Hadoop Course Content

Introduction to Big Data and Hadoop (HDFS & MapReduce)

- Need of BIG DATA
- Sources of BIG DATA
- Characteristics of BIG DATA
- Structure of BIG DATA
- Why Hadoop and Need of Hadoop
- History of Hadoop
- Uses of Hadoop
- Common Hadoop Distributions
- Setting up Hadoop Development

Hadoop 1.0 Architecture

- Hadoop Architecture
- Networking concepts
- Use cases – where Hadoop fits into

Hadoop 2.0 Architecture

- Limitations on Hadoop 1.0 architecture
- Features of Hadoop 2.0 architecture
- HDFS Federation
- High Availability of Name Node
- YARN
- Non MapReduce applications on top of Hadoop

Prerequisites for Hadoop Developer/ Data Analyst

LINUX

- UNIX architecture
- Linux basic to advanced commands
- Linux basic Admin activities
- Unix basic shell scripting
- Advanced shell scripting
- Scheduling jobs in unix

Java

- Introduction to Java. (JDK,JRE and JVM)
- Discussion on Object, Class and Methods
- OOPS concepts with examples
- Exception Handling
- Features and concepts of Core Java for developing MR jobs

Python

- Introduction to concepts of Python
- Complex data types in Python (Tuple, List, Dictionary)
- Inbuilt modules available in Python
- File handling functions using Python

Cluster Installation

- Hadoop Cluster Installation
- Types of Hadoop Cluster
- Installing Pseudo mode Cluster
- Walk through on inbuilt scripts, directories, config file and port numbers
- Discussion on Real time Cluster size

Understanding HDFS In-depth

- HDFS Design
- HDFS Commands
- Fundamental of HDFS (Blocks, NameNode, DataNode, Secondary Name Node)
- Rack Awareness from HDFS
- Read/Write from HDFS Command Line Interface
- Introduction to advanced HDFS commands

Understanding Map Reduce In-depth

- Introduction to Map Reduce architecture
- Detail discussion on different phases of MR
  - Mapper
  - Reducer
  - Splitting
  - Sorting
  - Shuffling
  - Combiner
  - Spilling
  - Partitioning
  - Merging
- Developing Map Reduce Application from Scratch
- Handling of MapReduce Job
  - - Task Failure / TaskTracker Failure / JobTracker Failure
- Discussion on difference between old MR API and new MR API
- Introduction to different file formats and their internal features (Sequential, Binary etc.,)
- Speculative Execution
- Job Scheduling

Map Reduce using Python

- Developing Map Reduce applications using Python
- Discussion on different features available in Streaming

Hadoop Eco System components

Deep Dive in Hive (DWH on top of Hadoop)

- What is Hive ?
- Introduction to HIVE architecture
- Configuring HIVE Metadata Store in different ways
- Basic queries in HIVE (DDL,DML..)
- how Hive Differs from Traditional RDBMS
- Introduction to HiveQL
- Data Types and File Formats in Hive
- Advanced features of HIVE
- JOINS (Mainly Map Side Join)
- UDF

PIG (Data Flow Language)

- What is Pig ?
- Basic commands in PIG
- Introduction to Pig Data Flow Engine
- When should be Pig Used ?
- Advanced features of PIG with real time scenarios
- Different ways of using PigStorage
- Dealing with unstructured data
- Developing regular expressions
- PigLatin Example in Detail

SQOOP (Import – Export Utility)

- Introduction to SQOOP
- Basic SQOOP commands
- Advanced Import Features
- Advanced Export Features
    - Upsert
    - Eval
    - Compressed formats

HBASE (NOSQL Database)

- NOSQL Landscape
- Introduction to HBASE and NOSQL
- Difference between row oriented and column oriented storage
- Basic HBASE commands
- Advanced HBASE features
    - Versions
    - Compression techniques

    - Bloom Filters
    - Sequential Scans
- Bulk Load to HBASE Features

SPARK

- What is Spark?
- Introduction to Spark and In-memory applications
- Get clear understanding of the limitations of MapReduce and role of Spark in overcoming these limitations
- Understanding RDD (Resilient Distributed Dataset)
- Spark Context and Spark SQL Context

IMPALA (InMemory Applicaition)

- What is IMPALA?
- Limitations of IMPALA?
- How Impala improve productivity for typical analysis tasks
- Basic Hive and Impala Query Language Syntax
- Differences Between Hive and Impala Query Syntax

FLUME

- What is Flume?
- When should Flume be used?
- Configuring Flume Components
- Basic Config File building
- Building Flume Config files for different scenarios
- Config file for connecting to different File Servers

KAFKA

- Introduction to Kafka architecture
- Single and Multi-Broker configuration
- Java Sample Producer
- Integration with Hadoop (Flume) and Kafka

Schedulers (OOZIE and Autosys)

Interview question and answer discussion

By Anil