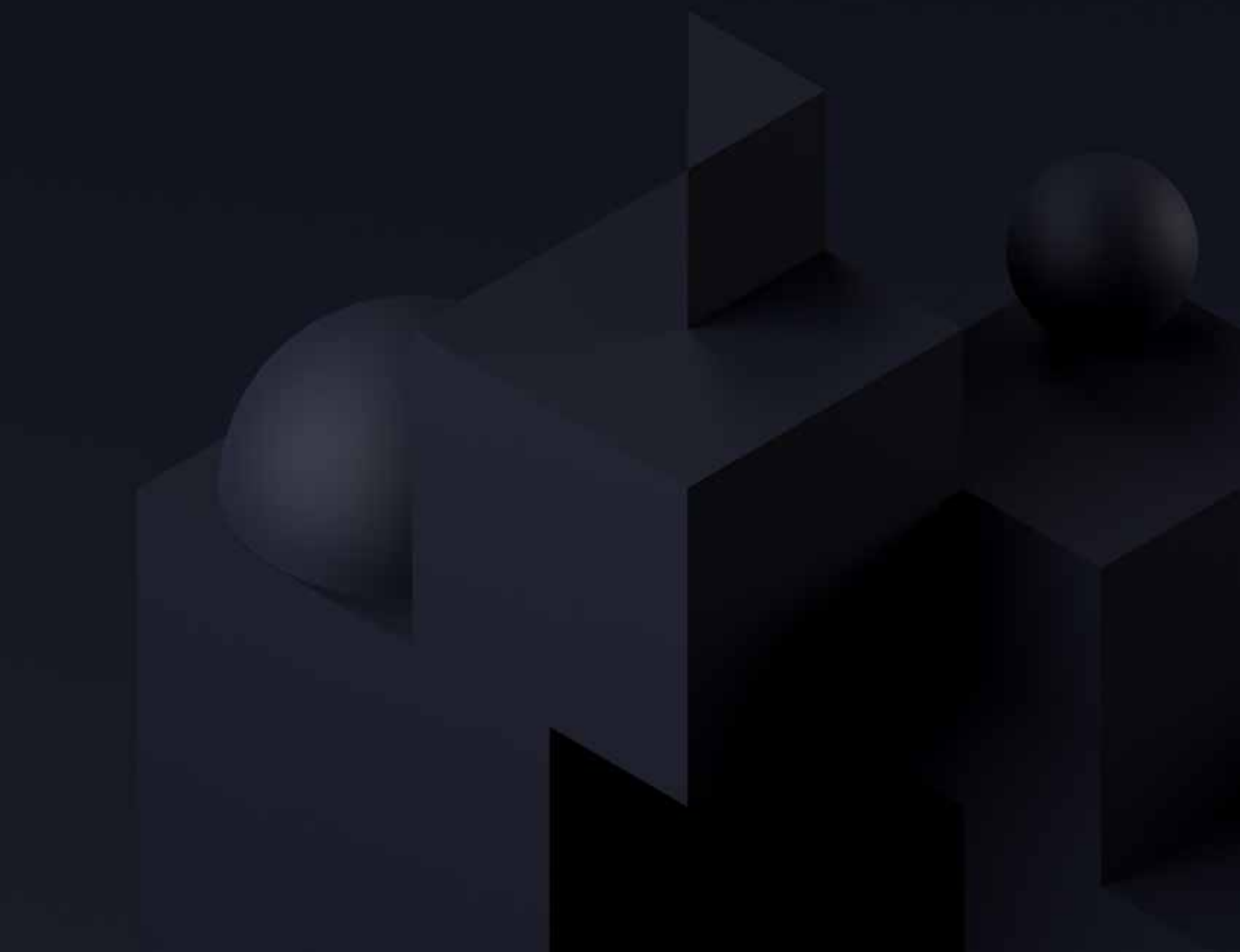
DATA+AI SUMMIT

BY databricks

2023 Trainings

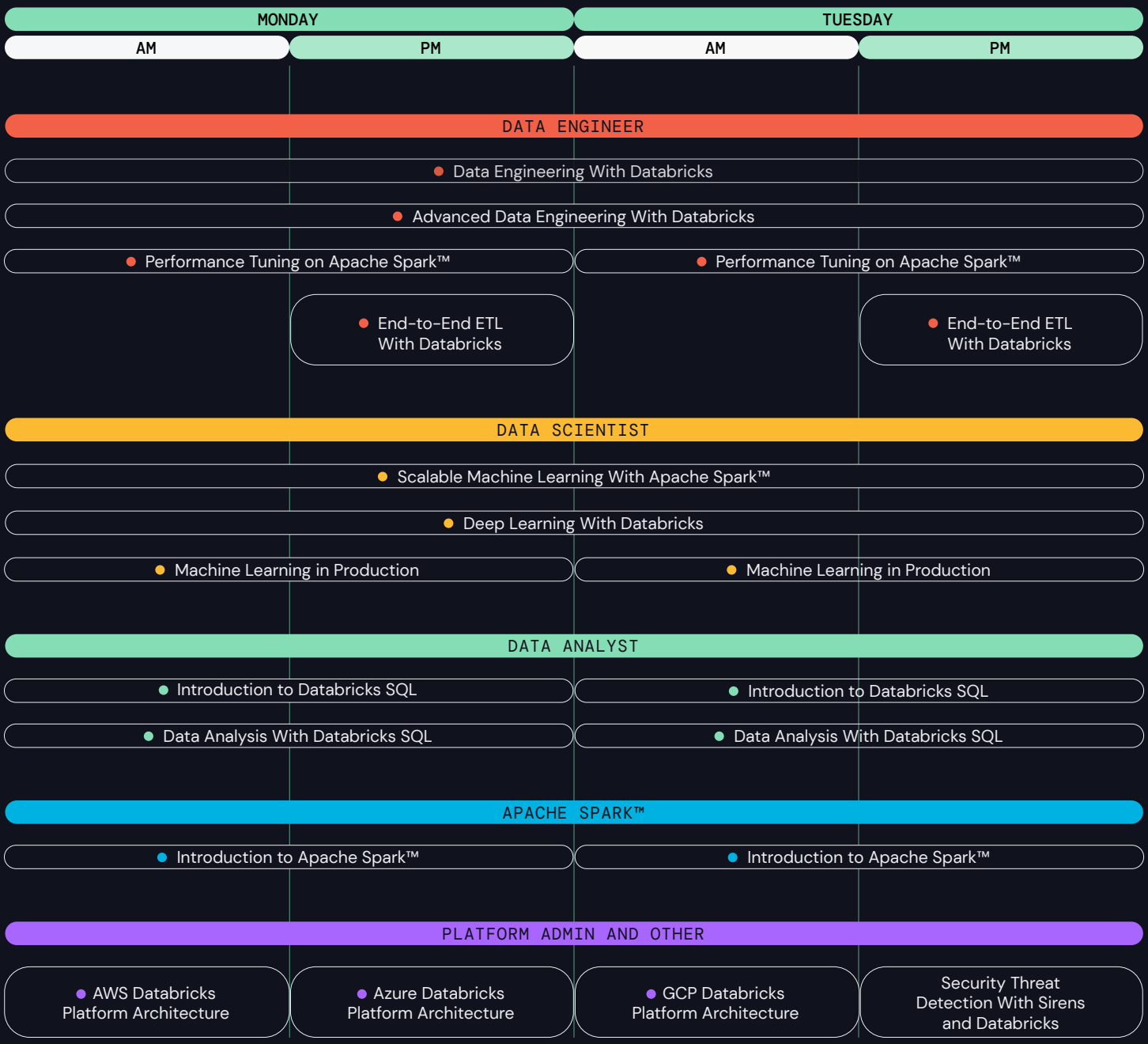


Contents

Training Courses

Certification Exams

Schedule


MONDAY
TUESDAY
AM
PM
AM
PM
DATA ENGINEER
Data Engineering With Databricks
Advanced Data Engineering With Databricks
Performance Tuning on Apache Spark™
Performance Tuning on Apache Spark™
End-to-End ETL With Databricks
End-to-End ETL With Databricks
DATA SCIENTIST
Scalable Machine Learning With Apache Spark™
Deep Learning With Databricks
Machine Learning in Production
Machine Learning in Production
DATA ANALYST
Introduction to Databricks SQL
Introduction to Databricks SQL
Data Analysis With Databricks SQL
Data Analysis With Databricks SQL
APACHE SPARK™
Introduction to Apache Spark™
Introduction to Apache Spark™
PLATFORM ADMIN AND OTHER
AWS Databricks Platform Architecture
Azure Databricks Platform Architecture
GCP Databricks Platform Architecture
Security Threat Detection With Sirens and Databricks

Training Courses

Data Engineering With Databricks

AUDIENCE:
Data Engineers

DURATION:
2 full days

HANDS-ON LABS:
Yes

CERTIFICATION PATH:
Databricks Certified Data Engineer Associate

/ Note: This course has options for taking it in either SQL or Python. /

Description:

This course prepares data professionals to leverage the Databricks Lakehouse Platform to productionalize ETL pipelines. Students will use Delta Live Tables with Spark SQL and Python to define and schedule pipelines that incrementally process new data from a variety of data sources into the Lakehouse. Students will also orchestrate tasks with Databricks Workflows and promote code with Databricks Repos.

