Big Data & Hadoop

About

Hadoop Development Training course teaches experienced / knowledge peoples on purpose of Hadoop Technology, how to setup Hadoop Cluster, how to store BigData using Hadoop (HDFS) and how to process/analyze the BigData using Map-Reduce Programming or by using other Hadoop ecosystems. ## Prerequisites * Basic Linux Commands * Core Java (OOPS Concepts, Collections , Exceptions) — For Map-Reduce Programming * SQL Query knowledge – For Hive Queries

Hardware and Software Requirements

- Virtual Box 5.1.14
- Host operating system can be Windows 10 /Mac/ Ubuntu 16.10 (Preferable)
- Java 1.8
- Eclipse IDE Neon.2 Release (4.6.2)
- Vagrant 1.9.2

Fundamentals of Hadoop

Introduction to Big Data

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

Introduction to Hadoop

- High Availability
- Scaling
- Advantages and Challenges
- MapReduce
- HDFS (Hadoop Distributed File System)
- YARN
- I/O

Installation

- Standalone Mode
- Pseudodistributed Mode
- Fully Distributed Mode (Cluster)

MapReduce Development

Developing a MapReduce Application

- Configuration API
- Setting Up the Development Environment
- Unit Test with MRUnit
- Running Locally and on Cluster
- Tuning a Job

MapReduce Workflow

- Decomposing a problem
- JobControl
- Oozie

MapReduce Type,Formats and Features

- Types
- Input Formats
- Output Formats

Hadoop Ecosystem

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).
- TTL, Bloom Filters and Compensation.

HBase

- HBase Installation
- HBase concepts
- HBase Data Model and Comparison between RDBMS and NOSQL.
- Master & Region Servers.
- 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.
- Bulk Loading and Coprocessors (Endpoints and Observers with programs).
- Real world use case consisting of HDFS,MR and HBASE.

Pig

- Installation
- 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 Exercises

Hive

- Installation
- Introduction and Architecture.
- Hive Services, Hive Shell, Hive Server and Hive Web Interface (HWI)
- Meta store
- Hive QL
- OLTP vs. OLAP
- 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 Exercises

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
- Use case of Flume: Flume the data from twitter in to HDFS and HBASE. Do some analysis using HIVE and PIG

Sqoop

- Installation
- Import Data.(Full table, Only Subset, Target Directory, protecting Password, file format other than CSV,Compressing,Control Parallelism, All tables Import) Incremental Import(Import only New data, Last Imported data, storing Password in Metastore, Sharing Metastore between Sqoop Clients)
- Free Form Query Import
- Export data to RDBMS,HIVE and HBASE
- Hands on Exercises.

Spark

- Overview
- Linking with Spark

- Initializing Spark
- Using the Shell
- Resilient Distributed Datasets (RDDs)
- Parallelized Collections
- External Datasets
- RDD Operations
- Basics, Passing Functions to Spark
- Working with Key-Value Pairs
- Transformations
- Actions
- RDD Persistence
- Which Storage Level to Choose?
- Removing Data
- Shared Variables
- Broadcast Variables
- Accumulators
- Deploying to a Cluster
- Unit Testing

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
- Proof of concept (POC).

Zookeeper

- Installing
- Running
- Zookeeper as service
- Example
- Building Application with Zookeeper