Hadoop Developer/Admin Training Course Content

Introduction to Big Data

- What is Big data
- Big Data opportunities
- Big Data Challenges
- Characteristics of Big data

Hadoop Architecture

- Introduction to Hadoop
- Parallel Computer vs. Distributed Computing
- Comparing Hadoop & SQL.
- Hadoop and Datawarehouse - When to use which?
- Industries using Hadoop.
- HDFS Design & Concepts
- Blocks, Name nodes and Data nodes
- HDFS High-Availability and HDFS Federation.
- Hadoop DFS The Command-Line Interface
- Basic File System Operations
- Anatomy of File Read
- Anatomy of File Write
- Block Placement Policy and Modes
- More detailed explanation about Configuration files.
- Metadata, FS image, Edit log, Secondary Name Node and Safe Mode.
- How to add New Data Node dynamically.
- How to decommission a Data Node dynamically (Without stopping cluster).
- FSCK Utility. (Block report).
- How to override default configuration at system level and Programming level.
- ZOOKEEPER Leader Election Algorithm.
- How to install Hadoop on your system
- How to install Hadoop cluster on multiple machines
- Hadoop Daemons introduction: **NameNode, DataNode, JobTracker, TaskTracker**
- Exploring HDFS (Hadoop Distributed File System)
- Exploring the HDFS Apache Web UI
- NameNode architecture (EditLog, FsImage, location of replicas)
- Secondary NameNode architecture
- DataNode architecture

MapReduce Architecture

- Exploring JobTracker/TaskTracker
- How to run a Map-Reduce job
- Exploring Mapper/Reducer/Combiner
- Shuffle: Sort & Partition
- Input/output formats
- Exploring the Apache MapReduce Web UI
- Distributed Cache and Hadoop Streaming (Python, Ruby and R).
- YARN.

Hadoop Developer Tasks

- Writting a Map-Reduce programme
- Reading and writing data using Java

- Hadoop Eclipse integration
- Mapper in details
- Reducer in details
- Using Combiners
- Reducing Intermediate Data with Combiners
- Writing Partitioners for Better Load Balancing
- Sorting in HDFS
- Searching in HDFS
- Hands-On Exercise

Hadoop Administrative Tasks

- Routine Administrative Procedures
- Understanding dfsadmin and mradmin
- Block Scanner, Balancer
- Health Check & Safe mode
- Monitoring and Debugging on a production cluster
- NameNode Back up and Recovery
- DataNode commissioning/decommissioning
- ACL (Access control list)
- Upgrading Hadoop

NOSQL

- ACID in RDBMS and BASE in NoSQL.
- CAP Theorem and Types of Consistency.
- Types of NoSQL Databases in detail.
- Columnar Databases in Detail (HBASE and CASSANDRA).

HBase Architecture

- Introduction to HBase
- Installation of HBase on your system
- Exploring HBase Master & Region server
- Exploring Zookeeper
- Column Families and Regions
- Basic HBase shell commands.
- HBase Data Model and Comparison between RDBMS and NOSQL.
- HBase Operations (DDL and DML) through Shell and Programming and HBase Architecture.
- Catalog Tables.
- Block Cache and sharding.
- SPLITS.
- DATA Modeling (Sequential, Salted, Promoted and Random Keys).
- JAVA API’s and Rest Interface.
- Client Side Buffering and Process 1 million records using Client side Buffering.
- HBASE Counters.
- Enabling Replication and HBASE RAW Scans.
- HBASE Filters.
- Hands-On Exercise

Hive Architecture

- Introduction to Hive
- HBase vs Hive
- Installation of Hive on your system
- HQL (Hive query language)
- Basic Hive commands
- Hive Services, Hive Shell, Hive Server and Hive Web Interface (HWI)
- Meta store
- Working with Tables.
- Primitive data types and complex data types.
- Working with Partitions.
- User Defined Functions
- Hive Bucketed Tables and Sampling.
- External partitioned tables, Map the data to the partition in the table, Writing the output of one query to another table, Multiple inserts
- Dynamic Partition
- Differences between ORDER BY, DISTRIBUTE BY and SORT BY.
- Bucketing and Sorted Bucketing with Dynamic partition.
- RC File.
- INDEXES and VIEWS.
- MAPSIDE JOINS.
- Compression on hive tables and Migrating Hive tables.
- Dynamic substation of Hive and Different ways of running Hive
- How to enable Update in HIVE.
- Log Analysis on Hive.
- Access HBASE tables using Hive
- Hands-On Exercise

Pig Architecture

- Introduction to Pig
- Installation of Pig on your system
- Basic Pig commands
- Execution Types
- Grunt Shell
- Pig Latin
- Data Processing
- Schema on read
- Primitive data types and complex data types.
- Tuple schema, BAG Schema and MAP Schema.
- Loading and Storing
- Filtering
- Grouping & Joining
- Debugging commands (Illustrate and Explain).
- Validations in PIG.
- Type casting in PIG.
- Working with Functions
- User Defined Functions
- Types of JOINS in pig and Replicated Join in detail.
- SPLITS and Multiquery execution.
- Error Handling, FLATTEN and ORDER BY.
- Parameter Substitution.
- Nested For Each.
- User Defined Functions, Dynamic Invokers and Macros.
- How to access HBASE using PIG.

- How to Load and Write JSON DATA using PIG.
- Piggy Bank.
- Hands-On Exercise

Sqoop Architecture

- Introduction to Sqoop
- Installation of Sqoop on your system
- Import/Export data from RDBMS to HDFS
- Import/Export data from RDBMS to HBase
- Import/Export data from RDBMS to Hive
- Hands-On Exercise
- Incremental Import(Import only New data, Last Imported data, storing Password in Metastore, Sharing Metastore between Sqoop Clients)
- Free Form Query Import

FLUME

- Installation
- Introduction to Flume
- Flume Agents: Sources, Channels and Sinks
- Log User information using Java program in to HDFS using LOG4J and Avro Source
- Log User information using Java program in to HDFS using Tail Source
- Log User information using Java program in to HBASE using LOG4J and Avro Source
- Log User information using Java program in to HBASE using Tail Source
- Flume Commands

MongoDB

- Introduction
- CRUD
- MongoDB Shell
- Indexing and Schema design
- Replication
- Sharding
- GridFS
- Aggressions

Oozie

- Workflow (Action, Start, Action, End, Kill, Join and Fork), Schedulers, Coordinators and Bundles.
- Workflow to show how to schedule Sqoop Job, Hive, MR and PIG.
- Real world Use case which will find the top websites used by users of certain ages and will be scheduled to run for every one hour.
- Zoo Keeper
- HBASE Integration with HIVE and PIG.
- Phoenix