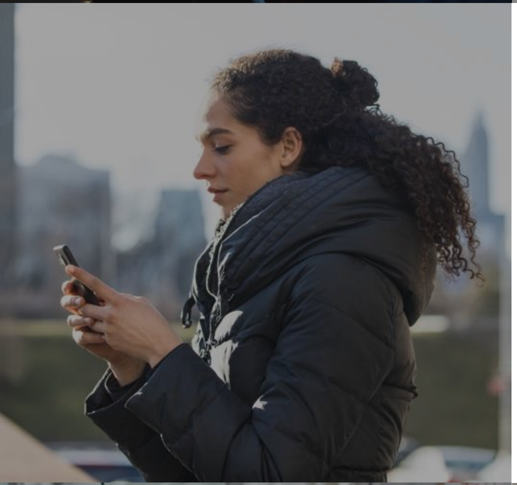
Microsoft

DP-200T01: Enabling Team Based Data Science with Azure Databricks

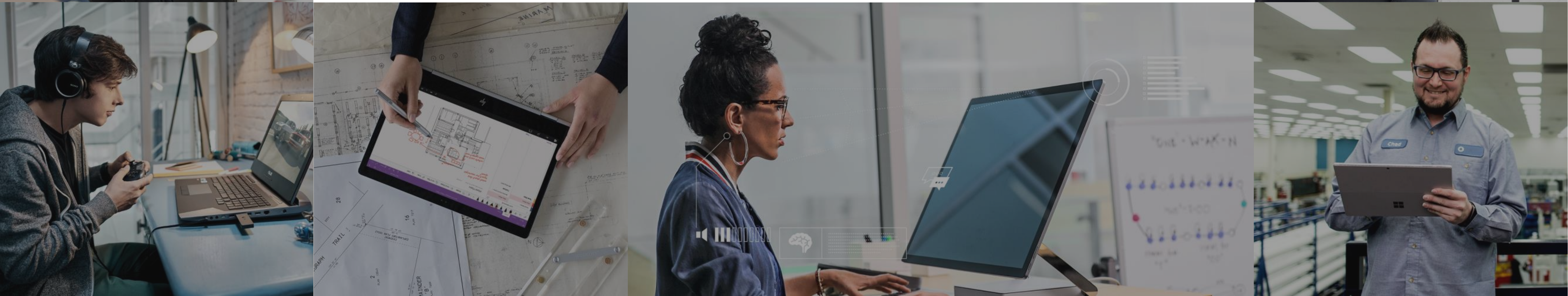
Agenda

- L01 – Describe Azure Databricks
- L02 - Provision Azure Databricks and Workspaces
- L03 - Read data using Azure Databricks
- L04 - Perform transformations with Azure Databricks



