General Cloud Concepts

Becoming cloud-native in scientific computing

Sergio Alvarez-Napagao - Techno Week 29/05/2023

Summary of contents

Introduction

Who am I?

- BSc in Computer Science and PhD in Artificial Intelligence (Universitat Politècnica de Catalunya)
- Currently:
  - Lecturer at Universitat Politècnica de Catalunya
  - Senior Researcher at Barcelona Supercomputing Center
- Previously:
  - Researcher in Grid computing, service-oriented architectures and intelligent agents
  - Senior Team Lead at ThoughtWorks
  - Developer at SkyTV
  - Collaboration with startups (Abiquo, Lernmi)

Objectives of the session

- Understanding the basics of cloud computing
- Learning about the impact of cloud computing on scientific computing
- Familiarizing with common tools and platforms
- Understanding best practices and architectures enabled by cloud computing

Understanding Cloud Computing

A metaphor

Devon Spinner, photograph by Barry Lewis under CC BY 2.0 (https://www.flickr.com/photos/16179216@N07/8582838851)

A metaphor

Photograph by Jakob Barfod under CC SA 3.0 (https://commons.wikimedia.org/wiki/File:20061007_3P_N_PE_CEE_connections.jpg)

A metaphor

Dedicated resources	Shared resources
Physical resources	Virtual resources
In-house maintenance	Outsourced maintenance
Limited scalability	Elastic scalability
Upfront investment	Flexible pricing

A metaphor

Dedicated resources	Shared resources
Physical resources	Virtual resources
In-house maintenance	Outsourced maintenance

Limited scalability	Elastic scalability
Upfront investment	Flexible pricing

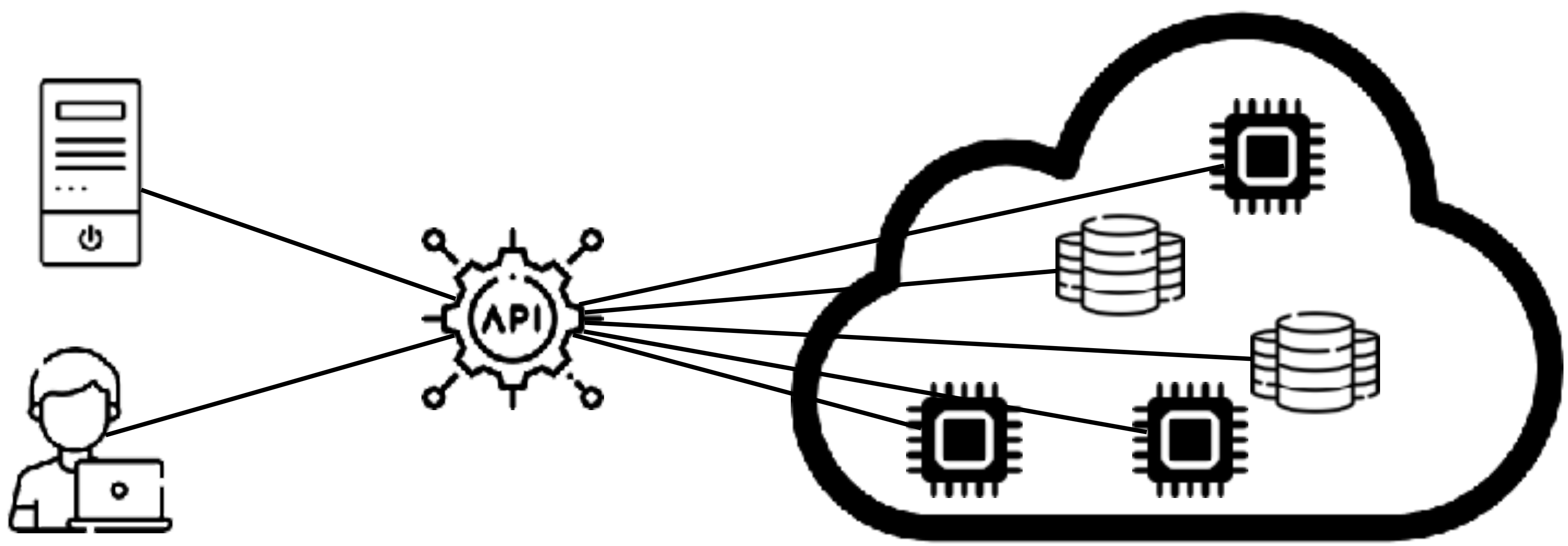
A definition

On-demand delivery of **computational resources and services**

Usually **maintained** by specialized providers

With **pay-as-you-go** pricing

Through **simplified provision interfaces**



History

Early milestones

1950s	Mainframe computing: shared access to large centralized computers
1961	John McCarthy's Idea - "Computation may someday be organized as a public utility.”
1969	ARPANET - The precursor to the modern Internet.
1980s	Rise of personal computing: affordable home computers
1990s	Massive supercomputing: first supercomputers with thousands of processors
1996	“Cloud computing” is used as a term for the first time (Compaq Computer Corp internal report)
2000s	Grid computing: using heterogeneous, multi-tenant resources to distribute complex problems
2004	Amazon Web Services launches as a collection of APIs and tools to access the catalog

