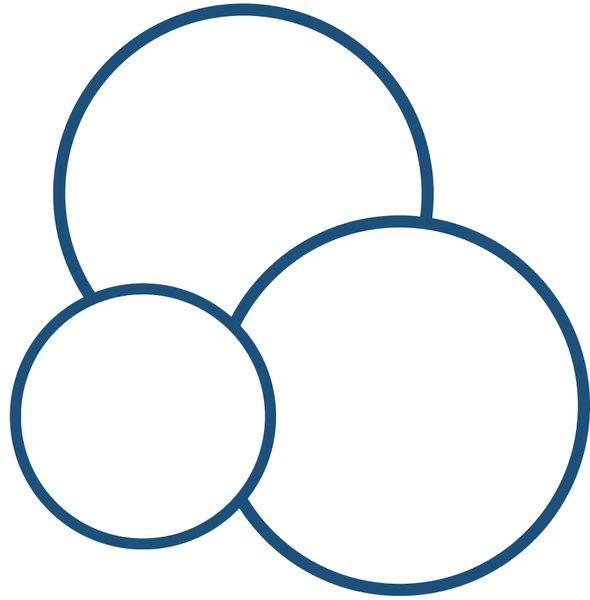
DATATECH LABS.

Data Engineering Course.

Week 6: Introduction to Cloud-Based Data Engineering Services

Outline: Week

- Introduction to cloud computing definitions and benefits.
- Amazon Web Services (AWS) for data engineering: Tools and demos
- Microsoft Azure for data engineering: Tools and demos.



Cloud Computing for Data Engineering

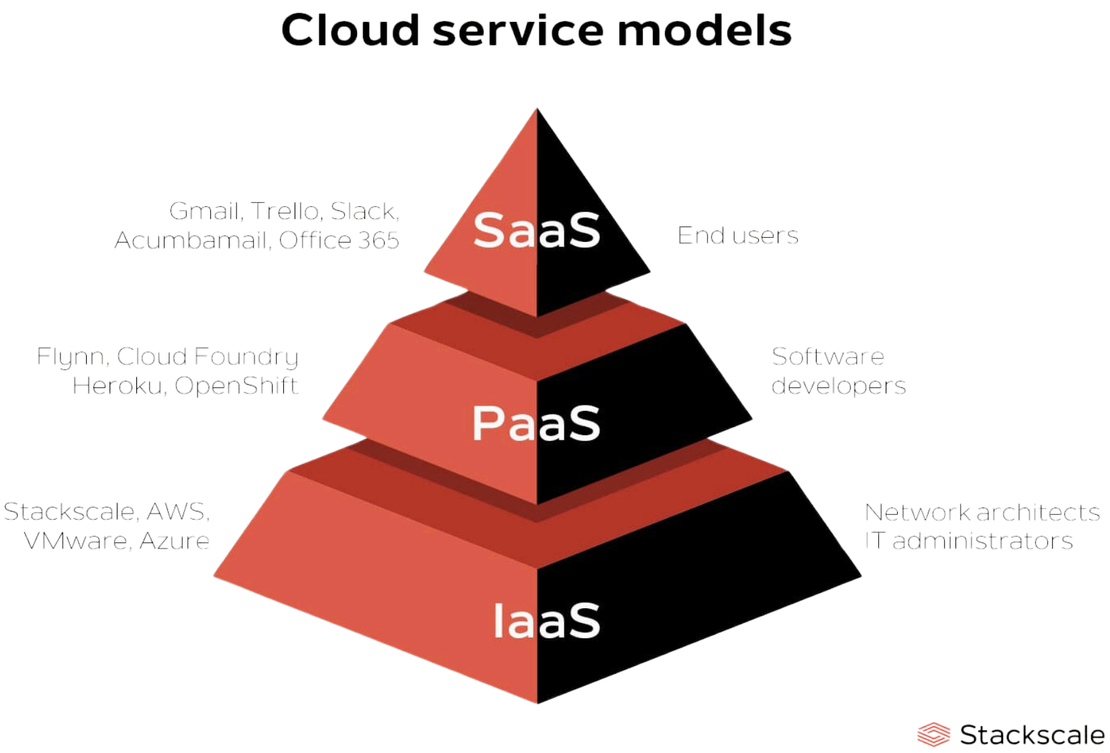
Cloud Computing Introduction

Introduction to Cloud Computing

- **Definition**: Cloud computing is the on-demand availability of computer resources, especially data storage and computing power, without direct active management by the user.
- Key characteristics:
  - On-demand self-service
  - Broad network access
  - Resource pooling
  - Rapid elasticity
  - Measured service

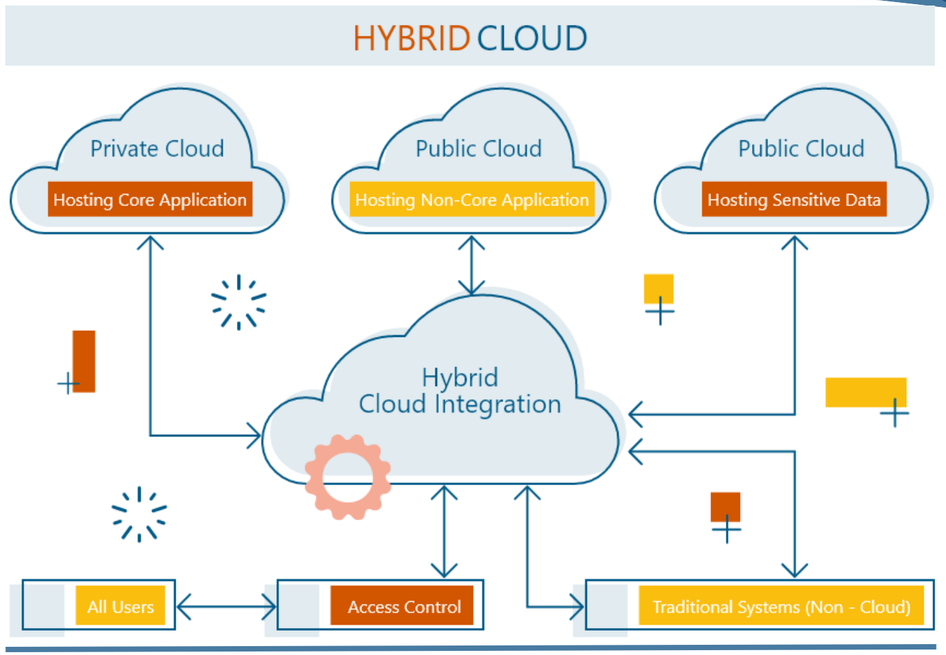
Cloud Service Models

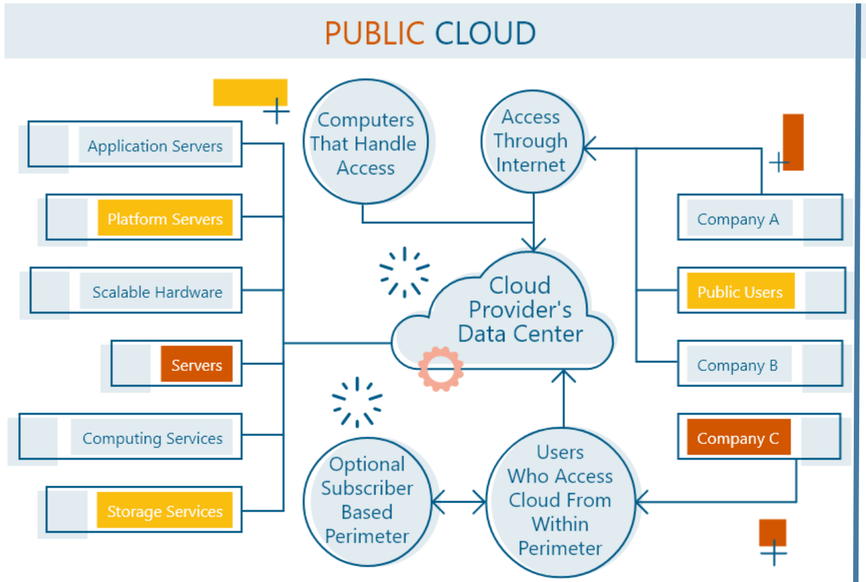
- Types of Cloud Services:
  - IaaS (Infrastructure as a Service)
  - PaaS (Platform as a Service)
  - SaaS (Software as a Service)

