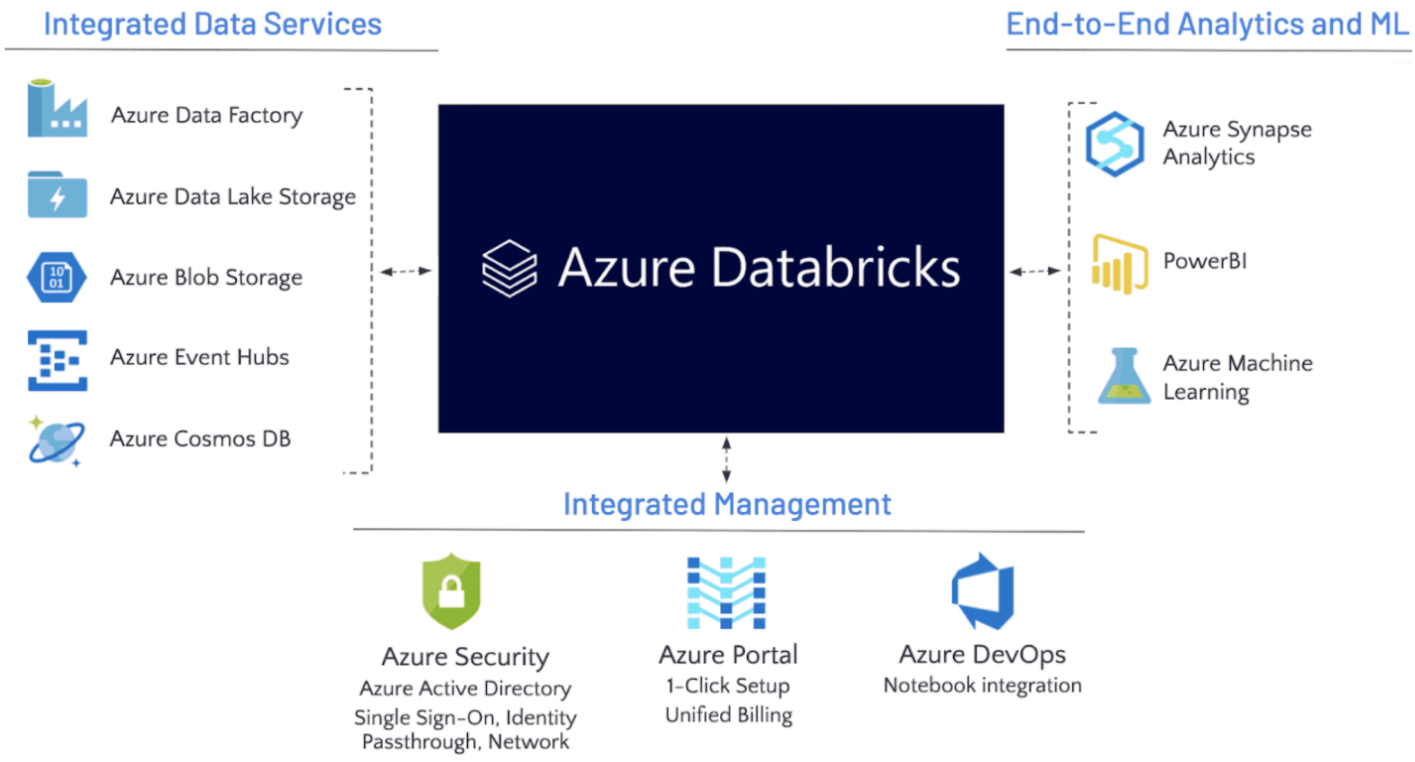
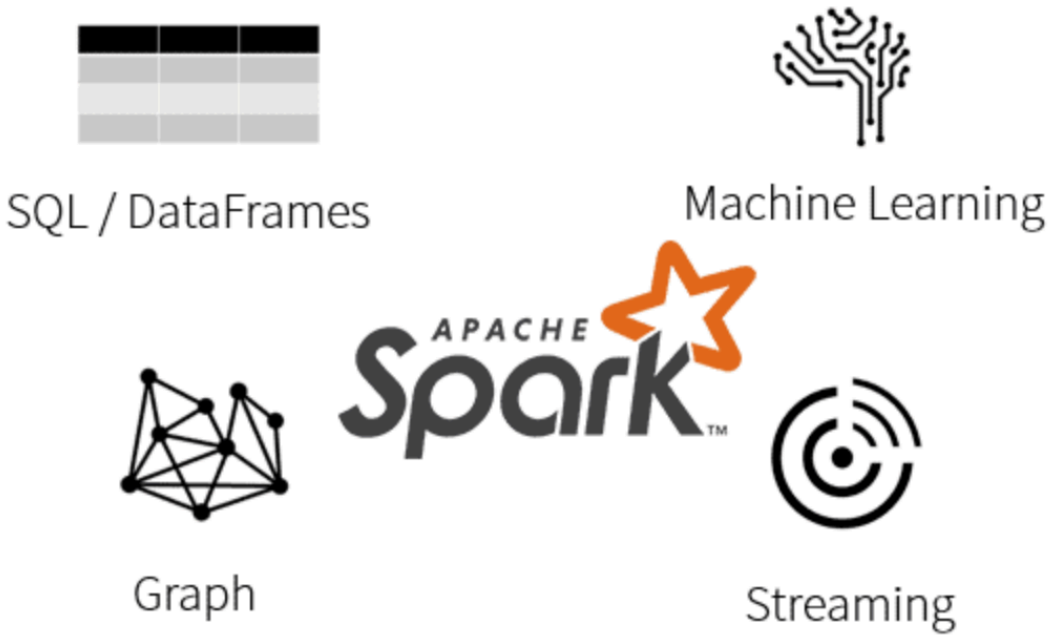
What Is DataBricks?

