BISP - PySpark Essential Training

Course Summary: In this course, you will learn Apache Spark framework and its components. Pyspark is an interactive layer of Spark built on python. You can leverage all Spark capabilities through Pyspark. During the training, We demonstrate how to build your data products over spark using Spark streaming, Spark RDDs, Spark SQL, Spark MLIB, Kafka and Flume.We also discuss in depth architecture of Spark and differences between Map Reduce and Spark.

1. Introduction to Spark

1.1. Origins of Spark
1.2. Understanding Spark
1.3. Where Spark Shines
1.4. Introduction to Notebooks

2. Spark Architecture and components/Installation

2.1. Overview of Spark Components
2.2. Spark Vs Hadoop
2.3. Challenges Spark addresses
2.4. Installation
2.5. Create and Configure Spark Cluster
2.6. Performance benchmarking – How Spark is faster than Hadoop

3. Working with RDDs – Execution on Spark Engine (Behind the scenes)

3.1. What is RDD
3.2. Spark transformations in RDD
3.3. Actions in RDD
3.4. Loading and Saving Data in RDD
3.5. RDD – key value pair
3.6. Broadcast Variables

4. Memory Management/Fault Tolerance/Lazy evaluation

4.1. RDD fault tolerance
4.2. In Memory computing

4.3. Lazy evaluation and its advantages

5. Dataframes – aggregate/filter/sort/transform (Actions/Transformations)

5.1. Introduction to Actions (take/collect/reduce/reduceByKey/foreach/histogram etc)
5.2. Introduction to Transformations(aggregate/.sql/joins/.distinct/temporary table creation etc)
5.3. Creating DataFrames
5.4. Specifying schema for a dataframe
5.5. Interacting and transforming dataframes

6. Introduction to Pyspark

6.1. Apache Spark Stack
6.2. Newest capabilities of Pyspark
6.3. Spark Execution process

7. Pyspark SQL and Dataframes

7.1. Spark SQL architecture
7.2. Interacting with RDDs and converting objects to DataFrames
7.3. SQL context in Spark SQL
7.4. Performance Tuning
7.5. Data processing with Spark Dataframes and UDFs
7.6. Select/Filter/aggregate/sort/presenting the data

8. Apache Kafka and Flume

8.1. Introduction to Kafka and Flume
8.2. Creating and configuring Kafka Cluster
8.3. Kafka architecture
8.4. Basic Kafka operations

9. Pyspark Streaming

9.1. Introduction to Spark Streaming
9.2. Transformations using Dstreams
9.3. Receiver based approach and Direct approach
9.4. Streaming context setup
9.5. Querying streaming data

10. Using Mlib in Spark for Machine Learning

10.1 Introduction to machine learning with Spark
10.2 Preparing data for Machine Learning
10.3 Building a linear regression model
10.4 Evaluating a linear regression model
10.5 visualizing a linear regression model

Hands-on :

1. Install Spark and Build Spark Applications
2. How to use Dataframes, functions
3. Load Data in RDDs, RDD transformations
4. RDD actions and functions, partitioning
5. Spark SQL, Spark-Hive Integration, Loading data and querying using Spark SQL
6. Create ETL pipelines using Pyspark
7. Setting up Kafka Cluster, Kafka - Flume Integration
8. Spark – Flume Integration
9. Applying ML models and ML workflow utilities