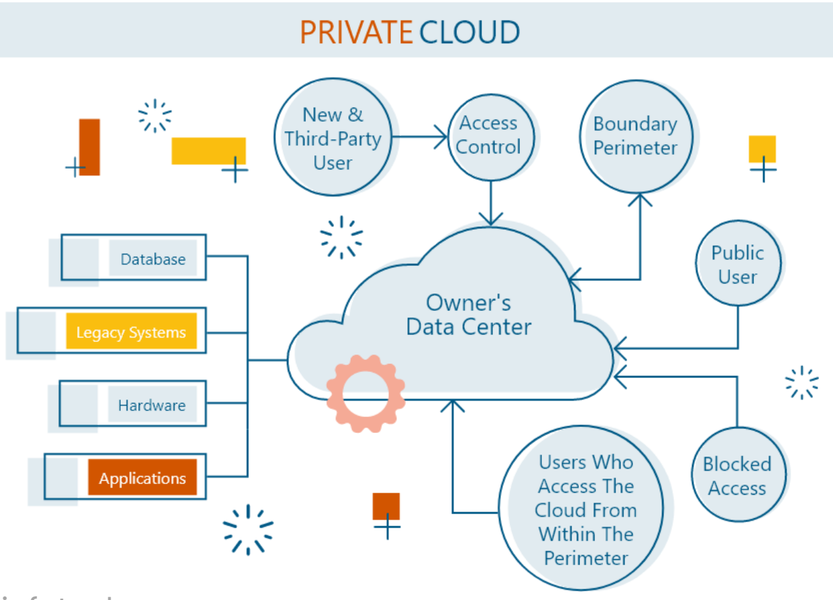


Cloud Deployment Models

- Public Cloud
- Private Cloud
- Hybrid Cloud



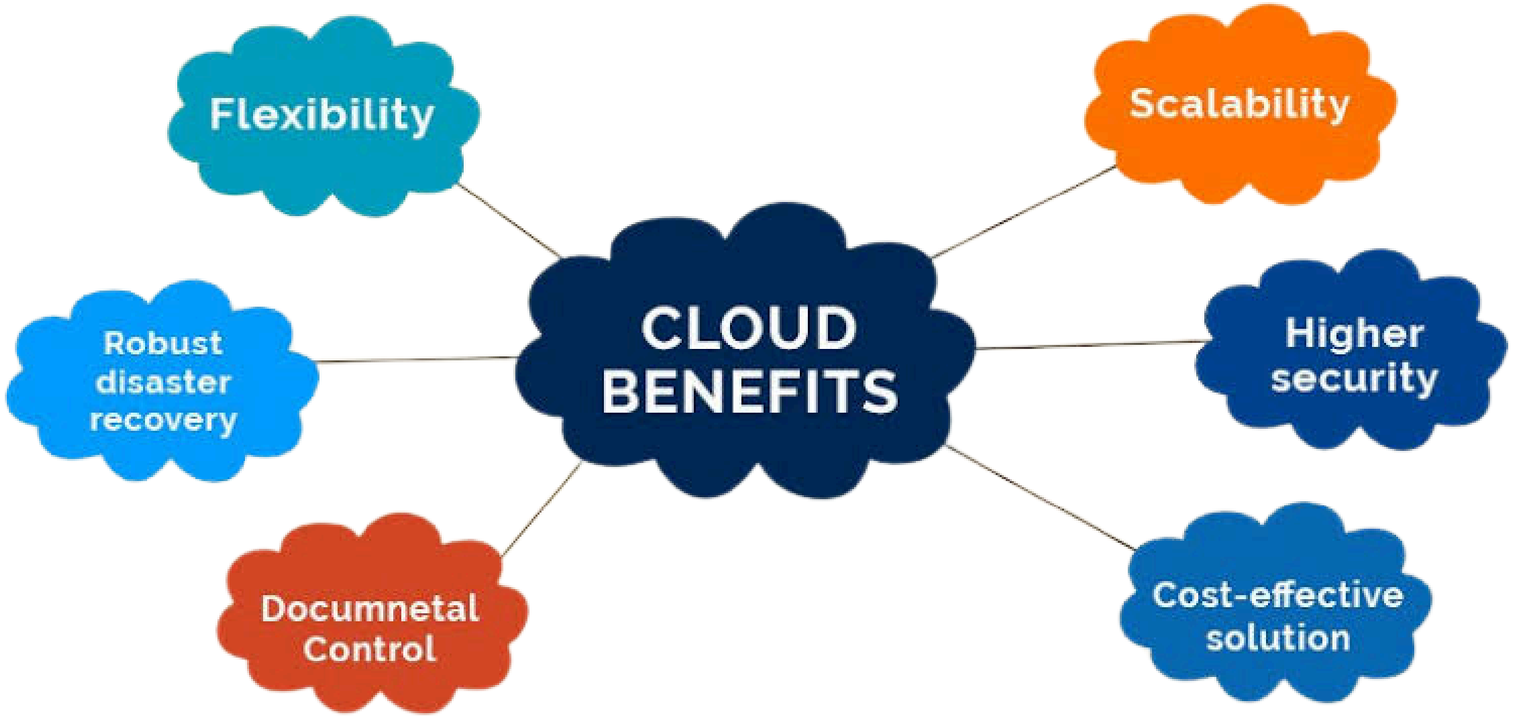




https://prakashinfotech.com

Benefits of Cloud Computing

- Cost efficiency
- Scalability
- Flexibility
- Disaster recovery
- Automatic updates
- Global reach



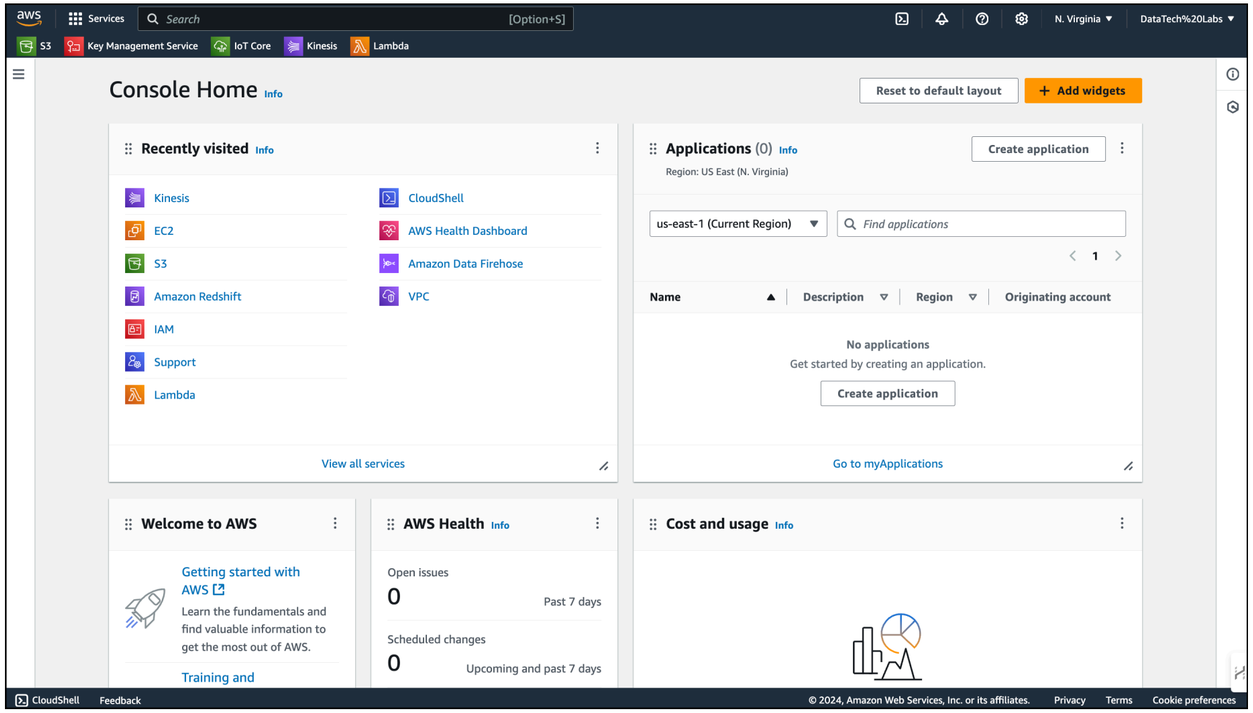
Cloud Service Providers Overview

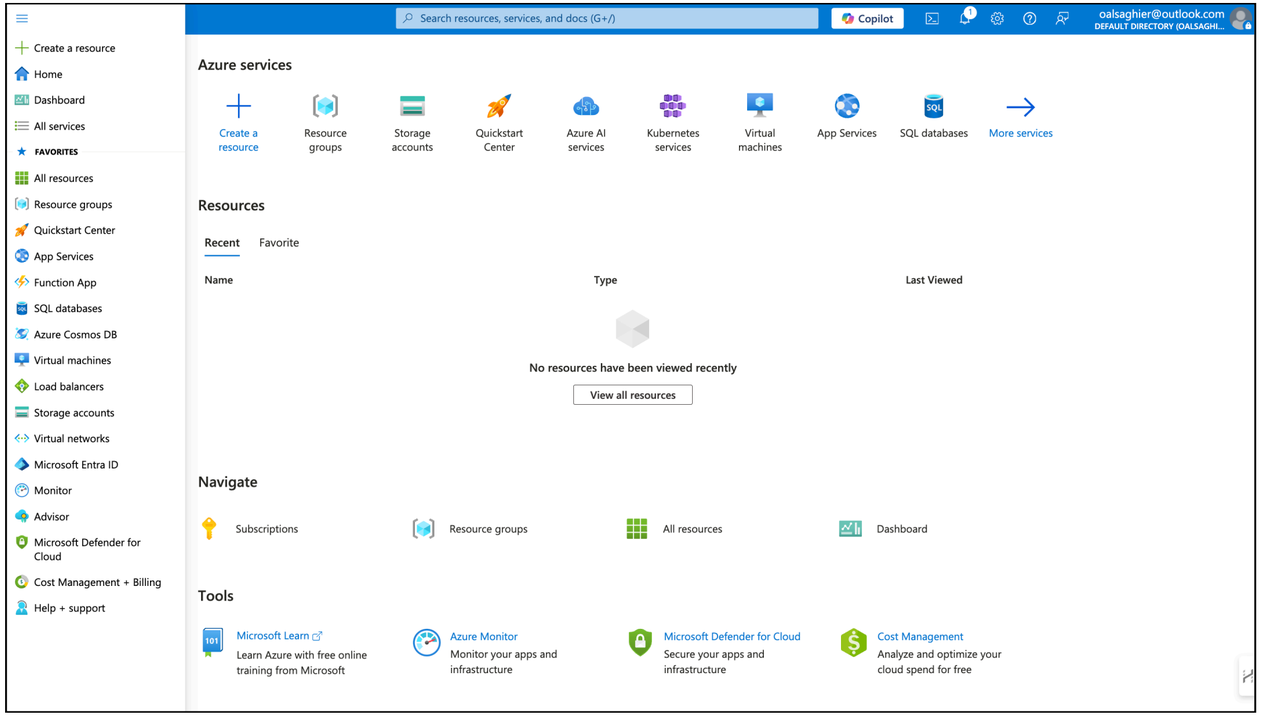
- Major players:
  - Amazon Web Services (AWS).
  - Microsoft Azure.
  - Google Cloud Platform (GCP).
- Our Focus: AWS and Azure are particularly strong in big data capabilities and services.

Google Cloud

Cloud Service Providers Overview

- Amazon Web Services (AWS).
- Microsoft Azure.





Amazon Web Services (AWS)

Introduction to AWS for Big Data

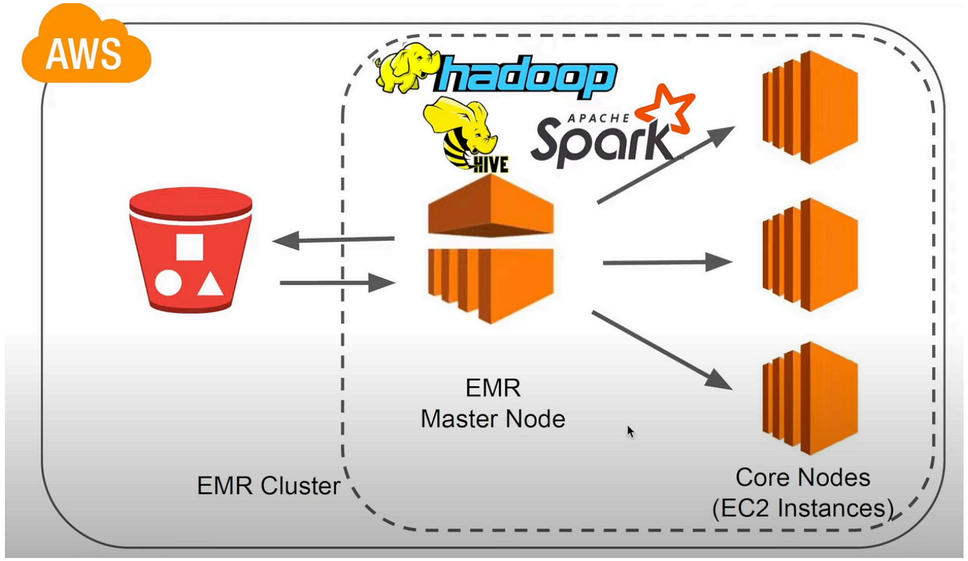
- AWS is a comprehensive cloud platform that offers a variety of services to build, deploy, and scale applications.
- Global infrastructure
  - AWS operates a robust, global network of infrastructure designed to deliver high availability and low latency.
- Regions and Availability Zones
  - AWS regions are globally distributed data centers
  - Each with multiple availability zones.

Core AWS Concepts

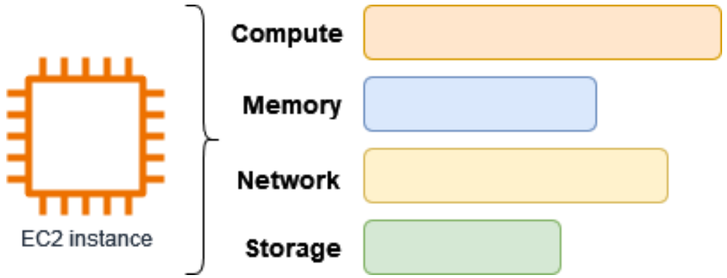
- AWS Management Console
  - A web-based interface for managing AWS resources, providing an easy-to-use dashboard for configuration and monitoring.
- AWS CLI
  - A command-line tool that enables users to interact with AWS services programmatically, ideal for automation and scripting.
- AWS SDK
  - Software Development Kits that allow developers to integrate AWS services into their applications in various programming languages.

