Website: hopetutors.com
Phone no: 7871012233

BIG DATA
COURSE SYLLABUS

Hadoop YARN Introduction

Hadoop YARN Setup

Programming in YARN framework j

Understanding big data and Hadoop

Big Data

Limitations and Solutions of existing Data Analytics Architecture

Hadoop Features

Hadoop Ecosystem

Hadoop 2.x core components

Hadoop Storage: HDFS

Hadoop Storage : Azure Data Lake Introduction

Hadoop Processing: MapReduce Framework

Hadoop Different Distributions.

Hadoop Mapreduce Framework & YARN

MapReduce Use Cases

Traditional way Vs MapReduce way

Why MapReduce

Hadoop 2.x MapReduce Architecture

Hadoop 2.x MapReduce Components

YARN MR Application Execution Flow

YARN Workflow

Anatomy of MapReduce Program

Demo on MapReduce. Input Splits

Relation between Input Splits and HDFS Blocks

MapReduce: Combiner & Partitioner

Demo on de-identifying Health Care Data set

Demo on Weather Data set.

Hadoop Architecture and HDFS

Hadoop 2.x Cluster Architecture – Federation and High Availability

A Typical Production Hadoop Cluster

Hadoop Cluster Modes

Common Hadoop Shell Commands

Hadoop 2.x Configuration Files

Single node cluster and Multi node cluster set up Hadoop Administration.

Advanced Mapreduce

Counters

Distributed Cache

Reduce Join

Custom Input Format

Sequence Input Format

Xml file Parsing using MapReduce.

Pig

About Pig

MapReduce Vs Pig

Pig Use Cases

Programming Structure in Pig

Pig Running Modes

Pig components

Pig Execution

Pig Latin Program

Data Models in Pig

Pig Data Types

Shell and Utility Commands

Pig Latin : Relational Operators

File Loaders

Group Operator

COGROUP Operator

Joins and COGROUP

Union

Diagnostic Operators

Specialized joins in Pig

Built In Functions (Eval Function

Load and Store Functions

Math function

String Function

Date Function

Pig UDF

Piggybank

Parameter Substitution (PIG macros and Pig Parameter substitution)

Pig Streaming

Testing Pig scripts with Punit

Aviation use case in PIG

Pig Demo on Healthcare Data set.

Hive
Hive Background
Hive Use Case
About Hive
Hive Vs Pig
Hive Architecture and Components
Metastore in Hive
Limitations of Hive
Comparison with Traditional Database
Hive Data Types and Data Models
Partitions and Buckets
Hive Tables(Managed Tables and External Tables)
Importing Data
Querying Data
Managing Outputs
Hive Script
Hive UDF
Retail use case in Hive

Hive Demo on Healthcare Data set.

Advanced Hive and Hbase

Hive QL: Joining Tables

Dynamic Partitioning

Custom Map/Reduce Scripts

Hive Indexes and views Hive query optimizers

Hive : Thrift Server

User Defined Functions

HBase: Introduction to NoSQL Databases and HBase

HBase v/s RDBMS

HBase Components

HBase Architecture

Run Modes & Configuration

HBase Cluster Deployment.

Advanced Hbase

HBase Data Model

HBase Shell

HBase Client API

Data Loading Techniques

ZooKeeper Data Model

Zookeeper Service

Zookeeper

Demos on Bulk Loading

Getting and Inserting Data

Filters in HBase.

Getting started with Sqoop

In this module, you will be introduced to Hadoop you will get to know the Traditional database's application. Also, you will get to know the basics of Sqoop.

Sqoop as an Import/Export tool

Sqoop Import Process

Basic Sqoop Commands

Importing Data in HDFS using Sqoop

Exporting Data from HDFS

:Import /Export Data between RDBMS and Hive/HBase

FLUME

Architecture

Flume events

Inceptors, channel ,sink processor

Twitter Data in HDFS

Telnet as source and HBase as a sink
Twitter Data in HBase

Oozie and Hadoop project
Oozie
Oozie Components
Oozie Workflow
Scheduling with Oozie
Demo on Oozie Workflow
Oozie Co-ordinator
Oozie Commands
Oozie Web Console
Oozie for MapReduce
PIG
Hive and Sqoop
Combine flow of MR
Hive in Oozie
Hadoop Project Demo
Hadoop Integration with Talend.

Understanding Apache Kafka and Kafka Cluster

Need for Kafka

Core Concepts of Kafka

Kafka Architecture

Where is Kafka Used

Processing Distributed data with Apache spark

What is Apache Spark

Spark Ecosystem

Spark Components

History of Spark and Spark Versions/Releases

Spark a Polyglot

What is Scala?

Why Scala?

SparkContext

RDD