CM2606 Data Engineering

Cloud Data Platforms

Week 09 | Piumi Nanayakkara

Learning Outcomes

- Covers L03 and LO4 for the Module

- On completion of this lecture, students are expected to be able to:
  - Analyze and select the most appropriate product / service in cloud that can be used to implement a designed data pipeline

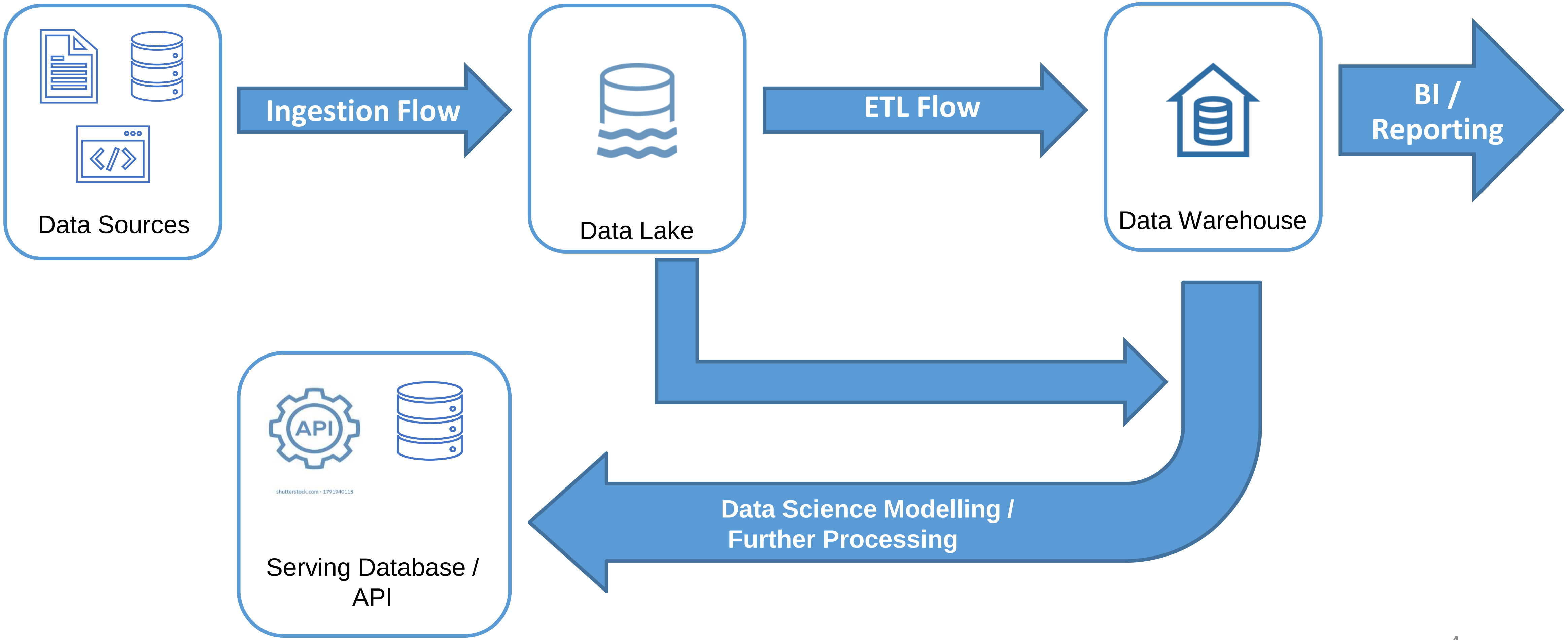
Content

- Cloud Computing
  - Concepts
  - Characteristics
  - Advantages

- Cloud Data Platforms
  - Benefits
  - On-Prem to Cloud

- Comparison of Cloud Services
  - Market Analysis
  - Tools and Services

Data Pipeline: Common Usage



Data Sources

Ingestion Flow

Data Lake

ETL Flow

Data Warehouse

BI / Reporting

Data Science Modelling / Further Processing

Serving Database / API

What is Cloud Computing

- In simple terms: Use of remote servers on the internet for your tasks.
- When you use someone else's machines that you do not own by you it is **cloud computing**. An organization that provides such resources is a **cloud service provider**.
- Little or no investment: You will only be paying only for the time you are using those resources, i.e., pay as you go policies.
- Core focus and less work force
  - Servers are managed by service providers including security, devops, auto scaling etc. enabling business to focus on core functionality.