History

Recent developments

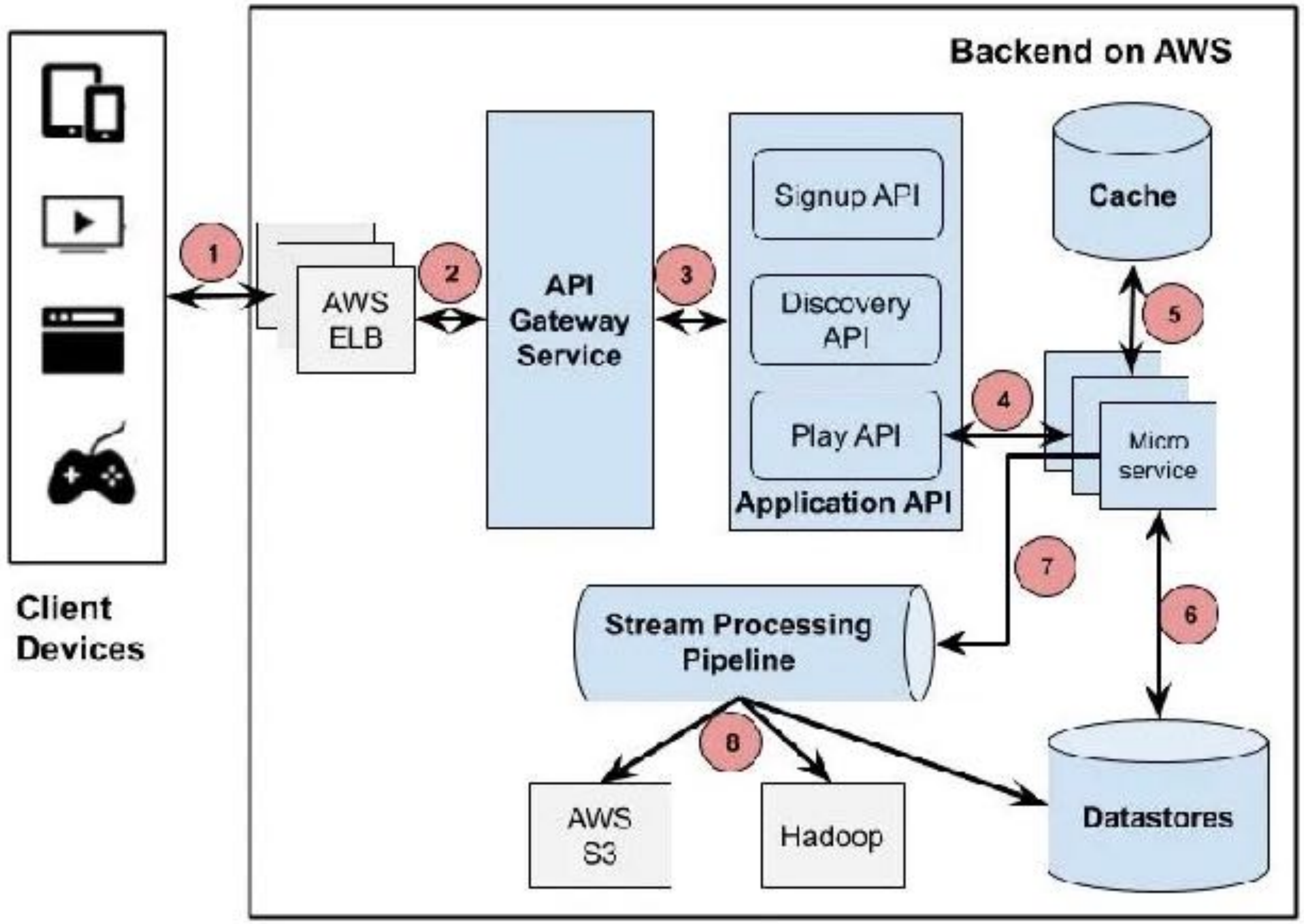
2006	Amazon Elastic Compute Cloud (EC2): First widely accessible cloud infrastructure service
2008	Dropbox: first fully integrated cloud storage service for end-users
2008	Google App Engine: Platform for developing and hosting web applications
2010	Microsoft Azure: strong focus on Platform-as-a-Service
2011	Google Cloud Platform: emphasis on data analytics and machine learning
2015	AWS Lambda: popularization of serverless computing
2017	Google Cloud adds Tensor Processing Units
2020	Widespread adoption of cloud computing
Present	Rise of edge computing and integration between hybrid clouds and HPC

Real-World example

Netflix

Netflix uses the cloud for:

- managing and serving massive volumes of content
- scaling requests to meet fluctuating demand
- processing data and requests at different levels of priority and computing requirements



https://medium.com/swlh/a-design-analysis-of-cloud-based-microservices-architecture-at-netflix-98836b2da45f

