Microsoft

Module: Introduction to Azure Databricks

Microsoft Services

Module Overview

- Lesson 1: Introduction to Databricks
- Lesson 2: Azure Databricks and Architecture
- Lesson 3: Price Tiers and Workloads
- Lesson 4: Azure Databricks Artifacts
- Lesson 5: Azure Databricks Clusters
- Lesson 6: Azure Databricks Workspace

Lesson 1: Introduction to Databricks

After completing this lesson, you will be able to:

- Understand the features of Apache Spark
- Learn the basic definition of Databricks and its evolution

What is Big Data?

- Data has become a critical asset for any organization.
- Today's digital age is generating data exponentially.
- A popular definition of big data is the data characterized by 3 Vs.

Volume	Variety	Velocity
Terabytes to Petabytes and more	Structured and Unstructured data: eg, cvs, json, image, IoT …	Accelerating rate of data ingestion and analysis

What is Big Data?

- The challenge is how to get the value out of this data.
- Big data technologies are used to ingest, process and analyze large volume of data at fast pace.
- Big data technologies run on distributed architectures, offering high availability at low cost.

What is Big Data Technology?

- Most of big data technologies are open-sources.
- They form an ecosystem of frameworks, applications and specialized tools.

Apache Hadoop

Apache Spark

Apache Kafka

Apache HBase

Apache Hive LLAP

Apache Storm

Apache Spark

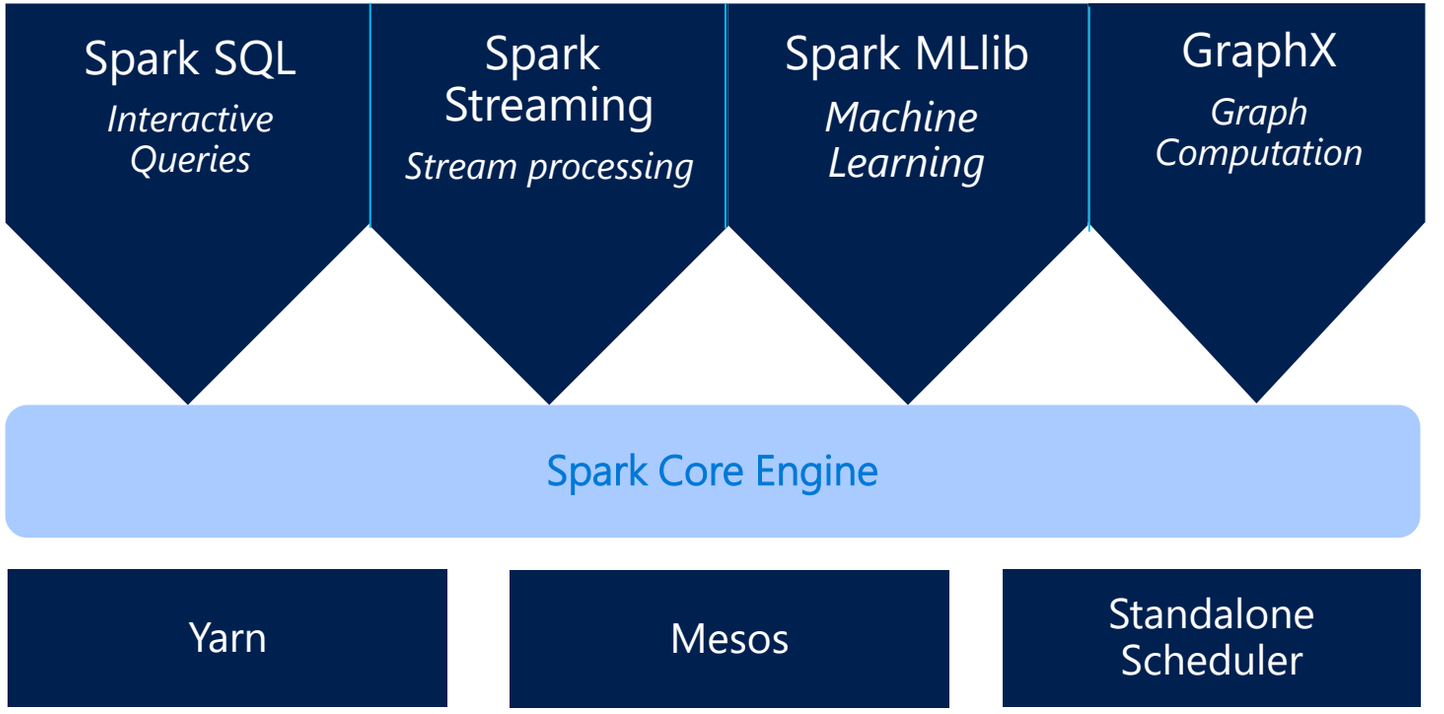
- Apache Software Foundation (ASF) open-source data processing project built around speed, ease of use, and sophisticated analytics.
- In-memory engine that is up to 100 times faster than Hadoop.
- Largest open-source data project with 1000+ contributors.
- Highly extensible with support for Scala, Java, Python, .NET, and R alongside Spark SQL, GraphX, Streaming and Machine Learning Library (Mllib).

Apache Spark

A unified, open source, and parallel, and in-memory data processing framework for Big Data Analytics

Spark Unifies:

- Batch Processing
- Interactive SQL
- Stream processing
- Machine Learning
- Deep Learning
- Graph Processing



What is Databricks

- It's a managed platform for running Apache Spark
- No cluster management
- No tedious maintenance tasks
- Point-and-click platform for developers that prefer a user interface
- Capabilities to automate aspects of data workloads with automated jobs
- Optimized autoscaling to resize a cluster intelligently

Knowledge Check

What is the major problem addressed by Databricks?

What are the capabilities of Spark SQL?

How’s Databricks being used?

