ASSESSMENT

NAME-SWARNA LATHA GUVVALA

DATE-28/12/2023

INTRODUCTION TO AZURE DATABRICKS:

AZURE DATABRICKS:

Azure Databricks is a cloud-based big data analytics platform provided by Microsoft Azure, in collaboration with Databricks. It integrates with Azure services to enable streamlined data processing, analytics, and machine learning workflows. Here are some key aspects of Azure Databricks:

1. **Unified Analytics Platform:** Azure Databricks provides a unified platform for big data and machine learning. It combines Apache Spark with Databricks' collaborative environment to facilitate data engineering, data science, and business analytics.
2. **Apache Spark:** Apache Spark is an open-source, distributed computing system that enables processing large-scale data sets quickly. Azure Databricks leverages Spark to provide a fast and general-purpose cluster-computing framework.
3. **Collaborative Environment:** Azure Databricks offers a collaborative workspace where data engineers, data scientists, and analysts can work together. It includes features like notebooks for code sharing, collaboration, and interactive data exploration.
4. **Integration with Azure Services:** Azure Databricks seamlessly integrates with other Azure services, such as Azure Storage, Azure SQL Data Warehouse, Azure Data Lake Storage, and Azure Active Directory. This integration allows users to easily move and process data across different Azure services.
5. **Data Import and Export:** You can import and export data from various sources and formats using Azure Databricks. It supports popular data formats like Parquet, Delta Lake, JSON, and more.

6. **Scalability:** Azure Databricks provides scalable clusters to handle large datasets and complex computations. Users can easily scale up or down based on their processing needs.
7. **Machine Learning:** The platform includes tools for building, training, and deploying machine learning models. It supports popular machine learning libraries like TensorFlow, PyTorch, and scikit-learn.
8. **Security and Compliance:** Azure Databricks includes features for securing data and complying with regulatory requirements. It integrates with Azure Active Directory for identity and access management.
9. **Job Scheduling and Automation:** Users can schedule and automate data processing jobs using Databricks Jobs. This allows for the execution of workflows at specified intervals or based on triggers.
10. **Monitoring and Logging:** Azure Databricks provides monitoring and logging features to track the performance of clusters, jobs, and workflows. This helps in troubleshooting and optimizing data processing tasks.

To get started with Azure Databricks, you can use the Azure portal to create a Databricks workspace and configure the necessary settings. From there, you can start using the collaborative environment to develop and run Spark-based applications and analytics.

APACHE SPARK:

Apache Spark is an open-source, distributed computing system that provides a fast and general-purpose cluster-computing framework for big data processing. It was originally developed at the University of California, Berkeley's AMPLab, and later donated to the Apache Software Foundation. Spark is designed to be fast, flexible, and easy to use, and it supports various programming languages such as Scala, Java, Python, and R.

Key features and components of Apache Spark include:

1. **Resilient Distributed Datasets (RDD):** RDD is the fundamental data structure in Spark, representing an immutable distributed collection of objects that can be processed in parallel. RDDs support fault tolerance through lineage information, allowing data to be reconstructed in case of node failures.
2. **Spark Core:** The core engine of Spark provides the basic functionality for distributed data processing. It includes task scheduling, memory management, fault recovery, and interacting with storage systems.

3. **Spark SQL:** Spark SQL allows querying structured data using SQL queries. It provides a DataFrame API for manipulating structured and semi-structured data and supports various data sources such as Parquet, Avro, JSON, and relational databases.
4. **Spark Streaming:** Spark Streaming enables the processing of real-time data streams. It breaks the continuous data stream into small, batch-sized intervals, allowing Spark to process and analyze the data in a stream-like fashion.
5. **MLlib (Machine Learning Library):** MLlib is Spark's machine learning library, offering a set of algorithms and tools for machine learning tasks such as classification, regression, clustering, and collaborative filtering.
6. **GraphX:** GraphX is a graph processing library for Spark, providing a distributed framework for graph computation and graph-parallel algorithms. It enables the analysis of large-scale graphs and networks.
7. **SparkR:** SparkR is an R package that allows R users to interact with Spark. It provides an R frontend to Spark, enabling data scientists and statisticians to leverage Spark's capabilities from within the R environment.
8. **Cluster Manager Integration:** Spark can run on various cluster managers, including Apache Mesos, Hadoop YARN, and its standalone built-in cluster manager. This flexibility makes it easy to integrate Spark with existing cluster management systems.
9. **Adaptive Query Execution:** Spark includes adaptive query execution to optimize query plans based on runtime statistics, improving the performance of Spark SQL queries.
10. **Community and Ecosystem:** Apache Spark has a large and active open-source community. It also has a rich ecosystem of libraries and tools built around it, making it a versatile choice for big data processing and analytics.

