

# Databricks Academy

# Course Catalog

UPDATED: JUNE 2022

<b>Welcome to the Databricks Academy</b>	<b>5</b>
About the Databricks Academy	5
About this course catalog	5
What's new/coming this quarter	6
What's being retired/replaced this quarter	7
<b>Databricks Academy offerings</b>	<b>9</b>
Training	9
Credentials	10
Learning paths	11
Databricks Lakehouse fundamentals	12
Data analysis	12
Data engineering	13
Machine learning	14
<b>Certification exam descriptions</b>	<b>17</b>
<b>Instructor-led course descriptions</b>	<b>17</b>
<b>Self-paced course descriptions</b>	<b>17</b>
Apache Spark Programming with Databricks	17
AWS Databricks Cloud Architecture and System Integration Fundamentals	17
AWS Databricks Cluster Usage Management	18
AWS Databricks Data Access Management	19

AWS Databricks Identity Access Management	20
AWS Databricks Security Fundamentals	21
AWS Databricks SQL Administration	22
AWS Databricks Workspace Deployment	22
Azure Databricks Cloud Architecture and System Integration Fundamentals	23
Azure Databricks Cluster Usage Management	24
Azure Databricks Data Access Management	25
Azure Databricks Identity Access Management	26
Azure Databricks Security Fundamentals	26
Azure Databricks SQL Administration	27
Azure Databricks Workspace Deployment	28
Certification Overview for the Databricks Certified Professional Data Engineer Exam	29
Certification Prep Course for the Databricks Certified Associate Developer for Apache Spark Exam	30
Configuring Workspace Access Control Lists (ACLs)	31
Data Analysis with Databricks SQL	31
Data Engineering with Databricks	32
Databricks Command Line Interface (CLI) Fundamentals	32
Databricks Datadog Integration	33
Databricks on Google Cloud: Architecture and Security Fundamentals	33
Databricks on Google Cloud: Cloud Architecture and System Integration	34
Databricks on Google Cloud: Cluster Usage Management	35
Databricks on Google Cloud: Workspace Deployment	36
Databricks with R	36
Delta Lake Rapid Start with Python	37
Deploying a Machine Learning Project with MLflow Projects	38
Easy ETL with Auto Loader	39

Enterprise Architecture with Databricks	39
Fundamentals of the Databricks Lakehouse Platform Accreditation	40
What is Databricks Machine Learning?	41
Getting Started with Databricks Data Science & Engineering Workspace	41
Getting Started with Databricks Machine Learning	42
Getting Started with Databricks SQL	43
Google Cloud Fundamentals	44
How to Ingest Data for Databricks SQL	44
Introduction to Apache Spark Architecture	47
Introduction to Applied Linear Models	48
Introduction to Applied Statistics	49
Introduction to Applied Tree-Based Models	49
Introduction to Applied Unsupervised Learning	50
Introduction to Cloning with Delta Lake	51
Introduction to Databricks Connect	52
Introduction to Databricks Repos	52
Introduction to Delta Lake	53
Introduction to Delta Live Tables	54
Introduction to Feature Engineering and Selection with Databricks	54
Introduction to Files in Databricks Repos	55
Introduction to Hyperparameter Optimization	56
Introduction to Jobs	57
Introduction to MLflow Model Registry	58
Introduction to MLflow Tracking	58
Introduction to Natural Language Processing	59
Introduction to Photon	60
Just Enough Python for Apache Spark	61

Migrating SAS Procedures to Databricks	61
Natural Language Processing at Scale with Databricks	62
New Capability Overview: Feature Store	64
Optimizing Apache Spark on Databricks	65
Propagating Changes with Delta Change Data Feed	66
Quick Reference: CI/CD	66
Quick Reference: Spark Architecture	67
Scaling Machine Learning Pipelines	68
Scalable Machine Learning with Apache Spark	69
Structured Streaming	69
Tracking Experiments with MLflow	70
What are Enterprise Data Management Systems?	70
What is Big Data?	71
What is Cloud Computing?	72
What is Databricks Machine Learning?	73
What is Databricks SQL?	73
What is Delta Lake?	74
What is Machine Learning?	75
What is Structured Streaming?	76
What is the Databricks Lakehouse Platform?	77
What's New in Apache Spark 3.0	78

# Welcome to the Databricks Academy

## About the Databricks Academy

Our mission at the Databricks Academy is to help our customers achieve their big data and analytics goals through engaging learning experiences. At Databricks, professionals from a wide variety of disciplines come together and use modern pedagogical techniques to develop training that showcases Databricks best practices. We offer our customers a wide range of materials to meet their diverse training needs – whether they want to study at home, participate in a traditional classroom setting, or engage with other Databricks users in public online courses – to grow professionally with cloud-native skills.

## About this course catalog

This course catalog is broken into the following categories:

- **Welcome to the Databricks Academy:** information about the Databricks Academy and the students we serve
- **What's new/being retired this quarter:** a list of the recently released training materials/materials being retired and removed from the Academy
- **Databricks Academy offerings:** an explanation of the types of learning content we offer
- **Course descriptions:** short descriptions for each course available through the Databricks Academy

# What's **new/coming** this quarter

## **MAY 2022 (release date)**

- Advanced Data Engineering with Databricks (SP beta version) (5/20)
- Data Analysis with Databricks (ILT/SP) (5/20)
- Machine Learning Associate Certification (Certification exam) (5/20)
- Machine Learning Professional Certification (Certification exam) (5/20)
- New Capability Overview: Graviton-Enabled Clusters (5/20)
- New Capability Overview: Unity Catalog (5/20)

## **JUNE 2022 (release date)**

- Advanced Data Engineering with Databricks (Version 2) (SP) (6/17)
- Data Analyst Associate Certification Overview (SP) (6/22)
- Data Engineer Associate Certification Overview (SP) (6/22)
- Data Engineer Professional Certification Overview (SP) (6/22)
- Data Engineering with Databricks (Version 2) (SP) (6/3)
- Databricks Lakehouse Platform Fundamentals: What is the Databricks Lakehouse Platform? (Version 2) (SP) (6/3)
- Databricks Lakehouse Platform Fundamentals: What is the Databricks Data Science and Data Engineering Workspace? (Version 2) (SP) (6/3)
- Databricks Lakehouse Platform Fundamentals: What is Databricks SQL? (Version 2) (SP) (6/3)
- Databricks Lakehouse Platform Fundamentals: What is Databricks Machine Learning? (Version 2) (SP) (6/3)
- Databricks Lakehouse Platform Fundamentals Accreditation (Version 2) (SP) (6/3)
- Databricks Academy Welcome Guide for Customers (SP) (6/10)
- Databricks Academy Welcome Guide for Partners (SP) (6/10)
- Databricks Academy Welcome Guide for Microsoft (SP) (6/10)
- How to Answer Business Questions with Statistics and Databricks SQL (SP) (6/10)
- Introduction to AutoML (Version 2) (SP) (6/10)
- Introduction to Feature Store (Version 2) (SP) (6/10)
- Machine Learning Associate Certification Overview (SP) (6/22)
- Machine Learning Professional Certification Overview (SP) (6/22)
- New Capability Overview: Delta Sharing (SP) (6/17)
- New Capability Overview: Databricks Workflows (SP) (6/17)
- New Capability Overview: Jobs in Repos (6/17)
- New Capability Overview: Nephos (6/17)
- Scalable Machine Learning with Apache Spark (SP) (6/17)
- Spark Associate Certification Overview (SP) (6/22)

## **JULY 2022 (release date)**

- Databricks Platform Administration with Unity Catalog (SP) (7/8)
- Databricks Platform Administration: Databricks Security Model (SP) (7/29)

- Databricks Platform Administration: Databricks Identity Management (SP) (7/29)
- Databricks Platform Administration: Databricks Account Administration (SP) (7/29)
- Databricks Platform Administration: Databricks Workspace Administration (SP) (7/29)
- Databricks Platform Administration: Databricks Repos (SP) (7/29)
- Databricks Platform Administration: Databricks Cluster Overview (SP) (7/29)
- Databricks Platform Administration: Databricks Cluster Governance (SP) (7/29)
- Databricks Platform Administration: Customizing Databricks Cluster Environment (SP) (7/29)

## What's being **retired/replaced** this quarter

**JUNE 2022 (retire date)**

### **Data analysis courses**

- Basic SQL for Databricks SQL (SP) (6/15)
  - Updated content now found in *Data Analysis with Databricks SQL (ILT/SP)*
    - Module: Basic SQL on Databricks SQL
    - Lesson: Delta Commands in Databricks SQL
- Databases, Tables, and Views on Databricks SQL (6/15)
  - Updated content now found in *Data Analysis with Databricks SQL (ILT/SP)*
    - Module: Getting Started with Databricks SQL
    - Lesson(s): Unity Catalog on Databricks SQL, Schemas, Tables, and Views on Databricks SQL
- Introduction to SQL on Databricks (6/15)
  - Updated content now found in *Data Analysis with Databricks SQL (ILT/SP)*
    - Module: Getting Started with Databricks SQL
    - Lesson: Getting Started with Databricks SQL, Navigating Databricks SQL
- Dashboards on Databricks SQL (6/15)
  - Updated content now found in *Data Analysis with Databricks SQL (ILT/SP)*
    - Module: Presenting Data Visually
    - Lesson: Dashboards on Databricks SQL
- Data Visualization on Databricks SQL (6/15)
  - Updated content now found in *Data Analysis with Databricks SQL (ILT/SP)*
    - Module: Presenting Data Visually
    - Lesson: Data Visualization on Databricks SQL
- How to Ingest Data for Databricks SQL (6/15)
  - Updated content now found in *Data Analysis with Databricks SQL (ILT/SP)*
    - Module: Basic SQL on Databricks SQL
    - Lesson: Ingesting Data
- SQL Analyst Accreditation

- Content will be replaced by *Data Analyst Associate Certification (Certification exam - release date 6/22)*
- SQL Coding Challenges (6/22)
  - Content will no longer be supported.

### Data engineering courses

- Databases, Tables, and Views on Databricks (6/15)
  - Updated content now found in *Data Engineering with Databricks (ILT/SP)*
    - Module: ELT with Spark SQL and Python
    - Lesson: Relational Entities on Databricks
- Delta Lake Rapid Start with Spark SQL (6/15)
  - Content replaced by *Data Engineering with Databricks (ILT/SP)*
    - Module: ELT with Spark SQL and Python
- ELT with Spark SQL (6/15)
  - Updated content now found in *Data Engineering with Databricks (ILT/SP)*
    - Module: ELT with Spark SQL and Python
- Lakehouse with Delta Lake Deep Dive (6/15)
  - Content replaced by *Data Engineering with Databricks (ILT/SP)*
- Quick Reference: CI/CD (6/15)
  - Content will be located here.
- What is Delta Lake (6/15)
  - Updated content will now be included in Introduction to Delta Lake

### Data science/ machine learning courses

- Data Science on Databricks Rapidstart (6/15)
  - Content will be replaced by *Getting Started with Databricks Machine Learning (SP)*
- Data Science with Databricks: The Bias Variance Tradeoff (6/22)
  - Content will no longer be supported.
- Deploying a Machine Learning Project with MLflow Projects (6/22)
  - Content will be replaced by *Scalable Machine Learning with Apache Spark (SP - release date 6/22)*
- Introduction to Applied Linear Models (6/15)
  - Content will no longer be supported.
- Introduction to Feature Engineering and Selection with Databricks (6/15)
  - Content will no longer be supported.
- Introduction to MLflow Tracking (6/22)
  - Content will be replaced by *Scalable Machine Learning with Apache Spark (SP - release date 6/22)*
- Tracking Experiments with MLflow



- Content will be replaced by *Scalable Machine Learning with Apache Spark (SP - release date 6/22)*

### Miscellaneous topics

- Enterprise Architecture with Databricks (6/15)
  - Content will no longer be supported.
- Introduction to Delta Live Tables (6/15)
  - Content will be included in *New Capability Overview: Databricks Workflows (SP - release date 6/15)*
- Introduction to Multi-Task Jobs (6/15)
  - Content will be replaced by *New Capability Overview: Databricks Workflows (SP - release date 6/15)*

## Databricks Academy offerings

### Training

**Self-paced online courses** – asynchronous virtual training available to individuals through the Databricks Academy website. This training is free for Databricks customers. Each course is typically 1-2 hours in length.

**Workshops** – live 1-3 hour trainings made available to groups, typically in a virtual format. Please reach out to a CSE / Databricks Account manager to request a Workshop.

**Instructor-led trainings** – one to two days of content offered over 2 or 4 half-days. Available to everyone – customers and the public, for a fee. Delivered virtually.

