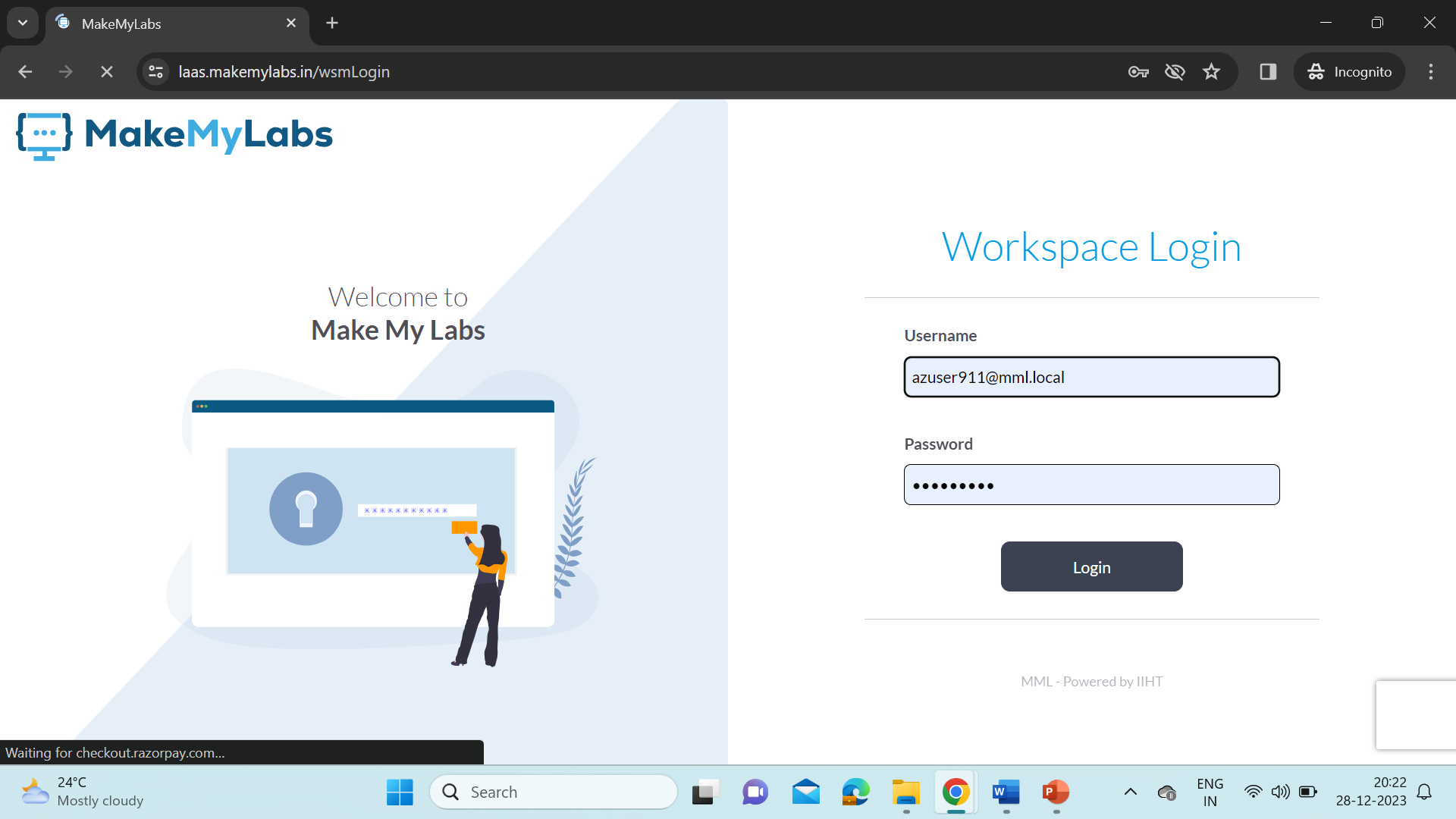
Kavya Gundala

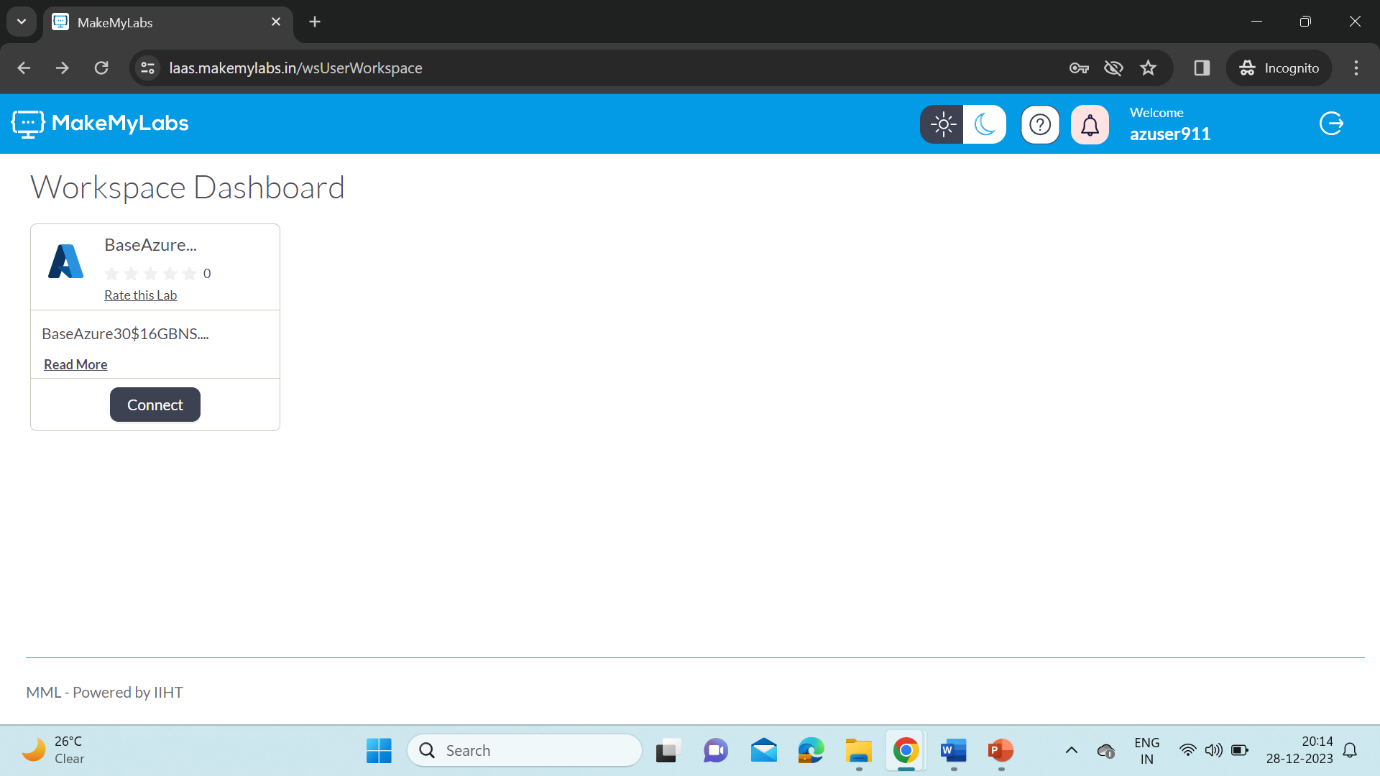
28-12-2023

