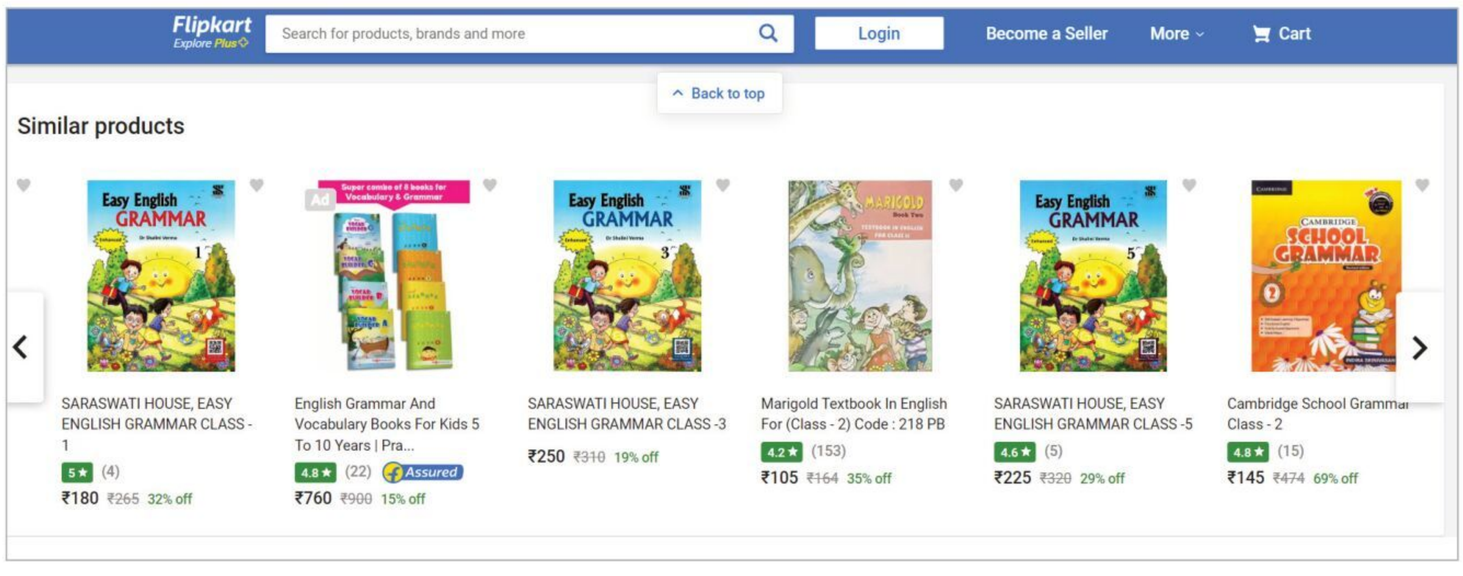


Figure 1: E-commerce sites show alternative products

- ***Alternative product recommendations:*** You've probably tried to buy something you wanted, only to find it wasn't available. But it's unlikely you left the website after that. Instead, you may come across something similar, which you may or may not purchase. Companies utilise a recommendation system to keep customers online by providing them with alternatives to the product they want.
- ***Related product recommendations:*** Another important application of recommender systems is related product recommendations. When customers make a purchase, it's only natural to offer them something complementary. If a customer purchases a wrist watch, the system may suggest a wallet, which goes well with it.
- ***Real-time true personalisation:*** With every second of customer activity on e-commerce sites, recommender systems can provide product or content or lifestyle recommendations. They can serve as a shopping assistant, using all the information the customer gives to offer improved recommendations, thus creating a better shopping experience.

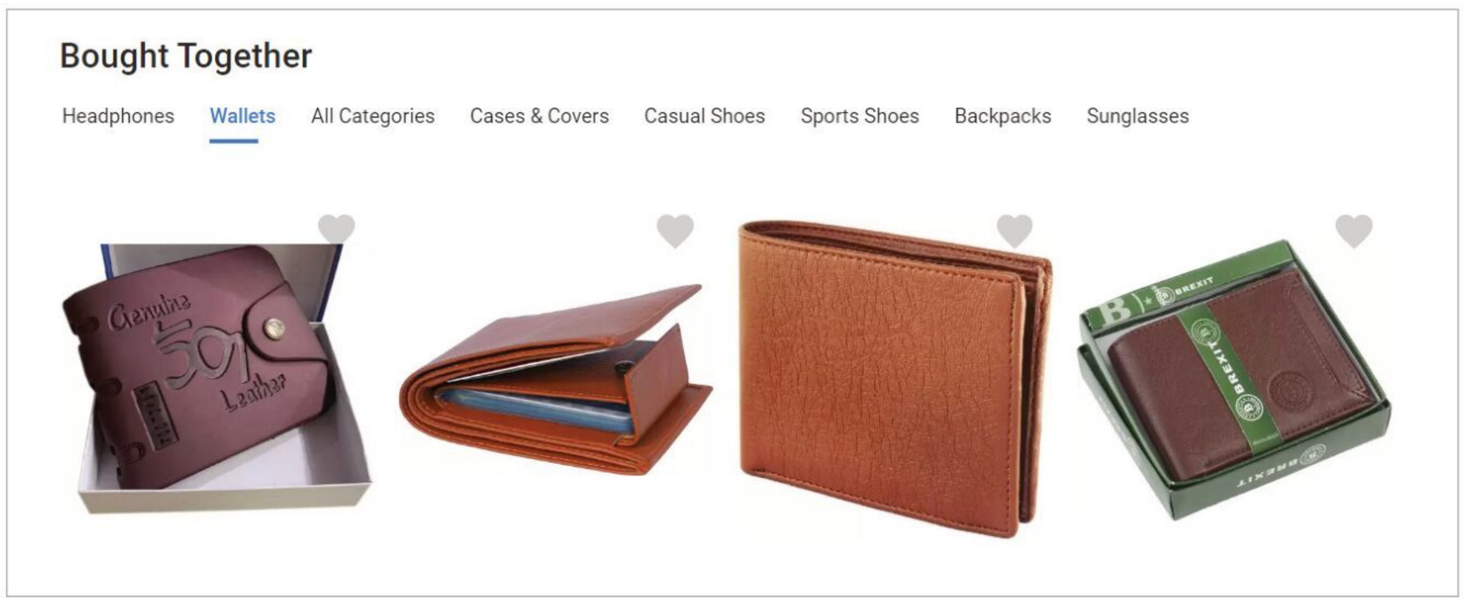


Figure 2: Related product recommendations

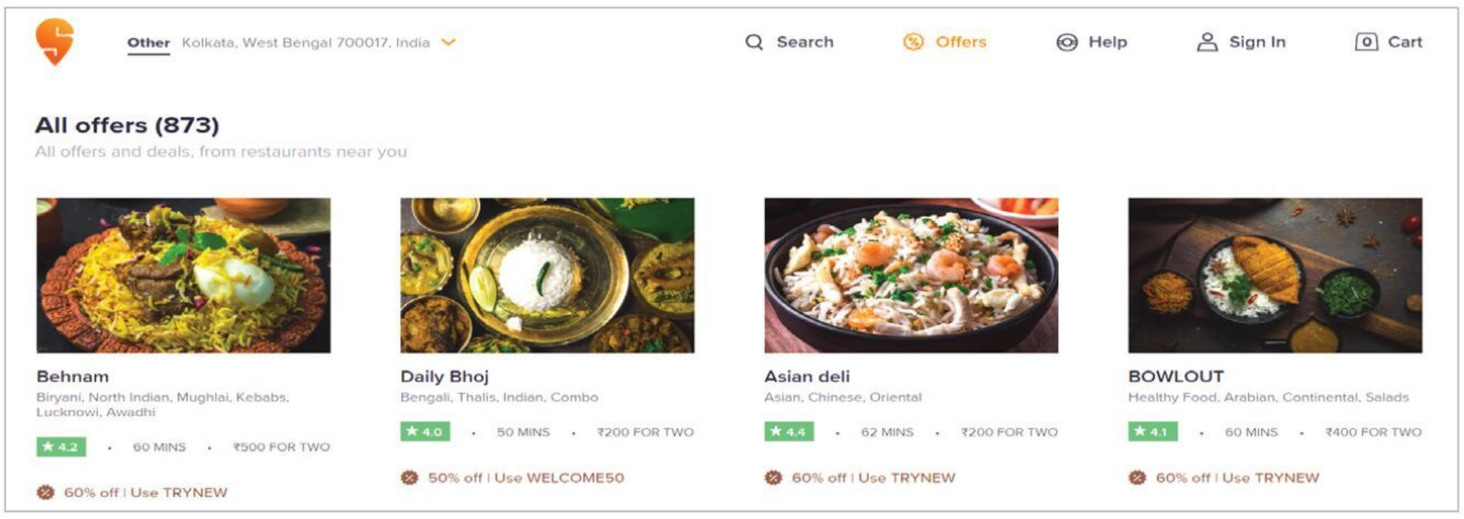


Figure 3: Location specific recommendations

Recommender systems: A broad categorisation

Content-based filters

Content-based recommender systems give recommendations based on objects or user metadata. The user's previous purchases are scrutinised. For example, if a user has previously read a book by a particular author or purchased a product from a particular brand, it is considered that the user has a preference for that author or brand and is likely to purchase a similar product in the future. For example, if a user likes the book 'The Adventures of Feluda' by Satyajit Ray, then the system recommends more books by the same author or other

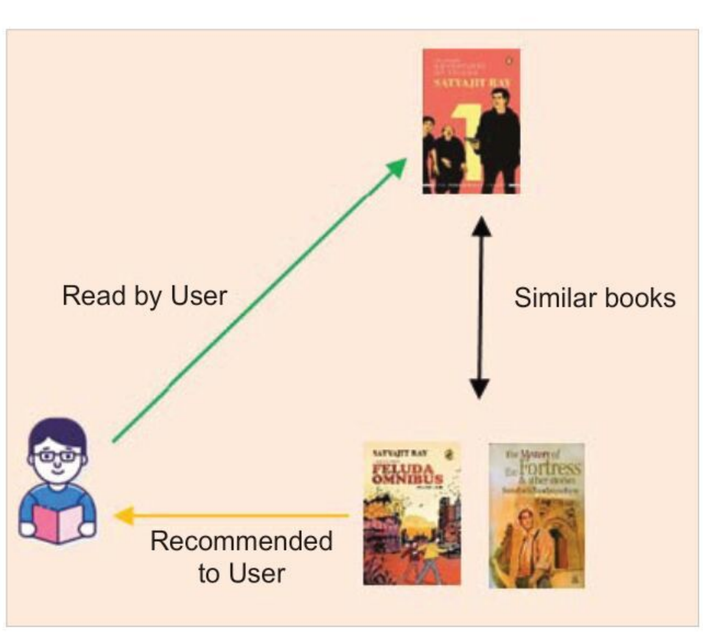


Figure 4: Content based filtering

detective books like 'The Mystery of the Fortress & Other Stories'.

Two types of data are used in this filtering. The first includes the user's preferences, interests, and personal information such as age or, in some cases, the user's history. The user vector represents this information. The second is an item vector, which is a collection of information about a product. The item vector contains all the information that can be used to calculate the similarity between things. Cosine similarity is used to determine the recommendations.

Collaborative filtering

This is regarded as one of the most intelligent recommender systems, and is based on the similarity between different users and things such as e-commerce websites and online movie websites. It makes recommendations based on the preferences of comparable users. Similarity is not limited to the user's preference; it can also be considered between distinct objects. If we have a big amount of knowledge about people and things, the algorithm will make more efficient recommendations.

Collaborative filtering is a method of recommending items that a user might enjoy based on the reactions of other users. It works by sifting through a huge group of people to locate a smaller group of users with likes similar to a specific user. It analyses their favourite things and creates a ranked list of recommendations.

A collaborative filtering system's workflow is usually as follows:

- A user expresses his or her choices by rating system objects (for example, music, books, or movies).

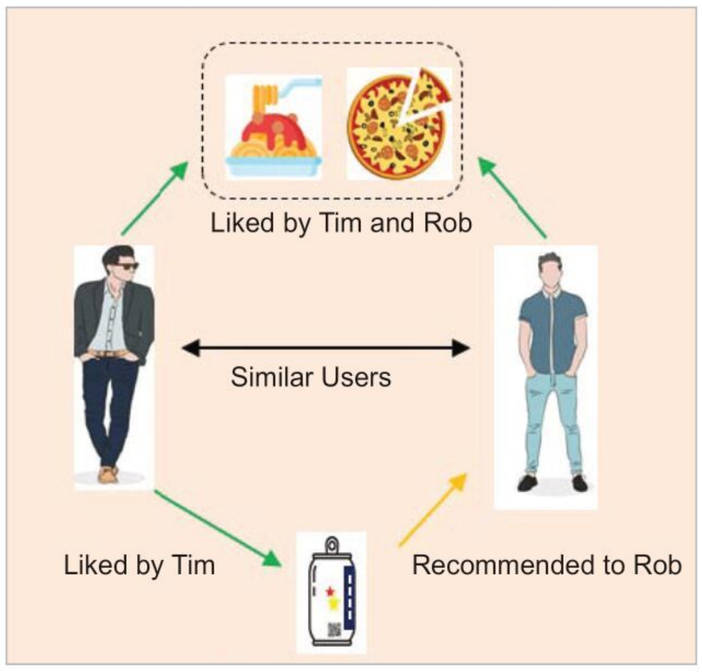


Figure 5: Collaborative filtering

These ratings might be considered as an approximation of the user's interest in the topic.

- The system matches this user's preferences against other user preferences and finds the people with the most 'similar' choice.
- The system recommends items that similar users have rated the most but have not yet been rated by this user (the absence of rating is often considered as unfamiliarity with an item).

The main challenge of collaborative filtering is combining and weighing the preferences of user neighbours. Users can sometimes rate the recommended items right away. As aresult, over time, the system obtains a more accurate depiction of consumer preferences.

The math behind finding the 'similarity' to recommend

So, how does a recommender system determine 'similarity'? The distance metric is used to do this. Similarity is a subjective concept that varies greatly depending on the domain and application. The points closest to each other are the most similar, while the points furthest apart are the least relevant. Two films, for example, may be comparable in terms of genre, length, or actors. When measuring distance between unrelated dimensions/ features, caution is advised. The relative values of each element must be normalised, else the distance calculation will be dominated by one feature.

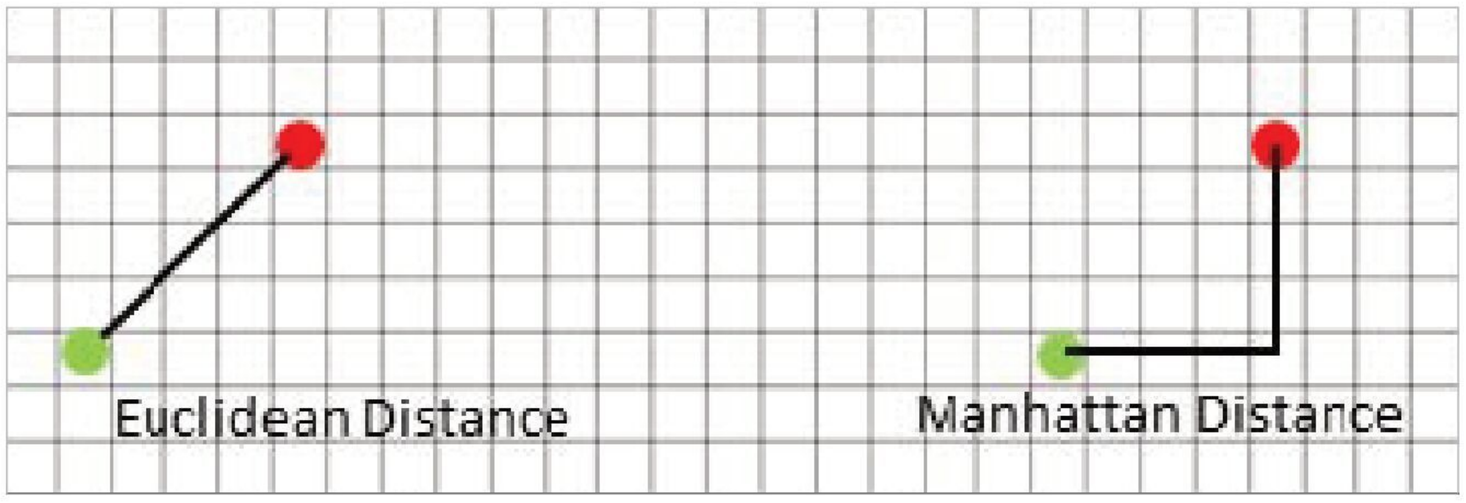


Figure 6: Euclidean distance and Manhattan distance

Minkowski distance: The Minkowski distance is the general form when the dimension of a data point is numeric. It's a metric for vector spaces with real values. Minkowski distance can only be calculated in a normed vector space, which is a space where distances can be represented as a vector with a length that cannot be negative. Manhattan(r=1) and Euclidean(r=2) distance measurements are generalisations of this generic distance metric.

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

The formula shown above for Minkowski distance is in a generalised form, and we can manipulate it to get different distance metrics. Here, the 'p' value can be manipulated to give us different distances like:

- p = 1; when p is set to 1, it is Manhattan distance
- p = 2; when p is set to 2, it is Euclidean distance

Manhattan distance: This distance is also known as taxicab distance or city block distance because of how it is calculated. The sum of the absolute differences of two places' Cartesian coordinates is the distance between them – it's the distance between two places when measured at right angles along axes. The formula is:

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)$$

Euclidean distance: The square root of the sum of squares of the difference between the coordinates is given by the Pythagorean theorem:

$$d(x, y) = \sqrt{\sum_{i}^{n} (x_i - y_i)^2}$$

Cosine similarity: This is a metric that is useful in determining how similar the data objects are, irrespective of their size. Cosine similarity calculates the cosine of the angle between two vectors to determine how similar they are. In cosine similarity, data objects in a data set are treated as a vector. The formula to find the cosine similarity between two vectors is:

```
Cos(x, y) = x . y / ||x|| *
||y||
```

where,

- x, y = product (dot) of the vectors 'x' and 'y'
- ||x|| and ||y|| = length of the two vectors 'x' and 'y'
- ||x|| * ||y|| = cross product of the two vectors 'x' and 'y'

The cosine similarity between two vectors is measured in 'θ'. If θ = 0°, the 'x' and 'y' vectors overlap, thus proving they are similar. If θ = 90°, the 'x' and 'y' vectors are dissimilar.

Word embedding in the context of building a recommender system

Word embeddings are texts that have been translated into numbers, and there may be multiple numerical

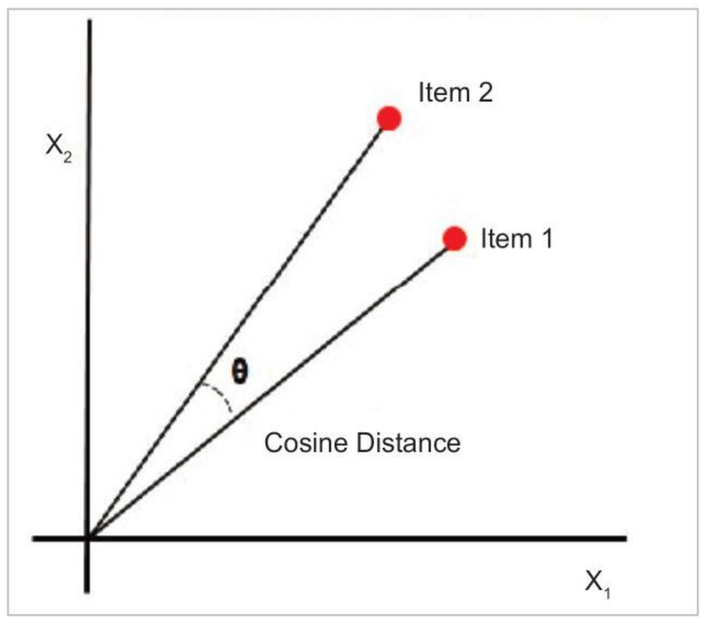


Figure 7: Cosine similarity

representations of the same text. But before we get into the specifics of word embedding, we need to consider the following question: Why do we need it?

