Course Content for Hadoop and Spark Batch

Introduction to BIGDATA and HADOOP

- What is Big Data?
- What is Hadoop?
- Relation between Big Data and Hadoop.
- What is the need of going ahead with Hadoop?
- Scenarios to apt Hadoop Technology in REAL TIME Projects
- Challenges with Big Data
    - Storage
    - Processing
- How Hadoop is addressing Big Data Changes
- Comparison with Other Technologies
    - RDBMS
    - Data Warehouse
    - TeraData
- Different Components of Hadoop Echo System
    - Storage Components
    - Processing Components
- Importance of Hadoop Echo System Components
- Other solutions of Big Data
    - Introduction to NO SQL

HDFS (Hadoop Distributed File System)

- What is a Cluster Environment?
- Cluster Vs Hadoop Cluster.
- Significance of HDFS in Hadoop
- Features of HDFS
- Storage aspects of HDFS
    - Block
    - How to Configure block size
    - Default Vs Configurable Block size
    - Why HDFS Block size so large?
    - Design Principles of Block Size

HDFS Architecture - 5 Daemons of Hadoop

- NameNode and its functionality
- DataNode and its functionality
- JobTracker and its functionality

- TaskTrack and its functionality
- Secondary Name Node and its functionality.

Replication in Hadoop – Fail Over Mechanism

- Data Storage in Data Nodes
- Fail Over Mechanism in Hadoop – Replication
- Replication Configuration
- Custom Replication
- Design Constraints with Replication Factor
- Can we change the replication factor in Hadoop?
- Can we change the block size for a file or directory in Hadoop?

Accessing HDFS

- CLI (Command Line Interface) and HDFS Commands
- Java Based Approach
- Hadoop Archives
- Configuration files in Hadoop Installation and the Purpose
- How to & Where to Configure Hadoop Daemons in a Hadoop Cluster?
- Difference between Hadoop 1.X.X and Hadoop 2.X.X version
    - Name Node HA (High Availability in Hadoop 2.X.X)

MapReduce

- Why Map Reduce is essential in Hadoop?
- Processing Daemons of Hadoop
- Job Tracker
    - Roles Of Job Tracker
    - Drawbacks w.r.to Job Tracker failure in Hadoop Cluster
    - How to configure Job Tracker in Hadoop Cluster
- Task Tracker
    - Roles of Task Tracker
    - Drawbacks w.r.to Task Tracker Failure in Hadoop Cluster

Input Split

- InputSplit
- Need Of Input Split in Map Reduce
- InputSplit Size
- InputSplit Size Vs Block Size
- InputSplit Vs Mappers

Map Reduce Life Cycle

- Communication Mechanism of Job Tracker & Task Tracker
- Input Format Class
- Record Reader Class
- Success Case Scenarios
- Failure Case Scenarios
- Retry Mechanism in Map Reduce

MapReduce Programming Model

- Different phases of Map Reduce Algorithm
- Different Data types in Map Reduce
    - Primitive Data types Vs Map Reduce Data types

How to write a basic Map Reduce Program

- Driver Code
- Mapper Code
- Reducer Code

Driver Code

- Importance of Driver Code in a Map Reduce program
- How to Identify the Driver Code in Map Reduce program
- Different sections of Driver code

Mapper Code

- Importance of Mapper Phase in Map Reduce
- How to Write a Mapper Class?
- Methods in Mapper Class

Reducer Code

- Importance of Reduce phase in Map Reduce
- How to Write Reducer Class?
- Methods in Reducer Class

IDENTITY MAPPER & IDENTITY REDUCER

Input Format's in Map Reduce

- TextInputFormat
- KeyValueTextInputFormat
- NLineInputFormat
- DBInputFormat

- SequenceFileInputFormat.
- How to use the specific input format in Map Reduce
- How to write Custom Input Format Class and Custom Record Reader

Output Format's in Map Reduce

- TextOutputFormat
- KeyValueTextOutputFormat
- NLineOutputFormat
- DBOutputFormat
- SequenceFileOutputFormat.
- How to use the specific Output format in Map Reduce
- How to write Custom Output Format Class and Custom Record Writer

Map Reduce API(Application Programming Interface)

