The InfoQ eMag / Issue #110 / December 2023

The InfoQ Trends Reports 2023

1. Culture & Methods
2. Software Architecture & Design
3. AI, ML, and Data Engineering
4. DevOps and Cloud
5. Java

The InfoQ Trends Reports 2023

IN THIS ISSUE

PRODUCTION EDITOR Ana Ciobotaru / COPY EDITOR Susan Conant / DESIGN Dragos Balasoiu, Ana Ciobotaru
GENERAL FEEDBACK feedback@infoq.com / ADVERTISING sales@infoq.com / EDITORIAL editors@infoq.com

CONTRIBUTORS

Shane Hastie

leads the Culture and Methods editorial team for InfoQ.com where he hosts the weekly InfoQ Culture Podcast. He is the Director of Community Development for ICAgile and is the founding chair of the Agile Alliance New Zealand.

Thomas Betts

is the Lead Editor for Architecture and Design at InfoQ, a co-host of the InfoQ Podcast, and a Laureate Software Architect at Blackbaud. For over two decades, his focus has always been on providing software solutions that delight his customers. He has worked in a variety of industries, including retail, finance, health care, defense and travel. Thomas lives in Denver with his wife and son, and they love hiking and otherwise exploring beautiful Colorado.

Blanca Garcia Gil

works as a lead analytics engineer at the Berlin-based fintech startup topi. She is focused on building and expanding the data engineering and analytics area. She spent the last 7 years at the BBC, contributing to the Data Platform team, where she built highly scalable data ingestion pipelines and led teams improving access to data in the data warehouse and data visualisation. Prior to the BBC, she held a variety of roles: web applications development, mobile prototyping and working on a content management system.

Eran Stiller

is a Principal Software Architect based in Melbourne, Australia. As a seasoned software architect and CTO, Eran designed, implemented and reviewed various software solutions across multiple business domains. Eran has many years of experience in the software development world and a track record of public speaking and community contribution. Microsoft recognized him as a Microsoft Regional Director (MRD) since 2018 and a Microsoft Most Valuable Professional (MVP) on Microsoft Azure between 2016-2022.

Vasco Veloso

has been developing and designing software for over twenty years. From assembly, through C, C++ and Prolog, to Java, Scala and Kotlin, on big and small computers, from floppy disks to SSDs, on-premises and cloud, he's been there, done that and used it. He brings teams together to produce well-crafted software. He still enjoys sharing knowledge by teaching and learning and continues to design software and connected devices. In his spare time, he's discovering the city of Amsterdam, loves photography and is very much interested in the aviation industry.

Daniel Bryant

works as an independent technical consultant and advisor to startups, and is the News Manager at InfoQ and Emeritus Chair for QCon London. His technical expertise focuses on 'DevOps' tooling, cloud/container platforms, and microservice implementations. Daniel is a long-time coder, platform engineer, and Java Champion who contributes to several open source projects. He also writes for InfoQ, O'Reilly, and The New Stack and regularly presents at international conferences such as KubeCon, QCon, and JavaOne. In his copious amounts of free time, he enjoys running, reading, and travelling.

Tanmay Deshpande

Enterprise Architect @ Schlumberger

Tanmay Deshpande blogs at https://deshpandetanmay.medium.com/

Matt Campbell

leads the DevOps editoral team for InfoQ. He is the VP of Cloud Platform at D2L, an education technology company, and is responsible for their Infrastructure and Cloud platform teams. His area of focus is DevOps and SRE and implementing these at enterprise scale.

Helen Beal

is chair of the Value Stream Management Consortium and co-chair of the OASIS Value Stream Management Interoperability Technical Committee. She is chief ambassador at DevOps Institute and chair of the DevNetwork DevOps Advisory Board. She also provides strategic advisory services to DevOps and VSM industry leaders. Helen is the author of the annual State of VSM Reports from the VSMC and the State of Availability Report from Moogsoft. She is a co-author of the book about DevOps and governance, Investments Unlimited, published by IT Revolution. She is a DevOps editor for InfoQ, and also writes for a number of other online platforms.

Stefan-Jan Wiggers

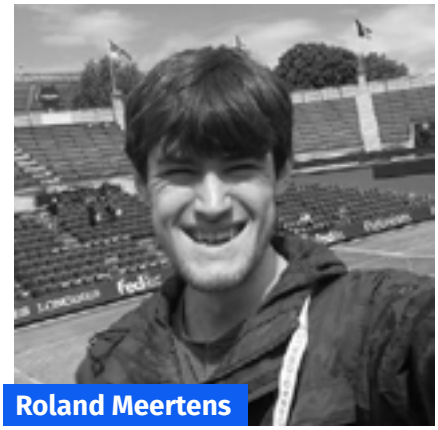
is one of InfoQ's senior cloud editors and works as a Technical Integration Architect at HSO in The Netherlands. His current technical expertise focuses on integration platform implementations, Azure DevOps, and Azure Platform Solution Architectures. Steef-Jan is a board member of the Dutch Azure User Group, a regular speaker at conferences and user groups, writes for InfoQ, and Serverless Notes. Furthermore, Microsoft has recognized him as Microsoft Azure MVP for the past eleven years.

Abby Bangser

is a Principal Engineer at Syntasso delivering Kratix, an open-source cloud-native framework for building internal platforms on Kubernetes. Her keen interest in supporting internal development comes from over a decade of experience in consulting and product delivery roles across platform, site reliability, and quality engineering. Abby is an international keynote speaker, co-host of the #CoffeeOps London meetup, and supports SLOConf as a global captain. Outside of work, Abby spoils her pup Zino and enjoys playing team sports.

Anthony Alford

is a Director, Development at Genesys where he is working on several AI and ML projects related to customer experience. He has over 20 years experience in designing and building scalable software. Anthony holds a Ph.D. degree in Electrical Engineering with specialization in Intelligent Robotics Software and has worked on various problems in the areas of human-AI interaction and predictive analytics for SaaS business optimization.



Roland Meertens

is a data scientist working on Computer Vision. Previously he worked on self driving cars, deep learning approaches for natural language processing (NLP) problems, social robotics, and computer vision for drones. Roland Meertens blogs here.

Srini Penchikala

currently works as Senior Software Architect in Austin, Texas. He is also the Lead Editor for AI/ML/Data Engineering community at InfoQ. Srini has over 22 years of experience in software architecture, design and development. He is the author of "Big Data Processing with Apache Spark. He is also the co-author of "Spring Roo in Action" book from Manning Publications.

Sherin Thomas

is a Software Engineer with over 12 years of experience at companies like Google, Twitter, Lyft, Netflix and Chime. She works in the field of Big Data, Streaming, ML/AI and Distributed Systems. Currently, she's building a shiny new data platform at Chime. Sherin has presented on the topic of ML and Streaming at various reputable conferences including a keynote address and has judged various awards such as SXSW Innovation awards and CES. Recently she advised NASA's SpaceML program and helped build a platform for processing petabytes of satellite imagery for detecting weather patterns and labelling raw data for climate science related AI research.

Daniel Dominguez

is an engineer with experience in software product development for companies ranging from Silicon Valley startups to Fortune 500. He is passionate about cloud computing to deliver innovative software solutions. In addition to software product management, Daniel is also involved in artificial intelligence and machine learning projects. Daniel is part of the Editorial Team for AI/ML/Data Engineering community at InfoQ.

Shaaron A Alvares

has a global work experience in technology and organizational transformation and she introduced lean agile product and software development practices within various global Fortune 500 companies in Europe, such as BNP-Paribas, NYSE-Euronext, ALCOA Inc. She led significant lean agile and DevOps practice adoptions and transformations at Amazon.com, Expedia, Microsoft, T-Mobile, Disney, while focusing on introducing the Agile mindset and customized value-driven practices aligned with organizational performance goals.

Rebecca Parsons

before joining Thoughtworks in 1999, I was a researcher and college lecturer in computer science. After completing my degree, Master's and Ph.D, I conducted research in compilers, program optimization, distributed computation, programming languages, theory of computation, machine learning and computational biology. In 2007 I became CTO, responsible for driving Thoughtworks' excellence in technology. In addition to deep technology, I am also a strong advocate for diversity and inclusion in the industry, particularly increasing the number of women in coding and STEM.

Alina Yurenko

is a developer advocate for GraalVM at Oracle Labs, a research & development organization at Oracle. She loves both programming and natural languages, compilers, and open source.

Michael Redlich

has been an active member within the Java community for the past 25 years. He founded the Garden State Java User Group (formerly the ACGNJ Java Users Group) in 2001 that remains in continuous operation. Since 2016, Mike has served as a Java community news editor for InfoQ where his contributions include monthly news items, technical writing and technical reviews. He has presented at venues such as Oracle Code One, Emerging Technologies for the Enterprise, Trenton Computer Festival (TCF), TCF IT Professional Conference, and numerous Java User Groups. Mike serves as a committer on the Jakarta NoSQL and Jakarta Data specifications.