Introduction to AWS for Big Data

- Big Data Services in AWS:
  - **Storage**: Amazon S3
  - **Compute**: EC2, EMR (Elastic MapReduce)
  - **Data Analytics**: Amazon Athena, Redshift
  - **Machine Learning**: Amazon SageMaker



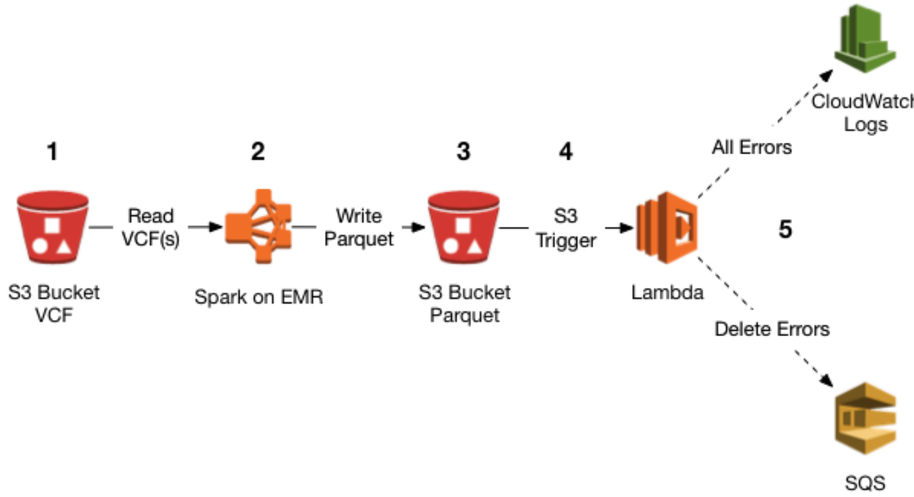
amazon S3

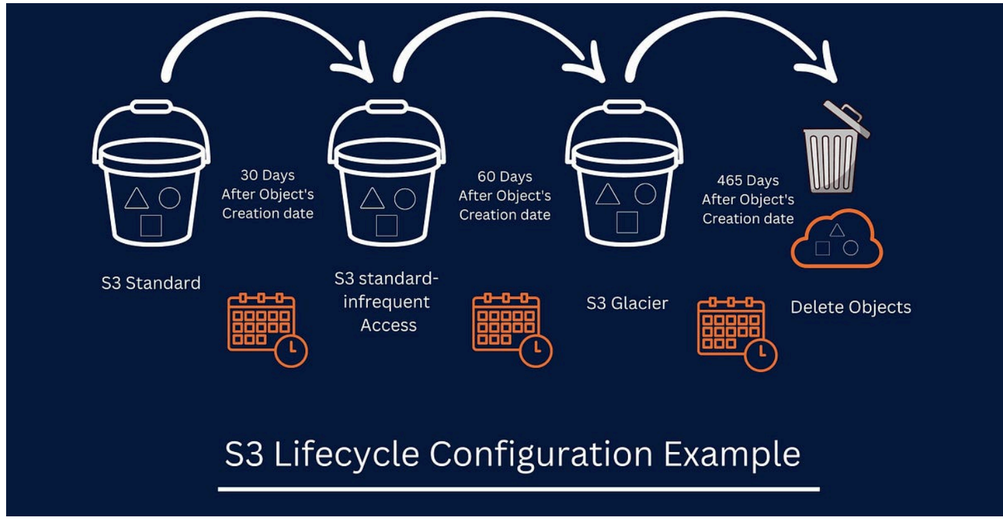


Amazon **Redshift**

AWS Storage Services

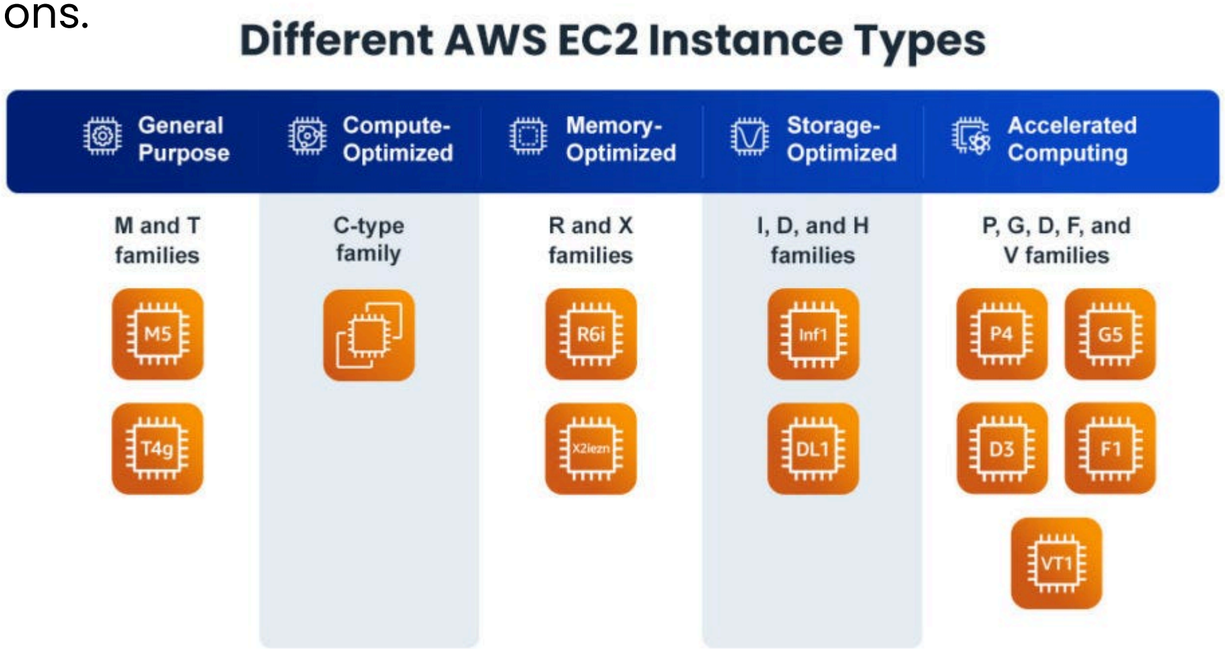
- **Amazon S3** (Simple Storage Service)
  - **Definition**: An object storage service known for scalability, durability, and ease of use.
  - Features:
    - **Buckets and objects**: Buckets are containers for data objects in S3.
    - **Storage classes**: Options like Standard, Intelligent-Tiering, and Glacier offer varying cost and access levels.
    - **Access control**: Supports fine-grained permissions for security.
    - **Versioning**: Enables keeping multiple versions of objects.
  - Use Cases: Storing large datasets, data lakes.





AWS Computing Services

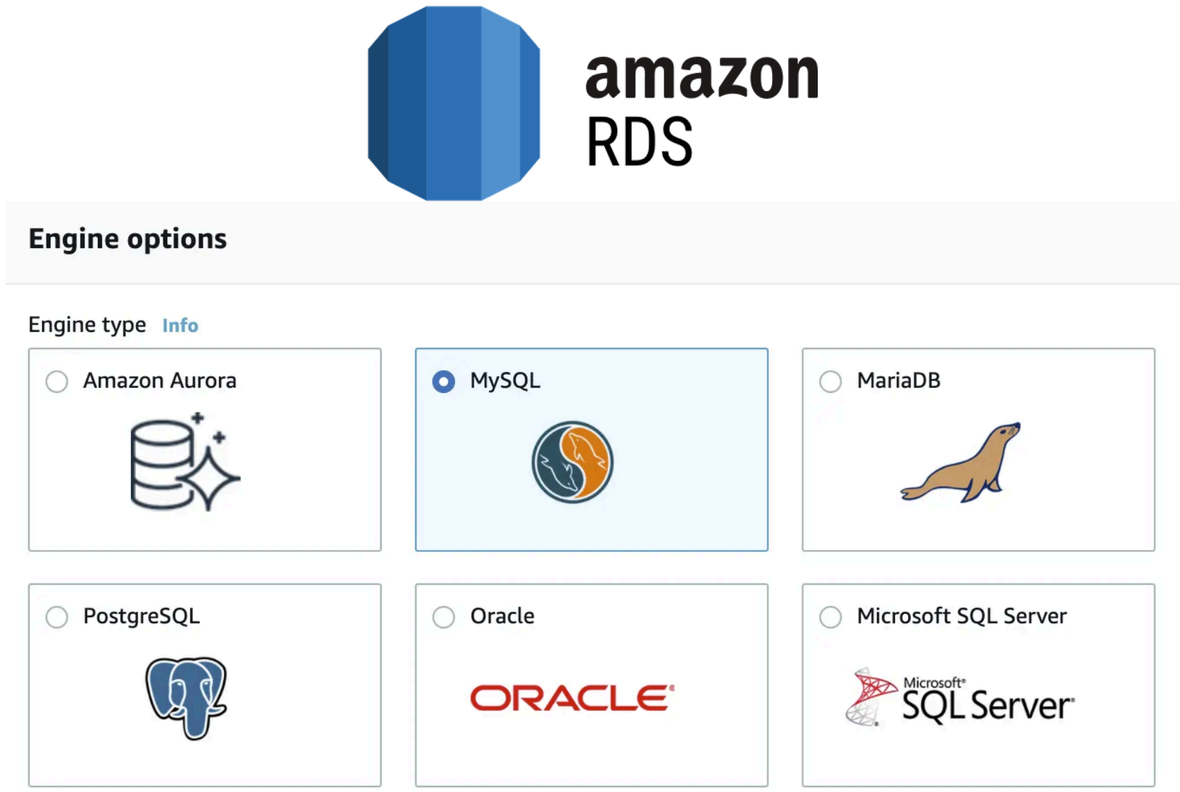
- **Amazon EC2** (Elastic Compute Cloud):
  - Definition: Scalable virtual servers in the cloud.
  - Features:
    - On-demand instances
    - Auto-scaling.
  - Use Cases: Running custom big data applications.