Many machine learning algorithms and almost all deep learning architectures are unable to interpret strings or plain text in their raw form, as we all know. They need numbers as inputs to accomplish any task, including classification, regression, and so forth. Word embedding is nothing but the process of converting text data to numerical vectors. These vectors are then utilised to create a variety of machine learning models. In a sense, we could describe this as extracting features from text in order to create numerous natural language processing models.

We can transform text data into numerical vectors in a variety of ways. We'll learn about numerous word embedding strategies with examples in this post, as well as how to apply them in Python.

Techniques for word embedding

The common and easy word embedding methods for extracting characteristics from text are:

- TF-IDF
- Bag of words
- Word2vec

TF-IDF

A popular word embedding technique for extracting features from corpus or language is TF-IDF. It is a statistical way of finding how important a word is to a document over other documents.

TF stands for 'term frequency'. In TF, we assign a score to each word or symbol according to how frequently it appears. The frequency of a word is determined by the document's length. This means that a word appears more frequently in huge documents than in small or medium documents. To solve this problem, we can normalise the frequency of a word by dividing it by the length of the document (total number of words).

Using this method, we may create a sparse matrix with the frequency of each word.

```
TF=Number of times a term occurs in a document/total number of words in a document
```

IDF

The full form of IDF is 'inverse document frequency'. Here, we are assigning a score value to a word, which explains how infrequent a word is across all documents. Infrequent words have higher IDF scores.

```
IDF = log base e (total number of documents/number of documents that have the term)
```

```
TF - IDF = TF * IDF
```

TF-IDF value is increased based on the frequency of the word in a document. This approach is mostly used for document/text classification, and also successfully used by search engines like Google as a ranking factor for content.

When compared to the words that are relevant to a document, some common words (like 'a', 'the', 'is') appear relatively frequently in a huge text corpus. These popular words convey very little information about the document's actual contents. If we fed the count data of these commonly occurring words directly to a classifier, the frequencies of rarer but more intriguing phrases would be overshadowed. So, ideally, what we want is to give lesser weightage to the common words occurring in almost all documents and give more importance to words that appear in a subset of documents.

The TF-IDF transform is frequently used to re-weigh count features into floating point values suitable for use by a classifier. This technique considers the occurrence of a word over the entire corpus, not just in a single document. Consider a business article — it will have more business-related terms like stock markets, prices, shares, and so on, than any other article, but terms like 'a', 'an', and 'the' will also appear frequently in it. As a result, this technique will punish certain high-frequency words.

Consider the table given below, which shows the total number of terms (tokens/words) in two papers.

Document 1		Document 2	
TERM	COUNT	TERM	COUNT
this	1	this	1
is	1	is	2
about	2	about	1
pandemic	4	technology	1

Let's define a few terminologies associated with TF-IDF now.

```
Term Frequency (TF) = (Number of times term t appears in a document)/(Number of terms in the document)
So, TF (This, Document1) = 1/8
TF (This, Document2) =1/5
```

This denotes the word's contribution to the document, i.e., terms that are relevant to the document should appear frequently. For example, a document discussing a pandemic should use the word 'pandemic' frequently.

```
IDF = log (N/n), where, N is
```

```
the number of documents and n
is the number of documents a
term t has appeared in.
So, IDF (This) = log (2/2) = 0.
```

So, what's the best way to explain why IDF exists. Ideally, if a term appears multiple times in a paper, it is unlikely that it is significant to that document. However, if it appears in a subset of papers, it is likely that the word is relevant to the documents in which it appears.

Let us compute IDF for the word 'pandemic':

```
IDF (pandemic) = log (2/1) = 0.301.
```

Now, let us compare the TF-IDF for a common word 'this' and a word 'pandemic' which seems to be of relevance to Document 1.

```
TF-IDF (this, Document1) =
(1/8) * (0) = 0
TF-IDF (this, Document2) =
(1/5) * (0) = 0
TF-IDF (pandemic, Document1) =
(4/8)*0.301 = 0.15
```

As you can see, for Document 1, the TF-IDF method heavily penalises the word 'this', but assigns greater weight to 'pandemic'. So, this may be understood as 'pandemic' is an important word for Document 1 from the context of the entire corpus.

Bag of words

Humans can understand a language in a fraction of a second. Machines, on the other hand, are unable to comprehend a language. We need to translate the text into a numerical format that the system can interpret. Because most machine learning and statistical models deal with quantitative data, we must transform all the texts to numbers. Bag of words is one of the simplest techniques to do this. In this approach, we perform two operations:

- Tokenization
- Vectors creation

Tokenization: This is the process of breaking down each sentence into words or smaller chunks. Each word or symbol is referred to as a token in this context. We can take unique words from the corpus after tokenization. The tokens we have from all the documents we're examining for the bag of words construction are referred to as 'corpus'.

Vectors creation: Here, the size of the vector is equal to the number of unique words of the corpus. We can fill each position of a vector with the corresponding word frequency in each sentence. Let's look at an example.

- I like to drink tea.
- Tea is good for your health.
- Tea is more popular than coffee in India.

If we want to analyse the above sentences or build any ML model, we first need to convert them into numbers, as I mentioned above.

The first step is to tokenize the sentences. This is nothing but splitting the above three sentences into individual words.

The next step is to find all the unique words from the above sentences and build the vocabulary. We'll have to make a dictionary with keys and values. The words in our corpus serve as keys, and the frequency of those terms throughout the corpus serves as values. To put it another way, simply count the number of times each word appears in each sentence.

From Table 1, we can see the numbers corresponding to the respective sentences. The first row represents the vector for sentence 1. Given below are the vectors for:

```
Sentence 1 =
[1,1,1,1,1,0,0,0,0,0,0,0,0,0,0]
Sentence 2 =
[0,0,0,0,1,1,1,1,1,0,0,0,0,0,0]
Sentence 3 =
[0,0,0,0,1,1,0,0,0,1,1,1,1,1,0]
```

The bag of words paradigm is based on this concept. The above vector representation can be used as an input for ML or statistical models.

Word2vec

In a sentence, every word is dependent on another word or words. Word dependencies must be captured if we are to uncover commonalities and relationships between words.

We can't capture the meaning or relationship of words from vectors using the bag of words and TF-IDF approaches. Word2vec creates embeddings, which are vectors. The word2vec model takes a big corpus as input and outputs it in vector space. The dimensionality of this vector space could be in the hundreds. This vector space will be used to position each word vector.

In vector space, all words in a corpus that are closer to each other share a common context. In this space, a word vector contains the positions of associated words.

Continuous bag of words (CBOW) and skip-gram are two neural network-based versions offered by word2vec. The neural network model in CBOW takes a variety of words as

	I	like	to	drink	tea	is	good	for	health	more	popular	in	India	than	coffee
Sentence 1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
Sentence 2	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0
Sentence 3	0	0	0	0	1	1	0	0	0	1	1	1	1	1	0

Table 1

input and predicts a target word that is closely connected to the input words' context. The skip-gram design, on the other hand, takes one word as input and anticipates its closely related context terms.

CBOW is faster and finds better numerical representations for common words, although skip-gram can represent rare words effectively. Word2vec models excel at capturing semantic connections between words. Consider the relationship between a country and its capital, such as France's capital, Paris, and Germany's capital, Berlin. These models are best at doing semantic analysis, which is useful for making recommendations.

Applicability of word2vec models in building product recommendations: Word2vec uses the sequential nature of the text as a fundamental property of a natural language to produce vector representations of text. There is a word sequence in every sentence or phrase. We would have great difficulty understanding the text without this sequence.

One such example is the purchases made by consumers on e-commerce websites online. There is generally a pattern in the buying behaviour of consumers. For instance, a person involved in sports-related activities might have an online buying pattern similar to what is shown in Figure 8.

We can quickly locate related products if we can represent each of these things as a vector. So, if a user is looking at a product online, we can quickly offer related things to him or her based on the product's vector similarity score. However, how can we obtain a vector representation of these items? Is it possible to obtain these vectors using the word2vec model? Yes, it is! Consider a customer's purchasing history as a sentence, with the products as the words, as shown in Figure 9.

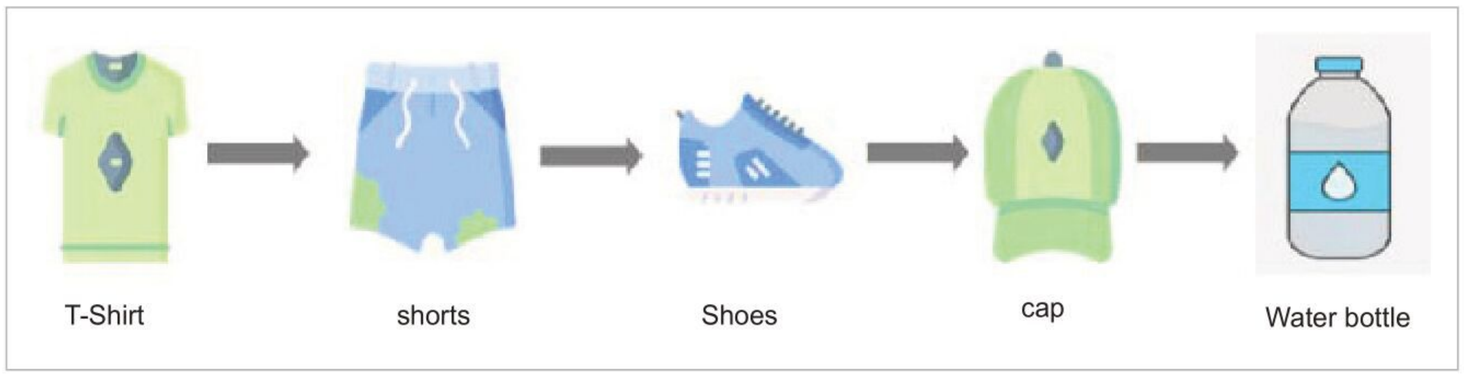


Figure 8: Purchase history

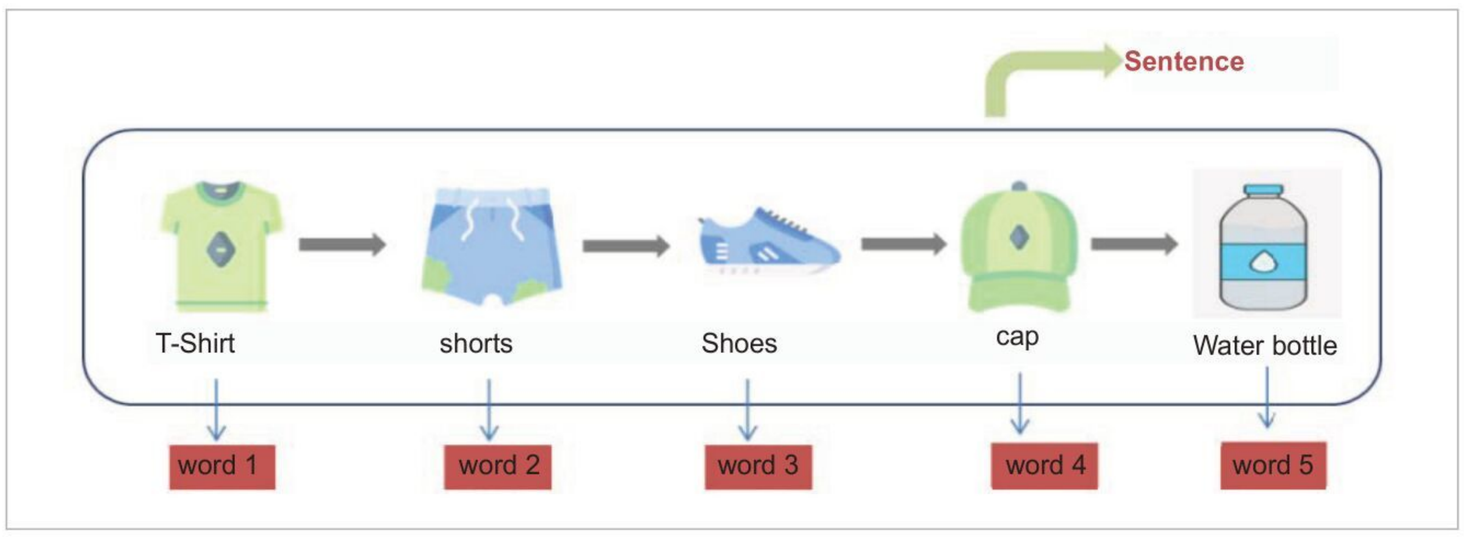


Figure 9: Customer's purchasing history as a sentence

So we have seen how a recommender system works and the mathematics behind it. We have also covered some basics of the ML being used in such a system. In the next part of this two-part series of articles, we will see how we can apply the knowledge learnt here to build a recommender system in Python. END

References

[1] *https://en.wikipedia.org/wiki/Recommender_system*
[2] *https://www.tensorflow.org/tutorials/text/word2vec*
[3] *https://www.opensourceforu.com/2020/08/natural-language-processing-simplified*
[4] *https://www.opensourceforu.com/2021/11/the-significance-of-neural-networks-in-nlp*

By: Dibyendu Banerjee

The author is a senior architect at Cognizant's AI and analytics practice, with more than 15 years of IT experience. His areas of expertise are Python, R, Java and other open source technologies.

THE COMPLETE MAGAZINE ON OPEN SOURCE
OpenSource ForYou
The latest from the Open Source world is here.
OpenSourceForU.com
Join the community at facebook.com/opensourceforu
Follow us on Twitter @OpenSourceForU

Data Visualisation in R: Graphs

In this tenth article in the R series, we will continue to explore data visualisation in R with the lattice and ggplot2 packages.



We will be using the R version 4.1.2 installed on Parabola GNU/Linux-libre (x86-64) for the example code snippets in this article.

```
$ R --version
R version 4.1.2 (2021-11-01) -- "Bird Hippie"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)
```

R is free software and comes with absolutely no warranty. You are welcome to redistribute it under the terms of the GNU General Public License versions 2 or 3. For more information about these matters, see *https://www.gnu.org/licenses/*.

Lattice

Line chart

Consider the consumer prices (annual per cent) inflation data for India between 1960 and 2022 available from the World Bank. You can use the years in the x-axis, and the inflation on the y-axis to produce a line chart using the *xyplot* function, as shown below:

```
> x<-c(1960:2020)

> y<-c(1.77,1.69,3.63,2.94,13.35,9.47,10.80,13.06,3.23,-
0.58,5.09,3.07,6.44,16.94,28.59,5.74,
-7.63,8.30,2.52,6.27,11.34,13.11,7.89,11.86,8.31,5.55,8.72,8.
80,9.38,7.07,8.97,13.87,11.78,6.32,10.24,10.22,8.97,7.16,13.2
3,4.66,4.00,3.77,4.29,3.80,3.76,4.24,5.79,6.37,8.34,10.88,11.
98,8.85,9.31,11.06,6.64,4.90,4.94,3.32,3.94,3.72,6.62)

> d <- data.frame(x,y)

> xyplot(y~x, data=d, type="l", main="Inflation, consumer
prices (annual %)")
```

The line chart is shown in Figure 1.

The *xyplot* accepts the following arguments:

Argument	Description
data	A data frame containing values
groups	A grouping variable in the data
main	The title of the chart
scales	A list that specifies how the axes are to be drawn
strip	A logical condition on whether to draw strips
x	The primary numeric variable
xlab	The label for x-axis
xlim	A numeric vector that specifies left and right limits for x-axis
ylab	The label for y-axis
ylim	A numeric vector of length two that mentions lower and upper limits for y-axis

The barchart function

The *bar chart* function produces a bar chart for the given data. In the following example, we specify a function to the axis argument to use the year on the x-axis.

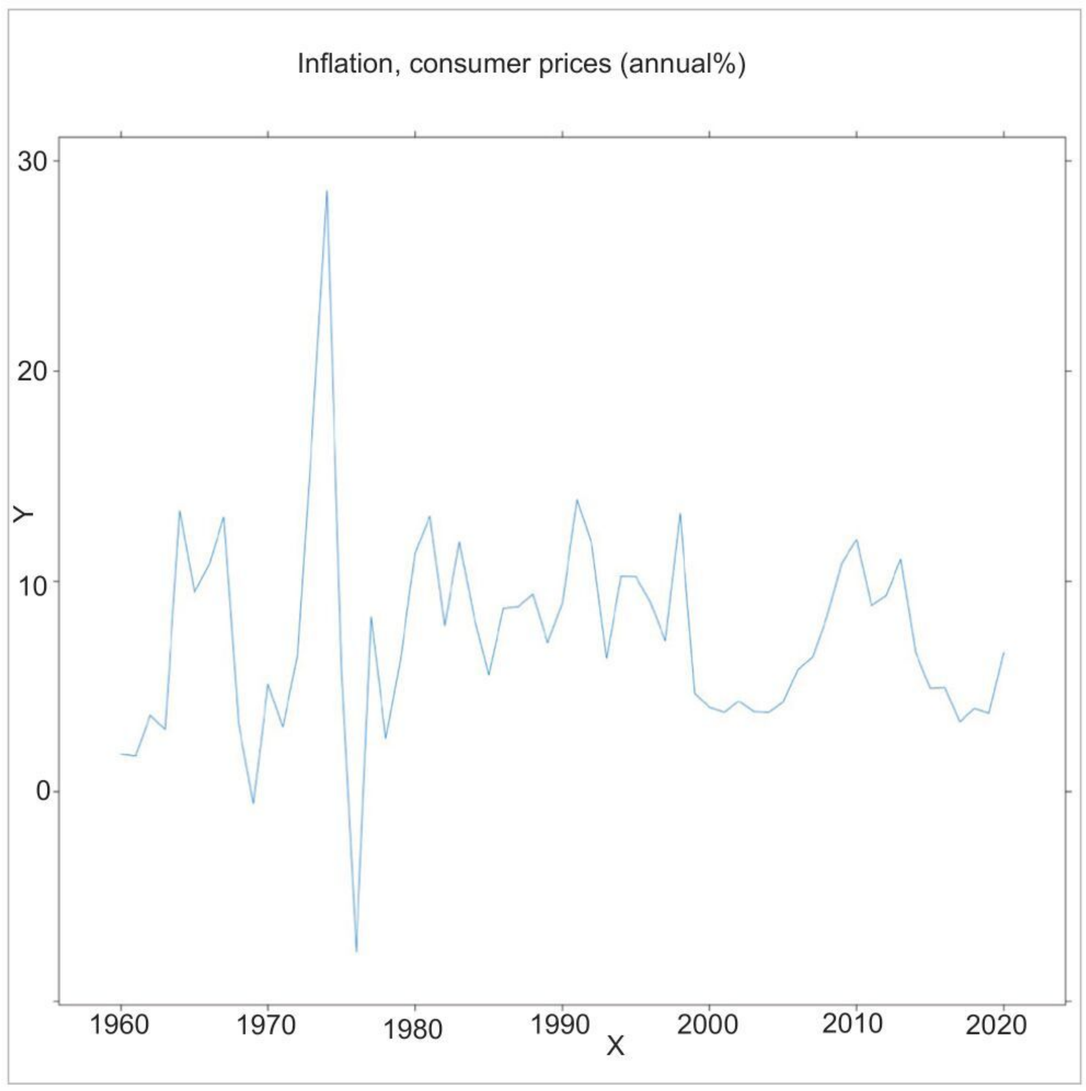


Figure 1: Line chart

```
> barchart(y~x|x, data=d, horizontal=FALSE,
axis=function(side, ...) { if (side==”bottom”) panel.
axis(at=seq_along(d$x), label=d$x, outside=TRUE, rot=0,
tck=0) else axis.default(side, ...)}, main=”Inflation,
consumer prices (annual %)”)
```