Lesson 01

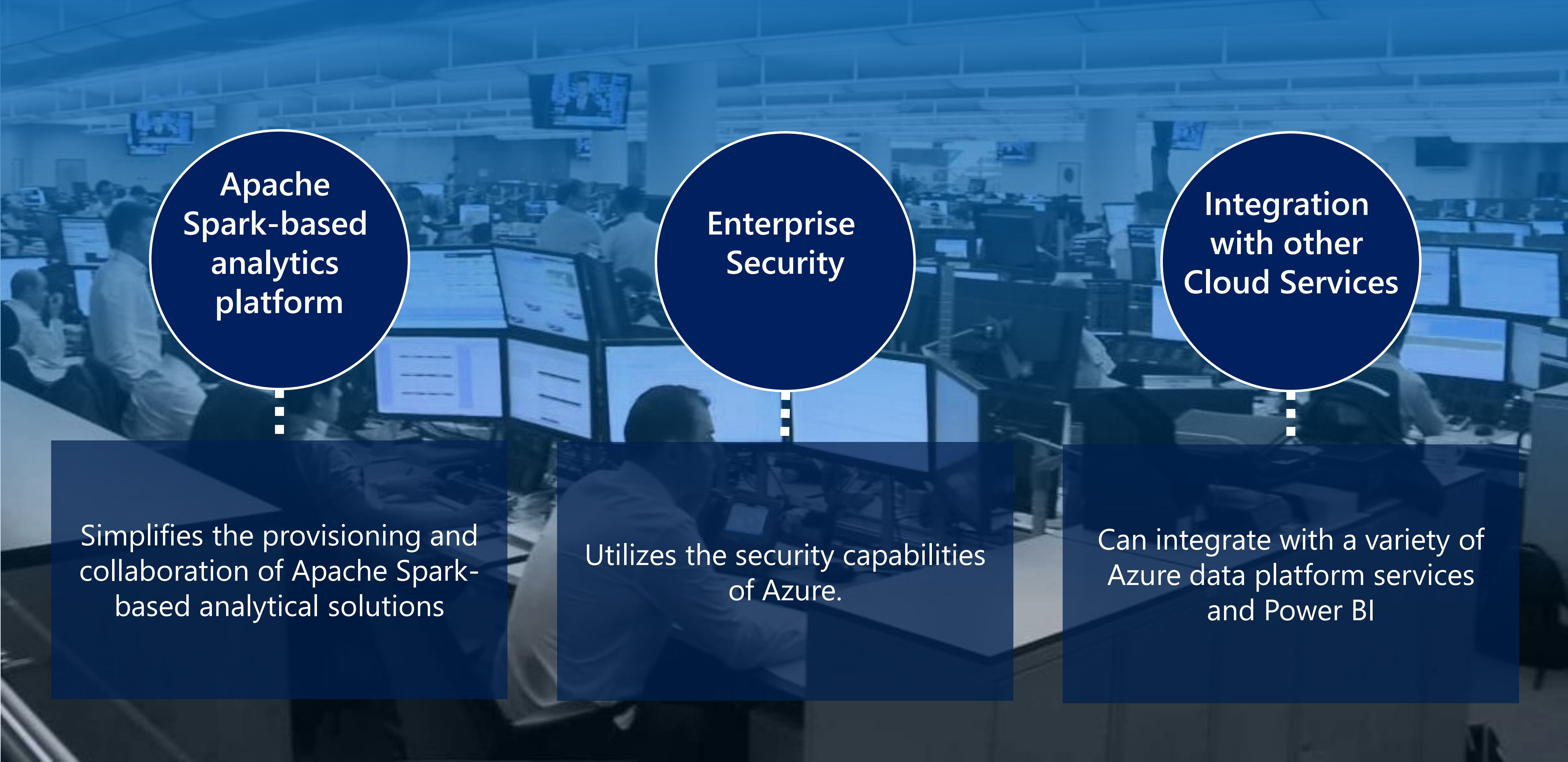
Describe Azure Databricks



Lesson Objectives

- What is Azure Databricks
- What are Spark based analytics platform
- How Azure Databricks integrates with enterprise security
- How Azure Databricks integrates with other cloud services

What is Azure Databricks



Apache Spark-based analytics platform

Simplifies the provisioning and collaboration of Apache Spark-based analytical solutions

Enterprise Security

Utilizes the security capabilities of Azure.

Integration with other Cloud Services

Can integrate with a variety of Azure data platform services and Power BI

What is Apache Spark

Apache Spark emerged to provide a parallel processing framework that supports in-memory processing to boost the performance of big-data analytical applications on massive volumes of data.

Interactive Data Analysis
Used by business analysts or data engineers to analyze and prepare data

Streaming Analytics
Ingest data from technologies such as Kafka and Flume to ingest data in real-time

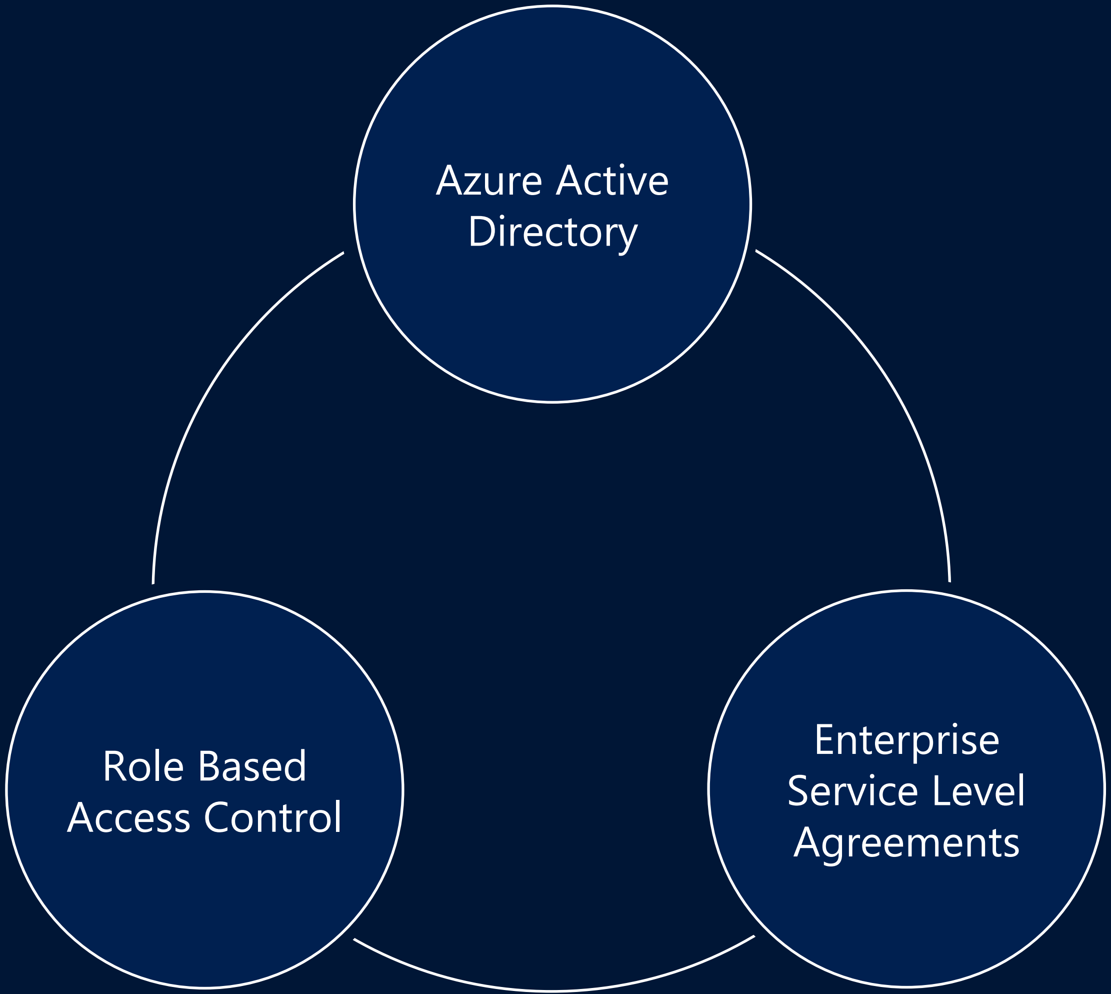
Machine Learning
Contains a number of libraries that enables a Data Scientist to perform Machine Learning

