**Assignment**

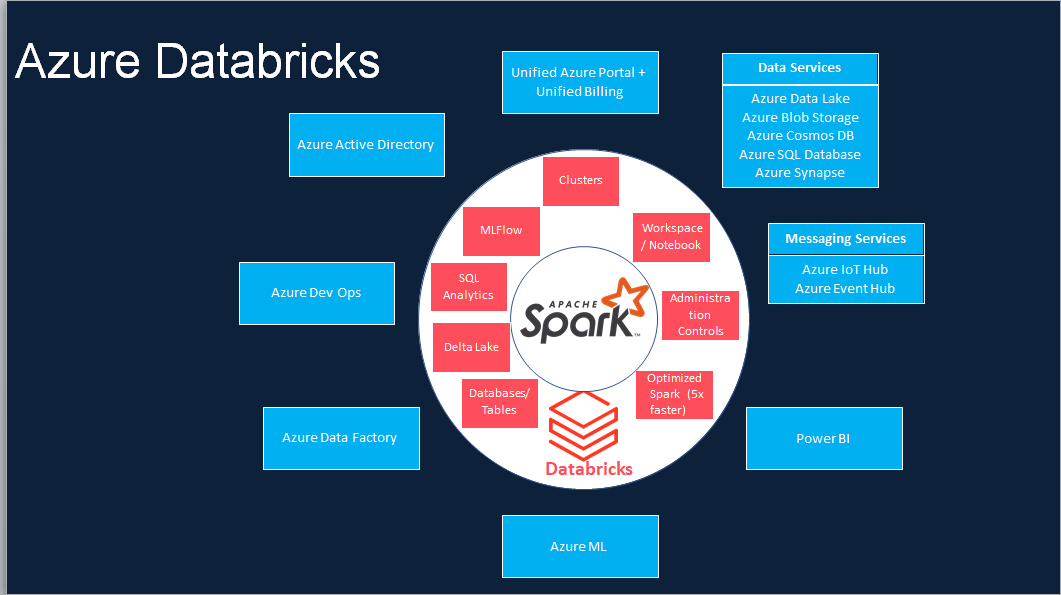
**Date : 28-12-23**

**Azure DataBricks**

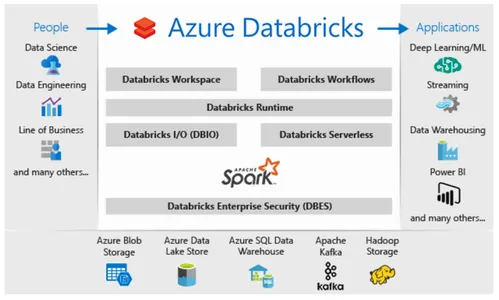
\* Databricks is a cloud-based platform that provides a unified analytics and data engineering workspace.

\* It is built on top of Apache Spark, a powerful open-source, distributed computing System.





\*Databricks simplifies big data processing and enables collaboration between data scientists, data engineers, and business analysts in a scalable and efficient manner.



**some key aspects and features of Databricks:**

* **Scalability and Reliability**: It is built on top of Apache Spark, which is a fast and scalable distributed computing framework.
* **Integration with Azure Services**: It is tightly integrated with the Microsoft Azure cloud, which means that users can easily integrate it with other Azure services such as Azure Blob Storage, Azure Data Lake Storage, and Azure SQL Database.
* **Automation**: It offers automated cluster provisioning, auto-scaling, and job scheduling features.
* **Unified Environment**:This means that teams can work collaboratively and seamlessly across different tasks and projects.
* **Scalable Analytics:** It is built on Apache Spark, which is a distributed computing framework that can process large amounts of data in parallel.
* **Machine Learning**: Azure Databricks provides various tools and frameworks for building, training, and deploying machine learning models.
* **Integrations**: This allows teams to easily build end-to-end data pipelines to ingest, process, and analyze data in real time.

**Azure Databricks Components**

The platform consists of various components that work together to facilitate data processing,analysis, and machine learning. Here are key components of Azure Databricks:

**1.Workspace:**

\* The Databricks Workspace is the central collaborative environment where users can create and manage notebooks, clusters, libraries, and dashboards.

\* It provides a web-based interface for users to interact with and manage their Databricks resources.

**2. Clusters:**

\* Clusters in Azure Databricks are computational resources that execute the code contained in notebooks.

\* Users can create and configure clusters with different specifications based on their processing and memory requirements.

\* Clusters are automatically managed by Databricks, providing scalability and resource optimization.

**3. Notebooks:**

\* Notebooks are interactive documents that combine live code, visualizations, and narrative text.

\* Users can create and run notebooks for data exploration, analysis, and code development.

\*Notebooks support multiple programming languages, including Python, Scala, SQL,and R.

**4. Jobs:**

\* Jobs in Databricks allow users to schedule and automate the execution of notebooks or JAR files.

\* Users can specify the frequency and parameters for job runs, making it useful for recurring tasks and ETL (Extract, Transform, Load) processes.

**5. Libraries:**

\* Libraries are packages or dependencies that users can attach to clusters to extend functionality.

\* Users can install and manage libraries to access additional tools, APIs, or custom