Cloud Vs On- Prem

Feature	On-Premise	Cloud
Computing Environment	• Needs physical servers, network infrastructure, and storage. • The equipment must have power and cooling. • A server needs at least one operating system (OS) installed	• Provide the physical and logical infrastructure to host services • Within minutes, an organization can provision anything from virtual servers to clusters of containerized apps
Maintenance	• Require periodic maintenance for the hardware, drivers, BIOS, operating system, software, and antivirus software by qualified personnel.	• CSP manages key infrastructure services such as physical hardware, computer networking, firewalls and network security, datacenter fault tolerance, compliance, and physical security of the buildings.
Availability	• The more uptime the SLA requires, higher the cost.	• Duplicates customer content for redundancy and high availability.

Cloud Vs On- Prem

Feature	On-Premise	Cloud
Scalability	• To scale an on-premises server horizontally, server administrators add another server node to a cluster.	• Can be as simple as a mouse click.
Support	• Server administrators might need to know how to use many different platforms	• Easy to support because the environments are standardized.
Total cost of ownership	• Hardware + Software licensing + Labor (installation, upgrades, maintenance) + Data Center overhead (power, telecommunications, building, heating and cooling)	• A subscription based on usage that's measured in compute units, hours, or transactions. • Because of economies of scale, an on-premises system can rarely compete with the cloud

Advantages of Cloud Computing

- **High availability:** Depending on the service-level agreement (SLA) that you choose, your cloud-based apps can provide a continuous user experience with no apparent downtime, even when things go wrong.

- **Scalability:** Apps in the cloud can scale vertically and horizontally:

- **Elasticity:** You can configure cloud-based apps to take advantage of autoscaling, so your apps always have the resources they need.

- **Fault Tolerance:** Ability of the system to remain up and running in case of component or service failures.

- **Geo-distribution:** You can deploy apps and data to regional data centers around the globe, thereby ensuring that your customers always have the best performance in their region.

Advantages of Cloud Computing

- Operational Expenditure Over Capital Expenditure
  - Cloud services are categorized as an OpEx, because of their consumption model.
  - There's no asset for the business to amortize, and its cloud service provider manages the costs that are associated with the purchase and lifespan of the physical equipment.

  - On an On-premise setting cost will be capitalized since server systems are very expensive.
  - This means that on financial statements, costs are spread out across the expected lifetime of the server equipment.
  - This restricts an IT manager's ability to buy upgraded server equipment as for demand during the expected lifetime of a server.

- Benefits of a consumption-based model
  - No upfront costs.
  - No need to purchase and manage costly infrastructure that users might not use to its fullest.
  - The ability to pay for additional resources when they are needed.
  - The ability to stop paying for resources that are no longer needed.

Deployment Models

- Public Cloud
  - Single machine is shared among multiple users
  - Cloud providers buy machines with huge configs (e.g., 300 GB RAM) and create multiple VMs in such a machine based on user demand
- Private Cloud
  - Machines are not shared among users, IT services are provisioned over private IT infrastructure
  - IT services are provisioned over private IT infrastructure
- Hybrid Cloud
  - Combination of public and private cloud
  - Storage can be hosted in private cloud and front-end applications in public cloud where they need to be exposed to the internet.
  - Down the line, the organization will be paying less.

What is CDP?

- A cloud data platform(CDP) is the implementation / migration of an organization’s **data ecosystem** and enterprise data in / to the cloud and away from traditional on-premises data centers or warehouses.

- For some organizations, a cloud data platform can take the form of multi-cloud environments.