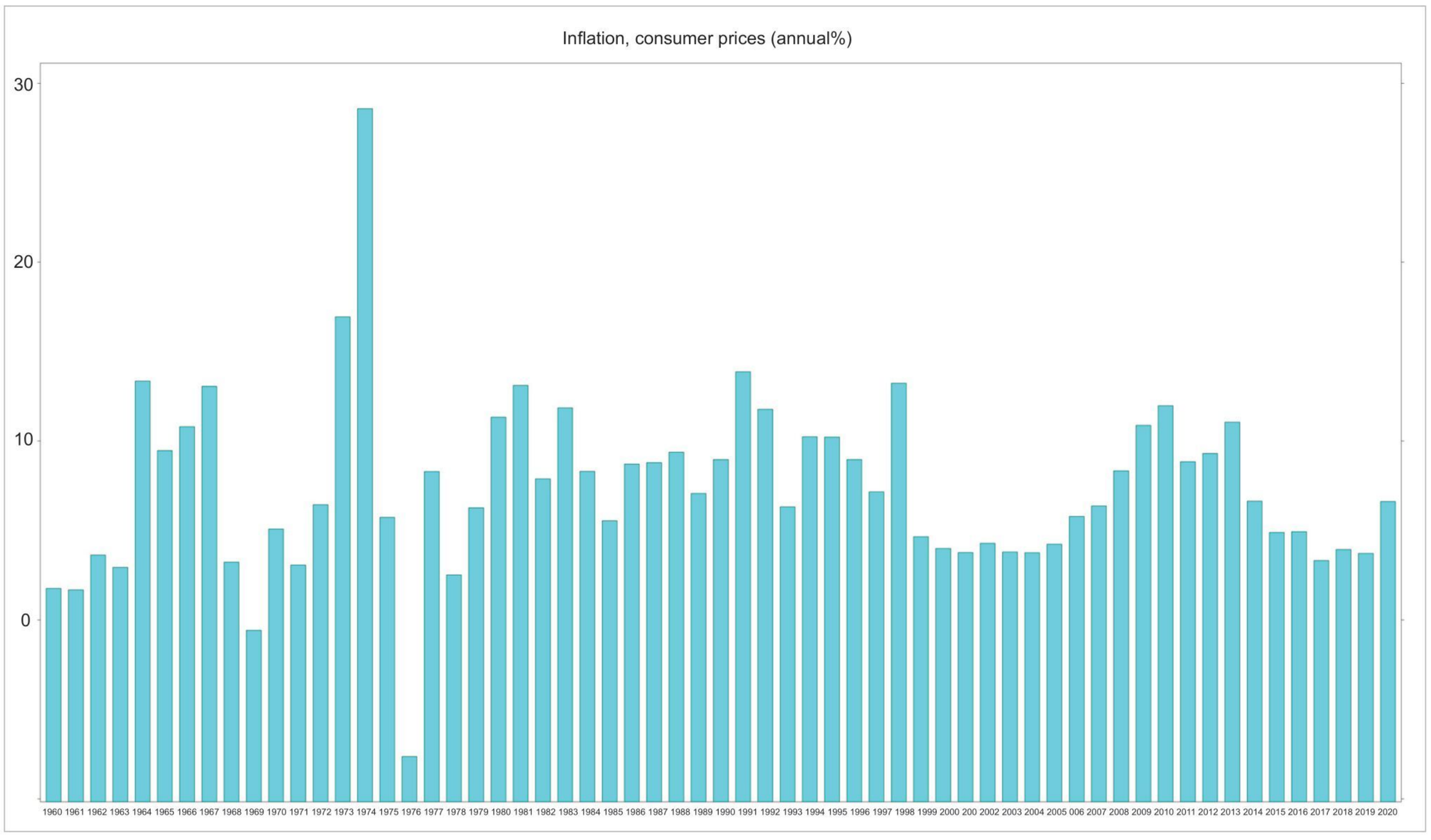


Figure 2: Bar chart

The additional set of arguments available to the *xyplot* and *barchart* are listed below:

Argument	Description
box.ratio	Specifies the ratio of the width of rectangles in barchart
panel	Plots x and y variables in each panel
default.prepanel	A default function as a fallback to the prepanel function
auto.key	Used to produce a suitable legend
aspect	The physical aspect ratio of the panels
axis	A function responsible for drawing the axis annotation
horizontal	The orientation of the bar chart
subscripts	A logical flag to pass a 'subscripts' vector to the panel function
subset	A set of rows from the data is used in the plot

Scatter plot

You can also display individual charts on a panel grid. For example, the all India consumer price index (rural/urban) data set up to November 2021 is available from *https://data.gov.in/catalog/all-india-consumer-price-index-ruralurban-0* for the different states in India. We can read the data from the downloaded file using the *read.csv* function, as shown below:

```
> cpi <- read.csv(file="CPI.csv", sep=",")

> head(cpi)
       Sector Year     Name Andhra.Pradesh Arunachal.Pradesh
Assam Bihar
1       Rural 2011  January    104              NA   104    NA
2       Urban 2011  January    103              NA   103    NA
3 Rural+Urban 2011  January    103              NA   104    NA
4       Rural 2011 February    107              NA   105    NA
5       Urban 2011 February    106            NA   106    NA
6 Rural+Urban 2011 February    105            NA   105    NA
  Chattisgarh Delhi Goa Gujarat Haryana Himachal.Pradesh
Jharkhand Karnataka
1      105    NA 103     104     104    104    105        104
2      104    NA 103     104     104    103    104        104
3      104    NA 103     104     104    103    105        104
4      107    NA 105     106     106     05     107        106
5      106    NA 105     107     107    105    107        108
6      105    NA 104     105     106    104    106        106
```

The *aggregate* function can be used to obtain the values for the state of Andhra Pradesh as follows:

```
ap <- aggregate(x=cpi$Andhra.Pradesh, by=list(cpi$Year),
FUN=sum)

> head(ap)
  Group.1       x
1    2011 3911.28
2    2012 4255.40
3    2013 4516.60
4    2014 4673.60
5    2015 4822.20
6    2016 4921.50
```

A simple scatter plot can be displayed for the consumer price indexes using the following arguments to the *xyplot* function:

```
> xyplot(x~Group.1, ap, main="Andhra Pradesh Consumer Price
Index upto November 2021", xlab="Year", ylab="Consumer Price
Index")
```

The corresponding scatter plot illustration is shown in Figure 3.

Panel grid

You can also visualise the values per year (Group.1) using the *xyplot:*

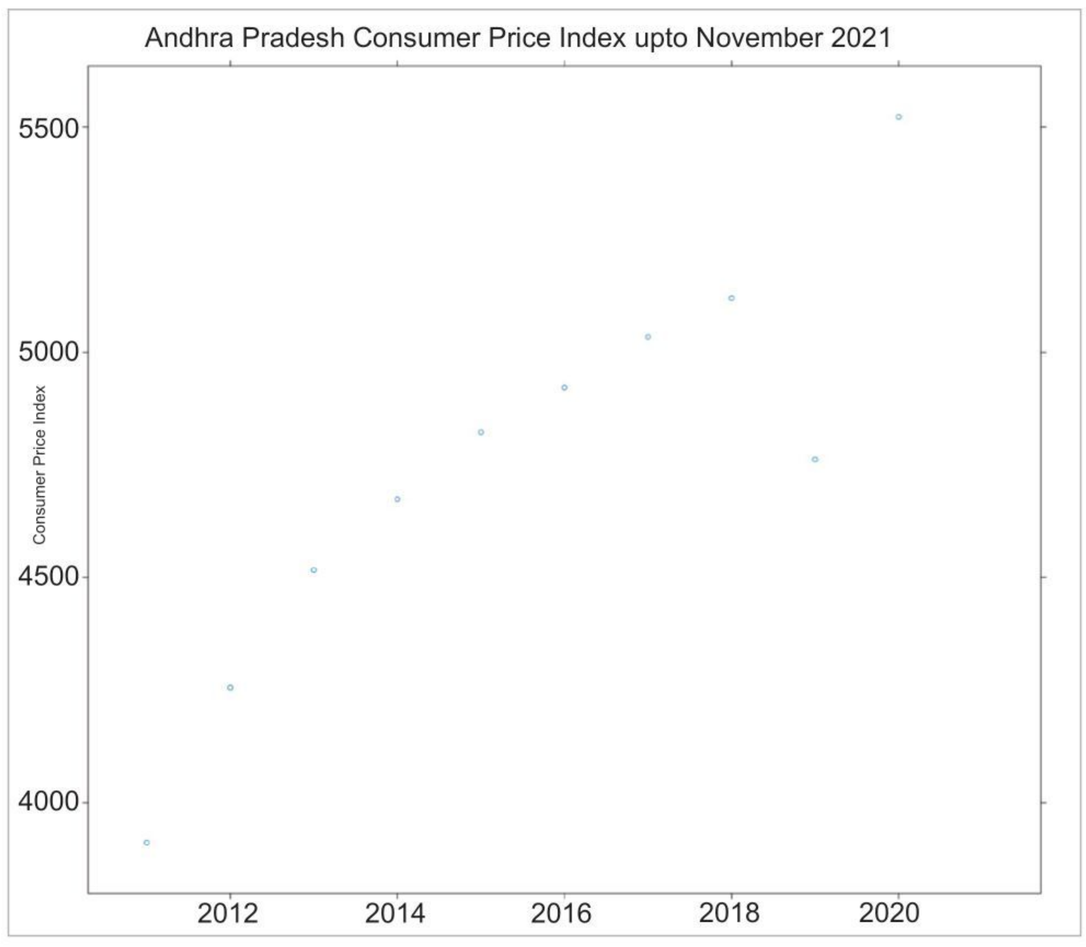


Figure 3: Scatter plot

```
> xyplot(x~Group.1|Group.1, ap, groups=Group.1, main="Andhra
Pradesh Consumer Price Index upto November 2021",
xlab="Year", ylab="Consumer Price Index", auto.key=TRUE)
```

The output chart produced by R is as shown in Figure 4.

In addition to the above listed plotting functions, lattice provides the *bwplot* function for box-and-whisker plots, and the *stripplot* function for one-dimensional scatter plots.

ggplot2

The *ggplot2* R package implements a grammar of graphics that specifies how to plot data. You can install the package using the following command:

```
> install.packages("ggplot2")
```

```
 *** installing help indices
 *** copying figures
 ** building package indices
 ** installing vignettes
 ** testing if installed package can be loaded from temporary
location
 ** testing if installed package can be loaded from final
location
 ** testing if installed package keeps a record of temporary
installation path
 * DONE (ggplot2)
```

The library needs to be loaded into the R session before you can use its functions:

```
> library(ggplot2)
```

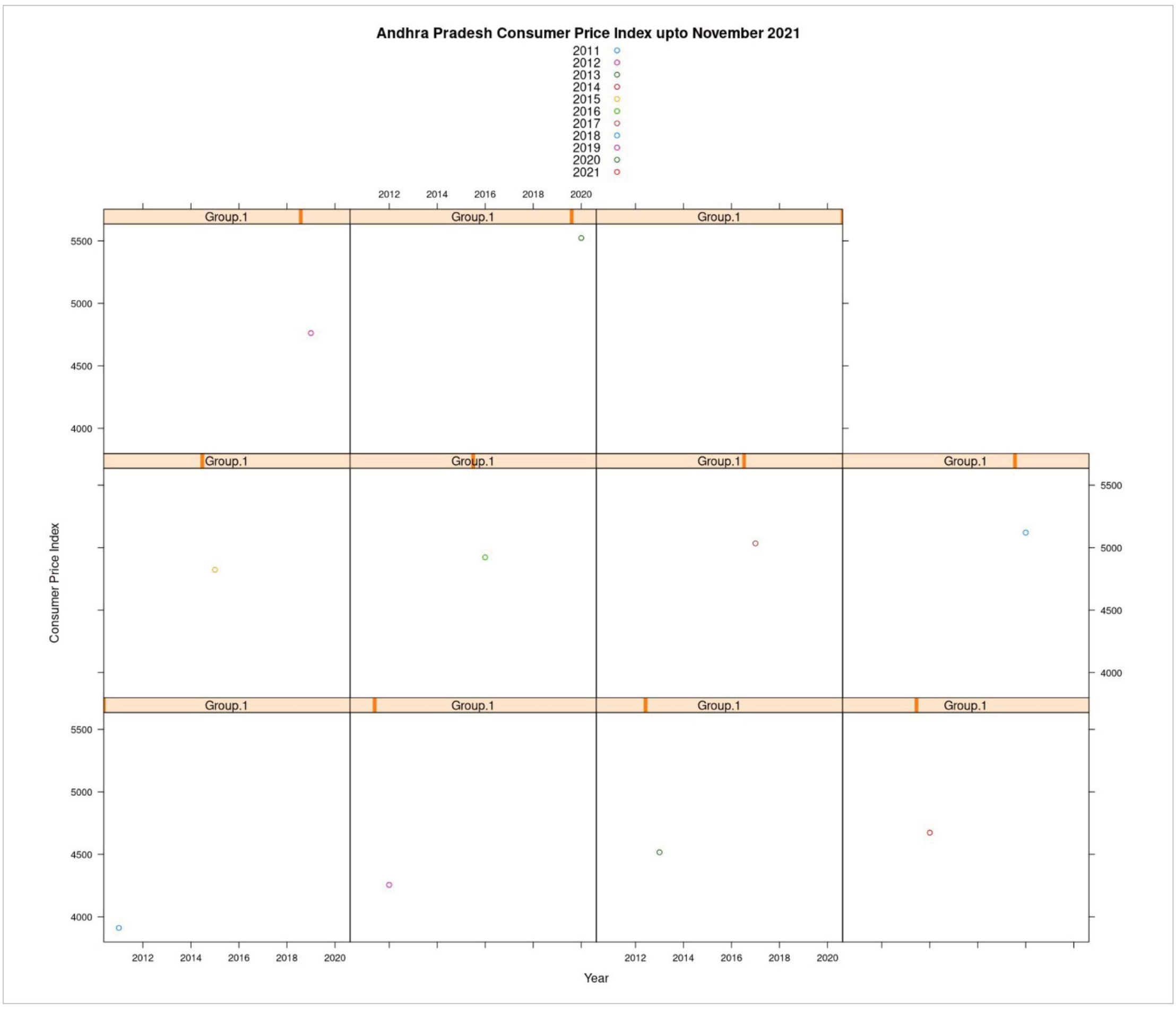


Figure 4: Grouping chart

Scatter plot

The same consumer prices (annual per cent) inflation data for India can be plotted using the quick plot or *qplot* function from the *ggplot2* package in R. For example:

```
> x<-c(1960:2020)
> y<-c(1.77,1.69,3.63,2.94,13.35,9.47,10.80,13.06,3.23,-
0.58,5.09,3.07,6.44,16.94,28.59,5.74,-
7.63,8.30,2.52,6.27,11.34,13.11,7.89,11.86,8.31,5.55,8.72,8.8
0,9.38,7.07,8.97,13.87,11.78,6.32,10.24,10.22,8.97,7.16,13.23
,4.66,4.00,3.77,4.29,3.80,3.76,4.24,5.79,6.37,8.34,10.88,11.9
8,8.85,9.31,11.06,6.64,4.90,4.94,3.32,3.94,3.72,6.62)
> d <- data.frame(x,y)
> qplot(x=x, y=y, data=d, xlab="Year", ylab="Inflation",
main="Inflation, consumer prices (annual %)")
```

The simple scatter plot is shown in Figure 5.

We can also store the results of the plot to a variable and ask R to provide a summary of the same, as shown below:

```
> ex1 <- qplot(x=x, y=y, data=d)
> summary(ex1)
data: x, y [61x2]
mapping:  x = ~x, y = ~y
faceting: <ggproto object: Class FacetNull, Facet, gg>
    compute_layout: function
    draw_back: function
    draw_front: function
    draw_labels: function
    draw_panels: function
    finish_data: function
    init_scales: function
    map_data: function
    params: list
    setup_data: function
    setup_params: function
    shrink: TRUE
    train_scales: function
```

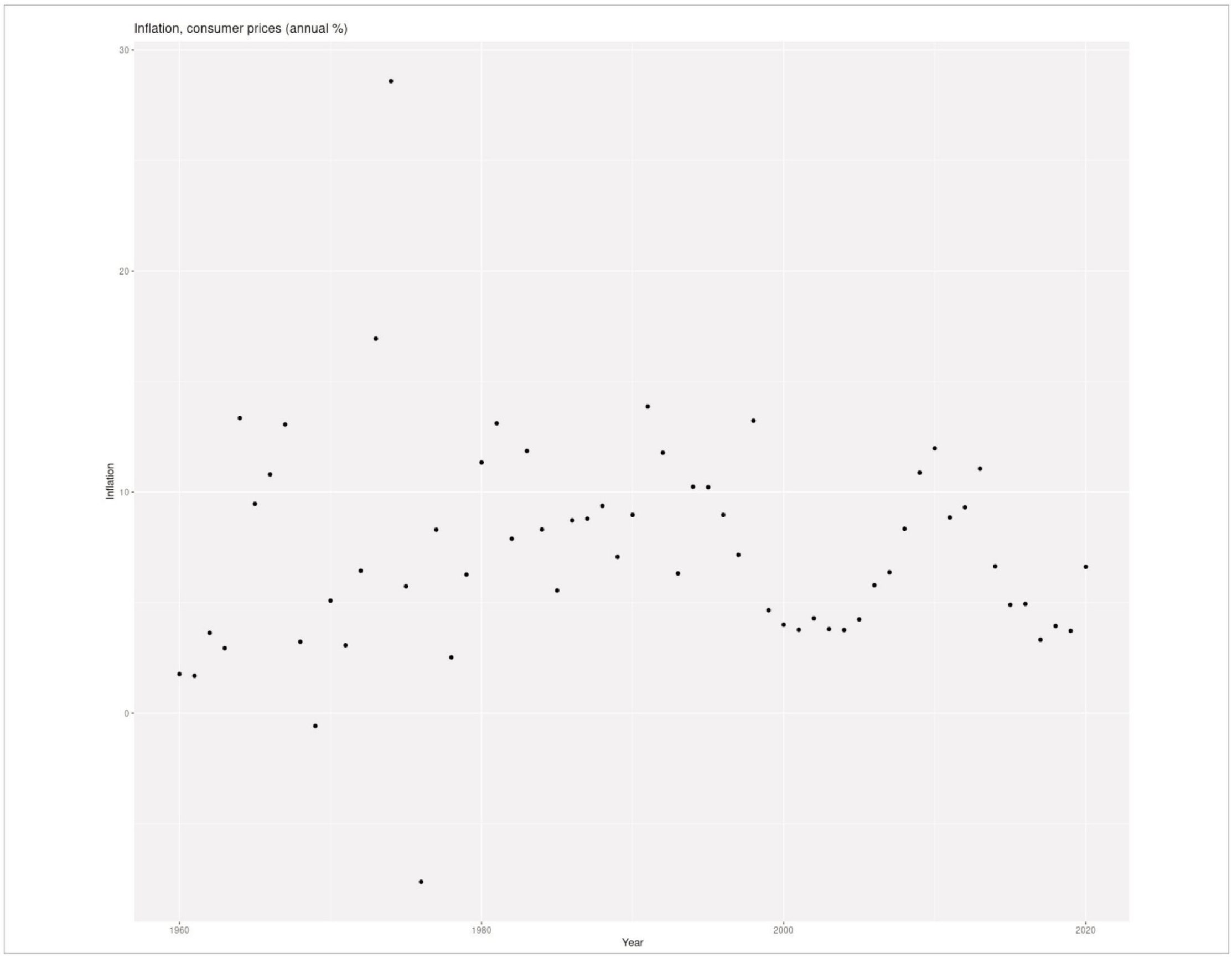


Figure 5: Simple *qplot*

```
    vars: function
    super:  <ggproto object: Class FacetNull, Facet, gg>
-----------------------------------
geom_point: na.rm = FALSE
stat_identity: na.rm = FALSE
position_identity
```

Line chart

We can generate a line chart by specifying the *geom* attribute as 'line', as shown below:

```
> qplot(x=x, y=y, data=d, xlab="Year", ylab="Inflation",
main="Inflation, consumer prices (annual %)", geom="line")
```

The corresponding line graph is shown in Figure 6.