Benefits of Cloud Computing

Cost Efficiency: Cloud computing eliminates the need for hardware investment and reduces operating costs

Collaboration Efficiency: Cloud services enable efficient collaboration, allowing teams to work together from anywhere

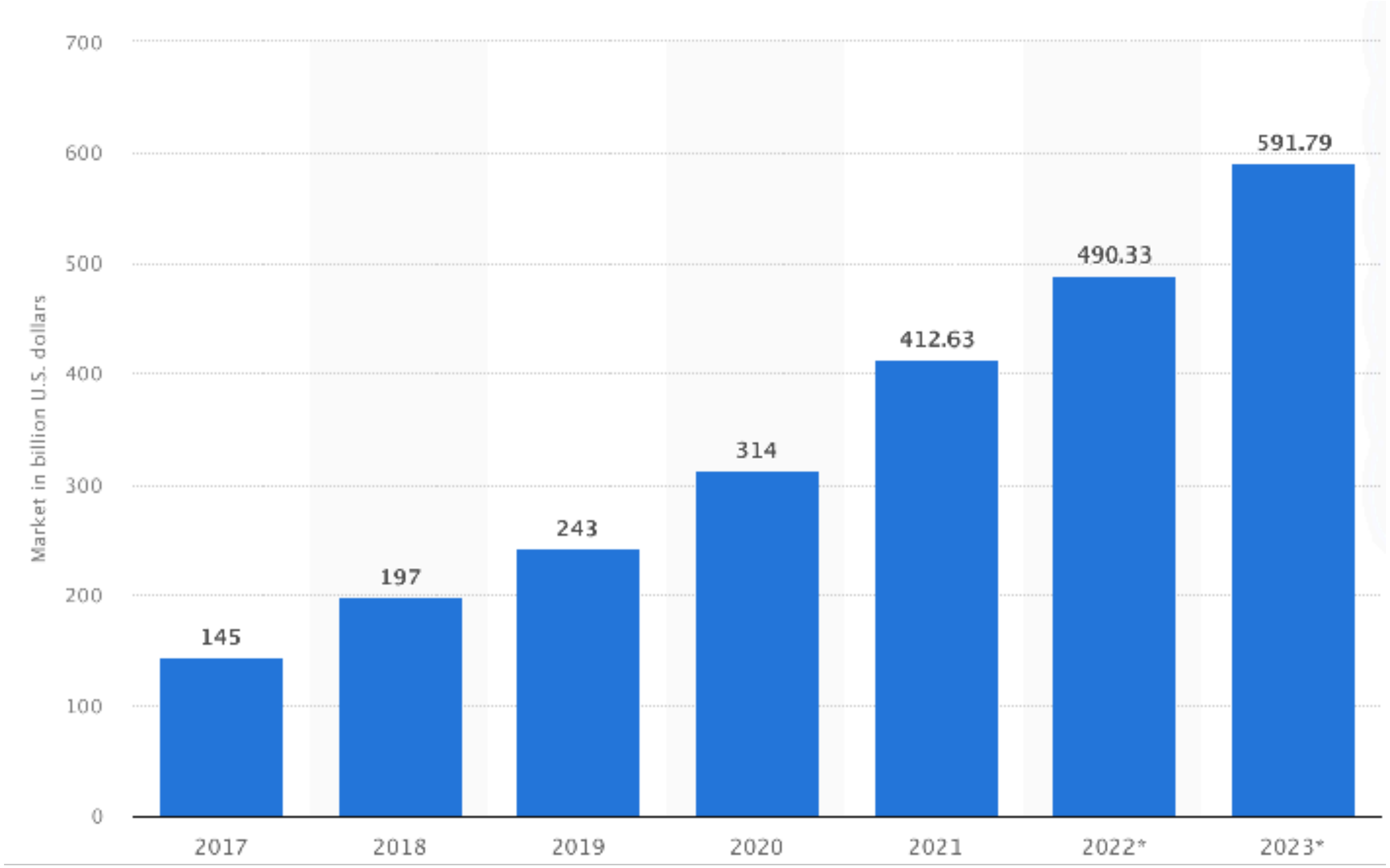
Scalability: Cloud services can be easily scaled up or down to match business needs

Disaster Recovery: Cloud-based backup systems are cost-effective and reliable, reducing the risk of data loss

Mobility: Access data and applications from anywhere, on any device with internet connectivity