**Accreditations/Certifications** – 30 minute unproctored quizzes to 2 hour proctored exams

## Training modalities

	Self-paced	Workshops	Instructor-led training	Certification
Associate-level	★	★	★	★
Professional-level	★	★	★	★
Technology focus	★	★	★	★
Role-focus	★		★	★
Live instructor		★	★	
Free to customers	★	★		
Customer size	All	All	All	All
Duration	Minutes – Hours	Hours	Days	Hours

©2022 Databricks Inc. — All rights reserved



## Credentials

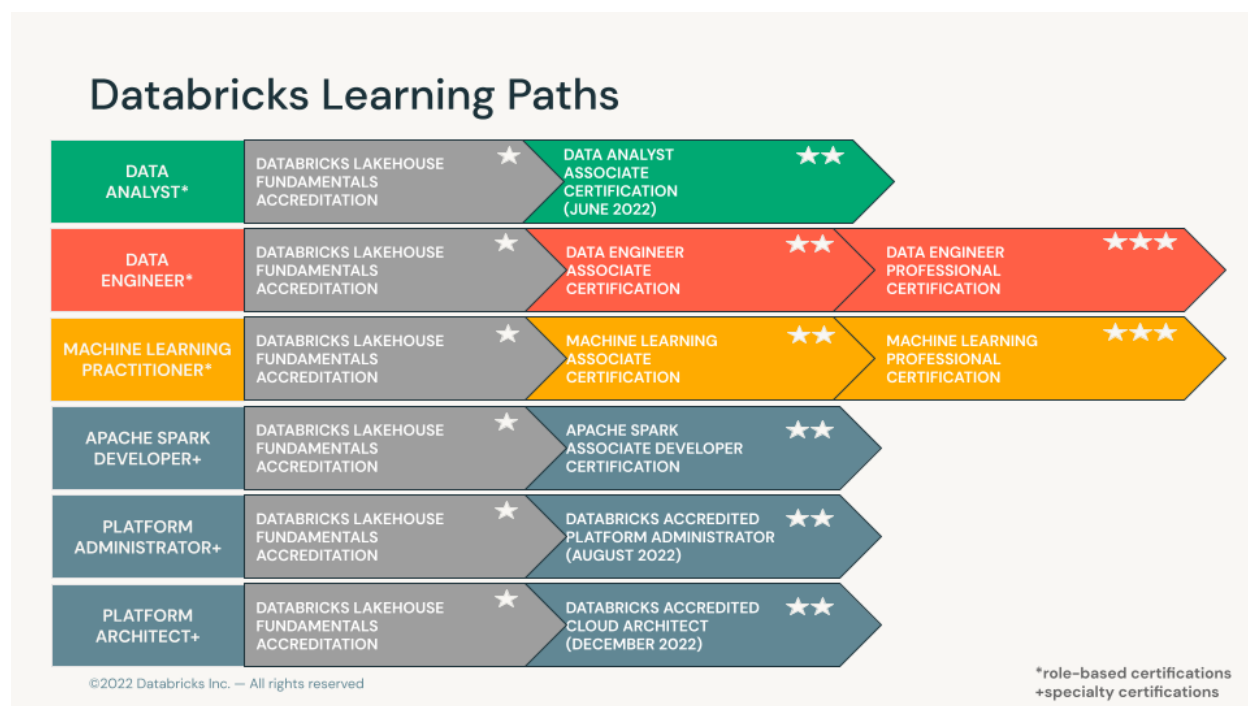
**Accreditations** – low stakes credentials resulting from an unproctored online exam administered through the Databricks Academy website. They are earned after demonstrating mastery of technology areas at the introductory level, and are in alignment with self-paced training.

**Certifications** – higher stakes credentials resulting from a proctored exam administered through a testing vendor. They are earned after demonstrating mastery of intermediate and advanced technical areas. They are in alignment with instructor-led training, and are role-based. Unlike accreditations, which are prepared for a general audience, certifications are designed to align with data practitioner roles (for example, a data engineer or a data analyst role).

# Learning paths

Learning paths are designed to help guide users to the courses most relevant to them.


Current pathways are available for Databricks fundamentals, data analysts, data engineers, machine learning practitioners, and Apache Spark. The credential milestones for each step within these pathways are shown in the images below.



Below, you'll find a breakdown of the courses required for each of these steps. We will update these regularly, as new courses are released.

## Databricks Lakehouse fundamentals

# Databricks Lakehouse Fundamentals




**DATABRICKS LAKEHOUSE FUNDAMENTALS ACCREDITATION**

- What is the Databricks Lakehouse Platform?
- What is Delta Lake?
- What is Databricks SQL?
- What is Databricks Machine Learning?
- Databricks Lakehouse Fundamentals Accreditation**



**Note:** Courses and accreditation available as self-paced content in the Databricks Fundamentals Learning Path via Databricks Academy.

©2022 Databricks Inc. — All rights reserved



## Data analysis

# Data Analyst

**DATABRICKS LAKEHOUSE FUNDAMENTALS ACCREDITATION** ★


**DATA ANALYST ASSOCIATE CERTIFICATION (JUNE 2022)** ★★

- What is the Databricks Lakehouse Platform? (SP)
- What is Delta Lake? (SP)
- What is Databricks SQL? (SP)
- What is Databricks Machine Learning? (SP)
- Databricks Lakehouse Fundamentals Accreditation (SP)**

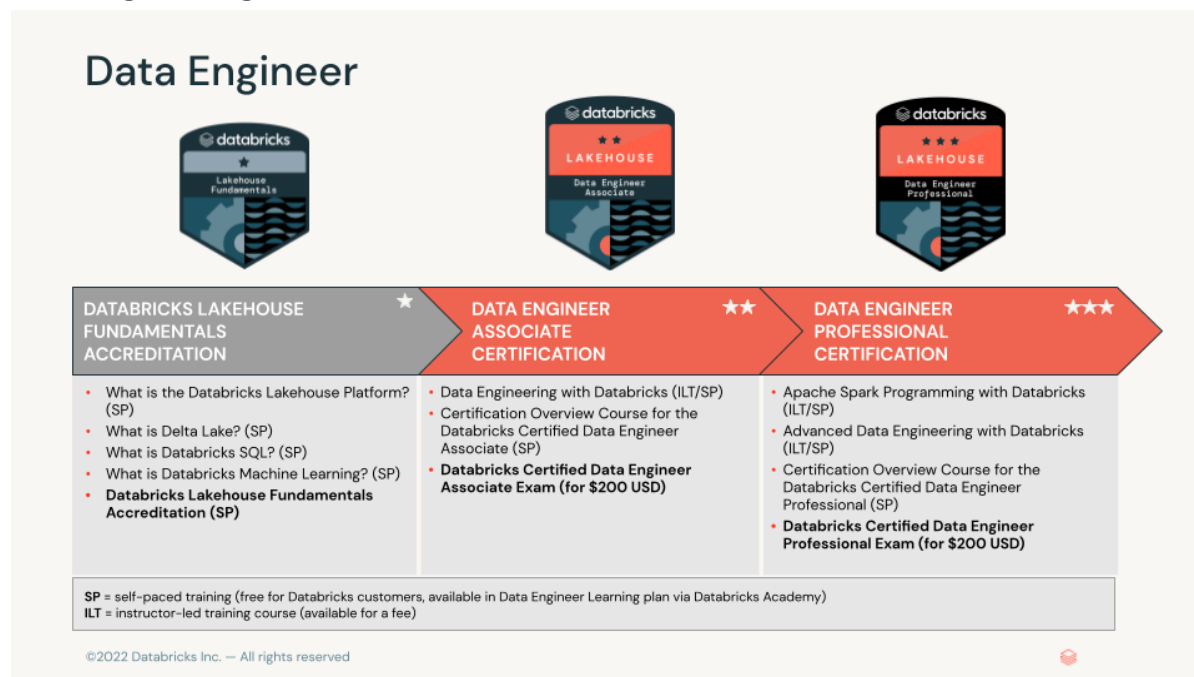
- Data Analysis with Databricks SQL (ILT/SP)
- Certification Overview Course for the Databricks Certified Data Analyst Associate (SP coming 27-June-2022)
- Databricks Certified Data Analyst Associate Exam**  
(Exam coming 27-June-2022 for \$200 USD)

SP = self-paced training (free for Databricks customers, available in Data Analyst Learning plan via Databricks Academy)  
 ILT = instructor-led training course (available for a fee)

©2022 Databricks Inc. — All rights reserved






## Data engineering



## Machine learning

### Machine Learning Practitioner



DATABRICKS LAKEHOUSE FUNDAMENTALS ACCREDITATION ★	MACHINE LEARNING ASSOCIATE CERTIFICATION ★★	MACHINE LEARNING PROFESSIONAL CERTIFICATION ★★★
<ul style="list-style-type: none"> <li>What is the Databricks Lakehouse Platform? (SP)</li> <li>What is Delta Lake? (SP)</li> <li>What is Databricks SQL? (SP)</li> <li>What is Databricks Machine Learning? (SP)</li> <li><b>Databricks Lakehouse Fundamentals Accreditation (SP)</b></li> </ul>	<ul style="list-style-type: none"> <li>Scalable Machine Learning with Apache Spark <i>(available now as ILT or as self-paced starting 22-June-2022)</i></li> <li>Certification Overview Course for the Databricks Certified Machine Learning Associate Exam <i>(SP coming 27-June-2022)</i></li> <li><b>Databricks Certified Machine Learning Associate Exam (\$200 USD)</b></li> </ul>	<ul style="list-style-type: none"> <li>Machine Learning in Production <i>(available now as ILT or as self-paced starting 22-June-2022)</i></li> <li>Certification Overview Course for the Databricks Certified Machine Learning Professional Exam <i>(SP coming 27-June-2022)</i></li> <li><b>Databricks Certified Machine Learning Professional Exam (\$200 USD)</b></li> </ul>

SP = self-paced training (free for Databricks customers, available in Machine Learning Practitioner Learning plan via Databricks Academy)  
 ILT = instructor-led training course (available for a fee)

©2022 Databricks Inc. — All rights reserved

## Apache Spark developer

### Apache Spark Developer

DATABRICKS LAKEHOUSE FUNDAMENTALS ACCREDITATION ★	APACHE SPARK ASSOCIATE DEVELOPER CERTIFICATION ★★
<ul style="list-style-type: none"> <li>What is the Databricks Lakehouse Platform? (SP)</li> <li>What is Delta Lake? (SP)</li> <li>What is Databricks SQL? (SP)</li> <li>What is Databricks Machine Learning? (SP)</li> <li><b>Databricks Lakehouse Fundamentals Accreditation (SP)</b></li> </ul>	<ul style="list-style-type: none"> <li>OPTIONAL: Just Enough Python for Apache Spark (ILT/SP)</li> <li>Apache Spark Programming with Databricks (ILT/SP)</li> <li>Certification Overview Course for the Databricks Associate Developer for Apache Spark Exam (SP)</li> <li><b>Databricks Certified Associate Developer for Apache Spark 3.0 (for \$200 USD)</b></li> </ul>

SP = self-paced training (free for Databricks customers, available in Apache Spark Learning plan via Databricks Academy)  
 ILT = instructor-led training course (available for a fee)

©2022 Databricks Inc. — All rights reserved

## Platform administration – (NOTE: NEW CONTENT COMING AUGUST 2022)

## Currently available Platform Admin content (Azure)

DATABRICKS LAKEHOUSE FUNDAMENTALS ACCREDITATION ★	AZURE DATABRICKS PLATFORM ADMINISTRATOR ASSOCIATE CERTIFICATION* ★★
<ul style="list-style-type: none"><li>• What is the Databricks Lakehouse Platform?</li><li>• What is Delta Lake?</li><li>• What is Databricks SQL?</li><li>• What is Databricks Machine Learning?</li></ul>	<ul style="list-style-type: none"><li>• Azure Databricks Cloud Architecture and System Integration Fundamentals</li><li>• Azure Databricks Workspace Deployment</li><li>• Azure Databricks Security Fundamentals</li><li>• Azure Databricks Identity Access Management</li><li>• Azure Databricks Data Access Management</li><li>• Azure Databricks Cluster Usage Management</li><li>• Azure Databricks SQL Administration</li></ul>
<b>Note:</b> These courses are available as self-paced only.	