The 'Bank Marketing Data Set' for a Portuguese banking institution is available from the UCI machine learning repository available at *https://archive.ics.uci.edu/ml/datasets/Bank+Marketing*. The data can be used for public research use.

There are four data sets available, and we will use the *read.csv()* function to import the data from a 'bank.csv' file into a data frame.

```
> bank <- read.csv(file="bank.csv", sep=";")

> bank[1:3,]
  age        job marital education default balance housing
loan  contact day
1  30 unemployed married   primary      no    1787      no
no cellular  19
2  33   services married secondary      no    4789     yes
yes cellular  11
3  35 management  single  tertiary      no    1350     yes
no cellular  16
  month duration campaign pdays previous poutcome  y
1   oct       79        1    -1        0  unknown no
2   may      220        1   339        4  failure no
3   apr      185    1  330    1 failure no
```

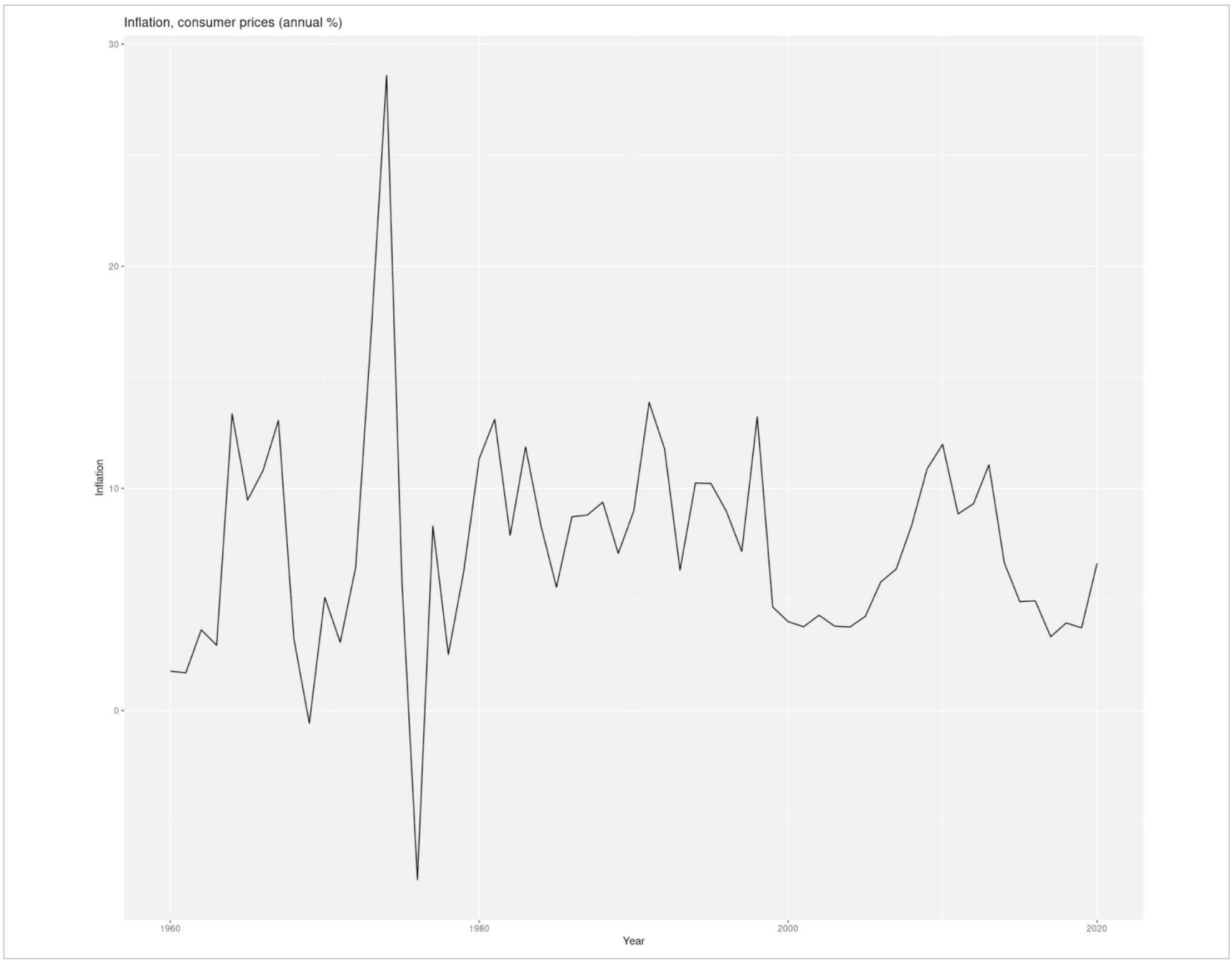


Figure 6: *qplot* line graph

Bar chart

The geometry argument can be specified as 'bar' to produce a bar chart, as indicated below:

```
> qplot(x=job, data=bank, geom="bar", weight=balance,
ylab="Balance", xlab="Category")
```

The produced bar chart is shown in Figure 7.

We can also list a summary of the chart by storing the results of the plot to a variable, and invoking the *summary* function on the same. For example:

```
> barchart <- qplot(x=job, data=bank, geom="bar",
weight=balance, ylab="Balance", xlab="Category")

> summary (barchart)
data: age, job, marital, education, default, balance,
housing, loan,
  contact, day, month, duration, campaign, pdays, previous,
poutcome, y
  [4521x17]
mapping:  x = ~job, weight = ~balance
faceting: <ggproto object: Class FacetNull, Facet, gg>
    compute_layout: function
    draw_back: function
    draw_front: function
    draw_labels: function
    draw_panels: function
    finish_data: function
    init_scales: function
    map_data: function
    params: list
    setup_data: function
    setup_params: function
    shrink: TRUE
    train_scales: function
    vars: function
    super:  <ggproto object: Class FacetNull, Facet, gg>
-----------------------------------
geom_bar: width = NULL, na.rm = FALSE, orientation = NA
stat_count: width = NULL, na.rm = FALSE, orientation = NA
position_stack
```

The *qplot* function accepts the following arguments:

Argument	Description
asp	The y/x aspect ratio
data	Optional data frame that contains x and y
geom	The geometry to use
main	The title of the chart
margin	Display margins
position	The adjustments to specify the position
x	X values
xlab	The x-axis label
xlim	The limits for the x-axis
y	Y values
ylab	The y-axis label
ylim	The limits for the y-axis

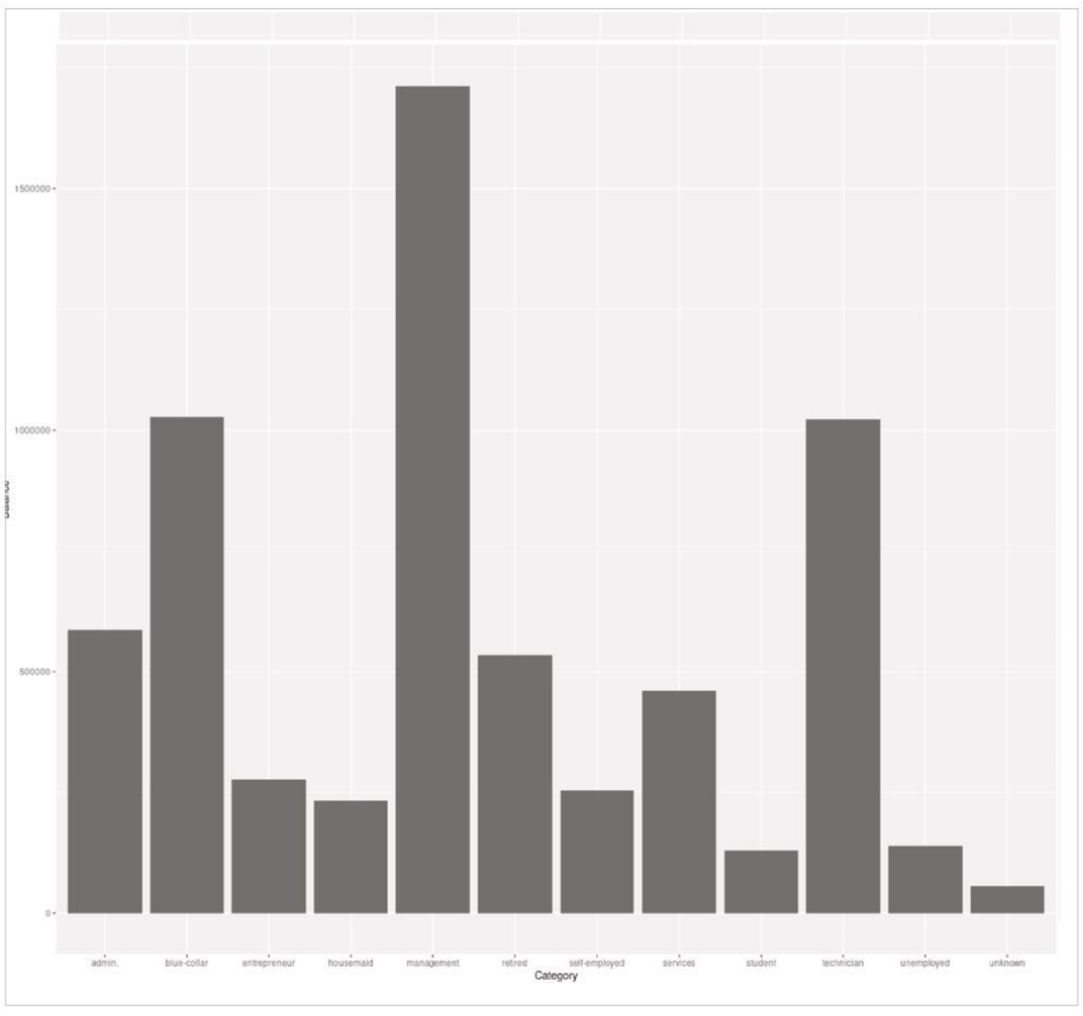


Figure 7: Bar chart

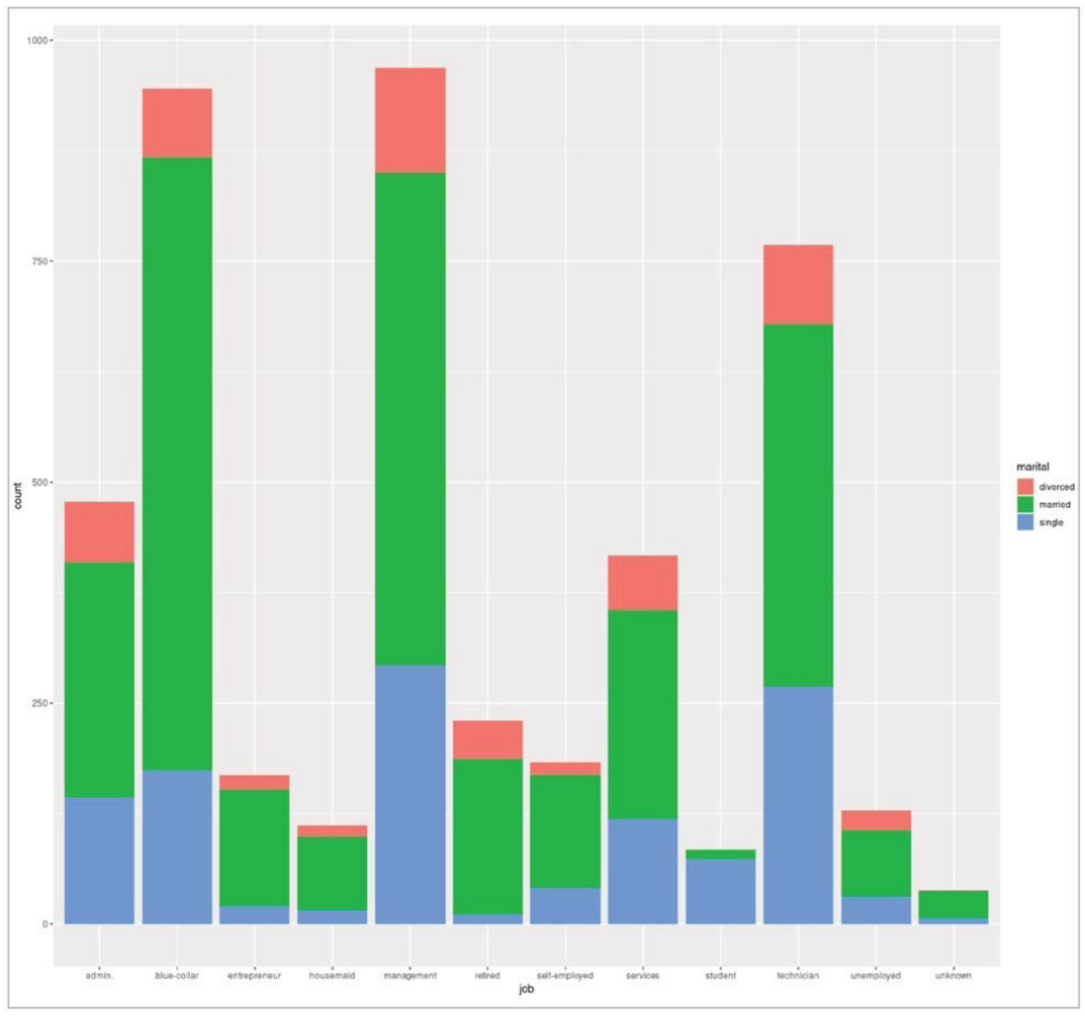


Figure 8: *ggplot* categorical graph

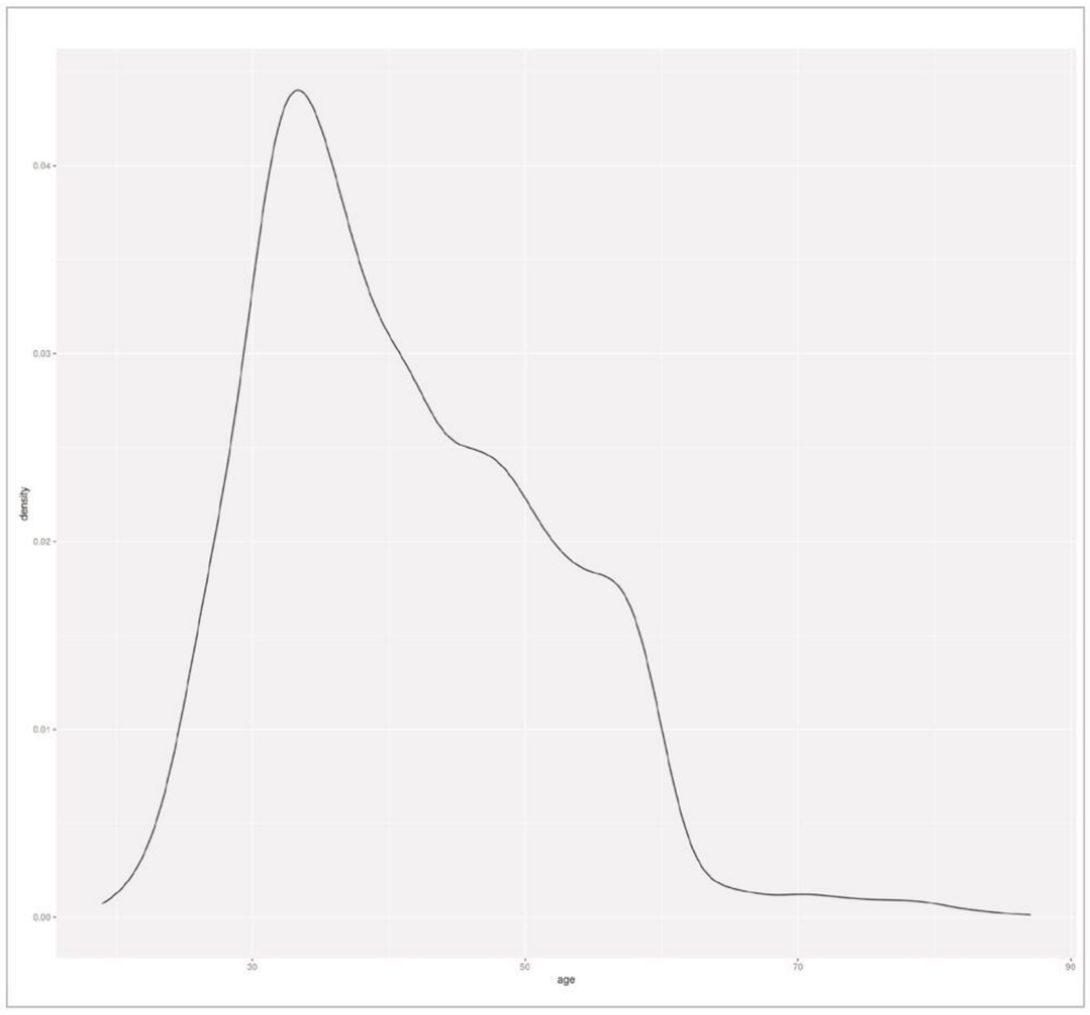


Figure 9: *ggplot* density graph

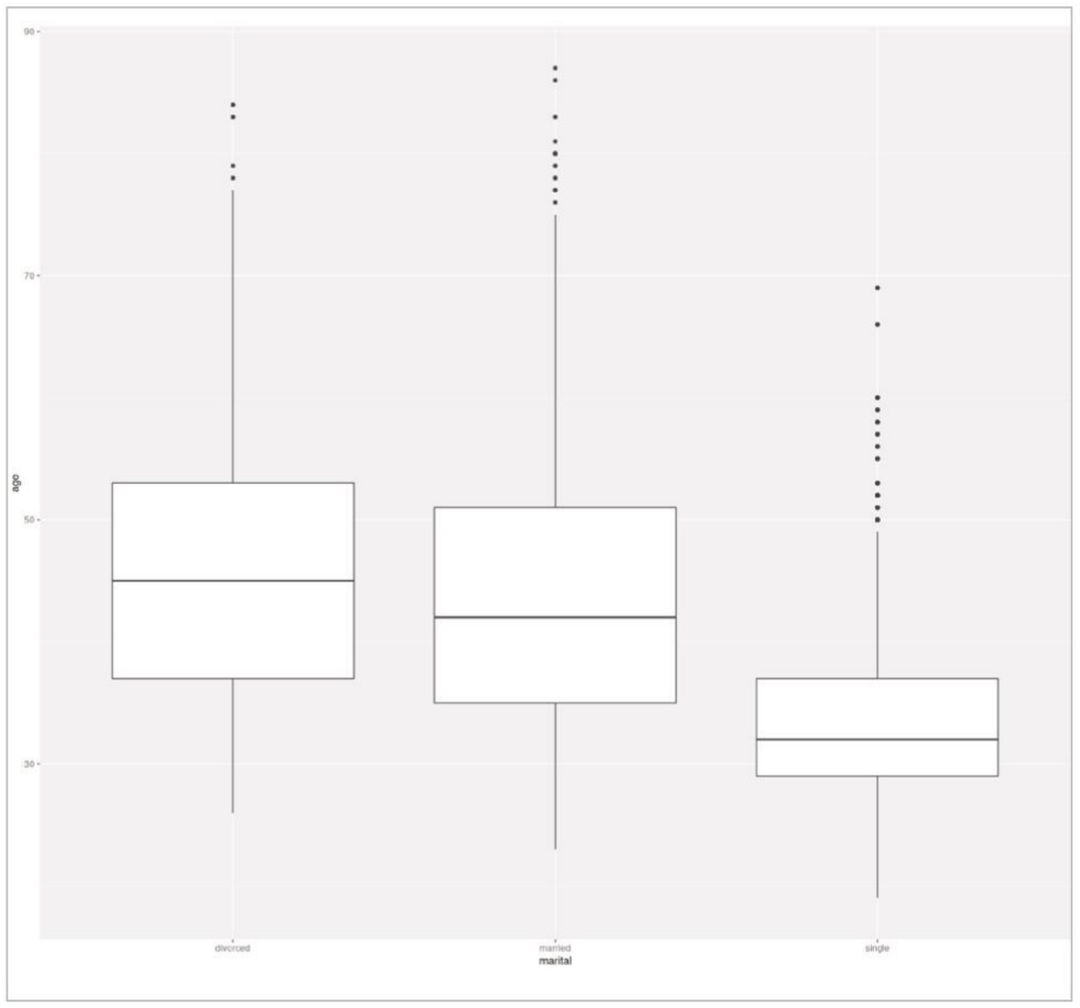


Figure 10: *ggplot* boxplot graph

ggplot

The *ggplot* function can be used to create a new *ggplot* object for input data, and also specify aesthetic mappings for the same.