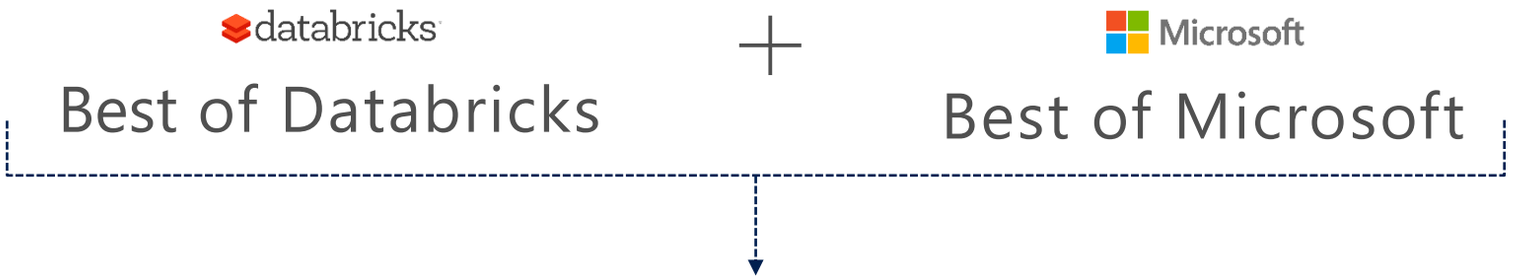
Lesson 2: Azure Databricks and Architecture

After completing this lesson, you will be able to:

- Understand Azure Databricks features and capabilities
- Understand the basics of Azure Databricks architecture

What is Azure Databricks?

A fast, easy and collaborative Apache® Spark™ based analytics platform optimized for Azure



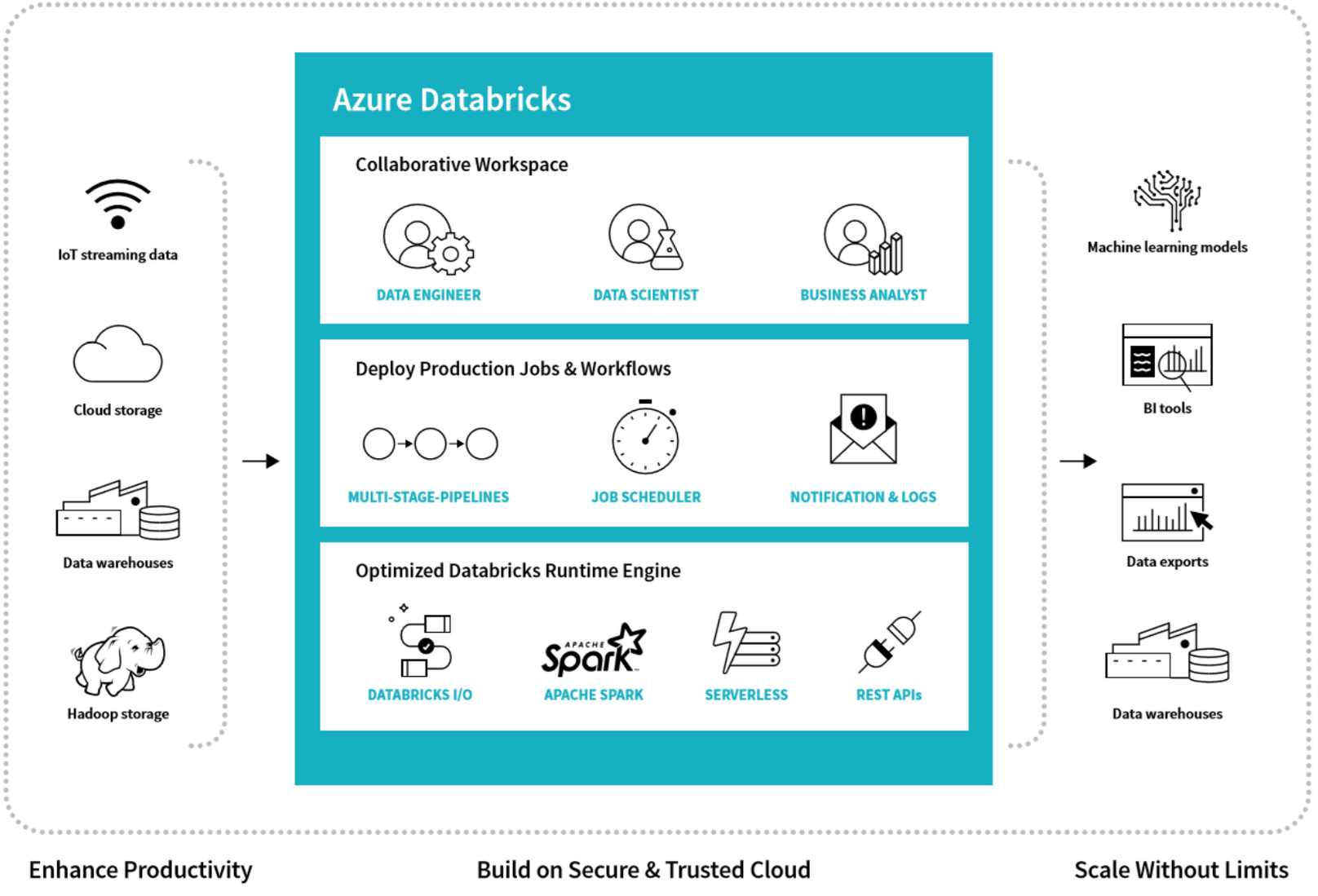
- Designed in collaboration with the founders of Apache Spark
- One-click set up; streamlined workflows
- Interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.
- Native integration with Azure services (Power BI, Synapse Analytics, Cosmos DB, Blob Storage)
- Enterprise-grade Azure security (Active Directory integration, compliance, enterprise-grade SLAs)

Azure Databricks

- Azure Databricks is a “first party” service on Azure
  - Unlike with other cloud services, it is not an Azure Marketplace, or a 3rd party hosted service
- Azure Databricks is integrated seamlessly with Azure services
  - Azure Portal: Services can be launched directly from Azure Portal
  - Azure Storage Services
  - Azure Active Directory (AAD): For user authentication
  - Azure Synapse and others
- Eliminates the need to create a separate account with Databricks

Azure Databricks

- The Azure Databricks service sits inside the Azure cloud
- Access all your Azure data sources to apply the power of the Azure Databricks analytics engine
- Distribute your results by writing to visual dashboards or back to data warehouses for analytics



Scale

- Operate at Massive Scale without Limits, Globally

Databricks enables your analytics processes to **scale up and down automatically**, enabling you to process all your data at once.

- Optimized Performance

Improve performance by as much as 10-100x over traditional Apache Spark deployments with performance optimizations including caching, indexing, and advanced query optimization.

Security

- Simplify Security and Identity Control

  Built-in integration with Azure Active Directory takes advantage of your existing roles and security settings.

- Build with Confidence

  Azure Databricks is backed by support, compliance and SLAs on the most-trusted cloud platform.