## Currently available Platform Admin content (AWS)

DATABRICKS LAKEHOUSE FUNDAMENTALS ACCREDITATION	★ CONTENT CURRENTLY AVAILABLE FOR AWS DATABRICKS PLATFORM ADMIN
<ul style="list-style-type: none"><li>• What is the Databricks Lakehouse Platform?</li><li>• What is Delta Lake?</li><li>• What is Databricks SQL?</li><li>• What is Databricks Machine Learning?</li></ul>	<ul style="list-style-type: none"><li>• AWS Databricks Cloud Architecture and System Integration Fundamentals</li><li>• AWS Databricks Workspace Deployment</li><li>• AWS Databricks Security Fundamentals</li><li>• AWS Databricks Identity Access Management</li><li>• AWS Databricks Data Access Management</li><li>• AWS Databricks Cluster Usage Management</li><li>• AWS Databricks SQL Administration</li></ul>
<b>Note:</b> These courses are available as self-paced only.	



## Currently available Platform Admin content (Google Cloud)

DATABRICKS LAKEHOUSE FUNDAMENTALS ACCREDITATION	★ CONTENT CURRENTLY AVAILABLE FOR DATABRICKS ON GOOGLE CLOUD PLATFORM
<ul style="list-style-type: none"><li>• What is the Databricks Lakehouse Platform?</li><li>• What is Delta Lake?</li><li>• What is Databricks SQL?</li><li>• What is Databricks Machine Learning?</li></ul>	<ul style="list-style-type: none"><li>• Databricks on Google Cloud: Workspace Deployment</li><li>• Google Cloud Fundamentals</li><li>• Databricks on Google Cloud: Architecture and Security Fundamentals</li><li>• Databricks on Google Cloud: Cloud Architecture and System Integration</li><li>• Databricks on Google Cloud: Cluster Usage Management</li></ul>
<b>Note:</b> These courses are available as self-paced only.	





## Certification exam descriptions

For a full list of available certification exams, along with their descriptions, please [click here](#).

## Instructor-led course descriptions

For a full list of available instructor-led courses, along with their descriptions, please [click here](#).

## Self-paced course descriptions

Note: All self-paced courses are free for Databricks customers. Non-customers can purchase some courses through role-based learning plans available via the Databricks Academy.

### Apache Spark Programming with Databricks

[Click here](#) for the customer enrollment link.

Duration: 12 hours

NOTE: This is an e-learning version of the Apache Spark Programming with Databricks instructor-led course. It is an on-demand recording available via the Databricks Academy and covers the same content as the instructor-led course. For more information about what's in the course itself, please [visit this link](#).

### AWS Databricks Cloud Architecture and System Integration Fundamentals

[Click here](#) for the customer enrollment link.

Duration: 1 hour

Course description: While the Databricks Unified Analytics Platform provides a broad range of functionality to many members of data teams, it is through integrations with other services that most cloud-native applications will achieve results desired by customers. This course is a series of demos designed to help students understand the portions of cloud workloads appropriate for Databricks. Within these demos, we'll highlight integrations with first-party services in the AWS cloud to build scalable and secure applications.

Prerequisites:

- Beginning knowledge of Spark programming (reading/writing data, batch and streaming jobs, transformations and actions)
- Beginning-level experience using Python or Scala to perform basic control flow operations.
- Familiarity with navigation and resource configuration in the AWS Console.

Learning objectives:

- Describe use cases for Databricks in an enterprise cloud architecture.
- Configure secure connections from Databricks to data in S3.
- Configure connections between Databricks and various first-party tools in an enterprise cloud architecture, including Redshift and Kinesis.
- Deploy an MLflow model to a Sagemaker endpoint for serving online model predictions.
- Configure Glue as an enterprise data catalog

## AWS Databricks Cluster Usage Management

[Click here](#) for the customer enrollment link.

Duration: 1.5 hours

Course description: In this course, you will first define computation resources (clusters, jobs, and pools) and determine which resources to use for different workloads. Then, you will learn cluster provisioning strategies for several use cases to maximize usability and cost-effectiveness. You will also identify best practices for

cluster governance, including cluster policies. This course also covers capacity limits, cost management, and chargeback analysis.

Prerequisites:

- Beginning experience using the Databricks workspace
- Beginning experience with Databricks administration

Learning objectives:

- Define computation resources (clusters, jobs, and pools) and determine which resources to use for different workloads.
- Describe cluster provisioning strategies for several use cases to maximize usability and cost effectiveness.
- Identify best practices for cluster governance, including cluster policies.
- Describe capacity limits on Azure Databricks.
- Describe how to manage costs and perform chargeback analysis.

## AWS Databricks Data Access Management

[Click here](#) for the customer enrollment link.

Duration: 1 hour

Course description: In this course, you will learn about the Databricks File System and Hive Metastore concepts. Then, you will apply best practices to secure access to Amazon S3 from Databricks. Next, you will configure access control for data objects including tables, databases, views, and functions. You will also apply column and row-level permissions and data masking with dynamic views for multiple users and groups. Lastly, you will identify methods for data isolation within your organization on Databricks.

Prerequisites:

- Beginning experience with AWS Databricks security, including deployment architecture and encryptions
- Beginning experience with AWS Databricks administration, including identity management and workspace access control
- Beginning experience using the Databricks workspace

- Databricks Premium Plan

Learning objectives:

- Describe fundamental concepts about the Databricks File System and Hive Metastore.
- Apply best practices to secure access to Amazon S3 from Databricks.
- Configure access control for data objects including tables, databases, views, and functions.
- Apply column and row-level permissions and data masking with dynamic views for multiple users and groups.
- Identify methods for data isolation within your organization on Databricks.

## AWS Databricks Identity Access Management

[Click here](#) for the customer enrollment link.

Duration: 1.5 hours

Course description: In this course, you will learn how to manage user accounts and groups in the Admin Console. You will also learn how to manage token-based authentication and settings for your workspace, such as workspace storage and additional cluster configurations. Lastly, this course covers access control for workspace objects, such as notebooks and folders, in addition to clusters, pools, and jobs.

Prerequisites:

- Experience using a web browser.
- Note: To perform the tasks shown in this course, you will need a Databricks workspace deployment with administrator rights.

Learning objectives:

- Manage user accounts and groups in the Admin Console.
- Generate and manage personal access tokens for authentication.
- Enable additional cluster configurations and purge deleted objects from workspace storage.

- Configure access control for workspace objects, such as notebooks and folders.
- Configure access control for clusters, pools, and jobs.

# AWS Databricks Security Fundamentals

[Click here](#) for the customer enrollment link.

Duration: 1 hour

Course description: This course provides an overview of essential security features to consider when managing your AWS Databricks workspace. You will start by identifying components of the Databricks platform architecture and deployment models. Then, you will define several features regarding network security including no public IPs, Bring Your Own VPC, VPC peering, and IP access lists. After recognizing IdP integrations, you will explore access control configurations for different workspace assets. You will then identify encryptions and permissions available for data protection, such as IdP authentication, secrets, and table access control. Lastly, you will describe security standards and configurations for compliance, including cluster policies, Bring Your Own Key, and audit logs.

Prerequisites:

- Beginning-level knowledge of basic AWS cloud computing terms (ex. S3, VPC, IAM, etc.)
- Beginning-level knowledge of basic Databricks concepts (ex. workspace, clusters, notebooks, etc.)

Learning objectives:

- Describe components of the AWS Databricks platform architecture and deployment model.
- Explain network security features including no public IP address, Bring Your Own VPC, VPC peering, and IP access lists.
- Describe identity provider integrations and access control configurations for an AWS Databricks workspace.
- Explain encryptions and permissions available for data protection, such as identity provider authentication, secrets, and table access control.

- Describe security standards and configurations for compliance, including cluster policies, Bring Your Own Key, and audit logs.

## AWS Databricks SQL Administration

[Click here](#) for the customer enrollment link.

Duration: 1 hour

Course description: In this course, you will learn how to set up and configure access to the Databricks SQL Analytics user interface. The administrative tasks in this course will be done using the Databricks Workspace and Databricks SQL Analytics UI, and will not include instruction for API access. By the end of this course, you will be able to set up computational resources for users, grant and revoke access to specific data, manage users and groups, and set up alert destinations.

Prerequisites:

- Intermediate knowledge of Databricks
- Databricks account on the Premium plan (with SQL Analytics enabled)
- Administrator credentials to your organization's Databricks Workspace

Learning objectives:

- Describe how Databricks SQL Analytics is used by data practitioners.
- Manage user and group access to Databricks SQL Analytics.
- Configure and monitor SQL Endpoints to maximize performance, control costs, and track usage on Databricks SQL Analytics.
- Set up access to data storage through SQL endpoints or external data stores in order for users to access data on Databricks SQL Analytics.
- Control user access to data objects (e.g. tables, databases, and views) by programmatically setting privileges for specific users and/or groups on Databricks SQL Analytics.
- Create and configure Databricks SQL Analytics alert destinations for users.

## AWS Databricks Workspace Deployment

[Click here](#) for the customer enrollment link.

Duration: 1 hour

Course description: This course will walk you through setting up your Databricks account including setting up billing, configuring your AWS account, and adding users with appropriate permissions. At the end of this course, you'll find guidance and resources for additional setup options and best practices.

Prerequisites:

- Experience using a web browser.
- Note: To follow along with this course, you will need access to a Databricks account with Account Owner permissions.

Learning objectives:

- Access the Databricks account console and set up billing.
- Configure an AWS account using cross-account role or access keys.
- Configure AWS storage and deploy the Databricks workspace.
- Add users and assign admin or cluster creation rights.
- Identify resources for additional setup options and best practices.

## Azure Databricks Cloud Architecture and System Integration Fundamentals

[Click here](#) for the customer enrollment link.

Duration: 1 hour

Course description: While the Databricks Unified Analytics Platform provides a broad range of functionality to many members of data teams, it is through integrations with other services that most cloud-native applications will achieve results desired by customers. This course is designed to help students understand the portions of cloud workloads appropriate for Databricks, and highlight integrations with first-party services in the Azure cloud to build scalable and secure applications.

Prerequisites:

- Beginning knowledge of Spark programming (reading/writing data, batch and streaming jobs, transformations and actions)

- Beginning-level experience using Python or Scala to perform basic control flow operations.
- Familiarity with navigation and resource configuration in the Azure Portal.

Learning objectives:

- Describe use-cases for Azure Databricks in an enterprise cloud architecture.
- Configure secure connections to data in an Azure storage account.
- Configure connections from Databricks to various first-party tools, including Synapse, Key Vault, Event Hubs, and CosmosDB.
- Configure Azure Data Factory to trigger production jobs on Databricks.
- Trigger CI/CD workloads on Databricks assets using Azure DevOps.

## Azure Databricks Cluster Usage Management

[Click here](#) for the customer enrollment link.

Duration: 1 hour

Course description: In this course, you will first define computation resources (clusters, jobs, and pools) and determine which resources to use for different workloads. Then, you will learn cluster provisioning strategies for several use cases to maximize usability and cost effectiveness. You will also identify best practices for cluster governance, including cluster policies. This course also covers capacity limits, cost management, and chargeback analysis.

Prerequisites:

- Beginning experience with the Databricks workspace UI
- Beginning experience with Databricks administration

Learning objectives:

- Define computation resources (clusters, jobs, and pools) and determine which resources to use for different workloads.
- Describe cluster provisioning strategies for several use cases to maximize usability and cost effectiveness.
- Identify best practices for cluster governance, including cluster policies.
- Describe capacity limits on Azure Databricks.



- Describe how to manage costs and perform chargeback analysis.

# Azure Databricks Data Access Management

[Click here](#) for the customer enrollment link.

Duration: 1 hour

Course description: In this course, you will learn about the Databricks File System and Hive Metastore concepts. Then, you will apply best practices to secure access to Azure data storage from Azure Databricks. Next, you will configure access control for data objects including tables, databases, views, and functions. You will also apply column and row-level permissions and data masking with dynamic views for multiple users and groups. Lastly, you will identify methods for data isolation within your organization on Azure Databricks.

Prerequisites:

- Beginning experience with Azure Databricks security, including deployment architecture and encryptions
- Beginning experience with Azure Databricks administration, including identity management and workspace access control
- Beginning experience using the Azure Databricks workspace
- Azure Databricks Premium Plan

Learning objectives:

- Describe the Databricks File System and Hive Metastore concepts.
- Apply best practices to secure access to Azure data storage from Azure Databricks.
- Configure access control for data objects including tables, databases, views, and functions.
- Apply column and row-level permissions and data masking with dynamic views for multiple users and groups.
- Identify methods for data isolation within your organization on Azure Databricks.

# Azure Databricks Identity Access Management

[Click here](#) for the customer enrollment link.

