Cloud Computing Essentials with Azure and AWS

# **4.1 Core AWS Services for Data Analytics**

## Storage - S3

* **Purpose**: Stores unstructured and structured data at scale.
* **Key Features**:
  + **Scalability**: Supports petabytes of data.
  + **Access Tiers**: Standard (frequent access), Infrequent Access, and Glacier (archival).
  + **Integration**: Works with AWS services like Athena, Redshift, and Glue.
* **Use Cases**:
  + Data lakes for analytics.
  + Backup and restore.
  + Staging area for big data processing.
* **Azure Equivalent**: Azure Blob Storage.

## Setting Up and Using RDS

* **Purpose**: Managed relational database service.
* **Key Features**:
  + Supports multiple engines: MySQL, PostgreSQL, SQL Server, MariaDB, and Oracle.
  + Automated backups, patching, and scaling.
  + High availability with Multi-AZ deployments.
* **Use Cases**:
  + Hosting transactional applications.
  + Analytics on structured data.
* **Azure Equivalent**: Azure SQL Database.

## Amazon Athena

* **Purpose**: Serverless query service for analyzing data in S3 using SQL.
* **Key Features**:
  + Pay-per-query pricing model.
  + Supports structured and semi-structured data (e.g., CSV, JSON, Parquet).
  + No infrastructure setup required.
* **Use Cases**:
  + Querying data in data lakes.
  + Ad-hoc SQL analysis.
* **Azure Equivalent**: Azure Synapse Analytics (Serverless SQL Pool).

## AWS Glue - DataBrew

* **Purpose**: Data preparation tool for cleaning and normalizing data.
* **Key Features**:
  + Visual, no-code interface for transformation.
  + Automated profiling for identifying data anomalies.
  + Supports data export to analytics tools.
* **Use Cases**:
  + Cleaning messy data for analytics.
  + Building pipelines for machine learning models.
* **Azure Equivalent**: Azure Data Factory (Mapping Data Flows).

## AWS Glue - Crawlers

* **Purpose**: Automatically discover and catalog data in S3, Redshift, and other sources.
* **Key Features**:
  + Builds a metadata catalog for querying in Athena or Redshift.
  + Handles schema inference for diverse data types.
* **Use Cases**:
  + Automating schema discovery in data lakes.
  + Integrating with ETL pipelines.
* **Azure Equivalent**: Azure Data Factory (Data Flow Metadata).

## AWS Glue - ETL

* **Purpose**: Extract, Transform, Load (ETL) service for preparing and moving data.
* **Key Features**:
  + Serverless, scalable data transformation.
  + Supports Python and Scala for custom ETL scripts.
  + Integrates with S3, Redshift, and DynamoDB.
* **Use Cases**:
  + Data migration and transformation.
  + Populating data lakes or warehouses.
* **Azure Equivalent**: Azure Data Factory.

## Redshift

* **Purpose**: Cloud-based data warehouse for large-scale analytics.
* **Key Features**:
  + Massively Parallel Processing (MPP) for high-performance queries.
  + Columnar storage for faster analytics.
  + Integrates with BI tools like Tableau and Power BI.
* **Use Cases**:
  + Running complex queries on structured data.
  + Building dashboards for business insights.
* **Azure Equivalent**: Azure Synapse Analytics (Dedicated SQL Pool).

## QuickSight

* **Purpose**: BI tool for creating interactive dashboards and reports.
* **Key Features**:
  + Machine learning-powered insights.
  + Integrates with AWS services (e.g., Redshift, Athena).
  + Pay-per-session pricing model.
* **Use Cases**:
  + Visualizing data from S3 or Redshift.
  + Building real-time dashboards.
* **Azure Equivalent**: Power BI.

## IAM (Identity and Access Management)

* **Purpose**: Securely manage access to AWS resources.
* **Key Features**:
  + Granular permissions using policies.
  + Multi-factor authentication for added security.
  + Supports roles for cross-service and cross-account access.
* **Use Cases**:
  + Controlling access to data analytics services.
  + Enforcing security compliance.
* **Azure Equivalent**: Azure Active Directory (Azure AD).