AWS Database Services

- **Amazon RDS** (Relational Database Service)
  - **Supported engines**: Includes popular databases like MySQL, PostgreSQL, SQL Server, and Oracle.
  - **Automated backups**: Automatic backups for recovery.
  - **Multi-AZ deployment**: For ensures high availability

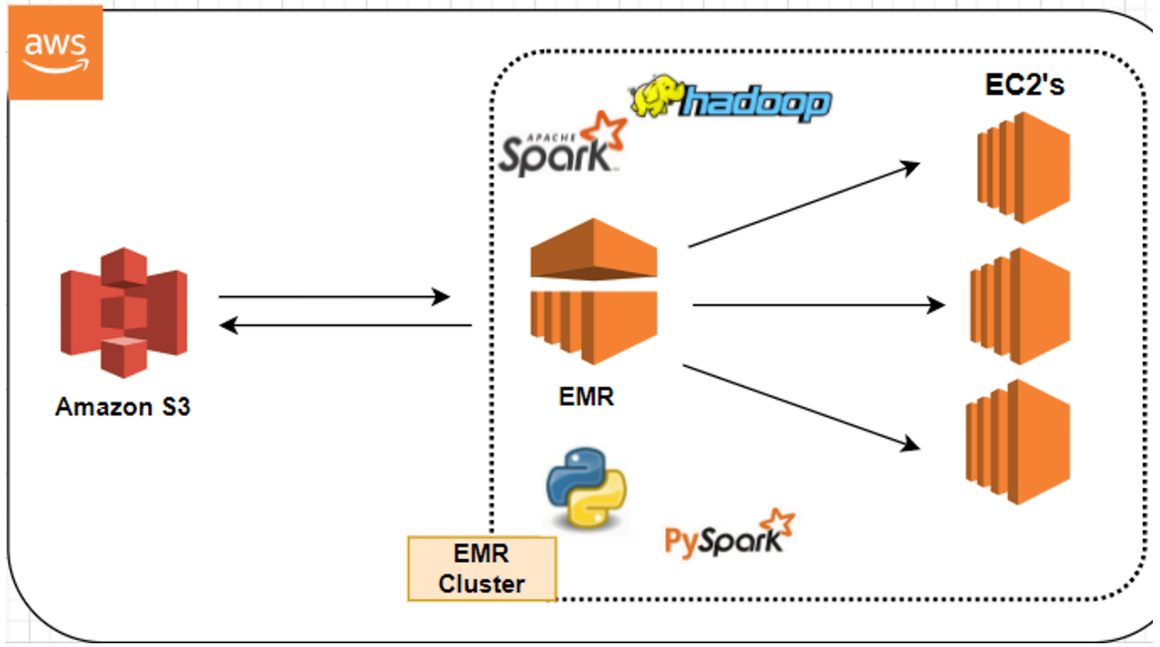
- **Amazon DynamoDB** (NoSQL database)
  - **Supports data models**: both key-value and document.
  - **Auto-scaling**: Dynamically adjusts throughput capacity to maintain performance.



AWS Big Data Services

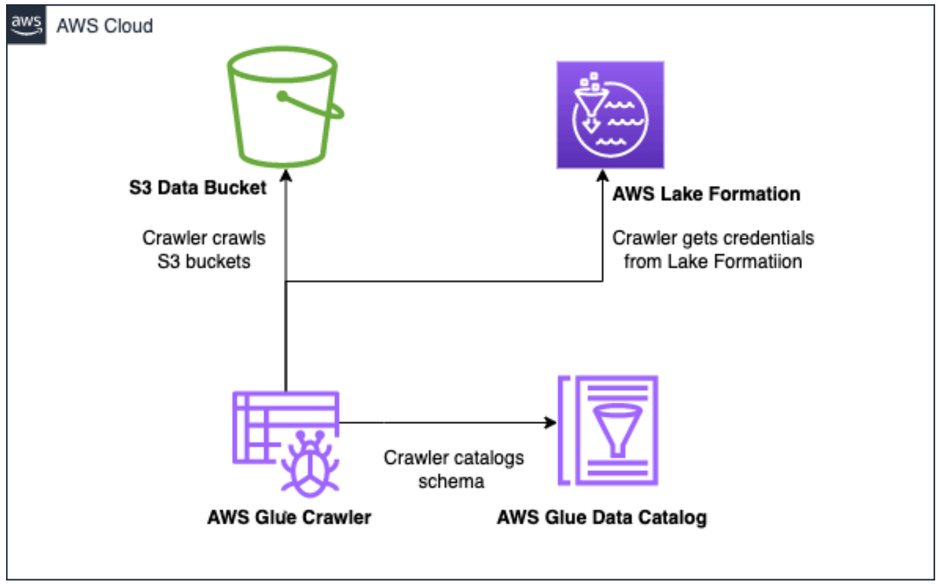
- **Amazon EMR** (Elastic MapReduce)
  - **Purpose**: Big data processing and analytics.
  - **Components**: Hadoop ecosystem, HDFS, Spark, Hive, Integration with other AWS services
  - **Benefits**: Easy setup for Hadoop and Spark clusters.

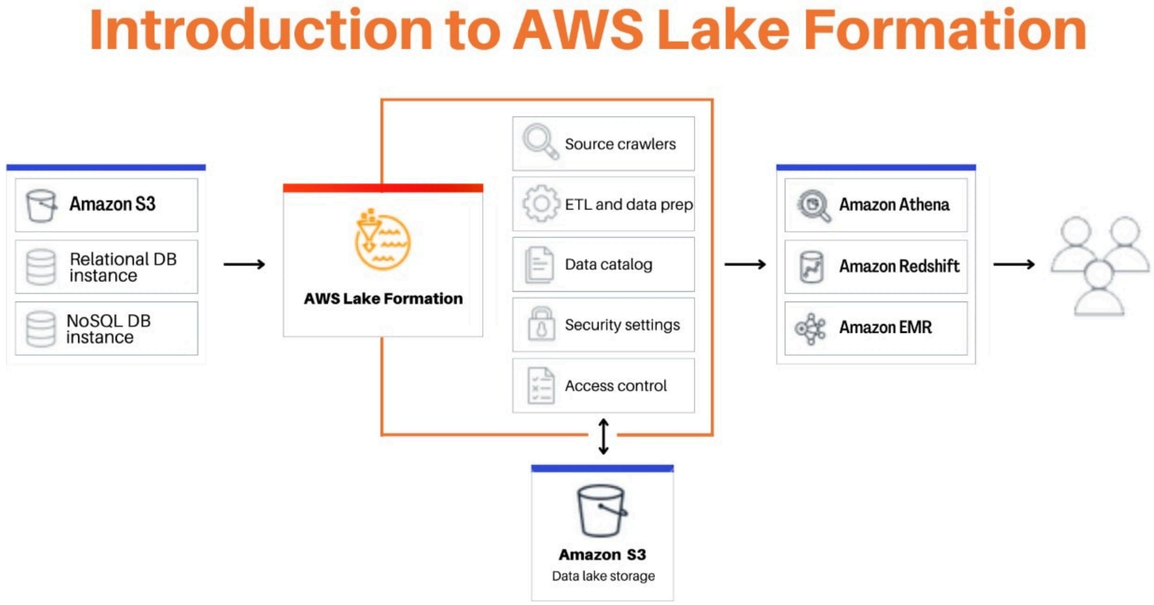
- Amazon EMR Architecture:
  - Master node
  - Core nodes
  - Task nodes
  - EMRFS (EMR File System)



AWS Big Data Services

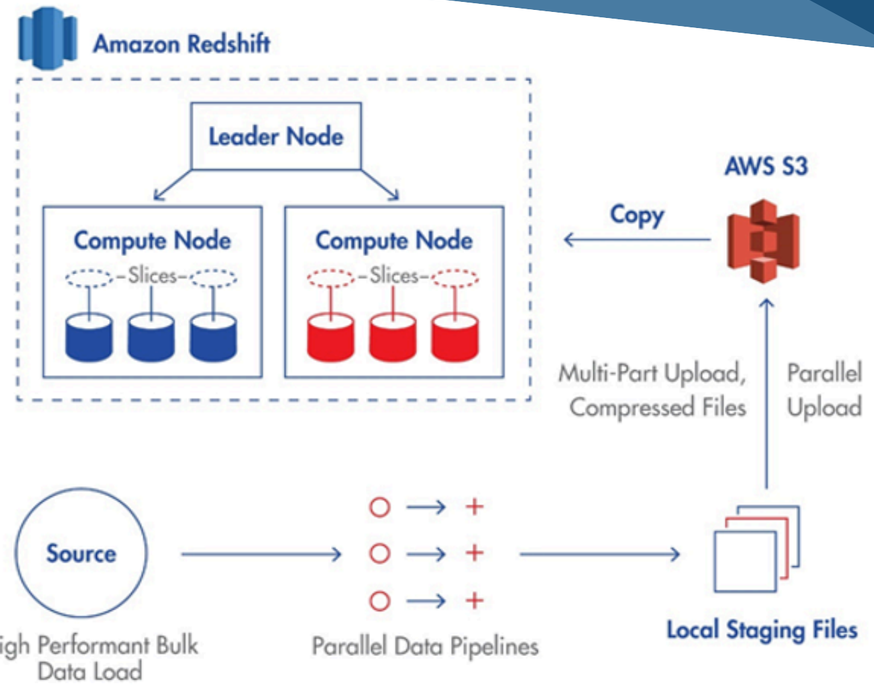
- **AWS Data Lake Formation**
  - **Components**: Services for ingesting, cataloging, and securing data for analysis.
  - **Lake Formation permissions**: Manages data access control for users and applications.
  - **Data catalog**: A centralized catalog storing metadata about the data lake's content.

