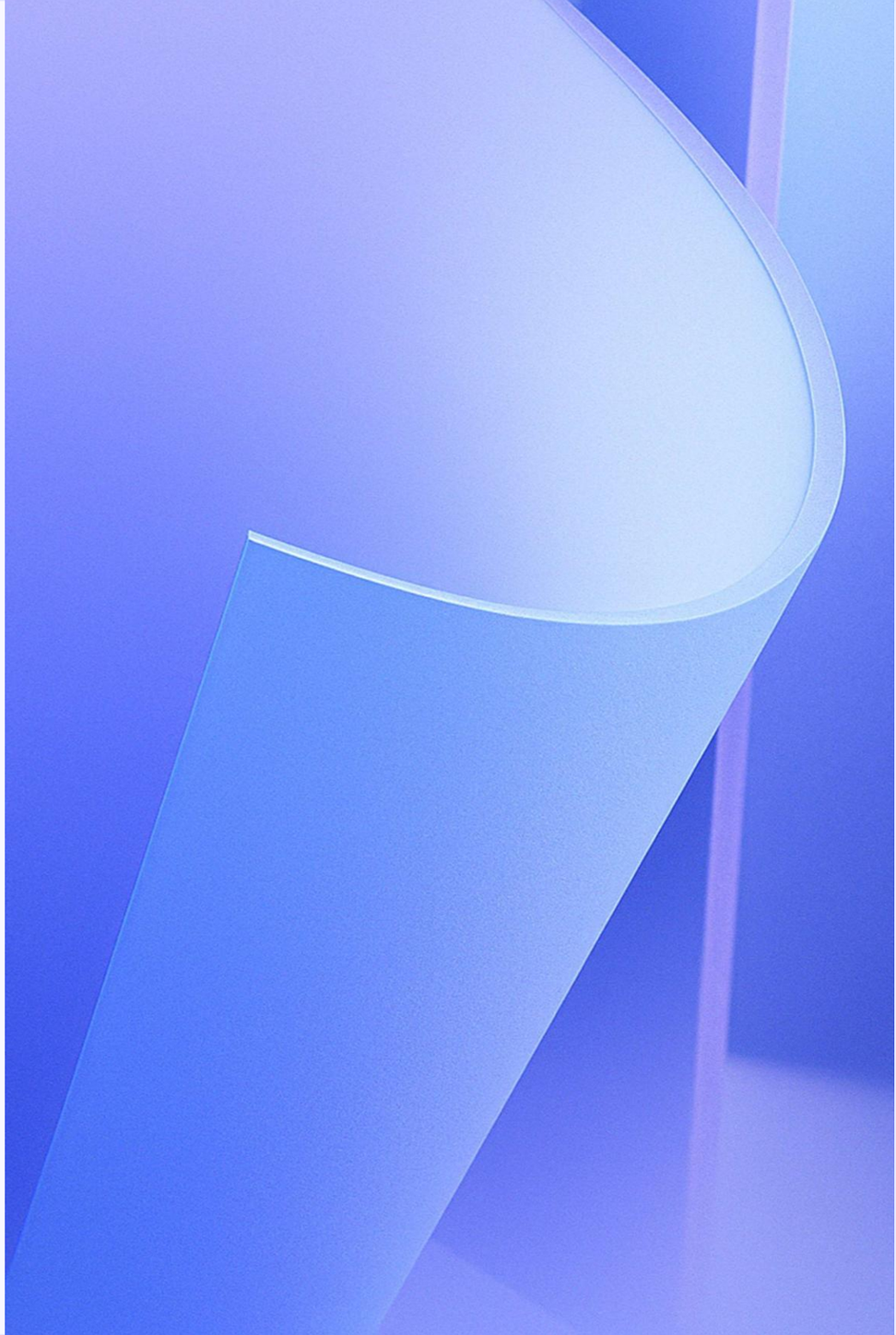
Why Data Engineering?

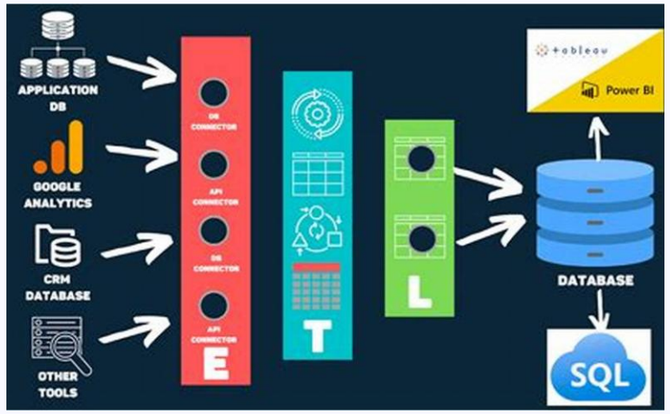
Presenter: Gad August



The Rise of Big Data

Expansion of Data Sources

Big data growth is driven by diverse sources like social media, IoT, e-commerce, sensors, and APIs generating vast data daily.



