Introduction to Big Data and Hadoop

- Big Data
    - What is Big Data
    - Why all industries are talking about Big Data
    - What are the issues in Big Data
        - Storage
            - What are the challenges for storing big data
        - Processing
            - What are the challenges for processing big data
    - What are the technologies support big data
        - Hadoop
        - Data Bases
            - Traditional
            - NO SQL
- Hadoop
    - What is Hadoop
    - History of Hadoop
    - Why Hadoop
    - Why Hadoop
        - Advantages and Disadvantages of Hadoop
        - Importance of Different Ecosystems of Hadoop
        - Importance of Integration with other BigData solutions
        - Big data real time Use Cases.

HDFS (Hadoop Distributed File System)

- HDFS Architecture
    - Name Node
        - Importance of Name Node
        - What are the roles of Name Node
        - What are the drawbacks in Name Node
    - Secondary Name Node
        - Importance of Secondary Name Node
        - What are the roles of Secondary Name Node
        - What are the drawbacks in Secondary Name Node
    - Data Node
        - Importance of Data Node
        - What are the roles of Data Node
        - What are the drawbacks in Data Node
- Data Storage in HDFS
    - How blocks are storing in DataNodes
    - How replication works in Data Nodes
    - How to write the files in HDFS
    - How to read the files in HDFS
- HDFS Block size
    - Importance of HDFS Block size
    - Why Block size is so large
    - How it is related to MapReduce split size
- HDFS Replication factor
    - Importance of HDFS Replication factor in production environment
    - Can we change the replication for a particular file or folder
    - Can we change the replication for all files or folders
- Accessing HDFS

    - CLI(Command Line Interface) using hdfs commands
    - Java Based Approach
- HDFS Commands
    - Importance of each command
    - How to execute the command
    - Hdfs admin related commands explanation
- Configurations
    - Can we change the existing configurations of hdfs or not
    - Importance of configurations
- How to overcome the Drawbacks in HDFS>
    - Name Node failures
    - Secondary Name Node failures
    - Data Node failures
    - Where does it fit and Where doesn't fit
    - Exploring the Apache HDFS Web UI
- How to configure the Hadoop Cluster
    - How to add the new nodes (Commissioning)
    - How to remove the existing nodes (De-Commissioning)
    - How to verify the Dead Nodes
    - How to start the Dead Nodes
- Hadoop 2.x.x version features
    - Introduction to Namenode fedoration
    - Introduction to Namenode High Availabilty
    - Difference between Hadoop 1.x.x and Hadoop 2.x.x versions

MAPREDUCE

- Map Reduce architecture
    - Jobtracker
        - Importance of JobTracker
        - What are the roles of TaskTracker
        - What are the drawbacks in TaskTracker

        - TaskTracker
            - Importance of TaskTracker
            - What are the roles of TaskTracker
            - What are the drawbacks in TaskTracker
            - Map Reduce Job execution flow
- Data Types in Hadoop
        - What are the Data types in Map Reduce
        - Why these are importance in Map Reduce
        - Can we write custom Data Types in MapReduce
- Input Format's in Map Reduce
        - Text Input Format
        - Key Value Text Input Format
        - Sequence File Input Format
        - Nline Input Format
        - DBInputFormat
        - Importance of Input Format in Map Reduce
        - How to use Input Format in Map Reduce
        - How to write custom Input Format's and its Record Readers
- Output Format's in Map Reduce
        - Text Output Format
        - Sequence File Output Format
        - DBOutputFormat
        - Importance of Output Format in Map Reduce
        - How to use Output Format in Map Reduce
        - How to write custom Output Format's and its Record Writers
- Mapper
        - What is mapper in Map Reduce Job

    - Why we need mapper?
    - What are the Advantages and Disadvantages of mapper
    - Writing mapper programs
    - Working with identity mapper
- Reducer
    - What is reducer in Map Reduce Job
    - Why we need reducer
    - What are the Advantages and Disadvantages of reducer
    - Writing reducer programs
    - Working with identity reducer
- Combiner
    - What is combiner in Map Reduce Job
    - Why we need combiner?
    - What are the Advantages and Disadvantages of Combiner
    - Writing Combiner programs
- Partitioner
    - What is Partitioner in Map Reduce Job
    - Why we need Partitioner
    - What are the Advantages and Disadvantages of Partitioner
    - Writing Partitioner programs
    - Working with default Partitioner
- Distributed Cache
    - What is Distributed Cache in Map Reduce Job
    - Importance of Distributed Cache in Map Reduce job
    - What are the Advantages and Disadvantages of Distributed Cache
    - Writing Distributed Cache programs
- Counters

    - What is Counter in Map Reduce Job
    - Why we need Counters in production environment
    - How to Write Counters in Map Reduce programs
- Importance of Writable and Writable Comparable Api's
    - How to write custom Map Reduce Keys using WritableComparable
    - How to write custom Map Reduce Values using Writable
- Joins
    - Map Side Join
        - What is the importance of Map Side Join
        - Where we are using it
    - Reduce Side Join
        - What is the importance of Reduce Side Join
        - Where we are using it
        - What is the difference between Map Side join and Reduce Side Join
- Compression techniques
    - Importance of Compression techniques in production environment
    - Compression Types
        - NONE, RECORD and BLOCK
    - Compression Codecs
        - Default, Gzip, Bzip, Snappy and LZO
    - Enabling and Disabling these techniques for all the Jobs
    - Enabling and Disabling these techniques for a particular Job
- Map Reduce Schedulers
    - FIFO Scheduler
    - Capacity Scheduler
    - Fair Scheduler
    - Importance of Schedulers in production environment