By the end of the course, you will be able to:

- Use the Databricks Data Science and Engineering Workspace to perform common code development tasks in a data engineering workflow
- Use Spark SQL or PySpark to extract data from a variety of sources, apply common cleaning transformations, and manipulate complex data with advanced functions
- Define and schedule data pipelines that incrementally ingest and process data through multiple tables in the Lakehouse using Delta Live Tables in Spark SQL or Python
- Orchestrate data pipelines with Databricks Workflow Jobs and schedule dashboard updates to keep analytics up-to-date
- Configure permissions in Unity Catalog to ensure that users have proper access to databases for analytics and dashboarding

Prerequisites for both versions of this course (Spark SQL and PySpark):

- Beginner-level familiarity with cloud computing concepts (virtual machines, object storage, etc.)
- Production experience working with data warehouses and data lakes
- Intermediate-level experience with basic SQL concepts (select, filter, groupby, join, etc.)

Additional prerequisites for the Python version of this course (PySpark):

- Beginner-level programming experience with Python (syntax, conditions, loops, functions)
- Beginner-level programming experience with the Spark DataFrame API:
- Configure DataFrameReader and DataFrameWriter to read and write data

Advanced Data Engineering With Databricks

AUDIENCE:
Data Engineers

DURATION:
2 full days

HANDS-ON LABS:
Yes

CERTIFICATION PATH:
Databricks Certified Data Engineer Professional

Description:

In this course, students will build upon their existing knowledge of Apache SparkTM, Structured Streaming and Delta Lake to unlock the full potential of the data lakehouse by utilizing the suite of tools provided by Databricks. This course places a heavy emphasis on designs favoring incremental data processing, enabling systems optimized to continuously ingest and analyze ever-growing data. By designing workloads that leverage built-in platform optimizations, data engineers can reduce the burden of code maintenance and on-call emergencies, and quickly adapt production code to new demands with minimal refactoring or downtime. The topics in this course should be mastered prior to attempting the Databricks Certified Data Engineer Professional exam.

By the end of the course, you will be able to:

- Design databases and pipelines optimized for the Databricks Lakehouse Platform
- Implement efficient incremental data processing to validate and enrich data-driven business decisions and applications
- Leverage Databricks-native features for managing access to sensitive data and fulfilling right-to-be-forgotten requests
- Manage error troubleshooting, code promotion, task orchestration, and production job monitoring using Databricks tools

Prerequisites:

- Comfort using PySpark APIs to perform advanced data transformations
- Familiarity implementing classes with Python
- Experience using SQL in production data warehouse or data lake implementations
- Experience working in Databricks notebooks and configuring clusters
- Familiarity with creating and manipulating data in Delta Lake tables with SQL
- Ability to use Spark Structured Streaming to incrementally read from a Delta table

Performance Tuning on Apache Spark™

AUDIENCE:
Data Engineers

DURATION:
1 full day

HANDS-ON LABS:
Yes

Description:

Complete guided challenges as you learn to diagnose and fix poorly performing queries. Using Python/Scala, participants will review performance problems to uncover solutions and best practices to be applied to your queries.

By the end of the course, you will be able to:

- Deconstruct the Spark UI to aid in performance analysis, application debugging, and tuning of Spark applications
- Summarize some of the most common performance problems associated with data ingestion and how to mitigate them

Prerequisites:

- 6+ months of experience working with the Spark DataFrame API is recommended
- Intermediate-level programming experience in Python or Scala
- Intermediate-level experience using the Databricks platform (e.g., can describe cluster configuration, ingesting data into Databricks)

End-to-End ETL With Databricks

AUDIENCE:
Data Engineers and Analytics Professionals

DURATION:
1 half-day

HANDS-ON LABS:
Yes

*/ **Note:** The course "Data Engineering With Databricks" covers these concepts with additional hands-on time and a broader introduction to Databricks and is more suitable for students preparing to complete the Databricks Certified Associate Data Engineer exam. /*

Description:

This course prepares data professionals to leverage the Databricks Lakehouse Platform to productionalize ETL pipelines. Students will use Delta Live Tables with Spark SQL and Python to define and schedule pipelines that incrementally process new data from a variety of data sources into the Lakehouse. Students will also orchestrate tasks with Databricks Workflows and promote code with Databricks Repos.

