Big Data / Hadoop – Course Outline

1 Duration

- 32 Hours

2 Objectives

At end of this workshop, participants will able to :

- Get an overall understanding of various technologies involved in Big Data space / Hadoop Ecosystem
- Understand internals of Hadoop HDFS, YARN and Map Reduce
- Get knowledge on Pig, Hive, HBase and Sqoop
- Get a feel of some of the technologies in action with hands-on and real time use cases
- Troubleshoot, debug, fine tune Hadoop components and learn usage patterns and best practices

3 Audience

Java Developers, Enterprise Warehouse Professionals and QA Professionals who wanted to get themselves familiarized with Big Data and technologies around it.

4 Pre-requisite

- Programming knowledge on Java
- Good knowledge on Unix commands
- Good knowledge on SQL
- Familiarity with data warehousing concepts / ETL tools

5 Hardware & Network Requirements

- Desktop with minimum 4GB RAM (8 GB RAM is recommended)
- Internet connection (to access AWS cloud for lab)

6 Software Requirements

- Windows / Linux / Mac OS
- VirtualBox / VM Player to run Hadoop Image being shared

7 Outline

7.1 Day 1

1) Introduction to Big Data & Hadoop

a) What is Big Data?
b) Challenges of Big Data World
c) Big Data in Industry
d) Limitations of traditional BI architecture
e) What is Hadoop?
f) Hadoop Key Characteristics
g) Hadoop Ecosystem
h) Hadoop Core Components

2) Hadoop Setup, Configuration and Data Loading

a) Hadoop setup
b) Common Hadoop Shell Commands
c) Hadoop Configuration Files
d) HDFS Monitoring over Web
e) Hadoop 1.x vs Hadoop 2.x vs Hadoop 3.x
f) Data Loading Techniques in Hadoop

3) HDFS Internals

a) HDFS Architecture
b) Components of HDFS - Name Node, Secondary Name Node, Data Node
c) HDFS File Write Anatomy
d) HDFS File Read Anatomy
e) Hadoop File Formats and Compression Techniques

7.2 Day 2

1) Hadoop MapReduce

a) MapReduce Framework, Anatomy and Flow
b) MapReduce concepts - Splits, Mappers, Reducers, Partitioners, Combiners and Counters
c) Input / Output File Formats
d) Map side join / Reduce side join
e) Distributed cache
f) MapReduce programs demo

2) Hadoop YARN

a) YARN Architecture
b) Resource Manager
c) Job Scheduler
d) Best Practices

7.3 Day 3

1) Pig

a) Introduction to Pig
b) MR vs Pig
c) Pig Setup and Configuration
d) Pig Data Types
e) Pig Execution Environments
f) Writing Pig scripts
g) Troubleshooting and Debugging Pig

2) Hive

a) Introduction to Hive
b) Pig vs Hive
c) Overview of Hive2
d) Hive Setup, Configuration and Commands
e) Hive Components, Architecture, Metastore
f) Hive Data Types
g) Hive Data Models
h) Hive Managed Tables, External Tables, Partitioned Tables, Clustered Tables concepts
i) Hive UDFs and UDAFs
j) Troubleshooting and Debugging Hive

7.4 Day 4

1) HBase

a) Introduction to HBase
b) HBase Setup and Configuration
c) HBase Components and Architecture
d) Zookeeper Overview
e) Using the HBase Shell
f) HBase General Commands
g) HBase Schema Design
h) HBase Data Model
i) Create and Manage HBase tables
j) Load data into HBase tables
k) Query data from HBase tables
l) Access HBase tables from Hive
m) Monitoring and Troubleshooting HBase

2) Sqoop

a) Introduction to Sqoop
b) Overview of Sqoop2
c) Sqoop Setup and Configuration
d) Sqoop examples to import / export data

3) **Hadoop Distributions and Latest Trends in Big Data Analytics**

   a) Overview of various Hadoop distributions
   b) Overview of Cluster Administration, Troubleshooting and Monitoring
   c) Overview of Latest Trends in Big Data Analytics space