- Regulate Access

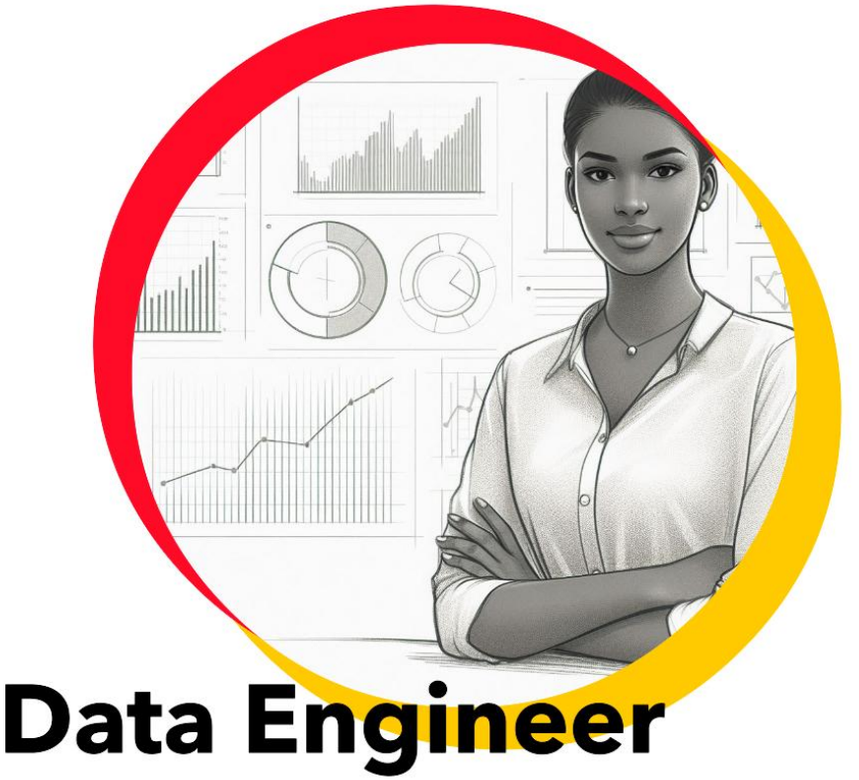
  Set fine-grained user permissions to Azure Databricks Notebooks, clusters, jobs, and data with different levels of permissions.

Personas in Azure Databricks

- Data Engineering and Data Science
- Machine Learning
- Business Analyst

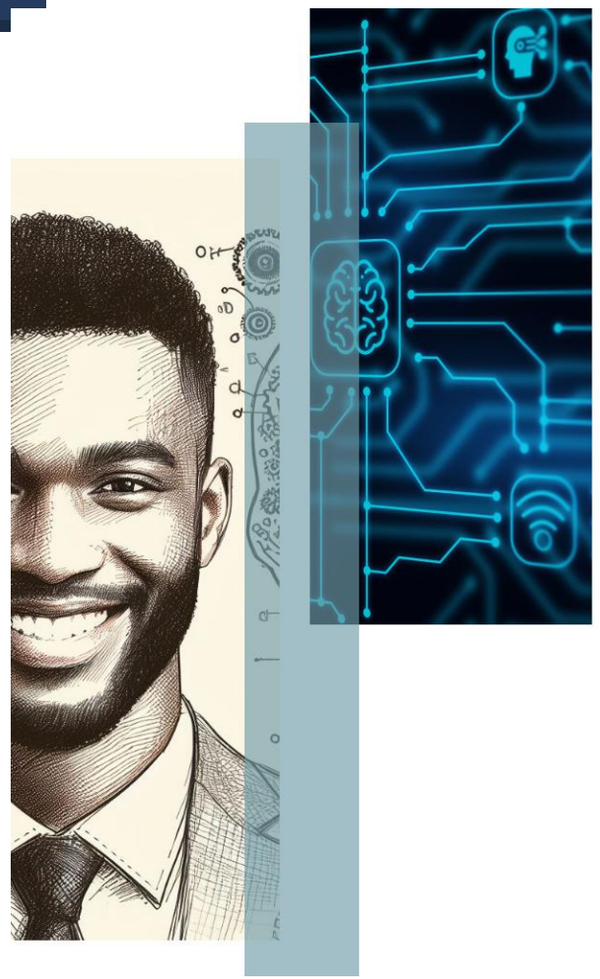
Responsibilities of a Data Engineer in Databricks

- Design, build, and maintain robust and scalable data pipelines
- Clean, transform and manipulate large data sets
- Ensure the timely and accurate delivery of data
- Collaborate with data scientists to optimize data sets for machine learning
- Proactively monitor data quality and system health
- Stay up-to-date with new features and best practices



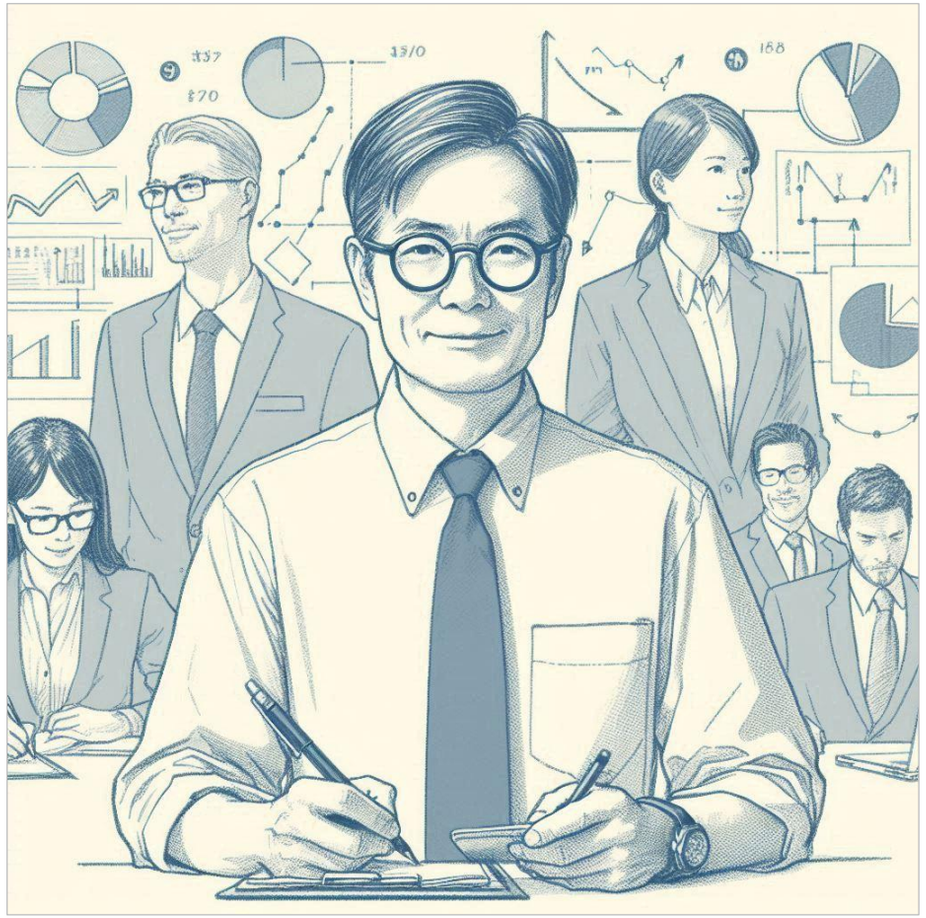
Responsibilities of a Machine Learning Engineer in Databricks