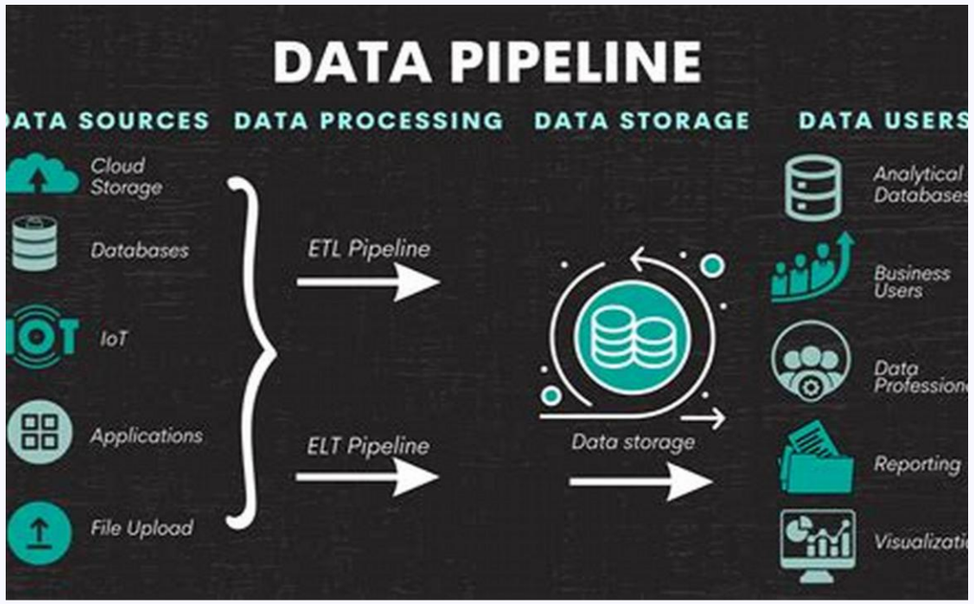
Variety and Volume of Data

Organizations handle structured, semi-structured, and unstructured data measured in quintillions of bytes, creating integration challenges.

Importance of Real-Time Processing

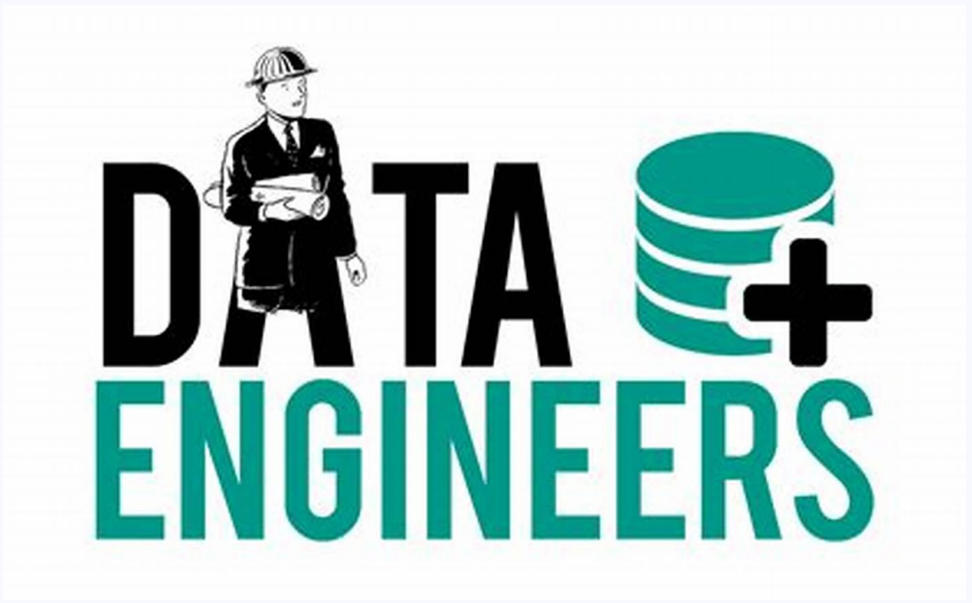
Real-time data processing reduces latency, enabling immediate analysis for applications like fraud detection and personalized experiences.

What is Data Engineering?



Definition of Data Engineering

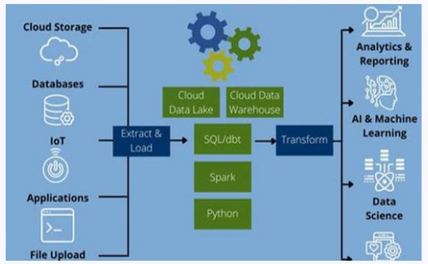
Data engineering involves designing and maintaining data pipelines and infrastructure to ensure data is accessible and reliable for analysis.



Key Responsibilities of Data Engineers

Develop scalable data architecture, streamline data acquisition from various sources, Ingestion, storage, transformation, integration, processing, orchestration, and quality governance.

The difference between a data engineer, data scientist, machine learning engineer and data analytics

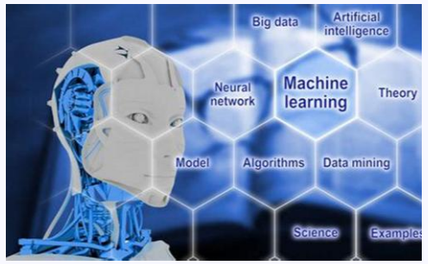


Role of Data Engineer

Data engineers design and maintain data pipelines and infrastructure to ensure reliable data flow for analysis and modeling.