Duration: 45 minutes

Course description: In this course, you will learn how to manage user accounts and groups in the Admin Console. You will also learn how to manage token-based authentication and settings for your workspace, such as workspace storage and additional cluster configurations. Lastly, this course covers access control for workspace objects, such as notebooks and folders, in addition to clusters, pools, and jobs.

Prerequisites:

- Beginning experience using the Databricks workspace.

Learning objectives:

- Manage user accounts and groups in the Admin Console.
- Generate and manage personal access tokens for authentication.
- Enable additional cluster configurations and purge deleted objects from workspace storage.
- Configure access control for workspace objects, such as notebooks and folders.
- Configure access control for clusters, pools, and jobs.

# Azure Databricks Security Fundamentals

[Click here](#) for the customer enrollment link.

Duration: 1.5 hours

Course description: This course provides an overview of essential security features to consider when managing your Azure Databricks workspace. You will start by identifying components of the Azure Databricks platform architecture and deployment model. Then, you will define several features regarding network security including no public IPs, Bring Your Own VNET, VNET peering, and IP access lists. After

recognizing IdP and AAD integrations, you will explore access control configurations for different workspace assets. You will then identify encryptions and permissions available for data protection, such as IdP authentication, secrets, and table access control. Lastly, you will describe security standards and configurations for compliance, including cluster policies, Bring Your Own Key, and audit logs.

Prerequisites:

- Beginning-level knowledge of basic Azure cloud computing terms (ex. Blob storage, ADLS, VNET, Azure Active Directory, etc.)
- Beginning-level knowledge of basic Databricks concepts (ex. workspace, clusters, notebooks, etc.)

Learning objectives:

- Describe components of the Azure Databricks platform architecture and deployment model.
- Explain network security features including no public IP address, Bring Your Own VNET, VNET peering, and IP access lists.
- Describe identity provider and Azure Active Directory integrations and access control configurations for an Azure Databricks workspace.
- Explain encryptions and permissions available for data protection, such as identity provider authentication, secrets, and table access control.
- Describe security standards and configurations for compliance, including cluster policies, Bring Your Own Key, and audit logs.

## Azure Databricks SQL Administration

[Click here](#) for the customer enrollment link.

Duration: 1 hour

Course description: In this course, you will learn how to set up and configure access to the Databricks SQL Analytics user interface. The administrative tasks in this course will be done using the Databricks Workspace and Databricks SQL Analytics UI, and will not include instruction for API access. By the end of this course, you will be able to set up computational resources for users, grant and revoke access to specific data, manage users and groups, and set up alert destinations.

Prerequisites:

- Intermediate knowledge of Databricks
- Databricks account on the Premium plan (with SQL Analytics enabled)
- Administrator credentials to your organization's Databricks Workspace

Learning objectives:

- Describe how Databricks SQL Analytics is used by data practitioners.
- Manage user and group access to Databricks SQL Analytics.
- Configure and monitor SQL Endpoints to maximize performance, control costs, and track usage on Databricks SQL Analytics.
- Set up access to data storage through SQL endpoints or external data stores in order for users to access data on Databricks SQL Analytics.
- Control user access to data objects (e.g. tables, databases, and views) by programmatically setting privileges for specific users and/or groups on Databricks SQL Analytics.
- Create and configure Databricks SQL Analytics alert destinations for users.

## Azure Databricks Workspace Deployment

[Click here](#) for the customer enrollment link.

Duration: 10 minutes

Course description: In this course, you will identify the prerequisites for creating an Azure Databricks workspace, deploy an Azure Databricks workspace in the Azure portal, launch the workspace, and access the Admin Console.

Prerequisites:

- To complete the actions outlined in this course, you must have access to an Azure subscription.

Learning objectives:

- Identify prerequisites for launching an Azure Databricks workspace.
- Deploy an Azure Databricks workspace in the Azure portal.
- Launch the deployed Azure Databricks workspace.
- Access the Admin Console in the deployed Azure Databricks workspace.

# Certification Overview for the Databricks Certified Professional Data Engineer Exam

[Click here](#) for the customer enrollment link.

Duration: 2 hours

Course description: This course will cover the format and structure of the exam, skills needed for the exam, tips for exam preparation, and the topics covered in the exam.

Prerequisites:

- Describe how to use and the benefits of the Databricks platform and developer tools
- Build optimized and cleaned data processing pipelines using the Spark and Delta Lake APIs
- Model data into a Lakehouse using knowledge of general data modeling concepts
- Ensure data pipelines secure, reliable, monitored, and tested before deployment

Learning objectives:

- Describe logistical information about registering and sitting for the Databricks Certified Professional Data Engineer exam.
- Describe the format and structure of the Databricks Certified Professional Data Engineer exam.
- Describe the topics covered in the Databricks Certified Professional Data Engineer exam.
- Recognize the types of questions provided in the Databricks Certified Professional Data Engineer exam and apply test-taking strategies to answer example questions.
- Identify resources that can be used to learn the material covered in the Databricks Certified Professional Data Engineer exam.

# Certification Prep Course for the Databricks Certified Associate Developer for Apache Spark Exam

[Click here](#) for the customer enrollment link.

Duration: 2 hours

Course description: Prepare to take the Databricks Certified Associate Developer for Apache Spark Exam. This course will cover the format and structure of the exam, skills needed for the exam, tips for exam preparation, and the parts of the DataFrame API and Spark architecture covered in the exam.

Prerequisites:

- Describe the basics of the Apache Spark architecture.
- Perform basic data transformations using the Apache Spark DataFrame API using Python or Scala.
- Perform basic data input and output using the Apache Spark DataFrame API using Python or Scala.
- Perform custom data actions using user-defined functions using Python or Scala.
- Perform data transformations using Spark SQL.
- Note: While the above skills are not necessary for this course, the course will be far more helpful in preparing students if they have these skills.

Learning objectives:

- Summarize the learning context behind the Databricks Certified Associate Developer for Apache Spark exam.
- Describe the topics covered in the Databricks Certified Associate Developer for Apache Spark exam.
- Describe the format and structure of the Databricks Certified Associate Developer for Apache Spark exam.
- Apply practical test-taking strategies to answer example questions similar to those of the Databricks Certified Associate Developer for Apache Spark exam.

- Identify resources that can be used to learn the material covered in the Databricks Certified Associate Developer for Apache Spark exam.

## Configuring Workspace Access Control Lists (ACLs)

[Click here](#) for the customer enrollment link.

Duration: 1 hour

Course description: Databricks has extensive access control lists (ACLs) for workspace assets to help administrators restrict and grant access to appropriate users. This course includes a set of instructions and caveats for configuring many of these settings, as well as a video walkthrough showing this configuration and the resultant user experience.

Prerequisites:

- Basic knowledge of the Databricks workspace

Learning objectives:

- Manage permissions for groups of users.
- Control access to notebooks and folders.
- Restrict access for cluster creation and editing.
- Change ownership of configured jobs.

## Data Analysis with Databricks SQL

[Click here](#) for the customer enrollment link.

Duration: 6 hours

NOTE: This is an e-learning version of the Data Analysis with Databricks instructor-led course. It is an on-demand recording available via the Databricks Academy and covers the same content as the instructor-led course. For more information about what's in the course itself, please [visit this link](#).

# Data Engineering with Databricks

[Click here](#) for the customer enrollment link.

Duration: 12 hours

NOTE: This is an e-learning version of the Data Engineering with Databricks instructor-led course. It is an on-demand recording available via the Databricks Academy and covers the same content as the instructor-led course. For more information about what's in the course itself, please [visit this link](#).

## Databricks Command Line Interface (CLI) Fundamentals

[Click here](#) for the customer enrollment link.

Duration: 45 minutes

Course description: While the Databricks platform web-based graphical user interface provides powerful functionality for data teams, many use cases call for programmatic command line access. The Databricks command line interface (CLI) provides access to a variety of powerful workspace features. This module is not intended as a comprehensive overview of all the CLI can do, but rather an introduction to some of the common features users may desire to leverage in their workloads.

Prerequisites:

- Familiarity with Apache Spark concepts
- Familiarity with the data engineering capabilities of the Databricks Platform
- Intermediate experience using the Databricks platform for data engineering (creating clusters, loading notebooks, scheduling jobs, etc.)

Learning objectives:

- Install and configure the Databricks CLI to securely interact with the Databricks Workspace.
- Configure workspace secrets using the CLI for more secure sharing and use of string-based credentials in notebooks.



- Sync notebooks and libraries between the Databricks workspace and other environments using the CLI.
- Perform a variety of tasks including interacting with clusters, jobs, and runs using the CLI.

## Databricks Datadog Integration

[Click here](#) for the customer enrollment link.

Duration: 1 hour

Course description: Datadog provides customizable integration scripts and dashboards to integrate your Databricks logs into your larger monitoring ecosystem. This lesson goes through basic configuration, as well as extending this configuration to add additional security and custom tagging.

Prerequisites:

- Basic familiarity with the Databricks workspace
- Basic familiarity with cluster configuration

Learning objectives:

- Configure both ends of the Databricks Datadog integration.
- Add custom variables to your monitored clusters.
- Use Databricks secrets to redact API tokens.

## Databricks on Google Cloud: Architecture and Security Fundamentals

[Click here](#) for the customer enrollment link.

Duration: 1.5 hours

Course description: This course dives into the platform architecture and key security features of Databricks on Google Cloud. You will start with an overview of Databricks on Google Cloud and how it integrates with the Google Cloud ecosystem. Then, you will define core components of the platform architecture and deployment model on Databricks on Google Cloud. You will also learn about key security features

to consider when provisioning and managing workspaces, as well as guidelines on network security, identity and access management, and data protection.

### Prerequisites & Requirements

- Prerequisites
  - Basic familiarity with Databricks concepts (workspace, notebooks, clusters, DBFS, etc)
  - Basic familiarity with Google Cloud concepts (projects, IAM, GCS, VPC, subnets, GKE, etc)

### Learning objectives

- Describe how Databricks integrates with the Google Cloud ecosystem.
- Identify components of the Databricks on Google Cloud platform architecture and deployment model.
- Recognize best practices for network security when deploying workspaces.
- Describe identity management and access control features in Databricks on Google Cloud.
- Identify storage locations and data protection features in Databricks on Google Cloud.

# Databricks on Google Cloud: Cloud Architecture and System Integration

[Click here](#) for the customer enrollment link.

Duration: 1.5 hours

Course description: This course is a series of demos designed to help students understand the portions of cloud workloads appropriate for Databricks. Within these demos, we'll highlight integrations with first-party services in Google Cloud to build scalable and secure applications.

### Prerequisites & Requirements

- Prerequisites
  - Familiarity with the Databricks on Google Cloud workspace
  - Beginning knowledge of Spark programming (reading/writing data, batch and streaming jobs, transformations and actions)

- Beginning-level experience using Python or Scala to perform basic control flow operations.
- Familiarity with navigation and resource configuration in the Databricks on Google Cloud Console.

### Learning objectives

- Describe where Databricks fits into a cloud-based architecture on Google Cloud.
- Authenticate to Google Cloud resources with service account credentials.
- Read and write data to Cloud Storage using Databricks secrets.
- Mount a GCS bucket to DBFS using cluster-wide service accounts.
- Configure a cluster to read and write data to BigQuery using credentials in DBFS.

# Databricks on Google Cloud: Cluster Usage Management

[Click here](#) for the customer enrollment link.

Duration: 30 minutes

Course description: This course covers essential cluster configuration features and provisioning guidelines for Databricks on Google Cloud. In this course, you will start by defining core computation resources (clusters, jobs, and pools) and determine which resources to use for different workloads. Then, you will learn cluster provisioning strategies for several use cases to maximize manageability. Lastly, you will learn how to manage cluster usage and cost for your Databricks on Google Cloud workspaces.

### Prerequisites & Requirements

- Prerequisites
  - Beginning experience using the Databricks workspace
  - Beginning experience with Databricks administration

### Learning objectives

- Describe the core computation resources in Databricks, clusters, jobs, and pools.

- Recognize best practices for configuring cluster resources for different workloads.
- Identify cluster provisioning use cases and strategies for manageability.
- Describe how to manage cluster usage and cost for Databricks on Google Cloud.

# Databricks on Google Cloud: Workspace Deployment

[Click here](#) for the customer enrollment link.

Duration: 20 minutes

Course description: This is a short course that shows new customers how to set up a Databricks account and deploy a workspace on Google Cloud. This will cover accessing the Account Console and adding account admins, provisioning and accessing workspaces, and adding users and admins to a workspace.

