PROGRAMMING APACHE SPARK

By Dr. Vishwanath Rao

Day 1

What is Big Data?
Big Data Customer Scenarios
Limitations and Solutions of Existing Data Analytics Architecture with Uber Use Case
How Hadoop Solves the Big Data Problem?
What is Hadoop?
Hadoop's Key Characteristics
Hadoop Ecosystem and HDFS
Hadoop Core Components
Rack Awareness and Block Replication
YARN and its Advantage
Hadoop Cluster and its Architecture
Hadoop: Different Cluster Modes
Hadoop Terminal Commands
Big Data Analytics with Batch & Real-time Processing
Why Spark is needed?
What is Spark?
How Spark differs from other frameworks?
Spark at Yahoo!

What is Scala?
Why Scala for Spark?
Scala in other Frameworks
Introduction to Scala REPL
Basic Scala Operations
Variable Types in Scala
Control Structures in Scala
Foreach loop, Functions and Procedures
Collections in Scala- Array
ArrayBuffer, Map, Tuples, Lists, and more

Functional Programming
Higher Order Functions
Anonymous Functions
Class in Scala
Getters and Setters
Custom Getters and Setters
Properties with only Getters
Auxiliary Constructor and Primary Constructor
Singletons
Extending a Class
Overriding Methods
Traits as Interfaces and Layered Traits

Day 2

Spark's Place in Hadoop Ecosystem
Spark Components & its Architecture
Spark Deployment Modes
Introduction to Spark Shell
Writing your first Spark Job Using SBT
Submitting Spark Job
Spark Web UI
Data Ingestion using Sqoop

Challenges in Existing Computing Methods
Probable Solution & How RDD Solves the Problem
What is RDD, It's Operations, Transformations & Actions
Data Loading and Saving Through RDDs
Key-Value Pair RDDs
Other Pair RDDs, Two Pair RDDs
RDD Lineage
RDD Persistence
WordCount Program Using RDD Concepts
RDD Partitioning & How It Helps Achieve Parallelization
Passing Functions to Spark

Day 3

Need for Spark SQL
What is Spark SQL?
Spark SQL Architecture
SQL Context in Spark SQL
User Defined Functions
Data Frames & Datasets
Interoperating with RDDs
JSON and Parquet File Formats
Loading Data through Different Sources
Spark – Hive Integration

Why Machine Learning?
What is Machine Learning?
Where Machine Learning is Used?
Face Detection: USE CASE
Different Types of Machine Learning Techniques
Introduction to MLlib
Features of MLlib and MLlib Tools
Various ML algorithms supported by MLlib

Day 4

Supervised Learning - Linear Regression, Logistic Regression, Decision Tree, Random Forest Unsupervised Learning - K-Means Clustering & How It Works with MLlib

- Analysis on US Election Data using MLlib (K-Means)
- Machine Learning MLlib
- K- Means Clustering
- Linear Regression
- Logistic Regression
- Decision Tree
- Random Forest

Need for Kafka
What is Kafka?
Core Concepts of Kafka
Kafka Architecture
Where is Kafka Used?

Understanding the Components of Kafka Cluster
Configuring Kafka Cluster
Kafka Producer and Consumer Java API
Need of Apache Flume
What is Apache Flume?
Basic Flume Architecture
Flume Sources
Flume Sinks
Flume Channels
Flume Configuration
Integrating Apache Flume and Apache Kafka

Day 5

Drawbacks in Existing Computing Methods
Why Streaming is Necessary?
What is Spark Streaming?
Spark Streaming Features
Spark Streaming Workflow
How Uber Uses Streaming Data
Streaming Context & DStreams
Transformations on DStreams
Describe Windowed Operators and Why it is Useful
Important Windowed Operators
Slice, Window and ReduceByWindow Operators
Stateful Operators

Apache Spark Streaming: Data Sources
Streaming Data Source Overview
Apache Flume and Apache Kafka Data Sources
Example: Using a Kafka Direct Data Source
Perform Twitter Sentimental Analysis Using Spark Streaming