Why use Azure Databricks
Azure Databricks is a wrapper around Apache Spark that simplifies the provisioning and configuration of a Spark cluster in a GUI interface

Azure Databricks components:

- Spark SQL and DataFrames
- Streaming
- Mlib
- GraphX
- Spark Core API

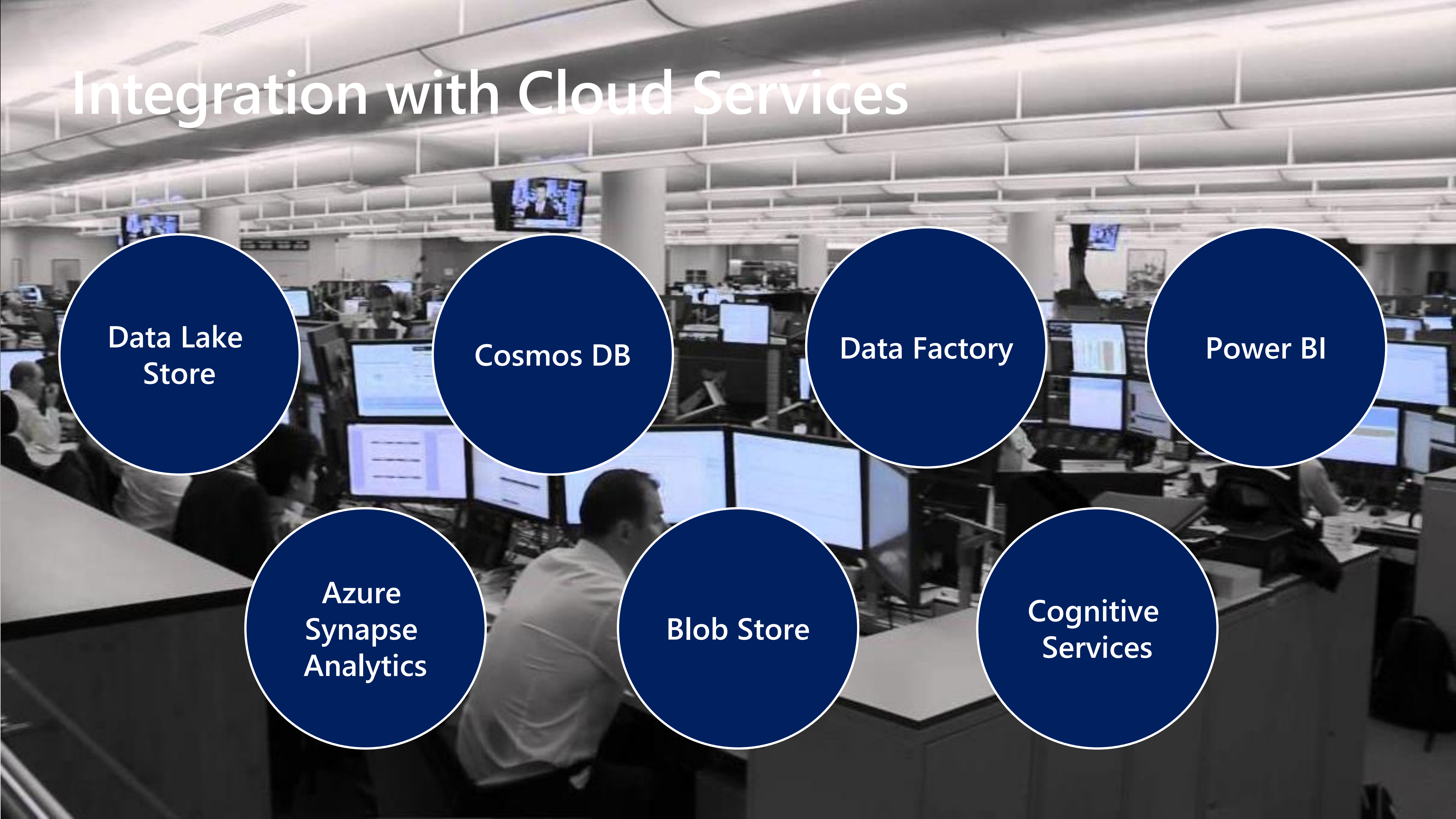
Enterprise Security


Azure Active Directory

Role Based Access Control

Enterprise Service Level Agreements

Integration with Cloud Services



Data Lake Store

Cosmos DB

Data Factory

Power BI

Azure Synapse Analytics