## Prerequisites & Requirements

- Prerequisites
  - Basic familiarity with Databricks concepts (Databricks account, workspace, DBFS, etc)
  - Basic familiarity with Google Cloud concepts (Cloud console, project, GCS, IAM, VPC, etc)

## Learning objectives

- Access the Databricks Account Console.
- Add account admins in the Account Console.
- Provision and access a Databricks workspace.
- Access the Admin Console for a Databricks workspace.
- Add workspace users and admins in the Admin Console.

# Databricks with R

[Click here](#) for the customer enrollment link.

Duration: 7 hours

Course description: In this seven-hour course, you will analyze clickstream data from an imaginary mattress retailer called Bedbricks. In this case study, you'll explore the fundamentals of Spark Programming with R on Databricks, including Spark architecture, the DataFrame API, and Machine Learning.

#### Prerequisites & Requirements

- Prerequisites
  - Beginning experience working with R.

#### Learning objectives

- Identify core features of Spark and Databricks.
- Describe how DataFrames are created and evaluated in Spark.
- Apply the DataFrame transformation API to process and analyze data.

## Delta Lake Rapid Start with Python

[Click here](#) for the customer enrollment link.

Duration: 2 hours

Course description: Apache Spark™ is the dominant processing framework for big data. Delta Lake is a robust storage solution designed specifically to work with Apache Spark™. It adds reliability to Spark so your analytics and machine learning initiatives have ready access to quality, reliable data. Delta Lake makes data lakes easier to work with and more robust. It is designed to address many of the problems commonly found with data lakes. This course covers the basics of working with Delta Lake, specifically with Python, on Databricks.

#### Prerequisites:

- Beginning level experience using Databricks to upload and visualize data
- Intermediate level experience using Apache Spark including the CTAS pattern and use of popular pyspark.sql functions
- Beginning level knowledge of Delta Lake

#### Learning objectives:

- Use Delta Lake to create a new Delta table.

- Convert an existing Parquet-based data lake table into a Delta table.
- Differentiate between a batch update and an upsert to a Delta table.
- Use Delta Lake Time Travel to view different versions of a Delta table.
- Execute a MERGE command to upsert data into a Delta table.

## Deploying a Machine Learning Project with MLflow Projects

[Click here](#) for the customer enrollment link.

Duration: 2 hours

Course description: In this course, we'll show you how to train and deploy a large scale machine learning model using MLflow and Apache Spark. This course is the third in a series of three courses developed to show you how to use Databricks to work with a single data set from experimentation to production-scale machine learning model deployment. We recommend taking the first two courses in this series before continuing with this course:

- Building and Deploying Machine Learning Models: The Bias-Variance Tradeoff
- Tracking Experiments with MLflow

Prerequisites:

- Beginning-level experience running data science workflows in the Databricks Workspace
- Beginner-level experience with Apache Spark
- Intermediate-level experience with the Scipy Numerical Stack
- Intermediate-level experience with the command line

Learning objectives:

- Summarize Databricks best practices for deploying machine learning projects with MLflow.
- Explain local development strategies for writing software with Databricks.
- Use Databricks to write production-grade machine learning software.

## Easy ETL with Auto Loader

[Click here](#) for the customer enrollment link.

Duration: 1 hour

Course description: Databricks Auto Loader is the preferred method for ingesting incremental data landing in cloud object storage into your Lakehouse. This course introduces Auto Loader and demonstrates some of the newer features added to this product. Included are recommended patterns for data ingestion with Auto Loader.

Prerequisites:

- Basic experience with Spark APIs
- Basic knowledge of Delta Lake
- Basic experience with Structured Streaming

Learning objectives:

- Describe the basic functionality and features of Auto Loader.
- Use Auto Loader to ingest data to Delta Lake without losing data.
- Configure automatic schema detection and evolution.
- Rescue unexpected data arriving in well-structured datasets.

## Enterprise Architecture with Databricks

[Click here](#) for the customer enrollment link.

Duration: 7 hours

Course description: In this course you'll learn about how business leaders, admins, and architects use Databricks in their architecture . We'll cover fundamental concepts about key players: Data Engineers, Data Scientists, Platform Administrator; raw data forms: structured and unstructured data, batch and streaming data, to help set the stage for our discussion on how end users help businesses create data assets like machine learning models, reports, and dashboards. Then, we'll discuss where components of Databricks Azure fit into an organization's big data ecosystem. Finally, we'll review real-world business use cases and create enterprise level architecture infrastructure diagrams.

#### Prerequisites:

- Beginning knowledge about characteristics that define big data (3 of the Vs of big data – velocity, volume, variety)
- Beginning knowledge about how organizations process and manage big data (Relational/SQL vs NoSQL, cloud vs. on-premise, open-source database vs. closed-source database as a service)
- Beginning knowledge about the roles that data practitioners play on data science teams (can distinguish between database administrators and data scientists, data analysts and machine learning engineers, data engineers and platform administrators)

#### Learning objectives:

- Create a requirements document which profiles the data needs of an organization.
- Translate business needs related to data analytics into technical requirements used for drawing an architectural diagram.
- Translate the Databricks Lakehouse Architecture with Delta to a technical requirements document.
- Design Azure Databricks architectures that includes integration with Azure services, for real-world scenarios.
- Evaluate, analyze, and validate detailed infrastructure designs.
- Create infrastructure designs.

## Fundamentals of the Databricks Lakehouse Platform Accreditation

[Click here](#) for the customer accreditation enrollment link.

Cost: Free for Databricks customers

Duration: .5 hours

Accreditation description: This is a 30-minute assessment that will test your knowledge about fundamental concepts related to the Databricks Lakehouse Platform. Questions will assess how well you know about the platform in general, how familiar you are with the individual components of the platform, and your ability to describe how the platform helps organizations accomplish their data engineering,



data science/machine learning, and business/SQL analytics use cases. Please note that this assessment will not test your ability to perform tasks using Databricks functionality. Instead, it will test how well you can explain components of the platform and how they fit together.

After successfully completing this assessment, you will be awarded a Databricks Lakehouse Platform badge.

This accreditation is the beginning step in most of the Databricks Academy learning plans – SQL Analysts, Data Scientists, Data Engineers, and Platform Administrators. Business leaders are also welcome to take this assessment.

Prerequisites:

- We recommend that you take the following courses to prepare for this accreditation exam:
  - What is the Databricks Lakehouse Platform?
  - What are Enterprise Data Management Systems? (particularly the section on Lakehouse architecture)
  - What is Delta Lake?
  - What is Databricks SQL?

What is Databricks Machine Learning?

## Getting Started with Databricks Data Science & Engineering Workspace

[Click here](#) for the customer enrollment link.

Duration: 1.5 hours

Course description: The Databricks Data Science and Engineering Workspace (Workspace) provides a collaborative analytics platform to help data practitioners get the most out of Databricks when it comes to data science and engineering tasks. This course guides practitioners through fundamental Workspace concepts and components necessary to achieve a basic development workflow.

Prerequisites & Requirements

- Prerequisites

- Beginning-level knowledge of the Databricks Lakehouse platform (high-level knowledge the structure and benefits of the Lakehouse platform)
- Intermediate-level knowledge of Python (good understanding of the language as well as ability to read and write code)
- Beginning-level knowledge of SQL (ability to understand and construct basic queries)

### Learning objectives

- Describe the Databricks architecture and the services it provides.
- Navigate the Databricks Data Science and Engineering Workspace.
- Create and manage Databricks clusters for running code.
- Manage data using the Databricks File System and Delta Lake.
- Create and run Databricks Notebooks.
- Schedule non-interactive execution of Databricks Notebooks using Databricks Jobs.
- Integrate a hosted Git service for revision control using Databricks Repos.

# Getting Started with Databricks Machine Learning

[Click here](#) for the customer enrollment link.

Duration: 1 hour

Course description: Databricks Machine Learning offers data scientists and other machine learning practitioners a platform for completing and managing the end-to-end machine learning lifecycle. This course guides practitioners through a basic machine learning workflow using Databricks Machine Learning. Along the way, students will learn how each of Databricks Machine Learning's features better enable data scientists and machine learning engineers to complete their work effectively and efficiently.

### Prerequisites & Requirements

- Prerequisites
  - Beginning-level knowledge of the Databricks Lakehouse platform
  - Intermediate-level knowledge of Python

- Intermediate-level knowledge of machine learning workflows

### Learning objectives

- Describe a basic overview of Databricks Machine Learning.
- Create a feature table for downstream modeling using Feature Store.
- Automatically develop a baseline model using AutoML.
- Manage the model lifecycle using Model Registry.
- Perform batch inference using the registered model and feature table.
- Schedule a monthly model refresh using Databricks Jobs and AutoML.

## Getting Started with Databricks SQL

[Click here](#) for the customer enrollment link.

Duration: 1 hour

Course description: This course is an introductory course for SQL analysts that demonstrates the entire data analysis process on Databricks SQL, from introducing the Databricks SQL workspace (Workspace) to creating a dashboard. The course will focus on what analysts can do, as opposed to what administrators can do, and it will use the Workspace without administrator permissions.

### Prerequisites & Requirements

- Prerequisites
  - Beginning knowledge of SQL
  - Access to Databricks SQL
  - Access to an empty database set up by an administrator
  - Access to a SQL endpoint set up by an administrator

### Learning objectives

- Describe the basics of the Databricks SQL service.
- Describe the benefits of using Databricks SQL to perform data analyses.
- Describe how to complete a basic query, visualization, and dashboard workflow using Databricks SQL.

# Google Cloud Fundamentals

[Click here](#) for the customer enrollment link.

Duration: 1.5 hours

Course description: Learn the basics of Google Cloud and how to configure various resources using the Cloud Console. This course begins with an overview of the platform, key terminology, and core services. You will then learn essential IAM concepts and how service accounts can be used to manage resources. You will also learn about the function and use cases of several storage services, such as Cloud Storage, Cloud SQL, and BigQuery. This course also covers virtual machine and networking concepts in Compute Engine and VPC services. The course ends with an overview of GKE clusters and Kubernetes concepts.

Prerequisites:

- Familiarity with basic cloud computing concepts (cloud computing, cloud storage, virtual machine, database, data warehouse)

Learning objectives:

- Define basic concepts and core services in the Google Cloud Platform.
- Describe IAM concepts and how service accounts can be used to manage resources.
- Identify use cases for storage services, such as Cloud Storage, Cloud SQL, and BigQuery.
- Define virtual machine and networking concepts in Compute Engine and VPC services.
- Describe Google Kubernetes Engine and the core components of Kubernetes clusters.

# How to Ingest Data for Databricks SQL

[Click here](#) for the customer enrollment link.

Duration: .5 hour

Course description: Before an analyst can analyze data, that data needs to be ingested into the Lakehouse. This course shows three different ways to ingest data: 1. Using the Data Science & Engineering UI, 2. Using SQL, and 3. Using Partner Connect.

#### Prerequisites & Requirements

- Prerequisites
  - Intermediate knowledge of Databricks SQL
  - Administrator privileges

#### Learning objectives

- Upload data using the Data Science & Engineering UI
- Import data using Databricks SQL
- Provide proper data access privileges to users
- Import data using Partner Connect

## How to Schedule a Job and Automate a Workload with the Databricks Data Science and Data Engineering Workspace

[Click here](#) for the customer enrollment link.

Duration: 6 minutes

Course description: Onboarding for AWS Databricks Customers providing an introduction to Notebook automation within the Workspace

#### Prerequisites:

- Knowledge of Data Engineering concepts
- Beginner experience with Databricks Data Science & Engineering Workspace
- Basic familiarity with Python

#### Learning objectives:

- Create a Job that runs a Notebook
- Schedule a Job
- Manage Jobs and Alerts
- Review execution results

# How to Use Databricks' COPY INTO for Incremental ETL with Databricks SQL

[Click here](#) for the customer enrollment link.

Duration: 7 minutes

Course description: Onboarding for AWS Databricks Customers providing an overview of COPY INTO for performing incremental ETL in Databricks SQL

Prerequisites:

- Knowledge of Data Engineering concepts
- Beginner experience with Databricks Data Science & Engineering Workspace and Databricks Notebooks
- Basic familiarity with SQL

Learning objectives:

- Explain benefits of incremental ETL
- Create an SQL Endpoint
- Create a Delta table
- Import and run a Notebook
- Create a DBSQL Query

# How to Use Databricks' Auto Loader for Incremental ETL with the Databricks Data Science and Data Engineering Workspace

[Click here](#) for the customer enrollment link.