Benefits of a CDP

- Quick processing time: Cloud data platforms can quickly ingest and process structured and unstructured data. This allows for quicker availability of the data and analyses.
- Scalable: Rather than committing to a large amount of storage space, cloud platforms allow for businesses to scale their usage as necessary. If a large amount of data is quickly accumulated, organizations can simply request more space.
- Improved Access: Facilitate creation of a data lake to democratize data and share it anywhere and anytime, among both onsite and remote users
- Improved Security: Manual controls for access/encryption vs features provided by cloud service offering

OnPrem to CDP: Approaches

- Lift and Shift
  - Replicate the existing on-premises design
  - (+) Proven on-premises solution
  - (-) Carrying all technical debt of legacy systems to cloud/future
  - (-) Not utilizing the full potential offered by cloud services

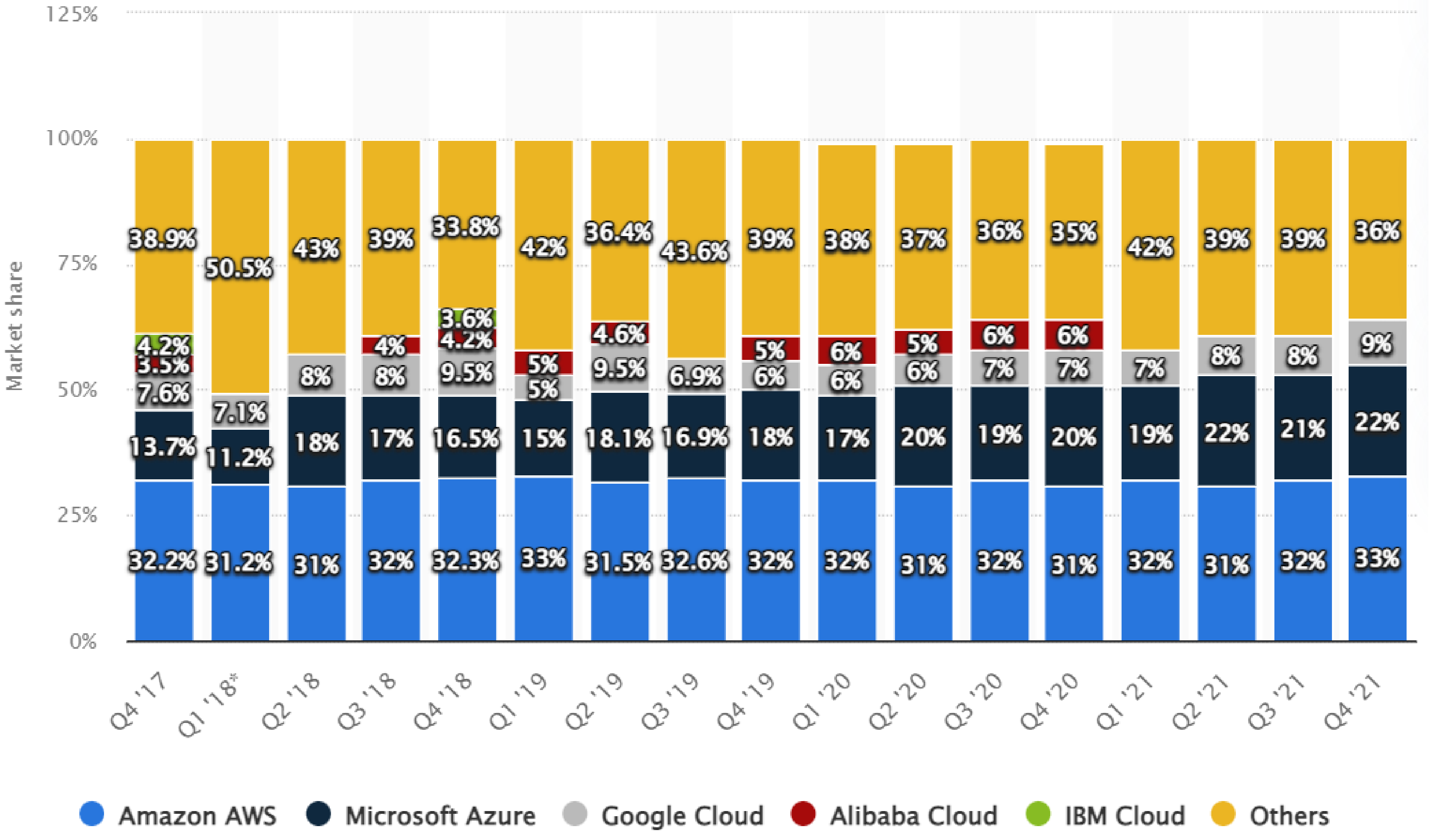
- Green Field
  - Start from scratch, leveraging the full potential of cloud architecture
  - (+) Opportunity to make use of latest features and have an updated design
  - (-) Unforeseen challenges and surprises with an unfamiliar technology