Status of Cloud computing

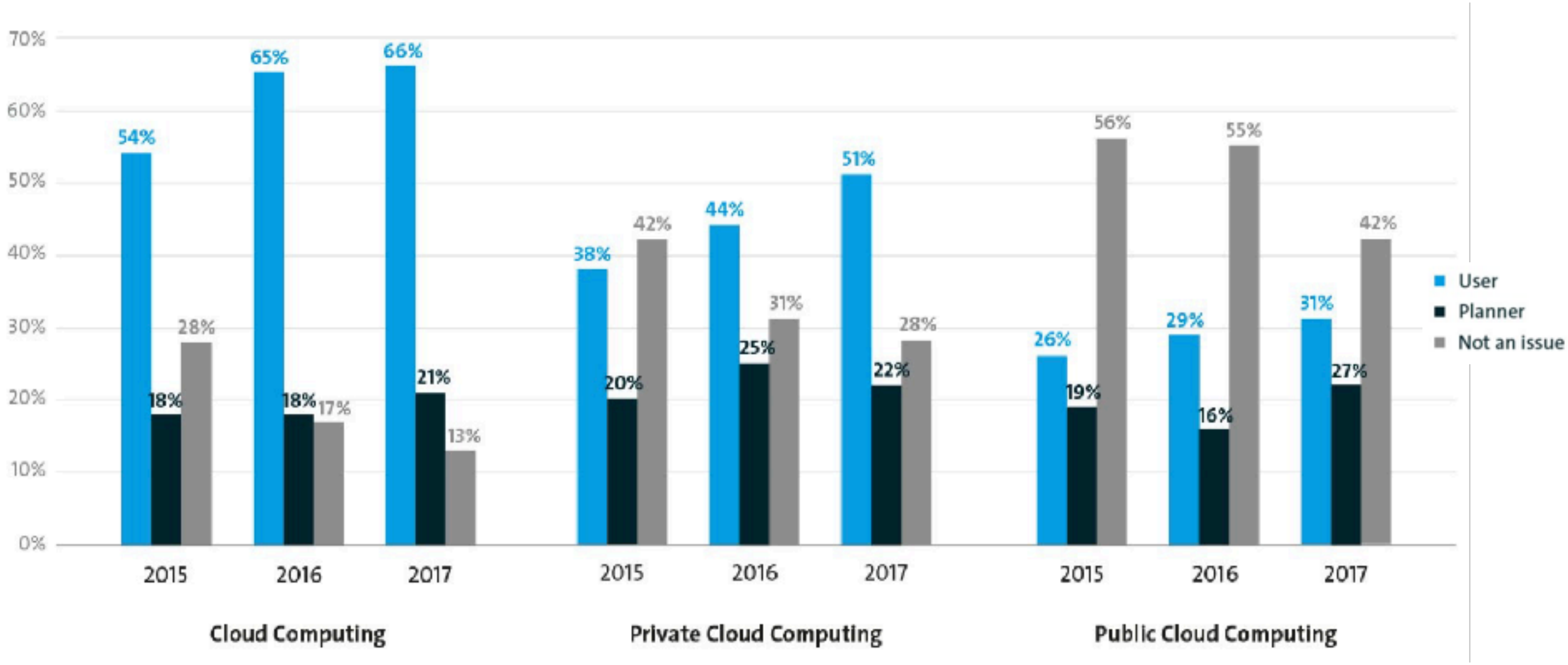
Market size



Source: “Forecast: Public Cloud Services, Worldwide, 2021-2027, 1Q23 Update.”, Gartner

Status of Cloud computing

Adoption



Source: "Cloud Monitor 2019", BitKom research,
https://www.bitkom.org/sites/default/files/2019-06/bitkom_kpmg_pk_charts_cloud_monitor_18_06_2019.pdf

Status of Cloud computing

Organizational performance

“Cloud usage is predictive of organizational performance”

Main predictive factors:

Software initially built on and for the cloud

Not using on-premises servers alone

Using multiple public clouds

Adoption of security best practices

“2022 Accelerate – State of DevOps Report”, Google Cloud & DORA,
https://services.google.com/fh/files/misc/2022_state_of_devops_report.pdf

Basic concepts

Virtualization

Cloud computing building blocks

Virtualization

Creation of simulated virtual computing environments

Local testing of infrastructure

Partition of large resource pools

Containerization

Virtualization method that packages an application and its dependencies together

Containers allow abstraction from infrastructure

Orchestration

Lifecycle management of containers

Automated configuration, provisioning, deployment and scaling

Cloud computing leverages:

virtualization to create isolated, scalable environments,

containerization for consistent and efficient application deployment, and

orchestration for automating and managing services at scale,

thereby providing a **flexible, cost-effective, and highly available computing platform**

Deployment models

Public Cloud

Resources are owned and operated by a third-party cloud service provider and delivered over the Internet

Private Cloud

Computing resources are used exclusively by a single business or organization

Hybrid Cloud

Combines public and private clouds, bound together by technology that allows data and applications to be shared between them

Also: **Multi-Cloud**

Service models: IaaS

Infrastructure as a Service

On-demand computing resources

- Virtual machines
- Virtual disks
  - Block storage
  - File storage
- Network infrastructure
  - IP addresses
  - VPNs
  - Firewalls
  - Load balancers

Billing based on provision and consumption of virtual hardware

Users have control over the infrastructure demanded

Amazon EC2

Google Compute Engine

Azure Virtual Machine

Service models: PaaS

Platform as a Service

- Provide access to platform, tools, frameworks, environments that enable users to:
  - **Develop**
  - **Run**
  - **Manage**

  applications

- In many cases, they transparently integrate common services with low maintenance:
  - Databases
  - Data analysis
  - Monitoring
  - Security & authentication

