Spark v2 For Data Analysts

Course Outline:

- **Scala primer**
  - A quick introduction to Scala
  - Labs : Getting know Scala
- **Spark Basics**
  - Big Data , Hadoop, Spark
  - Spark concepts and architecture
  - Spark eco system (core, spark sql, mlib, streaming)
  - Labs : Installing and running Spark
- **First Look at Spark**
  - Spark shell
  - Spark web UIs
  - Analyzing dataset – part 1
  - Labs: Spark shell exploration
- **RDDs (condensed coverage)**
  - RDDs concepts
  - Partitions
  - RDD Operations / transformations
  - Labs : Unstructured data analytics using RDDs
- **Dataframes / Datasets**
  - Understanding newer Dataset API
  - Dataframes
  - Loading structured data using Dataframes
  - Caching and persistence

- Labs : Dataframes, Datasets, Caching

- **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) (day – 3)**
  - Machine Learning primer
  - Machine Learning in Spark : MLib / ML
  - Spark ML overview (newer Spark2 version)
  - Algorithms : Clustering, Classifications, Recommendations
  - Labs : Writing ML applications
- **GraphX (day – 3)**
  - GraphX library overview
  - GraphX APIs
  - Labs : Processing graph data using Spark