Ixchel Ruiz

has developed software applications and tools since 2000. Her research interests include Java, dynamic languages, client-side technologies and testing. Ixchel is a Java Champion, Oracle ACE pro, Testcontainers Community Champion, CDF Ambassador, Hackergarten enthusiast, Open Source advocate, public speaker and mentor.

Johan Janssen

Architect at ASML, loves to share knowledge mainly around Java. Spoke at conferences such as Devoxx, Oracle Code One, Devnexus, and many more. Assisted conferences by participating in program committees and invented and organized JVMCON. Received the JavaOne Rock Star and Oracle Code One Star awards. Wrote various articles both for digital and printed media. Maintainer of various Java JDK/JRE packages for Chocolatey with around 100 thousand downloads a month.

Rustam Mehmandarov

Passionate computer scientist • Public speaker • Java Champion and Google Developers Expert (GDE) for Cloud • Oracle ACE Pro for Java • Chief Engineer and Architect • Software development, data integration, AI • Community lead • GDG Cloud Oslo organiser • Ex-leader of JavaZone and Norwegian JUG – javaBin.

Rafiq Gemmail

is a New Zealand based DevOps consultant and the Head of Technology at Developers Institute. He is a technical coach, teacher and polyglot who has helped leading product teams achieve faster-feedback and greater empiricism. He is a passionate advocate for Good Remote ways of working, disrupting traditional models for engineering education, and is also a champion for DevOps culture, who helps with NZ's DevOps days. He is also an ICAgile certified coach and a Java professor for Open Classroom.

Craig Smith

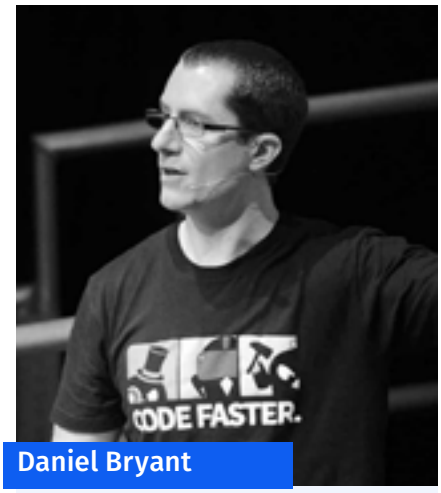
has been a software developer for over 15 years, specialising in a large number of technologies in that time. He has been an Agile practitioner for over 10 years, is a Certified Scrum Master and Certified ICAgile Professional and a member of both the Scrum Alliance and Agile Alliance and currently works as an Agile Coach, fulfilling technical lead, iteration manager and Agile coaching roles on technology and business projects. In his spare time, Craig is an avid motorsport fan.

Ben Linders

is an Independent Consultant in Agile, Lean, Quality and Continuous Improvement, based in The Netherlands. Author of Getting Value out of Agile Retrospectives, Waardevolle Agile Retrospectives, What Drives Quality, The Agile Self-assessment Game, Problem? What Problem?, and Continuous Improvement. Creator of many Agile Coaching Tools, for example, the Agile Self-assessment Game. As an adviser, coach and trainer he helps organizations by deploying effective software development and management practices.


2023


2024



Daniel Bryant

works as an independent technical consultant and advisor to startups, and is the News Manager at InfoQ and Emeritus Chair for QCon London. His technical expertise focuses on 'DevOps' tooling, cloud/container platforms, and microservice implementations. Daniel is a long-time coder, platform engineer, and Java Champion who contributes to several open source projects. He also writes for InfoQ, O'Reilly, and The New Stack and regularly presents at international conferences such as KubeCon, QCon, and JavaOne. In his copious amounts of free time, he enjoys running, reading, and travelling.

A LETTER FROM THE EDITOR

Welcome to this special edition of The InfoQ eMag, which contains a comprehensive collection of our popular InfoQ Trends Reports from 2023. This year has been a remarkable journey through the ever-evolving landscapes of technology, software development trends, and organizational practices. Our reports have delved deep into the latest advancements, challenges, and future directions. We have aimed to provide insight into topics such as hybrid working, monoliths vs microservices vs moduliths, platform engineering, large language models (LLMs), and the state-of-the-art within the Java ecosystem.

As you explore these pages, you'll encounter insights and analyses that are pivotal for understanding the current state of the software development landscape. The 2023 InfoQ Trends Reports, meticulously compiled by our expert contributors and other practitioners from the software engineering community, encapsulate a wealth of knowledge and experience.

This collection is not just a reflection of the past year's technological trends. We aim for it to be a beacon for future exploration and innovation. Each report serves as a lens through which we view the world of software development and delivery. Whether you're a developer, architect, technology leader, or an enthusiast, these reports offer valuable perspectives that will help you plan your future roadmaps and explore emerging technologies and practices.

We start the collection with the InfoQ Culture & Methods Trends Report. Our contributors explored how asynchronous work is becoming much more accepted, and organisations adopting the practices are gaining real benefits. In an accompanying podcast, they also discussed the impact of AI and how "responsible tech" is more than just regulatory compliance; companies are increasingly having to be more socially responsible in order to attract and retain customers and employees. For readers wanting to keep up to date with these topics, the weekly InfoQ Engineering Culture Podcast acts as a perfect guide.

In our flagship report, the InfoQ Software Architecture and Design Trends Report, we explore all the technologies, techniques, and trends relevant to architects. Much of the analysis here helps to set the scene for our monthly InfoQ Architects' Newsletter, the weekly InfoQ Podcast, and our QCon conferences. This year the team explored how architects are increasingly looking for improvements in how to document, communicate, and understand decisions. As explored in the accompanying podcast, this may be an area where large language models will play a role in the future, acting as forensic archaeologists to comb through Architecture Decision Records (ADRs) and git history.

The InfoQ DevOps and Cloud Trends Report has rapidly become very popular amongst our readers, as the impact of cloud technologies touches practically every organisation. Here the team explored how cloud innovation has transitioned from a "revolutionary phase to an evolutionary one", focusing on migrating and re-architecting workloads. There was also an increase in the adoption of managed services, emphasizing simplifying interactions and reducing cognitive load for teams. The accompanying podcast to this report discusses the question, "Is DevOps Dead?" (spoiler alert: no!), the impact of AI on operations and platform engineering, and the rise of FinOps.

Next up is the InfoQ AI, ML, and Data Engineering Trends Report. It would be an understatement to say that in 2023 Generative AI, powered by Large Language Models (LLMs) like GPT-3 and GPT-4, has gained significant prominence in the AI and ML industry. There has been widespread adoption driven by technologies like ChatGPT, and this technology is visible in all of our trend reports this year. In the accompanying podcast, industry experts scratch below the surface of this revolution, and explore foundational models, vector databases, and embedding stores. The team also discuss how modern data engineering is shifting towards decentralized and flexible approaches, such as Data Mesh.

Closing out the emag is a subject close to my technological heart, the InfoQ Java Trends Report. As a long-time Java developer, I was stoked to be able to sit down with the lead editor for this report for an InfoQ podcast episode, InfoQ Java Trends Report 2023 - Discussing Insights with Mike Redlich.

We explored how Java Virtual Threads was finalized in the recently released JDK 21, and why we believe that adoption of this feature will continue to grow as the latest edition of application frameworks, such as Helidon Níma and Vert.x, have already taken advantage of this. We also discussed the evolution of the Spring framework, including the very interesting Spring Modulith project.

As you read this collection of our trend reports, we encourage you to engage with the content actively. Reflect on how these trends resonate with your experiences, challenge your perspectives, and inspire your future roadmap. We aim for this emag to become a guide and a catalyst for your continuous learning and adaptation in the ever-changing realm of technology.

Thank you for joining us on this journey. We hope that the collection of InfoQ Trends Reports from 2023 will be an invaluable resource for you, sparking new ideas, conversations, and innovations.

Thanks for reading! And, as always, please send any feedback to us at editors@infoq.com.

InfoQ Culture & Methods Trends Report

by **Shane Hastie**

The two biggest factors influencing organisational culture in early 2023 are the highly publicised and hugely disruptive layoffs across the tech industry and the emergence of Large Language Models such as ChatGPT.

Layoffs in tech have damaged psychological safety

The biggest impact of the layoffs has not been the number of people out of work, but the widely publicised reports of the inhumane approach many of the large employers took to announcing and communicating the layoffs. Stories of people finding out they no longer have a job when they are not able to log on to their corporate account, or getting a text message from an unidentified number telling them to drop off their company equipment with no other communication have a very negative impact on the industry as a whole and psychological safety is down in almost every high tech company.

There is a realistic argument that the layoffs were a sensible response to the changing economic climate and that for many companies it is about right-sizing rather than down-sizing, however the way they were communicated to the impacted staff and to the world at large has resulted in a loss of trust and reduction of employee engagement across the industry. Rebuilding a generative culture will be difficult, but can be done with care and deliberate design.