Infrastructure is managed by the cloud provider

Billing based on actual costs: computing cost, network usage, data transfer

App Engine

amazon SQS

Azure Cosmos DB

Azure DevOps

Service models: FaaS

Function as a Service

- Cornerstone of *“serverless computing”*
- The cloud provides a platform for developing **functions** that react to specific events

*Infrastructure **and software stack** are managed by the cloud provider*

Billing based on actual costs: computation cost of each invocation

AWS Lambda

Google Cloud Run

Service models: SaaS

Software as a Service

- The cloud provider delivers applications over the Internet, as a service
- Zero development, zero maintenance for the user

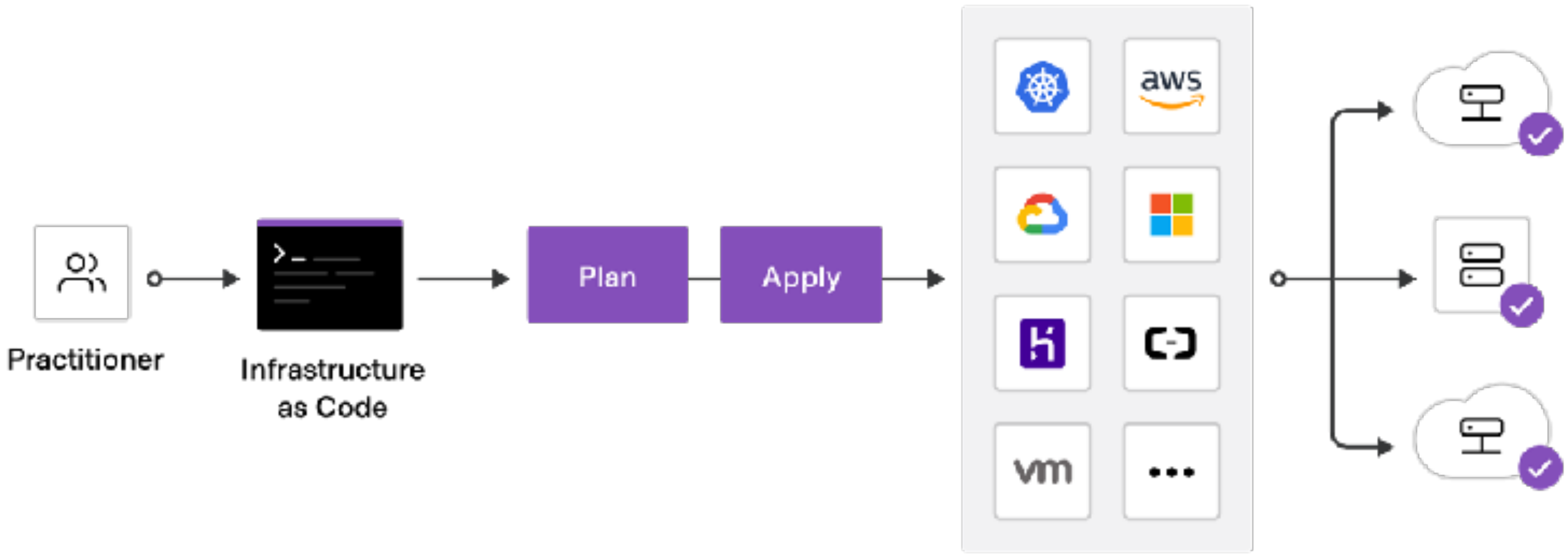
Infrastructure, development/ deployment is managed by the cloud provider

Billing based on service usage, usually by subscription

Infrastructure as Code

IaC is the process of managing and provisioning cloud infrastructure through machine-readable definition files **instead of interactive, physical hardware configuration or configuration tools**