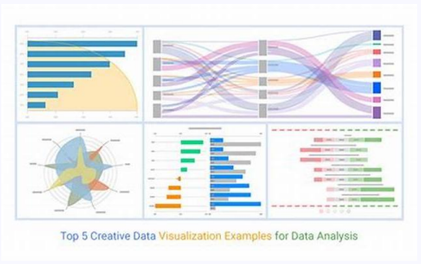
Role of Data Scientist

Data scientists analyze complex data, build predictive models, and generate insights to support business decisions.



Role of Machine Learning Engineer

Machine learning engineers deploy and maintain scalable ML models, bridging data science and software engineering.



Role of Data Analyst

Data analysts create reports and visualizations from structured data to help stakeholders make informed decisions.

Real-World Applications

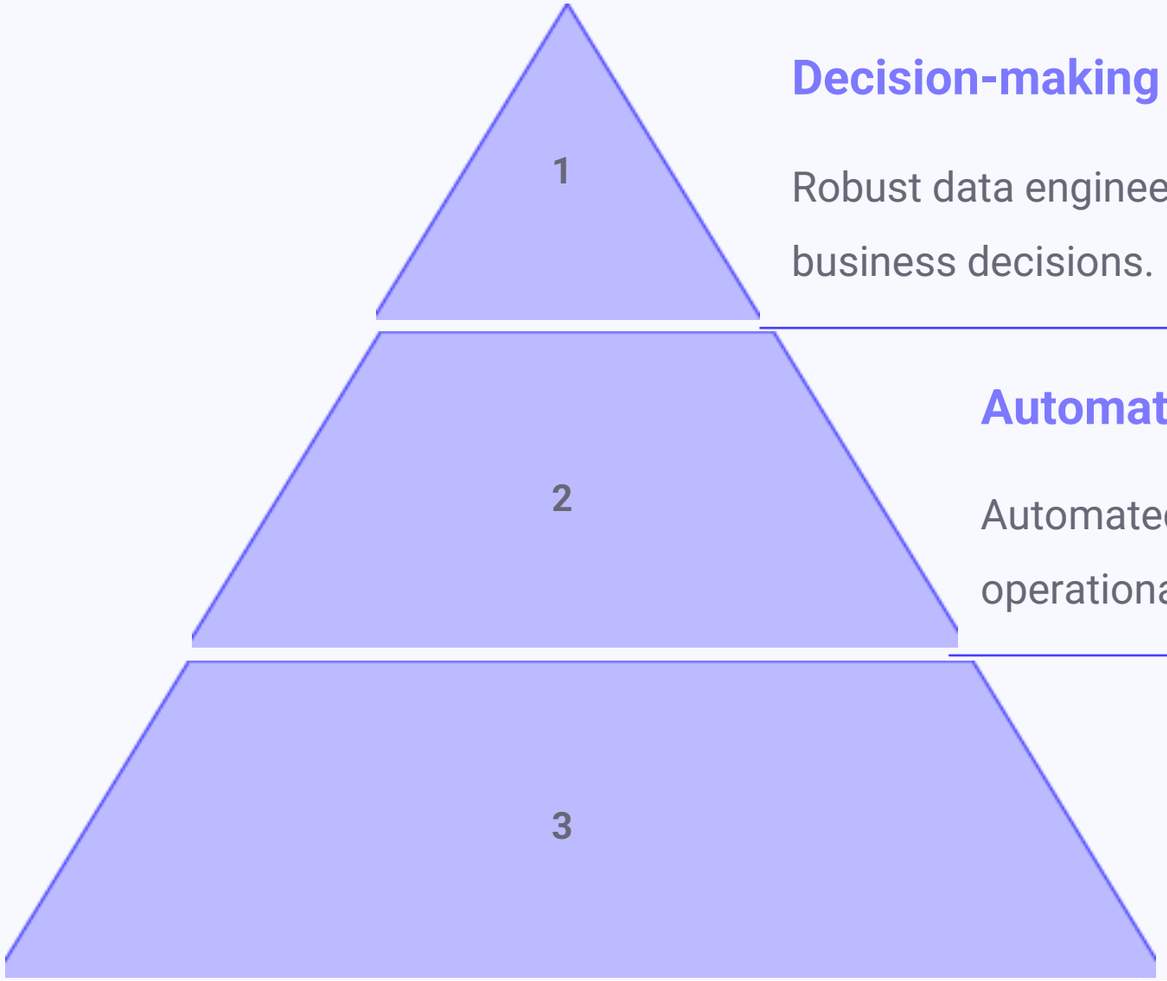
Role of Data Engineering in Industry

Data engineering supports major companies by enabling efficient handling of large data volumes and real-time processing for personalized customer experiences.

Examples from Leading Companies

Netflix, Amazon, and Uber use data engineering for real-time analytics, dynamic pricing, and optimized service delivery to enhance user engagement and profitability.

Importance in Modern Organizations



Decision-making Support

Robust data engineering ensures accessible, reliable data for rapid, informed business decisions.