Despite the large numbers who were laid off, there are still large shortages of tech workers around the world and most of those impacted are finding new roles relatively quickly.

Rebecca Parsons, CTO of Thoughtworks commented on the podcast accompanying this trend report that:

There is a strong business case for having a strong employee value proposition, because turnover is expensive and replacing people is expensive. If you treat people well, you can make all of these wonderful business case arguments, but you can also look at it as it's just the right thing to do.

Employers who treat people badly in layoffs are going to struggle to attract people when the economy turns and they look to grow again.

Large language models - friend or foe for developers?

The launch of large language models such as ChatGPT, the premium service ChatGPT Plus and the enhanced GitHub Copilot have resulted in headlines about jobs that are at risk of being replaced by AI, with programmers at the top of some lists. The question of how real the threat is, and how do the potentially displaced adapt is very open at the moment as it is early days in the evolution of the tools. Legal battles are shaping up around open source licensing and copyright both of the source material used to generate content and the copyright status of generated content.

On the other hand there are reports of how using these tools can significantly increase developer productivity - handling many of the more mundane tasks of coding, reducing the cognitive load and allowing the creative human to apply their energies to the areas of development where creativity is best applied. Developer tools already include AI assistance, and this will continue to increase.

What has become clear is that there is a skill to crafting the questions to ask; prompt engineering is the name most used to describe this skill. These code assist tools work best if you know the answer that you want. Users need to verify the results and validate that the responses actually address the issue raised. The models lack context and don't know what they don't know and that makes using them potentially risky.

ChatGPT recently had a temporary shutdown of its service due a bug which enabled users to view one another's prompts. This illustrates the inherent risks of sharing confidential data with the providers of large language models. OpenAI's terms of service clearly state that they collect information which is "included in the input" or in "file uploads." As effective prompt engineering relies on providing contextual information, it is important that organisations safeguard themselves against inadvertent data privacy breaches. While some organisations have been putting in place governance, JPMorgan recently blocked all staff from submitting data to ChatGPT.

Unskilled cybercriminals are already using ChatGPT to generate malware; as technologists we need to be aware of the potential impacts and actively mitigate the risks.

While these two factors have dominated the headlines there has been a lot happening in other areas that is impacting organisational cultures and people operations.

Movements in responsible tech

Technologists are becoming more aware of the broader societal impacts of the work they do and the importance of ethical considerations in the products we build and the way we build them. Climate impact, social good, privacy, safety, diversity, equity and inclusion and the impact of bias need to be considered in our development processes.

Responsible technology goes beyond compliance with rules and regulations. Taking a responsible view means considering the potential impacts of the products we're building before we start building, asking the hard questions about how could this product be applied in ways we don't intend, who are the communities and stakeholders that we haven't consulted or considered, how

can we actively prevent harm, rather than just complying with the minimum regulatory requirements?

It has a financial benefit from enhanced reputation, from employees and customers wanting to be associated with ethical companies and from customers being prepared to pay more for ethical products. Responsible technology doesn't just happen - it needs deliberate practice and clear goals covering areas such as supply chain, operations, employees and clients, in addition to the more obvious issues around product and service development.

DevSusOps is more than just being carbon neutral - to be effectively sustainable an organisation needs to look to the needs of all stakeholders, social, environmental and economic.

Ethical standards and guidelines for software engineering are evolving and emerging in new areas, such as agile coaching.

Project to product gaining traction

The ideas around project to product/#NoProjects are gaining traction as more organisations shift from seeing their IT departments as an overhead cost to truly partnering as a value adding service essential to running the business effectively and meeting customers' needs.

Value stream management is being used more frequently to visualise and optimise the development process in end to end DevOps implementations.

As organisations adopt these approaches, project based working breaks down and product management becomes necessary and effective. Approaches such as Evidence Based Management provide tools to help maximise value delivered and focus on business outcomes.

The worth of an architect

The frequent conflict between development teams and architects has caused harm and some organisations have found ways to effectively reconcile the potentially competing goals. Development teams are tasked with delivering features as quickly as possible to add value and delight customers. Architects have a longer term view of maximising the long-term asset value of the IT estate. Aligning these two perspectives requires a shift in communication and engagement.

Making architecture a core part of the development process requires constant collaboration and communication, not through presentations and ivory-tower decrees, but through working code and practical application that show the value the big picture thinking has on the product both in the short and long term.

Organisation are increasingly distributing the architectural function across teams, through the use of cross-team reviewed Design Docs or RFCs. Grygoriy Gonchar recently published an article on InfoQ about the use of a Simple Framework for Architectural Decisions, which includes the use of RFC like Architectural Decision Records (ADRs). Gonchar wrote that such approaches ensure alignment with "business goals" and allow teams to make "informed" architectural decisions.

Team topologies for rapidly changing environments

Team topologies are emerging that support rapidly evolving organisations and the remote/hybrid workforce. Reteaming is a competency that organisations need to build - enabling teams to rapidly form around new challenges or opportunities and become fully productive and well gelled quickly.

Understanding the impact of Conway's Law, and using the Reverse Conway approach enable organisations to deliberately align communication to be more effective. Team Topologies has provided patterns to guide organisations in restructuring teams to optimise for value delivery and minimising cognitive load. As

Twitter's mass layoffs have recently shown, the downside of mass redundancies can be the loss of moral, sustainability and stability due to increases in cognitive load for remaining teams.

The pace of change and the dynamic nature of employment mean that organisations need to continue reviewing the changing cognitive load on teams, when faced with large scale organisational change. Teams must also become good at onboarding new people to help them become productive.

Staff Plus and technical career pathways

The Staff Plus role is gaining more attention and a clear recognition that there are career paths which don't require going into people management positions. Deliberately shifting into and out of people leadership roles on a pendulum is a viable option that both deepens and broadens skillsets.

Hybrid work will continue to evolve

The shift from fully remote to hybrid working continues steadily, with a few companies mandating complete back to the office and losing people as a result. Hand in hand with the shift in working comes a far greater acceptance of asynchronous work in remote teams and an uptick in the need for more effective written communication and the need for process and guidelines on how to (for instance) communicate architectural decisions so they are understood across the whole community who need to know about them. Finding the right balance between freedom & flexibility and documentation & process is an important factor when much of the work is done asynchronously.

Workplace design is changing as organisations retool for hybrid working. We see designs with fewer desks and more collaboration spaces with technology to support teamwork, pods and quiet places for individual work and quiet conversations and reflection.

Wellbeing and helping support mental wellness need to be taken into account when establishing ways of working. Understanding the diverse needs of the organisational community and ensuring inclusiveness and respect enables more effective collaboration.

The collaboration tools available today have improved significantly, but they still have gaps and frustrations. A hybrid meeting with some in-person and some remote participants is often an exercise in frustration with poor sound quality and missed interactions, exacerbating the us-and-them divide which can easily emerge in distributed teams.

TL;DR

- The layoffs in big tech have had a dampening effect on psychological safety across the whole industry
- Generative AI tools like Co-pilot or ChatGPT can have a significant impact on developer productivity but are also significantly flawed
- Responsible tech is more than just regulatory compliance and companies are having to be more socially responsible in order to attract and retain customers and employees
- Asynchronous work is becoming much more accepted and organisations adopting the practices are gaining real benefits
- Hybrid workplaces are now the norm, it is very important to be deliberate about the reasons for bringing people together in person, rather than just having a fixed schedule comb through ADRs and git history.

InfoQ Software Architecture and Design Trends Report

by **Thomas Betts, Bianca Garcia Gil, Eran Stiller, Vasco Veloso, Daniel Bryant, Tanmay Deshpande**

The InfoQ Trends Reports provide InfoQ readers a high-level overview of the topics to pay attention to, and also help the InfoQ editorial team focus on innovative technologies. In addition to this report and the trends graph, an accompanying podcast features some of the editors discussing these trends.

Updates to the trends graph

More details follow later in the report, but first it is helpful to summarize the changes from last year's trends graph.

Three new items were added to the graph this year. Large language models and software supply chain security are new innovator trends, and "architecture as a team sport" was added under early adopters.

Trends which gained adoption, and therefore moved to the right, included "design for portability," data-driven architecture, and serverless. eBPF was removed as it has niche applications, and is not likely to be a major driver in architectural decisions.

A few trends were renamed and/or combined. We consider Dapr as an implementation of the "design for portability" concept, so it was removed as a separate trend. Data-driven architecture is the combination of "data + architecture" and data mesh. Blockchain was replaced with the broader idea of decentralized apps, or dApps. WebAssembly now notes both server-side and client-side, as these are related but separate ideas and may evolve independently in the future.

Design for portability

The portability aspect of "design for portability" is not about being able to pick up your code and move it. Rather, it creates a clean abstraction from the

infrastructure. As InfoQ editor Vasco Veloso says, "whoever is designing and building the system can focus on what brings value, instead of having to worry too much with the platform details that they are going to be running on."