- Green-Shift
  - Combine the two into a hybrid mix
  - Lift and Shift the main model and re design the connecting models

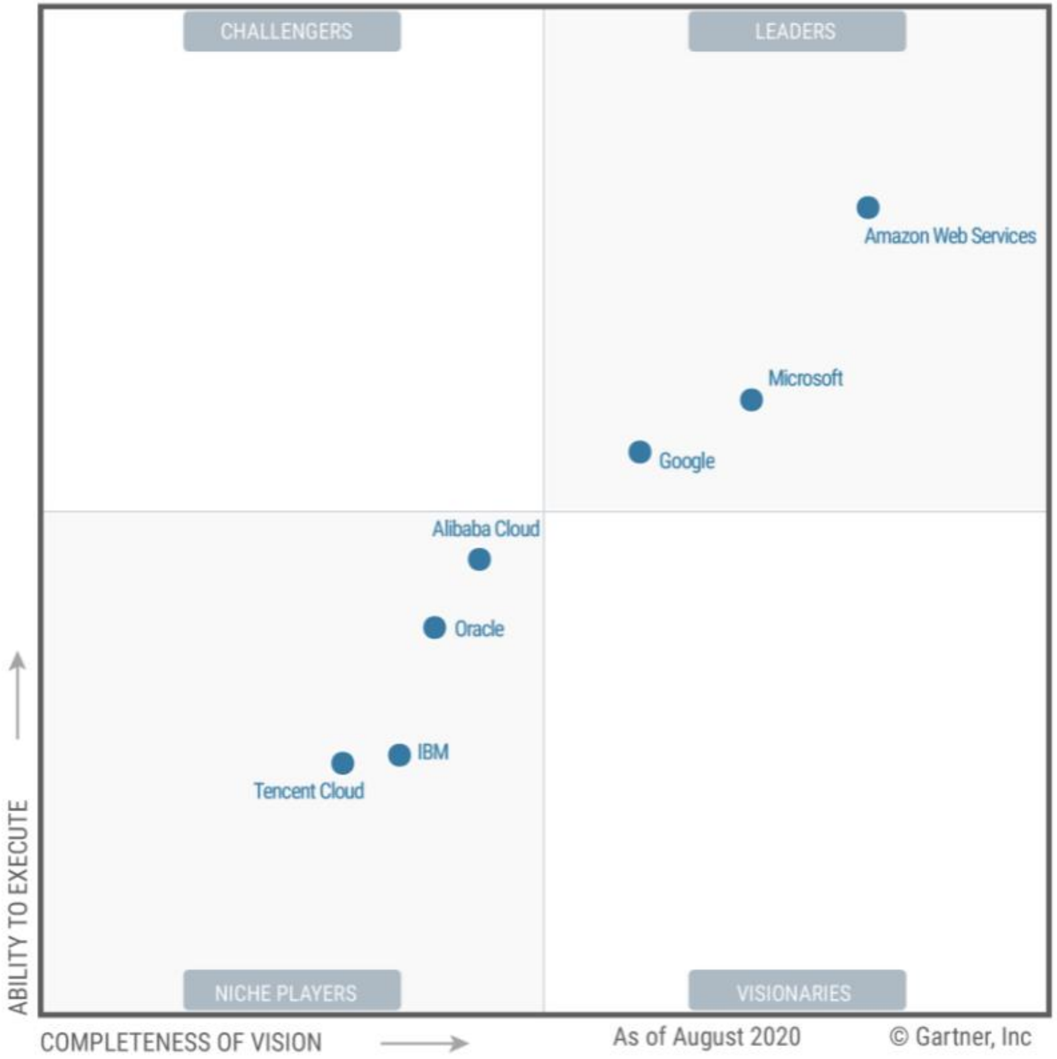
Comparison of Cloud Services

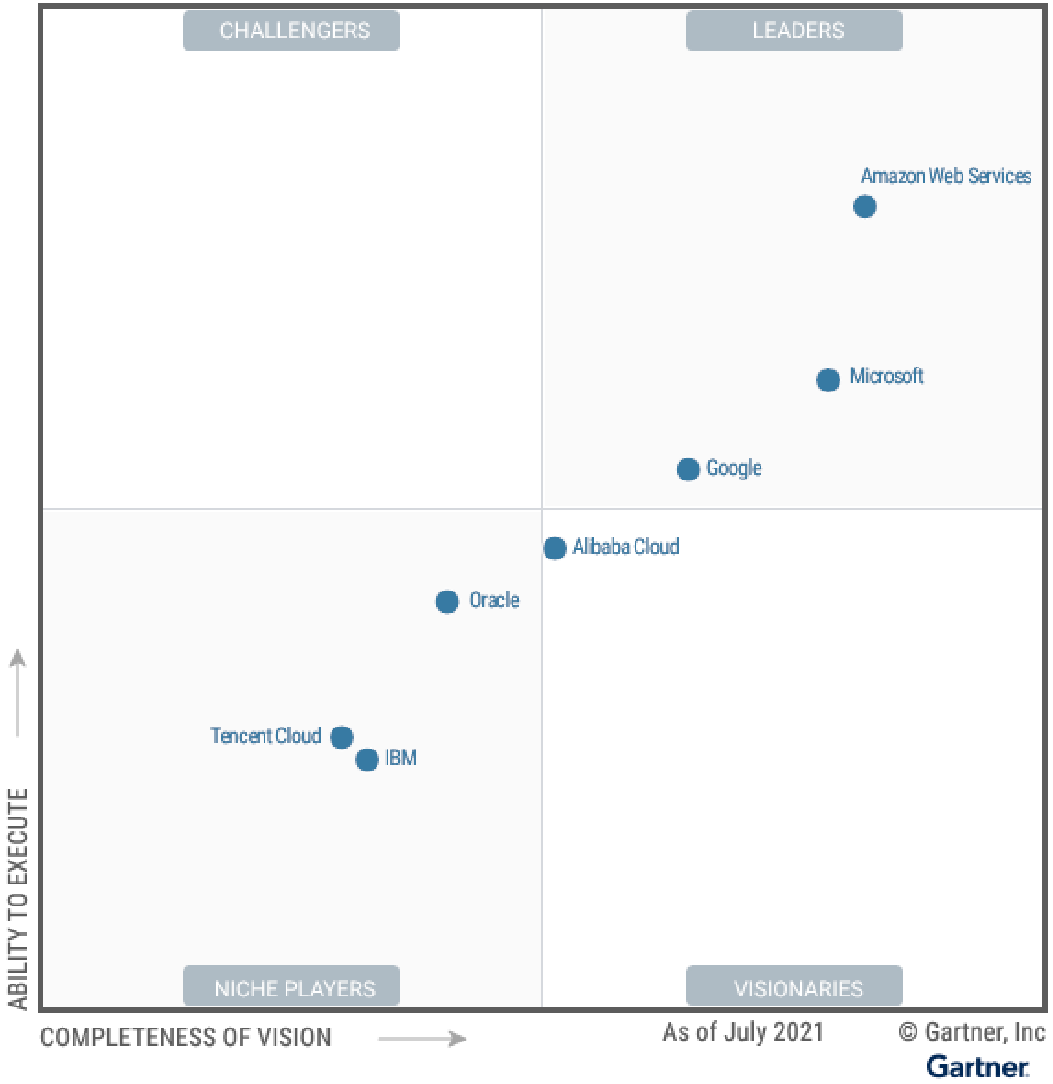
AWS vs. Azure vs. GCP

Market Share



Market Segmentation





Overview

AWS	Azure	GCP
• 26 geographic regions • 84 availability zones	• 60+ regions • Minimum of three availability zones in each region	• 29 cloud regions • 88 zones