**Microsoft Azure Portal:**

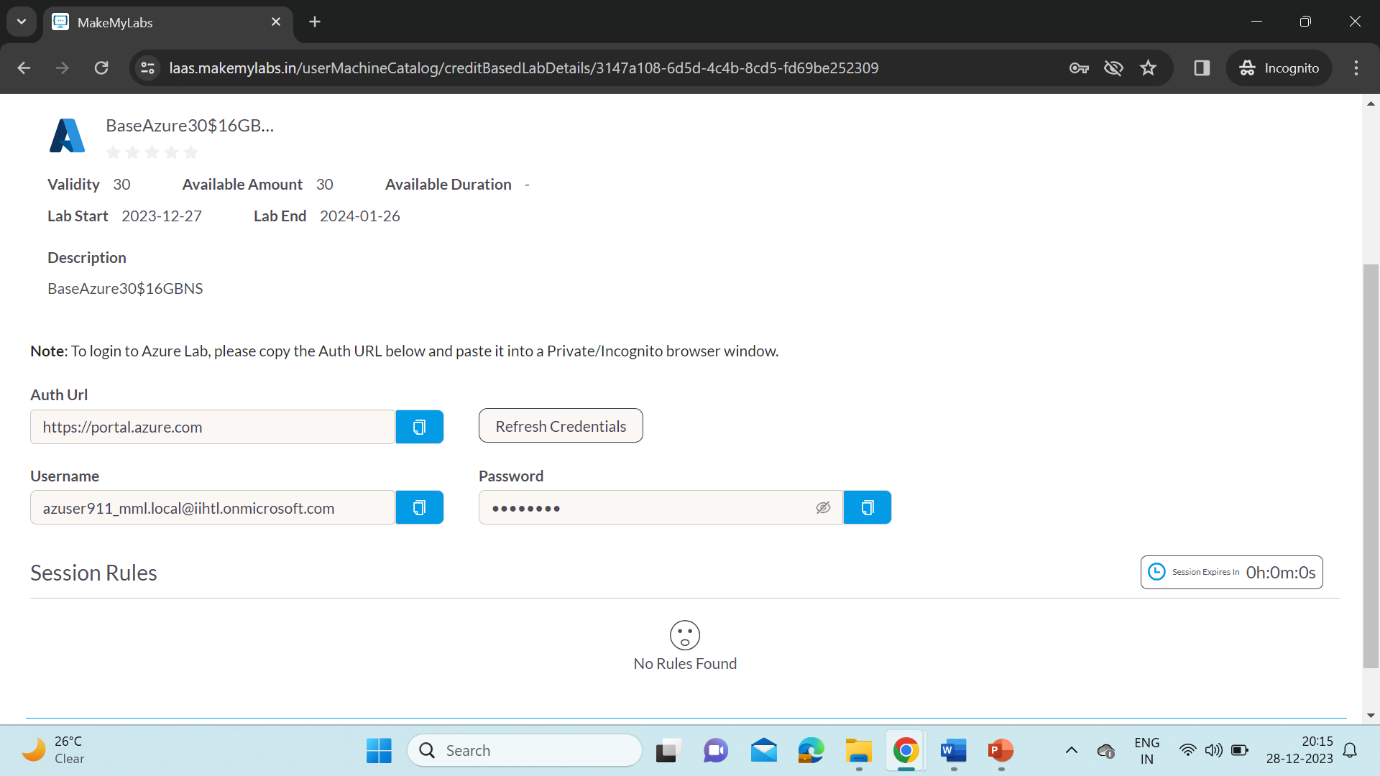
* Logging in MakeMyLabs:



