INFOTEKH

BIG DATA AND HADOOP

COURSE CURRICULUM

ENROLL NOW

☎ CALL US
+91 630 171 9270

Module 1: Big Data Overview

- What is Big Data?
- Characteristics of Big Data (Volume, Velocity, Variety, etc.)
- Real-World Use Cases of Big Data

Module 2: Hadoop Distributions & Ecosystem

- Hadoop Distributions: Cloudera, Hortonworks, MapR, Amazon EMR
- Introduction to Apache Hadoop
- Flavors of Hadoop (BigInsights, Google Query, etc.)
- Overview of the Hadoop Ecosystem Components

Module 3: Hadoop Architecture

- Understanding Hadoop Clusters
- Hadoop Core Components:
  NameNode
  DataNode
  ResourceManager / JobTracker
  NodeManager / TaskTracker
  SecondaryNameNode

- HDFS Architecture:
  Block Size (Why 64MB?)
  Block Replication Factor (Why 3?)
  Network Topology and Rack Awareness
  Block Assignment and Heartbeats
  Block Management Service
  Block Reports
  Anatomy of File Write & Read

Module 4: Hadoop High Availability

- Hadoop Federation
- Hadoop HA Concepts

Module 5: MapReduce Programming

- Why MapReduce?
- Use Cases of MapReduce
- MapReduce Components
- Execution Phases: Shuffle, Sort & Merge
- Input and Output File Formats
- Advanced Concepts:
  Joins
  Multi Outputs
  Counters
  Distributed Cache
- Failure Scenarios & Speculative Execution
- Configuration Files:
  core-default.xml
  hdfs-default.xml
  mapred-default.xml
  yarn-site.xml
  hadoop-env.sh
  slaves & masters files

Module 6: YARN (Yet Another Resource Negotiator)

- Introduction to Hadoop 2.x
- YARN Architecture
- Comparison: Hadoop Classic vs YARN

Module 7: Sqoop

- Sqoop Architecture
- Importing & Exporting Data
- Integration with Hive and HBase
- Sqoop Practical Exercises

Module 8: Hive

- What is Hive?
- Hive vs Pig vs MapReduce
- Hive Architecture and Execution
- Hive Table Types: Managed, External, Native, Non-native
- Partitions: Static & Dynamic
- Hive Data Model and Data Types
- Hive Queries:
  Create, Load, Insert
  Joins: Inner, Outer, Skew
  Multi-table Inserts
  SerDe and UDFs
- Hive Best Practices & Optimization Techniques
- Hive Practical Labs

Module 9: Pig

- Introduction and Need for Pig
- Pig vs MapReduce
- Pig Operators:
  Load, Store, Dump, Filter
  Distinct, Group, Join, Limit, Union, Split, Cross
  Diagnostic Operators: Describe, Explain, Illustrate
- Pig Data Types: Primitive & Complex (Bag, Tuple, Map)
- UDFs, Macros, Storage Handlers
- Pig Debugging Tools
- Pig Stats & Practical Exercises

Module 10: HBase & NoSQL Databases

- Introduction to NoSQL
- HBase vs RDBMS
- HBase Architecture:
  HMaster, RegionServer, Zookeeper, Region
  HBase Client and Shell
- Create Tables and Perform Writes
- Row Key Design Principles
- HBase Practical Exercises

Module 11: Apache Spark

- History of Big Data & Spark
- Introduction to Spark Shell & Environment
- Spark RDDs, DataFrames, and SQL
- Lazy Evaluation and Actions
- Reading Data from Parquet, HDFS, Local FS
- Spark Architecture and Internals
- Accumulators & Broadcast Variables
- Debugging and Performance Tuning
- Caching, Persistence, and Memory Management
- Advanced RDD Programming (Shuffle, Partitioning, etc.)

Module 12: Transformers - 1

- DataWeave basics & syntax
- Preview & sample data usage
- Externalize DWL expressions
- Transform XML, JSON, Java
- Use message variables & properties
- LAB: Data Transformations

Module 13: Transformers - 2

- Work with complex data structures
- Use collections, map, $, and DWL operators
- Formatting, conditions, custom data types
- LAB: Advanced DWL usage

Module 14: Handling Errors

- System exception handling
- OnErrorContinue vs OnErrorPropagate
- Flow, app, processor level handling
- Custom error types, validations, reconnection
- LAB: Error Handling

Module 15: MUnit Testing

- Functional testing using MUnit
- Auto-generate test flows
- Asserts, setup, teardown
- LAB: MUnit test cases

Module 16: API-Led Connectivity

- Experience, Process, and System layers

Module 17: Designing APIs

- Use API Designer with RAML
- Mocking, request/response
- Add API to Exchange
- LAB: Design API using RAML

Module 18: Managing APIs

- Deploy to CloudHub
- Create API Proxy
- Add Policies, SLAs, and client ID enforcement
- LAB: Deploy & manage API proxy

Module 19: CloudHub Deployment

- Deploy apps on CloudHub
- On-premise to CloudHub migration changes

Thank you

INFOTEKH

Thank You for Going Through Big Data and Hadoop Curriculum
We hope this guide has provided a clear and structured learning path to strengthen your skills in Big Data and Hadoop.

📌 **NEXT STEPS**

- Start practicing with real-world use cases and hands-on exercises
- Build personal or client-based projects for your portfolio
- Keep exploring updates and best practices in the industry
- Join discussions and stay connected with the community

📞 Need Help or Guidance?

Feel free to contact our course support team:

Course Coordinator

GS Infotekh

✉️ **contact@gsinfotekh.com**

🌐 **www.gsinfotekh.com**

📞 **+91 630 171 9270**