

## **Databricks and Lakehouse Platform (5 days)**

**By Dr. Vishwanath Rao**

### **Introduction to Databricks**

- Intro to Data Lakehouse
- Intro to Databricks Lakehouse Platform
- Intro to Databricks Lakehouse Platform Architecture and Security Fundamentals
- Intro to Supported Workloads on the Databricks Lakehouse Platform
- What is Databricks
- Create the Databricks cluster and workspace
- Understanding Databricks Notebooks
- Working with Datastores
- Working with Datasets

### **Spark Architecture Basics**

- Overview of Spark Architecture
- The architecture of Databricks spark cluster
- The architecture of Spark Job

### **Read and Write Data in the Databricks**

- Read Data in CSV and JSON Format
- Read Data in the Parquet Format
- Read Data stored in views and tables
- Write the data

### **Dataframes in Databricks**

- What is a Dataframe?
- Using the Common Dataframe Methods
- Display Function
- Create a DataFrame from a CSV file
- Configure options to read a CSV file
- How to reference columns of a DataFrame
- Understand the DataFrame Schema: Part 1
- Understand the DataFrame Schema: Part 2
- Specify a DataFrame Schema using a DDL-formatted string : Part 1
- Specify a DataFrame Schema using a DDL-formatted string : Part 2
- Spark Architecture: The Organization of a DataFrame

## Dataframe APIs

- Adding columns to a DataFrame
- Renaming columns of a DataFrame
- Removing columns from a DataFrame
- Filtering rows from a DataFrame
- Joining multiple DataFrames: Part 1
- Joining multiple DataFrames: Part 2
- Aggregation: Count
- Aggregation: Count Distinct
- Aggregation: Get the Min value
- Aggregation: Get the Max value
- Aggregation: Get the Sum and SumDistinct
- Aggregation: Average and Mean
- Aggregation: Grouping data - Part 1
- Aggregation: Grouping data - Part 2
- Apache Spark Architecture: How Apache Spark Transforms data Internally
- User Defined Function

## Spark SQL

- Run SQL on a DataFrame: TempView
- Run SQL on a DataFrame: GlobalView
- Databases: List, Create, Delete, Select
- Tables: Unmanaged
- Tables: Managed
- SQL Fundamentals: Select Clause & Select Expression
- SQL Fundamentals: Where Clause, Equality Checks
- SQL Fundamentals: Handling NULLs in Where Clause
- SQL Fundamentals: Aggregations - Sum, Count, AVG, Mean
- SQL Fundamentals: Group By Clause
- SQL Fundamentals: Having Clause
- SQL Fundamentals: Order By Clause
- SQL Fundamentals: Inner Joins
- SQL Fundamentals: Left Outer Joins
- SQL Fundamentals: Right Outer Joins
- SQL Fundamentals: Predicates and Operators, like predicate
- SQL Fundamentals: Case Expressions

## **Lazy Evaluation and Performance features of Databricks**

- Differentiate the Lazy and Eager function
- How does the Catalyst Optimizer work?
- Identify Transformations and Actions
- Performance Enhancements Enabled by the Tungsten and Shuffle operations

## **DataFrames Columns in Databricks**

- Column Class
- Working with the Column Expressions

## **Dataframes Advanced Methods in Databricks**

- Perform time and data manipulation
- Using Aggregate functions

## **Platform Architecture, Data protection in the Databricks**

- Databricks platform architecture
- Perform data protection

## **Building and Querying a Data Lake**

- Open Source Delta Lake
- How Databricks manages Delta Lake

## **Process the Streaming Data with the Databricks structured streaming**

- Databricks structured streaming
- Performing the Stream Processing through the structured streaming
- Working with the Time Windows
- Process the data from the Event Hubs with the structured streaming

## **Databricks Delta Lake Architecture**

- Bronze, Gold, and Silver Architecture
- Performing the Batch and stream processing

## **Creating the production workloads on Databricks with the Data factory**

- Scheduling the Databricks jobs in the data factory pipeline
- Passing the Parameters in and out of the Databricks jobs in the data factory