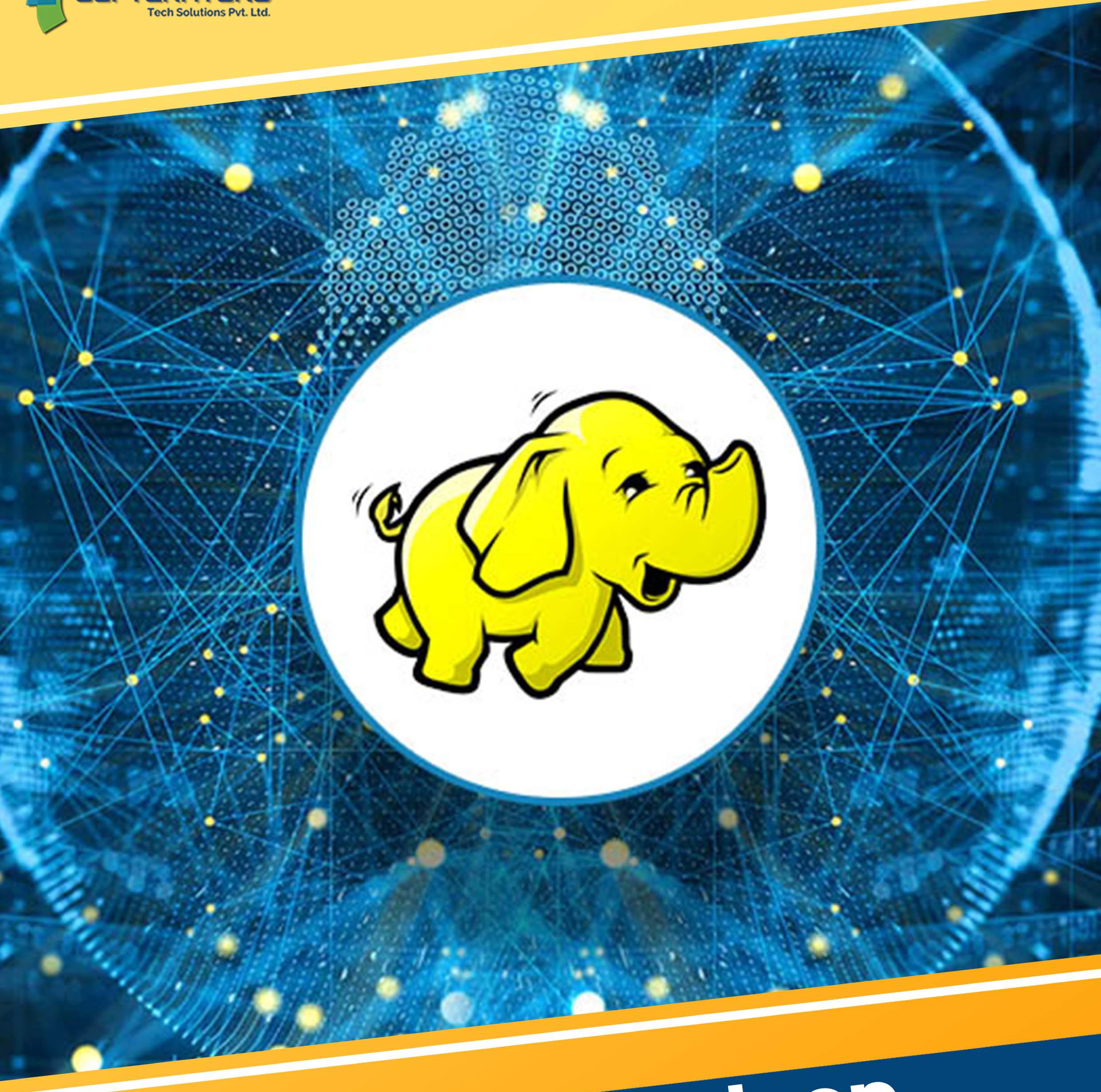
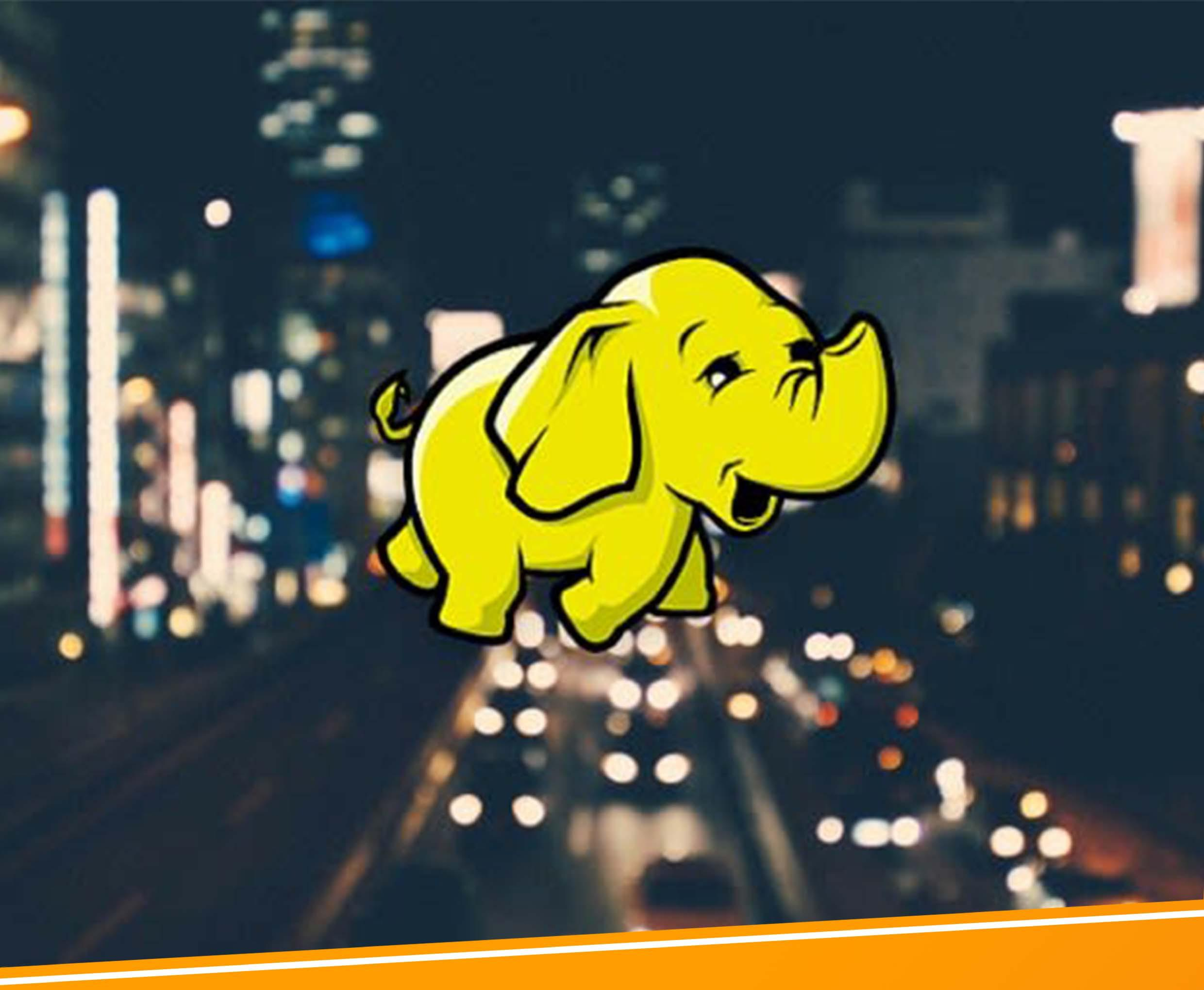
SOFTCRAYONS
Tech Solutions Pvt. Ltd.