By the end of the course, you will be able to:

- Ingest and enrich data for production applications
- Use Python and Spark SQL to build and deploy production data engineering pipelines
- Leverage the Databricks platform for code development and workload orchestration

Prerequisites:

- Experience building and maintaining production ETL pipelines with SQL or PySpark
- Beginner-level familiarity with cloud computing concepts (virtual machines, object storage, etc.)
- Production experience working with data warehouses and data lakes
- Beginner-level knowledge of the Databricks Workspace

Scalable Machine Learning With Apache Spark™

AUDIENCE:
Data Scientists and Machine Learning Engineers

DURATION:
2 full days

HANDS-ON LABS:
Yes

CERTIFICATION PATH:
Databricks Certified Machine Learning Associate

Description:

This course teaches you how to scale ML pipelines with Spark, including distributed training, hyperparameter tuning, and inference. You will build and tune ML models with SparkML while leveraging MLflow to track, version and manage these models. This course covers the latest ML features in Apache Spark, such as pandas UDFs, pandas Functions, and the pandas API on Spark, as well as the latest ML product offerings, such as Feature Store and AutoML.

Prerequisites:

- Intermediate-level experience with Python
- Familiarity with PySpark DataFrame API
- Experience building machine learning models

Deep Learning With Databricks

AUDIENCE:
Data Scientists and Machine Learning Engineers

DURATION:
2 full days

HANDS-ON LABS:
Yes

Description:

This course begins by covering the basics of neural networks and the tensorflow.keras API. We will then focus on using Spark to scale our models, including distributed training, hyperparameter tuning, and inference, and leverage MLflow to track, version and manage these models. You will apply model interpretability libraries to explain model predictions. Further, you will learn the concepts behind Convolutional Neural Networks (CNNs) and transfer learning, and apply them to solve image classification tasks. We will wrap up the course by covering Recurrent Neural Networks (RNNs) and attention-based models for natural language processing (NLP) applications.

Prerequisites:

- Intermediate-level experience with Python and pandas
- Familiarity with Apache Spark
- Working knowledge of machine learning and data science

Machine Learning in Production

AUDIENCE:
Data Scientists and Machine Learning Engineers

DURATION:
1 full day

HANDS-ON LABS:
Yes

CERTIFICATION PATH:
Databricks Certified Machine Learning Professional

Description:

In this course, you will learn MLOps best practices for putting machine learning models into production. The first half of the course uses a feature store to register training data and uses MLflow to track the machine learning lifecycle, package models for deployment, and manage model versions. The second half of the course examines production issues including deployment paradigms, monitoring and CI/CD. By the end of this course, you will have built an end-to-end pipeline to log, deploy and monitor machine learning models.

Prerequisites:

- Intermediate-level experience with Python and pandas
- Familiarity with Apache Spark
- Working knowledge of machine learning and data science

Introduction to Databricks SQL

AUDIENCE:
SQL Analysts, Data Analysts, Business Analysts

DURATION:
1 half-day

HANDS-ON LABS:
Yes

*/ **Note:** The course "Data Analysis With Databricks SQL" covers these concepts with additional hands-on time and a broader introduction to Databricks and is more suitable for students preparing to complete the Associate Data Analysis With Databricks certification exam. /*

Description:

Meet Databricks SQL and find out how you can achieve high performance while querying directly on your organization's data lake. Using Databricks SQL, learners will practice writing and visualizing queries. Students will leave this course with the ability to use Databricks SQL to write a variety of queries, create various visualizations, and combine their visualizations into a dashboard that can be shared with others.

By the end of the course, you will be able to use Databricks SQL to:

- Create a query
- Create a visualization from a query
- Combine multiple visualizations into a dashboard

The course, Data Analysis with Databricks SQL, is more detailed and is more suitable for students preparing to complete the Associate Data Analysis With Databricks certification exam.

Prerequisites:

- Basic familiarity with ANSI SQL

Data Analysis With Databricks SQL

AUDIENCE:
SQL Analysts, Data Analysts, Business Analysts

DURATION:
1 full day

HANDS-ON LABS:
Yes

CERTIFICATION PATH:
Databricks Certified Data Analyst Associate

Description:

Meet Databricks SQL and find out how you can achieve high performance while querying directly on your organization's data lake. Using Databricks SQL, learners will practice writing and visualizing queries. Students will leave this course having created a personal dashboard, complete with parameterized queries and automated alerts.