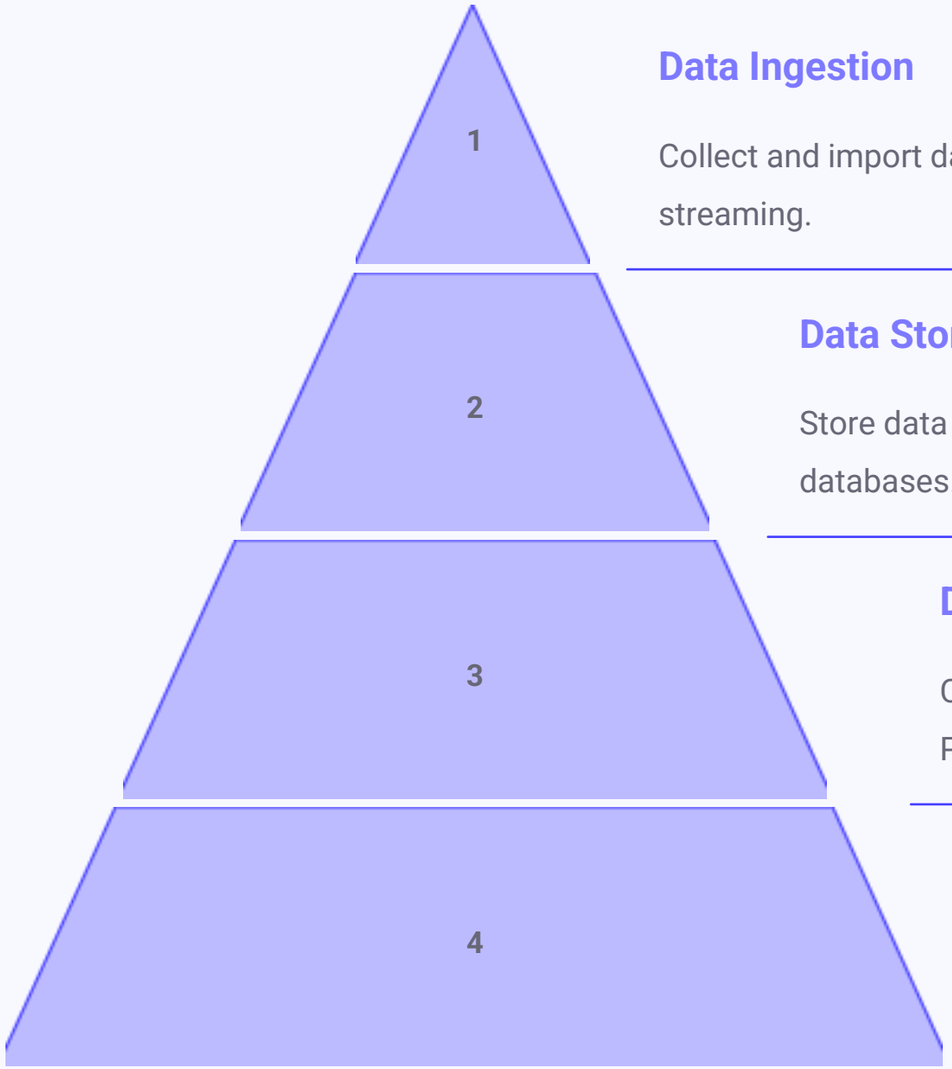
Automation Benefits

Automated data pipelines reduce errors, save time, and enhance operational efficiency.

Scalability Advantages

Scalable architectures handle growing data volumes using cloud and distributed systems.

Data Engineering Lifecycle



Data Ingestion

Collect and import data from diverse sources using tools like, ADF, Airflow, Apache Kafka for real-time streaming.

Data Storage

Store data in scalable repositories such as data lakes, lake houses warehouses, or NoSQL databases.

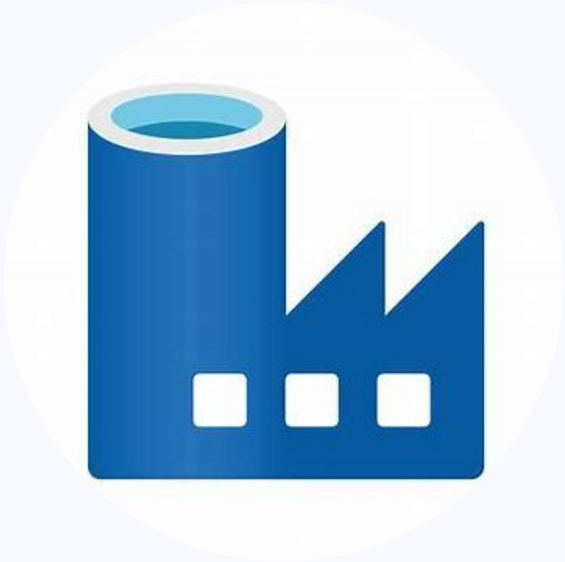
Data Processing and Transformation

Clean, enrich, and transform raw data using ETL/ELT pipelines with tools like Python.

Data Serving

Make processed data accessible to users via dashboards, data marts, and OLAP systems.

Tools of the Trade (Microsoft Fabric Data Engineer)

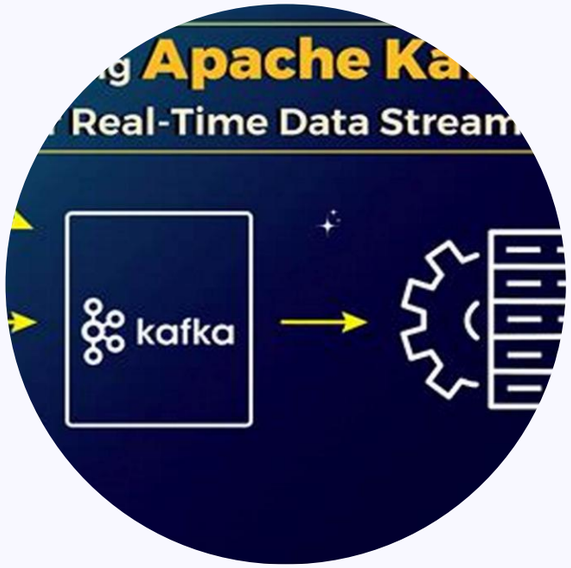


Core Workflow Orchestration

Azure data factory is widely used for authoring, scheduling, and monitoring complex data pipelines with flexibility and scalability.

