Advanced Data Engineering with Databricks Training

Duration: **2 Day**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

Overview

What Can You Achieve with This Training?

- Strengthen your ability to develop and optimise **advanced data engineering workflows using Databricks, Apache Spark, and Delta Lake**.
- Learn how to process streaming data with **Spark Structured Streaming**, including aggregations, time windows, watermarks, and data quality controls.
- Explore **Unity Catalog, data anonymisation, and Change Data Feed (CDF)** to support data privacy and change management.
- Investigate performance bottlenecks using **Spark UI**, and understand techniques such as **Data Skipping, Liquid Clustering, and Shuffle optimisation**.
- Learn how to deploy Databricks projects across environments using **Databricks Asset Bundles (DABs)** and **CI/CD workflows**.
- Become familiar with the role of **Visual Studio Code, Databricks CLI, and GitHub Actions** in modern data engineering delivery.

The **Advanced Data Engineering with Databricks Training** is designed for professionals who want to build on their existing Databricks knowledge and develop more advanced data engineering capabilities.

The programme consists of **four modules, each lasting four hours**, providing a total of **16 hours of training**. It focuses on four areas: **streaming data processing, data privacy, performance optimisation, and automated deployment**.

The first module introduces **Apache Spark Structured Streaming** and **Delta Lake** for streaming workloads. Participants explore streaming queries, event-time processing, aggregations, time windows, watermarks, and **Streaming ETL** patterns. Data ingestion and quality enforcement are also covered.

The second module focuses on **data privacy and security within Databricks**. Topics include **Unity Catalog**, regulatory compliance, data isolation, pseudonymisation, anonymisation, and using **Change Data Feed (CDF)** to process and propagate data changes.

The third module examines **performance optimisation for Spark and Delta Lake workloads**. Learners use **Spark UI** concepts to investigate performance issues and explore how **Data Skipping, Liquid Clustering, Skew, Shuffles, Spill, and Serialization** affect application performance.

The final module introduces **DevOps, DataOps, and CI/CD practices** for Databricks projects. Participants learn how **Databricks Asset Bundles (DABs)** support project configuration, validation, deployment, and execution across multiple environments. The module also introduces **Databricks CLI, Visual Studio Code, and GitHub Actions** as tools for development and deployment workflows.

Advanced Data Engineering with Databricks Training

Duration: **2 Day**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

Throughout the course, **demonstrations and hands-on labs** reinforce the concepts and help participants develop practical approaches to creating efficient, secure, and maintainable Databricks data engineering solutions.

Prerequisites

This course is intended for participants with existing experience in Databricks and data engineering.

Learners should have:

- Experience using **PySpark APIs** for advanced data transformations
- Familiarity with implementing **Python classes**
- Experience using **SQL in production environments**, such as data warehouses or data lakes
- Hands-on experience with **Databricks notebooks and cluster configuration**
- Understanding of **creating and manipulating Delta Lake tables using SQL**

These prerequisites may be met by completing the following courses and earning the related certifications:

- **Data Engineering with Databricks (DBDE)**
- **Apache Spark Programming with Databricks (DBASP)**
- **Databricks Certified Data Engineer Associate**
- **Databricks Certified Associate Developer for Apache Spark**

For participants who do not meet one or more prerequisites, the following preparatory courses are recommended. They may be taken **in either order**:

- **DBDE – Data Engineering with Databricks**
- **DBASP – Apache Spark Programming with Databricks**

Advanced Data Engineering with Databricks Training

Duration: **2 Day**

BILGINÇ IT Academy
info@bilginc.com
bilginc.com

Who Should Attend

This course is particularly suitable for:

- **Data engineers** who want to optimise and scale data processing workloads in Databricks
- **Big data professionals** working with streaming and batch data
- **Cloud engineers** focusing on Delta Lake architectures and Structured Streaming
- Individuals preparing for the **Databricks Certified Data Engineer Professional** examination

Advanced Data Engineering with Databricks Training

Duration: **2 Day**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

What You Will Learn

By the end of this course, participants will be able to:

- Explain the role of **Spark Structured Streaming** and **Delta Lake** in streaming data processing.
- Understand **streaming queries** and the configuration of streaming data reads.
- Apply **Event Time, Time Windows, Aggregation, and Watermark** concepts to streaming scenarios.
- Use data ingestion and quality enforcement approaches in **Streaming ETL** workflows.
- Recognise **data privacy** and regulatory compliance considerations within Databricks.
- Understand the role of **Unity Catalog** in securing and managing access to data.
- Explain **pseudonymization, anonymization, and PII security** approaches.
- Understand how **Change Data Feed (CDF)** supports processing and propagating data changes.
- Use **Spark UI** concepts to analyse workload performance and investigate application issues.
- Explain the role of **Data Skipping and Liquid Clustering** in performance optimisation.
- Recognise performance issues associated with **Skew, Shuffles, Spill, and Serialization**.
- Assess how **cluster configuration and instance type selection** influence workload performance.
- Explain how **DevOps, DataOps, CI/CD, and testing practices** apply to Databricks projects.
- Describe the structure, core components, and templates of **Databricks Asset Bundles (DABs)**.
- Use the **Databricks CLI** to configure, validate, deploy, and execute DAB projects.
- Prepare **DAB deployments for multiple target environments**.
- Understand local development and testing workflows using **Visual Studio Code**.
- Recognise how **GitHub Actions** can support automated deployment pipelines.
- Apply concepts that support **automated Databricks project deployment and performance improvement**.