Blob Store

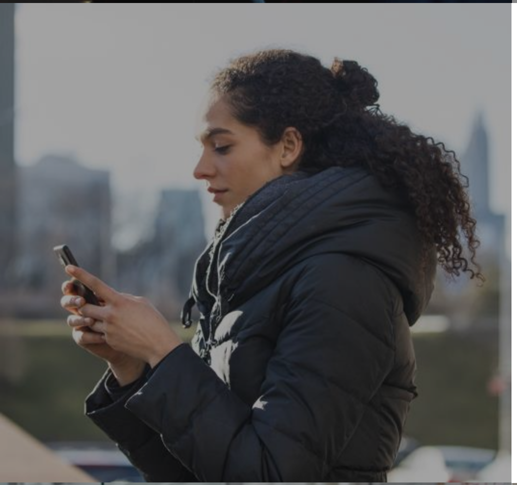
Cognitive Services

Review Questions

- Q01 – Azure Databricks encapsulates which Apache technology?
- A01 – Apache Spark.

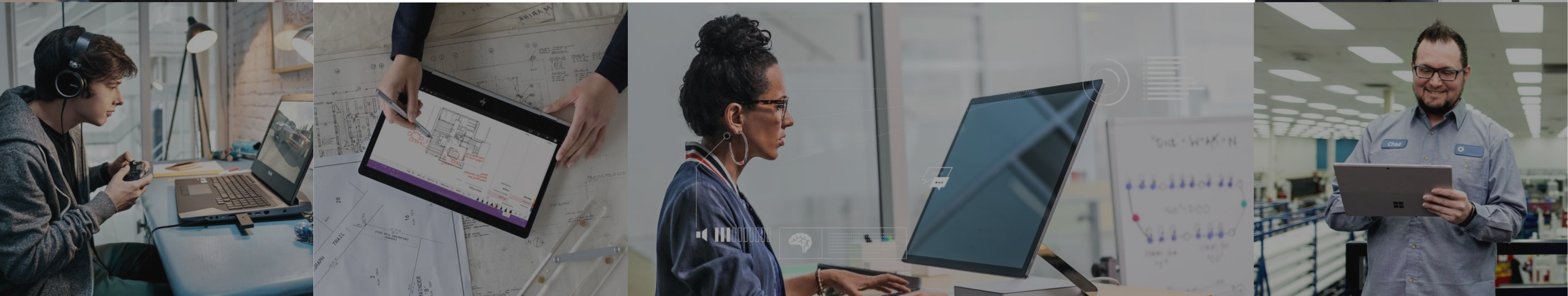
- Q02 - Which security features does Azure Databricks not support?
- A02 – Shared Access Keys.

- Q03 - Which of the following Azure Databricks is used for support for R, SQL, Python, Scala, and Java?
- A03 – Spark Core API.



Lesson 02

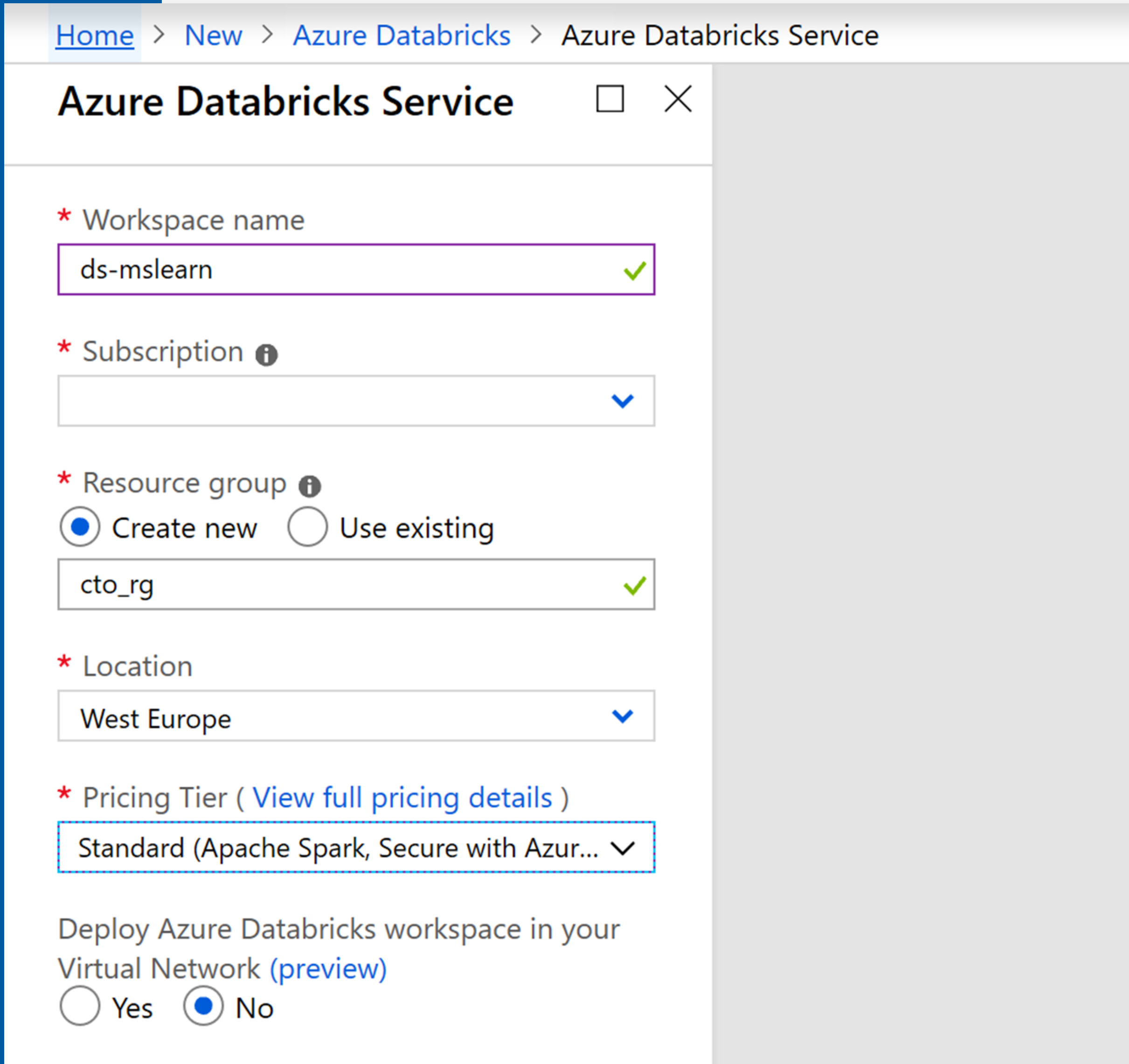
Provision Azure Databricks and Workspaces



