



Azure Databricks

Overview

- Module 1: Describe Azure Databricks

In this module, you will:

- Understand the Azure Databricks platform
- Create your own Azure Databricks workspace
- Create a notebook inside your home folder in Databricks
- Understand the fundamentals of Apache Spark notebook
- Create, or attach to, a Spark cluster
- Identify the types of tasks well-suited to the unified analytics engine Apache Spark

- Module 2: Spark architecture fundamentals

In this module, you will:

- Understand the architecture of an Azure Databricks Spark Cluster
- Understand the architecture of a Spark Job

- Module 3: Read and write data in Azure Databricks

In this module, you will:

- Use Azure Databricks to read multiple file types, both with and without a Schema.
- Combine inputs from files and data stores, such as Azure SQL Database.
- Transform and store that data for advanced analytics.

- Module 4: Work with DataFrames in Azure Databricks

In this module, you will:



Azure Databricks

- Use the `count()` method to count rows in a DataFrame
- Use the `display()` function to display a DataFrame in the Notebook
- Cache a DataFrame for quicker operations if the data is needed a second time
- Use the `limit` function to display a small set of rows from a larger DataFrame
- Use `select()` to select a subset of columns from a DataFrame
- Use `distinct()` and `dropDuplicates` to remove duplicate data
- Use `drop()` to remove columns from a DataFrame
- **Module 5: Describe lazy evaluation and other performance features in Azure Databricks**

In this module, you will:

- Describe the difference between eager and lazy execution
- Define and identify transformations
- Define and identify actions
- Describe the fundamentals of how the Catalyst Optimizer works
- Differentiate between wide and narrow transformations
- **Module 6: Work with DataFrames columns in Azure Databricks**

In this module, you will:

- Learn the syntax for specifying column values for filtering and aggregations
- Understand the use of the Column Class
- Sort and filter a DataFrame based on Column Values
- Use `collect()` and `take()` to return records from a Dataframe to the driver of the cluster



Azure Databricks

- Module 7: Work with DataFrames advanced methods in Azure Databricks

In this module, you will:

- Manipulate date and time values in Azure Databricks
- Rename columns in Azure Databricks
- Aggregate data in Azure Databricks DataFrames
- Module 8: Describe platform architecture, security, and data protection in Azure Databricks

In this module, you will:

- Learn the Azure Databricks platform architecture and how it is secured.
- Use Azure Key Vault to store secrets used by Azure Databricks and other services.
- Access Azure Storage with Key Vault-based secrets.
- Module 9: Build and query a Delta Lake

In this module, you will:

- Learn about the key features and use cases of Delta Lake.
- Use Delta Lake to create, append, and upsert tables.
- Perform optimizations in Delta Lake.
- Compare different versions of a Delta table using Time Machine.
- Module 10: Process streaming data with Azure Databricks structured streaming

In this module, you will:

- Learn the key features and uses of Structured Streaming.



Azure Databricks

- Stream data from a file and write it out to a distributed file system.
- Use sliding windows to aggregate over chunks of data rather than all data.
- Apply watermarking to throw away stale old data that you do not have space to keep.
- Connect to Event Hubs read and write streams.
- **Module 11: Describe Azure Databricks Delta Lake architecture**
In this module, you will:
 - Process batch and streaming data with Delta Lake.
 - Learn how Delta Lake architecture enables unified streaming and batch analytics with transactional guarantees within a data lake.
- **Module 12: Create production workloads on Azure Databricks with Azure Data Factory**
In this module, you'll:
 - Create an Azure Data Factory pipeline with a Databricks activity.
 - Execute a Databricks notebook with a parameter.
 - Retrieve and log a parameter passed back from the notebook.
 - Monitor your Data Factory pipeline.
- **Module 13: Implement CI/CD with Azure DevOps**
In this module, you will:
 - Learn about CI/CD and how it applies to data engineering.
 - Use Azure DevOps as a source code repository for Azure Databricks notebooks.



Azure Databricks

- Create build and release pipelines in Azure DevOps to automatically deploy a notebook from a development to a production Azure Databricks workspace.
- Module 14: Integrate Azure Databricks with Azure Synapse
In this module, you will:
 - Access Azure Synapse Analytics from Azure Databricks by using the - SQL Data Warehouse connector.
- Module 15: Describe Azure Databricks best practices
In this module, you will learn best practices in the following categories:
 - Workspace administration
 - Security
 - Tools & integration
 - Databricks runtime
 - HA/DR
 - Clusters