Outline

1. Databricks Streaming and Delta Live Tables

This module focuses on **Spark Structured Streaming** and **Delta Lake**, including streaming queries, aggregations, and data quality enforcement.

Advanced Data Engineering with Databricks Training

Duration: **2 Day**

BİLGİNÇ IT Academy
info@bilginc.com
bilginc.com

Topics Covered

- **Streaming Data Concepts**
- Introduction to **Structured Streaming**
- **Demo:** Reading from a Streaming Query
- **Streaming from Delta Lake**
- **Lab:** Streaming Query Lab
- **Aggregation, Time Windows, and Watermarks**
- **Event Time** and aggregations over time windows
- **Lab:** Stream Aggregation Lab
- **Demo:** Windowed Aggregation with Watermark
- **Data Ingestion Patterns**
- **Demo:** Auto Load to Bronze
- **Demo:** Stream from Multiplex Bronze
- **Quality Enforcement Patterns**
- **Demo:** Quality Enforcement
- **Lab:** Streaming ETL Lab

Participants examine **Structured Streaming computation models**, configuration for streaming reads, and approaches to maintaining **data quality in streaming environments**.

2. Databricks Data Privacy

This module examines data privacy, security, regulatory considerations, and data change management on the Databricks platform.

Topics Covered

- **Regulatory Compliance**
- **Data Privacy**
- **Key Concepts and Components**
- **Auditing Your Data**
- **Data Isolation**
- **Demo:** Securing Data in Unity Catalog

BİLGİNÇ IT Academy

info@bilginc.com

bilginc.com

- **Pseudonymization and Anonymization**
- **Summary and Best Practices**
- **Demo:** PII Data Security
- **Capturing Changed Data**
- **Deleting Data in Databricks**
- **Demo:** Processing Records from CDF and Propagating Changes
- **Lab:** Propagating Changes with CDF Lab

This module helps data engineering professionals understand how to perform Databricks-related tasks while considering **privacy, security, and controlled data changes**.

3. Databricks Performance Optimization

This module explores methods for improving**Apache Spark and Delta Lake workload performance**.

Topics Covered

- Introduction to **DevOps Spark UI**
- Introduction to **Designing Foundation**
- **Demo:** File Explosion
- **Data Skipping and Liquid Clustering**
- **Lab:** Data Skipping and Liquid Clustering
- **Skew**
- **Shuffles**
- **Demo:** Shuffle
- **Spill**
- **Lab:** Exploding Join
- **Serialization**
- **Demo:** User-Defined Functions
- **Fine-Tuning:** Choosing the Right Cluster
- Selecting the **Best Instance Types**

Participants explore how to use **Spark UI** to assess performance and investigate application issues. The module also considers performance-related topics such as **physical data layout, caching, and Photon**.

BİLGİNÇ IT Academy

info@bilginc.com

bilginc.com

4. Automated Deployment with Databricks Asset Bundles

This module examines how **Databricks Asset Bundles (DABs)** support the development, configuration, and automated deployment of data engineering projects.

Topics Covered

- **DevOps Review**
- Review of **Continuous Integration and Continuous Deployment/Delivery (CI/CD)**
- **Demo:** Course Setup and Authentication
- **Deploying Databricks Projects**
- Introduction to **Databricks Asset Bundles (DABs)**
- **Demo:** Deploying a Simple DAB
- **Lab:** Deploying a Simple DAB
- **Variable Substitutions in DABs**
- **Demo:** Deploying a DAB to Multiple Environments
- **Lab:** Deploy a DAB to Multiple Environments
- **DAB Project Templates Overview**
- **Lab:** Use a Databricks Default DAB Template
- **CI/CD Project Overview with DABs**
- **Demo:** Continuous Integration and Continuous Deployment with DABs
- **Lab:** Adding ML to Engineering Workflows with DABs
- **Developing Locally with Visual Studio Code (VSCode)**
- **Demo:** Using VSCode with Databricks
- **CI/CD Best Practices for Data Engineering**
- **Next Steps:** Automated Deployment with GitHub Actions

The module begins by reviewing **DevOps and DataOps principles**, along with CI/CD and testing practices relevant to data engineering pipelines.

Participants are introduced to deployment tools such as the **Databricks REST API, SDK, and CLI**, before exploring the structure and key components of **Databricks Asset Bundles**.

They learn how DABs can simplify deployment across different target environments and how to use the **Databricks CLI** to modify, validate, deploy, and run bundles with environment-specific

BİLGİNÇ IT Academy

info@bilginc.com

bilginc.com

configurations and variable substitutions.

The course also introduces **Visual Studio Code** as an integrated development environment for building, testing, and deploying DAB projects locally.

The module concludes with an introduction to **GitHub Actions** for automating deployment pipelines and strengthening CI/CD workflows.