For the *bank.csv* data, we can tabulate the job and marital status together using the *with* function as follows:

```
> with(bank, table(job, marital))
               marital

job             divorced married single
  admin.              69     266    143
  blue-collar         79     693    174
  entrepreneur        16     132     20
  housemaid           13      84     15
  management         119     557    293
  retired             43     176     11
  self-employed       15     127     41
  services            62     236    119
  student              0      10     74
  technician          89     411    268
  unemployed          22      75     31
  unknown              1      30      7
```

You can now plot the above categorical data using *ggplot*, as follows:

```
> ggplot(bank, aes(x = job, fill = marital)) + geom_bar()
```

The resultant graph is shown in Figure 8.

The age distribution can be plotted as a density using the *geom_density* function as follows:

```
> ggplot(bank, aes(x = age)) + geom_density()
```

The corresponding graph is shown in Figure 9.

A box plot for the age and marital status can be visualised using the following arguments to *ggplot:*

```
> ggplot(bank, aes(x = age, y = marital)) + geom_boxplot() +
coord_flip()
```

The output graph is as shown in Figure 10.

The *ggplot* function accepts the following arguments:

Argument	Description
data	The data frame for the plot
mapping	The aesthetic mappings to be used in the plot
environment	The *globalenv()* environment for the aesthetics

Do try and explore more functions and charts in the graphics packages available in R. END

References

[1] India Inflation, consumer prices (annual %), *https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG?locations=IN*

[2] [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

By: Shakthi Kannan

The author is a free software enthusiast.

An Introduction to teradata.



Teradata is one of the most popular relational database management systems (RDBMS) appropriate for huge data warehousing applications. It is highly scalable and capable of handling large volumes of data. This article provides a basic understanding of Teradata architecture, and the types of spaces and tables it has.

As per official Teradata documentation: "Teradata database is a massively parallel analytics engine that can help businesses achieve breakthrough results across all industries. Whether companies need to improve profitability, mitigate risks, innovate products, enhance the customer experience, or achieve other objectives, analytics paves the way to success."

Teradata architecture

Node: This is an essential building block of the Teradata database system and is made up of hardware and software components. A server can also be considered as a node.

PE (parsing engine): This is a type of vproc (virtual processor) for session control, task dispatching and SQL parsing in the multi-tasking and possibly parallel-processing environment of the Teradata database. Vproc is a software process running in an SMP (symmetric multiprocessing) environment or a node. The vproc or

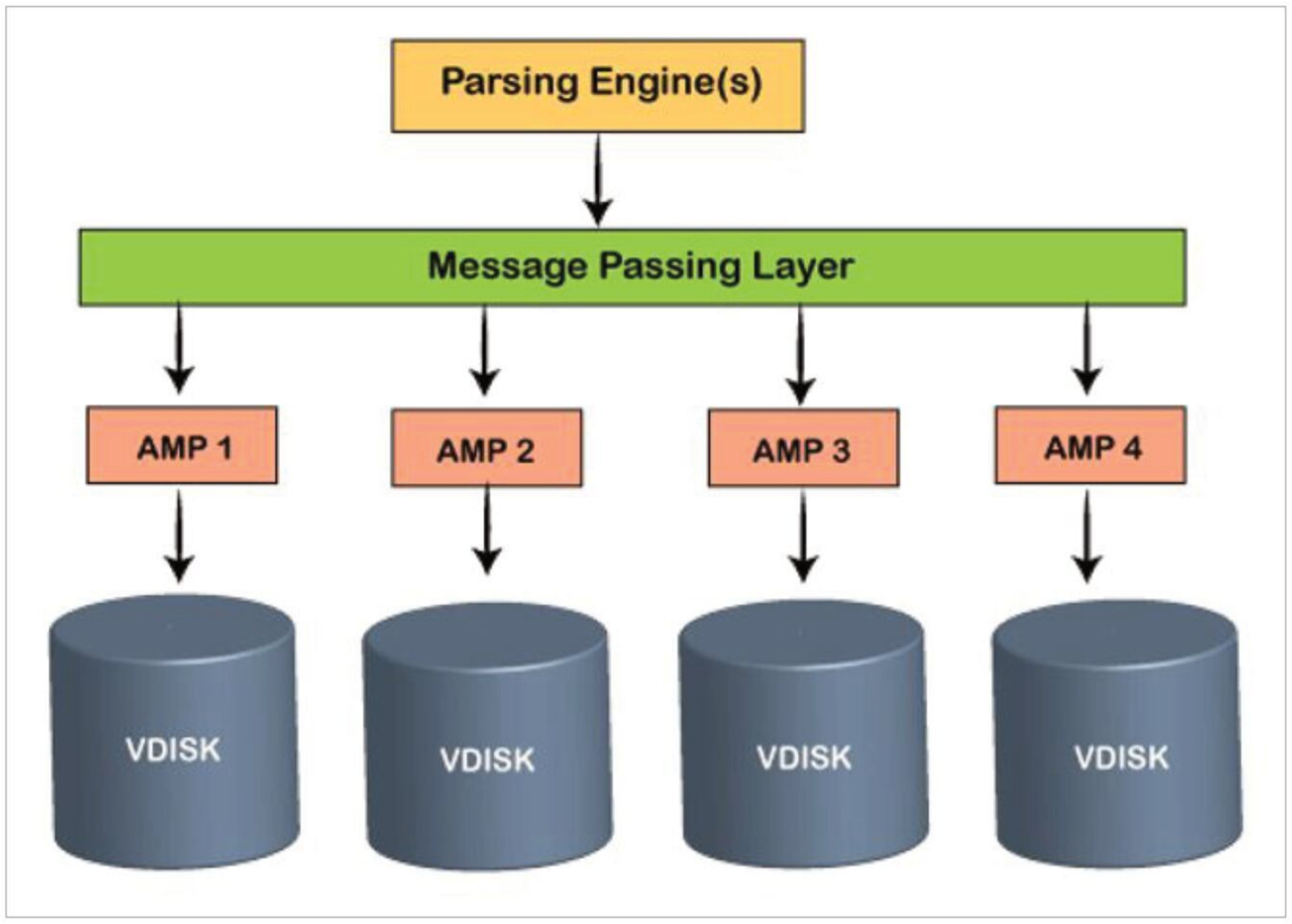


Figure 1: Teradata architecture

the access module processor (AMP) are responsible for storing and retrieving data.

The different components of a parsing engine are listed below.

- *Parser:* This decomposes SQL queries into relational data management processing steps.
- *Query optimizer:* This determines the most efficient path to access data.
- *Step generator:* This produces processing steps and encapsulates them into packages.
- *Dispatcher:* This transmits the encapsulated steps from the parser to the relevant AMPs, and performs monitoring and error-handling functionalities during step processing.
- *Session controller:* This manipulates session-control activities (e.g., log on, log off and authentication) and restores a session only after client or server failures.

Message passing layer: This is also called BYNET (Banyan network), and is the networking layer in the Teradata system. It allows interaction between PE and AMP, and also between the nodes. It receives the execution plan from the PE and sends it to AMP. Similarly, it receives the outputs from the AMP and sends them to PE. In the Teradata system, there are always two BYNET systems. These are called BYNET 0 and BYNET 1. But this is generally referred to as a single BYNET system. If one BYNET fails, the second one takes its place. But when data is large, both BYNETs can be made functional, which enhances the communication between PE and AMPs.

Disk: This component organises and stores the data while performing data manipulation activities. Teradata database employs redundant array of independent disks (RAID) storage technology to provide data security at the disk level. When AMPs are associated with disks, it is called Vdisks (virtual disks).

Note: Each AMP is allowed to read and write on its own disk only. That is why it is also called shared-nothing architecture.

Teradata acts as a single data store that receives a large number of concurrent requests from various client applications. It is defined as an information repository supported by tools and utilities that make it, as part of Teradata Warehouse, a complete and active RDBMS.

Types of spaces in Teradata

Permanent space: This is the maximum amount of space allocated to the user/database to hold data rows. It is used for database objects (like permanent tables, indexes, etc) creation, and to hold their data. The total permanent space is divided among the total number of AMPs. Whenever the per AMP limit exceeds the allocated space of AMP, a 'No more room in database' error is generated. Permanent tables, journals, fallback tables, and secondary index sub-tables use permanent space. All databases have a predefined upper limit of permanent space. Teradata doesn't physically preallocate permanent space for databases and users when they are defined during object definition time.

Spool space: This is the unused permanent space that is used by the system to keep the intermediate results of the SQL query. Once the query is complete, the spool space is released. The volatile tables use the spool space. Users without spool space can't execute any query. Data is active up to the current session only. It is divided among the number of AMPs. Whenever the per AMP limit exceeds the allocated space, the user will get a spool space error.

Temporary space: This is the unused permanent space used by global temporary tables (GTT). It is divided by the number of AMPs. Data is active up to the current session only. It is reserved prior to spool space for any user defined operations. Temp space is allocated at the database level or user level, but not at the table level. This is the amount of space used for GTT, and these results

remain available to the user until the session is terminated.

Tables created in temp space will survive a restart. A query example is given below:

```
CREATE DATABASE new_db FROM existing_db
AS PERMANENT = 2000000, SPOOL =
5000000, TEMP = 2000000;
```

A new database must be created from an existing database. Permanent space can be used for tables in the database. Spool space is allocated for the maximum amount of workspace available for requests. Temp space is allocated for temp tables in the database.

Types of tables in Teradata

Permanent tables: These remain in the system until they are dropped. The table definition is stored in the data dictionary. Data and structure can be shared across multiple sessions and users. Data is stored in the permanent space, and 'collect statistics' is supported. Indexes can be created. COMPRESS columns, as well as DEFAULT and TITLE clauses are supported. Partition primary index (PPI) is also supported. If the primary index clause is not defined in the 'Create table' statement, then Teradata will create the first column as primary by default.

Global temporary tables (GTT): This is a kind of temporary table. The table definition is stored in the data dictionary. It requires a minimum of 512 bytes from the permanent space to store table definitions in the data dictionary. The structure can be shared across multiple users but data will remain private to the session (and its user). Data is stored in the temporary space. Data will be purged for that session once the session ends. 'Collect statistics' is supported. Indexes can be created. COMPRESS columns, as well as DEFAULT and TITLE clauses are supported. PPI is also supported.

This table can be identified from the data dictionary table *(dbc.tables)* using the 'CommitOpt' column. If its value is 'D' (on *commit Delete* rows) or 'P' (on *commit Preserve* rows), then it's a GTT. If the primary index clause is not specified while creating the table, then Teradata will create the first column as primary by default. One session can generate up to 2000 global temporary tables at a time.

Volatile tables: These are created in the spool space for temporary use, and their life span extends for the duration of the session only. The table definition is not stored in the data dictionary. Structure and data is private to the session and its user. Data is stored in the spool space. The table gets dropped once the session ends, i.e., when 'login again…no volatile tables' shows up. 'Collect statistics' is supported. Indexes cannot be created. The COMPRESS column is supported, but the DEFAULT and TITLE clauses are not supported. PPI is supported. If the primary index clause is not specified while creating the table, then Teradata will create the first column as primary by default.

```
ON COMIT DELETE ROWS (Default)
ON COMIT PRESERVE ROWS (If we want to
retain rows)
```

Derived tables: This temporary table is derived from one or more other tables as the result of a sub-query. Derived tables are local to the query and exist only for the duration of the query. The table is automatically deleted once the query is done. Data is stored in the spool space. Table definition is not stored in the data dictionary. The first column in a derived table acts like a PI column for it.

Properties of Teradata tables

- ***Fallback:*** This protects data by storing a second copy of each row of a table on a different AMP in the same cluster. If an AMP fails, the system accesses the fallback rows to meet requests.
- ***Journals:*** BEFORE JOURNAL holds the image of impacted rows before any changes are made. AFTER JOURNAL holds the image of affected rows after changes are done. In DUAL BEFORE/AFTER JOURNAL, two images are taken and are stored in two separate AMPs.
- ***Checksum:*** This detects some forms of lost writes. A lost write is a write of a file system block that received successful status on completion, but the write either never actually occurred or was written to an incorrect location. As a result, subsequent reads of the same block return the old data. NONE, LOW, MEDIUM, and HIGH levels are available in the checksum.

Teradata is a powerful RDBMS and can help in transforming how businesses work. We have just discussed its basics in this article and a lot still needs to be unearthed. We hope you find this basic information useful to explore the features and capabilities of Teradata further. END

By: Saurabh Kumar and Surabhi Kumari

Saurabh Kumar is a senior data architect in the data science division of American Express. He has worked in several domains like MIS, data analytics, business intelligence, fraud detection, and data anonymization.

Surabhi Kumari is well experienced in the field of IT. She is a quality analyst at a leading IT company. Her areas of interest are software testing with automation, APIs, and SQL.

Manual Renewal of SSL Certificates: A Simple Guide



Image source:
http://www.freepik.com

In the April 2022 issue of this magazine, we read about the different types of SSL certificates and their applications. This article explains how to manually renew an existing SSL certificate, so that it stays updated with the latest security requirements.

When visitors interact with your website and share critical details like credit card numbers, they trust that their information has been secured against misuse. So it becomes your responsibility to respect that trust and offer complete protection to all the visitors on your site. Failing to do so can not only cost you the loyalty of customers but may also put you in a legal soup. There have been many instances where websites that couldn't protect their customers' data against leakage, theft or misuse were forced to pay hefty penalties, and also lost their reputation.

How does an SSL certificate secure customers' data?

One of the best ways to protect sensitive customer information is to secure your site with an SSL (secure sockets layer) certificate. Without

going into the technical nitty-gritties, an SSL certificate encrypts the communication between a Web server and the visitor's browser, and thus makes it technically impossible for hackers or threat actors to steal data in transit. SSL establishes a secure handshake process to decrypt the encrypted information — a process that is too complex to be cracked by humans or even software.

Why do you need to update your SSL certificate?

While an SSL certificate does offer security against data theft or misuse, you need to update it periodically to ensure the most effective security against the latest threats. This article will list the step-by-step instructions for renewing your SSL certificate in the right way.

There are quite a few benefits of updating an SSL certificate:

- Timely renewal authenticates your website's identity.
- You get updated security.
- A one-year validity promotes the healthy practice of periodically renewing/upgrading the scope of protection, thus eliminating the risks associated with outdated versions.

Note: The best practice is to select an auto renewal option that relieves you from the stress of remembering the renewal date or manually following the related steps.

A bit off topic …a pure open source way to build your own SSL certificate

Yes, it is absolutely true! With some simplified and compact steps you can actually build your own SSL certificate from scratch! While the entire process is out of the scope of this article, here are a few key open source components and tools that you can use to create your SSL certificate.

- **OpenSSL:** This is a highly credible tool to implement TLS and crypto libraries.
- **EasyRSA:** This command-line tool enables you to build PKI CA and manage it efficiently.
- **CFSSL:** Cloudflare has finally built a multi-purpose, multi-capability tool for PKI and TLS.
- **Lemur:** A fair enough TLS producer developed by Netflix.

How to renew your SSL certificate

While the broad process of SSL renewal remains the same, there could be some minor tweaks and variations depending upon your specific SSL provider.

The renewal process follows three main steps — CSR (certificate signing request) generation, certificate activation and, finally, the certificate installation.

Generating CSR: For cPanel hosting panels, you can click on the *Security* tab and search for the *SSL/TLS* option. It will display a screen that shows a link just under the *CSR* option. This section facilitates new CSR generation for your desired domain name.

You will be asked detailed contact information for confirming that you are the genuine domain owner. After filling the form, you will get a CSR code that is needed for certificate reactivation.

Activating an SSL certificate: In your dashboard, you can see a quick overview of the SSL certificate, domains and other digital infrastructure products that you own. Click this button to start the SSL renewal process. Enter the CSR generated a while ago and confirm the accuracy of the information. You can now validate the SSL renewal process.

Validating the SSL certificate: You will once again be prompted to confirm domain ownership. Enter your domain-associated email. Upload a file on the Web server where the certificate needs to be installed. Validate the SSL certificate with the help of CNAME records. While there are multiple options, it is best and easiest to validate it with an email. Once you enter the domain associated email, you will get an email containing a specific link followed by another mail that comprises the new certificate file with a *.crt* extension.

Installing the SSL certificate: Your Web hosting provider will either give you an option for communicating with the support team for installing renewed files or provide you with the detailed instructions on how you can do it manually through your cPanel. Keep in mind that different hosts offer different options for renewal. That said, if you are a non-technical person then contacting the support team will be the best option for you.

For manual updating, visit the *SSL/TLS* section of your cPanel and find the *Manage SSL sites* option. It contains the entire domain list that you own. Corresponding to each domain name you can see the *Certificate Renewal* option.

In the screen next to it, enter the details in *Private Key* using the *Autofill by Domain* option. Under the *Certificate* option, fill the details of your *.crt* file. You are almost done. Just click the button that reads *Install Certificate.*

Along with saving the critical data and sensitive information of your users, the SSL certificate also builds trust by reaffirming that data shared on your site is secure. It can also affect your SEO profile positively as Google considers SSL certification a healthy practice. However, to continue enjoying the best security with this certificate, you need to update it periodically. This ensures that your site is fully secured against the data in transit attacks as per the latest security requirements. END

By: Jitendra Bhojwani

The author is an open source enthusiast and a technical blogger.

Analyse Logs with Elasticsearch



We often need to search and analyse the logs of our Windows based computer or server to get an insight into the system or application. Such analysis gets easier if the logs are sent to Elasticsearch. This article is a step-by-step guide on how to integrate Winlogbeat and Sysmon with Elasticsearch.

Winlogbeat is a tool that runs as a service on a MS Windows based computer and is capable of sending Windows logs to Elasticsearch for analysis. It's an open source and free tool to keep a tab on what's happening in your Windows based system. It streams Windows event logs to Elasticsearch and Logstash.

Installation of Winlogbeat

For installing Winlogbeat on a Windows computer, follow the steps given below.

Step 1: Go to the download link and download the Winlogbeat zip file by selecting your system architecture. The link is *https://www.elastic.co/downloads/beats/winlogbeat/*.

Step 2: Once the download process is done, extract the contents in the zip file to *C:\Program Files*. Rename the folder as *Winlogbeat* for later use.

Step 3: Now open PowerShell as *Administrator* and navigate to the directory *C:\Program Files\winlogbeat*.

Step 4: Enter the following command in PowerShell:

```
.\install-service-winlogbeat.ps1
```

The output of the above command should look like what's shown in Figure 1.

If you get an error like 'Script execution is disabled', then run the following command:

```
PowerShell.exe -ExecutionPolicy UnRestricted -File .\install-service-winlogbeat.ps1
```

```
PS C:\Program Files\winlogbeat> powershell.exe -ExecutionPolicy UnRestricted -File .\install-service-winlogbeat.ps1

Security warning
Run only scripts that you trust. While scripts from the internet can be useful, this script can potentially harm your
computer. If you trust this script, use the Unblock-File cmdlet to allow the script to run without this warning
message. Do you want to run C:\Program Files\winlogbeat\install-service-winlogbeat.ps1?
[D] Do not run  [R] Run once  [S] Suspend  [?] Help (default is "D"): R

Status   Name               DisplayName
------   ----               -----------
Stopped  winlogbeat         winlogbeat
```

Figure 1: Service installation in PowerShell

Configuring Winlogbeat

Once the installation process is completed, we need to configure the *config* file (*.yml* file) to send the captured logs from Windows to Elasticsearch. Open the configuration file, *winlogbeat.yml*, which is present in the directory where the zip file is extracted.

