**Day 14 Notes**

**Name : Podutur Lahari - DE126**

**Date : 27-11-2024**

#### **Overview:**

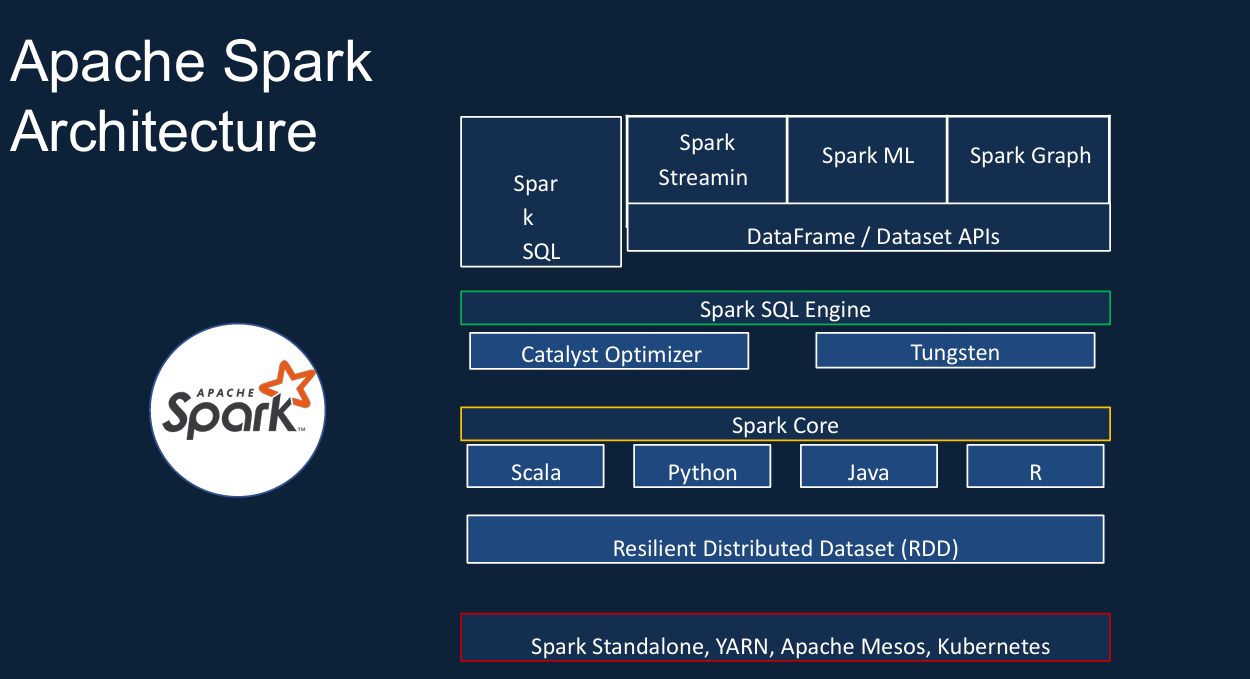
* Azure Databricks integrates with Microsoft Azure for a collaborative and scalable data platform.
* Built on Apache Spark, providing a unified analytics engine for big data and machine learning.
* Key features include:
  + SQL, Streaming, Machine Learning (ML), and Graph Processing.
  + Open-source with simple APIs (Scala, Python, Java, R).

#### **Core Components:**

* Apache Spark Architecture:
  + Key modules: Spark Core, SQL, Streaming, ML, and Graph.
  + Features: In-memory processing, RDD, Catalyst Optimizer, Tungsten execution engine.
* Databricks Features:
  + Optimized Spark (5x faster).
  + Integration with Delta Lake for reliability and performance.
  + Supports MLFlow for machine learning experiments.
  + Seamlessly integrates with Azure services like Power BI, Azure ML, and Azure Data Factory.

#### **Azure Databricks Architecture:**

* Control Plane:
  + Manages Databricks UX and Azure Resource Manager.
* Data Plane:
  + Contains customer data and virtual networks (VNet), Virtual Machines (VMs), and clusters.



#### **Clusters:**

1. Types:
   * Job Cluster: Automated workloads; terminated post-job.
   * All-Purpose Cluster: Persistent for interactive workloads; shared among users.
2. Configuration:
   * Single/Multi-node.
   * Auto-termination and auto-scaling options.
   * Variety of VM types (Compute, Memory, Storage optimized, and GPU-accelerated).
3. Cluster Pools:
   * Reuse idle instances to minimize start-up costs and improve efficiency.

#### 

#### **Pricing:**

* Pricing factors include:
  + Workload type (SQL, Photon, ML, etc.).
  + Tier (Standard vs. Premium).
  + VM configuration.
* Databricks Units (DBU): Measure of processing power influencing cost.
* Example cost: $0.76/hour for a single-node cluster in Premium Tier.

#### **Best Practices for Cost Control:**

* Use auto-termination to shut clusters after inactivity.
* Set budget alerts and monitor usage via Azure subscriptions.
* Optimize workloads by choosing appropriate cluster types and VM sizes.

#### **Tools and Integrations:**

* Integrates with Azure services like Azure Data Lake, Blob Storage, Event Hub, Cosmos DB, Synapse Analytics.
* Supports notebooks, SQL analytics, and administrative controls for collaborative environments**.**

### 

### **1. Database**

* Purpose: Stores current data required for powering applications; supports Online Transaction Processing (OLTP).
* Characteristics:
  + Can be relational (fixed schema) or non-relational (flexible schema).
  + Stores structured and semi-structured data.
  + Provides features like security, indexing, and query languages.
* Use Cases:
  + Applications requiring real-time data operations (e.g., patient records, e-commerce inventories).
  + Examples: MySQL, MongoDB, Redis.

### **2. Data Warehouse**

* Purpose: Combines and analyzes structured, historical, and current data from multiple sources to support Online Analytical Processing (OLAP).
* Characteristics:
  + Structured data in a pre-defined schema (schema-on-write).
  + Data updated periodically via ETL processes.
  + Supports business intelligence (BI) tools for reporting and dashboards.
* Use Cases:
  + Long-term analysis and BI reporting.
  + Examples: Amazon Redshift, Snowflake, Google BigQuery.

### **3. Data Lake**

* Purpose: Stores raw, unstructured, and structured data, often used for future analytics and machine learning.
* Characteristics:
  + Flexible data storage with no predefined schema (schema-on-read).
  + Cost-effective for large-scale data storage.
  + Supports multiple data types (e.g., JSON, CSV, images).
* Use Cases:
  + Storing diverse data for machine learning and predictive analytics.
  + Examples: AWS S3, Azure Data Lake, Google Cloud Storage.

### 

### **4. Delta Lake**

* Enhancement of Data Lake: Introduces ACID transactions, schema enforcement, and versioning to data lakes.
* Advantages: Supports structured and unstructured data with higher reliability for analytical workloads.