Lesson Objectives

- Create your own Azure Databricks workspace
- Create a cluster and notebook in Azure Databricks

Create an Azure Databricks Workspace.



Home > New > Azure Databricks > Azure Databricks Service

Azure Databricks Service

* Workspace name

ds-mslearn ✓

* Subscription

* Resource group

● Create new ○ Use existing

cto_rg ✓

* Location

West Europe

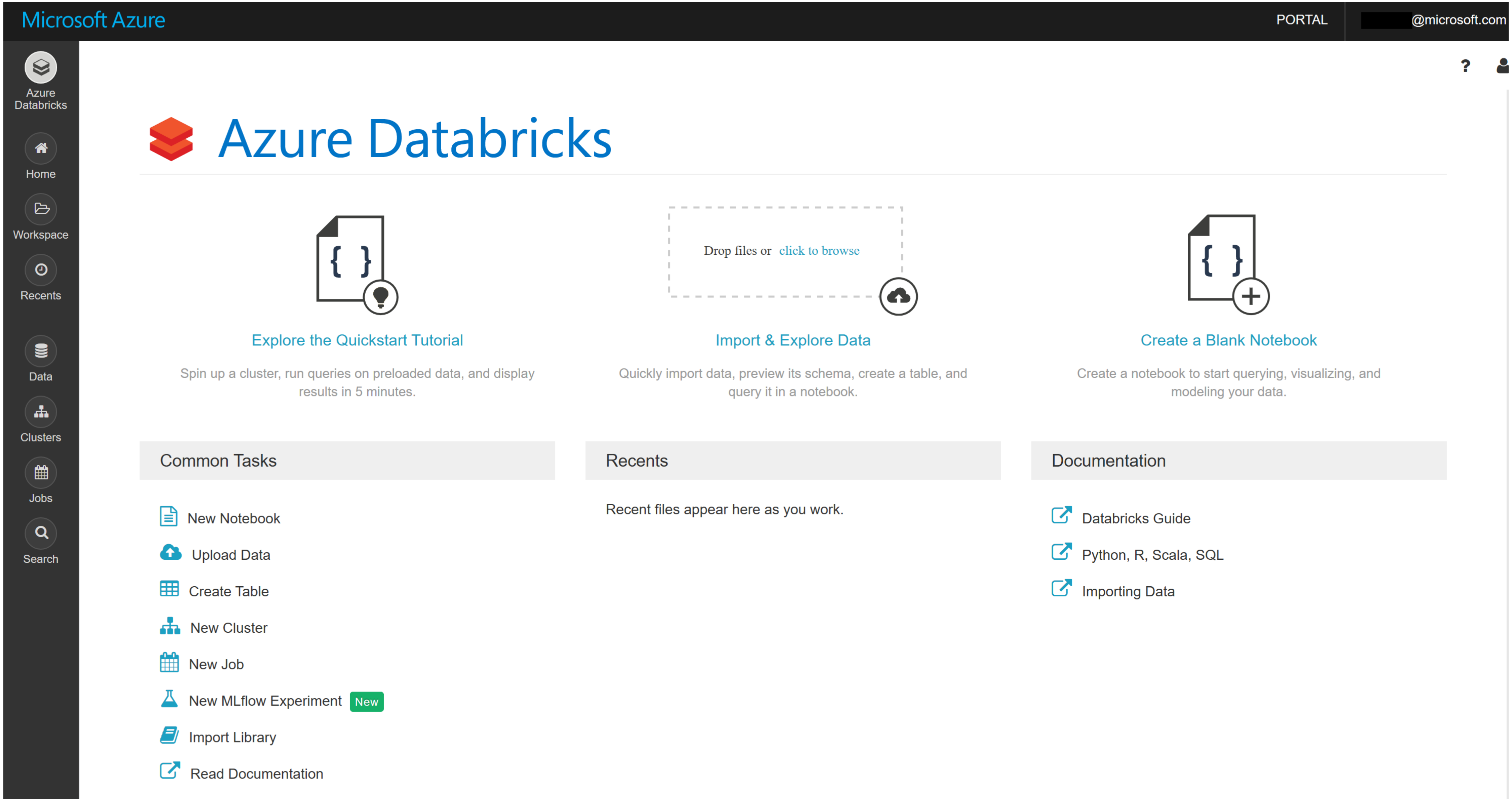
* Pricing Tier (View full pricing details)

Standard (Apache Spark, Secure with Azur...