In that file, scroll down to the *output elasticsearch* section and uncomment that part (*output.elasticsearch*, hosts, user name and password).

Make sure your host is set to what you configured while configuring Elasticsearch. By default, it is *localhost:9200*. Enter the user name and password of the 'elastic' user, which you might have generated after installing Elasticsearch.

After configuring, the output should look similar to what is shown in Figure 2.

Now we need to configure the *setup.kibana* part in the same file. Uncomment that part, and make sure the host is set properly to the one you configured while installing Kibana. For getting all the logs in a neat visualised format, make sure *setup.dashboards.enabled* is set to true in the same file. Refer to Figure 3.

Next, we will modify the *winlogbeat.event_logs* section in the same file. We need to mention which services are to be monitored. By default, it has three services — application, security and system. You can leave it as default or you can add any if needed according to your use case.

Once everything is done, run the command given below in PowerShell (same directory) to save the configurations:

```
.\winlogbeat.exe test config -c .\
winlogbeat.yml -e
```

Once you have successfully executed the command to save, load the assets for data visualisation, and run the command:

```
.\winlogbeat.exe setup -e
```

```
# ---------------------------- Elasticsearch Output ----------------------------
output.elasticsearch:
  # Array of hosts to connect to.
  hosts: ["localhost:9200"]

  # Protocol - either `http` (default) or `https`.
  #protocol: "https"

  # Authentication credentials - either API key or username/password.
  #api_key: "id:api_key"
  username: "[redacted]"
  password: "[redacted]"
```

Figure 2: Output of Elasticsearch configuration

```
# ================================== Dashboards ==================================
# These settings control loading the sample dashboards to the Kibana index. Loading
# the dashboards is disabled by default and can be enabled either by setting the
# options here or by using the `setup` command.
setup.dashboards.enabled: true

# The URL from where to download the dashboards archive. By default this URL
# has a value which is computed based on the Beat name and version. For released
# versions, this URL points to the dashboard archive on the artifacts.elastic.co
# website.
#setup.dashboards.url:

# =================================== Kibana ===================================

# Starting with Beats version 6.0.0, the dashboards are loaded via the Kibana API.
# This requires a Kibana endpoint configuration.
setup.kibana:

  # Kibana Host
  # Scheme and port can be left out and will be set to the default (http and 5601)
  # In case you specify and additional path, the scheme is required: http://localhost:5601/path
  # IPv6 addresses should always be defined as: https://[2001:db8::1]:5601
  host: "localhost:5601"

  # Kibana Space ID
  # ID of the Kibana Space into which the dashboards should be loaded. By default,
  # the Default Space will be used.
  #space.id:
```

Figure 3: Kibana dashboard configuration

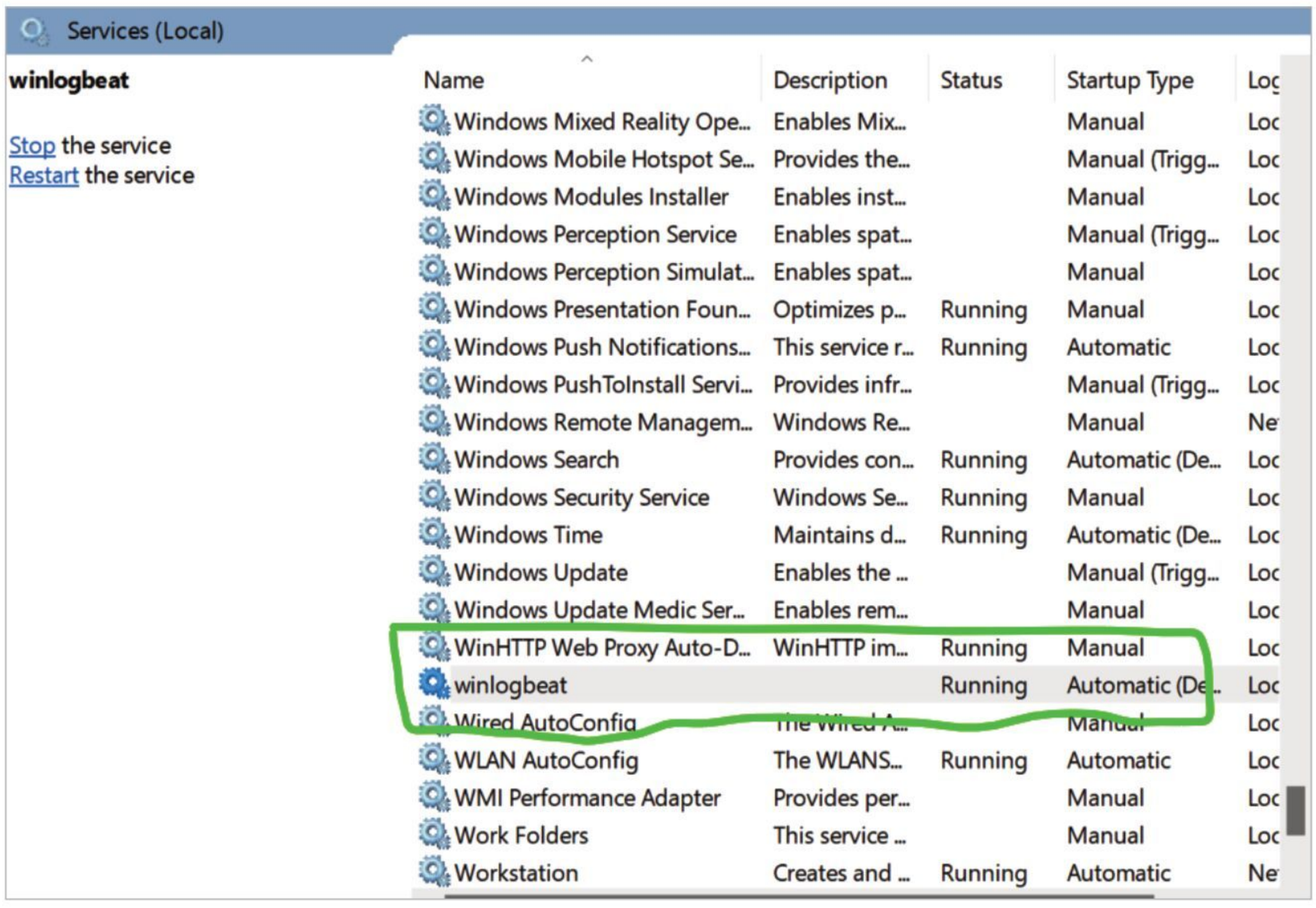


Figure 4: Winlogbeat service status

This completes the setup part. It's time to start the *winlogbeat*. Use the following command to start Winlogbeat (in the same directory in PowerShell):

```
Start-Service winlogbeat
```

To check whether the service is running or not, start searching for services. This will prompt a window; scroll down till you find Winlogbeat. Make sure it is running. If the above configuration is done properly, then the service should be running (refer to Figure 4).

Note: The Winlogbeat service will start running only if Elasticsearch and Kibana are running.

Login to your Kibana dashboard, and in the *Dashboard* tab, select *Winlogbeat*. Once the dashboard is loaded, you can see all the windows events and logs in a visualised format (refer to Figure 5).

This completes our integration of Windows logs with Elasticsearch. Next, we will see how to configure Sysmon.

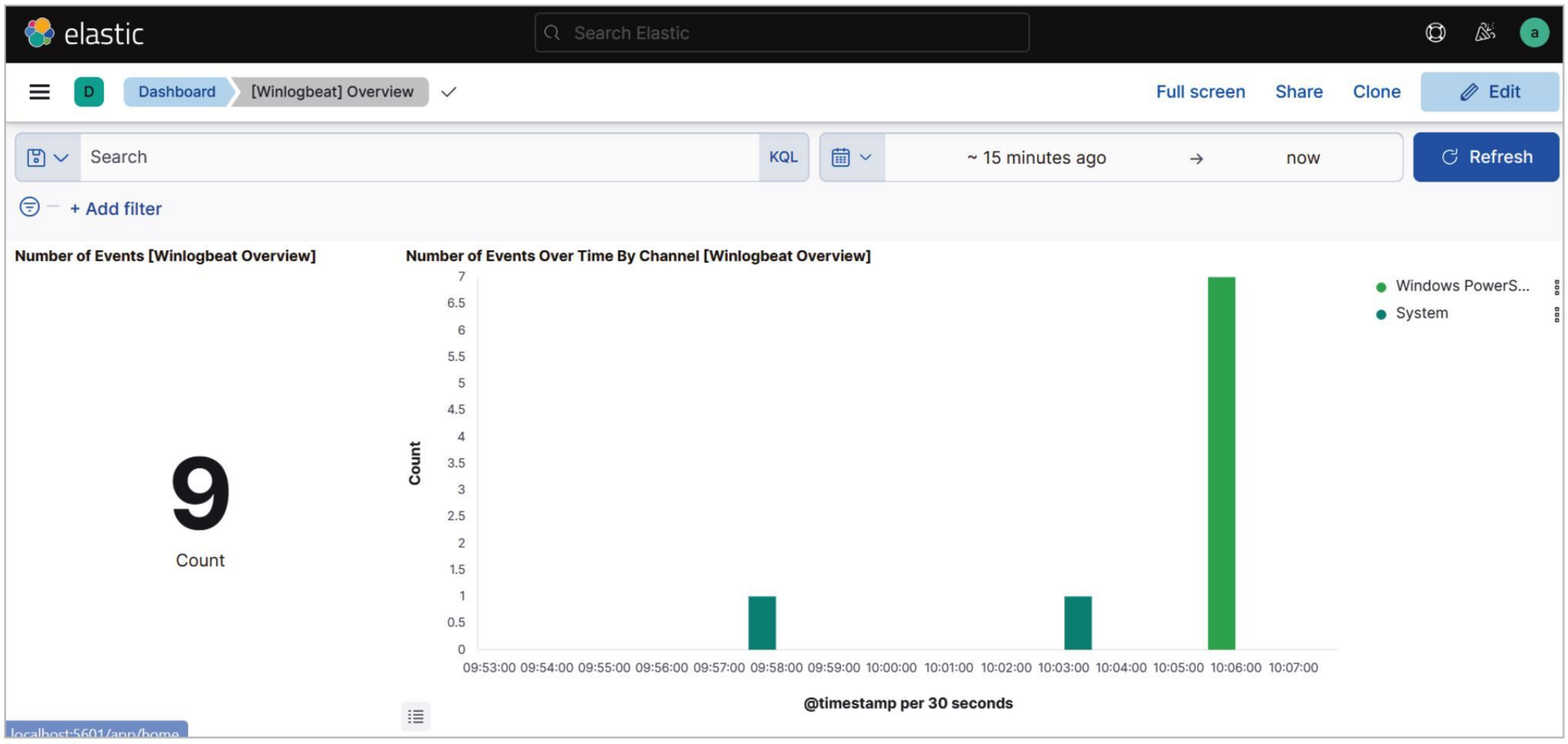


Figure 5: Winlogbeat dashboard overview

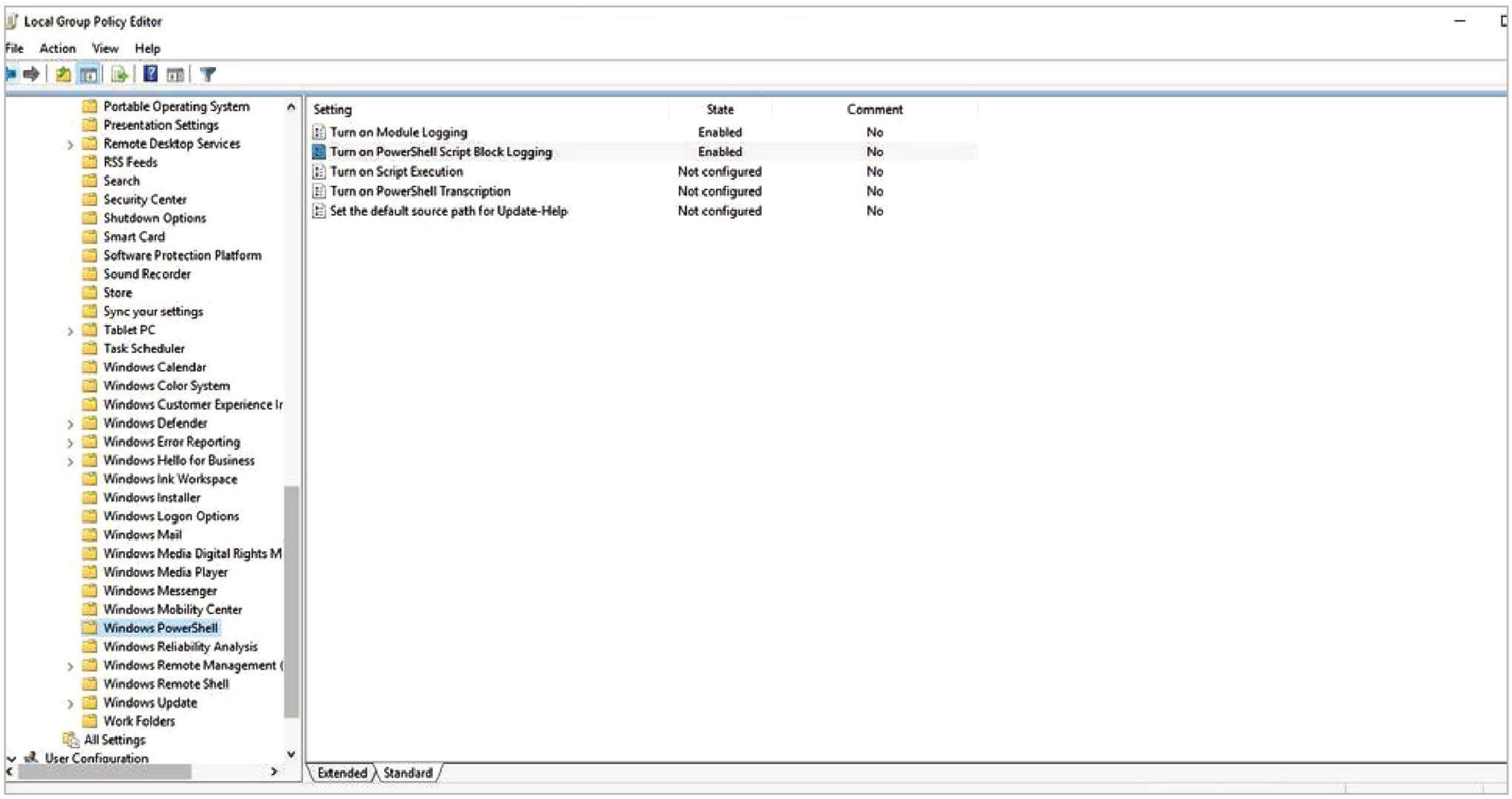


Figure 6: Local group policy settings

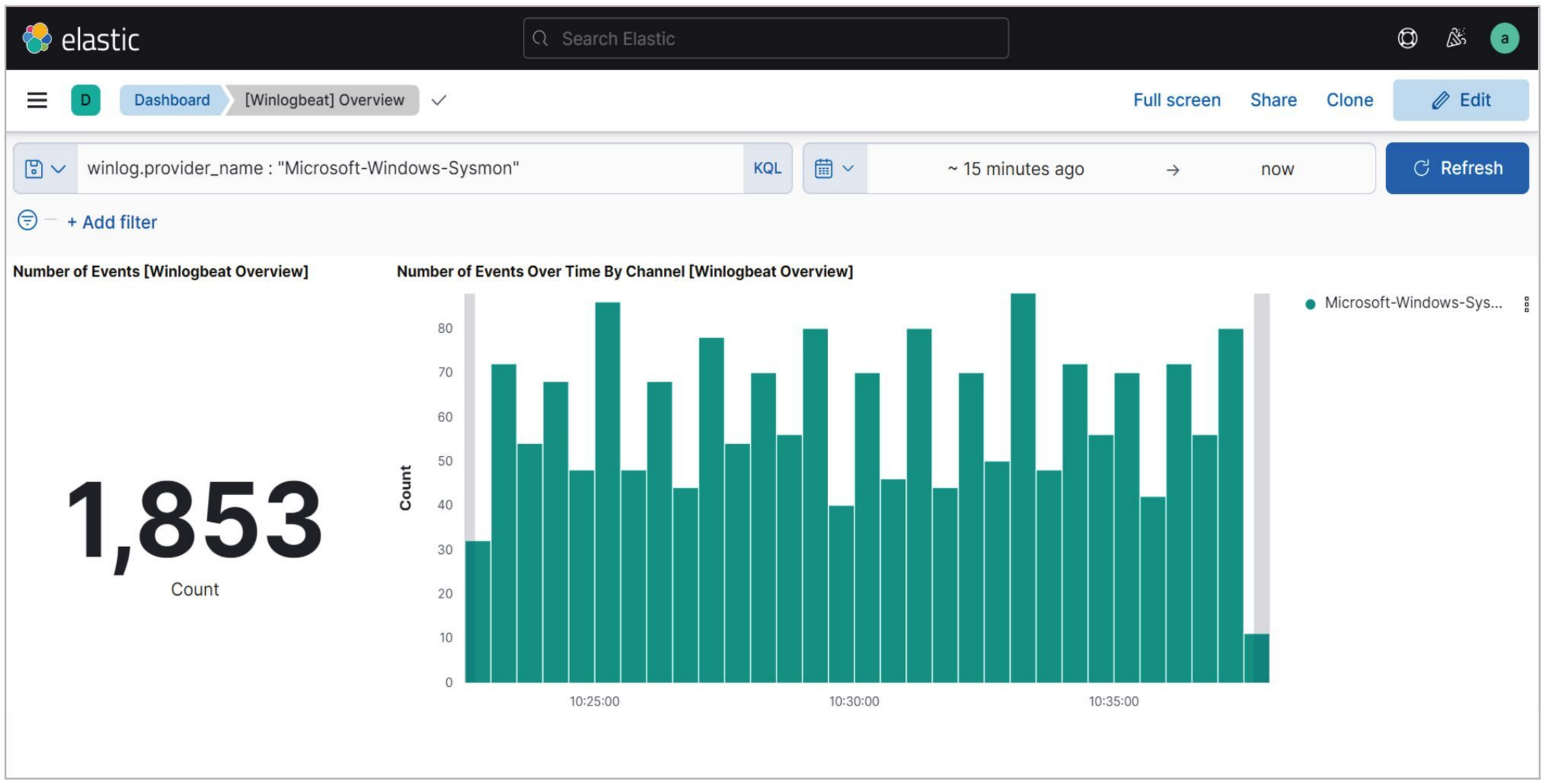


Figure 7: Sysmon logs in Kibana

Sysmon

System Monitor (Sysmon) is a Windows service. Once installed and configured it remains persistent, and logs system activity to the Windows event log. This module provides very detailed information about process creation, network connections and any file changes.

Installation of Sysmon

For complete setup of Sysmon, two things must be installed. One is a Sysmon tool available in the zip file and the other is the configuration file from GitHub. The download links are given below:

- *https://docs.microsoft.com/en-us/sysinternals/downloads/Sysmon*
- *https://github.com/SwiftOnSecurity/Sysmon-config*

Create a new folder in *C:/Program Files/*. Extract both the files in that folder. Now open PowerShell as administrator and navigate to that directory. Run the following command:

```
.\sysmon64.exe -accepteula -i
sysmonconfig-export.xml
```

Open *Local Group Policy* (search in S*tart*), and navigate to *Computer Configuration* > *Policies* > *Administrative Templates* > *Windows Components* > *Windows PowerShell*.

On the right side, enable 'Turn on Module Logging' and 'Turn on PowerShell Script Block Logging' (refer to Figure 6).

