



Hadoop Developer COURSE

At Completion of this Course, you will learn:

- ✓ Introduction to **Big Data and Hadoop**
- ✓ Hadoop ecosystem concepts
- ✓ Hadoop MapReduce concepts and features
- ✓ Developing MapReduce applications
- ✓ Pig concepts
- ✓ Hive concepts
- ✓ Impala
- ✓ Oozie workflow concepts
- ✓ Sqoop Data Ingestion
- ✓ Flume Agents
- ✓ HBase concepts
- ✓ Real Time tools like Hue, Putty, FileZilla, Cloudera Manager
- ✓ **Real Time Projects**

Training Duration – **35-42 Hours**



Introduction to Big Data & Hadoop:

What is Big Data?
What are the challenges for processing big data?
What is Hadoop?
Why Hadoop?
History of Hadoop
Hadoop ecosystem
HDFS
MapReduce

Training Duration – **02 Hours**

Understanding the Cluster:

- Hadoop 2.x Architecture
- Typical workflow
- HDFS Commands
- Writing files to HDFS
- Reading files from HDFS
- Rack awareness
- Hadoop daemons

Training Duration – **02 Hours**



Approach about MapReduce:

Training Duration – 02 Hours

Before MapReduce
MapReduce overview
Word count problem
Word count flow and solution
MapReduce flow

Developing the MapReduce Application:

Training Duration – 03 Hours

- Data Types
- File Formats
- Explain the Driver, Mapper and Reducer code
- Configuring development environment - Eclipse
- Writing unit test
- Running locally
- Running on cluster
- Hands on exercises

How MapReduce Works:

Training Duration – 02 Hours

- Anatomy of MapReduce job run
- Job submission
- Job initialization
- Task assignment
- Job completion
- Job scheduling
- Job failures
- Shuffle and sort
- Hands on exercises

MapReduce Types & Formats:

Training Duration – 02 Hours

- File Formats – Sequence Files
- Compression Techniques
- Input Formats - Input splits & records, text input, binary input
- Output Formats - text output, binary output, lazy output



MapReduce Features:

Training Duration – 02 Hours

- Counters
- Side data distribution
- MapReduce combiner
- MapReduce partitioner
- MapReduce distributed cache
- Hands exercises

Hive Architecture:

Training Duration – 08 Hours

- Hive Architecture
- Types of Metastore
- Hive Data Types
- HiveQL
- File Formats – Parquet, ORC, Sequence and Avro Files Comparison
- Partitioning & Bucketing
- Hive JDBC Client
- Hive UDFs
- Hive on Tez
- Hands-on exercises
- Integration with Tableau

Pig Architecture:

Training Duration – 04 Hours

- Pig Architecture
- Pig Data Types
- Load/Store Functions
- PigLatin
- Pig Udfs

Hbase Concepts & Data Model:

Training Duration – 04 Hours

- HBase architecture and concepts
- Hbase Data Model
- Hbase Shell Interface
- Hbase Java API



Sqoop **Architecture:**

- Sqoop Architecture
- Sqoop Import Command Arguments, Incremental Import
- Sqoop Export
- Sqoop Jobs
- Hands-on exercises

Training Duration – 02 Hours

Flume **Architecture & Setup:**

- Flume Architecture
- Flume Agent Setup
- Types of sources, channels, sinks Multi Agent Flow
- Hands-on exercises

Training Duration – 02Hours

Oozie **Workflows & Concepts:**

- Oozie Fundamentals
- Oozie workflow creations
- Oozie Job submission, monitoring, debugging
- Concepts on Coordinators and Bundles
- Hands-on exercises

Training Duration – 02 Hours

Two Projects & **Case Studies**

- ✓ **Case studies Discussion**
- ✓ Log File Analysis covering Flume, HDFS, MR/Pig, Hive, Tableau
- ✓ Hadoop Use Cases in Retail Domain

Training Duration – 04 Hours