**Apache Spark Programming with Databricks**

**Duration: 40 Hours**

**Prerequisites:**

* Basic programming knowledge in Python or Scala.
* Familiarity with SQL and relational databases.
* Understanding of data analytics concepts.
* Basic knowledge of distributed systems (preferred but not mandatory)

**Scope of the Training:**

* Build a solid foundation in Spark programming and Databricks.
* Enable participants to work with real-world datasets for ETL and analytics.
* Provide hands-on experience in building scalable data pipelines and streaming applications.
* Prepare participants for roles such as Data Engineer, Big Data Developer, and ETL Engineer.

**Module 1: Introduction to Apache Spark and Databricks**

* Overview of Apache Spark
* Key features of Databricks Unified Analytics Platform
* Advantages of using Databricks for Apache Spark
* Setting up Databricks workspace

**Module 2: Spark Core Concepts**

* Spark Architecture and Cluster Overview
* RDDs (Resilient Distributed Datasets)
* Transformations and Actions in RDDs
* Lazy Execution and DAG

**Module 3: DataFrames and Datasets**

* Understanding DataFrames and Datasets
* Schema Management and Optimization
* Operations: Joins, Aggregations, and Filters
* Working with JSON, Parquet, and Avro files

**Module 4: Spark SQL**

* SQL in Spark: Overview and Usage
* Creating and Managing Tables with Spark SQL
* Performance Optimization with Catalyst Optimizer
* Using Databricks SQL for Queries and Visualization

**Module 5: Streaming with Spark Structured Streaming**

* Real-Time Data Processing Concepts
* Sources: Kafka, Delta Lake, and Files
* Window Functions and Watermarking
* Fault Tolerance and Checkpointing

**Module 6: Machine Learning with Spark MLlib**

* MLlib Overview and Architecture
* Feature Engineering and Transformation
* Building Classification and Regression Models
* Model Persistence and Deployment in Databricks

**Module 7: Performance Tuning and Optimization**

* Understanding Spark Execution Plans
* Managing Partitioning and Caching
* Advanced Performance Tuning Techniques
* Debugging and Monitoring Spark Applications

**Module 8: Hands-On Projects**

* Building ETL Pipelines with Spark
* Real-Time Analytics Use Cases
* Implementing Machine Learning Models