Duration: 7 minutes

Course description: Onboarding for AWS Databricks Customers providing an overview of COPY INTO for performing incremental ETL in Databricks SQL

Prerequisites:

- Knowledge of Data Engineering concepts
- Beginner experience with Databricks Data Science & Engineering Workspace and Databricks Notebooks
- Basic familiarity with SQL

Learning objectives:

- Explain benefits of incremental ETL
- Create an SQL Endpoint
- Create a Delta table
- Import and run a Notebook
- Create a DBSQL Query

# Introduction to Apache Spark Architecture

[Click here](#) for the customer enrollment link.

Duration: 1 hour

Course description: In this course, you will explore how Apache Spark executes a series of queries. Examples will include simple narrow transformations and more complex wide transformations.

This course will give developers a working understanding of how to write code that leverages the power of Apache Spark for even the simplest of queries.

Prerequisites:

- Familiarity with basic information about Apache Spark (what it is, what it is used for)

Learning objectives:

- Explain how Apache Spark applications are divided into jobs, stages, and tasks.
- Explain the major components of Apache Spark's distributed architecture.

## Introduction to Applied Linear Models

[Click here](#) for the customer enrollment link.

Duration: 1 hour

Course description: Linear modeling is a popular starting point for machine learning studies for a number of reasons. Generally, these models are relatively easy to interpret and explain, and they can be applied to a broad range of problems. In this course, you will learn how to choose, apply, and evaluate commonly used linear modeling techniques. As you work through the course, you can put your new skills to practice in 5 hands-on labs.

### Prerequisites & Requirements

- Prerequisites
  - Intermediate experience with machine learning (experience using machine learning and data science libraries like scikit-learn and Pandas, knowledge of linear models).
  - Intermediate experience using the Databricks Workspace to perform data analysis (using Spark DataFrames, Databricks notebooks, etc.).
  - Beginning experience with statistical concepts commonly used in data science.

### Learning objectives

- Describe and evaluate linear regression for regression problems.
- Describe how to ensure machine learning models generalize to out-of-sample data.
- Describe and evaluate logistic regression for classification problems.
- Practice using linear modeling techniques using the Databricks Data Science Workspace.



# Introduction to Applied Statistics

[Click here](#) for the customer enrollment link.

Duration: 1 hour

Course description: In this course you'll learn, both in theory and in practice, about statistical techniques that are fundamental to many data science projects. Throughout the course, videos will guide you through the conceptual information you need to know about these statistical concepts, and hands-on lab activities will give you the chance to apply the concepts you learn using the Databricks Workspace. This course is divided into three modules: Introduction to Statistics and Probability, Probability Distributions, and Applying Statistics to Learn from Data.

## Prerequisites & Requirements

- Prerequisites
  - Beginning experience using the Databricks Data Science Workspace (familiarity with Spark SQL, experience importing files into the Databricks Data Science Workspace)
  - Beginning experience using Python (ability to follow guided use of the SciPy library)

## Learning objectives

- Contrast descriptive statistics and inferential statistics.
- Explain fundamental concepts behind discrete probability.
- Compare and contrast discrete and continuous probability distributions.
- Explain how discrete and continuous probability distributions can be used to model data.
- Apply hypothesis testing techniques to learn from data.

# Introduction to Applied Tree-Based Models

[Click here](#) for the customer enrollment link.

Duration: 3 hours

Course description: In this course, you'll learn how to solve complex supervised learning problems using tree-based models. First, we'll explain how decision trees can be used to identify complex relationships in data. Then, we'll show you how to develop a random forest model to build upon decision trees and improve model generalization. Finally, we'll introduce you to various techniques that you can use to account for class imbalances in a dataset. Throughout the course, you'll have the opportunity to practice concepts learned in hands-on labs.

Prerequisites:

- Intermediate level knowledge about machine learning/machine learning workflows (feature engineering and selection, applying tree-based models, etc.)
- We recommend that you take the following courses prior to taking this course: Fundamentals of Machine Learning, Introduction to Feature Engineering and Selection with Databricks, Applied Unsupervised Learning with Databricks.

Learning objectives:

- Describe how decision trees are used to solve supervised learning problems.
- Identify complex relationships in data using decision trees.
- Develop a random forest model to build upon decision trees and improve model generalization.
- Employ common techniques to account for class imbalances in a dataset.

## Introduction to Applied Unsupervised Learning

[Click here](#) for the customer enrollment link.

Duration: 3 hours

Course description: In this course, we will describe and demonstrate how to learn from data using unsupervised learning techniques during exploratory data analysis. The course is divided into two sections – one of which will focus on K-means clustering and the other will describe principal components analysis, commonly

referred to as PCA. Each section includes demonstrations of important concepts, a quiz to solidify your understanding, and a lab to practice your skills.

Prerequisites:

- Intermediate experience with machine learning (experience using machine learning and data science libraries like scikit-learn and Pandas, knowledge of linear models)
- Intermediate experience using the Databricks Workspace to perform data analysis (using Spark DataFrames, Databricks notebooks, etc.)
- Beginning experience with machine learning concepts.

Learning objectives:

- Identify relationships between data records using K-means clustering.
- Identify patterns in a high-dimensional feature space using principal components analysis.
- Learn from data using unsupervised learning techniques during exploratory data analysis.

## Introduction to Cloning with Delta Lake

[Click here](#) for the customer enrollment link.

Duration: 30 minutes

Course description: The addition of clone to Delta Lake empowers data engineers and administrators to easily replicate data stored in the Lakehouse. Organizations can use deep clone to archive versions of their production tables for regulatory compliance. Developers can easily create development datasets isolated from production data with shallow clone. In this course, you'll learn the basics of cloning with Delta Lake and get hands-on experience working with the syntax.

Prerequisites:

- Hands-on experience working with Delta Lake
- Intermediate experience with Spark and Databricks

Learning objectives:

- Describe the basic execution of deep and shallow clones with Delta Lake.
- Use deep clones to create full incremental backups of tables.

- Use shallow clones to create development datasets.
- Describe strengths, limitations, and caveats of each type of clone.

## Introduction to Databricks Connect

[Click here](#) for the customer enrollment link.

Duration: 40 minutes

Course description: In this course, participants will be introduced to DB Connect through various presentations and demos. Participants will start by contrasting how DB Connect works to other development patterns. Then we will explore the simplicity by which DB Connect is installed and configured. And then we will conclude with a real-time demonstration of an application running on a developer's local machine while executing its Spark jobs against a cluster in the Databricks workspace.

Prerequisites:

- Intermediate experience using the Databricks Workspace

Learning objectives:

- Explain how Databricks Connect is used by data practitioners working with Databricks.
- Install and configure Databricks Connect.

## Introduction to Databricks Repos

[Click here](#) for the customer enrollment link.

Duration: 30 minutes

Course description: Repos aims to make Databricks simple to use by giving data scientists and engineers the familiar tools of git repositories and file systems. These tools enable a more laptop-like developer experience for customers. Repos is the new, top-level, customer-facing feature that packages these tools together in the Databricks user interface. This course teaches how to get started with Repos.

Prerequisites:

- Familiarity with Git and Git commands
- Familiarity with Databricks workspaces

Learning objectives:

- Describe the motivations for Databricks Repos.
- Configure workspace integration with Github.
- Sync local and remote notebook changes using Repos

## Introduction to Delta Lake

[Click here](#) for the customer enrollment link.

Duration: 1 hour, 15 minutes

Course description: Delta Lake is a powerful tool created by Databricks. Delta Lake is an open, reliable, performant and secure data storage and management layer for your data lake that enables you to create a true single source of truth. Since it is built upon Apache Spark, you're able to build high performance data pipelines to clean your data from raw ingestion to business level aggregates. Finally, given the open format – it allows you to avoid unnecessary replication and proprietary lock-in.

Ultimately – it provides the reliability, performance, and security you need to serve your downstream data use cases.

Prerequisites:

- Intermediate SQL skills (e.g. can do CRUD statements in SQL)
- Beginner experience with working on Databricks in the Data Science & Engineering workspace or the Machine Learning workspace (e.g. can import DBC files, can access workspaces). Also note: although this course relies heavily on SQL as a language, this is not intended for learners who primarily use the Databricks SQL workspace products.
- Beginner experience with working with data pipelines is helpful

Learning objectives:

- Describe the basic features and technical implementation of Delta Lake.
- Ingest data and manage Delta tables to keep data complete, up-to-date, and organized.
- Optimize Delta performance using common strategies.

# Introduction to Delta Live Tables

[Click here](#) for the customer enrollment link.

Duration: 30 minutes

Course description: Delta Live Tables enables data teams to innovate rapidly with simple development, using declarative tools to build and manage batch or streaming data pipelines. Built-in quality controls and data quality monitoring ensure accurate and useful BI, Data Science, and ML built on top of quality data. Delta Live Tables is designed to scale with rapidly growing companies and provides clear observability into pipeline operations and automatic error handling. This course will cover the basics of this new product, including syntax, configuration, and deployment.

Prerequisites:

- Beginner experience working with PySpark or Spark SQL
- Basic familiarity with the Databricks workspace

Learning objectives:

- Describe the motivations for Delta Live Tables.
- Use PySpark or SQL syntax to declare Delta Live Tables.
- Schedule and deploy pipelines with the Databricks UI.
- Review pipeline logs and metrics.

# Introduction to Feature Engineering and Selection with Databricks

[Click here](#) for the customer enrollment link.

Duration: 2.5 hours

Course description: As data practitioners work on supervised machine learning solutions, they often need to manipulate data to ensure that it is compatible with machine learning algorithm requirements and the model is meeting its objective. This process is known as feature engineering, and the end result is to improve the output of machine learning solutions. Once features are engineered, data

practitioners also commonly need to determine the best way to select the best features to use in their machine learning projects.

In this course, you'll learn how to perform both of these tasks. This course is divided into two modules – in the first, you'll explore feature engineering. In the second, you'll explore feature selection. Both modules will start with an introduction to these topics – what they are and why they're used. Then, you'll review techniques that help data practitioners perform these tasks. Finally, you'll have the chance to perform two hands-on lab activities – one where you will engineer features and another where you will select features for a fictional machine learning scenario.

Prerequisites:

- Intermediate experience with machine learning (experience using machine learning and data science libraries like scikit-learn and Pandas, knowledge of linear models)
- Intermediate experience using the Databricks Workspace to perform data analysis (using Spark DataFrames, Databricks notebooks, etc.)
- Beginning experience with statistical concepts commonly used in data science

Learning objectives:

- Explain popular feature engineering techniques used to improve supervised machine learning solutions.
- Explain popular feature selection techniques used to improve supervised machine learning solutions.
- Engineer meaningful features for use in a supervised machine learning project using the Databricks Data Science Workspace.
- Select meaningful features for use in a supervised machine learning project using the Databricks Data Science Workspace.

## Introduction to Files in Databricks Repos

[Click here](#) for the customer enrollment link.

Duration: 30 minutes

Course description: This course teaches how to add non-notebook files to Databricks Repos. Learners will connect a Databricks workspace to a hosted Git

repository. Next, they will import and store non-DBC and non-notebook files using Databricks Repos. Then, they will import a markdown file and sync changes between a Databricks Repo and a Git provider.

Prerequisites:

- Familiarity with Git and Git commands
- Familiarity with Databricks workspaces
- Completion of of the Introduction to Databricks Repos course

Learning objectives:

- Connect a Databricks workspace to a hosted Git repository using Databricks Repos
- Import and store non-DBC and non-notebook files using already-configured Databricks Repos with a Git provider
- Import a markdown file imported into workspace
- Sync changes within Databricks to a Git provider

# Introduction to Hyperparameter Optimization

[Click here](#) for the customer enrollment link.

Duration: 2 hours

Course description: In this course, you'll learn how to apply hyperparameter tuning strategies to optimize machine learning models for unseen data. First, you'll work within a balanced binary classification problem setting where you'll use random forest to predict the correct class. You'll learn to tune the hyperparameters of a random forest to improve a model. Then, you'll again work with a binary classification problem using random forest and a technique known as cross-validation to generalize a model.

Prerequisites:

- Intermediate level knowledge about machine learning/machine learning workflows (feature engineering and selection, applying tree-based models, etc.)



- We recommend that you take the following courses prior to taking this course: Fundamentals of Machine Learning, Introduction to Feature Engineering and Selection with Databricks, Introduction to Applied Tree-based Models with Databricks.

Learning objectives:

- Explain common machine learning techniques that are used to optimize machine learning models for unseen data.
- Apply machine learning techniques to improve the fit of machine learning models.
- Apply machine learning techniques to improve the generalization of machine learning models.

## Introduction to Jobs

[Click here](#) for the customer enrollment link.