Spark is commonly used for large-scale data processing tasks, including ETL (Extract, Transform, Load) operations, data analytics, machine learning, and graph processing. Its ability to perform in-memory data processing and support for multiple programming languages contribute to its popularity in the big data landscape.

DATABRICKS:

Databricks is a cloud-based platform for big data analytics and machine learning. It was founded by the creators of Apache Spark and provides a collaborative environment that simplifies the process of building, managing,

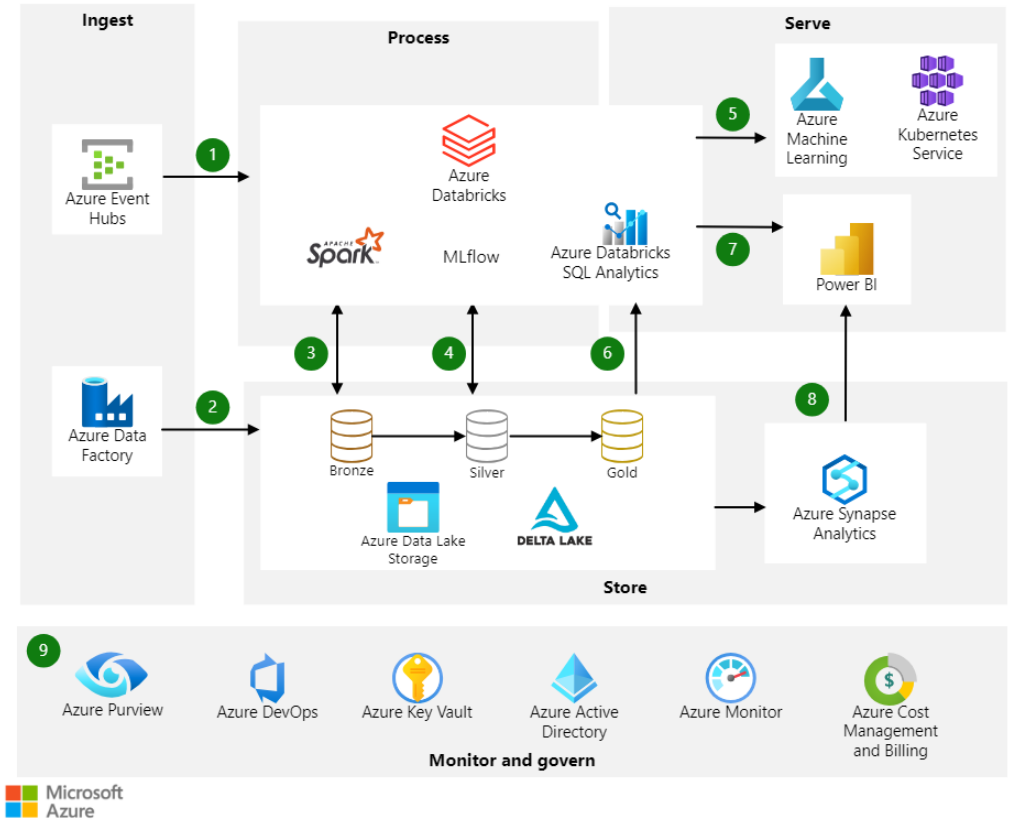
and deploying Spark-based applications. Databricks is designed to work seamlessly with Apache Spark and extends its capabilities by offering additional features and services. Here are key aspects of Databricks:

1. **Unified Analytics Platform:** Databricks provides a unified platform for data engineering, data science, and business analytics. It integrates Apache Spark with a collaborative workspace, allowing users from different roles to work together in a shared environment.
2. **Apache Spark Integration:** Databricks is tightly integrated with Apache Spark, a fast and general-purpose cluster-computing framework. It allows users to leverage the power of Spark without dealing with the complexities of cluster management.
3. **Collaborative Workspace:** The platform includes a collaborative workspace that supports interactive data exploration, collaborative coding, and the creation of notebooks. Notebooks can contain a mix of code, visualizations, and narrative text, making it easy to share and communicate findings.
4. **Automated Cluster Management:** Databricks automates cluster management, making it easier to provision, configure, and scale Spark clusters. Users can dynamically adjust the size of clusters based on workload requirements.
5. **Data Import and Export:** Databricks supports seamless integration with various data sources and formats. It can read and write data from and to popular data storage solutions, including Azure Data Lake Storage, AWS S3, and Delta Lake.
6. **Job Scheduling:** Users can schedule and automate the execution of jobs using Databricks Jobs. This allows for the automation of data processing workflows, machine learning model training, and other tasks.
7. **Libraries and APIs:** Databricks supports a wide range of libraries and APIs, making it easy to extend its functionality. This includes support for machine learning libraries such as MLlib, TensorFlow, and scikit-learn.
8. **Machine Learning Workflows:** Databricks provides tools for building, training, and deploying machine learning models. It supports collaborative model development and includes features for model tracking and experimentation.
9. **Security and Compliance:** The platform includes features for securing data and ensuring compliance with regulatory requirements. It integrates with identity providers such as Azure Active Directory for authentication and access control.

10. **Integration with Cloud Providers:** Databricks is available on major cloud platforms, including Microsoft Azure, Amazon Web Services (AWS), and Google Cloud Platform (GCP). It can take advantage of cloud-based storage and other services.

Databricks is commonly used for a variety of big data and analytics tasks, including data preparation, exploratory data analysis, machine learning, and real-time analytics. Its collaborative and integrated nature makes it a popular choice for teams working on complex data projects in the cloud.

ARCHITECTURE OF AZURE DATABRICKS:



DATAFLOW:

1. Azure Databricks ingests raw streaming data from Azure Event Hubs.
2. Data Factory loads raw batch data into Data Lake Storage Gen2.
3. For data storage:

- Data Lake Storage Gen2 houses data of all types, such as structured, unstructured, and semi-structured. It also stores batch and streaming data.
- Delta Lake forms the curated layer of the data lake. It stores the refined data in an open-source format.
- Azure Databricks works well with a medallion architecture that organizes data into layers:
    - Bronze: Holds raw data.
    - Silver: Contains cleaned, filtered data.
    - Gold: Stores aggregated data that's useful for business analytics.

4. The analytical platform ingests data from the disparate batch and streaming sources. Data scientists use this data for these tasks:
    - Data preparation.
    - Data exploration.
    - Model preparation.
    - Model training.

    MLflow manages parameter, metric, and model tracking in data science code runs. The coding possibilities are flexible:

    - Code can be in SQL, Python, R, and Scala.
    - Code can use popular open-source libraries and frameworks such as Koalas, Pandas, and scikit-learn, which are pre-installed and optimized.
    - Practitioners can optimize for performance and cost with single-node and multi-node compute options.
5. Machine learning models are available in several formats:
    - Azure Databricks stores information about models in the MLflow Model Registry. The registry makes models available through batch, streaming, and REST APIs.
    - The solution can also deploy models to Azure Machine Learning web services or Azure Kubernetes Service (AKS).
6. Services that work with the data connect to a single underlying data source to ensure consistency. For instance, users can run SQL queries on the data lake with Azure Databricks SQL Analytics. This service:
    - Provides a query editor and catalog, the query history, basic dashboarding, and alerting.

- Uses integrated security that includes row-level and column-level permissions.
- Uses a Photon-powered Delta Engine to accelerate performance.

7. Power BI generates analytical and historical reports and dashboards from the unified data platform. This service uses these features when working with Azure Databricks:
    - A built-in Azure Databricks connector for visualizing the underlying data.
    - Optimized Java Database Connectivity (JDBC) and Open Database Connectivity (ODBC) drivers.
8. Users can export gold data sets out of the data lake into Azure Synapse via the optimized Synapse connector. SQL pools in Azure Synapse provide a data warehousing and compute environment.
9. The solution uses Azure services for collaboration, performance, reliability, governance, and security:
    - Microsoft Purview provides data discovery services, sensitive data classification, and governance insights across the data estate.
    - Azure DevOps offers continuous integration and continuous deployment (CI/CD) and other integrated version control features.
    - Azure Key Vault securely manages secrets, keys, and certificates.
    - Microsoft Entra ID provides single sign-on (SSO) for Azure Databricks users. Azure Databricks supports automated user provisioning with Microsoft Entra ID for these tasks:
        - Creating new users.
        - Assigning each user an access level.
        - Removing users and denying them access.
    - Azure Monitor collects and analyzes Azure resource telemetry. By proactively identifying problems, this service maximizes performance and reliability.
    - Azure Cost Management and Billing provide financial governance services for Azure workloads.