- Develop machine learning models to solve business problems.
- Implement, evaluate, and optimize machine learning algorithms.
- Collaborate with Data Scientists and Business Analysts to understand business needs.
- Stay up-to-date with the latest machine learning trends and techniques.



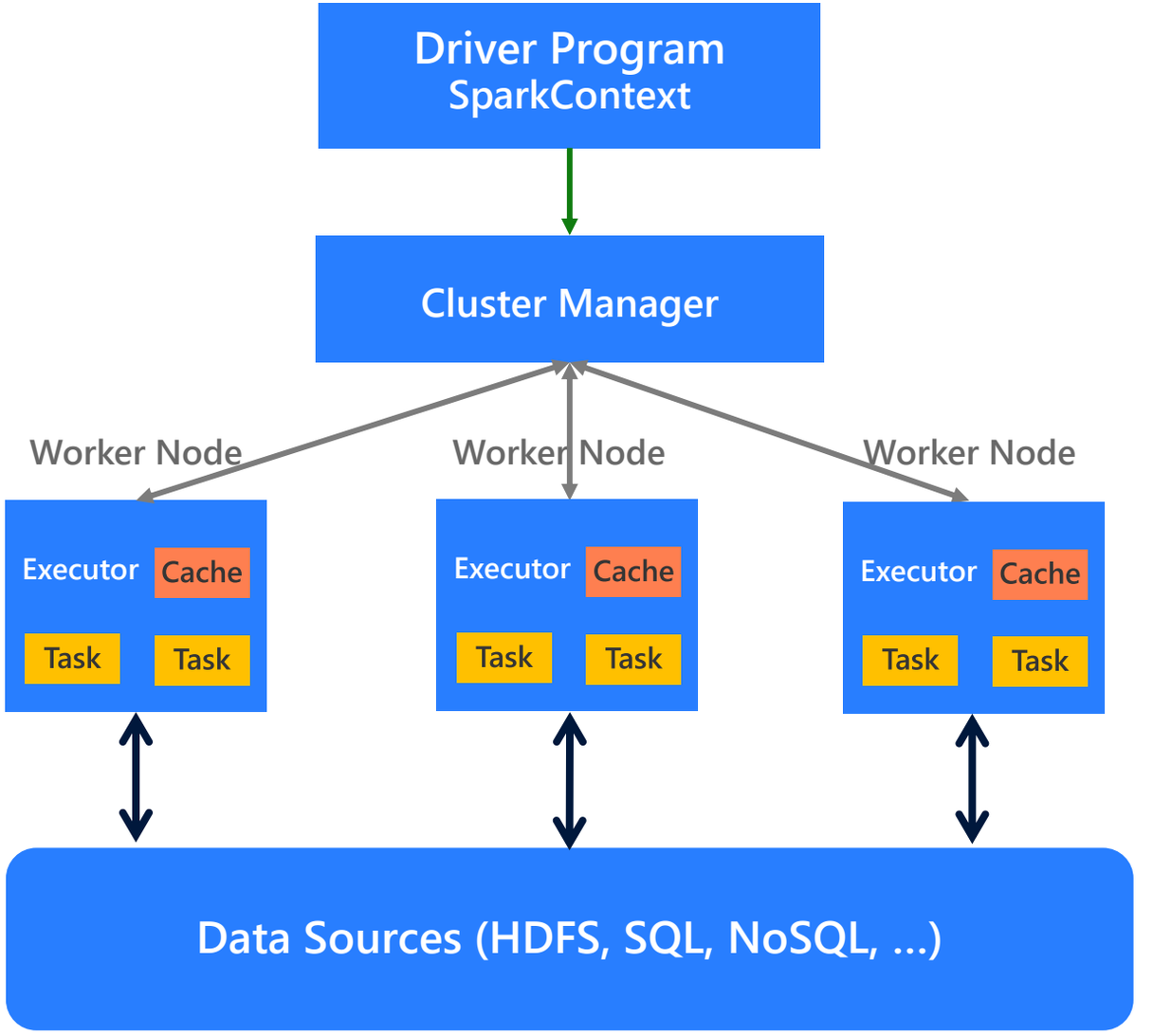
Responsibilities of a Business Analyst in Databricks

- Analyze and interpret data to inform business decisions.
- Work closely with stakeholders to understand their needs and objectives.
- Translate business requirements into technical specifications.
- Collaborate with technical teams to develop and implement solutions.