    - How to use Schedulers in production environment
- Map Reduce Programming Model
    - How to write the Map Reduce jobs in Java
    - Running the Map Reduce jobs in local mode
    - Running the Map Reduce jobs in pseudo mode
    - Running the Map Reduce jobs in cluster mode
- Debugging Map Reduce Jobs
    - How to debug Map Reduce Jobs in Local Mode.
    - How to debug Map Reduce Jobs in Remote Mode.
- YARN (Next Generation Map Reduce)
    - What is YARN
    - What is the importance of YARN
    - Where we can use the concept of YARN in Real Time
    - What is difference between YARN and Map Reduce
- Data Locality
    - What is Data Locality
    - Will Hadoop follows Data Locality
- Speculative Execution
    - What is Speculative Execution
    - Will Hadoop follows Speculative Execution
- Map Reduce Commands
    - Importance of each command
    - How to execute the command
    - Mapreduce admin related commands explanation
- Configurations
    - Can we change the existing configurations of mapreduce or not
    - Importance of configurations

- Writing Unit Tests for Map Reduce Jobs
- Configuring hadoop development environment using Eclipse
- Use of Secondary Sorting and how to solve using MapReduce
- How to Identify Performance Bottlenecks in MR jobs and tuning MR jobs.
- Map Reduce Streaming and Pipes with examples
- Use of Total Order Partitioner in mapreduce

Apache PIG

Introduction to Apache Pig

Map Reduce Vs Apache Pig

SQL Vs Apache Pig

Different data types in Pig

- Modes Of Execution in Pig
  - Local Mode
  - Map Reduce Mode
- Execution Mechanism
  - Grunt Shell
  - Script
  - Embedded
- UDF's
  - How to write the UDF's in Pig
  - How to use the UDF's in Pig
  - Importance of UDF's in Pig
- Filter's
  - How to write the Filter's in Pig
  - How to use the Filter's in Pig
  - Importance of Filter's in Pig
- Load Functions
  - How to write the Load Functions in Pig

    - How to use the Load Functions in Pig
    - Importance of Load Functions in Pig
- Store Functions
    - How to use the Store Functions in Pig
    - Importance of Store Functions in Pig
    - Transformations in Pig
    - How to write the complex pig scripts
    - How to integrate the Pig and Hbase

Apache Hive

- Hive Introduction
- Hive commands
- Hive architecture
    - Driver
    - Compiler
    - Semantic Analyzer
    - Hive Integration with Hadoop
    - Hive Query Language(Hive QL)
    - SQL VS Hive QL
    - Hive Installation and Configuration
    - Hive, Map-Reduce and Local-Mode
    - Hive DLL and DML Operations
- Hive Services
    - CLI
    - Hiveserver
    - Hwi
- Metastore

    - embedded metastore configuration
    - external metastore configuration

- UDF's
    - How to write the UDF's in Hive
    - How to use the UDF's in Hive
    - Importance of UDF's in Hive
- UDAF's
    - How to use the udaf's in hive
    - How to write the UDAF's in Hive
    - Importance of udaf's in hive
- UDTF's
    - How to use the UDTF's in Hive
    - Importance of UDTF's in Hive
    - Transformations in Pig
    - How to write a complex Hive queries
    - What is Hive Data Model
- Partitions
    - Importance of Hive Partitions in production environment
    - Limitations of Hive Partitions
    - How to write Partitions
- Buckets
    - Importance of Hive Buckets in production environment
    - How to write Buckets
- SerDe
    - Importance of Hive SerDe's in production environment
    - How to write SerDe programs

- How to integrate the Hive and Hbase

Apache Zookeeper

Introduction to zookeeper

Pseudo mode installations

Zookeeper cluster installations

Basic commands execution

Apache HBase

- Hbase introduction
- Hbase use cases
- Hbase basics
  - Column families
  - Scans
- Hbase installation
  - Local mode
  - Psuedo mode
  - Cluster mode
- Hbase Architecture
  - Storage
  - WriteAhead Log
  - Log Structured MergeTrees
  - Block cache
- Mapreduce integration
  - Mapreduce over Hbase
- Hbase Usage

    - Key design
    - Bloom Filters
    - Versioning
    - Coprocessors
    - Filters
- Hbase Clients
    - REST
    - Thrift
    - Hive
    - Web Based UI
- Hbase Admin
    - Schema definition
    - Basic CRUD operations

Apache Sqoop

Introduction to Sqoop

MySQL client and Server Installation

Sqoop Installation

How to connect to Relational Database using Sqoop

Sqoop Commands and Examples on Import and Export commands

Sqoop incremental imports

Sqoop with hive integration

Sqoop with hbase integration

Apache Flume

Introduction to flume

Flume installation

Flume agent usage and Flume examples execution

Apache OOZIE

Introduction to oozie

Oozie installation

Executing oozie workflow jobs

Monitering Oozie workflow jobs

Hadoop Ecosystems Integrations:

- Hadoop and Hive Integration
- Hadoop and Pig Integration
- Hadoop and HBase Integration
- Hadoop and Sqoop Integration
- Hadoop and Oozie Integration
- Hadoop and Flume Integration
- Hive and Pig Integration
- Hive and HBase integration
- Pig and HBase integration
- Sqoop and RDBMS Integration

Hadoop Distributions :

- Working with hortonworks distribution
- Working with cloudera distribution

Real Time Project