This design philosophy is being enabled by frameworks such as Dapr. Daniel Bryant, InfoQ news manager, sees the benefit of the CNCF project as providing a clearly defined abstraction layer and API for building cloud-native services. Bryant said, "[with integration] it's all about the APIs and [Dapr] provides abstractions without doing the lowest common denominator".

A recent article by Bilgin Ibryam described the evolution of cloud-native applications into cloud-bound applications. Instead of designing a system with logical components for application logic and compute infrastructure, cloud-bound applications focus on the integration bindings. These bindings include external APIs as well as operational needs such as workflow orchestration and observability telemetry.

Another technology that supports designing for portability is WebAssembly, specifically server-side WebAssembly. Often WebAssembly is thought of as a client-side capability, for optimizing code running in the browser. But using WebAssembly has significant benefits for server-side code. InfoQ Editor Eran Stiller described the process for creating WebAssembly-based containers.

Instead of compiling it to a Docker container and then needing to spin up an entire system inside that container on your orchestrator, you compile it to WebAssembly and that allows the container to be much more lightweight. It has security baked in because it's meant to run the browser. And it can run anywhere—in any cloud, or on any CPU, for that matter. – Eran Stiller

More information about Dapr and WebAssembly can be found by following those topics on InfoQ.

Large language models

The news around AI, specifically large language models such as GPT-3 and GPT-4, has been impossible to ignore. This is not simply a tool used by software professionals as the adoption by everyday people and the coverage in all forms of media has demonstrated. But what does it mean to software architects? In some ways, it is too early to know what will happen.

With ChatGPT and Bing, we're just beginning to see what is possible with large language models like GPT-3. This is the definition of an innovator trend. I don't know what will come of it, but it will be significant, and something I look forward to seeing evolve in the next few years. – Thomas Betts

While the future is uncertain, we have optimism that these AI models will generally have a positive benefit on the software we build and how we build it. The code-generation capabilities of ChatGPT, Bing chat, and GitHub Copilot are useful for writing code and tests and allowing developers to work faster. Architects are also using the chatbots to discuss design options and analyze trade-offs.

While these improvements in efficiency are useful, care must be taken to understand the limitations of AI models. They all have built-in biases which may not be obvious. They also may not understand your business domain, despite sounding confident in their responses.

This will definitely be a major trend to watch in 2023, as new products are built on large language models and companies find ways to integrate them into existing systems.

Data-driven architecture

Last year, we discussed the idea of "data + architecture" as a way to capture how architects are considering data differently when designing systems. This year we are combining that idea with Data Mesh under the heading of "data-driven architecture."

The structure, storage, and processing of data are up-front concerns, rather than details to be handled during implementation. Blanca Garcia-Gil, a member of the QCon London programming committee, said, "when designing cloud architectures there is a need to think from the start about data collection, storage, and security, so that later on we can derive value from it, including the use of AI/ML." Garcia-Gil also pointed out that data observability is still an innovator trend, at least compared to the state of observability of other portions of a system.

Data Mesh was a paradigm shift, with teams aligned around the ownership of data products. This fits the idea of data-driven architecture, as well as incorporating Conway's Law into the overall design of a system.

Design for sustainability

While there has been more adoption in designing for sustainability, we chose to leave it as an innovator trend because the industry is just starting to really embrace sustainable systems and designing for a low carbon footprint. We need to consider sustainability as a primary feature, not something we achieve secondarily when trying to reduce costs. Veloso said, "I have noticed that there is more talk about sustainability these days. Let's be honest that probably half of it is because energy is just more expensive and everybody wants to reduce OPEX."

One of the biggest challenges is the difficulty in measuring the carbon footprint of a system. Until now, cost has been used as a stand-in for environmental impact, because there is a correlation between how much compute you use and how much carbon you use. But this technique has many limitations.

The Green Software Foundation is one initiative trying to help create tools to measure the carbon consumed. At QCon London, Adrian Cockcroft gave an overview of where the three major cloud vendors (AWS, Azure, GCP) currently stand in providing carbon measurements.

As the tooling improves, developers will be able to add the carbon usage to other observability metrics of a system. Once those values are visible, the system can be designed and modified to reduce them.

This also ties into the ideas around portability and cloud-native frameworks. If our systems are more portable, that means we will more easily be able to adapt them to run in the most environmentally-friendly ways. This could mean moving resources to data centers that use green energy, or processing workloads during times when the energy available is more green. We can no longer assume running at night, when the servers are less busy is the best option, as solar power could mean the middle of the day is the greenest time.

Decentralized apps (dApps)

Blockchain and a distributed ledger is often the technology behind decentralized apps. However, blockchain remains a technology that solves a problem most people do not see as a problem.

The use of decentralization outside of blockchain and web 3.0 has also emerged elsewhere within the software development ecosystem. For example, Mastodon emerged as a decentralized social network that is an alternative to Twitter. However, consumer uptake of this platform has yet to be as successful as the existing centralized offerings.

Because of this niche applicability, decentralized applications remain classified as an innovator trend.

Architecture as a team sport

Architects no longer work alone, and architects can no longer think only about technical issues. The role of an architect varies greatly across the industry, and some companies have eliminated the title entirely, favoring "principal engineers" as the role primarily responsible

for architectural decisions. This corresponds to a more collaborative approach, where architects work closely with the engineers who are building a system to continually refine the system design.

Architects have been working collaboratively with software teams to come up with and iterate designs. I continue to see different roles here (especially in larger organizations), but communication and working together through proof of concepts to try out designs if needed is key. – Blanca Garcia-Gil

Architecture Decision Records (ADRs) are now commonly recognized as a way to document and communicate design decisions. They are also being used as a collaboration tool to help engineers learn to make technical decisions and consider trade-offs.

Listen to the Trends Report Discussion on the InfoQ Podcast

The Architecture & Design editorial team met remotely to discuss these trends and we recorded our discussion as a podcast. You can listen to the discussion and get a feel for the thinking behind these trends.

TL;DR

- Design for Portability is gaining adoption, as frameworks like Dapr focus on a cloud-native abstraction model and allow architects to separate business logic from implementation details.
- Large language models are going to have a significant impact, from helping understand architectural trade-offs to empowering a new generation of low-code and no-code developers.
- Sustainability of software will be a major design consideration in the coming years. Work is being done to better measure and then reduce the carbon footprint of software systems.
- Decentralized apps are taking blockchain beyond cryptocurrency and NFTs, but a lack of consumer demand will keep this as a niche pattern.
- Architects are always looking for improvements on how to document, communicate, and understand decisions. This may be another area where large language models will play a role in the future, acting as forensic archeologists to comb through ADRs and git history.

SPONSORED ARTICLE

InfoQ Dev Summit in Boston: Two Days of Talks for Senior Developers

As a senior software developer, understanding and adapting to the latest technologies and techniques is crucial for making informed decisions within your development projects. The need to keep learning, identify shifting trends, and understand how to prioritize software decisions is a constant challenge.

InfoQ is delighted to announce a new two-day conference, InfoQ Dev Summit Boston 2024, taking place June 24-25, 2024. This event is designed to help senior developers navigate their immediate development challenges, focusing exclusively on the technical aspects that matter right now.

Why Attend InfoQ Dev Summit Boston 2024?

Practitioner-Curated Topics

The InfoQ Dev Summit features an agenda **curated by senior developers** focusing on critical technical challenges encountered by today's software development teams.

Actionable Technical Talks

The summit offers over **20 technical sessions** by senior developers across two days, providing practical, actionable advice for immediate application to their current development decisions and project prioritization.

Content You Can Trust

InfoQ Dev Summit provides **genuine learning experiences with no hidden marketing**, offering real-world talks from senior developers with valuable advice.

'We Care' Experience

The InfoQ Dev Summit ensures an inclusive, safe space with a strict code of conduct, valuing everyone's contributions.

Ready to level up?

Early bird savings and team discounts are available until February 13th. For teams of three or more attending from the same company, email info@devsummit.infoq.com, mentioning the group size, to receive a personalized discount.

The InfoQ Dev Summit creates a unique opportunity for senior software developers to learn transformative insights that will directly impact their immediate development roadmap and decision-making processes.

The InfoQ Team looks forward to connecting with you in Boston, June 24-25, 2024, for the InfoQ Dev Summit.

InfoQ DevOps and Cloud Trends Report

by **Steef-Jan Wiggers, Matt Campbell, Daniel Bryant, Helen Beal, Abby Bangser**

The InfoQ Trends Reports provide InfoQ readers with an opinionated high-level overview of the topics we believe architects and technical leaders should pay attention to. In addition, they also help the InfoQ editorial team focus on writing news and recruiting article authors to cover innovative technologies.

In addition to this report and the updated DevOps and Cloud InfoQ trends graph, an accompanying podcast is available that features several editors and friends of InfoQ discussing these trends.

Updates on the trends graph

More details follow later in the report, but first, it is helpful to summarize the changes from last year's trends graph.