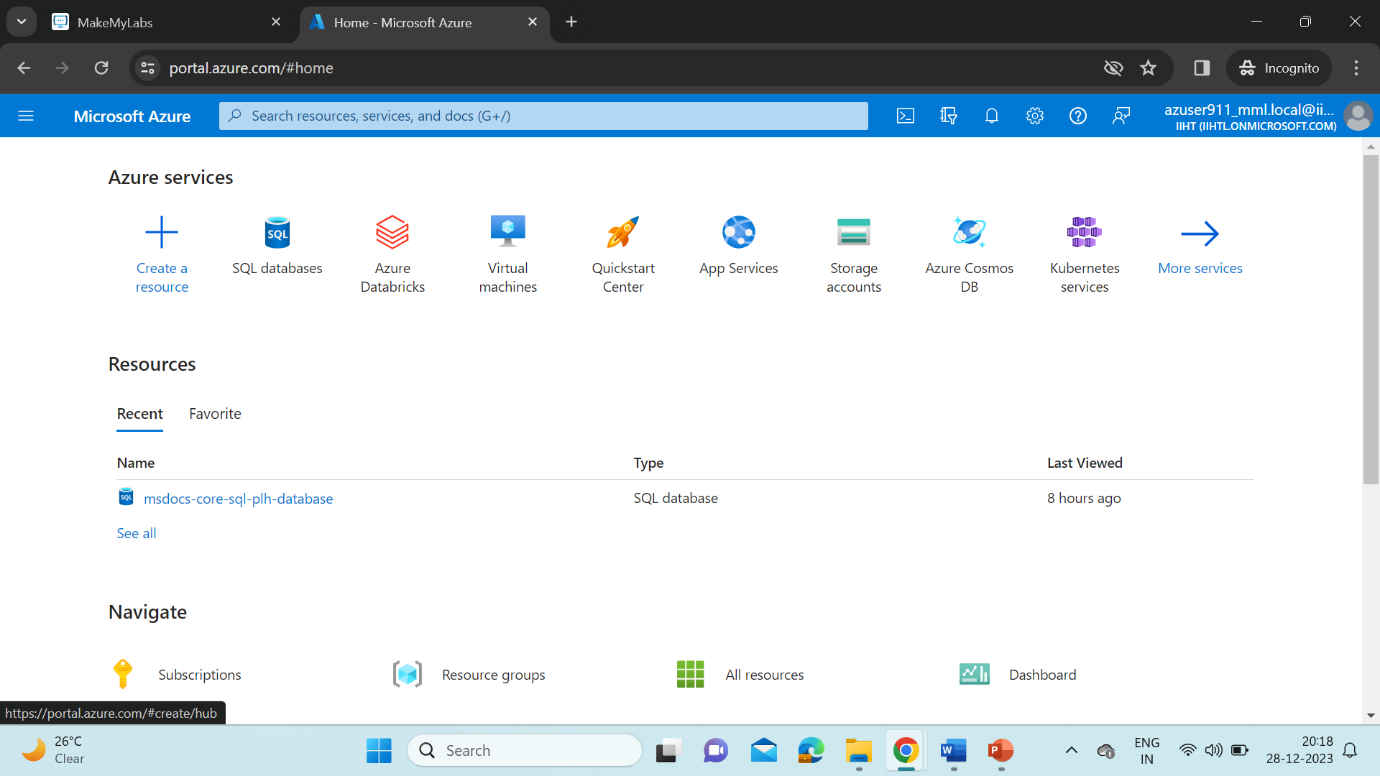
* Connect to the Lab



* Use the Credentials for logging into Azure Portal.