High-Performance Data Processing

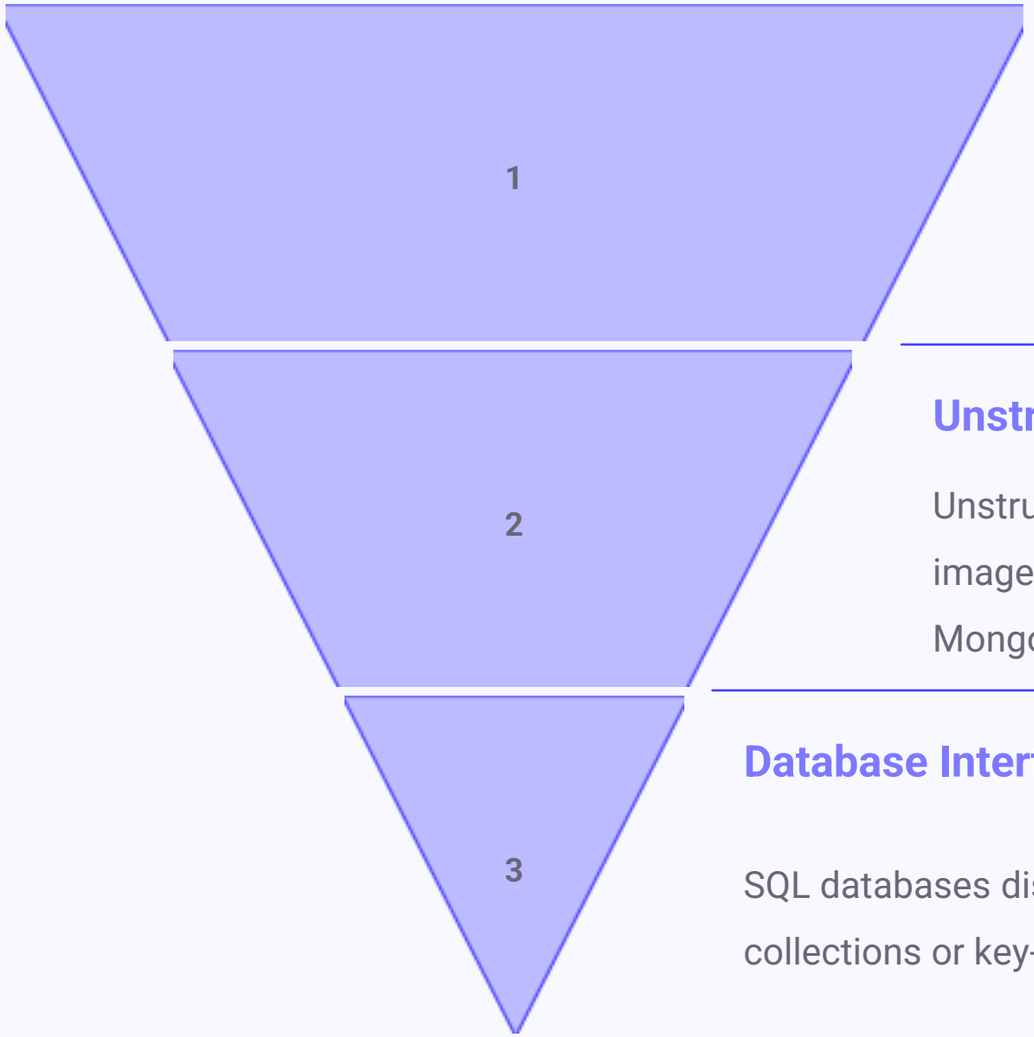
Apache Spark supports batch, streaming, machine learning, and graph processing with in-memory computing for fast data computation.



Real-Time Data Streaming

Apache Kafka enables high-throughput, fault-tolerant real-time data feeds, essential for streaming analytics and event-driven architectures.

Databases: SQL vs NoSQL



Structured Data Characteristics

Structured data is organized in predefined schemas using tables, rows, and columns, supporting complex queries with SQL databases like SQL Server and PostgreSQL.

Unstructured Data Characteristics

Unstructured data lacks fixed schemas and includes formats like text, images, and videos, managed by NoSQL databases such as Neo4J, MongoDB and Cassandra with flexible schemas.

Database Interface Examples

SQL databases display structured tables, while NoSQL dashboards show document collections or key-value pairs, illustrating their different data management approaches.