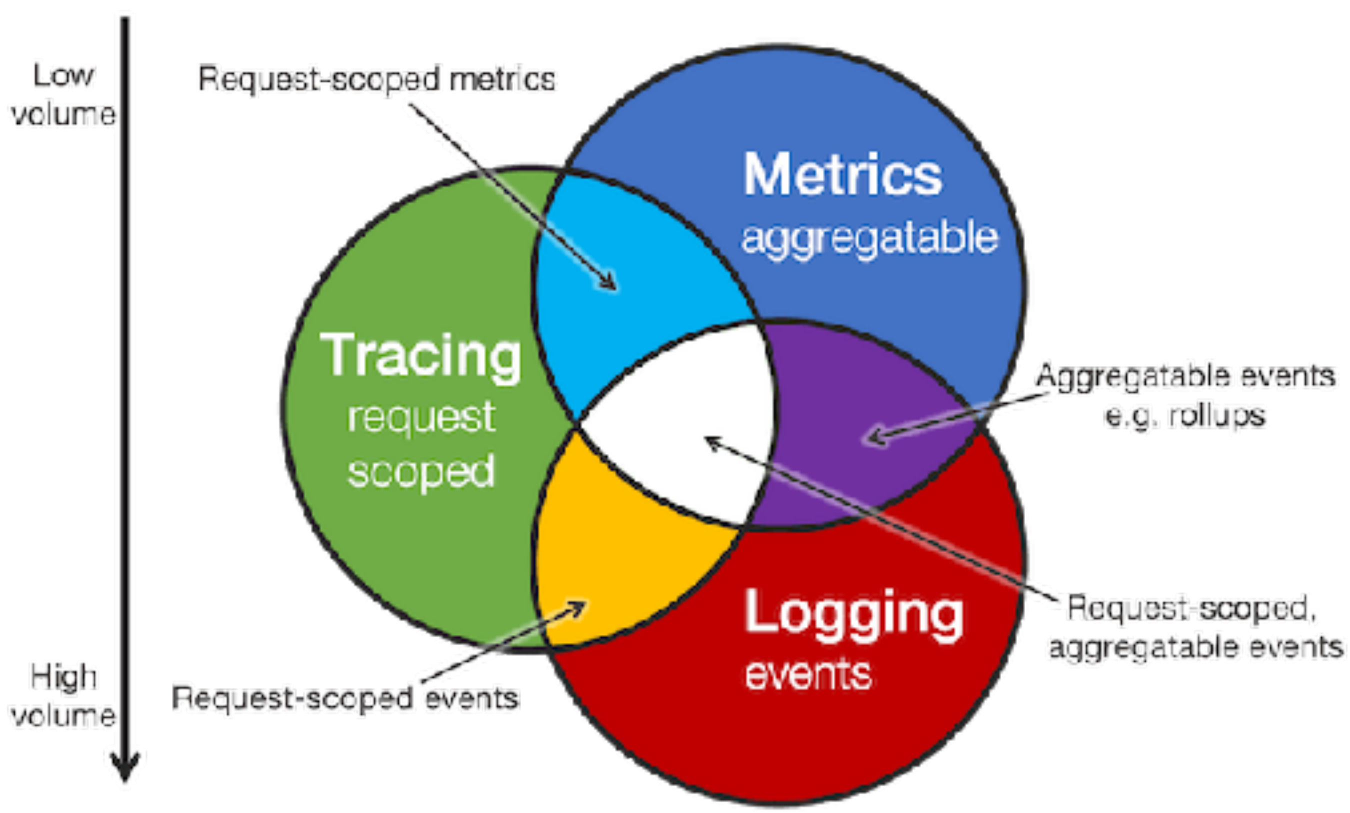
Minimizing the potential for human error, increases efficiency, and ensures consistency and compliance



https://developer.hashicorp.com/terraform/tutorials/aws-get-started/infrastructure-as-code

Observability

- Observability: the ability to understand the internal state of the system by examining its outputs
- Being able to ask it questions and getting clear answers
- Fundamental requirement for ensuring reliability and diagnosing problems



https://tisgoud.nl/2020/02/config-management-camp-2020-day-1/

Cloud for Scientific Computing

Relevant cloud services

- Compute services for heavy data processing:
  - Virtual machines
  - Container deployment
  - Serverless computing
- Storage services for vast amounts of data:
  - Durability, redundancy, scalability
  - Full cycle: usage, archival, backup, recovery
  - Data lakes
- Big Data/Machine learning services:
  - (Pre-)processing of large datasets
  - Collaborative coding – notebooks
  - Data warehousing
  - Stream or batch processing
- Networking services:
  - Fast, secure, reliable communication between services and regions
  - Load balancing

Case studies

- CERN - Cloud Computing for Large Hadron Collider
  - https://cloud.google.com/customers/cern
  - Cloud Storage: handling of petabytes of data, workloads of dozens of terabytes at 200GB per second
  - Google Kubernetes Engine: automated scaling of computing power using on-demand repeatable virtual machine instances
  - Leveraging the public cloud to make data available as open source for researchers, scientists, and educators
- NASA - Mars Rover data processed using AWS
  - https://aws.amazon.com/blogs/publicsector/perseverance-lands-mars-cloud-ready-explore/

