G. Rupa Manasa 28-12-2023

Day 21 Assessment

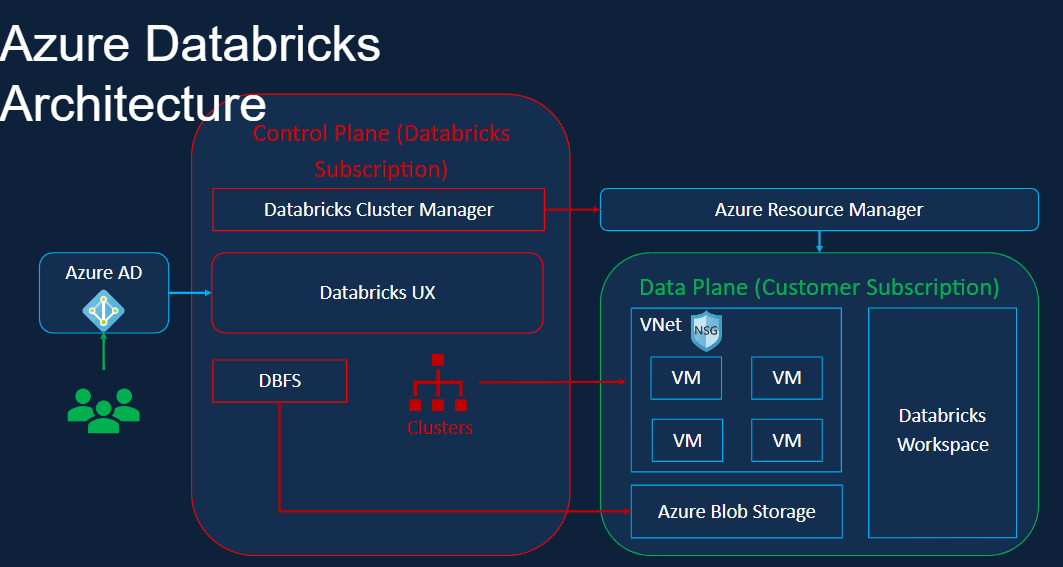
**Introduction to Databricks:**

Azure Databricks is a cloud-based big data analytics platform provided by Microsoft Azure in collaboration with Databricks.

It combines the power of Apache Spark with a set of Azure services to offer a unified analytics platform.

Allows organizations to process large volumes of data, build machine learning models, and derive insights from diverse data sources.

**Azure Databricks Architecture:**

****

**Azure Active Directory (AAD):**

Azure Databricks integrates with Azure Active Directory for authentication and authorization.

Users can access Databricks using their Azure AD credentials, and role-based access control (RBAC) is used to manage permissions.

**Cluster Management:**

Clusters in Databricks are managed through the control plane.

This includes the ability to create, configure, and terminate clusters that are used for processing large-scale data and running Apache Spark jobs.

**Databricks UX:**

The user experience (UX) in Databricks is designed to provide a collaborative and streamlined environment for data engineers, data scientists, and analysts to work on big data analytics and machine learning tasks.

**DBFS:**

Databricks File System (DBFS) is a distributed file system that comes integrated with the Databricks Unified Analytics Platform.

It provides a layer of abstraction over various data storage systems, allowing users to interact with and manage their data in a simplified and unified manner.

**Clusters:**

In the context of Databricks, a cluster refers to a set of virtual machines (VMs) that work together to process data and run computations.

These clusters are used to execute Apache Spark jobs and provide the computational power needed for big data analytics and machine learning tasks.

Users can create and configure clusters based on their specific needs.

Configuration options include selecting the Spark version, specifying the number and type of VMs (driver and worker nodes), and setting cluster policies.

**Databricks Workspace:**

Databricks Workspace provides a unified platform where users can perform data engineering, data science, and machine learning tasks.

It integrates various tools and features into a single environment.

Notebooks are a central feature of the workspace, providing an interactive and collaborative environment for writing and executing code.

Users can create notebooks in multiple languages, including Scala, Python, SQL, R.

Databricks Workspace supports real-time collaboration, allowing multiple users to work on the same notebook simultaneously.

Changes made by one user are reflected instantly for others, facilitating teamwork.

Jobs in Databricks Workspace allow users to schedule and automate the execution of notebooks or scripts.

This is useful for running regular ETL (Extract, Transform, Load) processes or other recurring tasks.

**Azure Blob Storage:**

It is designed to store and manage large amounts of unstructured data, such as documents, images, videos, and log files.

**Databricks Workspace Components:**

Notebooks

Jobs

Cluster

Data

Models

**Cluster Configuration:**

Single or multi node.

Access Mode: Single user, Shared, no isolation shared, Custom.

Databricks Runtime: Spark, scala, java, ubuntu, Python, R.

Auto Termination: Terminates the cluster after X minutes of inactivity. Default value for Single Node and Standard clusters is 120 minutes. Users can specify a value between 10 and 10000 mins as the duration.

Auto Scaling: User specifies the min and max work nodes. Auto scales between min and max based on the workload. Not recommended for streaming workloads