ASM Educational Center

ASM Educational Center (ASM)

25 YEARS Service Excellence

Spark V2 For Developers

Course Outline:

Scala primer

- A quick introduction to Scala
- Labs : Getting know Scala

Spark Basics

- Big Data, Hadoop, Spark
- What’s new in Spark v2
- Spark concepts and architecture
- Spark eco system (core, spark sql, mlib, streaming)
- Labs : Installing and running Spark

Spark Shell

- Spark shell
- Spark web UIs
- Analyzing dataset – part 1
- Labs: Spark shell exploration

RDDs (Condensed coverage)

- RDDs concepts
- Partitions
- RDD Operations / transformations

- More detailed coverage if required : RDD types, Key-Value pair RDDs, MapReduce on RDD
- Labs : Unstructured data analytics using RDDs

Spark Dataframes & Datasets

- Learning about Dataframe / Dataset
- Programming in Dataframe / Dataset API
- Loading structured data using Dataframes
- Caching and persistence
- Labs : Dataframes, Datasets, Caching

Spark API programming (Scala / Python)

- Introduction to Spark API
- Submitting the first program to Spark
- Debugging / logging
- Configuration properties
- Labs : Programming in Spark API, Submitting jobs

Spark SQL

- Spark SQL concepts and overview
- Defining tables and importing datasets
- Querying data using SQL
- Handling various storage formats : JSON / Parquet / ORC
- Labs : querying structured data using SQL; evaluating data formats

Spark and Hadoop

- Hadoop Primer : HDFS / YARN
- Hadoop + Spark architecture
- Running Spark on Hadoop YARN
- Processing HDFS files using Spark
- Spark & Hive

Machine Learning (ML / MLib)

- Machine Learning primer
- Machine Learning in Spark : MLib / ML
- Spark ML overview (newer Spark2 version)
- Algorithms : Clustering, Classifications, Recommendations
- Labs : Writing ML applications

GraphX

- GraphX library overview
- GraphX APIs
- Labs : Processing graph data using Spark

Spark Streaming

- Streaming overview
- Evaluating Streaming platforms
- Streaming operations
- Sliding window operations
- Structured Streaming

- Labs : Writing spark streaming applications

Spark Performance and Tuning

- Broadcast variables
- Accumulators
- Memory management & caching