Case studies

Processing of sub-studies for COVID-19 data

Private cloud with virtual secure network infrastructure

Versioned containerized SQL databases

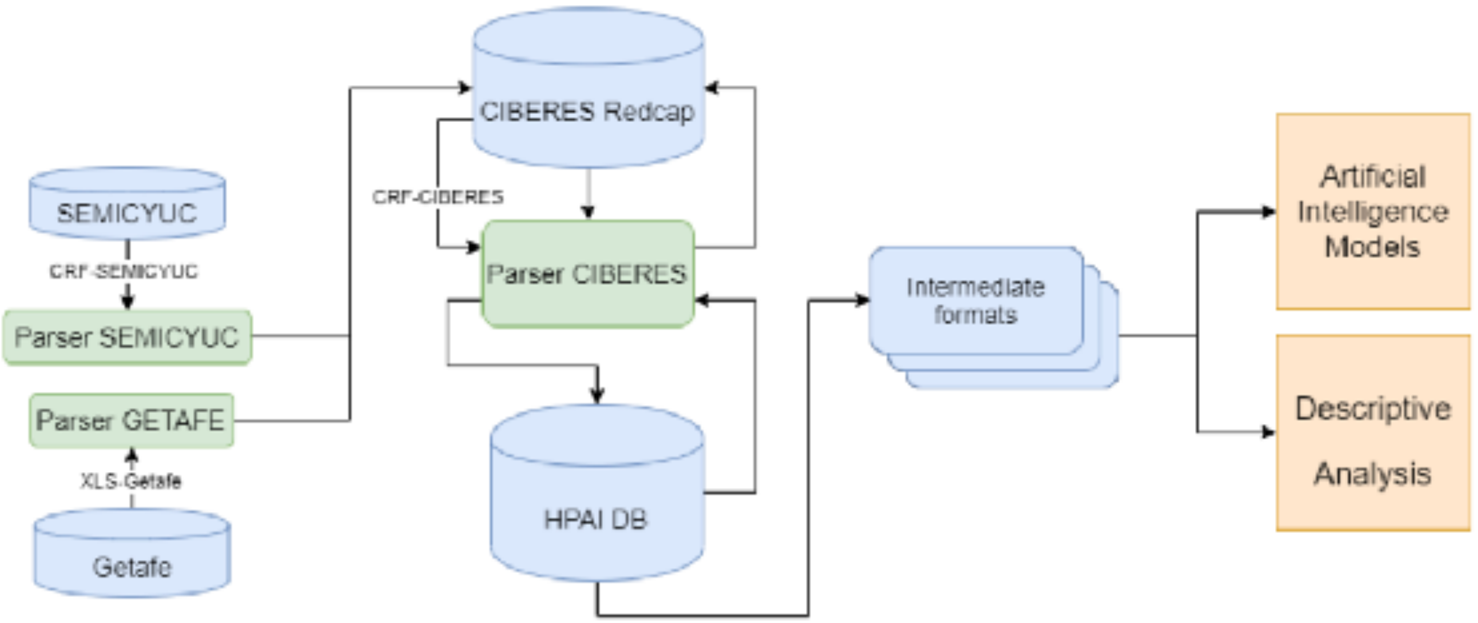
Versioned containerized pre-processing pipelines to keep source of truth intact

Versioned containerized AI models running in virtual computing infrastructure

All containers running on virtual distributed redundant storage

Experiments automatically triggered by version control

Outputs of experiments (PDFs with statistical analysis) uploaded back to the cloud



Marta Barroso et al, *"Collection, Processing and Analysis of Heterogeneous Data Coming from Spanish Hospitals in the Context of COVID-19"*, Artificial Intelligence Research and Development Issue 254, 2021

The combination of containerized databases, pre-processing pipelines and AI models **guaranteed full reproducibility** for each sub-study

Best practices

Reproducibility of Experiments

- **Use Infrastructure as Code (IaC):** Replicate environments easily
- **Version control:** Track changes to data, code, and configurations
- **Automate workflows:** Ensure consistency and reduce human error
- **Document extensively:** Make it easier for others to understand and reproduce your work

Best practices

Managing and Sharing Large Datasets

- **Use cloud storage services:** Benefit from scalability and global accessibility
- **Optimize data transfer:** Compress data, use dedicated transfer tools or services
- **Metadata management:** Maintain rich metadata for easier data discovery and use
- **Establish clear data policies:** Define who can access and modify the data, and under what circumstances
- **Backup and recovery:** Establish clear policies and implement automated processes

Best practices

Data Security and Privacy

- **Use robust encryption:** Protect data both at rest and in transit
- **Implement strong access control:** Restrict who can access the data
- **Regularly audit and monitor:** Detect any unauthorized access or anomalies
- **Ensure regulatory compliance:** Follow industry standards and regulations for data security

Best practices

Cost Optimization

- **Right sizing:** Choose the right services for your needs
- **Elasticity:** Scale resources up or down based on demand
- **Spot Instances:** Leverage unused cloud capacity at a significant discount
- **Monitoring and Reporting:** Keep track of your usage and spending"

Other common challenges

- Technical knowledge and training
- Integrating with existing IT Infrastructure
- Vendor lock-in

Becoming Cloud-Native

Cloud-Native

Benefits

- Approach to building and running applications that fully exploit the advantages of the cloud computing delivery model
- Main objective: to have loosely coupled systems
    - Resilient, manageable, observable, scalable
    - Frequent changes deployed through automation and minimal boilerplate
- Keywords: containers, virtualization, service mesh, immutable infrastructure, continuous delivery
- CNCF: Cloud Native Computing Foundation
    - https://github.com/cncf/toc/blob/main/DEFINITION.md
    - Aims to propose open-source, vendor-neutral pieces of the stack