FinOps, the practice of managing cloud costs effectively, is moving toward the early majority of adoption. The FinOps Foundation and cloud companies like Microsoft, AWS, and Google promote adopting FinOps practices, which align with sustainability goals and optimize resource usage. Recently, Google became certified as a FinOps Certified Service Provider, and Microsoft joined the FinOps organization as a premier member.

The continued evolution of WebAssembly (Wasm) is delivering on the promise toward achieving "write once, run anywhere" in the cloud, offering reusability and interoperability across different languages and platforms. eBPF (Extended Berkeley Packet Filter) is gaining traction in areas like observability and security at the kernel level.

We have observed that the concept of generic Function as a Service (FaaS) and Backend as a Service (BaaS) is gaining traction

among the «late majority.» The adoption of serverless technologies and techniques has become commonplace. The statement «we are 100% serverless» no longer carries the same level of astonishment it once did, as serverless has become a mainstream approach in the industry.

Is the cloud domain moving from revolution to evolution? And is DevOps dead?

In the accompanying cloud and DevOps trends podcast discussion, the participants address the state of cloud innovation and DevOps. They agree that cloud innovation has slowed down, moving from "revolution" to "evolution". While large numbers of organizations have adopted cloud technologies, there are many enterprises that want to migrate and re-architect workloads.

As for DevOps, it is still alive but has reached a stage of stagnancy in some organizations. The concept of DevOps, which aims to provide access and autonomy to create business value, is still alive, but the implementation has faced challenges. The panelists mentioned their interest in Value Stream management to unlock DevOps's flow and value realization.

The public cloud vendors have evolved from their original goal of providing on-demand access to scalable resources to focus more on offering managed services. This evolution has made cloud computing more ubiquitous. However, technology is changing rapidly around existing services, new business requirements are being discovered, and new challenges are emerging. Teams must balance adopting and updating technology stacks while continually delivering business value. InfoQ Lead DevOps Editor and panelist Matthew Campbell said:

Businesses want to evolve and adapt quickly as well. [...] we're now in a phase where we're trying to figure out how do we sustainably leverage all of the cool stuff that we've invented and created and all these ways of interacting with each other and move it to a place where we can innovate comfortably going forward.

In addition, cloud services are now significantly adopted throughout small and large organizations, even late adopters, and the COVID-19 pandemic was often a forcing function. For instance, the evolution of automated setup environments, such as the ability to quickly set up complete development and test environments, is now commonplace. However, challenges still exist in bridging the gap between development and operations. Identity and access management issues create a perceived boundary between the dev and ops teams.

What is the current impact of AI and LLMs on the domains of Cloud and DevOps?

The panelists discussed cognitive overload and how AI can alleviate it by addressing cognitive load limits. A specific application of AI called AIOps, which focuses on AI for IT operations, is highlighted for its effectiveness in instant management and ticketing systems. Large language Models (LLMs) have tangible benefits, such as using ChatGPT to validate information, generate teachin notes, and aid in writing and creative processes. For instance, Microsoft has integrated AI into its products and services, showcasing its significant investment in AI technology. InfoQ Lead Cloud Editor and panelist Steef-Jan Wiggers stated:

A lot of the services Microsoft offers, even something recent, the Microsoft Fabric, a complete SaaS Data Lake or Lakehouse solution they have, is thoroughly infused with AI.

Microsoft Fabric is just an example, and other public cloud providers offer AI-infused services like Amazon Sagemaker and Google's Vertex AI and AutoML). Additional information on OpenAI can be found by following this topic on InfoQ.

How do AI-based and ChatGPT-like products impact the low-code and no-code domains?

Integrating AI into low-code tools is a business opportunity, with AI supporting business users by providing safe and valuable knowledge. This challenges previous concerns about shadow IT and encourages collaboration between product management and software engineering teams.

Furthermore, there is the idea of "ClickOps," where low-code platforms enable users to interact through clicking while generating version-controlled, declarative, and adaptable code. For instance, improving the code generation capabilities of AI tools like GitHub Copilot and Codeium allows the generation of readable code that adheres to organizational standards and can evolve. The evolution of LLMs and AI-driven code generation will bring exciting advancements in the low-code domain.

Lastly, governance and data access in the low-code landscape is essential. It comes with challenges associated with providing business users with power and access to data while ensuring proper governance and compliance. Campbell points out that there is a need for a "DevOpsy" layer of governance within low-code augmented platform engineering that provides guardrails and prevents certain configuration boundaries being crossed.

In addition, Helen Beal, DevOps Strategic Advisor, Chief Ambassador at DevOps Institute, and panelist, added:

I found it quite interesting that we are turning a corner because AI is actually supporting businesspeople, giving them an amount of knowledge that is probably safe.

How will Platform Engineering evolve?

The evolution of platform engineering involves a shift toward simplification, focusing on value delivery, and adopting a platform-as-a-service mindset. This change entails providing self-service platforms that hide complexity and reduce cognitive load for application developers. The role of platform engineering teams is evolving from being the keepers of complex infrastructure to becoming service providers to the rest of the organization. They now focus on developer relations, marketing, and customer engagement to delight users and drive value.

Abby Bangser, a Principal Engineer at Syntasso and panelist, explains:

Platform engineering teams are learning what developer relations and marketing look like to engage with customers, get feedback, and have a roadmap that meets their needs.

Adopting technologies like Kubernetes is being pushed down the stack, with a growing emphasis on API interfaces and streamlining interactions. Additionally, there is a heightened focus on observability, including service level and key performance indicators, as well as the financial aspects of platform usage and cost justification. Overall, the future of platform engineering lies in building platforms that add value and create delightful experiences for users while addressing the evolving needs and constraints of the business.

More information on Platform Engineering can be found by following this topic on InfoQ.

Is FinOps moving to the early majority of adoption?

FinOps, managing cloud costs effectively, is moving toward the early majority of adoption. More companies are joining the FinOps Foundation, and many tools are available to support FinOps processes. However, it's important to note that FinOps is not just about tools but also about the process and understanding of the value derived from spending. The FinOps Foundation and cloud companies like Google and Microsoft actively participate in this journey and promote adopting FinOps practices.

The awareness of FinOps is growing, leading to discussions

about why certain cloud resources are being provisioned and run, and whether they are being effectively utilized. Sustainability and GreenOps are also related to FinOps, as the focus on optimizing costs aligns with the broader goal of resource efficiency. AI plays a role in FinOps by identifying unused data and helping optimize storage, contributing to financial savings and environmental benefits.

When building cloud-based applications or adopting DevOps practices, are architects and developers overloaded with security concerns?

Architects and developers face an increasing list of security concerns when building cloud-based applications or adopting DevOps practices. Developers, in particular, can feel overwhelmed by the shift left approach, where they are expected to identify and prioritize security issues throughout the development process.

While there is a growing awareness of the importance of security and a push from leadership to address these concerns, developers often need help to balance security requirements against the pressure to deliver new features.

The evolving nature of security tooling is also a factor in this landscape. Early solutions were designed by experts for experts, making them less user-friendly for developers. However, there is a growing recognition of the need for more accessible and user-friendly security tools. The goal is to make security an enablement function and build platforms that simplify security implementation while providing education and support to development teams. This approach aims to bridge the gap between expert-driven security implementations and the practical needs of developers working on the code.

Is WebAssembly (Wasm) a final realization of "write once, run anywhere" in the cloud?

Wasm is a significant step toward achieving the "write once, run anywhere" vision in the cloud. It promises reusability and interoperability, allowing developers to build libraries in one language (such as Go) and seamlessly call them from applications written in other languages that can compile down to Wasm (such as Rust).

This component model within the cloud enables the creation of applications for multiple platform targets, including ARM-based CPUs, which have gained popularity in cloud infrastructure due to their performance and cost advantages. The adoption of Wasm extends beyond application development and into cloud platform extension formats. It is used to extend cloud-native proxies, API gateways, and service meshes.

In addition to WebAssembly, eBPF is gaining traction as a platform component developer's tool. While application engineers may not extensively use this, eBPF can be found extensively in projects that contain networking and security use cases. It allows developers to access kernel-level information and gain insights into container system operations, enhancing observability and security capabilities.

Overall, WebAssembly and eBPF offer intriguing possibilities for achieving greater portability, reusability, and performance in cloud-based applications.

More information on Wasm and eBPF can be found by following these topics on InfoQ.

How widely adopted is OpenTelemetry for collecting metrics and event-based observability data?

OpenTelemetry, a framework for collecting metrics and event-based observability data, has seen rapid adoption and is becoming the de facto standard in the industry. The collaborative efforts of many talented individuals and vendors have contributed to its cross-vendor support and cross-language compatibility, making it an essential component of applications. OpenTelemetry's widespread adoption has been accelerated by its inclusion within major cloud vendor offerings, such as AWS (AWS

Distro for OpenTelemetry), Microsoft Azure (Monitoring service), and Google Cloud Platform (Google Cloud OpenTelemetry).