Duration: 30 minutes

Course description: Databricks Jobs allow users to run applications in a non-interactive way on a cluster. Jobs allow users to manage and orchestrate production tasks, making it simple to promote notebooks from interactive development to scheduled workloads. In this course, you'll explore various features of the Jobs UI as you orchestrate a simple pipeline.

Prerequisites:

- Intermediate knowledge of Python or SQL
- Beginning knowledge of software development principles (e.g. code modularity, code scheduling, code orchestration)
- Beginning knowledge of navigating Databricks UI

Learning objectives:

- Describe jobs and motivations for using jobs in the workflow of data practitioners.
- Create single task jobs with a scheduled trigger.
- Orchestrate multiple notebook tasks with the Jobs UI.
- Discuss common use cases and patterns for Jobs.

# Introduction to MLflow Model Registry

[Click here](#) for the customer enrollment link.

Duration: 30 minutes

Course description: This course will introduce you to MLflow Model Registry. Model Registry is a centralized model management tool that allows you to track metrics, parameters, and artifacts as part of experiments, package models and reproducible ML projects, and deploy models to batch or real-time serving platforms. You will learn how your team can use Model Registry as a central place to share ML models, collaborate on moving them from experimentation to testing and production, and implement approval and governance workflows.

Prerequisites:

- Beginner-level experience with machine learning.
- Beginner-level experience with MLflow Model Tracking.
- Beginner-level experience with Python.
- Beginner-level experience with Apache Spark on Databricks.

Learning objectives:

- Describe the components and functionalities of Model Registry.
- Explain the benefits of using Model Registry for machine learning model management.
- Describe how Model Registry fits into the ML lifecycle with Databricks Machine Learning.
- Demonstrate how to use Model Registry to perform essential tasks in the ML workflow.

# Introduction to MLflow Tracking

[Click here](#) for the customer enrollment link.

Duration: 1 hour

Course description: MLflow is an open-source platform for managing the end-to-end machine learning lifecycle. In this course, we're going to explore one of the four primary functions of MLflow: tracking. Tracking in MLflow is an API and UI for

logging parameters, code versions, metrics, and output files when running your machine learning code and for later visualizing the results. MLflow Tracking lets you log and query experiments using Python, REST, R, and Java APIs.

### Prerequisites & Requirements

- Prerequisites
  - Experience developing machine learning models in SciKit-Learn
  - Experience and comfortability using Python and Data Science related libraries (e.g. writing functions, using attributes and methods, instantiating classes, basic file I/O with Pandas)
  - Comfortability with building classification and regression models in SciKit-Learn

### Learning objectives

- Describe the basics of Databricks-managed MLflow Tracking.
- Identify the best run using the MLflow Experiments UI and the Tracking UI.
- Identify the best run using the MLflow Client API.
- Manually and automatically log metrics, artifacts, and models in an MLflow Run.

# Introduction to Natural Language Processing

[Click here](#) for the customer enrollment link.

Duration: 4 hours

Course description: This course will introduce you to natural language processing with Databricks. You will learn how to generate term-frequency-inverse-document-frequency (TFIDF) vectors for your datasets and how to perform latent semantic analysis using the Databricks Machine Learning Runtime.

### Prerequisites:

- Intermediate experience performing machine learning/data science workflows

- Intermediate experience using the Databricks Data Science Workspace to perform machine learning workflows

Learning objectives:

- Describe foundational concepts about how latent semantic analysis is used to analyze text data.
- Perform latent semantic analysis using the Databricks Machine Learning Runtime with the Databricks Workspace.
- Generate TFIDF vectors to reduce the noise in a dataset being used for latent semantic analysis in a Databricks Workspace.

## Introduction to Photon

[Click here](#) for the customer enrollment link.

Duration: 30 minutes

Course description: In this course, you'll learn how Photon can be used to reduce Databricks total cost of ownership (TCO) and dramatically improve query performance. You'll also learn best practices for when to use and not use Photon. Finally, the course will include a demonstration of a query run with and without Photon to show improvement in query performance.

Prerequisites:

- Administrator privileges
- Introductory knowledge about the Databricks Lakehouse Platform (what the Databricks Lakehouse Platform is, what it does, main components, etc.)

Learning objectives:

- Explain fundamental concepts about Photon on Databricks.
- Describe the benefits of enabling Photon on Databricks.
- Identify queries that would benefit from using Photon
- Describe the performance differences between a query run with and without Photon enabled

# Just Enough Python for Apache Spark

[Click here](#) for the customer enrollment link.

Duration: 6 hours

NOTE: This is an e-learning version of the Just Enough Python for Apache Spark instructor-led course. It is an on-demand recording available via the Databricks Academy and covers the same content as the instructor-led course. For more information about what's in the course itself, please [visit this link](#).

# Migrating SAS Procedures to Databricks

[Click here](#) for the customer enrollment link.

Duration: 30 minutes

Course description: This course will enable experienced SAS developers to quickly learn how to translate familiar SAS statements and functions into code that can be run on Databricks. It begins with an introduction to the Databricks environment and the different approaches to coding in Databricks, followed by an overview of how SAS PROC and DATA steps can be performed in Databricks. You will learn about how you can use Spark SQL, PySpark, and other tools to read .sas7bdat files and perform common operations. Finally, you will see code examples and gain hands-on practice performing some of the most common SAS operations in Databricks.

Prerequisites:

- Intermediate to advanced SAS programming experience
- Beginning knowledge of Python programming
- Beginning-level experience with SQL

Learning objectives:

- Read data stored in .sas7bdat files using Spark SQL and PySpark.
- Explain the conceptual and syntactical relationships between SAS DATA and PROC statements and their correlaries on Databricks.
- Learn how Python can be leveraged to augment ANSI SQL to create reusable Spark SQL code.
- Translate common PROC functions to Databricks.

- Translate common DATA steps to Databricks.

# Natural Language Processing at Scale with Databricks

[Click here](#) for the customer enrollment link.

Duration: 5 hours

Course description: This five-hour course will teach you how to do natural language processing at scale on Databricks. You will apply libraries such as NLTK and Gensim in a distributed setting as well as SparkML/MLlib to solve classification, sentiment analysis, and text wrangling tasks. You will learn how to remove stop words, when to lemmatize vs stem your tokens, and how to generate term-frequency-inverse-document-frequency (TFIDF) vectors for your dataset. You will also use dimensionality reduction techniques to visualize word embeddings with Tensorboard and apply and visualize basic vector arithmetic to embeddings.

Prerequisites:

- Experience working with PySpark DataFrames
- Mastery of concepts presented in the Databricks Academy "Apache Spark Programming" course
- Mastery of concepts presented in the Databricks Academy "Scalable Machine Learning with Apache Spark" course

Learning objectives:

- Explain the motivation behind using Natural Language Processing to analyze data.
- Identify distributed Natural Language Processing libraries commonly used when analyzing data.
- Perform a series of Natural Language Processing workflows in the Databricks Data Science Workspace

# New Capability Overview: Databricks AutoML

[Click here](#) for the customer enrollment link.

Duration: 30 minutes

Course description: This course will introduce data scientists to Databricks AutoML. AutoML is a tool that empowers data teams to quickly build and deploy machine learning models by automating the heavy lifting of preprocessing, feature engineering and model training/tuning. You will learn how to train, modify, and register models for classification or regression with the Databricks Machine Learning UI and the Python API.

## Prerequisites & Requirements

- Prerequisites
  - None

## Learning objectives

- Identify the core features and benefits of Databricks AutoML
- Describe where AutoML fits in the workflow of machine learning teams
- Configure and run AutoML experiments using the user interface and API
- Describe how AutoML integrates with other features in Databricks Machine Learning

# New Capability Overview: MLflow Model Serving

[Click here](#) for the customer enrollment link.

Duration: 40 minutes

Course description: MLflow Model Serving on Databricks is an exciting feature which makes model serving simple. Model serving is an important part of machine learning infrastructure. With serving, a model's prediction can be delivered in real-time all from the Databricks Machine Learning platform.

## Prerequisites & Requirements

- Prerequisites
  - Beginner level experience with machine learning
  - Beginner level experience with Python
  - Beginner level experience with Apache Spark on Databricks

## Learning objectives

- Learners will be able to describe common problems that model serving overcomes
- Learners will be able to utilize Databricks Model Serving to create a REST endpoint

# New Capability Overview: Feature Store

[Click here](#) for the customer enrollment link.

Duration: 1 hour

Course description: In this course, learners will practice using the Databricks Feature Store. From creating and updating feature store tables to searching across the Feature Store, functionality is accessible through Databricks notebooks and Jobs. Feature Store enables data practitioners to share and discover features across their organization, as well as ensure that the same feature computation code is used for model training and inference.

## Prerequisites & Requirements

- Prerequisites
  - Creating models with SciKit-Learn or ML Lib
  - Hardening for security concerns like handling data in flight, CORS or SQL injection
  - API architecture beyond REST (e.g. SOAP or graph models will not be discussed)
  - Optimizing clusters for serving (e.g. latency, SLAs, and throughput concerns)
  - How the MLflow Registry works. Better if learner can log models to the registry
  - Monitoring model drift and performance



### Learning objectives

- Describe common problems that Model Serving overcomes
- Utilize Databricks Model Serving to deploy a real-time model via a REST endpoint

## New Capability Overview: Time Series Forecasting with AutoML

[Click here](#) for the customer course enrollment link.

Duration: 31 minutes

Course description: Time series forecasting is an incredibly important component of any organization's portfolio of models. But the logic, set up, and tuning of time series models can be overwhelming. Databricks offers Time Series models in its AutoML product, which takes much of the time and headache out of creating these models! This course will introduce learners to the time series functionality in AutoML.

### Prerequisites:

- Beginning-level knowledge of and experience with AutoML.
- Beginning level knowledge of time series modeling

### Learning objectives:

- Locate key features of AutoML Time Series UI
- Employ knowledge of the Prophet time series model to create a forecast in AutoML via API

## Optimizing Apache Spark on Databricks

[Click here](#) for the customer course enrollment link.

Duration: 12 hours

NOTE: This is an e-learning version of the Optimizing Apache Spark on Databricks instructor-led course. It is an on-demand recording available via the Databricks

Academy and covers the same content as the instructor-led course. For more information about what's in the course itself, please [visit this link](#).

# Propagating Changes with Delta Change Data Feed

[Click here](#) for the customer course enrollment link.

Duration: 1 hour

Course description: A Delta change data feed represents row-level changes between versions of a Delta table. When enabled on a Delta table, the runtime records “change events” for all the data written into the table. This includes the row data along with metadata indicating whether the specified row was inserted, deleted, or updated. In this course, we'll examine some of the motivations and use cases for this feature and see it in action.

Prerequisites:

- Basic knowledge of Spark Structured Streaming APIs
- Basic knowledge of Delta Lake

Learning objectives:

- Describe how Delta Change Data Feed emits change data records.
- Use appropriate syntax and settings to set up Change Data Feed.
- Propagate inserts, updates, and deletes with Change Data Feed.

## Quick Reference: CI/CD

[Click here](#) for the customer course enrollment link.

Duration: 30 minutes

Course description: This quick-reference provides an overview of fundamental concepts behind CI/CD. While the Databricks tools and integrations mentioned in this course can be used by DevOps teams for CI/CD, this course was designed to summarize what happens during each stage of a CI/CD pipeline (not provide a

technical how-to into each of these stages). Future courses will dive into each of these stages in greater detail. Note: We will use Jenkins as an example automation system in this course.

Prerequisites:

- Beginning-level experience with CI/CD, DevOps and/or the software development lifecycle (not necessarily on Databricks)

Learning objectives:

- Summarize each stage in a traditional CI/CD pipeline.
- Outline the steps in configuring the Jenkins automation agent for use in CI/CD.

## Quick Reference: Spark Architecture

[Click here](#) for the customer course enrollment link.

Duration: 30 minutes

Course description: Apache Spark™ is a unified analytics engine for large scale data processing known for its speed, ease and breadth of use, ability to access diverse data sources, and APIs built to support a wide range of use-cases. Databricks builds on top of Spark and adds many performance and security enhancements. This course is meant to provide an overview of Spark's internal architecture.

Prerequisites:

- Beginning knowledge of big data and data science concepts.

Learning objectives:

- Describe basic Spark architecture and define terminology such as “driver” and “executor”
- Explain how parallelization allows Spark to improve speed and scalability of an application
- Describe lazy evaluation and how it relates to pipelining
- Identify high-level events for each stage in the Optimization process

# Scaling Machine Learning Pipelines

[Click here](#) for the customer course enrollment link.