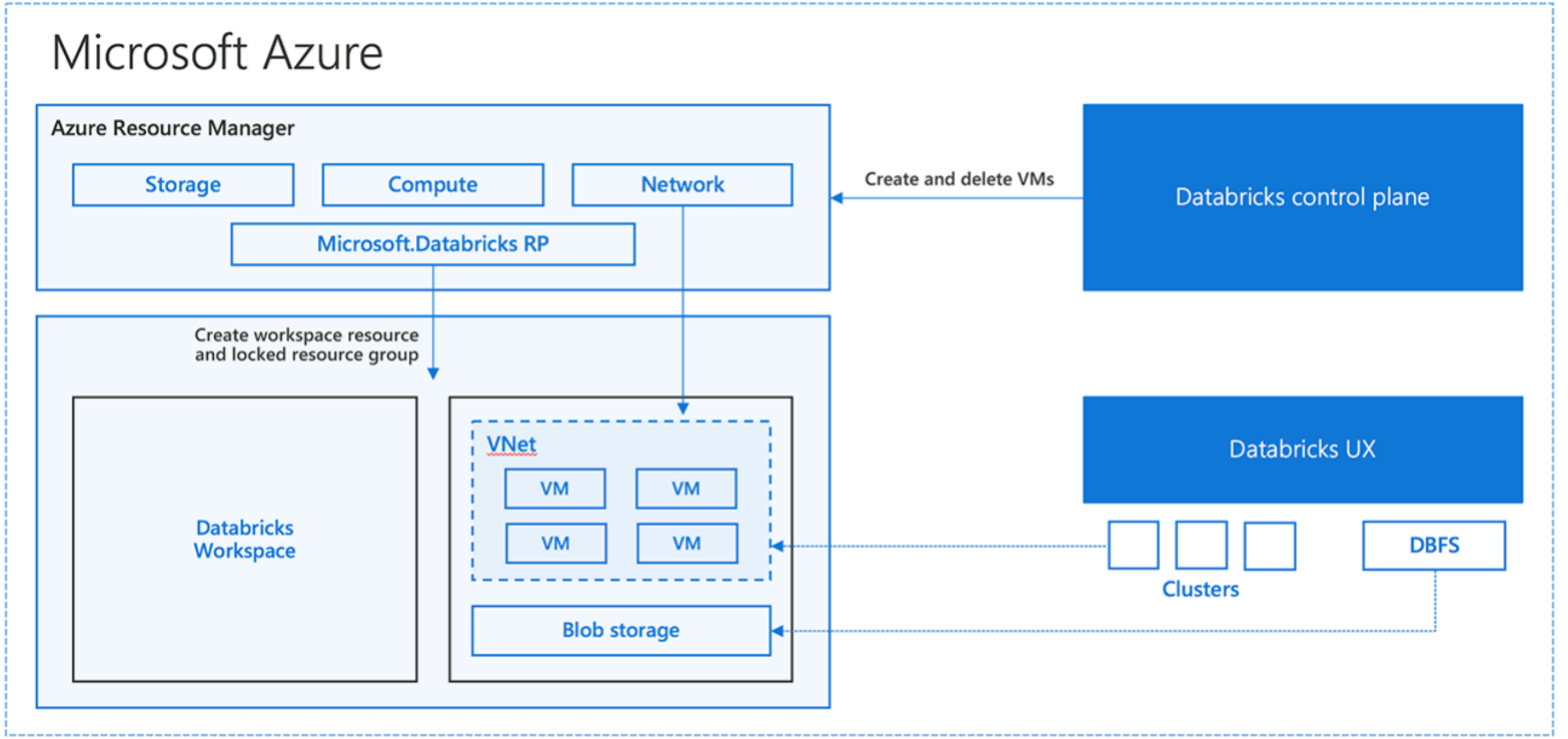


General Spark Architecture

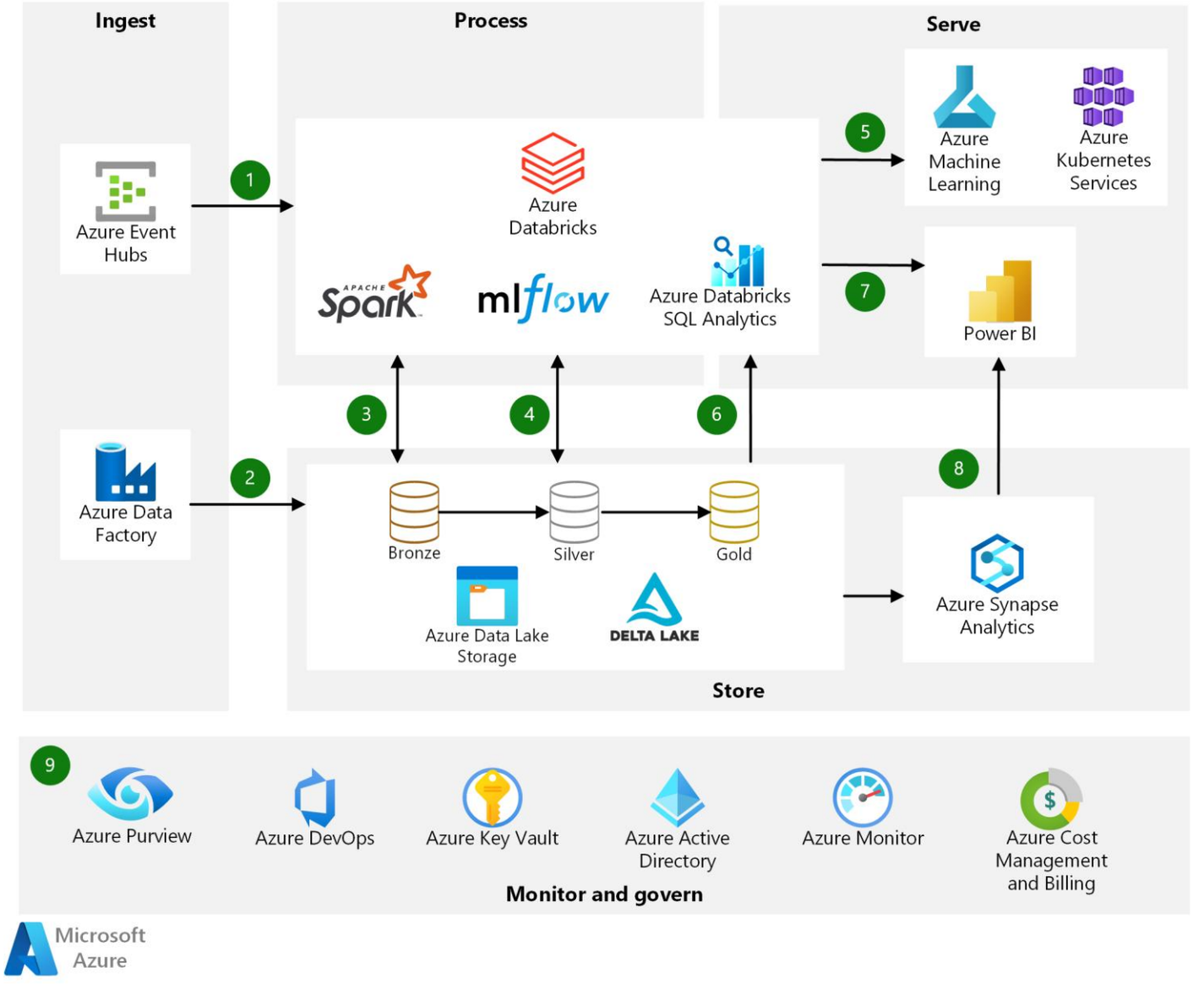
- 'Driver' runs the user's 'main' function and executes the various parallel operations on the worker nodes.
- The results of the operations are collected by the driver
- The worker nodes read and write data from/to Data Sources including HDFS, ADLS, others.
- Worker node also cache transformed data in memory.
- Worker nodes and the Driver Node execute as VMs in public clouds



Azure Databricks Cluster Architecture



Modern analytics architecture



Knowledge Check

What are the advantages of using Databricks on Azure?