Big Data Hadoop Spark

@softcrayons



About Course

Spark is a Hadoop enhancement to MapReduce. The primary difference between Spark and MapReduce is that Spark processes and retains data in memory for subsequent steps, whereas MapReduce processes data on disk. As a result, for smaller workloads, Spark's data processing speeds are up to 100x faster than MapReduce.

BIG DATA HADOOP SPARK

CURRICULUM

1 Exploring Scala

- Introducing Scala
- Installation and configuration of Scala
- Developing, debugging, and running basic Scala programs
- Various Scala operations
- Functions and procedures in Scala
- Scala APIs for common operations
- Loops and collections- Array, Map, List, Tuple
- Pattern-matching and Regex
- Eclipse with Scala plugin

2 Object-Oriented And Functional Programming

- Introduction to OOP - object oriented programming
- Different oops concepts
- Constructors, getters, setters, singletons; overloading and overriding
- Nested Classes and visibility Rules
- Functional Structures
- Functional programming constructs
- Call by Name, Call by Value

3 Big Data And The Need For Spark

- Problems with older Big Data solutions
- Batch vs Real-time vs in-Memory processing
- Limitations of MapReduce
- Apache Storm introduction and its limitations
- Need for Apache Spark

4 A Deep Dive Into Apache Spark

- Introduction to Apache Spark
- Architecture and design principles of Apache Spark
- Spark features and characteristics
- Apache Spark Ecosystem components and their insights

5 Deploying Spark In Local Mode

- Spark environment setup
- Installing and configuring prerequisites
- Installation of Spark in local mode
- Troubleshooting encountered problems

6 Apache Spark Deployment In Different Modes

- Spark installation and configuration in standalone mode
- Installation and configuration of Spark in YARN mode
- Installation and configuration of Spark on a real cluster
- Best practices for Spark deployment

7 Demystifying Apache Spark

- Working on the Spark shell
- Executing Scala and Java statements in the shell
- Understanding SparkContext and the driver
- Reading data from local file-system and HDFS
- Caching data in memory for further use
- Distributed persistence
- Spark streaming
- Testing and troubleshooting

8 Learning RDDs In Spark

- Introduction to Spark RDDs
- How RDDs make Spark a feature rich framework
- Transformations in Spark RDDs
- Spark RDDs action and persistence
- Lazy operations and fault tolerance in Spark
- Loading data and how to create RDD in Spark
- Persisting RDD in memory or disk
- Pairing operations and key-value in Spark
- Hadoop integration with Spark
- Apache Spark practicals and workshops

9 Spark Streaming

- The need for stream analytics
- Comparison with Storm and S4

- Real-time data processing using streaming
- Fault tolerance and checkpointing in Spark
- Stateful Stream Processing
- DStream and window operations in Spark
- Spark Stream execution flow
- Connection to various source systems
- Performance optimizations in Spark

10 Spark MLlib And Spark GraphX

- The need for Spark machine learning
- Introduction to Machine learning in Spark
- Various Spark libraries
- Algorithms for clustering, statistical analytics, classification etc.
- Introduction to Spark GraphX
- The need for different graph processing engine
- Graph handling using Apache Spark

11 Spark SQL

- Introduction to Spark SQL
- Apache Spark SQL Features and Data flow
- Architecture and components of Spark SQL
- Architecture and components of Spark SQL
- Hive and Spark together
- Data frames and loading data

12 Real Life Hadoop & Spark Project

Live Apache Spark & Hadoop project using Spark & Hadoop components to solve real-world Big Data problems in Hadoop & Spark.