Deploy Azure Databricks workspace in your Virtual Network (preview)

○ Yes ● No

Create a Cluster and Notebook in Azure Databricks



Microsoft Azure

PORTAL [redacted]@microsoft.com

Azure Databricks

Home

Workspace

Recents

Data

Clusters

Jobs

Search

Azure Databricks

Drop files or click to browse

Explore the Quickstart Tutorial

Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.

Import & Explore Data

Quickly import data, preview its schema, create a table, and query it in a notebook.

Create a Blank Notebook

Create a notebook to start querying, visualizing, and modeling your data.

Common Tasks

- New Notebook
- Upload Data
- Create Table
- New Cluster
- New Job
- New MLflow Experiment New
- Import Library
- Read Documentation

Recents

Recent files appear here as you work.

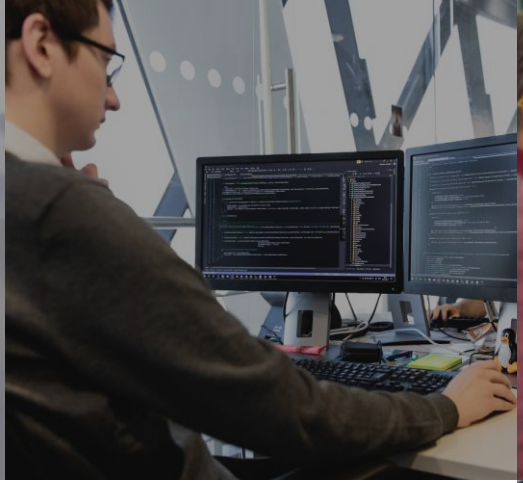
Documentation

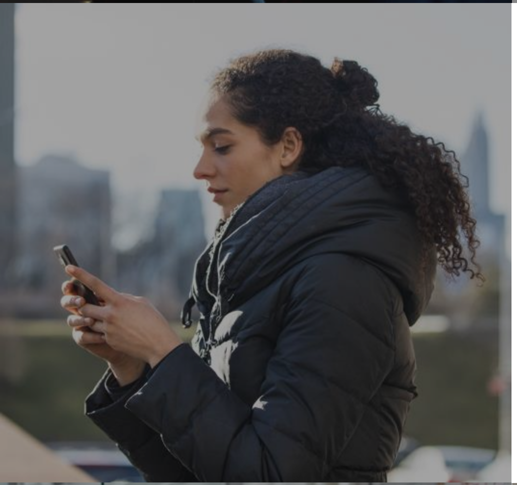
- Databricks Guide
- Python, R, Scala, SQL
- Importing Data

Review Questions

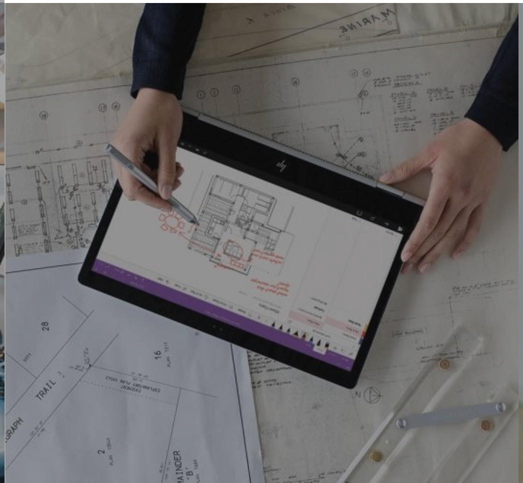
- Q01 - Which Notebook format is used in Databricks?
- A01 – DBC.

- Q02 - Which browsers are recommended for best use with Databricks Notebook?
- A02 – Chrome and Firefox.