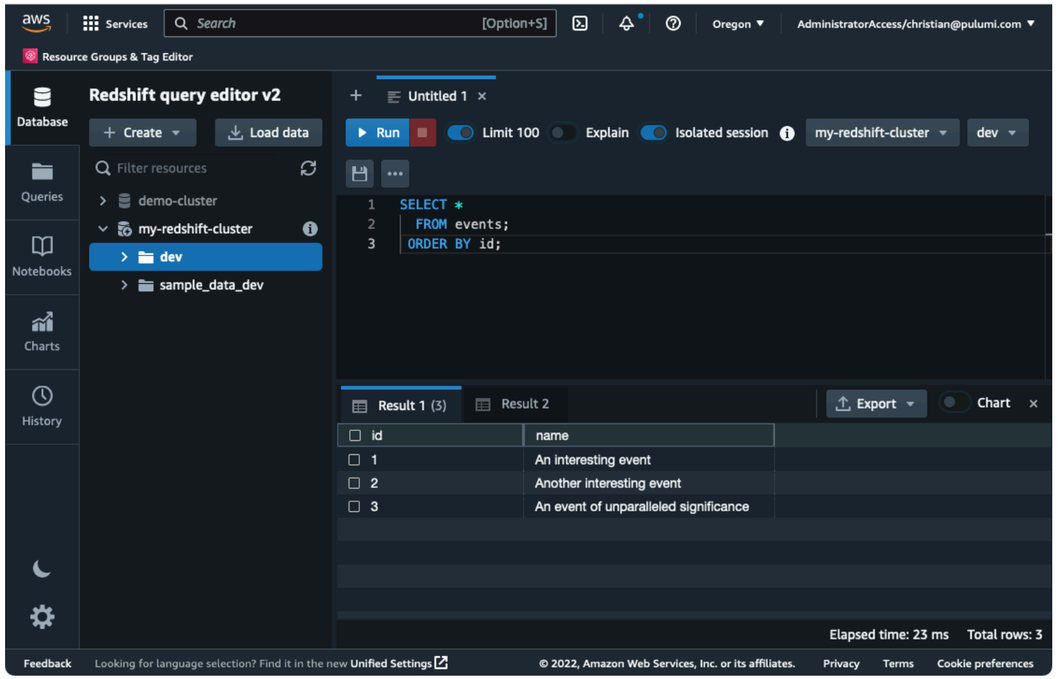




Data Analytics on AWS

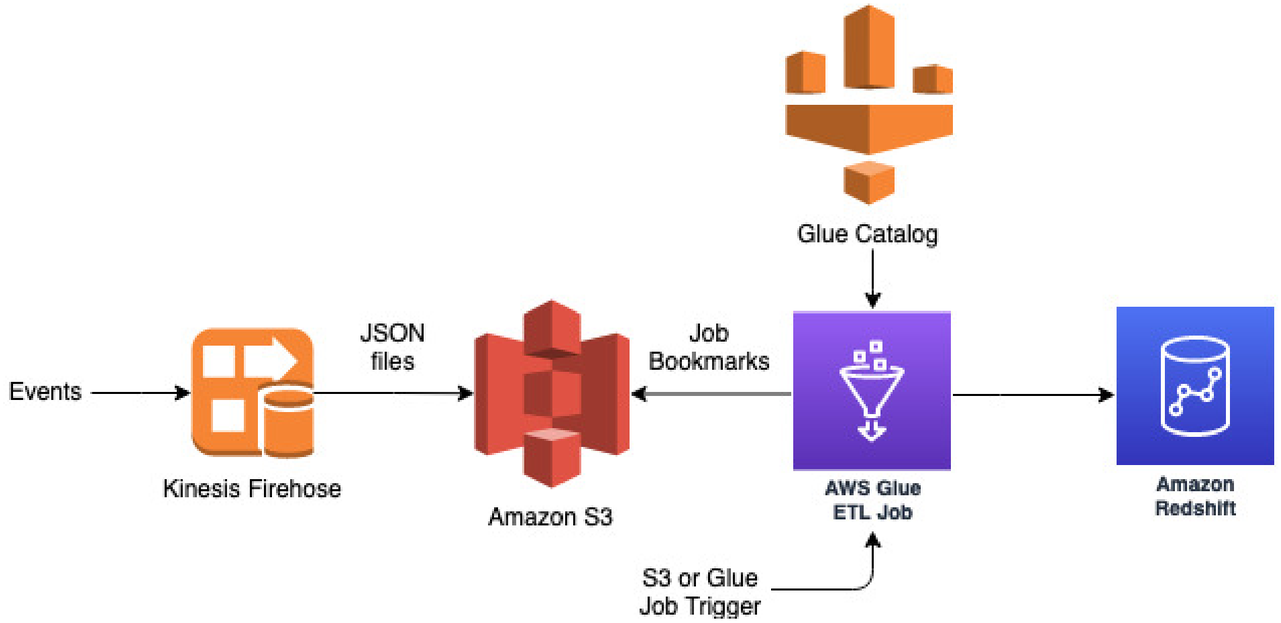
- **Amazon Redshift**:
  - Definition: Fully managed data warehouse service.
  - Features:
    - Massively parallel processing
    - Petabyte-scale data.
    - Columnar storage
  - Use Cases: Data warehousing, BI analytics (Integration with BI tools).





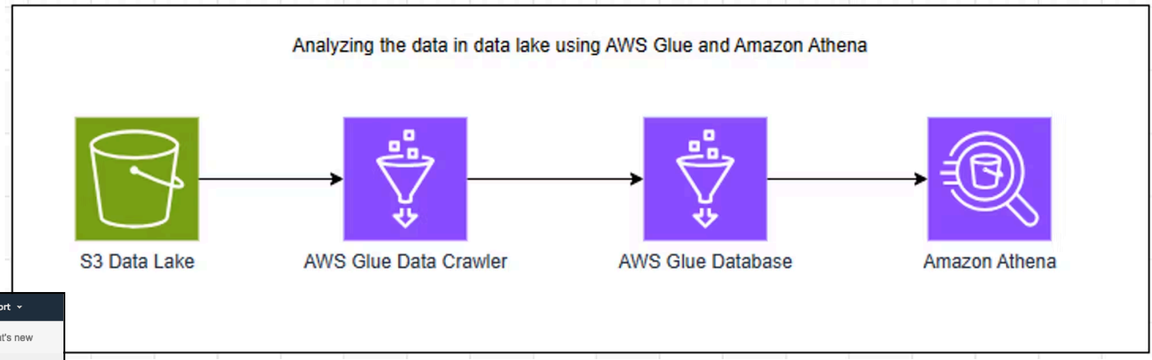
ETL Services on AWS

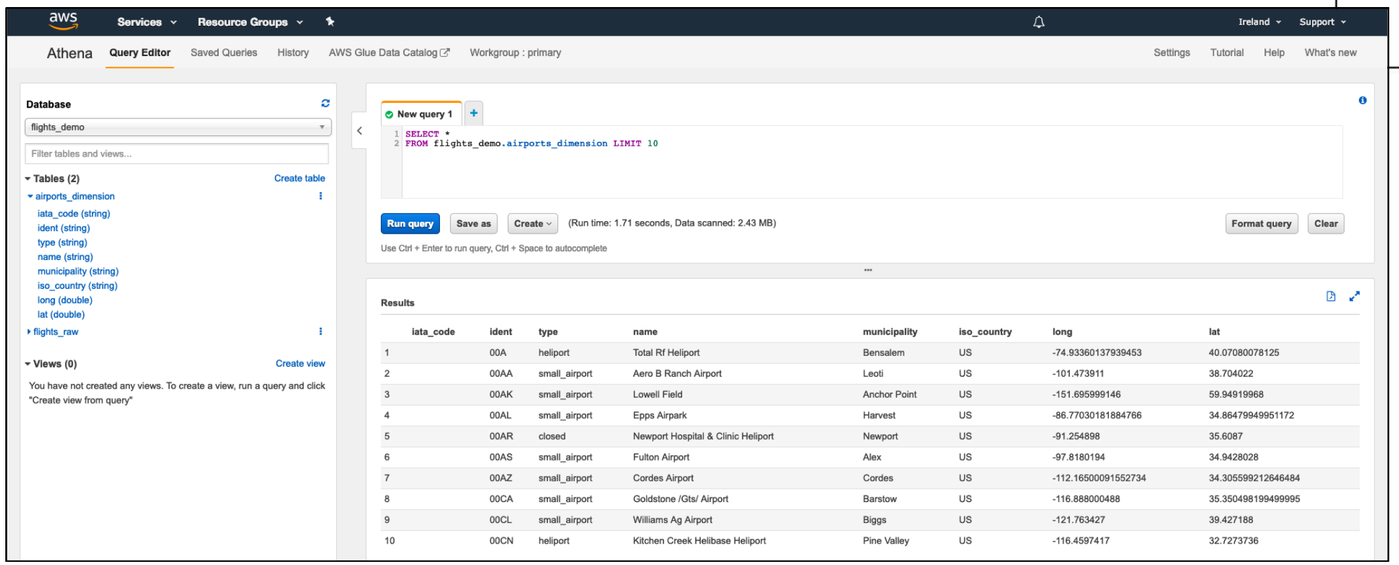
- **AWS Glue**:
  - **ETL service**: Automates extract, transform, and load processes.
  - **Data catalog**: Centralized metadata repository that keeps track of data location, schema, and other properties.
  - **Job scheduling**: Allows automated scheduling of ETL jobs.
  - **Development endpoints**: Provisioned environments for developing and testing code before production deployment.



Querying Data with AWS

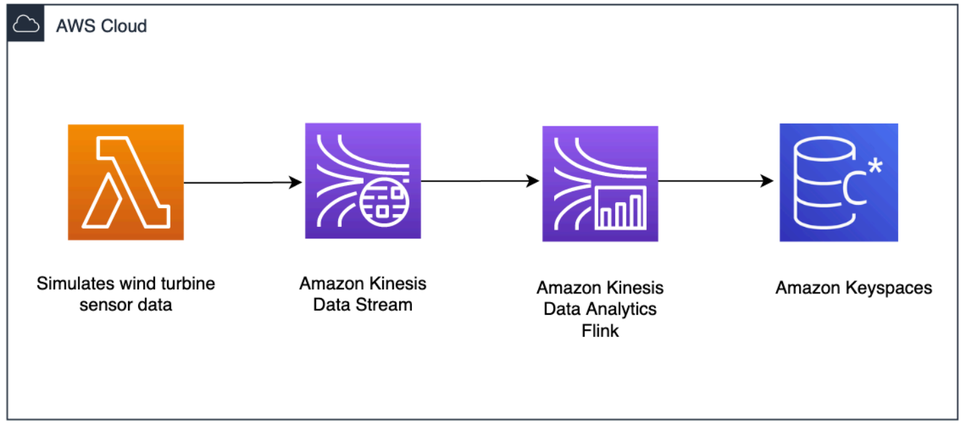
- **Amazon Athena**:
  - Definition: Interactive query service using SQL.
  - Key Feature: Serverless; no need to manage infrastructure.
  - Use Case: Query data directly from S3.

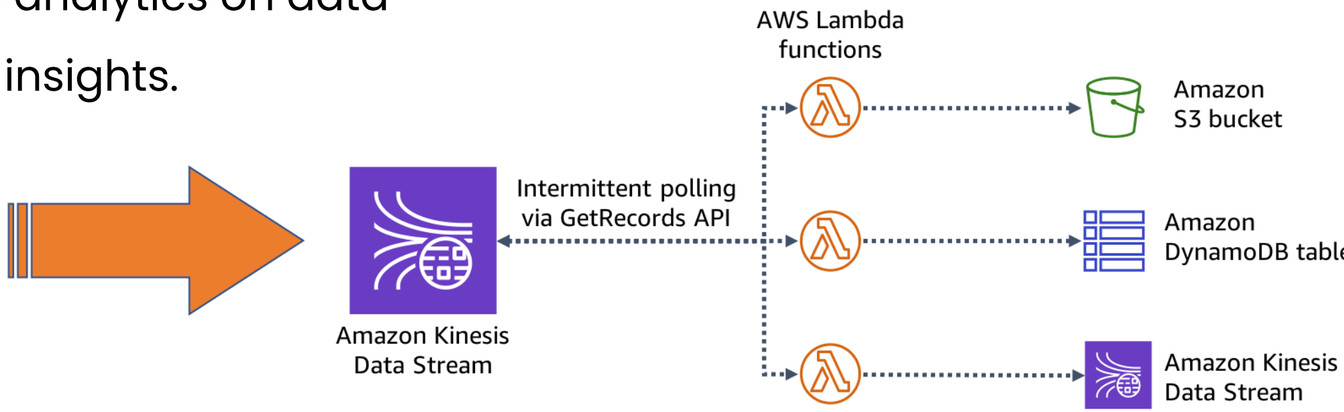




Data Streaming with AWS

- **Amazon Kinesis**: Real-time data streaming
- Processes and analyzes streaming data in real-time, useful for time-sensitive applications.
  - **Kinesis Data Streams**: Allows real-time ingestion and processing of high-volume streaming data.
  - **Kinesis Data Firehose**: Delivers streaming data to destinations like S3, Redshift, or Elasticsearch for storage and analysis.
  - **Kinesis Data Analytics**: Enables real-time analytics on data streams with SQL queries, providing quick insights.