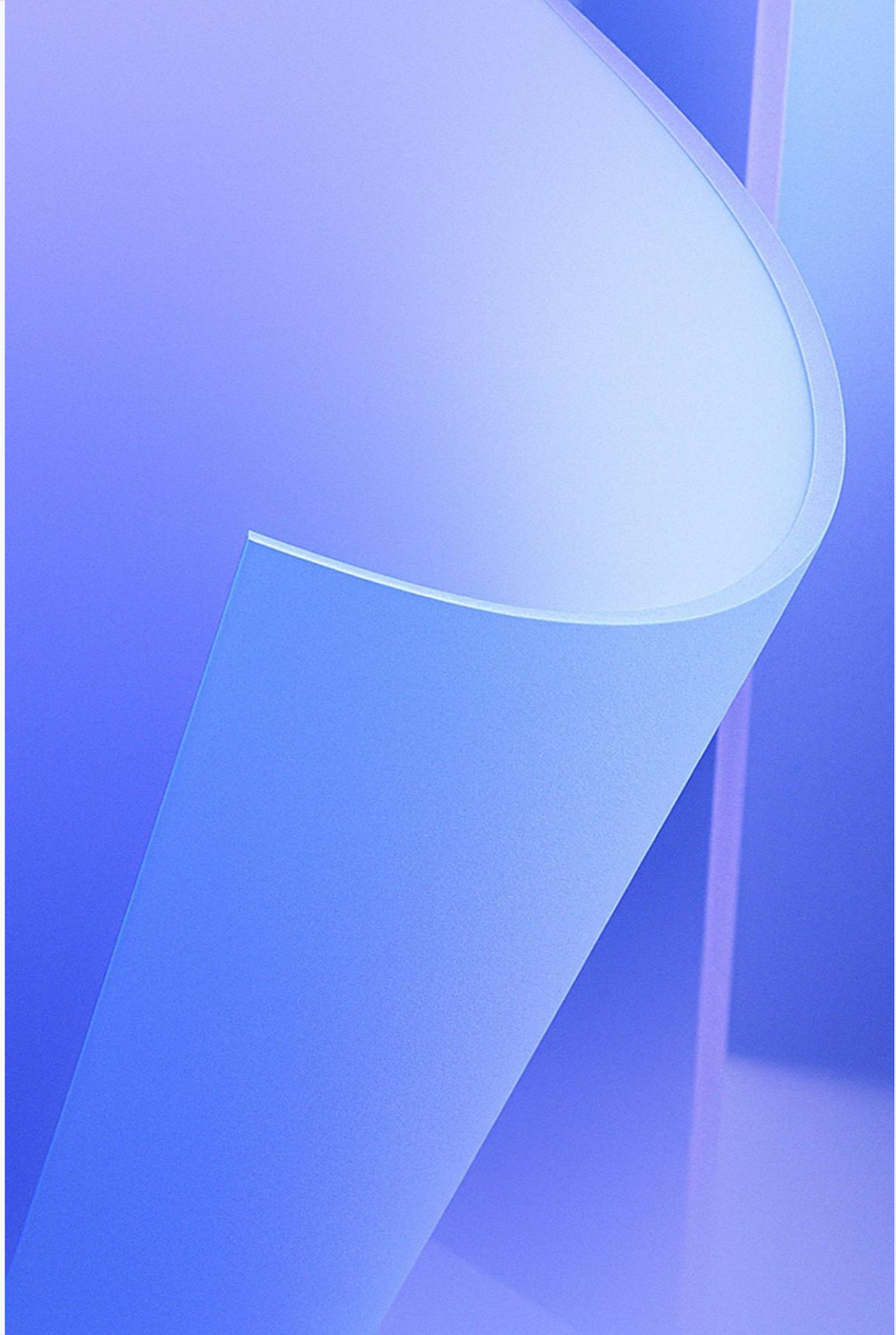
ETL vs ELT

Overview of ETL and ELT

ETL extracts, transforms, then loads data, while ELT extracts, loads, then transforms data within the target system.

Advantages and Challenges

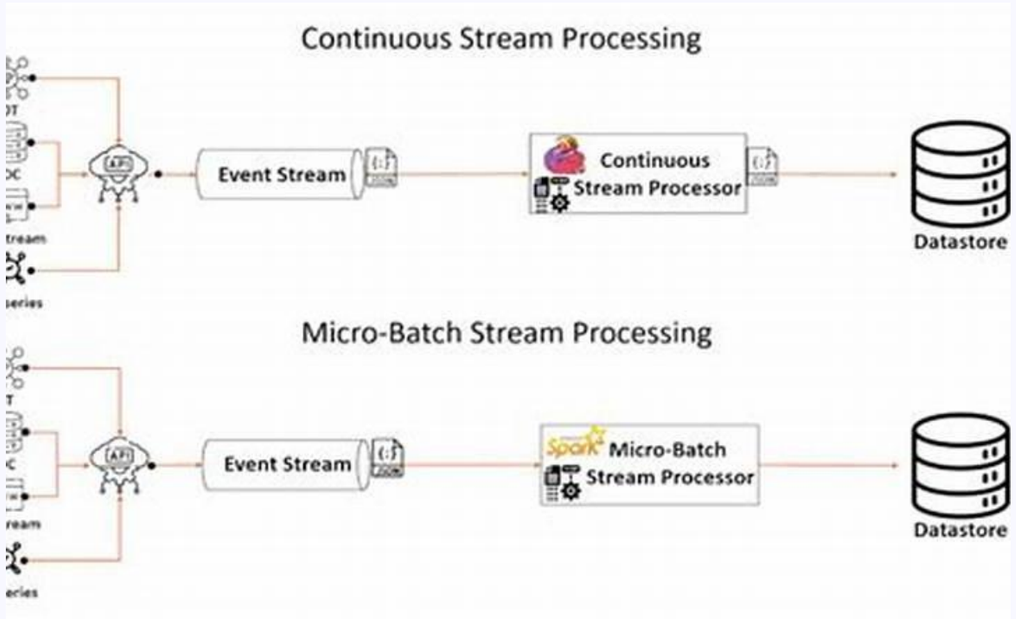
ETL ensures data quality before loading but may cause latency; ELT enables faster loading and flexible transformations but requires strong target system resources.