Duration: 3 hours

Course description: In this course, learners integrate machine learning solutions with scalable production pipelines backed by Apache Spark. Learners will start by investigating common inefficiencies in machine learning. Next, students will learn to scale the development and tuning of the machine learning workflow using tools like Spark ML and Hyperopt. Finally, learners will finish by using Pandas UDFs and the Pandas Function APIs to create and apply group-specific machine learning models. By the end of this course, learners will be capable of scaling the entirety of a machine learning pipeline.

Prerequisites:

- Intermediate level experience with Apache Spark (familiarity with Spark architecture and Spark DataFrame API).
- Intermediate level experience with Python and its single-node data science stack (familiarity with libraries, iteration, control flow, operators, and classes).
- Intermediate level knowledge of and experience in machine learning (supervised learning vs. unsupervised learning, regression vs. classification, clustering, and experience building models following a machine learning workflow with single-node libraries like Scikit-learn).

Learning objectives:

- Evaluate characteristics of machine learning pipelines to determine how to scale with Apache Spark.
- Run common machine learning data preparation techniques on big data using Apache Spark.
- Develop machine learning models for big data using Apache Spark.
- Accelerate the tuning of single-node machine learning models using Hyperopt and Apache Spark.
- Apply grouped machine learning model training and inference using Pandas UDFs and the Pandas Function APIs.
- Employ Databricks-recommended best practices to scale a machine learning pipeline using previously covered techniques.

# Scalable Machine Learning with Apache Spark

[Click here](#) for the customer course enrollment link.

Duration: 12 hours

NOTE: This is an e-learning version of the Scalable Machine Learning with Apache Spark instructor-led course. It is an on-demand recording available via the Databricks Academy and covers the same content as the instructor-led course. For more information about what's in the course itself, please [visit this link](#).

## Structured Streaming

[Click here](#) for the customer course enrollment link.

Duration: 1 hour

Course description: This hands-on self-paced training course targets data engineers who want to process big data using Apache Spark™ Structured Streaming. The course is a series of four self-paced lessons. Each lesson includes hands-on exercises. The course contains Databricks notebooks for both Azure Databricks and AWS Databricks; you can run the course on either platform.

Prerequisites:

- Completion of Apache Spark Programming on Databricks course strongly encouraged

Learning objectives:

- Use the interactive Databricks notebook environment
- Ingest streaming log file data
- Aggregate small batches of data with time windows
- Stream data from a Kafka connection
- Use Structured Streaming in conjunction with Databricks Delta
- Visualize streaming live data
- Use Structured Streaming to analyze streaming Twitter data

# Tracking Experiments with MLflow

[Click here](#) for the customer course enrollment link.

Duration: 2 hours

Course description: In this course, we'll show you how to design an MLflow experiment to identify the best machine model for deployment. This course is the second in a series of three courses developed to show you how to use Databricks to work with a single data set from experimentation to production-scale machine learning model deployment. The other courses in this series include:

- Data Science on Databricks: The Bias-Variance Tradeoff
- Deploying a Machine Learning Project with MLflow Projects

Prerequisites:

- Beginning-level experience running data science workflows in the Databricks Workspace
- Beginner-level experience with Apache Spark
- Intermediate-level experience with the Scipy Numerical Stack

Learning objectives:

- Create and explore an augmented sample from user event and profile data.
- Design an MLflow experiment and write notebook-based software to run the experiment to assess various linear models.
- Examine experimental results to decide which model to develop for production.

# What are Enterprise Data Management Systems?

[Click here](#) for the customer course enrollment link.

Duration: 1 hour

Course description: Whether your organization is moving to the cloud for the first time or reevaluating its current approach, making decisions about the technology

used when storing your data can have huge implications for costs and performance in downstream analytics. As a platform focused on computation and analytics, Databricks seeks to help our customers make choices that unlock new opportunities, reduce redundancies, and connect data teams. In this course, you'll start by exploring the characteristics of data lakes, and data warehouses, two popular data storage technologies. Then, you'll learn about the Lakehouse, a new data storage system invented and made popular by Databricks.

#### Prerequisites:

- Beginning knowledge about the Databricks Unified Data Analytics Platform.
- We recommend taking the courses: Fundamentals of Big Data and Fundamentals of Unified Data Analytics with Databricks prior to taking this course.

#### Learning objectives:

- Describe the strengths and limitations of data lakes, related to data storage.
- Describe the strengths and limitations of data warehouses, related to data storage.
- Contrast data lake and data warehouse characteristics.
- Compare the features of a Lakehouse to the features of popular data storage management solutions.

## What is Big Data?

[Click here](#) for the customer course enrollment link.

Duration: 1 hour

Course description: This course was created for individuals who are new to the big data landscape and want to become conversant with big data terminology. It will cover foundational concepts related to the big data landscape including: characteristics of big data; the relationship between big data, artificial intelligence, and data science; how individuals on data science teams work with big data; and how organizations can use big data to enable better business decisions.

#### Prerequisites:

- Experience using a web browser

### Learning objectives:

- Explain foundational concepts used to define big data.
- Explain how the characteristics of big data have changed traditional organizational workflows for working with data.
- Summarize how individuals on data science teams work with big data on a daily basis to drive business outcomes.
- Articulate examples of real-world use-cases for big data in businesses across a variety of industries.

## What is Cloud Computing?

[Click here](#) for the customer course enrollment link.

Duration: 30 minutes

Course description: This introductory-level course is designed to familiarize individuals new to the cloud computing landscape. It will cover foundational concepts related to cloud computing starting with the basics – what cloud computing is and why, since 2011, over 30% of organizations have moved their operations to the cloud. The course will also cover topics like cloud delivery models and deployment types.

Please note that this course is about cloud computing in general and does not focus on Databricks, specifically.

### Prerequisites:

- Experience using a web browser

### Learning objectives:

- Summarize foundational concepts about cloud computing.
- Describe major cloud computing components.
- Explain the three major cloud computing delivery models.
- Explain the three major cloud computing deployment models.
- Outline the benefits of moving an organization's operations to the cloud.



# What is Databricks Machine Learning?

[Click here](#) for the customer course enrollment link.

Duration: 30 minutes

Course description: Databricks Machine Learning offers data scientists and other machine learning practitioners a platform for completing and managing the end-to-end machine learning lifecycle. This course guides business leaders and practitioners through a basic overview of Databricks Machine Learning, the benefits of using Databricks Machine Learning, its fundamental components and functionalities, and examples of successful customer use.

Prerequisites:

- Beginning-level knowledge of the Databricks Lakehouse platform

Learning objectives:

- Describe the basic overview of Databricks Machine Learning.
- Identify how using Databricks Machine Learning benefits data science and machine learning teams.
- Summarize the fundamental components and functionalities of Databricks Machine Learning.
- Exemplify successful use cases of Databricks Machine Learning by real Databricks customers.

# What is Databricks SQL?

[Click here](#) for the customer course enrollment link.

Duration: 30 minutes

Course description: Databricks SQL offers SQL users a platform for querying, analyzing, and visualizing data in their organizations Lakehouse. This course explains how Databricks SQL processes queries and guides users through how to use the interface. Then, this course will explain how you can connect to Databricks SQL to your favorite business intelligence tool, so that you can query your Lakehouse without making changes to your analytical and dashboarding workflows.

Prerequisites:

- None.

Learning objectives:

- Summarize fundamental concepts for using Databricks SQL effectively.
- Identify tools and features in Databricks SQL for querying and analyzing data as well as sharing insights with the larger organization.
- Explain how Databricks SQL supports data analysis workflows that allow users to extract and share business insights

## What is Delta Lake?

[Click here](#) for the customer course enrollment link.

Duration: 30 minutes

Course description: Delta Lake is an open format storage layer that sits on top of your organization's data lake. It is the foundation of a cost-effective, highly scalable Lakehouse and is an integral part of the Databricks Lakehouse Platform.

In this course, we'll break down the basics behind Delta Lake – what it does, how it works, and why it is valuable from a business perspective, to any organization with big data and AI projects.

Note: This is an introductory-level course that will *\*not\** showcase in-depth technical Delta Lake demos nor provide hands-on technical training with Delta Lake. Please see the Delta Lake Rapidstart courses available in the Databricks Academy for technical training on Delta Lake.

Prerequisites:

- Beginning knowledge of the Databricks Lakehouse Platform. We recommended taking the course Fundamentals of the Databricks Lakehouse Platform prior to taking this course.

Learning objectives:

- Describe how Delta Lake fits into the Databricks Lakehouse Platform.
- Explain the four elements encompassed by Delta Lake.

- Summarize high-level Delta Lake functionality that helps organizations solve common challenges related to enterprise-scale data analytics.
- Articulate examples of how organizations have employed Delta Lake on Databricks to improve business outcomes.

## What is Machine Learning?

[Click here](#) for the customer course enrollment link.

Duration: 1 hour

Course description: In this course you'll learn fundamental concepts about machine learning. First, we'll review machine learning basics – what it is, why it's used, and how it relates to data science. Then, we'll explore the two primary categories that machine learning problems are categorized into – supervised and unsupervised learning. Finally, we'll review how the machine learning workflow fits into the data science process.

### Prerequisites & Requirements

- Prerequisites
  - Beginning knowledge about concepts related to the big data landscape helpful but not required (i.e. big data types, analysis techniques, processing techniques, etc.)
  - We recommend taking the Databricks Academy course "Introduction to Big Data" before taking this course.

### Learning objectives

- Explain how machine learning is used as an analysis tool in data science.
- Summarize the relationship between the data science process and the machine learning workflow.
- Describe the two primary categories that machine learning problems are categorized into.
- Describe popular machine learning techniques within the two primary categories of machine learning.
- Determine the machine learning technique that should be used to analyze data in a given real-world scenario.

# What is Structured Streaming?

[Click here](#) for the customer course enrollment link.

Duration: 1 hour

Course description: A common struggle that organizations face is how to accurately ingest and perform calculations on real-time data. This data is also referred to as streaming data, and the challenges behind working with it lie in its real-time nature – because it is constantly arriving, mechanisms must be put into place to process and write to a data store. In this course, you'll learn about Structured Streaming, an Apache Spark API that helps data practitioners overcome the challenges of working with streaming data. We'll cover fundamental concepts about batch and streaming data to help set the stage for our discussion on Structured Streaming. Then, we'll discuss where Structured Streaming fits into an organization's big data ecosystem. Finally, we'll review real-world Structured Streaming business use cases.

## Prerequisites & Requirements

- Prerequisites
  - Beginning knowledge about the Databricks Unified Data Analytics Platform (what it is, what it is used for)
  - Beginning knowledge about concepts related to the big data landscape (for example: structured streaming, batch processing, data pipelines)
  - Note: We recommend taking the following two Databricks Academy courses to help you prepare for this course: Fundamentals of Big Data and Fundamentals of Unified Data Analytics with Databricks.

## Learning objectives

- Explain the benefits of Structured Streaming for working with streaming data.
- Distinguish where Structured Streaming fits into an organization's big data ecosystem.
- Articulate examples of real-world business use cases for Structured Streaming.
- Describe popular machine learning techniques within the two primary categories of machine learning.

# What is the Databricks Lakehouse Platform?

[Click here](#) for the customer course enrollment link.

Duration: 30 minutes

Course description: This course is designed for everyone who is brand new to the Platform and wants to learn more about what it is, why it was developed, what it does, and the components that make it up.

Our goal is that by the time you finish this course, you'll have a better understanding of the Platform in general and be able to answer questions like: What is Databricks? Where does Databricks fit into my workflow? How have other customers been successful with Databricks?

NOTE: This course does not contain hands-on practice with the Databricks Lakehouse Platform.

Prerequisites:

- Experience using a web browser.

Learning objectives:

- Describe what the Databricks Lakehouse Platform is.
- Explain the origins of the Lakehouse data management paradigm.
- Outline fundamental problems that cause most enterprises to struggle with managing and making use of their data.
- Identify the most popular components of the Databricks Lakehouse Platform used by data practitioners, depending on their unique role.
- Give examples of organizations that have used the Databricks Lakehouse Platform to streamline big data processing and analytics.
- Describe security features that come built-in to the Databricks Lakehouse Platform.

# What's New in Apache Spark 3.0

[Click here](#) for the customer course enrollment link.

Duration: 30 minutes

Course description: This course was created to teach Databricks users about the major improvements to Spark in the 3.0 release. It will give an overview of new features meant to improve performance and usability. Students will also learn about backwards compatibility with 2.x and some of the considerations required for updating to Spark 3.0.

Prerequisites:

- Familiarity with Spark 2.x

Learning objectives:

- Describe major improvements to performance in Spark 3.0
- Identify major usability improvements in Spark 3.0
- Recognize relevant compatibility considerations for migrating to Spark 3.0