    - New API
    - Deprecated API
- Combiner in Map Reduce
    - Is combiner mandate in Map Reduce
    - How to use the combiner class in Map Reduce
    - Performance tradeoffs w.r.to Combiner
    - Real Time Use Cases
    - Where to Use & Where Not to Use Combiner
- Partitioner in Map Reduce
    - Importance of Practitioner class in Map Reduce
    - How to use the Partitioner class in Map Reduce
    - Different types of Practitioners in Map Reducer
    - Importance of hashPartitioner
    - How to write a custom Practitioner
    - Real Time Use Cases
- Compression Techniques in Map Reduce
    - Importance of Compression in Map Reduce
    - What is CODEC
    - Compression Types
    - GzipCodec
    - BzipCodec
    - LZOCodec
    - SnappuCodec
    - Configurations w.r.to Compression Techinques
    - How to customize the Compression per one job Vs all the job.
- Map Reduce Job Chaining

    - What is Map Reduce Job Chaining?
    - Use of MR Chaining in Real Time Hadoop Projects
    - Real Time Use case
    - Performance trade off's using MR Chaining
- Joins - in Map Reduce
    - Map Side Join
    - Reduce Side Join
    - Performance Trade Off
    - Distributed cache
- How to debug MapReduce Jobs in Local and Pseudo cluster Mode.
    - Introduction to MapReduce Streaming
    - Data locality in Map Reduce
    - Secondary Sorting Using Map Reduce

Apache PIG

- Introduction to Apache Pig
- Map Reduce Vs Apache Pig
- SQL Vs Apache Pig
- Different datat ypes in Pig
- Where to Use Map Reduce and PIG in REAL Time Hadoop Projects
- Modes Of Execution in Pig
    - Local Mode
    - Map Reduce OR Distributed Mode
- Execution Mechanism
    - Grunt Shell
    - Script
    - Embedded
- Transformations in Pig
- How to write a simple pig script
- Parameter substitution in PIG Scripts
- How to develop the Complex Pig Script
- Bags , Tuples and fields in PIG
- UDFs in Pig
    - Need of using UDFs in PIG
    - How to use UDFs
    - REGISTER Key word in PIG
- Techniques to improve the performance and efficiency of Pig Latin Programs

HIVE

- Hive Introduction
- Need of Apache HIVE in Hadoop
- When to choose PIG & HIVE in REAL Time Project
- Hive Architecture
    - Driver
    - Compiler
    - Executor(Semantic Analyzer)
- Meta Store in Hive
    - Importance Of Hive Meta Store
    - Embedded metastore configuration
    - External metastore configuration
    - Communication mechanism with Metastore
- Hive Integration with Hadoop
- Hive Query Language(Hive QL)
- Configuring Hive with MySQL MetaStore
- SQL VS Hive QL
- Data Slicing Mechanisms
    - Partitions In Hive
    - Buckets In Hive
    - Partitioning Vs Bucketing
    - Real Time Use Cases
- Collection Data Types in HIVE
    - Array
    - Struct
    - Map
    - Real Time Use Cases
- User Defined Functions(UDFs) in HIVE
    - UDFs
    - UDAFs
    - UDTFs
    - Need of UDFs in HIVE
- Hive Serializer/Deserializer - SerDe
- Semi Structured Data Processing Using Hive
- (XML/JSON)
- HIVE – HBASE Integration

SQOOP

- Introduction to Sqoop.
- MySQL client and Server Installation
- How to connect to Relational Database using Sqoop

- Different Sqoop Commands
    - Different flavors of Imports
    - Export
    - Hive-Imports

Hbase

- Hbase introduction
- HDFS Vs Hbase
- Hbase Vs RDBMS
- Hbase Vs NO SQL
- Hbase usecases
- Hbase Data modeling Elements
    - Column families
    - Column Qualifier Name
    - Row Key
- Hbase Architecture
- Clients
    - REST
    - Thrift
    - Java Based
    - Avro
- Map Reduce Integration
- Map Reduce over Hbase
- Hbase Admin
    - Schema Definition
    - Basic CRUD Operations
    - Client Side Buffering in Hbase

Flume

- Flume Introduction
- Flume Architecture
- Flume Master , Flume Collector and Flume Agent
- Flume Configurations
- Real Time Use Case using Apache Flume

Oozie

- Oozie Introduction
- Oozie Architectrure
- Oozie Configuration Files

- Oozie Job Submission
    - Workflow.xml
    - Coordinator.xml
    - job.coordinator.properties
    - Transit parameters in workflow.xml