Batch vs Stream Processing

Batch Processing Overview

Batch processing handles large data volumes collected over time, processing them at scheduled intervals for complex computations and in-depth analytics.



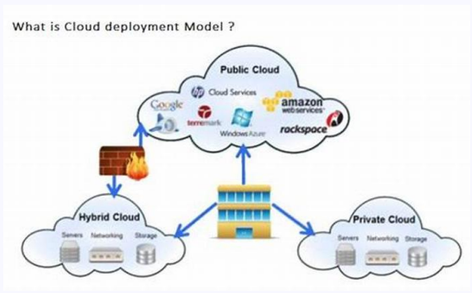
Stream Processing Overview

Stream processing continuously ingests and analyzes data in real-time, enabling instant insights and rapid decision-making for dynamic applications.

Cloud Computing?

Definition of Cloud Computing

Cloud computing delivers computing services over the internet, enabling faster innovation and flexible resource use without owning infrastructure.



Cloud Deployment Models

Public, private, and hybrid clouds offer different levels of resource sharing and control to meet diverse business needs.

Leading Cloud Providers

AWS, Microsoft Azure, and Google Cloud Platform dominate with unique strengths in services, integration, and analytics.

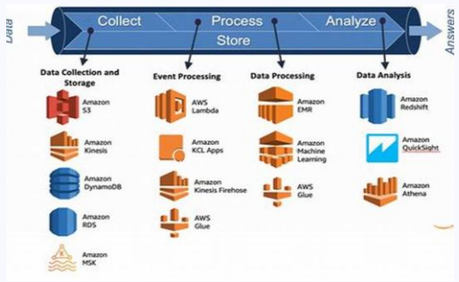
Cloud Data Engineering

Cloud platforms provide scalable tools for building data pipelines, warehousing, ETL, and real-time analytics.

Major Cloud Data Engineering Providers

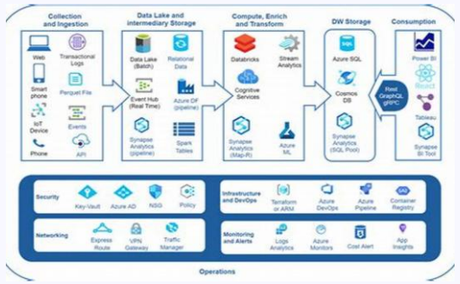
Leading Cloud Providers in Data Engineering

The top three cloud providers are AWS, Microsoft Azure, and Google Cloud Platform, each offering specialized data services.



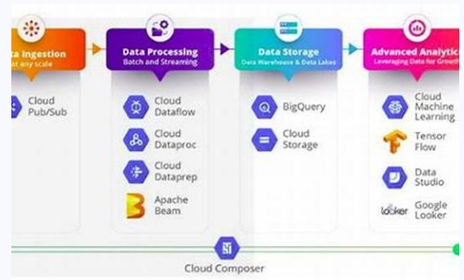
Amazon Web Services (AWS) Overview

AWS provides an ecosystem with services like EMR, Kinesis, Redshift, and Glue, excelling in scalability and real-time analytics.



Microsoft Azure Capabilities

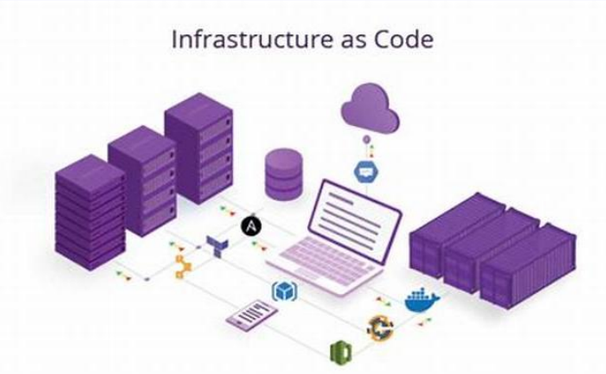
Azure focuses on enterprise and hybrid solutions, integrating well with Microsoft products and offering Microsoft fabric, and Power BI



Google Cloud Platform (GCP) Strengths

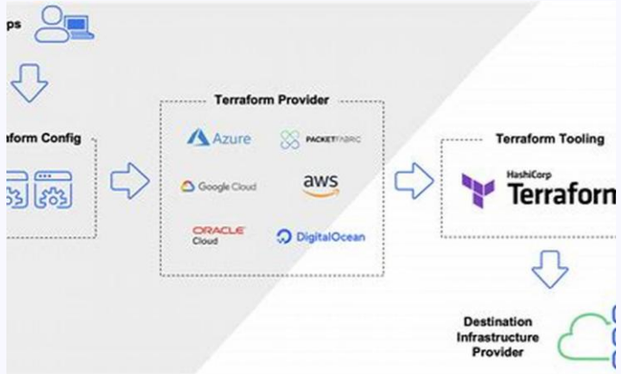
GCP is known for web analytics and machine learning tools such as Dataproc, Dataflow, BigQuery, and Cloud Composer, with a user-friendly interface.