- **Databricks + Apache Spark + enterprise cloud = Azure Databricks**
- It is a fully-managed version of the open-source Apache Spark data analytics and it features optimized connectors to storage platforms for the quickest possible data access.
- It offers a notebook-oriented Apache Spark as-a-service workspace environment which makes it easy to explore data interactively and manage clusters.
- It is secure cloud-based machine learning and big data platform.
- It is supporting multiple languages such as Scala, Python, R, Java, and SQL.



What is Apache Spark?

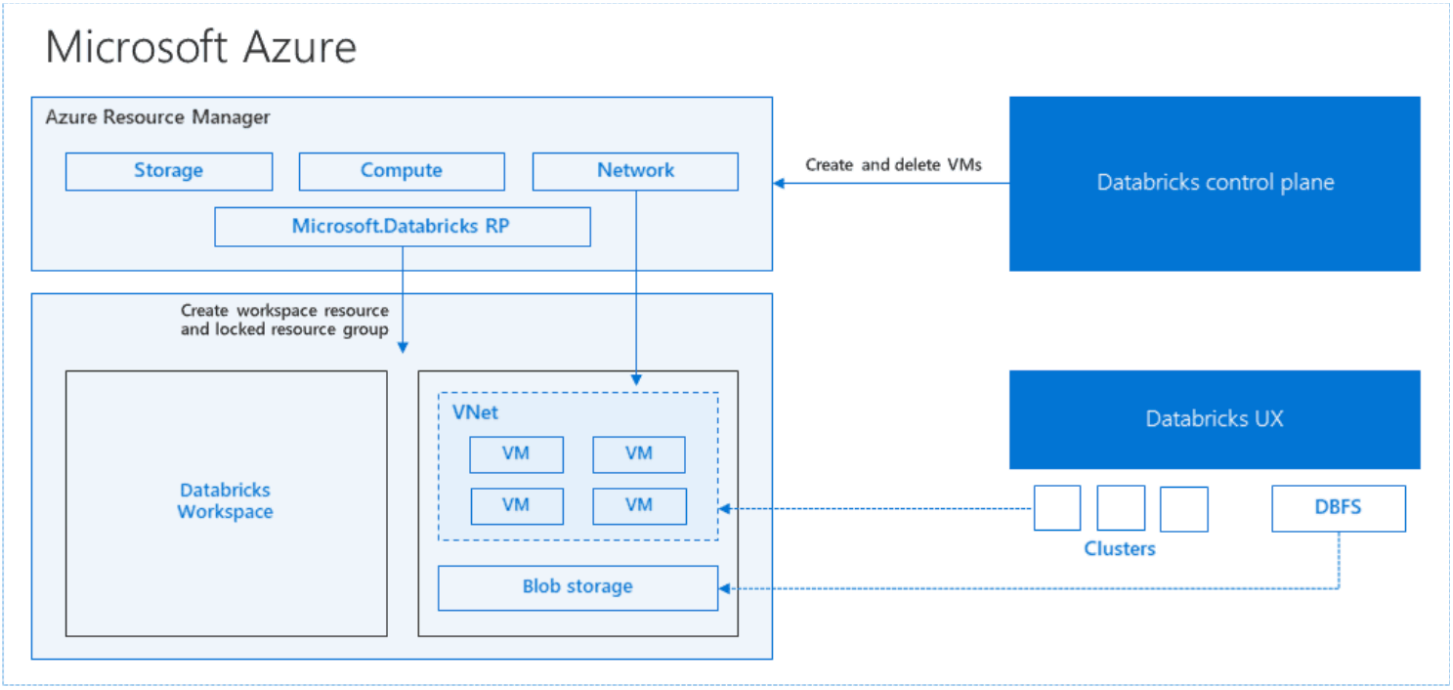
- Spark is an integrated processing engine that can analyze big data using SQL, graph processing, machine learning, or real-time stream analysis.
- Spark ML offers high class and finely tuned machine learning algorithms for handling big data.



Azure Databricks Architecture & Diagram

- When we launch a cluster via Databricks, a “Databricks appliance” is deployed as an Azure resource in our subscription.

- Then we specify the types of VMs to use and how many, but Databricks handle all other elements.
- A managed resource group is deployed into the subscription that we populate with a VNet, a storage account, and a security group.
- Once these services are ready, we will control the Databricks cluster over the Databricks UI.



Why Azure Databricks ?

1) Optimized Environment

2) Persistent collaboration

3) Simple to use

1) Optimized Environment

- Databricks Azure was optimized automatically from the ground up for cost-efficiency and performance in the cloud.
- Auto-scaling and auto-termination of Spark clusters, no doubt it minimizes costs automatically.
- Optimizations including indexing, caching, and advanced query optimization, which can enhance performance by as much as 10-100x over conventional Apache Spark deployments in the cloud.
-

2) Persistent collaboration

- Notebooks on Databricks are live and easy to share, with real-time teamwork.
- Dashboards allow business users to call a current job with new parameters.
- Databricks integrates closely with PowerBI for hand-on visualization.
-

3) Simple to use

- Azure Databricks comes with notebooks that let you run machine learning algorithms, connect to common data sources, and learn the basics of Apache Spark to get started rapidly.

- It also a unified debugging environment features to let you analyze the progress of your Spark jobs from under interactive notebooks, and powerful tools to examine past jobs.
- No need to install common analytics libraries, such as the Python and R data science stacks, which are preinstalled.