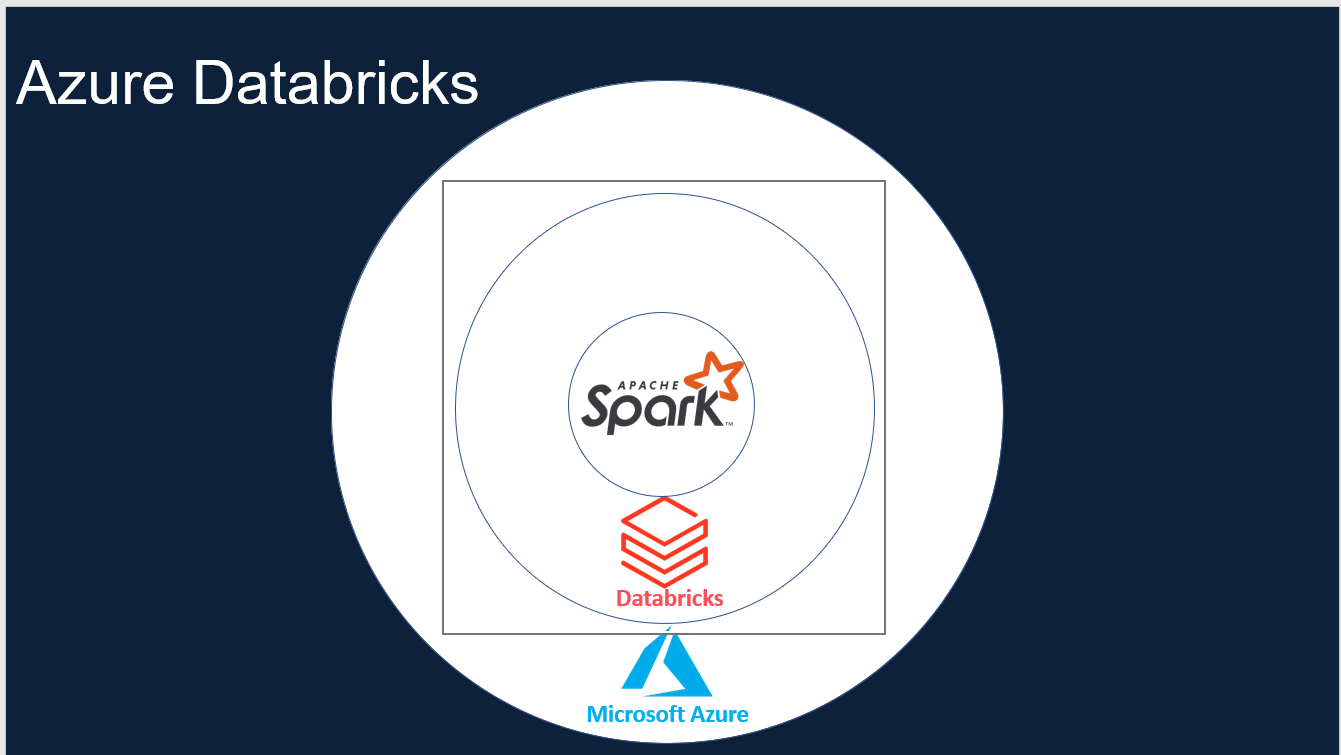


* After logging, we get a lot of services opened.



**Azure DataBricks for Data Engineering:**

* Azure Databricks is a fast, easy, and collaborative Apache Spark-based big data analytics service designed for data science and data engineering.
* Azure Databricks provides tools that help you connect your sources of data to one platform to process, store, share, analyze, model, and monetize datasets with solutions from BI to generative AI.
* Microsoft Azure implements Databricks and Apache Spark.