Now we need to update Winlogbeat so that it ships the logs from Sysmon too. For that, once again open the *winlogbeat.yml* file. Under the *winlogbeat.event_logs* section, make sure the following are added:

- *Microsoft-Windows-PowerShell/Operational*
- *Microsoft-Windows-Sysmon/Operational*

Once this is done, let's check whether the logs are being sent to Kibana by Winlogbeat or not. Enter the following search query to check:

.

```
winlog.provider_name : "Microsoft-
Windows-Sysmon"
```

The result of this query gives the details of logs sent by Sysmon (refer to Figure 7). If you didn't find any logs, try to increase the time period or generate some traffic.

This is how we can install and configure Winlogbeat and Sysmon to stream logs to Elasticsearch and visualise the data. Depending on your use cases, you can add the name of the event logs to be sent in *winlogbeat.yml* file and cutomise the logs visualisation as per your requirement. END

Acknowledgements

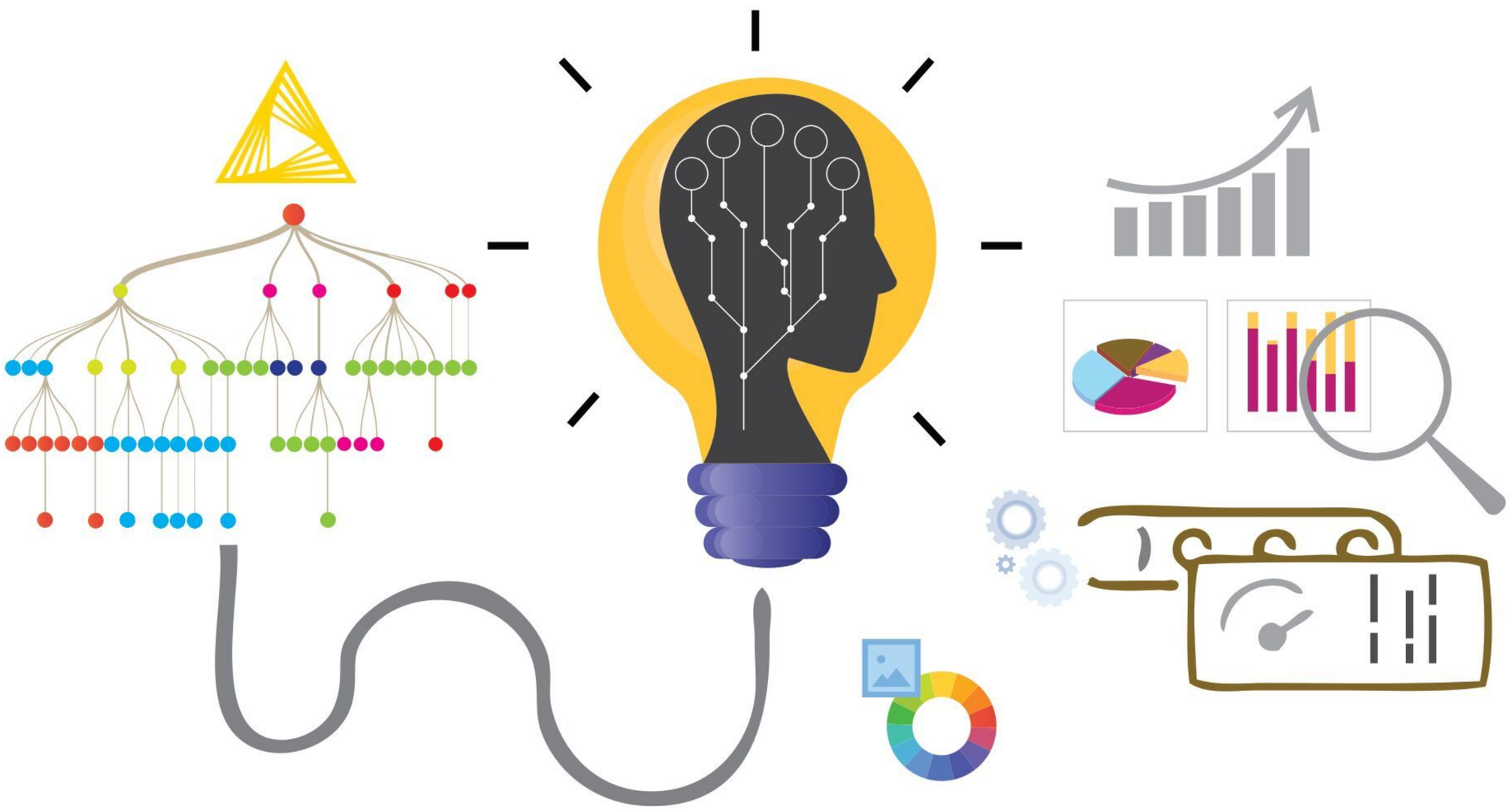
The authors are grateful to Sibi Chakkaravarthy Sethuraman, Karthikeyan S., Sumathi D. and Kumar Debasis at the School of Computer Science and Engineering, VIT-AP for their continuous guidance and support. A special thanks to the Centre for Excellence in Artificial Intelligence and Robotics.

By: Tridev Reddy and Aditya Mitra

Both the authors like to work in the field of network analysis, IoT and security.

Supervised Machine Learning for Car Evaluation Using KNIME

The KNIME (Konstanz Information Miner) analytics platform is perhaps the strongest and most comprehensive free platform for drag-and-drop analytics, machine learning, statistics, and ETL (extract, transform and load) models. This article introduces a supervised machine learning decision tree algorithm through the KNIME tool for a car evaluation data set.



Data pre-processing, modelling, analysis, and visualisation are all enabled within KNIME. The workflows can run both through the interactive interface and also in batch mode. These two setups allow for easy local job management and regular process execution. KNIME integrates various components for machine learning and data mining through its modular data pipelining 'Lego of Analytics' concept.

Let us look at a supervised machine learning decision tree algorithm through the KNIME tool for the car evaluation data set. Decision trees are created for the car model based on the technical features, pricing criteria (overall cost, maintenance cost), comfort, size of luggage boot, safety, passenger capacity, etc. The decision tree learner node and the predictor nodes are connected with both trained data and test data for training and evaluating the model with better output. Then the decision tree predictor is configured with the scorer node to generate the required evaluation performance metrics and to report data for user interface visualisation.

KNIME installation

To start with, we will first install KNIME. Download the KNIME desktop application from *https://www.knime.com/* and install it as per the instruction on the screen. You will have to provide the necessary permissions during the setup process.

Decision tree

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to construct a model that contains conditional control statements. Decision trees are used for handling non-linear data sets effectively. The decision tree tool is used in real life in many fields, such as engineering, civil planning, law, and business.

Figure 1 lists the nodes that we will be using to create the decision tree.

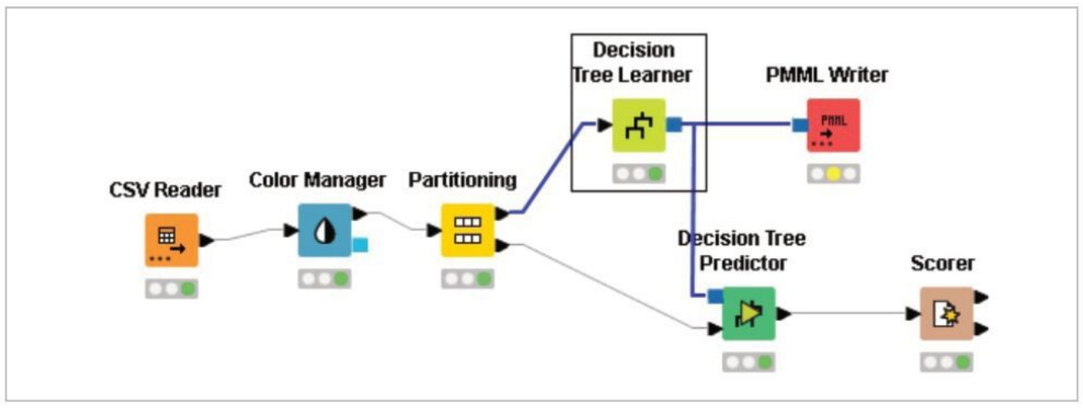


Figure 1: List of nodes used for creating the decision tree

Designing the supervised ML car evaluation decision tree model

Here are the steps needed to create a decision tree model for car evaluation.

- Create a new workflow for the car evaluation model as seen in Figure 2.
- Drag the *.csv* file into KNIME to create and initialise the *CSV* reader node.
- Connect it to a colour manager, where each genre can be assigned a specific colour under *swatches* or a preset under *palettes*.
- Connect the colour manager to the *partitioning* node, which is used to separate the given data set into two partitions — one consisting of 80 per cent of the data and the other consisting of the remaining 20 per cent of the data. Thus the data is partitioned into two sets for testing and training data in an 80:20 ratio. This partition is done by choosing the *randomly* option.
- Connect the *partitioning node* to the decision tree learner. Configure it as a class column as units, and change the records as per the required analysis.
- Connect the decision tree learner to the decision tree predictor. Send the smaller partition (test data) to the decision tree learner in order to classify the values, which can then be predicted by the decision tree predictor.
- Finally, send the output to the *scorer node* to measure the performance metric score. Compile and execute all the nodes at each stage of connection to get proper outputs.
- Now, to view the output as a report, connect it to the PMML writer node.

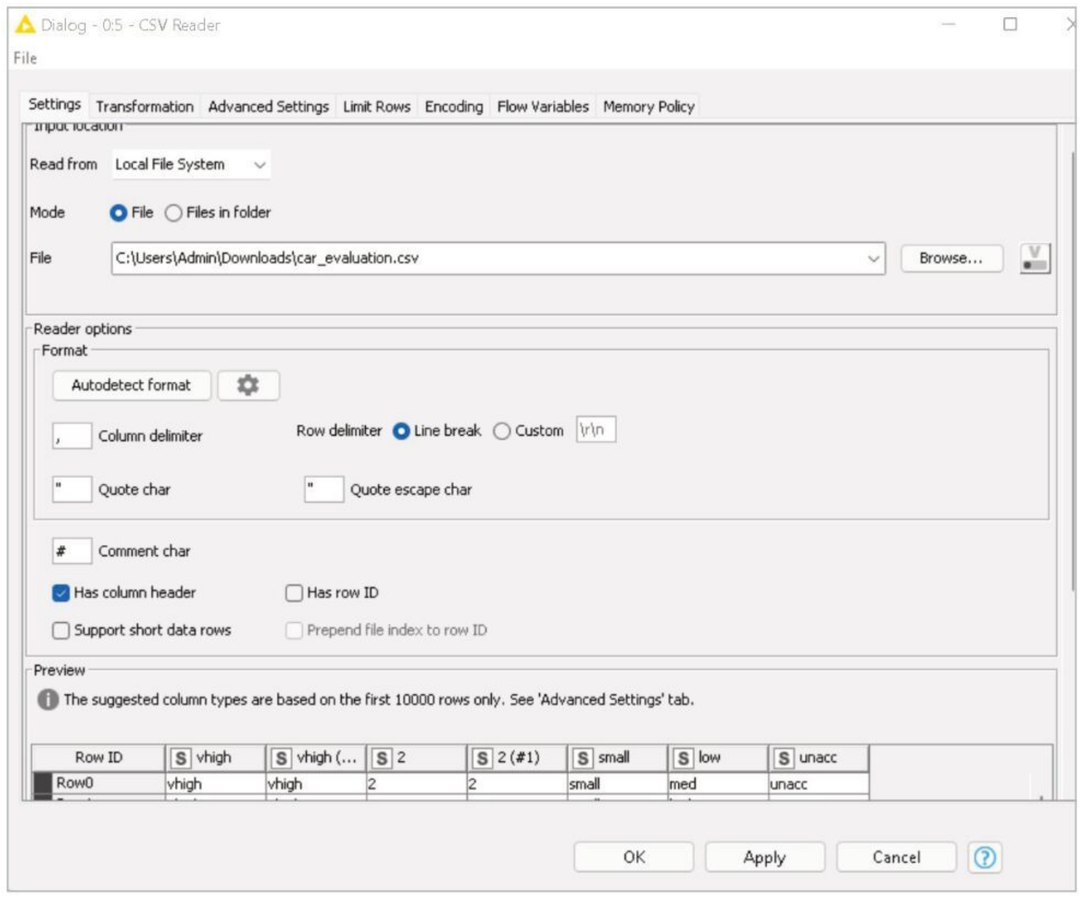


Figure 2: Configuration of CSV reader node

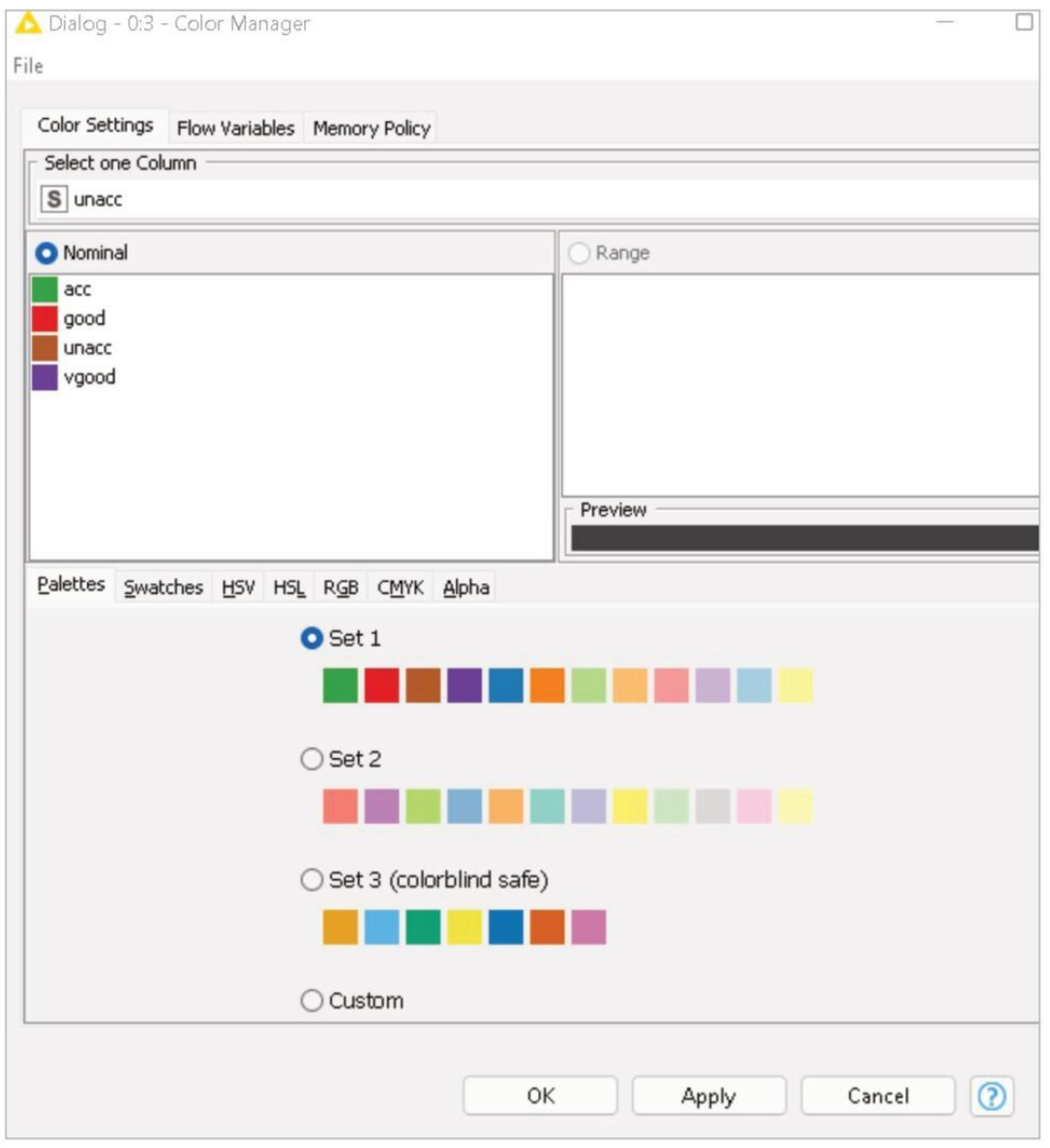


Figure 3: Configuration of colour manager node

Node configurations

CSV reader: This node reads CSV files. You can also use this node if the workflow is created in a server or batch environment, and where the structure of the input files changes between different invocations. The car evaluation data set is loaded using the CSV reader.

Colour manager node: Colours can be assigned for either nominal (possible values have to be available) or numeric columns (with lower and upper bounds). If these bounds are not available, a '?' is provided as a minimum and maximum value. The values are then computed during execution. If a column attribute is selected, the colour can be changed with the colour chooser.

Partitioning node: The input table is split into two partitions (i.e., row-wise), e.g., train and test data. In this case, it is 80 per cent training and 20 per cent testing.

Decision tree learner node: This node induces a classification decision tree in the main memory. The target attribute must be nominal. The other attributes used for decision making can be either nominal or numerical. Numeric splits are always binary (two outcomes), dividing the domain into two partitions at a given split point. Here the target variable is 'Evaluation'

