DVS SPARK Course Content

MODULE 1 - INTRODUCTION AND EVOLUTION OF APACHE SPARK

- What is Apache Spark & Why Spark?
- Spark History
- Unification in Spark
- Spark ecosystem Vs Hadoop
- Spark with Hadoop

MODULE 2 – SCALA BASICS (Object Oriented and Functional Programming)

- Introduction to Functional Programming
- Interactive Shell – REPL, Data types, Variables, Expressions, Conditional statements, Loops – For comprehension
- Pattern Matching in Scala with Match expression
- Simple Functions and their variants, Tail Recursion, Functions as Objects aka Anonymous functions, Higher Order Functions
- Scala Collections and the usage of higher order methods on Collections
- Classes and Objects, Class Constructors in Scala, Case classes, Abstract and Generic Class
- Exception Handling in Scala
- Traits in Scala, Properties of Traits
- Magic Apply method
- Singleton and Companion objects
- Implicits in Scala – Implicit parameters, def, classes

MODULE 3 - DOWNLOADING SPARK AND GETTING STARTED

- Installing Spark
- Introduction to Spark's Python and Scala Shells
- Spark Standalone Cluster Architecture and its application flow
- Spark on YARN and its application flow

MODULE 4 - PROGRAMMING WITH RDDS

- RDD Basics and its characteristics, Creating RDDs
- RDD Operations
- Transformations
- Actions
- RDD Types
- Lazy Evaluation
- Persistence (Caching)

MODULES 5 - ADVANCED SPARK PROGRAMMING

- Accumulators and Fault Tolerance

- Broadcast Variables
- Custom Partitioning

MODULE 6 - LOADING AND SAVING YOUR DATA

- Dealing with different file formats (Text, CSV, JSON files etc.)
- Hadoop Input and Output Formats
- Connecting to diverse Data Sources (HDFS, Hive, S3, RDBMS and NoSQL etc.)

MODULE 7 - SPARK SQL

- Linking with Spark SQL
- Initializing Spark SQL
- Data Frames & Caching
- Case Classes, Inferred Schema
- Loading and Saving Data
- Apache Hive
- Data Sources/Parquet
- JSON
- JDBC/ODBC Server
- Spark SQL User Defined Functions (UDFs)
- Hive UDFs

MODULE 8 - SPARK STREAMING

- Batch vs Streaming
- Architecture and Abstraction
- DStreams, DStreams vs RDD
- Transformations
- Input Streams (Socket, HDFS, Twitter, Kafka)
- Check pointing, Persist and Caching
- Batch and Window Sizes
- Level of Parallelism

MODULE 9 - MACHINE LEARNING WITH MLLIB

- Machine Learning Basics and terminology
- Apache Spark MLLib Algorithms
- Examples implementing Machine Learning algorithms using Spark MLLib – Linear Regression

Overview sessions on Cassandra, Kafka

Project with Spark SQL and Spark Streaming using Kafka & Cassandra