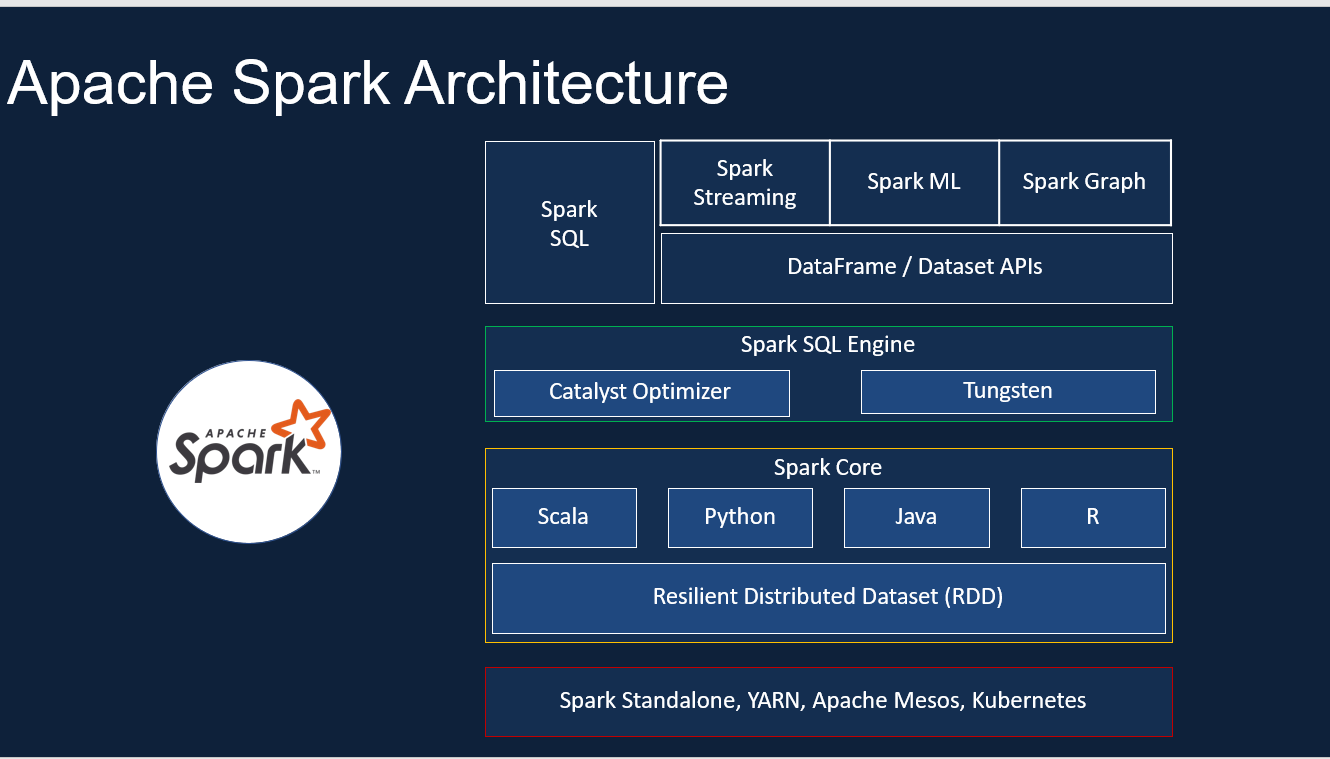


**Apache Spark:**

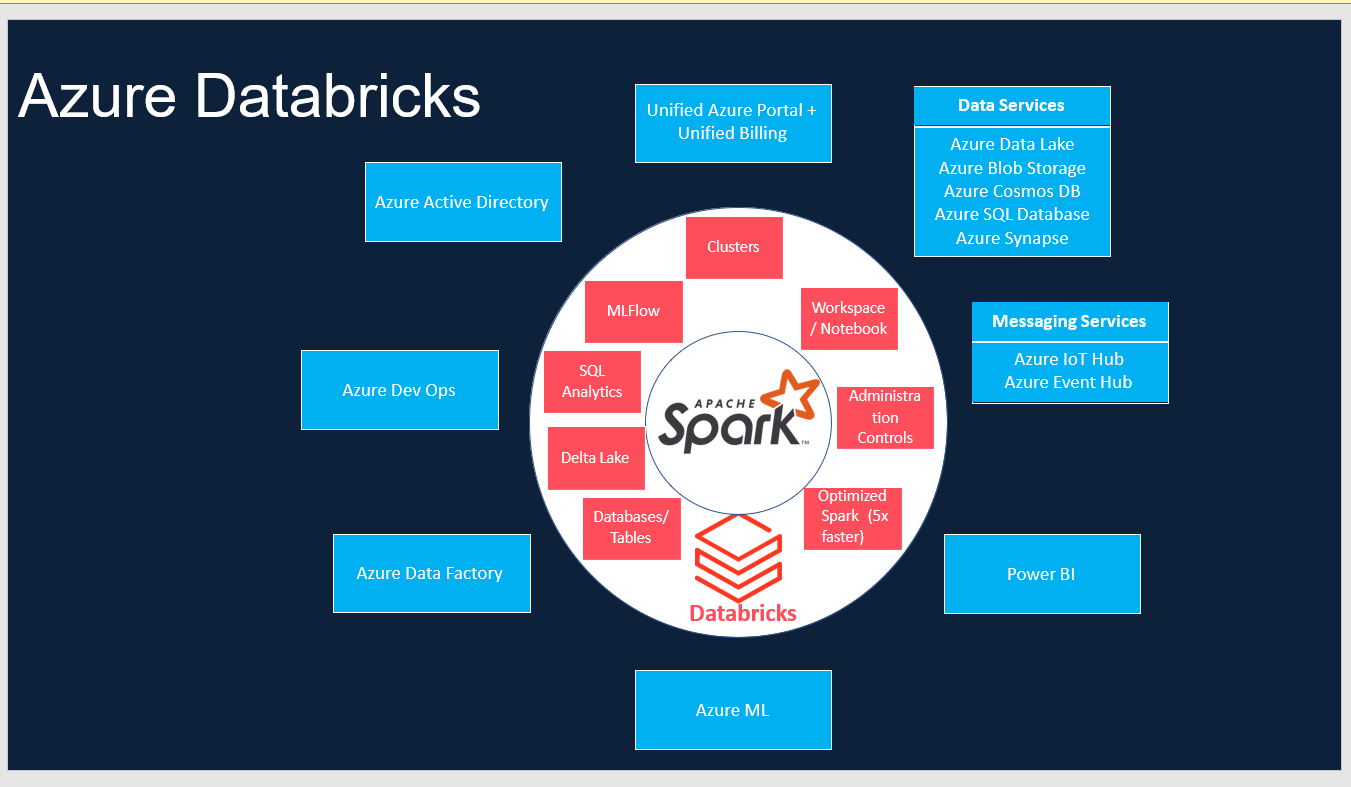
* Apache Spark is a lightning-fast unified analytics engine for big data processing and machine learning.
* Spark provides an interface for programming clusters with implicit data parallelism and fault tolerance.