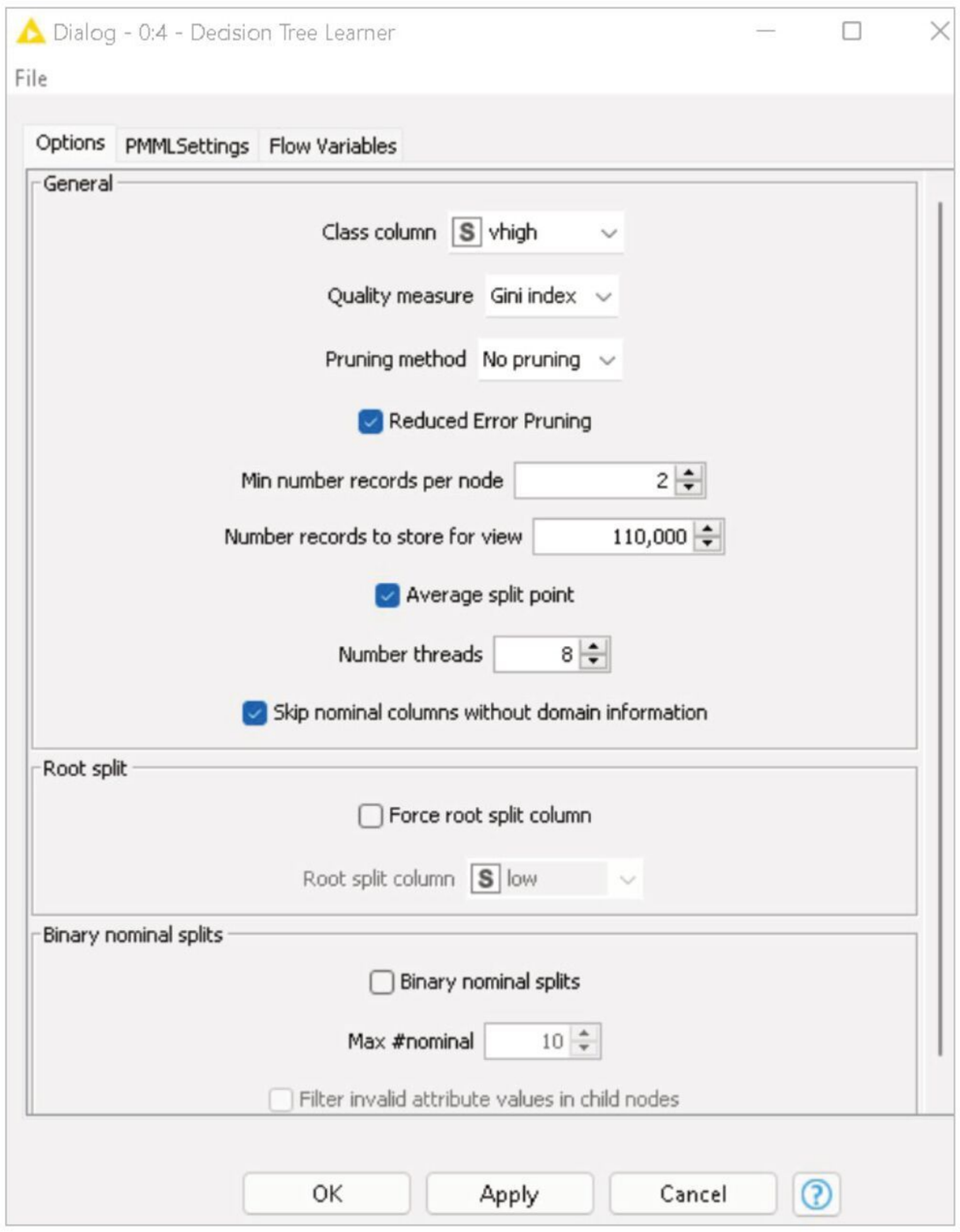


Figure 4: Configuration of partitioning node

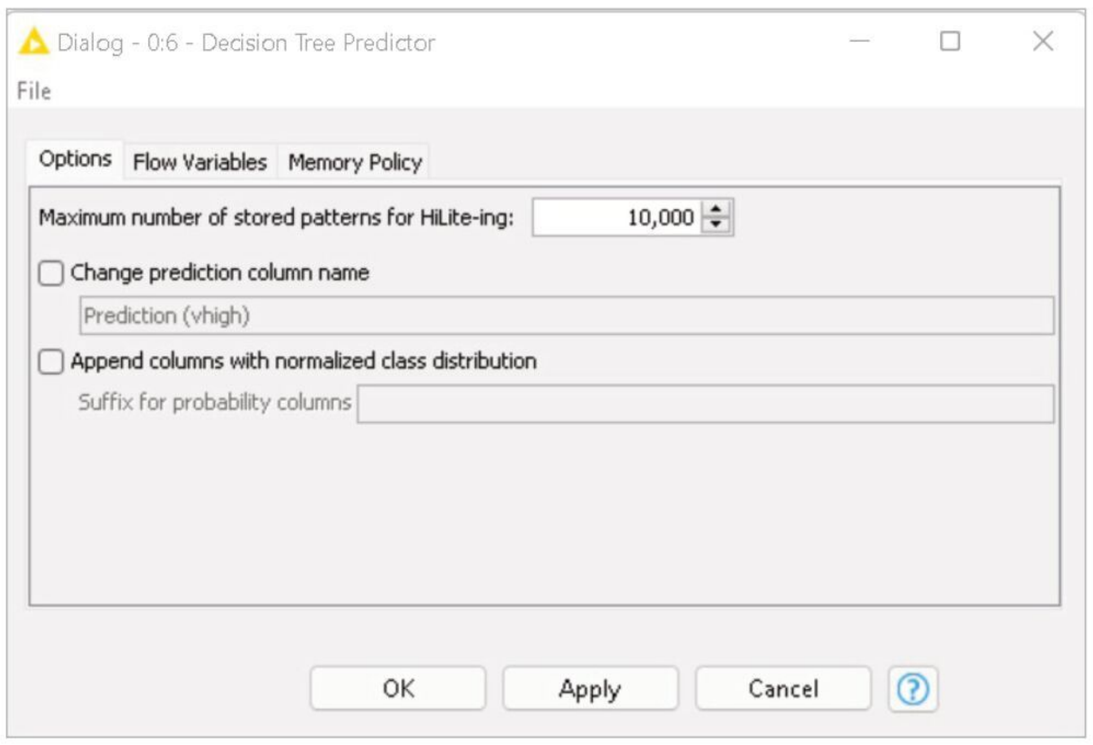


Figure 5: Configuration of decision tree learner node

Decision tree predictor node: This node uses an existing decision tree (passed in through the model port) to predict the class value for new patterns.

Scorer node: This node compares two columns by their attribute-value pairs and shows the confusion matrix, i.e., how many rows of which attribute and their classifications match. Additionally, it is possible to highlight cells of this matrix to determine the underlying rows. Here the accuracy achieved is 95.6 per cent.

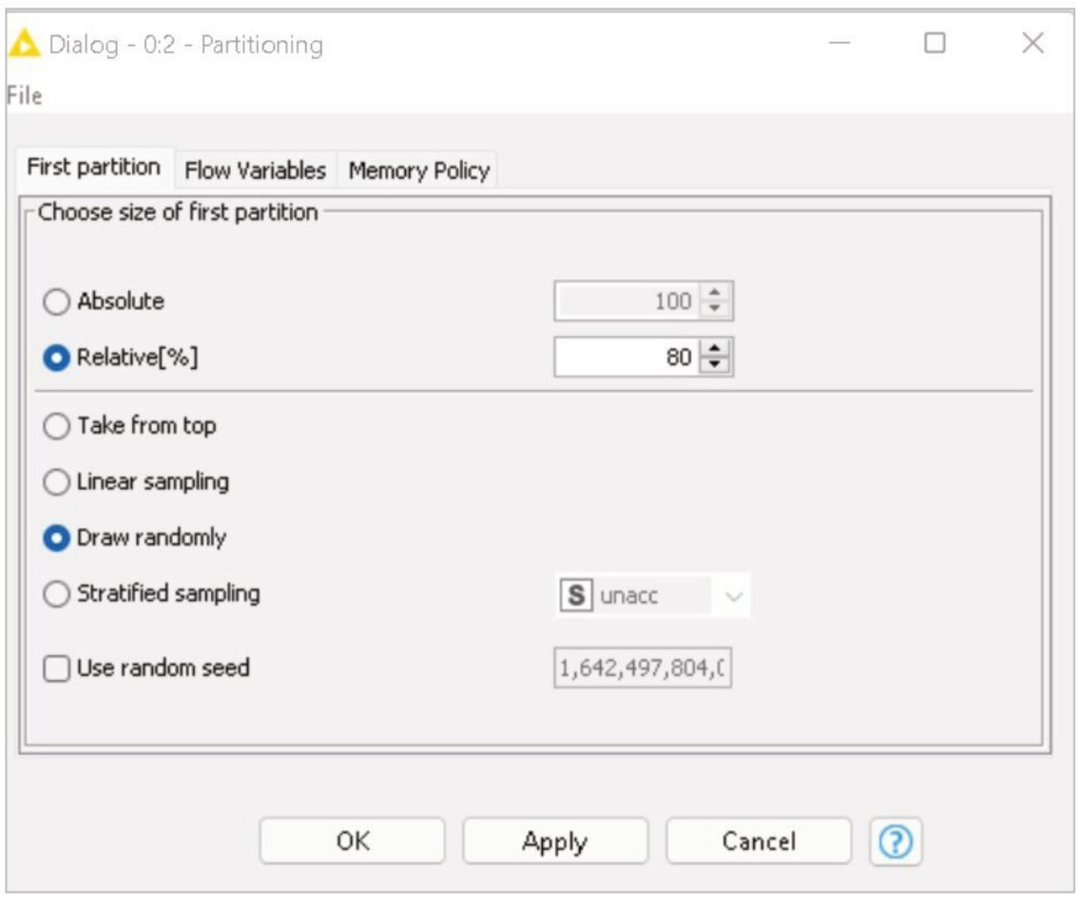


Figure 6: Configuration of decision tree predictor node

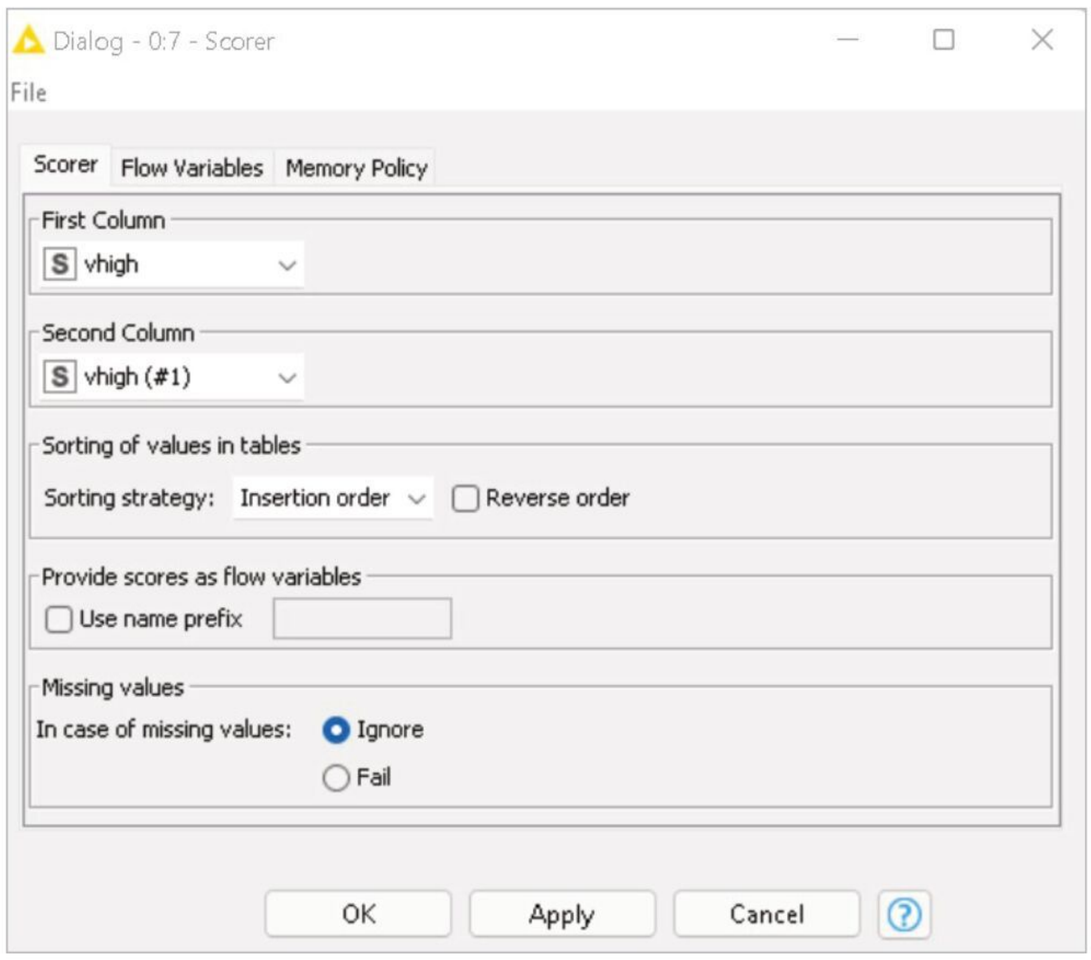


Figure 7: Configuration of scorer node

Visualisation output

Figure 8 demonstrates the decision tree output of the car evaluation model, which visualises the learned decision tree. Based on the selections, the tree can be expanded and collapsed with the plus/minus signs.

As with the splitting criteria, the attribute name of the parent node was split and its numeric value and nominal set value were represented to this child. The value in round brackets states (x of y), where x is the quantity of the majority class and y is the total count of examples in this model.

The PMML (predictive model markup language) node is used to transmit models between nodes. PMML ensembles models through reader and writer nodes for importing and

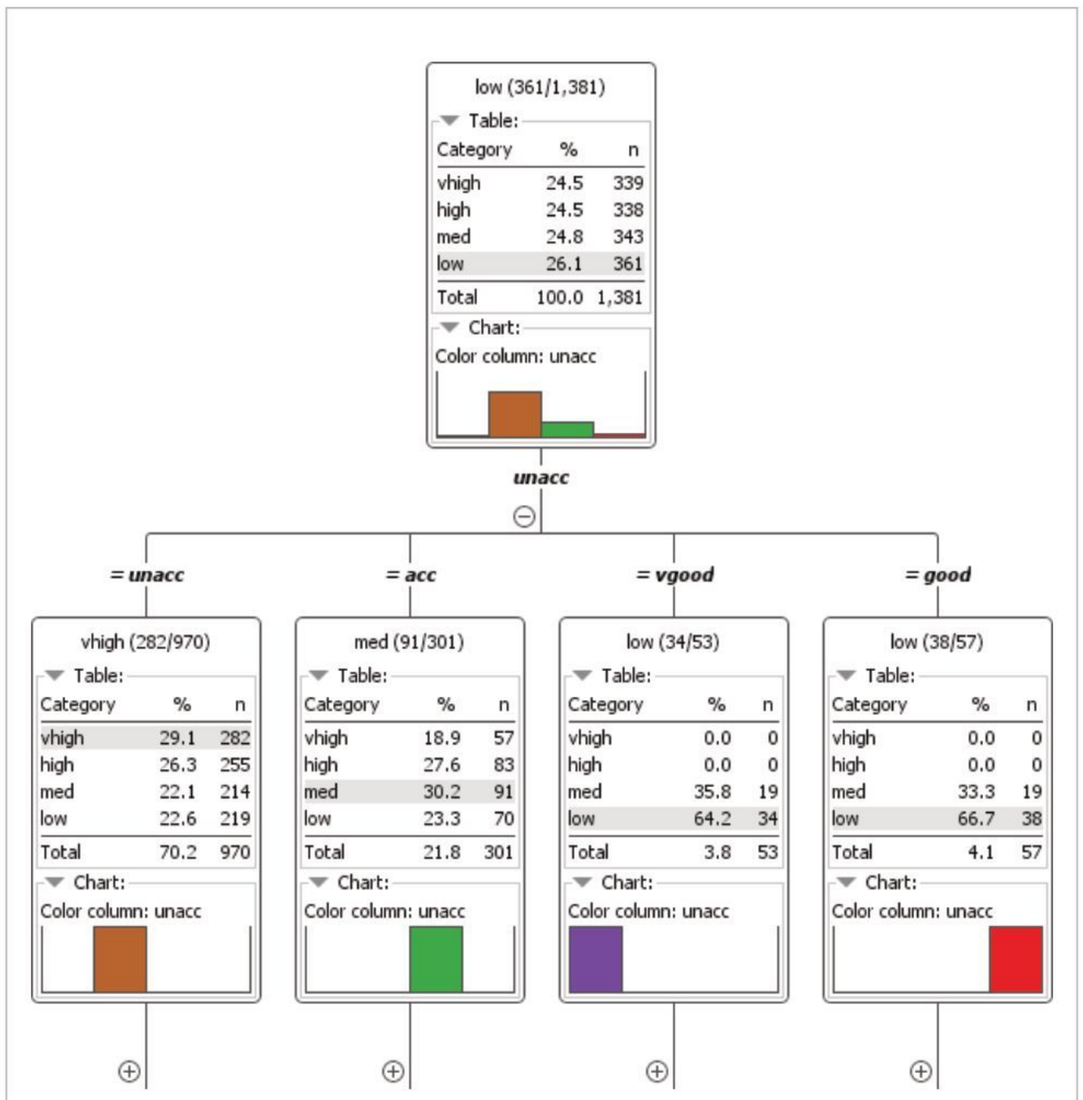


Figure 8: Decision tree output of the car evaluation model

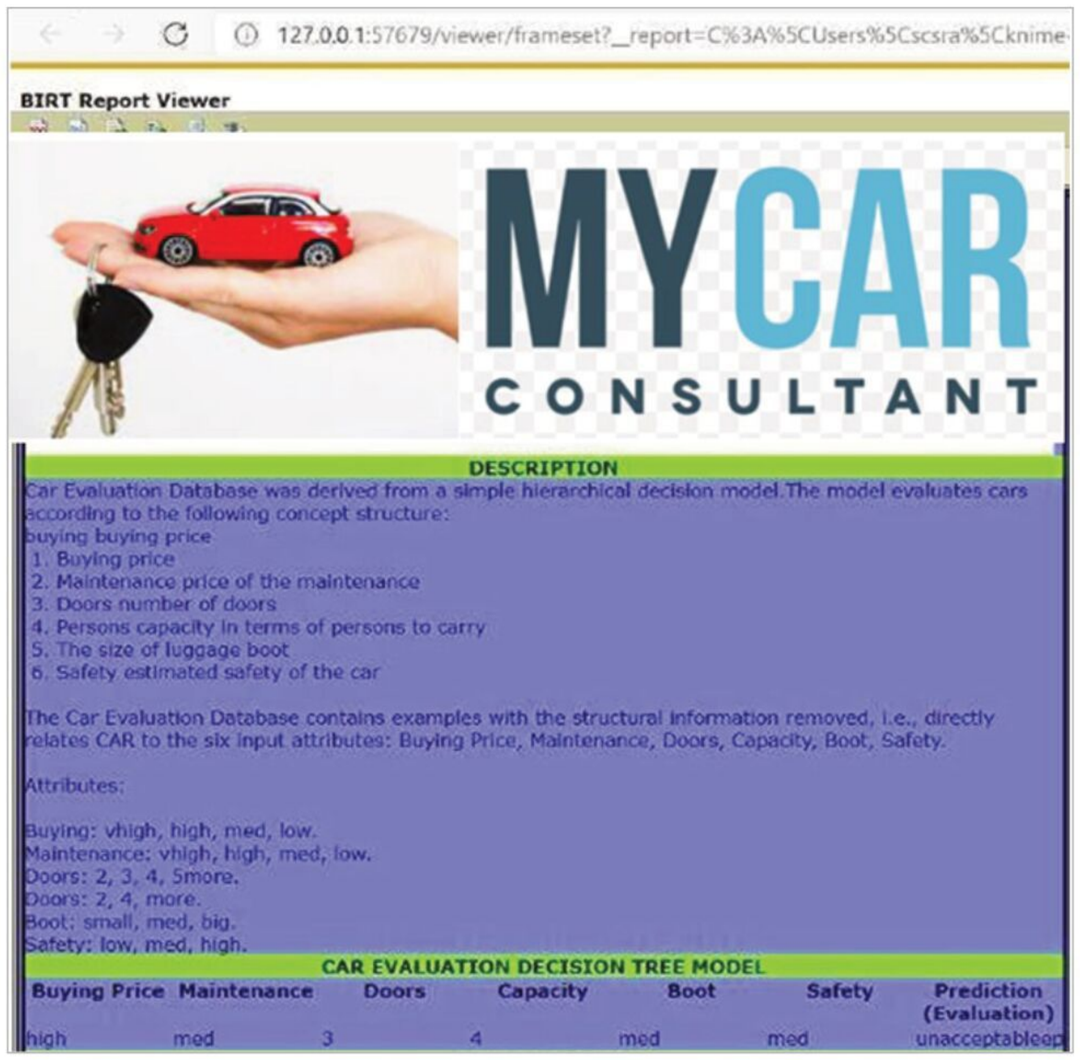


Figure 9: Visualisation report of car evaluation model

exporting them with other software. Figure 9 gives the visualisation report of the car evaluation model. END

By: Dr Maheswari R.

The author is an associate professor at the Vellore Institute of Technology, Chennai.



EFY GROUP
YOURS SINCE 1969

Thank You IIT Madras

It was in your hostel room that the idea was conceived



The first issue in Jan ‘69



The Jan ‘22 Issue

Celebrating 54 years in TECH media

Magazines
- Electronics For You
- Electronics Bazaar
- Open Source For You

Portals
- www.electronicsforu.com
- www.opensourceforu.com
- www.eleb2b.com
- www.electronicsb2b.com

Events
- India Electronics Week
  - IOTshow.in
  - EFY Expo
  - Test & Measurement India
- Open Source India
- Techworldcongress.com

EFY Enterprises Pvt Ltd
D-87/1, Okhla Industrial Area, Phase-1, New Delhi-110020

• Phone: +91-98111-55335
• Email: info@efy.in • Website: www.efy.in

R N I No. DELENG/2012/49440, Mailed on 27/28th of Advance month
Published on 27th April 2022
Delhi Postal Regd. No. DL(S)-01/3443/2022-24
Licenced to Post without Pre-Payment Licence No U(S)-56/2022-24
Posted at LPC Delhi RMS Delhi 110006

India's No.1 Electronics Magazine
RANKING AS PER ELCINA-IMRB REPORT

"I started reading it when I was a student... ...and I am still reading it, as a student"

—CEO, Design House



electronics FOR YOU
YOURS SINCE 1969

To Subscribe:
https://subscribe.efy.in

OR

Scan This Code

For any query, call: +91-98111-55335 or email: support@efy.in

