

INFOSOFT IT SOLUTIONS

Training | Projects | Placements

Revathi Apartments, Ameerpet, 1st Floor, Opposite Annapurna Block,

Info soft It solutions, Software Training & Development 905968394,918254087

HADOOP ECOSYSTEM TRAINING

1 Introduction to Big Data and Hadoop

- Understanding Big Data
- Challenges with Traditional Systems
- Introduction to Hadoop
- History of Hadoop
- Hadoop Architecture Overview

2 Hadoop Distributed File System (HDFS)

- HDFS Architecture
- HDFS Operations
- HDFS Commands
- HDFS Federation and High Availability
- HDFS Permissions and Security

3 MapReduce

- MapReduce Basics
- MapReduce Architecture
- Writing MapReduce Jobs
- MapReduce Phases (Map, Shuffle, Reduce)
- MapReduce Job Optimization
- MapReduce Best Practices

4 YARN (Yet Another Resource Negotiator)

- YARN Architecture
- YARN Components (Resource Manager, Node Manager, Application Master)
- YARN Scheduling

- YARN Capacity Scheduler
- YARN Fair Scheduler

5 Hadoop Ecosystem Tools

- Pig Latin Basics
- Pig Execution Modes
- Pig Optimization Techniques
- Apache Hive
 - Hive Query Language (HiveQL)
 - Hive Architecture
 - Hive Optimization Techniques
- Apache HBase
 - HBase Architecture
 - HBase Data Model
 - HBase Operations
- Apache Sqoop
 - Importing and Exporting Data with Sqoop
 - Sqoop Connectors
- Apache Flume
 - Flume Architecture
 - Flume Sources, Channels, and Sinks
- Apache Oozie
 - Oozie Workflow
 - Oozie Coordinator and Bundle

6 Data Ingestion and Integration

- Real-time and Batch Data Ingestion
- Data Integration Best Practices
- Data Serialization Formats (Avro, Parquet, Sequence File, etc.)

7 Data Analysis and Visualization

- Introduction to Data Analysis
- Introduction to Data Visualization
- Tools for Data Analysis (Apache Zeppelin, Jupyter Notebooks, etc.)
- Tools for Data Visualization (Tableau, Power BI, etc.)

8 Hadoop Security

- Kerberos Authentication

- Hadoop Security Best Practices
- Hadoop Encryption
- Hadoop Audit Logging

9 Monitoring and Management

- Hadoop Cluster Monitoring
- Log Management
- Hadoop Backup and Recovery
- Hadoop Cluster Upgrades

10 Advanced HDFS

- HDFS Federation and High Availability in depth
- NameNode Federation
- HDFS Snapshots
- HDFS Quotas
- Erasure Coding in HDFS

Advanced MapReduce

- Combiners and Partitioners
- Custom Input and Output Formats
- Secondary Sorting
- MapReduce Counters and Distributed Cache
- MapReduce Performance Tuning
- MapReduce Output Compression

11 YARN Optimization

- Node Labels and Node Attributes
- YARN Resource Allocation Tuning
- YARN Resource Localization
- YARN NodeManager Disk Health Checks
- YARN Timeline Service and History Server

Advanced Apache Hive

- Hive Transactions and ACID Support
- Hive Vectorization and Cost-Based Optimization
- Hive Indexes and Materialized Views
- Hive Windowing Functions
- Hive Authorization and Row-Level Security

12 Advanced Apache Spark

- Spark SQL Optimization
- Spark Structured Streaming
- Spark MLlib for Machine Learning
- Spark GraphX for Graph Processing
- Spark Performance Tuning

13 Apache Kafka Streams

- Kafka Streams DSL
- Stateful Stream Processing
- Exactly Once Semantics
- Kafka Connect Sink and Source Connectors
- Kafka Security with SSL and SASL

14 Advanced Data Analysis

- Advanced Analytics Techniques (Predictive Analytics, Prescriptive Analytics)
- Time Series Analysis with Hadoop
- Graph Analytics with Hadoop
- Text Analytics with Hadoop
- Spatial Analytics with Hadoop

15 Security Enhancements

- Apache Ranger for Authorization and Audit
- Apache Knox for Gateway Security
- Hadoop Encryption Zones
- Secure Hadoop Cluster Deployment Best Practices

16 Cloud Integration

- Hadoop on Cloud Platforms (AWS EMR, Azure HDInsight, Google Cloud Dataproc)
- Hybrid Cloud Deployment Strategies
- Hadoop Cluster Autoscaling

17 Real-world Use Cases and Best Practices

- Hadoop for IoT Analytics
- Hadoop for Financial Analytics
- Hadoop for Healthcare Analytics
- Hadoop for E-commerce Analytics
- Industry-specific Case Studies and Solutions

18 Performance Benchmarking and Troubleshooting

- Hadoop Cluster Performance Benchmarking
- Hadoop Cluster Troubleshooting Techniques
- Performance Monitoring with Grafana and Prometheus

19 Emerging Technologies and Trends

- Hadoop Cluster Performance Benchmarking
 - Hadoop Cluster Troubleshooting Techniques
 - Performance Monitoring with Grafana and Prometheus
- Serverless Hadoop with AWS Glue and Azure Databricks
- Edge Computing with Hadoop
- Quantum Computing and Hadoop