By the end of the course, you will be able to:

- Work with Delta Lake tables in Unity Catalog
- Implement a serverless SQL warehouse to save time and money
- Write queries that answer specific BI questions
- Visualize query output
- Produce a dashboard that combines multiple visualizations

Prerequisites:

- Basic familiarity with ANSI SQL

Introduction to Apache Spark™

AUDIENCE:
Data Engineers, Data Scientists, Data Architects

DURATION:
1 full day

HANDS-ON LABS:
Yes

CERTIFICATION PATH:
Databricks Certified Associate Developer for Apache SparkTM, when taken in conjunction with the subsequent courses in the Apache Spark Learning Pathway

/ Note: Unity Catalog is a Preview feature /

Description:

This course uses a case study–driven approach to explore the fundamentals of Spark Programming with Databricks, including Spark architecture, the DataFrame API, query optimization, and Structured Streaming. First, you will become familiar with Databricks and Spark, recognize their major components, and explore data sets for the case study using the Databricks environment. After ingesting data from various file formats, you will process and analyze data sets by applying a variety of DataFrame transformations, column expressions and built-in functions. Lastly, you will execute streaming queries to process streaming data and highlight the advantages of using Delta Lake.

By the end of the course, you will be able to:

- Define the major components of Spark architecture and execution hierarchy
- Describe how DataFrames are built, transformed and evaluated in Spark
- Apply the DataFrame API to explore, preprocess, join and ingest data in Spark
- Apply the Structured Streaming API to perform analytics on streaming data
- Navigate the Spark UI and describe how the catalyst optimizer, partitioning and caching affect Spark's execution performance

Prerequisites:

- Familiarity with basic SQL concepts (select, filter, group by, join and others)
- Beginner-level programming experience with Python (syntax, conditions, loops, functions, method chaining)

Databricks Platform Administration

AUDIENCE:
Platform Administrator

DURATION:
1 full day

HANDS-ON LABS:
No

Description:

This course instructs students in best practices for configuring Databricks and Unity Catalog, whether you administer a single workspace or an enterprise deployment spanning many cloud regions. Basic platform administration tasks around IAM, ACLs and workspace configuration will also be covered.

By the end of the course, you will be able to:

- Describe and configure Databricks identities
- Configure secure access to workspaces, compute resources and data
- Upgrade legacy data objects to Unity Catalog
- Describe and implement features to support continuous integration and deployment in the Databricks environment

Prerequisites:

- Basic familiarity with SQL
- Beginner-level knowledge of concepts related to identity access management
- Beginner-level knowledge of the Databricks Workspace
- Beginner-level familiarity with cloud computing concepts (virtual machines, object storage, etc.)

AWS Databricks Platform Architecture

AUDIENCE:
Platform Administrator

DURATION:
1 half-day

HANDS-ON LABS:
No

Description:

While the Databricks Lakehouse 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 you understand the cloud-specific portions of a Databricks deployment, highlighting integrations with first-party services in AWS to build scalable and secure applications.

Azure Databricks Platform Architecture

AUDIENCE:
Platform Administrator

DURATION:
1 half-day

HANDS-ON LABS:
No

Description:

While the Databricks Lakehouse 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 you understand the cloud-specific portions of a Databricks deployment, highlighting integrations with first-party services in Azure to build scalable and secure applications.

GCP Databricks Platform Architecture

AUDIENCE:
Platform Administrator

DURATION:
1 half-day

HANDS-ON LABS:
No

Description:

While the Databricks Lakehouse 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 you understand the cloud-specific portions of a Databricks deployment, highlighting integrations with first-party services in GCP to build scalable and secure applications.

Security Threat Detection With Sirens and Databricks

AUDIENCE:
Data Engineer, Security Engineer, Threat Hunter, Security Analyst, Security Data Scientist, Detection Engineer

DURATION:
1 half-day

HANDS-ON LABS:
Yes

Description:

Are you a security practitioner? Are you looking for better methods of threat detection, contextualization or threat hunting? Are you a data scientist wanting to work more closely with security ops teams? In this hands-on course, you will learn how easy it is to ingest data into Delta Lake, analyze DNS data, enrich it using threat intel, create detections using ML models and detect cyber criminals. You will use the new open source Sirens project from Databricks to ingest DNS data into a Cybersecurity Lakehouse followed by building and deploying a machine learning model on large-scale streaming data to detect threats. Not familiar with data science or Databricks? Not to worry. The course's live support staff has decades of security operations and data science experience.