Infrastructure as Code & Scalability in Microsoft fabric data engineering



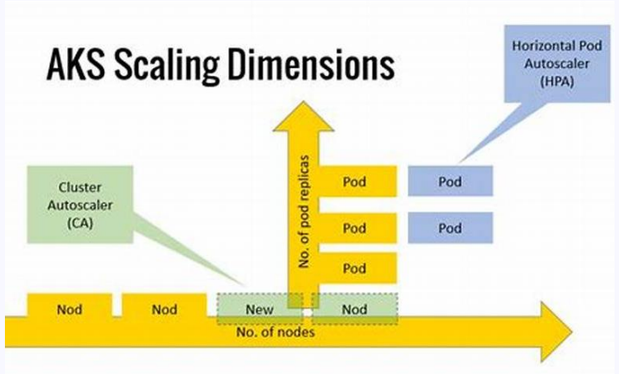
Role of Infrastructure as Code in Data Engineering

Infrastructure as Code (IaC) automates cloud infrastructure management, enhancing deployment efficiency and consistency in data engineering.



Tools and Techniques for IaC in Microsoft Fabric

Tools like Terraform and Azure Resource Manager enable automated deployment and management of data pipelines and resources in Microsoft Fabric.



Scalability through Automation and Autoscaling

Combining IaC with autoscaling features, such as Spark cluster scaling, ensures flexible resource adjustment to meet dynamic data workloads.

Security & Compliance Microsoft Fabric data engineering

Security Framework in Microsoft Fabric

Microsoft Fabric integrates robust security measures to protect data engineering workflows and assets.

Compliance Standards Adherence

The platform complies with industry standards and regulations to ensure data governance and privacy.

Data Protection Mechanisms

Advanced encryption, access controls, and monitoring tools safeguard data throughout its lifecycle.

Future Of Jobs Report 2025

2025 Job Market Trends

Data engineer demand grows with cloud, big data, and AI importance.

Key Skills Required

Programming in Python, SQL, Scala, Java; cloud skills in Azure, AWS.

Career Pathways

Roles include Data Engineer, Analytics Engineer, Platform Engineer.

1 2 3 4 5 6

Industry Demand

High demand in finance, healthcare, retail, and tech startups.

Additional Skills

ETL/ELT pipelines, SQL/NoSQL, real-time streaming, infrastructure as code.

Future Outlook

Strong growth driven by cloud computing and real-time analytics focus.

Skills in Demand

Core Programming Skills

Python programming is essential for scripting, automation, and integrating data processing frameworks.

1

2

Database Management

Advanced SQL proficiency enables efficient querying and management of structured data and data warehouses.

Workflow Orchestration

3

Expertise in Azure Data Factory supports automation and reliable scheduling of complex data pipelines.

4

Communication and Problem Solving

Strong communication and analytical problem-solving skills ensure alignment with business goals and effective issue resolution.

Career Pathways

Emergence of Data Engineering

Data engineering grows rapidly due to increasing data volumes and reliance on data-driven decisions.

Core Roles in Data Engineering

Includes Data Engineer, Analytics Engineer, Platform Engineer, and AI/Machine Learning Engineer.

Entry-Level Responsibilities

Focus on building data pipelines, ETL/ELT processes, and designing data architectures.

Mid-Level Progression

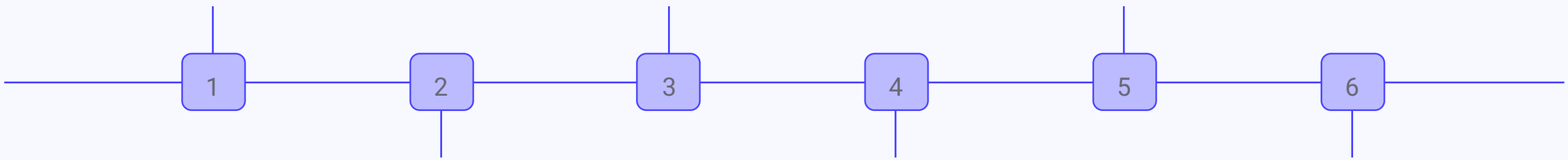
Advance to senior roles leading scalable platform design and complex data workflows.

Senior and Leadership Roles

Opportunities to become Data Architect, Lead Data Engineer, or Chief Data Officer.

Skills and Qualifications

Master programming, big data tools, cloud platforms, governance, and obtain certifications.



Final Thoughts & Resources

The Role of Data Engineering

Data engineering builds robust data architectures and pipelines, enabling reliable and high-quality data for analysis and decision-making.

Continuous Learning and Resources

Staying current with cloud, real-time processing, and AI requires ongoing learning through MOOCs, product documentation, blogs, and professional certifications.

Links:

Future of Jobs report 2025

Uk Salary Trends Report

Future Trends for Success

Microsoft Learn

Q&A and Contact Information

Q&A Sessions

1

2

Contact Information

What's app: 07778419886

Thank You