Demo:

AWS for Data Engineering

Microsoft Azure

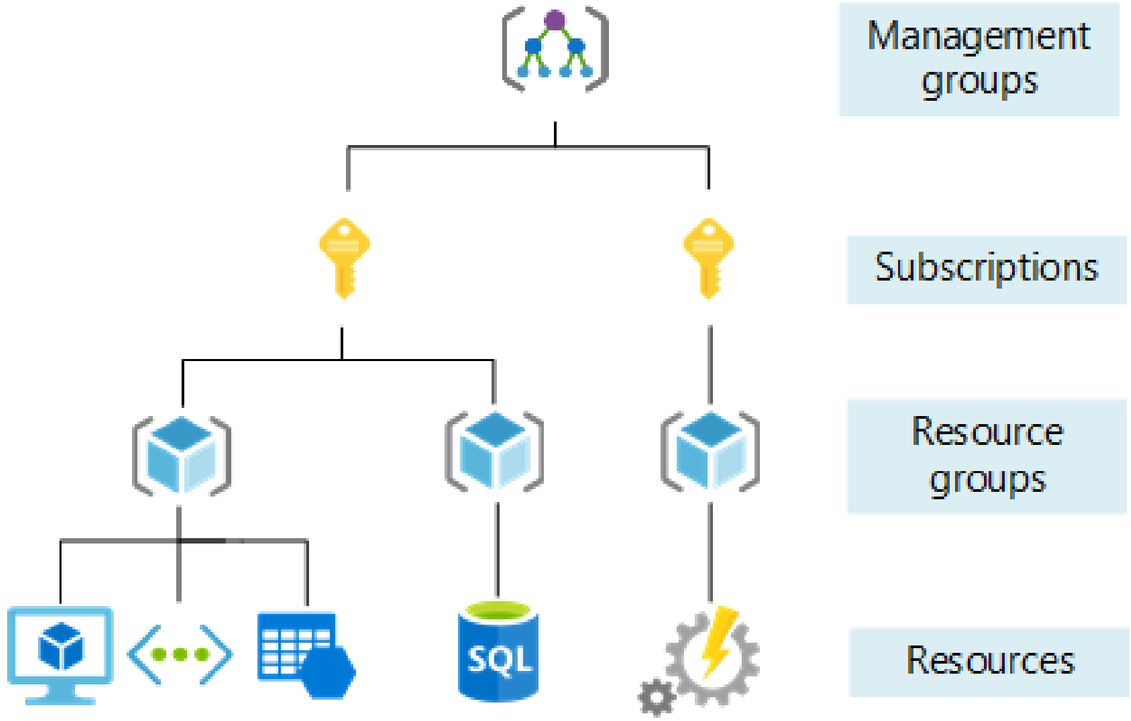
Introduction to Azure for Big Data

- **Azure**: is Microsoft's cloud platform, providing various services for computing, analytics, storage, and networking.
- **Global infrastructure**: A global network of data centers delivering secure and reliable services.
- **Regions and Availability Zones**: Each Azure region comprises multiple availability zones, ensuring high availability and resilience.

- Azure Big Data Services:
  - **Storage**: Azure Blob Storage, Data Lake
  - **Compute**: HDInsight, Azure Databricks
  - **Analytics**: Azure Synapse Analytics, Azure Stream Analytics

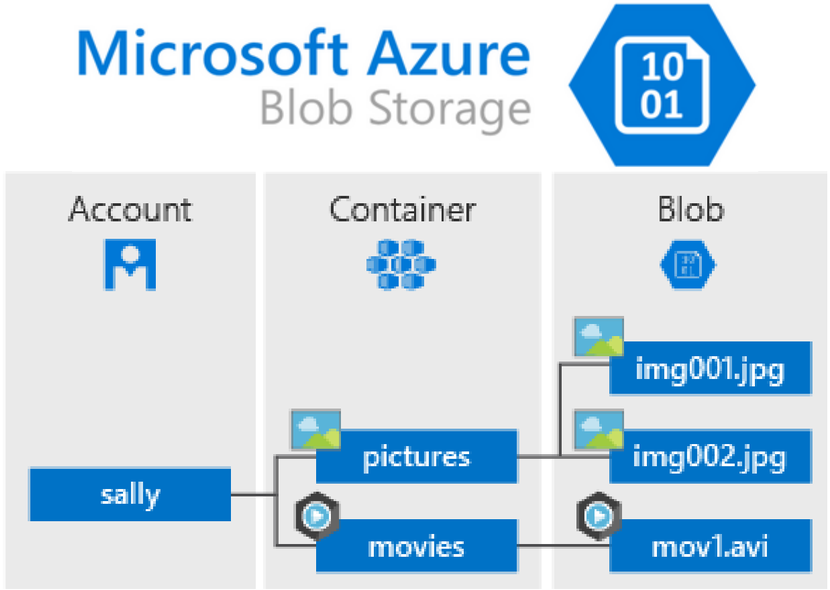
Core Azure Concepts

- **Resource Groups**: Logical containers for managing and organizing Azure resources for better access and billing management.
- **Subscriptions**: Azure subscriptions are the top-level containers for billing and access to Azure services.
- **Azure CLI**: Command-line tool for managing Azure resources, enabling automation and scripting.
- **Azure PowerShell**: A scripting tool for managing Azure services and infrastructure from the command line.
- **Azure Cloud Shell**: An online shell environment that provides both CLI and PowerShell for managing Azure resources.

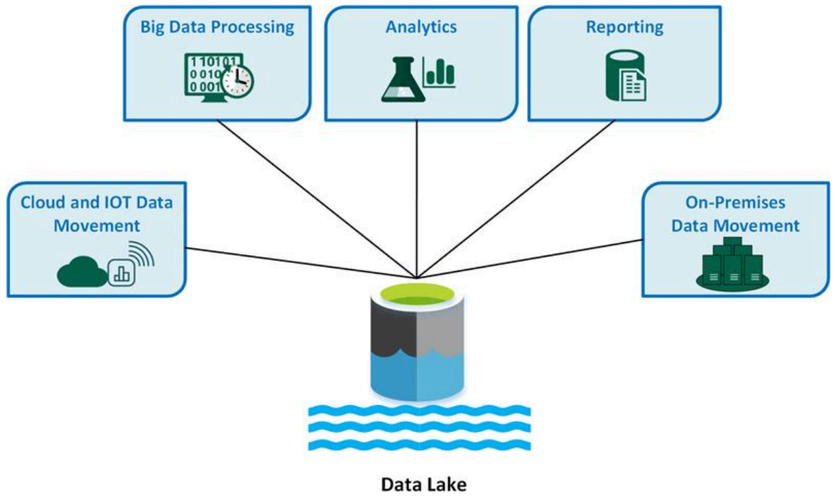


Azure Storage Services

- **Azure Blob Storage**:
  - Definition: Scalable storage for unstructured data in Containers and blobs.
  - Key Features: Access tiers (Hot, Cool, Archive) for cost optimization.
  - Use Case: Data lakes, archival data storage.

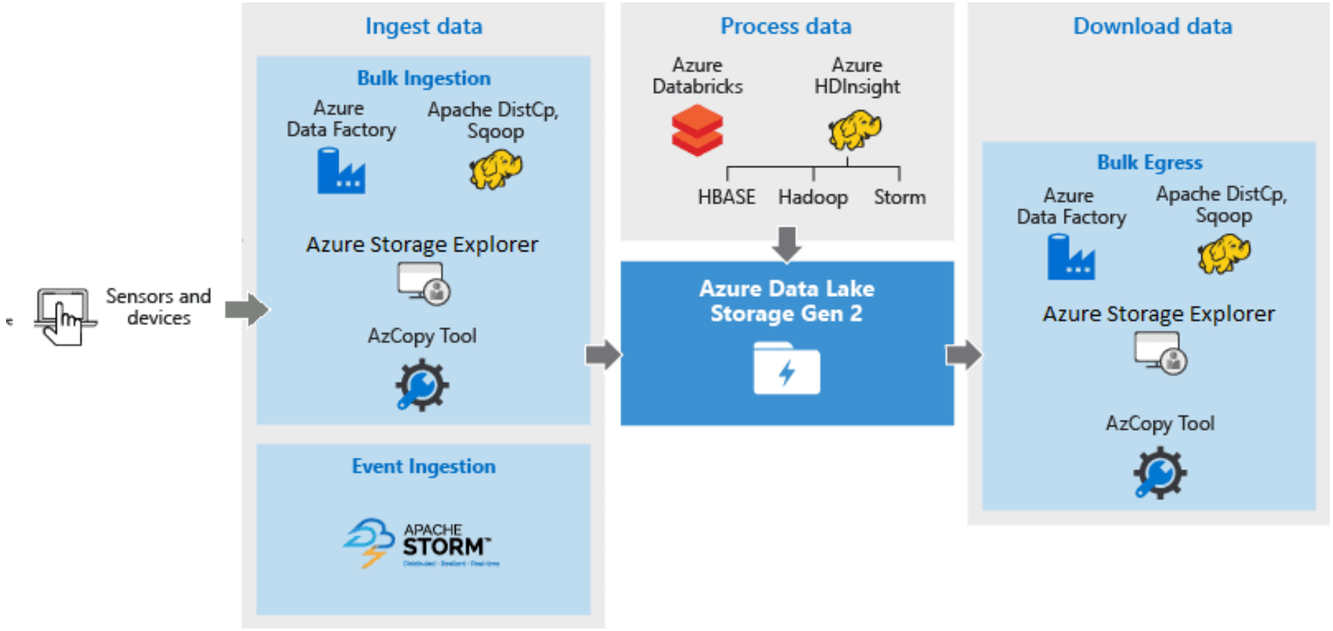


- **Azure Data Lake Storage**:
  - Definition: Optimized storage for big data analytics.
  - Features: High-performance, hierarchical namespace.
  - Use Case: Storing and processing large datasets.



Azure Storage Services

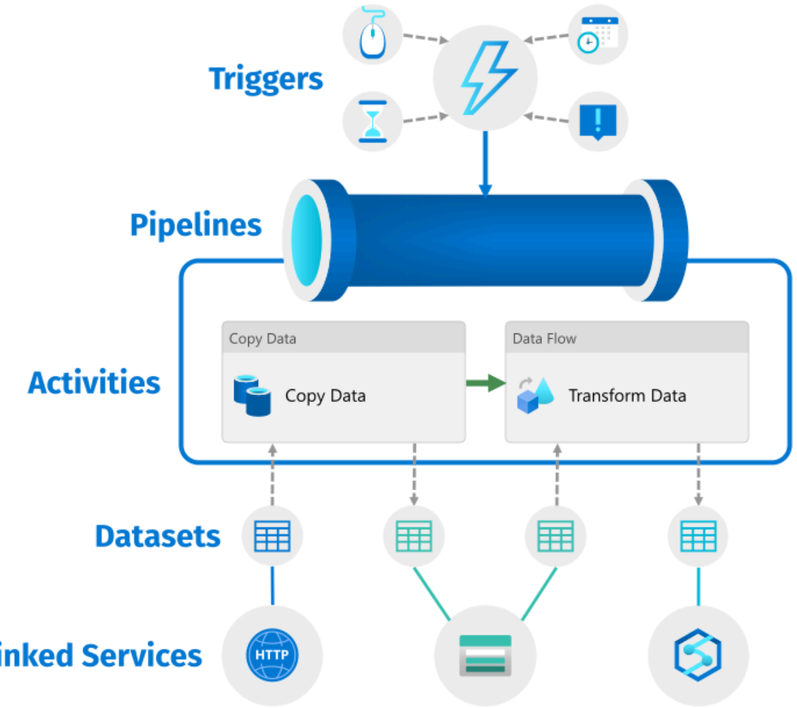
- **Azure Data Lake Storage Gen2**:
  - Hierarchical namespace
  - It is built on Azure Blob storage and has all the key features of ADLS Gen1.
  - Security and access control
  - Integration with big data services: ADLS Gen2 permits you to access and manage data just as you would with a HDFS.