Lesson 3: Price Tiers and Workloads

After completing this lesson, you will be able to:

- Understand different workloads available on Azure Databricks
- Understand the differences between 2 tiers called Standard and Premium

Azure Databricks Pricing

Price Tiers: Azure Databricks offers Standard and Premium tiers for various workload types with a difference in the features:

Standard Tier

Premium Tier

Workloads: Azure Databricks offers different compute types on several VM Instances tailored for your data analytics workflow

All-Purpose Compute

Meant for data scientists to explore, visualize, manipulate, and share data and insights interactively.

Jobs Compute

Meant for data engineers to build and execute jobs

SQL Warehouses

SQL commands on data objects in the SQL editor or interactive notebooks. You can create SQL warehouses using the UI, CLI, API.

Pricing Tiers in Azure Databricks

- Azure Databricks offers a range of pricing options.
- Premium, Standard, and Data Engineering Light Pricing tiers.
- Workloads can be optimized to reduce costs.

Lesson 4: Azure Databricks Artifacts

After completing this lesson, you will be able to:

- Understand the major components of Databricks

Core Artifacts

Clusters / Compute	Set of Azure Linux VMs that host the Spark Driver and Worker Nodes
Workspaces	Enables users to organize, and share, their Notebooks, Libraries and Dashboards
Notebooks	A popular way to develop, and run, Spark Applications
Libraries	Enables external code to be imported and stored into a Workspace
Jobs	Schedule mechanism to submit Spark application code for execution on the Databricks clusters
Secrets	A key-value pair that stores secret material, with a key name unique within a secret scope

Knowledge Check

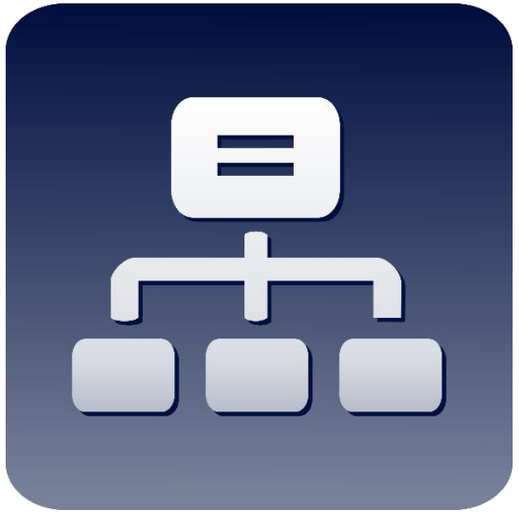
What are the core artifacts in Azure Databricks?

Lesson 5: Azure Databricks Clusters

After completing this lesson, you will be able to:

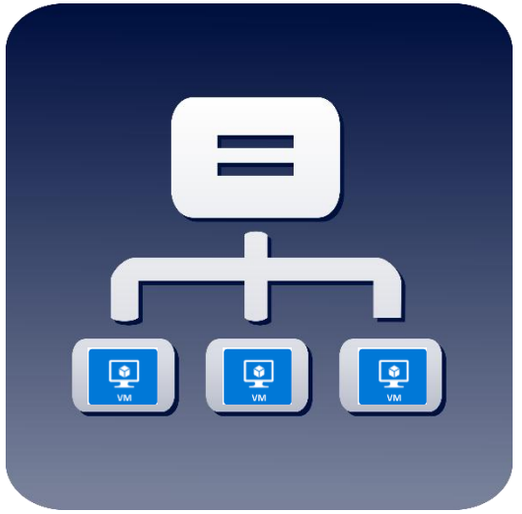
- Understand the Databricks Cluster and how it works?
- Learn about Access Control on Databricks Cluster

Compute / Cluster



- Clusters are the set of Azure Linux VMs that host the Spark Worker and Driver Nodes
- Spark application code (i.e. Jobs) runs on the provisioned clusters.
- Clusters are launched in your subscription but are managed through the Azure Databricks portal.
  - You can also manage clusters using CLI or API

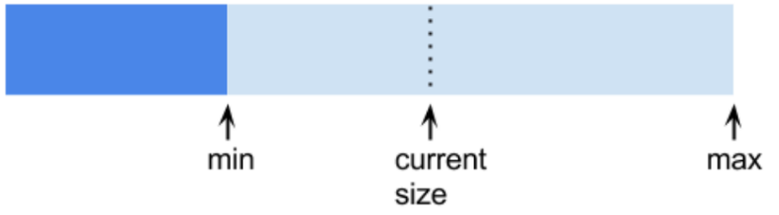
Compute / Cluster



- Azure Databricks provides a comprehensive set of graphical wizards to manage the complete lifecycle of clusters - from creation to termination.
- Types of Cluster:
  - **Interactive Clusters:** are used to analyze data collaboratively with interactive notebooks.
  - **Job Clusters:** are used to run fast and robust automated workloads using the UI or API.

Cluster Auto Scaling

- Autoscaling allow Databricks to automatically resize your cluster by providing the min and max range of workers.
- When you select autoscaling, the cluster size is automatically adjusted between the minimum and a maximum number of worker limits during the cluster's lifetime.



Cluster Auto Scaling

- During runtime Databricks will dynamically reallocate workers to account for the characteristics of your job.
- During computationally demanding phases, Databricks automatically adds additional workers and removes when they're no longer needed.

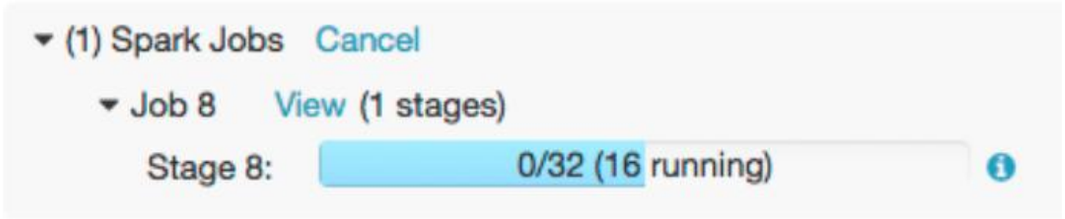
How Auto Scaling Works?