The standardized nature of OpenTelemetry brings numerous benefits; it is vendor-agnostic and it has the ability to export telemetry data and utilize various tools for analysis. This standardization encourages optimization and innovation among vendors as they strive to offer unique and advanced features beyond the baseline functionalities of data collection and visualization. OpenTelemetry's emergence as an open standard signifies the maturing of the industry and fosters healthy competition among vendors to provide compelling solutions and gain market share.

More information on OpenTelemetry can be found by following this topic on InfoQ.

What is the current state of the adoption level for Serverless?

Serverless technology has seen a shift in adoption levels where it is becoming a common choice rather than a distinct architectural concept. The term "Serverless" is less frequently used to discuss a standalone concept, as it has almost transformed into a synonym for managed services that offer scalability, micro-billing, and abstracted infrastructure. Major cloud providers like AWS, Google, and Microsoft have integrated Serverless components into their services, such as databases (DBaaS) and container runtimes (CaaS), emphasizing the benefits of auto-scaling and simplified billing structures. The focus has shifted from building architectures solely on Serverless functions to leveraging managed services, aligning with the platform engineering approach and reducing cognitive overload for developers.

The value of Serverless, such as scaling to zero and cost-per-request pricing, has found new expression beyond the traditional Serverless architecture. Organizations now recognize these benefits and are demanding them in various architectural decisions. While Serverless is one of many approaches to achieve these advantages, organizations are increasingly asking their engineering teams to deliver cost-effective solutions and optimize customer acquisition and support costs. This evolution highlights Serverless principles' growing influence and impact on the broader architectural landscape.

More information on Serverless can be found by following this topic on InfoQ.

How is the focus on sustainability and green computing impacting cloud and DevOps?

The focus on sustainability and green computing is having a significant impact on cloud and DevOps practices. There is a growing adoption of pricing models considering the environmental impact and resource consumption of applications and services. This trend encourages organizations to make architectural choices that prioritize efficiency and sustainability. Managed services are favored as they offer optimized resource utilization and scalability, allowing businesses to minimize their carbon footprint and reduce energy consumption. The consideration of pricing about architecture and the adoption of managed services aligns with sustainability and green computing goals.

Regarding responsibility, there is a recognition that addressing sustainability falls within the realm of Site Reliability Engineering (SRE) and related roles. These teams are well-positioned to analyze the environmental impact of technology decisions and drive initiatives that promote efficiency and sustainability. Architecture discussions now encompass considerations such as componentization, isolation, security, and cost-efficiency. Organizations are evaluating their requirements and seeking middle-ground solutions that

meet security needs without unnecessarily high costs. This reflects a shift toward more pragmatic approaches to security, finding the right balance between enterprise-grade features and cost-effectiveness.

What are our predictions for the future of the cloud and DevOps spaces?

The panelist' predictions for the cloud and DevOps spaces' future revolved around simplifying, reducing cognitive overload, and focusing on innovation. There is a desire to streamline processes and tools to allow teams to concentrate on their specific areas of expertise and maximize their impact.

The convergence of AIOps, platform engineering, sustainability, and FinOps is a positive shift that may lead to more focused, more effective, and happier teams. The challenge lies in distinguishing between hype and genuine opportunities, acknowledging the “nuggets of value” within emerging trends while remaining critical of the “overselling” and broad applicability claims.

Open-source adoption, the standardization facilitated by initiatives like OpenTelemetry and CloudEvents, and the potential of AI-infused services such as Copilots and ChatGPT are all points of excitement. Overall, there is a sense of enthusiasm for the ongoing developments and the learning opportunities they bring.

TL;DR

- Cloud innovation has transitioned from a revolutionary phase to an evolutionary one, focusing on migrating and re-architecting workloads. The cloud space has evolved toward providing on-demand access to scalable resources and managed services, emphasizing simplifying interactions and reducing cognitive load for teams.
- Artificial Intelligence (AI) and Large Language Models (LLMs) may play a significant role in the domains of cloud and DevOps by addressing cognitive overload and supporting tasks like instant management, ticketing systems, and code generation. Major cloud providers like Microsoft, Google, and AWS have integrated AI into their products and services, showcasing the industry's investment in AI technology.
- Low-code and no-code domains are impacted by AI-based and ChatGPT-like products, offering collaboration opportunities between business users and software engineering teams.
- Platform engineering is evolving toward simplification and value delivery, adopting a platform-as-a-service mindset. The role of platform engineering teams is shifting from complex infrastructure management to becoming service providers focused on user satisfaction and value creation. Observability, financial aspects, and sustainability considerations are becoming integral to platform engineering.
- OpenTelemetry is widely adopted for collecting metrics and event-based observability data, becoming the de facto standard in the industry. Its standardized nature encourages optimization and innovation among vendors.
- The focus on sustainability and green computing drives architectural choices toward efficiency and minimizing carbon footprints. Site Reliability Engineering (SRE) teams are crucial in analyzing environmental impact and promoting sustainability initiatives. comb through ADRs and git history.

InfoQ AI, ML, & Data Engineering Trends Report

by **Srini Penchikala, Roland Meertens, Sherin Thomas, Anthony Alford, Daniel Dominguez**

The InfoQ Trends Reports provide InfoQ readers with an opinionated high-level overview of the topics we believe architects and technical leaders should pay attention to. In addition, they also help the InfoQ editorial team focus on writing news and recruiting article authors to cover innovative technologies.

In this annual report, the InfoQ editors discuss the current state of AI, ML, and data engineering and what emerging trends you as a software engineer, architect, or data scientist should watch. We curate our discussions into a technology adoption curve with supporting commentary to help you understand how things are evolving.

In this year's podcast, InfoQ editorial team was joined by external panelist Sherin Thomas, software engineer at Chime. The following sections in the article summarize some of these trends and where different technologies fall in the technology adoption curve.

Generative AI

Generative AI, including Large Language Models (LLMs) like GPT-3, GPT-4, and Chat GPT, has become a major force in the AI and ML industry. These technologies have garnered significant attention, especially given the progress they made over the last year. We have seen wide adoption of these technologies by users, in particular driven by ChatGPT. Multiple players such as Google and Meta have announced their own generative AI models.

The next step we expect is a larger focus on LLMOps to operate these large language models in an enterprise setting. We are divided in whether prompt engineering will be a large topic in the future or whether the adoption will be so widespread

that everyone will be able to contribute to the prompts used.

Vector Databases and Embedding Stores

With the rise in LLM technology there's a growing focus on vector databases and embedding stores. One intriguing application gaining traction is the use of sentence embeddings to enhance observability in generative AI applications.

The need for vector search databases arises from the limitations of large language models, which have a finite token history. Vector databases can store document summaries as feature vectors generated by these language models, potentially resulting in millions or more feature vectors. With traditional databases, finding relevant documents becomes challenging as the dataset grows. Vector search databases enable efficient similarity searches, allowing users to locate the nearest neighbors to a query vector, enhancing the search process.

A notable trend is the surge in funding for these technologies, signaling investor recognition of their significance. However, adoption among developers has been slower, but it's expected to pick up in the coming years. Vector search databases like Pinecone, Milvus, and open-source solutions like Chroma are gaining attention. The choice of database depends on the specific application and the nature of the data being searched.

In various fields, including Earth observation, vector databases have demonstrated their potential. NASA, for instance, leveraged self-supervised learning and vector search technology to analyze satellite images of Earth, aiding scientists in tracking weather phenomena such as hurricanes over time.

Robotics and Drone Technologies

Cost of robots is going down. In the past legged balancing robots were hard to acquire, but there are already some models available for around 1,500 dollars. This allows more users to use robot technologies in their applications. The Robot Operating System (ROS) is still the leading software framework in this field, but companies like VIAM are also developing middleware solutions that make it easier to integrate and configure plugins for robotics development.

We expect that advances in unsupervised learning and foundational models will translate into improved capabilities. For example, by integrating a large language model into the path planning part of the robot to enable planning using natural language.

Responsible and Ethical AI

As AI starts to affect all of humanity there is a growing interest in responsible and ethical AI. People are simultaneously calling for stricter safety around large language models, as well as being frustrated by the output of such models reminding users of the safeguards in place.

It remains important for engineers to keep in mind to improve the lives of all people, not just a select few. We expect a similar impact from AI regulation as GDPR had a few years ao.

We have seen some AI fail because of bad data. Data discovery, operations, data lineage, labeling and good model development practices are going to take center stage. Data is crucial to explainability.

Data Engineering

The state of modern data engineering is marked by a dynamic shift towards more decentralized and flexible approaches to manage the ever-growing volumes of data. Data Mesh, a novel concept, has emerged to address the challenges posed by centralized data management teams becoming bottlenecks in data operations. It advocates for a federated data platform partitioned across domains, where data is treated as a product. This allows domain owners to have ownership and control over their data products,

reducing the reliance on central teams. While promising, Data Mesh adoption may face hurdles related to expertise, necessitating advanced tooling and infrastructure for self-service capabilities.