Prerequisites:

- Familiarity with Python programming
- Familiarity with SQL
- Familiarity with security concepts of threat intelligence, ransomware, phishing
- Familiarity with the security operation concepts of threat detection and threat hunting

Certification Exams

Data Engineer Associate Certification Exam

AUDIENCE:
Data Engineer

DURATION:
1.5 hours

Description:

The Databricks Certified Data Engineer Associate certification exam assesses an individual's ability to use the Databricks Lakehouse Platform to complete introductory data engineering tasks. This includes an understanding of the Lakehouse Platform and its workspace, its architecture and its capabilities. It also assesses the ability to perform multi-hop architecture ETL tasks using Apache Spark™ SQL and Python in both batch and incrementally processed paradigms. Finally, the exam assesses the student's ability to put basic ETL pipelines and Databricks SQL queries and dashboards into production while maintaining entity permissions. Individuals who pass this certification exam can be expected to complete basic data engineering tasks using Databricks and its associated tools.

Data Engineer Professional Certification Exam

AUDIENCE:
Data Engineer

DURATION:
2 hours

Description:

The Databricks Certified Data Engineer Professional certification exam assesses an individual's skills in using Databricks to perform common data engineering tasks. This includes an understanding of the Databricks platform and developer tools like Apache Spark™, Delta Lake, MLflow and the Databricks CLI and REST API. It also evaluates the ability to build optimized and cleaned ETL pipelines. Additionally, modeling data into the Lakehouse using knowledge of general data modeling concepts will be assessed. Finally, ensuring that data pipelines are secure, reliable, monitored and tested before deployment will be included in this exam. Individuals who pass this certification exam can be expected to complete data engineering tasks using Databricks and its associated tools.

Associate Developer for Apache Spark™ Certification Exam

AUDIENCE:
Data Engineer, Data Scientist

DURATION:
2 hours

Description:

The Databricks Certified Associate Developer for Apache SparkTM 3.0 certification exam assesses the understanding of the Spark DataFrame API and the ability to apply the Spark DataFrame API to complete basic data manipulation tasks within a Spark session. These tasks include selecting, renaming and manipulating columns; filtering, dropping, sorting and aggregating rows; handling missing data; combining, reading, writing and partitioning DataFrames with schemas; and working with UDFs and Spark SQL functions. In addition, the exam will assess the basics of the Spark architecture like execution/deployment modes, the execution hierarchy, fault tolerance, garbage collection and broadcasting.

Data Analyst Associate Certification Exam

AUDIENCE:
Data Engineer

DURATION:
1.5 hours

Description:

The Databricks Certified Data Analyst Associate certification exam assesses an individual's ability to use the Databricks SQL service to complete introductory data analysis tasks. This includes an understanding of the Databricks SQL service and its capabilities, an ability to manage data with Databricks tools following best practices, using SQL to complete data tasks in the Lakehouse, creating production-grade data visualizations and dashboards, and developing analytics applications to solve common data analytics problems.

Machine Learning Associate Certification Exam

AUDIENCE:
Data Scientist

DURATION:
1.5 hours

Description:

The Databricks Certified Machine Learning Associate certification exam assesses an individual's ability to use Databricks to perform basic machine learning tasks. This includes an ability to understand and use Databricks Machine Learning and its capabilities like AutoML, Feature Store and select capabilities of MLflow. It also assesses the ability to make correct decisions in machine learning workflows and implement those workflows using Spark ML. Finally, an ability to understand advanced characteristics of scaling machine learning models is assessed. Individuals who pass this certification exam can be expected to complete basic machine learning tasks using Databricks and its associated tools.

Machine Learning Professional Certification Exam

AUDIENCE:
Machine Learning Engineer, Data Scientist

DURATION:
2 hours

Description:

The Databricks Certified Machine Learning Professional certification exam assesses an individual's use of Databricks Machine Learning and its capabilities to perform advanced machine learning in production tasks. This includes the ability to track, version and manage machine learning experiments and the machine learning model lifecycle. In addition, the certification exam evaluates the implementation of strategies for deploying machine learning models. Finally, participants will be assessed on their ability to build monitoring solutions to detect data drift. Individuals who pass this certification exam can be expected to perform advanced machine learning engineering tasks using Databricks Machine Learning.