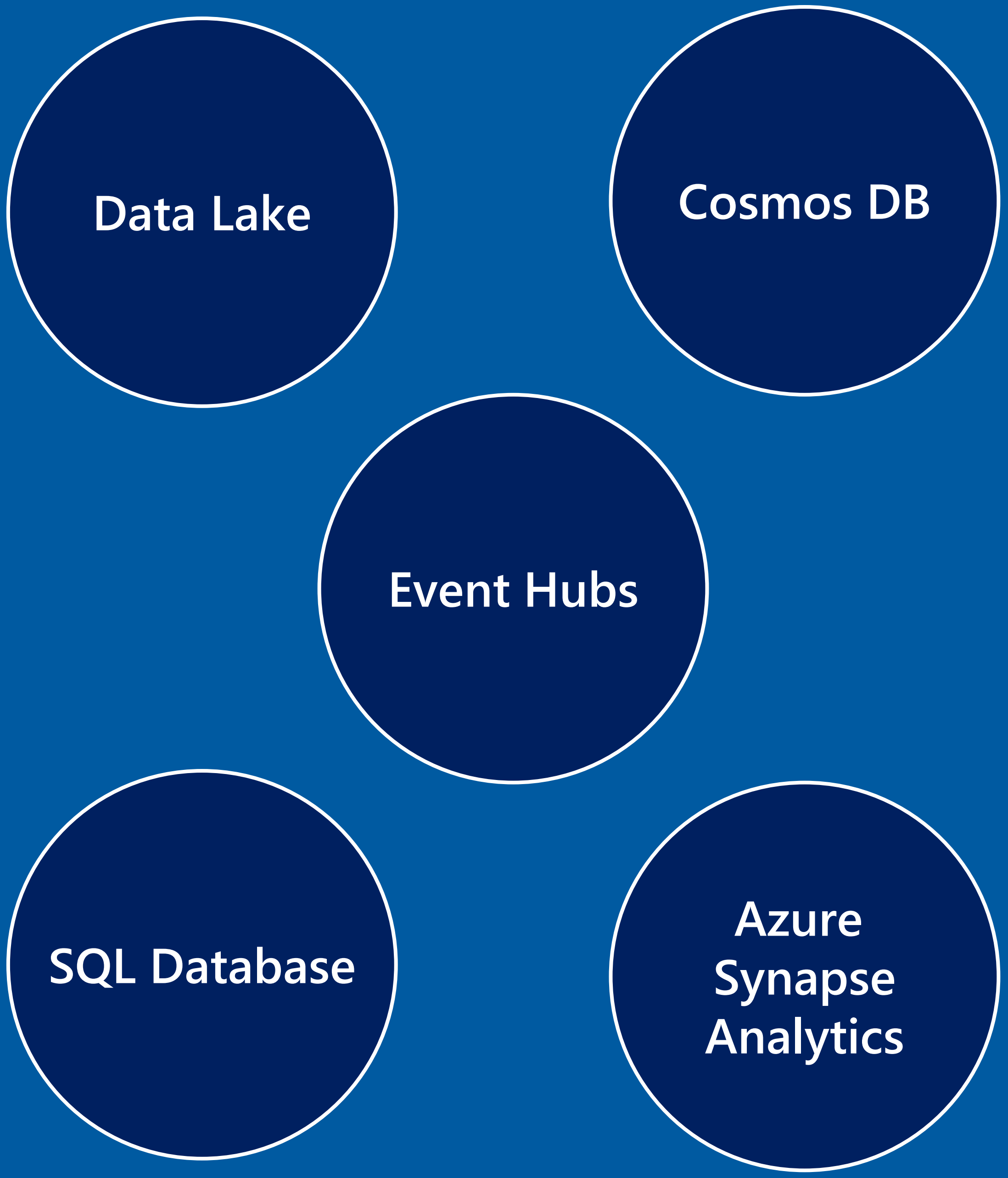
Lesson 03
Read Data using Azure Databricks



Lesson Objectives

- Use Azure Databricks to access data sources
- Reading Data in Azure Databricks

Use Azure Databricks to access data sources.



Data Lake

Cosmos DB

Event Hubs

SQL Database

Azure Synapse Analytics

Reading Data in Azure Databricks.

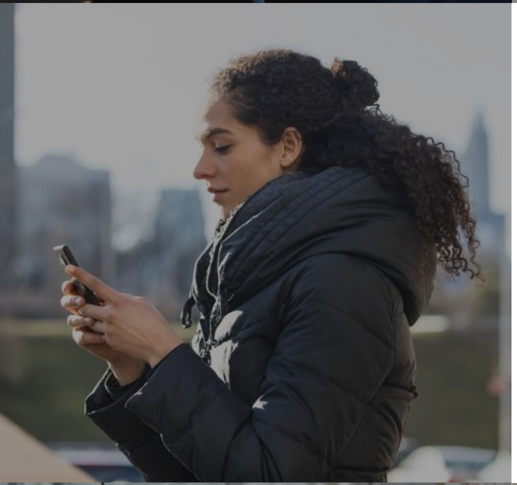
SQL	DataFrame (Python)
SELECT col_1 FROM myTable	df.select(col("col_1"))
DESCRIBE myTable	df.printSchema()
SELECT * FROM myTable WHERE col_1 > 0	df.filter(col("col_1") > 0)
..GROUP BY col_2	..groupBy(col("col_2"))
..ORDER BY col_2	..orderBy(col("col_2"))
..WHERE year(col_3) > 1990	..filter(year(col("col_3")) > 1990)
SELECT * FROM myTable LIMIT 10	df.limit(10)
display(myTable) (text format)	df.show()
display(myTable) (html format)	display(df)

Review Questions

- Q01 – How do you connect your Spark cluster to the Azure Blob?
- A01 – By mounting it