**Features:**

* 100% Open source under Apache License
* Simple and easy to use APIs
* In-memory processing engine
* Distributed computing Platform
* Unified engine which supports SQL, streaming, ML and graph processing

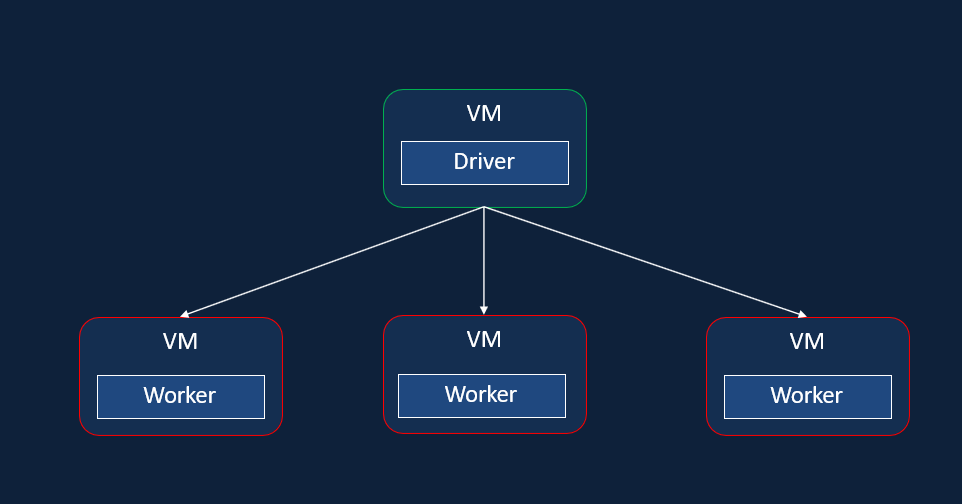


* Spark has clusters, Notebook, Administration on controls, Databases, SQL Analytics, ML Flow etc.



* Azure Databricks can implement with integration of Active Directory, Power BI, DevOps, Data Factory, Data Services, ML, etc.
* Databricks workspace components:
  + Notebook
  + Jobs
  + Clusters
  + Data
  + Models

**DataBricks Cluster:**



* Clusters are two types: All Purpose and Job Cluster
* Cluster configuration can be: Single/Multi node
* Access Mode in Clusters:
  + For single user(only one user access. Supports Python, SQL, Scala, R),
  + For Shared(multiple user access. Only available in Premium. Supports Python, SQL),
  + No Isolation Shared(Multiple User Access Supports Python, SQL),
  + Custom(Legacy Configuration).
* Databricks Runtime
* Auto Termination : The cluster terminates 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. Not recommended for streaming workloads.
* Cluster VM Type/Size: Memory Optimized, Compute Optimized, Storage Optimized, General Purpose, GPU Accelerated.
* Cluster Policy:
  + Simplifies the user interface
  + Enables standard users to create clusters
  + Achieves cost control
  + Only available on premium tier