Data Integration Services

- **Azure Data Factory**: A managed service for building, scheduling, and orchestrating ETL workflows across data sources.
- Components:
  - **Pipelines**: Defined workflows that perform data movement and transformation.
  - **Activities**: Specific tasks within a pipeline, such as data copying, transformations, or script execution.
  - **Datasets**: Representations of data stored in Azure or external sources, used within pipelines.
  - **Linked services**: Connections to data sources, storage, or compute environments.

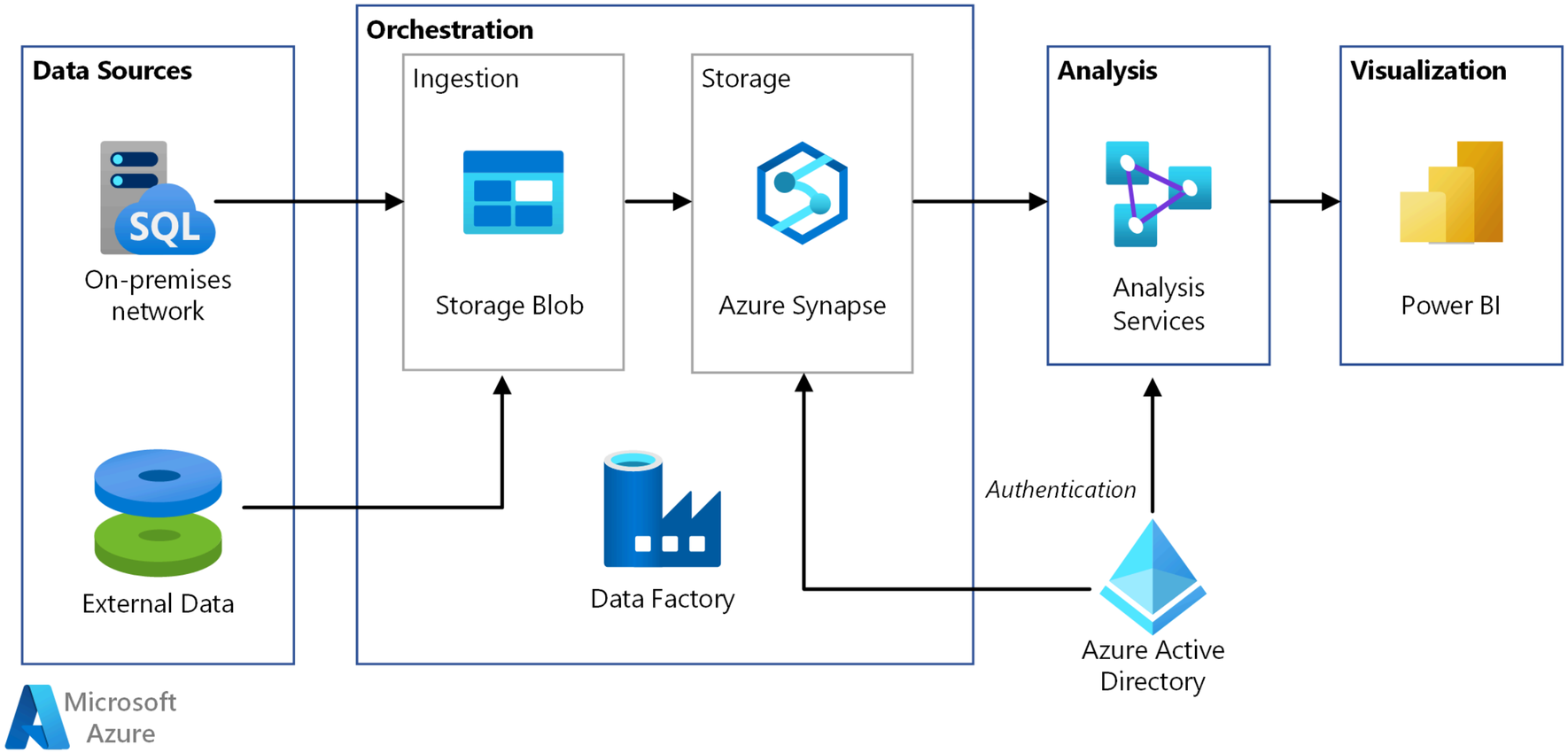
Azure Data Factory



© 2021 Cathrine Wilhelmsen (hi@cathrinew.net)

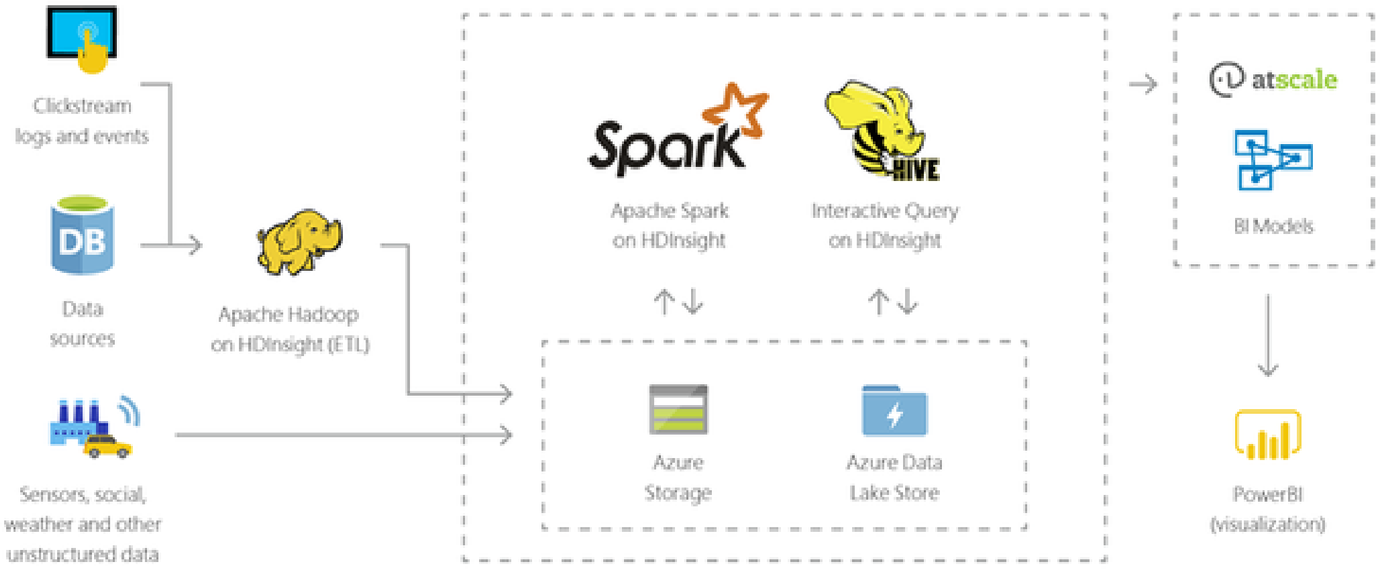
Data Integration Services

- **Azure Data Factory**



Data Processing on Azure

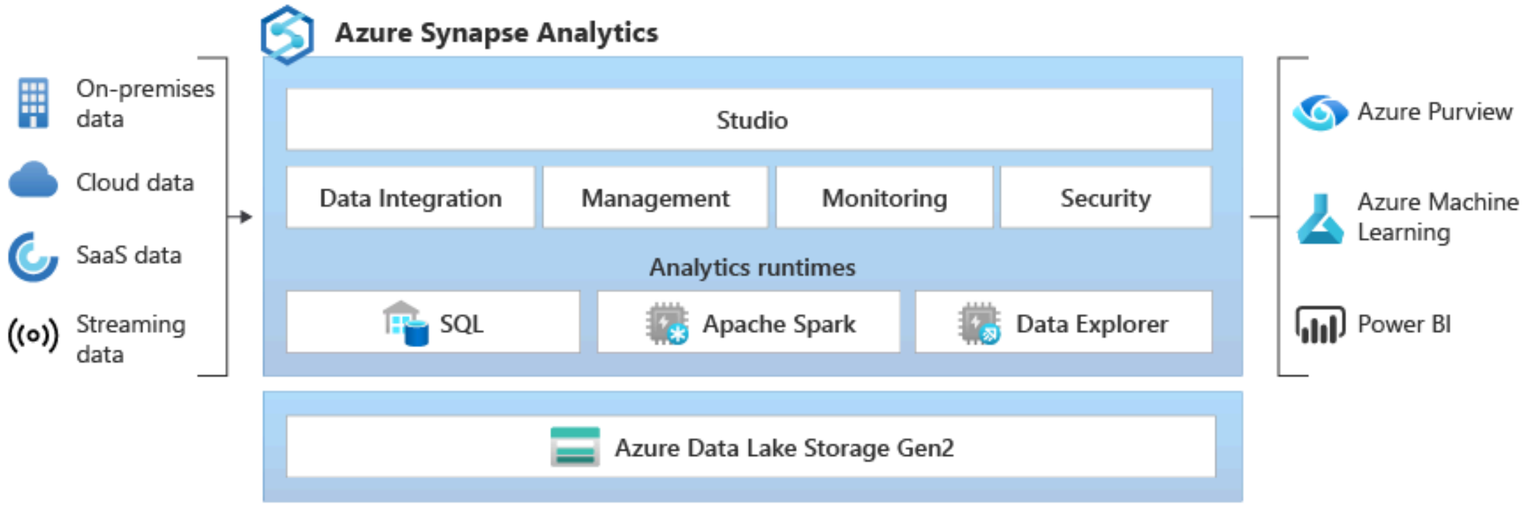
- **Azure HDInsight**:
  - Purpose: Managed big data services for Hadoop, Spark, Kafka.
  - Hadoop-based service. Supported cluster types:
    - Hadoop
    - Spark
    - HBase
    - Kafka
  - Benefits: Fast deployment, flexible cluster configurations.



