PySpark Basics

Duration – 16 Hours

Program Description

This basic PySpark training program provides a comprehensive understanding of PySpark architecture, including SparkSession, RDDs, DataFrames, and Datasets. Participants will gain hands-on experience in setting up the PySpark environment, loading, and exploring data from various formats (CSV, JSON, Parquet).

The course covers SQL queries on DataFrames, column operations, and advanced data transformations such as aggregations, handling missing data, and type casting.

Additionally, participants will learn how to perform joins and union operations on DataFrames. The program culminates with a Capstone Project to apply the learned concepts in real-world scenarios.

Learning Goals

- Understand the fundamentals of PySpark, including SparkSession, and DataFrames.
- Perform data ingestion, exploration, transformation, and cleaning on datasets.
- Build and evaluate an end-to-end PySpark workflow for actionable business insights.

Course Topics

- PySpark architecture, SparkSession, RDDs, DataFrames, and Datasets.
- Setting up the PySpark Environment
- Data Processing - Loading and Exploring DataFrames & SQL Queries on DataFrames
- Data Transformation - Column Operations, Aggregations and Grouping, Handling Missing Data and Type Casting, Joins and Union Operations
- Design and implement a complete PySpark-based data pipeline to process, clean, and analyze large datasets, enabling actionable insights & analysis in real-world scenarios.

PySpark Intermediate

Duration – 16 Hours

Program Description

This PySpark training program equips participants with advanced data transformation techniques, including handling nested data, exploding arrays, and optimizing performance for complex JSON structures. It covers window functions for advanced analytical operations and optimized join strategies such as broadcast and skewed joins.

The course focuses on improving data serialization and scaling data aggregations using approximation techniques. Participants will also gain insights into PySpark caching mechanisms for better performance.

Real-time data processing concepts are introduced through structured streaming. The program concludes with a Capstone Project, enabling participants to apply their learning in real-world data scenarios.

Learning Goals

- Perform advanced transformations on complex nested datasets using PySpark.
- Optimize PySpark workflows with caching, serialization, and advanced join strategies.
- Implement real-time data processing pipelines for streaming data analysis.

Course Topics

- Advanced Data Transformations and Nested Structure Processing in PySpark
- Advanced Join Strategies: Broadcast Joins, Skewed Joins
- Optimizing Data Serialization
- Aggregating at Scale with Approximation Techniques
- Understanding PySpark Caching Mechanisms
- RealTime Data Processing with PySpark
- Build a real-time data pipeline using PySpark Structured Streaming to process and analyze live data, enabling timely insights and proactive decision-making and optimization.

PySpark Advanced

Duration – 24 Hours

Program Description

This Spark training program focuses on optimizing Spark memory management by understanding its architecture, execution, and storage settings for peak performance. Participants will learn advanced performance management techniques, including partitioning, caching, and monitoring Spark SQL operations with Spark UI.
The course covers the features and advantages of Spark 3.x, such as Adaptive Query Execution (AQE), skew join optimization, and GPU acceleration for improved resource management. Debugging and troubleshooting Spark applications is addressed through logging, profiling, and debugging tools to resolve errors efficiently.
The program culminates with a Capstone Project, providing hands-on experience to apply learned concepts to real-world Spark applications.

Learning Goals

- Develop a comprehensive understanding of Spark's memory management architecture, including execution and storage memory, and learn how to configure and tune memory settings for diverse workloads.
- Master partitioning, caching, and persistence strategies to optimize Spark application performance, along with efficient utilization of broadcast variables and serialization techniques.
- Leverage the advanced features of Spark 3.x, such as Adaptive Query Execution, GPU acceleration, and enhanced structured streaming, to boost performance and scalability in data-intensive workflows.
- Gain expertise in debugging and troubleshooting Spark applications, using tools like Spark Web UI, logging levels, and advanced debugging frameworks to ensure seamless and error-free application execution.

Course Topics

- Spark Memory Architecture, Configuring Spark Memory Settings, Spark Execution Memory and Storage Memory, Tuning Memory Settings for Optimal Performance
- Partitioning, caching, and persistence strategies to optimize data distribution and performance in Spark applications. Learn to leverage broadcast variables, serialization techniques, and Spark UI for monitoring and tuning Spark SQL and DataFrame operations effectively.
- Advanced optimization techniques in Spark, including Adaptive Query Execution (AQE) with dynamic partition pruning, skew join optimization, and runtime bloom filter pushdown for efficient query execution. Explore enhanced structured streaming, improved ANSI SQL compatibility, and GPU acceleration for managing resources, scheduling, memory, and SQL operations to boost performance.
- Spark debugging and monitoring by utilizing logging levels, debugging tools, and the Spark Web UI for resolving common errors, profiling performance, and identifying bottlenecks. Gain expertise in debugging Spark Streaming, SQL, and DataFrame issues to ensure efficient and error-free application execution.