Data observability has become paramount in data engineering, analogous to system observability in application architectures. Observability is essential at all layers, including data observability, especially in the context of machine learning. Trust in data is pivotal for AI success, and data observability solutions are crucial for monitoring data quality, model drift, and exploratory data analysis to ensure reliable machine learning outcomes. This paradigm shift in data management and the integration of observability across the data and ML pipelines reflect the evolving landscape of data engineering in the modern era.

Explaining the updates to the curve

With this trends report also comes an updated graph showing what we believe the state of certain technologies is. The categories are based on the book "Crossing the Chasm", by Geoffrey Moore. At InfoQ we mostly focus on categories which have not yet crossed the chasm.

One notable upgrade from innovators to early adopters are the "AI Coding Assistants". Although they were very new last year and hardly used, we see more and more companies offering this as a service to their employees to make them more efficient. It's not a default part of every stack, and we are still discovering how to use them most effectively, but we believe that adoption will continue to grow.

Something which we believe is crossing the chasm right now is natural language processing. This will not come as a surprise to anyone as many companies are currently trying to figure out how to adopt generative AI capabilities in their product offering following the massive success of ChatGPT. We thus decided to make it cross the chasm already into the early majority category. There is still a lot of potential for growth here, and time will teach us more what the best practices and capabilities are for this technology.

There are some notable categories who did not move at all. These are technologies such as synthetic data generation, brain-computer interfaces and robotics. All of these seem to be consistently stuck in the innovators category. The most promising in this regard is the synthetic data generation topic, which is lately getting more attention with the GenAI hype. We do see more and more companies talking about generating more of their training data, but have not seen enough applications actually using it in their stack to warrant it moving to the early adopters category. Robotics has been getting a lot of attention for multiple years now, but its adoption rate is still too low for us to warrant a movement.

We also introduced several new categories to the graph. A notable one is vector search databases, something which comes as a byproduct of the GenAI hype. As we are gaining more understanding of how we can represent concepts as a vector there is also more need for efficient storing and retrieving said vectors. We also added explainable AI to the innovators category. We believe that computers explaining why they made a certain decision will be vital for widespread adoption to combat hallucinations and other dangers. However, we currently don't see enough work in the industry to warrant a higher category.

Conclusion

The field of AI, ML, and Data Engineering keeps on growing year over year. There is still a lot of growth in both the technological capabilities as well as the possible applications. It's exciting for us editors at InfoQ to be so close to the progress, and we are looking forward to making the same report next year.
In the podcast we make several

predictions for the coming year, which range from «there will be no AGI» to «Autonomous Agents will be a thing». We hope you enjoyed listening to the podcast and reading this article, and would love to see your predictions and comments below this article.

TL;DR

- Generative AI, powered by Large Language Models (LLMs) like GPT-3 and GPT-4, has gained significant prominence in the AI and ML industry, with widespread adoption driven by technologies like ChatGPT.
- Major tech players such as Google and Meta have announced their own generative AI models, indicating the industry's commitment to advancing these technologies.
- Vector databases and embedding stores are gaining attention due to their role in enhancing observability in generative AI applications.
- Responsible and ethical AI considerations are on the rise, with calls for stricter safety measures around large language models and an emphasis on improving the lives of all people through AI.
- Modern data engineering is shifting towards decentralized and flexible approaches, with the emergence of concepts like Data Mesh, which advocates for federated data platforms partitioned across domains.

InfoQ Java Trends Report

by **Michael Redlich, Ixchel Ruiz, Alina Yurenko, Rustam Mehmandarov, Johan Jenssen**

This report provides a summary of how the InfoQ Java editorial team currently sees the adoption of technology and emerging trends within the Java space. We focus on Java the language, as well as related languages like Kotlin and Scala, the Java Virtual Machine (JVM), and Java-based frameworks and utilities. We discuss trends in core Java, such as the adoption of new versions of Java, and also the evolution of frameworks such as Spring Framework, Jakarta EE, Quarkus, Micronaut, Helidon, MicroProfile and MicroStream.

You can also listen to the additional podcast discussion around Java Trends for 2023.

This report has two main goals:

- To assist technical leaders in making mid- to long-term technology investment decisions.
- To help individual developers in choosing where to invest their valuable time and resources for learning and skill development.

This is our fifth published Java trends report. However, this topic has received ample news coverage as we have been internally tracking Java and JVM trends since 2006.

To help navigate current and future trends at InfoQ and QCon, we make use of the "crossing the chasm" mental model for technology success pioneered by Geoffrey Moore in his book of the same name. We try to identify ideas that fit what Moore referred to as the early market, where «the customer base is made up of technology enthusiasts and visionaries who are looking to get ahead of either an opportunity or a looming problem.»

As we have done for the 2022, 2021, 2020 and 2019 Java trend reports, we present the internal topic graph for 2023.

Aside from several new technologies having been identified in the **Innovator** category, notable changes are described as follows.

Java 17+ has been recategorized as simply *Java 17* and remains in the **Early Adopter** phase as more frameworks have committed to Java 17 as a baseline. *Java 21* has been introduced in the **Innovator** category.

We have created a new label, *Fast JVM Startup*, with further refinements, *Fast JVM Startup (CRaC)*, placed in the **Innovators** category, and *Fast JVM Startup (GraalVM)*, placed in the **Early Adopters** phase. The rationale was to acknowledge the relatively new technologies that have been recently introduced to the Java community.

What follows is a lightly edited summary of the corresponding discussion on various topics among several InfoQ Java Queue editors and Java Champions:

- Michael Redlich, Director at Garden State Java User Group and Java Queue Lead Editor at InfoQ. Retired Senior Research Technician at ExxonMobil Technology & Engineering Company
- Johan Janssen, Software Architect at ASML and Java Queue Editor at InfoQ
- Ixchel Ruiz, CDF Foundation Ambassador at The Linux Foundation
- Alina Yurenko, Developer Advocate for GraalVM at Oracle Labs
- Rustam Mehmandarov, Chief Engineer at Computas AS

We also acknowledge the Java Queue editors who provided input on updating our "crossing the chasm" model for 2023:

- Ben Evans, Senior Principal Software Engineer at Red Hat, Java Queue Editor at InfoQ and Java Champion
- Erik Costlow, Senior Director of Product Management and Java Queue Editor at InfoQ
- Karsten Silz, Senior Full-Stack Java Developer and Java Queue Editor at InfoQ
- Olimpiu Pop, Chief Technology Officer at mindit.io
- Bazlur Rahman, Software Engineer and Java Champion
- Shaaf Syed, Senior Principal Technical Marketing Manager at Red Hat

We believe this summary provides more context for our recommended positioning of some of the technologies on the internal topic graph.

GraalVM/Cooridanted Restore at Checkpoint (CRaC)

Janssen: All the improvements in GraalVM and CRaC (Coordinated Restore at Checkpoint) to reduce the startup time of Java applications are impressive. It's great to see the continuous improvements to GraalVM, and the integration with many frameworks makes it easy to use native images within your application. Apart from native image support, GraalVM also offers a Java runtime engine which may be used instead of the JVM from your vendor and may result in better performance for your application by just changing the runtime.

Redlich: Apart from the releases of JDK 20 and JDK 21, I believe the most significant change comes from Oracle Labs and GraalVM. Over the past year, we've seen: applicable parts of GraalVM technology being contributed to OpenJDK; the creation of Project Galahad, a project that will initially focus on the continued development and integration of the Graal just-in-time (JIT) compiler as an alternative to the existing HotSpot JIT compiler for possible inclusion in a future OpenJDK release; releases of GraalVM aligned with the releases of OpenJDK; and the elimination of GraalVM Enterprise in favor of a new license.

Java 17 and Beyond

Ruiz: The release cadence is not only bringing in new features in a more digestible way, but it is also allowing different users to pick them up and try them out. So overall, I have seen a good attitude towards early testing and richer feedback.

In a way, it has also simplified the roadmap for updating the Java version in production. The predictability allows for better synchronisation of development teams.

Yurenko: I see the speed of adoption of the latest Java versions increasing. This is something I often see being discussed at conferences, reflected in questions I receive, and also observed in the GraalVM Community Survey the GraalVM team ran last year -- 63% of our users were already on Java 17 or higher.

Mehmandarov: We have Java 20 and Java 21 this year. Some of the most noticeable features are Record Patterns and Pattern Matching for switch that are finally out of preview. Those can be exciting features working with large amounts of data and can simplify the code.

Native Java (GraalVM/Spring Native/Project Leyden)

Yurenko: I might be biased, but I see a lot of projects and libraries adopting GraalVM, particularly Native Image. Spring Boot now supports Native Image out of the box along with other popular Java frameworks, and I see many libraries that have added support as well.

Java for Beginners

JEP 445, Unnamed Classes and Instance Main Methods (Preview), delivered in JDK 21, was inspired by the September 2022 blog post, Paving the on-ramp, by Brian Goetz, the Java language architect at Oracle. This feature will "evolve the Java language so that students can write their first programs without needing to understand language features designed for large programs."