Advanced Analytics on Azure

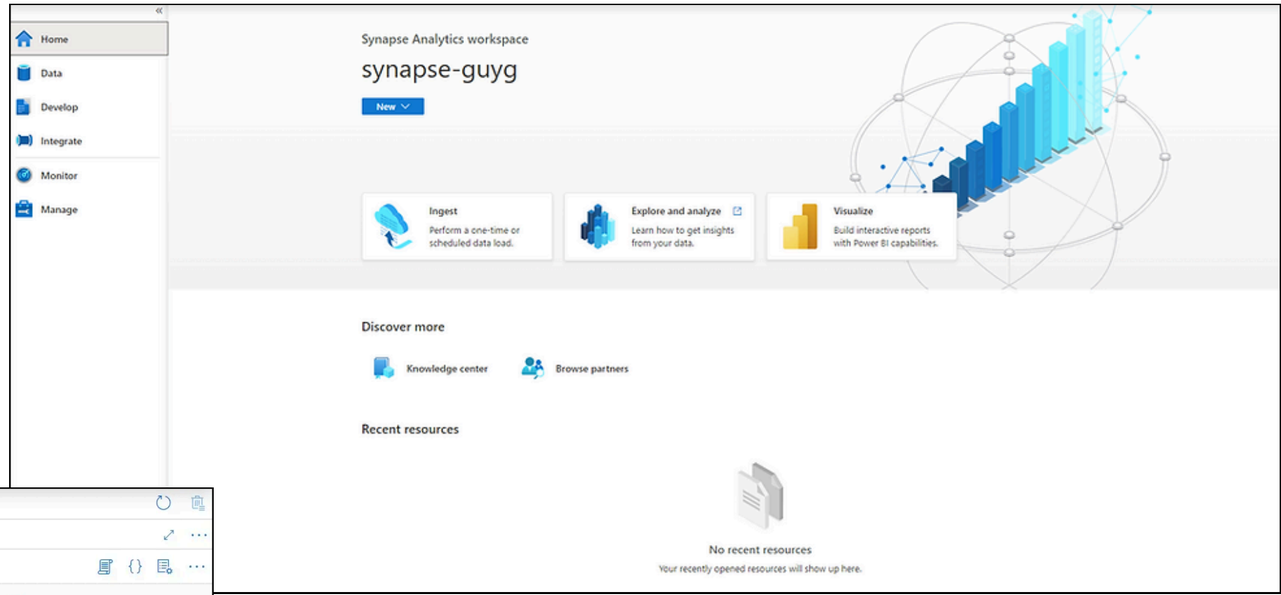
- **Azure Synapse Analytics**:
  - Definition: Unified analytics platform. Integrated data warehousing and big data analytics.
  - Key Features: Serverless and dedicated pools, Synapse Studio.
  - Use Cases: Data warehousing, real-time data processing.
  - Components:
    - SQL pools (dedicated and serverless)
    - Spark pools
    - Data integration
    - Studio experience

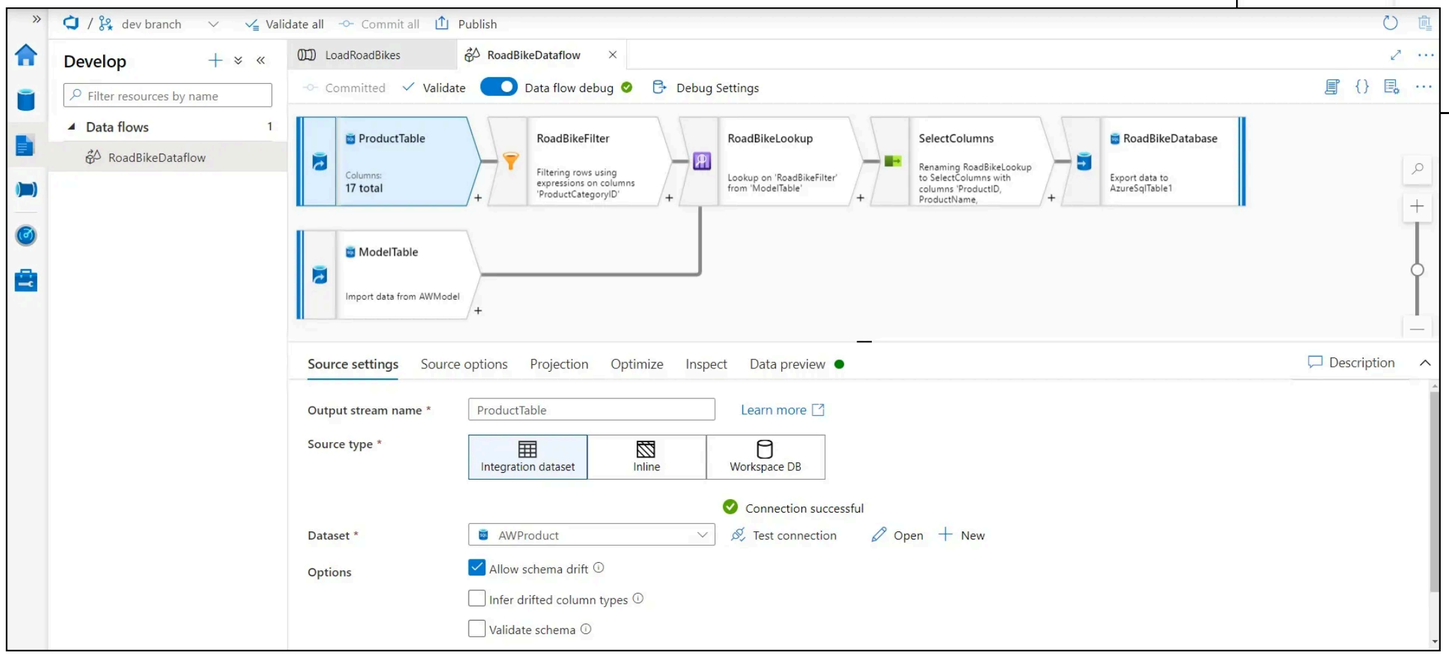
Azure
Synapse Analytics



Advanced Analytics on Azure

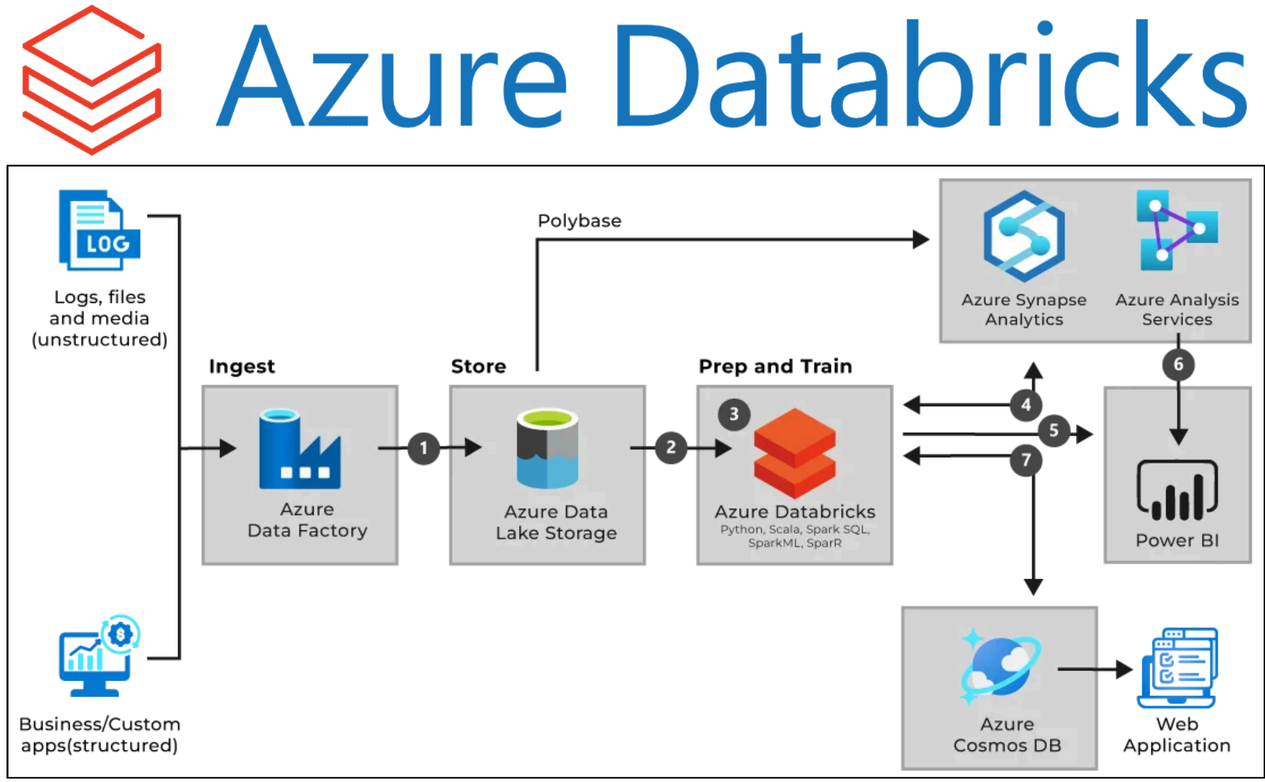
- **Azure Synapse Analytics**





Advanced Analytics on Azure

- **Azure Databricks**
  - A collaborative analytics platform optimized for Apache Spark on Azure, ideal for big data and AI.
- **Architecture overview**: Azure Databricks integrates with Azure's security, storage, and data services for seamless operation.
  - **Workspace components**: Includes notebooks, clusters, and jobs for interactive and batch data processing.
  - **Integration with other Azure services**: Works closely with services like Azure Data Lake, Blob Storage, and Synapse Analytics.

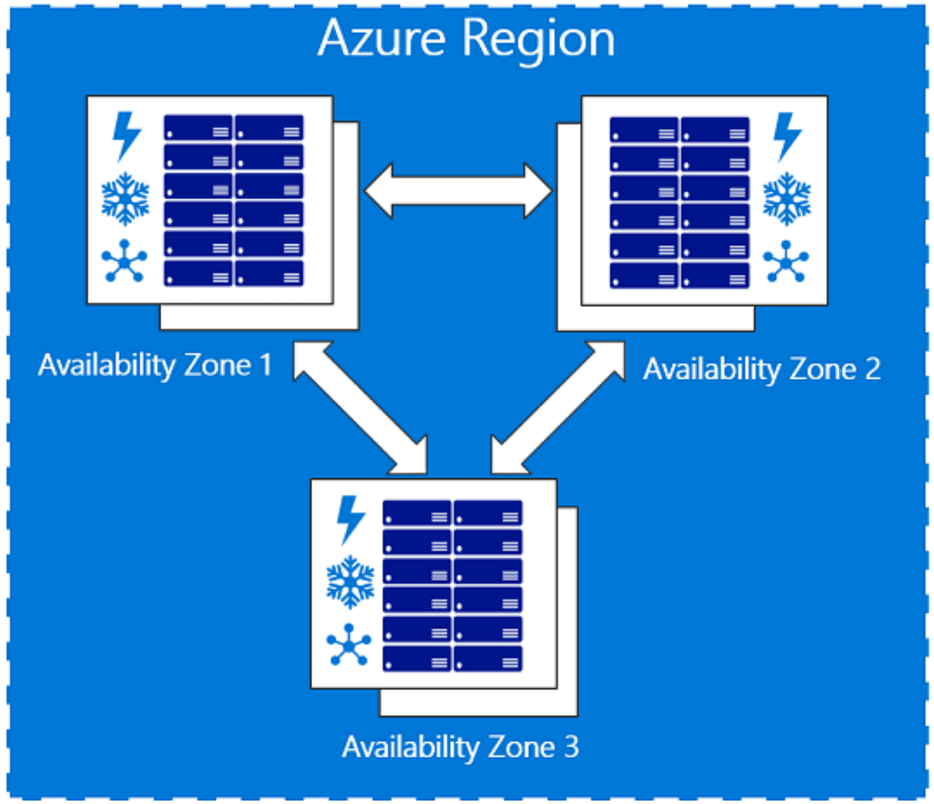


Real-Time Data Processing

- **Azure Stream Analytics**:
  - Definition: Real-time data processing service
  - Components:
    - Input sources
    - Output sinks
    - Stream processing capabilities
  - Features: SQL-based query language, integrations with Event Hub.
  - Use Case: IoT data streaming, event processing.

Best Practices for Cloud Big Data

- Optimize data storage costs.
- Automate resource scaling.
- Ensure high availability and fault tolerance.



Demo:

Azure for Data Engineering

Q&A and Discussion

MEET OUR TEAM

Omar AlSaghier

Sr. Data Engineer

DATATECH LABS.

THANK YOU

OUR CONTACT

DataTechLabs

datatechlabs.ai@gmail.com

datatechlabs.ai

Amman, Jordan