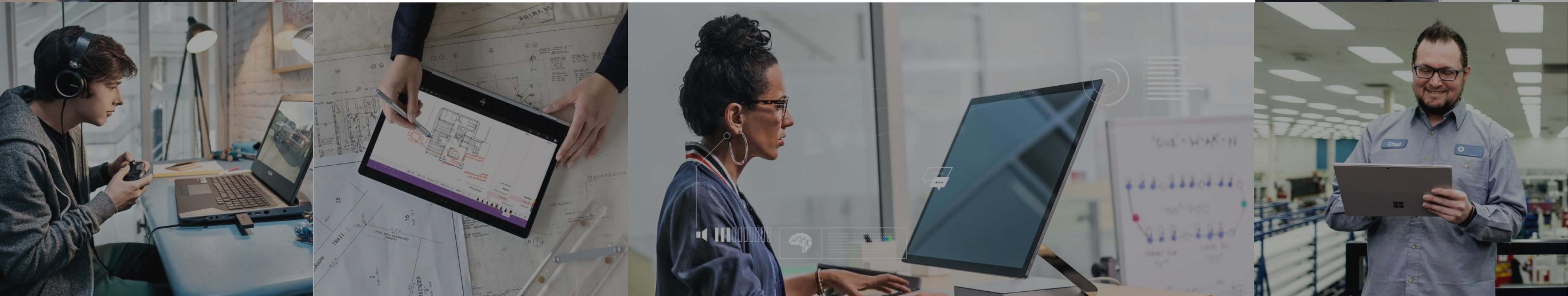
- Q02 - How does Spark connect to databases like MySQL, Hive and other data stores?
- A02 – JDBC

- Q03 - How do you specify parameters when reading data?
- A03 – Using .option() during your read allows you to pass key/value pairs specifying aspects of your read



Lesson 04

Perform Transformations with Azure Databricks



Lesson Objectives

- Performing ETL to populate a data model
- Perform basic transformations
- Perform advanced transformations with user-defined functions

Performing ETL to populate a data model

The goal of transformation in Extract Transform Load (ETL) is to transform raw data to populate a data model.

Extraction	Data Validation	Transformation	Corrupt Record Handling	Loading Data
Connect to many data stores: • Postgres • SQL Server • Cassandra • Cosmos DB • CSV, Parquet • Many more..	Validate that the data is what you expect.	Applying structure and schema to your data to transform it into the desired format.	Built-in functions of Databricks allow you to handle corrupt data such as missing and incomplete information.	Highly effective design pattern involves loading structured data back to DBFS as a parquet file.

Basic transformation

Normalizing Values

Missing/Null data

De-duplication

Pivoting Data frames

Advanced Transformations

Advanced data transformation using custom and advanced user-defined functions, managing complex tables and loading data into multiple databases simultaneously.

User-defined functions

This fulfils scenarios when you need to define logic specific to your use case and when you need to encapsulate that solution for reuse. UDFs provide custom, generalizable code that you can apply to ETL workloads when Spark's built-in functions won't suffice.

Joins and lookup tables

A standard (or shuffle) join moves all the data on the cluster for each table to a given node on the cluster. This is an expensive operation. Broadcast joins remedy this situation when one DataFrame is sufficiently small enough to duplicate on each node of the cluster, avoiding the cost of shuffling a bigger DataFrame.

Multiple databases

Loading transformed data to multiple target databases can be a time-consuming activity. Partitions and slots are options to get optimum performance from database connections. A partition refers to the distribution of data while a slot refers to the distribution of computation.

Review Questions

- Q01 – By default, how are corrupt records dealt with using spark.read.json()
- A01 – They appear in a column called "_corrupt_record"

- Q02 - What is the recommended storage format to use with Spark?
- A02 – Apache Parquet

Lab: Enabling Team Based Data Science with Azure Databricks

Lab overview

In this lab, By the end of this lab the student will be able to explain why Azure Databricks can be used to help in Data Science projects. The students will provision and Azure Databricks instance and will then create a workspace that will be used to perform a simple data preparation task from a Data Lake Store Gen II store. Finally, the student will perform a walk-through of performing transformations using Azure Databricks.

Lab objectives

After completing this lab, you will be able to:

1. Explain Azure Databricks
2. Work with Azure Databricks
3. Read data with Azure Databricks
4. Perform transformations with Azure Databricks

Lab scenario

In response to the Information Services (IS) department, you will start the process of building a predictive analytics platform by listing out the benefits of using the technology. The department will be joined by data scientists and they want to ensure that there is a predictive analytics environment available to the new team members.

You will stand up and provision an Azure Databricks environment, and then test that this environment works by performing a simple data preparation routine on the service by ingesting data from a pre-existing Data Lake Storage Gen II account. As a data engineer, it has been indicated to you that you may be required to help the data scientists perform data preparation exercises. To that end you have been recommended to walk-through a notebook that can help you perform basic transformations.

At the end of this lad, you will have:

1. Explain Azure Databricks
2. Work with Azure Databricks
3. Read data with Azure Databricks
4. Perform transformations with Azure Databricks

Lab review

- Exercise 1 – Why did you select the option to store the image files?
- Exercise 2 – Apart from the Azure Portal, are there other methods to automate the deployment of storage accounts?
- Exercise 3 – How is the storage structure of Data Lake Storage Gen II different to Storage Accounts?
- Exercise 4 – Where in the Azure Portal would you find Microsoft Azure Storage Explorer?

Module Summary ❯

In this module, you have learned about:

- Azure Databricks
- How to provision Azure Databricks and Workspaces
- How to read Data using Azure Databricks
- How to perform transformations with Azure Databricks

Next steps ❯

After the course, consider watching [this video with Yatharth Gupta that provides a deep dive into Azure Databricks deployment, networking and security]().