JEP 463, Implicitly Declared Classes and Instance Main Methods (Second Preview), was recently promoted from its **JEP Draft 8315398** to **Candidate** status, and we anticipate that it will be targeted for JDK 22. Formerly known as *Unnamed Classes and Instance Main Methods (Preview)*, *Flexible Main Methods and Anonymous Main Classes (Preview)* and *Implicit Classes and Enhanced Main Methods (Preview)*, this JEP incorporates enhancements in response to feedback from JEP 445. Gavin Bierman, a consulting member of the technical staff at Oracle, has published the first draft of the specification document for review by the Java community.

Yurenko: Another great trend related to this one which I love, is how Java becomes more accessible for beginners. I think it's very important for a community to stay open and welcoming for beginners, whether that's students or newcomers to the industry, and features such as Records, Pattern Matching, Unnamed Classes, and Instance Main Methods allow beginners to learn faster, develop their first app more easily, and become increasingly productive.

What is the Java Community Saying?

Ruiz: A mixed bag, as many people were focused on the 21st release of Java. Project Loom has been tantalising many developers since the first previews. Virtual Threads and Structured Concurrency.

Others have noted the progress of JVM advances that are not strictly tied to syntax changes in the Java language.

Plans for migration from Java 8 to a newer version of Java. A "now or never" attitude!

Yurenko: I recently saw a new interesting way to analyze the trends in our community – Marcus Hellberg, VP of Developer Relations at Vaadin, nalyzed talks presented at four major Java conferences in 2023. You can check out the article yourself, and here are a few of my conclusions:

- Microservices and Kubernetes are still the hottest topics
- AI and ML in the third spot confirm my observation about how hot this topic is now
- I always see a lot of interest in framework talks, in this report, it's Spring Boot and Quarkus
- What's rather surprising for me is that the security topic is in the fifth spot

Another trend that I see is conversations about reducing the startup time of JVM applications and the evergreen topic of performance; for me, that's a sign that we are doing the right thing :)

Mehmandarov: Lately, there have been numerous talks and general excitement around Virtual Threads, which is finally out of preview with Java 21. We also see more development and application of Generative AI, especially for code generation. It still needs more maturity, but it is an exciting start. We also see more development of SDKs or frameworks for AI and ML with Java, like Semantic Kernel, Deeplearning4J, djl, and Tribuo.

Janssen: The Java ecosystem is still booming with many new developments and improvements. Apart from the things already mentioned, there are interesting developments, mainly around AI, which is evolving quickly, and it's great to see projects such as Spring AI instead of all the Python-based solutions. Next to the big hype topics, there are many other interesting projects like Spring Modulith, which is now an official Spring project and allows the creation of better monoliths instead of microservices.

Redlich: Most of the buzz that I have been hearing and seeing throughout 2023 is Project Loom. In particular, Virtual Threads was a final feature with the release of JDK 21. Leading up to this anticipated release in September 2023, there were numerous presentations and YouTube videos on virtual threads. At a special Java Community Process (JCP) 25th anniversary event held in New York City in September 2023, a panel of JCP Executive Committee members revealed their favorite feature of JDK 21, in which they unanimously said virtual threads.

What is New and Exciting That We Didn't Expect?

Ruiz: Given all this attention to LLM, ML and AI, I would not be surprised to see new projects, libraries and APIs in Java to support use cases, workflows and products.

Mehmandarov: If I have to pick one thing, it is Generative AI, specifically focusing more on code generation. While it struggles with logical errors in the generated code, hallucinations, and other issues, it still fits nicely into the "unexpected, new, and exciting" category.

Janssen: It's awesome to see many new features in Java 21, such as virtual threads from Project Loom. I was a bit afraid that some of those features would be postponed to later Java releases. As Java 21 is the new Long Term Support (LTS) version, it's great that they are included, as many companies only use LTS versions. At first glance, it looks like nothing big has been removed, and hopefully this results in easy upgrades for our projects. Those upgrades are nowadays even easier with OpenRewrite, which allows automated upgrades not only for the Java language itself, but also for libraries such as JUnit.

Redlich: The new MicroProfile JWT Bridge specification, currently being developed, is a collaboration between Jakarta EE and MicroProfile Working Groups. This new specification will enable Jakarta Security applications to build on the MicroProfile JWT Authentication specification that will provide seamless integrations and eliminate duplication of effort and circular dependencies. The goal is to move the optional section of MicroProfile JWT Authentication to the new bridge specification together with TCKs and to have

this specification ready for MicroProfile 7.0.

What's Getting You, Personally, Really Excited in the Java Space?

Ruiz: The resurgence of CLI tools in the JVM space. We have seen examples of mature projects that solve or reduce the friction for developers to try, test, release and publish tools, projects and products. JBang and JReleaser are part of this set of resurgent tools.

Yurenko: I like seeing many new projects occurring every day. For example, artificial intelligence and machine learning are probably the hottest trends now, and there are many opportunities in this field for Java developers. One of my favorites is Tribuo, an open source ML library developed by my colleagues at Oracle Labs. OpenJDK projects Valhalla and Panama will also greatly benefit Java developers working with AI.

Mehmandarov: There are quite a few things to get excited about. Some of them are new and upcoming, like String Templates (still in preview) and various libraries supporting Machine Learning and Java, and some of them are more mature but still see improvements and new development, like the developments in the Cloud-Native stack for Java (like Jakarta EE and MicroProfile).

When it comes to working with large datasets, I am also excited to see more concepts like Data-Oriented Programming in Java as well as the improvements to more memory usage efficiency, like projects Lilliput and value objects from project Valhalla.

Also, the importance and the excitement I get from interacting with the Java community worldwide. It is a genuinely vibrant and supporting group eager to learn and share their knowledge.

Redlich: I have been putting together a presentation entitled "**Jakarta EE 11: Going Beyond the Era of Java EE**" that I will present numerous times starting in November 2023. It's amazing how Jakarta EE has evolved since 2018, and it's been awesome studying the Jakarta EE specifications.

The Java Community

Janssen: Of course, every year, we get two new Java releases packed with features and lots of improvements in all the tools, libraries and frameworks. Next to that, it's good to see that the Java or JVM conferences are again happening and attracting more attendees, so make sure to visit them to learn more about Java and have some great discussions with fellow developers.

Conclusion

Please note that the viewpoints of our contributors only tell part of the story. Different groups, segments, and locales of the Java ecosystem may have different experiences. Our report for 2023 should be considered as a starting point for debate rather than a definitive statement, and an invitation to an open discussion about the direction the industry is taking.

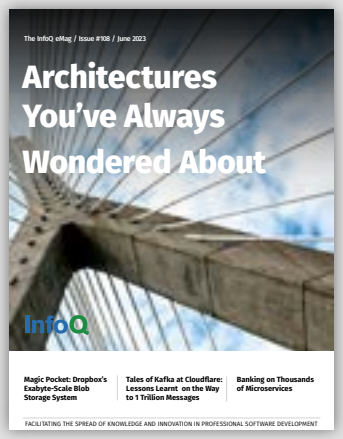
TL;DR

- Java Virtual Threads was finalized in the recently released JDK 21, and we believe that adoption of this feature will continue to grow as the latest edition of application frameworks, such as Helidon Níma and Vert.x, have already taken advantage of this.
- Oracle has made a commitment to evolve the Java language so that students and beginners can more easily learn to write their first “Hello, world!” applications without the need to understand more complex features of the language.
- Project Galahad, launched late last year, continues to aim to contribute Java-related GraalVM technologies to the OpenJDK Community and prepare them for possible incubation in a JDK mainline release.
- There is increasing community interest in learning about modern microservice frameworks, such as Spring Boot, Quarkus, and Jakarta EE. The Spring Modulith project is now an official Spring project and allows the creation of better monoliths instead of microservices.
- Since the release of Java 17, we've noticed faster adoption of newer Java versions than we did when Java 11 was released.
- We see increasing development and application of Generative AI in the Java space, especially for code generation. We also see more development of SDKs or frameworks for AI and ML with Java, like Semantic Kernel, Deeplearning4J, djl, and Tribuo.

Read recent issues

Data Engineering Innovations

In the InfoQ "Data Engineering Innovations" eMag, you'll find up-to-date case studies and real-world data engineering solutions from technology SME's and leading data practitioners in the industry.



Architectures You've Always Wondered About

This eMag showcases real-world examples of innovator companies pushing the limits with modern software systems. At recent conferences in San Francisco and London, the speakers clearly showed what "scalable" can really mean, from a trillion messages to exabytes of data. This eMag brings together several of these stories and hopefully provides advice and inspiration for your future projects.

The Platform Engineering Guide: Principles and Best Practices

Platform Engineering has quickly become one of the hottest topics in DevOps. The explosion of new technologies has made developing software more interesting but has substantially increased the number of things that development teams need to understand and own.

InfoQ

InfoQ @InfoQ InfoQ InfoQ