HADOOP

Hadoop Ecosystem Overview

- **Introduction to Big Data and Hadoop**
  - Understanding Big Data: Characteristics and Challenges
  - Hadoop Ecosystem Overview: Core Components and Tools
  - Hadoop Architecture: HDFS, YARN, and MapReduce
  - Hadoop Distributions: Cloudera, Hortonworks, and Apache Hadoop
  - Hadoop Cluster Setup: On-Premises vs Cloud Deployments

Hadoop Distributed File System (HDFS)

- **HDFS Architecture**
  - Data Storage and Replication in HDFS
  - HDFS Blocks and File System Design
  - Namenode, Datanode, and Secondary Namenode Roles
  - Fault Tolerance and High Availability in HDFS
  - Data Integrity and Security in HDFS
- **HDFS Commands and Operations**
  - Basic File Operations: Creating, Deleting, Moving Files
  - HDFS Access Control: Permissions and Ownership
  - HDFS Quotas and Snapshots
  - Using WebHDFS and HDFS REST API
  - Data Ingestion into HDFS: Sqoop, Flume, and Kafka

MapReduce Programming Model

- **MapReduce Fundamentals**
  - The MapReduce Paradigm: Mapper and Reducer
  - Data Flow in a MapReduce Job
  - Writing MapReduce Programs in Java and Python
  - Combiner and Partitioner in MapReduce
  - Distributed Cache and Counters in MapReduce
- **Advanced MapReduce Concepts**
  - Optimizing MapReduce Jobs for Performance
  - Handling Large Data Sets with MapReduce
  - Joins in MapReduce: Map-Side Join, Reduce-Side Join
  - Custom Input and Output Formats
  - Debugging and Monitoring MapReduce Jobs

Hadoop Ecosystem Tools

- **Apache Hive**
  - Data Warehousing with Hive: Architecture and Use Cases
  - HiveQL: SQL-like Query Language in Hive
  - Partitioning and Bucketing in Hive
  - Optimizing Hive Queries: Tez, LLAP, and Vectorization
  - Advanced Hive Features: UDFs, UDAFs, and Windowing Functions
- **Apache Pig**
  - Data Processing with Pig: Pig Latin Script
  - Pig Execution Modes: Local, MapReduce, and Tez
  - Data Transformation with Pig: Filters, Joins, and Grouping
  - Writing UDFs in Pig: Custom Functions in Java, Python
  - Integrating Pig with HCatalog for Metadata Management

HADOOP

- **Apache HBase**
  - NoSQL Databases and HBase Architecture
  - Data Modeling in HBase: Row Keys, Column Families
  - CRUD Operations in HBase: Get, Put, Scan, Delete
  - HBase with MapReduce: Integration and Use Cases
  - Optimizing HBase Performance: Tuning and Best Practices
- **Apache Sqoop**
  - Data Import and Export with Sqoop: Relational Databases to HDFS
  - Sqoop Commands: Import, Export, Eval, and Merge
  - Incremental Data Loads with Sqoop
  - Integrating Sqoop with Hive and HBase
  - Securing Sqoop Jobs: Kerberos and Password Encryption
- **Apache Flume**
  - Real-Time Data Ingestion with Flume
  - Flume Architecture: Sources, Channels, and Sinks
  - Configuring Flume Agents: Single and Multi-Hop Flows
  - Custom Flume Interceptors and Serializers
  - Integrating Flume with HDFS, Hive, and HBase
- **Apache Kafka**
  - Distributed Messaging with Kafka: Overview and Architecture
  - Kafka Producers and Consumers: Designing Data Pipelines
  - Kafka Topics, Partitions, and Offsets
  - Kafka Streams API: Real-Time Data Processing
  - Kafka Integration with Hadoop Ecosystem: Flume, Spark, HBase

Data Processing and Analytics with Hadoop

- **Apache Spark**
  - Introduction to Spark and its Ecosystem
  - RDDs, DataFrames, and Datasets in Spark
  - Spark SQL and Data Processing
  - Spark Streaming: Real-Time Data Processing
  - Optimizing and Tuning Spark Jobs
- **Apache Oozie**
  - Workflow Scheduling with Oozie: Overview and Components
  - Defining and Managing Oozie Workflows and Coordinators
  - Integrating Oozie with Hadoop Ecosystem: MapReduce, Hive, Pig
  - Oozie Actions: SSH, Email, Shell, and Java Actions
  - Monitoring and Troubleshooting Oozie Workflows
- **Apache Zookeeper**
  - Coordination Service with Zookeeper: Overview and Architecture
  - Zookeeper Data Model: Znodes, Ephemeral, and Sequential Nodes
  - Zookeeper in Hadoop: High Availability, Leader Election
  - Using Zookeeper for Distributed Coordination
  - Zookeeper API and Client Libraries

Advanced Hadoop Security

- **Security Mechanisms in Hadoop**
  - Hadoop Authentication: Kerberos Setup and Configuration
  - Authorization with Apache Ranger and Sentry
  - Data Encryption in HDFS: In-Transit and At-Rest
  - Auditing and Monitoring Hadoop Clusters
  - Securing Hadoop Ecosystem Tools: Hive, HBase, Kafka

Flat No: 506,5th Floor, Nilgiri Block, Aditya Enclave, Ameerpet ,Hyd -5000038.
Contact: +91 9032734343, Mail: info@vritsol.com www.vritsol.com

- **Hadoop in the Cloud**
  - Deploying Hadoop Clusters on AWS, Azure, GCP
  - Cloud-Based Storage Integration: S3, Azure Blob, GCS
  - Managed Hadoop Services: EMR, HDInsight, Dataproc
  - Cloud-Native Data Processing with Hadoop
  - Cost Optimization for Hadoop in the Cloud

Big Data Analytics and Machine Learning

- **Big Data Analytics with Hadoop**
  - Data Exploration and Visualization with Hadoop
  - Data Preprocessing and Feature Engineering
  - Building Machine Learning Models with Mahout and Spark MLlib
  - Real-Time Analytics with Hadoop and Spark Streaming
  - Use Cases: Predictive Analytics, Sentiment Analysis, Recommender Systems
- **Hadoop Integration with Modern Data Platforms**
  - Integrating Hadoop with Data Lake Architecture
  - Hadoop and Data Warehousing Solutions: Snowflake, Redshift
  - Hadoop with NoSQL Databases: Cassandra, MongoDB
  - Hybrid and Multi-Cloud Data Management with Hadoop
  - Modernizing Legacy Hadoop Workloads

Capstone Projects

- **End-to-End Data Processing Pipeline**
  - Designing and Implementing a Complete Data Pipeline
  - Data Ingestion, Processing, and Analysis with Hadoop Ecosystem
  - Integrating Machine Learning Models in Hadoop
  - Optimizing and Tuning the Pipeline for Performance
  - Deployment and Monitoring of Hadoop Workloads in Production
- **Industry-Specific Projects**
  - Finance: Fraud Detection, Risk Analysis
  - Healthcare: Patient Data Analysis, Genomics
  - Retail: Customer Segmentation, Demand Forecasting
  - Telecom: Network Optimization, Customer Churn Prediction
  - Social Media: Sentiment Analysis, Trend Prediction