Availability regions are the geographic locations of the cloud data centers

The **availability zone** refers to an isolated data center within a single region. Each availability zone includes multiple data centers,

Data Migration

AWS	Azure	GCP
• **AWS Database Migration Service:** support both homogenous and heterogenous databases • **AWS DataSync:** automates and accelerates moving data between on premise file systems and S3 • **AWS Snowball** - a physical hardware device that organizations can use to transfer petabytes of data in situations where internet transfer isn't practical. • **AWS Direct Connect:** establishes a dedicated network connection between on-premises internal network and AWS	**Azure Database Migration Service:** Migrate your database and server objects—including user accounts, agent jobs, and SQL Server Integration Services (SSIS) packages **Azure Migrate:** Discover, assess, right-size on-prem data **Azure Data Box:** Devices easily move data to Azure when busy networks aren't an option.	• **Database Migration Service:** Migrate databases to Cloud SQL(MySQL or PostgreSQL,) from on-premises, Compute Engine, and other clouds. • **Storage Transfer Service:** Complete large-scale online data transfers from online and on-premises sources to Cloud Storage. • **Transfer Appliance:** Securely migrate large volumes of data (from hundreds of terabytes up to one petabyte) to Google Cloud

Data Pipeline Design & Orchestration

AWS	Azure	GCP
• **AWS Data Pipeline:** web service that helps you reliably process and move data between different AWS compute and storage services, as well as on-premises data sources, at specified intervals. • **AWS Glue:** serverless data integration service that makes it easy to discover, prepare, and combine data for analytics, machine learning, and application development.	• **Azure Data Factory:** create data-driven workflows for orchestrating and automating data movement and data transformation.	• **Google Cloud Dataflow:** Unified stream and batch data processing that's serverless, fast, and cost-effective. • **Google Cloud Composer:** Contains predefined operators for standard tasks • **Cloud Data Fusion:** Visual point-and-click interface enabling code-free deployment of ETL/ELT data pipelines

Data Ingestion

AWS	Azure	GCP
• **Kinesis Streams:** for real-time data streaming • **Kinesis Firehose:** for large-scale data ingestion • **Amazon Managed Streaming for Apache Kafka**: ingest and process streaming data in real time with fully managed Apache Kafka. • **Amazon Simple Notification Service SNS:** trigger the processing pipelines when new content is updated • **Amazon Simple Queueing Service (SQS):** fully managed message queuing service	• **Azure Event Grid:** A pipeline that listens to Azure storage, and pull information when subscribed events occur • **Event hub:** A pipeline that transfers events from services to Azure Data Explorer. • **IoT Hub**: A pipeline that is used for the transfer of data from supported IoT devices to Azure Data Explorer	• **Pub/Sub:** Messaging and ingestion for event-driven systems and streaming analytics. Ingest events for streaming into BigQuery, data lakes or operational databases • **Streaming Insert:** stream and process data in near-real time, can be performed on a BigQuery table using the Cloud SDK or Google Dataflow

Data Lake & Warehousing

	AWS	Azure	GCP
Object storage	AWS Simple Storage Service (S3)	Azure Data Lake	Cloud Storage
Archival storage	Amazon S3 Glacier	Azure Archive Storage	Cloud Storage Archive:
Data Warehousing	Amazon Redshift	Azure Synapse Analytics	BigQuery

Databases

Type	AWS	Azure	GCP
SQL Databases	• AWS RDS • Amazon Aurora (mysql and postgreSQL compatible)	• Azure SQL • Database for MySQL • Database for PostgreSQL	• Cloud SQL • Cloud Spanner
Document DB	• Amazon DocumentDB	• Azure Cosmos DB	• Firestore
Key Value Pairs	• Amazon DynamoDB	• Azure Cosmos DB • Table Storage	• Big Table
Graph	• Neptune	• Gremlin API in Azure Cosmos DB	• Neo4j AuraDB
other NOSQL Databases	• Simple DB		• Cloud Datastore

Big Data Processing

AWS	Azure	GCP
• **Elastic MapReduce (EMR):** Managed Hadoop, Spark and Presto solution. • **AWS Athena:** interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL	• **Azure Data Explorer:** Fully managed, high-performance, big data analytics platform that makes it easy to analyze high volumes of data in near real time. • **Azure HDInsight:** Provision cloud Hadoop, Spark, R Server, HBase, and Storm clusters • **Azure Data Lake Analytics:** On-demand analytics job service which easily develop and run massively parallel data transformation and processing programs in U-SQL, R, Python, and .NET	**Dataproc:** Deploy open-source data and analytics processing services (Apache Hadoop, Apache Spark, etc.) with improved efficiency and security.