- Databricks monitors load on Spark clusters and decides whether to scale a cluster up or down and by how much.
- If a cluster has pending Spark tasks, the cluster scales up. If a cluster does not have any pending Spark tasks, the cluster scales down.
- The autoscaling algorithm ensure that users experience fast workloads while maintaining efficient cluster utilization.

How Auto Scaling Works?

- Clusters with no *pending* tasks *do not* scale up. This usually indicates that the cluster is fully utilized and adding more nodes will not make the processing faster.

For example, this cluster currently has 16 running tasks and 16 pending tasks (total tasks - running tasks) and will be scaled up.



Cluster Auto Termination

- You can set auto termination for a cluster.
- During cluster creation, you can specify an inactivity period in minutes after which you want the cluster to be terminated.
- If the time difference between the current time and the last command run on the cluster is more than the inactivity period specified, Azure Databricks automatically terminates that cluster.

AutoScaling and AutoTermination Benefits

- Need not worry about # of nodes
  - You don't need to guess or determine by trial and error, the correct number of nodes for the cluster.
- Dynamic Scaling
  - As the workload changes you do not have to manually tweak the number of nodes
- It's pay-per-use!
  - You do not have to worry about wasting resources when the cluster is idle.
- Easy management
  - You do not have to wait and watch for jobs to complete just so you can shutdown the clusters.

Demo:
Azure Databricks Workspace Creation and Cluster Configuration

Creating an Azure Databricks Workspace and Configure a Databricks Cluster

Lab: Workspace and Cluster Configuration

Learn about Workspace Creation and Cluster Configuration

Demo:
Azure Databricks Cluster Management

How to manage Azure Databricks Cluster

Lab: Cluster Management

Learn about managing Azure Databricks Clusters

Knowledge Check

What are the advantages of Auto Scaling?

What is the difference between Interactive Cluster and Job Cluster?

What is Auto Scaling for Jobs?

Lesson 6: Azure Databricks Workspace

After completing this lesson, you will be able to:

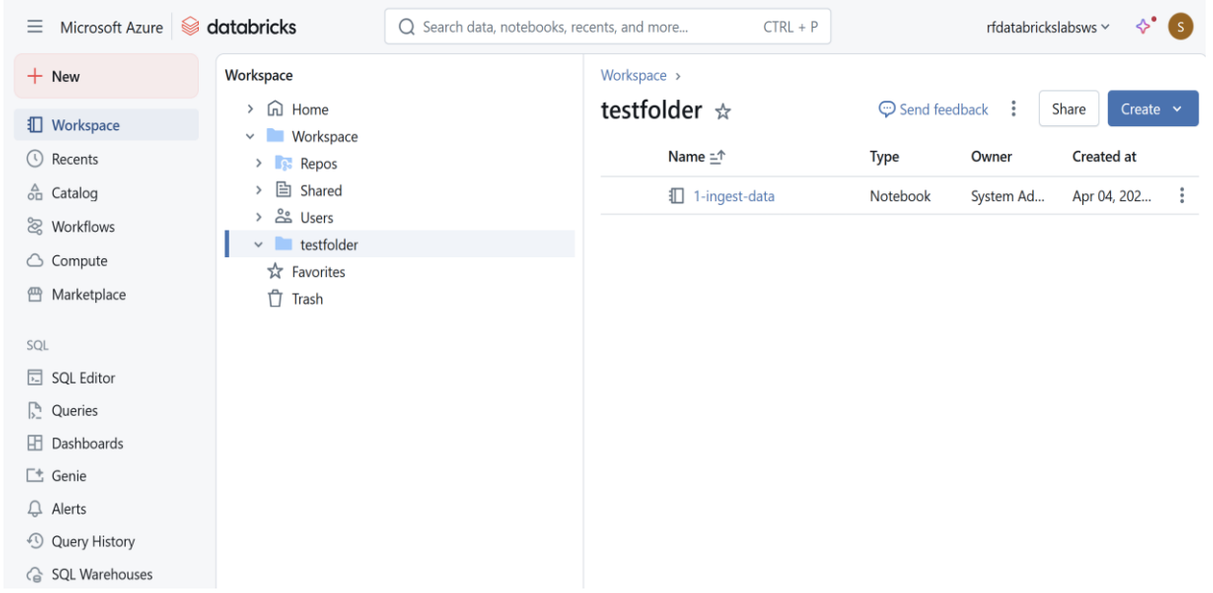
- Understand Workspace and its importance
- Manage Workspace Access Control
- Manage Azure Databricks Libraries