Cloud-Native

How to transition

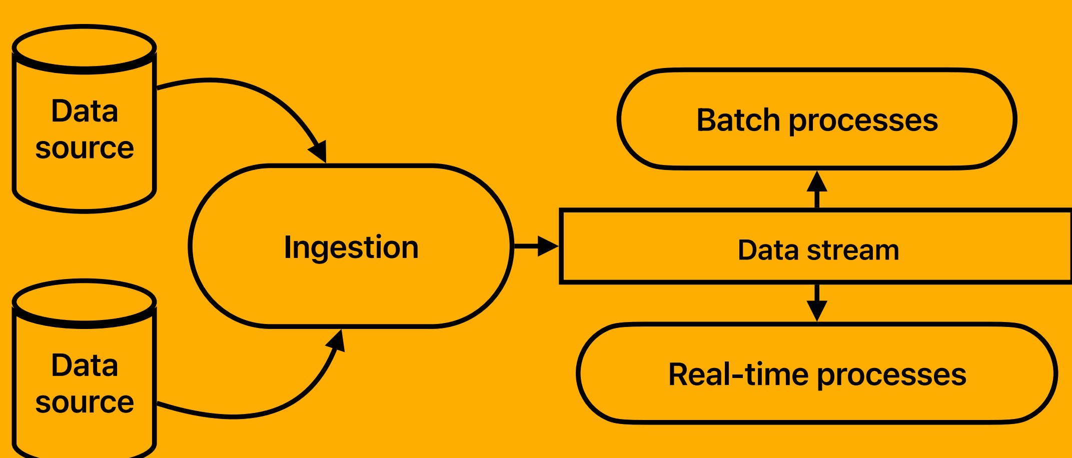
- Usually a strategic plan is required
  - Incremental migration of existing code
  - Deconstructing monoliths into small components
- Embrace DevOps practices
  - Automate **everything**
  - Continuous Integration/Deployment/Delivery
  - Bonus: traceability, reproducibility, auditability
- All of this requires specialized staff or re-training of existing staff

Cloud-enabled architectures

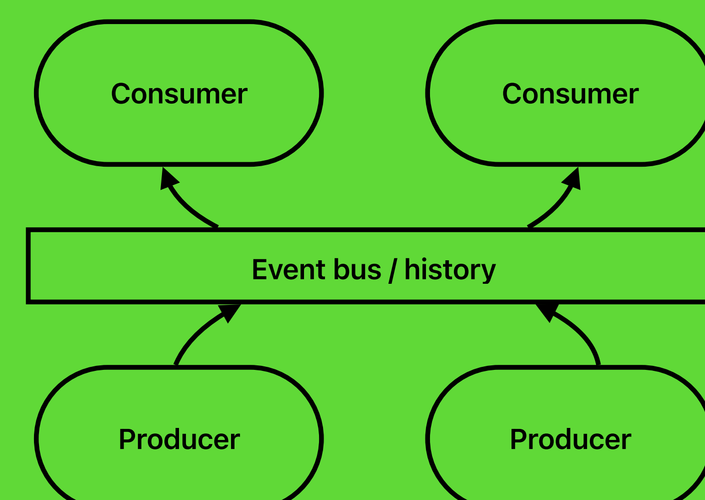
Microservices
Independent, loosely-coupled micro-components

Serverless
Zero-management of servers, total focus on functional logic

Lambda Architecture
Combine batch and (real-time) stream data processing into the same pipeline



Event Driven
Define system as a reaction to external events or state changes



Emerging roles

Cloud Architect
Plans, designs, and assists with all aspects of a company's cloud computing strategy

Cloud Engineer
Uses programming languages to design, develop, and implement cloud-based solutions

Cloud Consultant
Provides insights and practical guidance to businesses about how to best use and manage cloud-based resources

DevOps Engineer
Works with development and IT teams to oversee code releases and deployments in a cloud environment

Tools & Providers

Popular providers

Popular tools

Virtualization, containerization, configuration

HashiCorp Vagrant

HashiCorp Terraform

kubernetes

docker

Apache MESOS™

ANSIBLE

puppet

CHEF™

Popular tools

Continuous delivery, architecture and observability

GitLab CI/CD

APACHE kafka®

Spark

RabbitMQ

hadoop

kibana

Grafana

Prometheus

Conclusion

Recap

- Definition of what cloud computing is and its fundamentals, and its main benefits
- Some challenges that arise with its adoption
- We have described some key concepts, including what it means to be cloud-native
- Quick overview of relevant tools and providers

Looking ahead

- More growing adoption expected
- Edge Computing
- More complex, more capable Artificial Intelligence and Machine Learning workloads
- Widespread adoption of serverless computing?

Questions?
Topics for debate?

Image attribution

- Images have been attributed to their corresponding authors by the links included in their captions
- Logos are copyrighted to their corresponding owners and are used here for educational and information purposed
- All icons used are created by Freepick and Ahmad Roaayala (Flaticon): https://www.flaticon.com/free-icons