Machine Learning

Use Case	AWS	Azure	GCP
ML Infra	Amazon EC2 P3	Azure Data Science Virtual Machines	Deep Learning VM Images .
ML Platform	Amazon SageMaker	• Azure AI Platform • AutoML in Azure ML Studio	Vertex AI
Natural language processing	Amazon Comprehend	Azure Text Analytics	Natural Language AI
Video intelligence	Amazon Rekognition Video	Azure Video Indexer	Video Intelligence API
Image Recognition	Amazon Rekognition Image	Azure Computer Vision	Vision AI

Machine Learning

Use Case	AWS	Azure	GCP
Speech Recognition	Amazon Transcribe	Azure Speech to Text	Speech-to-Text
Speech Synthesis	Amazon Polly	Azure Text to Speech	Text-to-Speech
Translation	Amazon Translate	Azure Translator	Translation AI
Document Understanding	Amazon Textract	Azure Form Recognizer	Document AI .

Dashboarding and Analysis tools

Use Case	AWS	Azure	GCP
BI and Dashboarding	Quick Sight	Power BI	• Looker • Google Analytics
Data Discovery & Wrangling	• AWS Glue Data Catalog • Amazon SageMaker Data Wrangler	• Azure Purview • Azure Data Explorer	• Data Catalog • Dataprep by Trifecta

Pros and Cons

AWS	Azure	GCP
+ Most services available, from networking to robotics + Most mature + Considered the gold standard in cloud reliability and security + More compute capacity vs Azure & GCP + All major software vendors make their programs available on AWS	+ Easy integration and migrations for existing Microsoft services + Many services available, including best-in-class AI, ML, and analytics services + Relatively cheaper for most services vs AWS & GCP + Great support for hybrid cloud strategies	+ Plays nicely with other Google service and products + Excellent support for containerized workloads + Global fiber network
- Complex pricing strategy - Can overwhelm newcomers with the sheer number of services and options - Comparatively limited options for hybrid cloud	- Fewer service offerings vs AWS - Particularly geared towards enterprise customers - maintenance required for the platform and the high expertise needed to use Azure	- Limited services vs AWS & Azure - Limited support for enterprise use cases - Historically not as enterprise focused

Summary

AWS	Azure	GCP
• Most high-performance and flexible complex cloud software solution • Focus on Public Cloud • Complex Pricing • Medium to large scale customers	• Primary choice for Windows based enterprise customers • Supports hybrid cloud implementation • Small to large scale customers	• Good support for Big Data and AI • Still rising • Small Scale customers

Free Tiers (for Coursework)

<table>
<tr><th>AWS</th><th>Azure</th><th>GCP</th></tr>
<tr><td>• Applicable for 12 months and restrictions are defined at product level.
• No Free credits</td><td>• <b>$200 credit</b> to spend in the <b>first 30 days</b> after you sign up.
• In addition, you get free monthly amounts of two groups of services: popular services, which are free for 12 months, and more than 40 other services that are free always.</td><td><b>$300</b> in free credits to spend on Google Cloud products during the <b>first 90 days</b>
Additional restrictions at product/service level: e.g., For pub/sub - All customers get up to 10 GB/month for ingestion or delivery of messages, free of charge.</td></tr>
<tr><td colspan="3">• Always monitor your usage in the console of the cloud service provider
• Set budgetary controls for cost as well as for usage of the product/service that you are going to use, so that you will get notified when you are reaching / exceeding the limit.
• Make sure to shut down all your resources when not in use.</td></tr>
</table>

READING

- Cloud Data Platforms for dummies (2nd edition) by David Baum (Snowflake)
- Designing Cloud Data Platforms by Danil Zburivsky, Lynda Partner Released May 2021