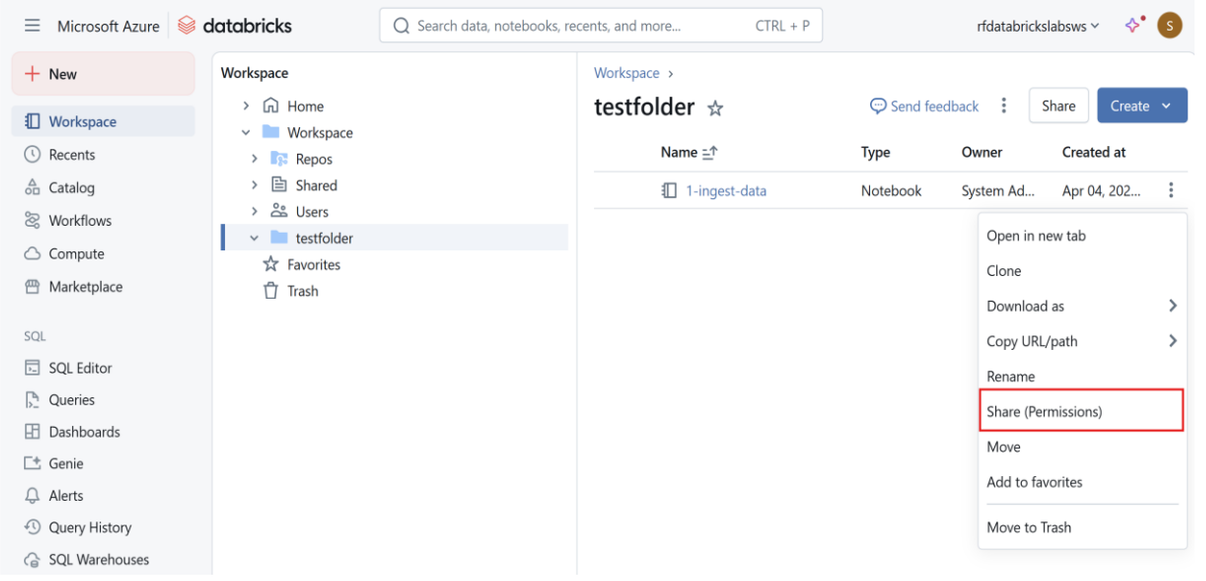
Workspace

- Workspaces enable users to organize—and share—their Notebooks, Libraries and Dashboards.
- Everything in a workspace is organized into hierarchical folders. Folders can hold Libraries, Notebooks, Dashboard or more (sub) folders.
  - Icons indicate the type of the object contained in a folder
- Every user has one directory that is private and unshared.
  - By default, the workspace and all its contents are available to users.

Workspace

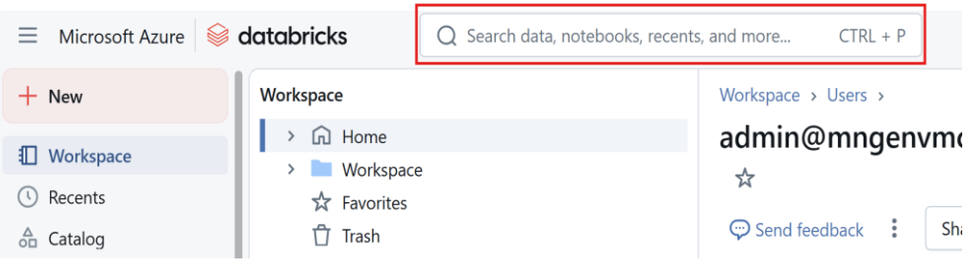
- Fine grained access control can be defined on workspaces to enable *secure collaboration with colleagues*.

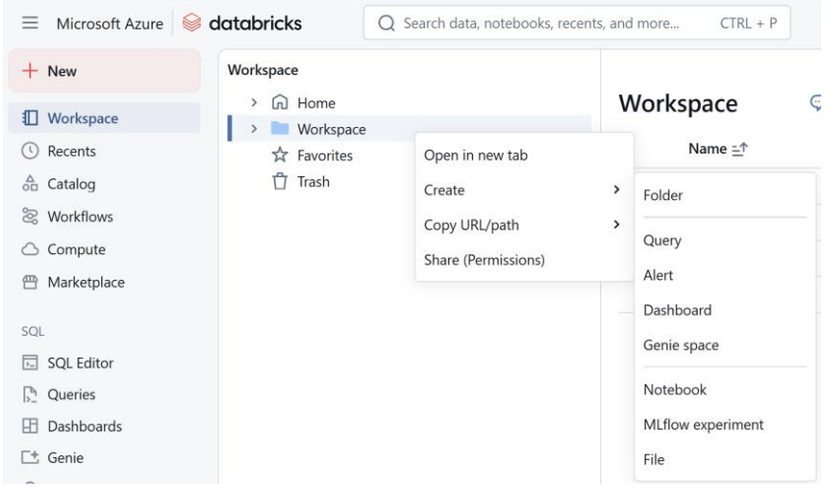


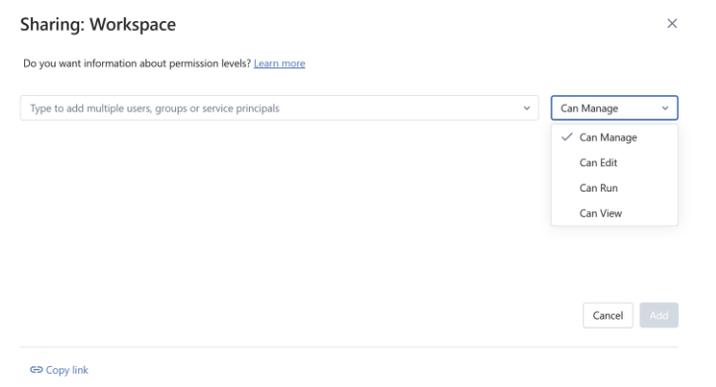


Workspace Operations

- You can search the entire workspace.
- In the Azure Databricks Portal, via the Workspaces drop down menu, you can:
  - Create Folders, Notebooks and Libraries.
  - Import Notebooks into the Workspace.
  - Export the Workspace to a database archive.
  - Set Permissions. You can grant 4 levels of permissions
    - Can Manage
    - Can Read
    - Can Edit
    - Can Run







Folder Operations and Access Control

- In the Azure Databricks Portal, via the Folder drop down menu, you can:
  - Create Folders, Notebooks and Libraries within the folder.
  - Clone the folder to create a deep copy of the folder.
  - Rename or delete the folder.
  - Move the folder to another location.
  - Export a folder to save it and its contents as a Databricks archive.
  - Import a saved Databricks archive into the selected folder.
  - Set Permissions for the folder. As with Workspaces you can set 5 levels of permissions: *No Permissions, Can Manage, Can Read, Can Edit, Can Run*.

Libraries

- Enables external code to be imported and stored into a Workspace.
- Libraries are containers to hold all your *Python*, *R*, *Java/Scala* libraries.
- Libraries resides within workspaces or folders.
- Libraries are created by importing the source code.
- Imported libraries are immutable—can be deleted or overwritten only.
- You can customize installation of libraries via Init Scripts by writing custom UNIX scripts (not available on Shared Cluster!).
- Libraries can also be managed via the Library API or Library CLI.

Secrets

- A secret is a key-value pair that stores secret material, with a key name unique within a secret scope
- Secrets can be read from a notebook or job using the Secrets Utility (dbutils.secrets)
- A scope is limited to **1000 secrets**. The **maximum** allowed **secret value size** is **128 KB**
- A scope can be Databricks-backed or Azure Key Vault based
- Secret can be created and managed using the Secrets CLI.
- Secrets can also be managed via the Secrets API

Module Summary

- Azure Databricks and it’s capabilities
- Azure Databricks Architecture
- Azure Databricks Clusters concepts
- Working with Workspace and Libraries