YARN (Yet another Resource Negotiator) – Next Gen. MapReduce

- What is YARN?
- Difference between Map Reduce & YARN
- YARN Architecture
    - Resource Manager
    - Application Master
    - Node Manager
- When should we go ahead with YARN
- YARN Process flow
- YARN Web UI
- Different Configuration Files for YARN
- Examples on YARN

Impala

- What is Impala?
- How can we use Impala for Query Processing?
- When should we go ahead with Impala
- HIVE Vs Impala
- REAL TIME Use Cases with Impala

MongoDB (As part of NoSQL Databases)

- Need of NoSQL Databases
- Relational VS Non-Relational Databases
- Introduction to MongoDB
- Features of MongoDB
- Installation of MongoDB
- Mongo DB Basic operations
- REAL Time Use Cases on Hadoop & MongoDB Use Cases

Apache Cassandra

- Introduction to Cassandra
- Mongo DB Vs Cassandra
- Basic Operation using Cassandra

Apache Kafka (A Distributed Message Queuing System)

- Introduction to Kafka
- Installation of Kafka
- Difference between MQ Vs Kafka
- Basic Operation using Kafka

Mahout (As a part of BIGDATA ANALYTICS)

- Introduction to Machine Learning (ML) Languages
- Types of Machine Learning
- Introduction to Apache MAHOUT
- Categories of Mahout Algorithms

Real Time Use case using Classifier Algorithm of Mahout
– Naives Bayes

SCALA (Object Oriented and Functional Programming)

- Getting started With Scala.
- Scala Background, Scala Vs Java and Basics.
- Interactive Scala – REPL, data types, variables,expressions, simple functions.
- Running the program with Scala Compiler.
- Explore the type lattice and use type inference
- Define Methodsand Pattern Matching.

Scala Environment Set up.

- Scala set up on Windows.
- Scala set up on UNIX.

Functional Programming.

- What is Functional Programming.
- Differences between OOPS and FPP.

Collections (Very Important for Spark)

- Iterating, mapping, filtering and counting
- Regular expressions and matching with them.
- Maps, Sets, group By, Options, flatten, flat Map
- Word count, IO operations,file access, flatMap

Object Oriented Programming.

- Classes and Properties.

- Objects, Packaging and Imports.
- Traits.
- Objects, classes, inheritance, Lists with multiple related types, apply

Integrations

- What is SBT?
- Integration of Scala in Eclipse IDE.
- Integration of SBT with Eclipse.

SPARK CORE.

- Batch versus real-time data processing
- Introduction to Spark, Spark versus Hadoop
- Architecture of Spark.
- Coding Spark jobs in Scala
- Exploring the Spark shell -> Creating Spark Context.
- RDD Programming
- Operations on RDD.
- Transformations
- Actions
- Loading Data and Saving Data.
- Key Value Pair RDD.
- Broadcast variables.

Persistence.

- Configuring and running the Spark cluster.
- Exploring to Multi Node Spark Cluster.
- Cluster management
- Submitting Spark jobs and running in the cluster mode.
- Developing Spark applications in Eclipse
- Tuning and Debugging Spark.

CASSANDRA (N0SQL DATABASE)

- Learning Cassandra
- Getting started with architecture
- Installing Cassandra.
- Communicating with Cassandra.
- Creating a database.
- Create a table
- Inserting Data
- Modelling Data.

- Creating an Application with Web.
- Updating and Deleting Data.

SPARK INTEGRATION WITH NO SQL (CASSANDRA) and AMAZON EC2

- Introduction to Spark and Cassandra Connectors.
- Spark With Cassandra -> Set up.
- Creating Spark Context to connect the Cassandra.
- Creating Spark RDD on the Cassandra Data base.
- Performing Transformation and Actions on the Cassandra RDD.
- Running Spark Application in Eclipse to access the data in the Cassandra.
- Introduction to Amazon Web Services.
- Building 4 Node Spark Multi Node Cluster in Amazon Web Services.
- Deploying in Production with Mesos and YARN.

SPARK STREAMING

- Introduction of Spark Streaming.
- Architecture of Spark Streaming
- Processing Distributed Log Files in Real Time
- Discretized streams RDD.
- Applying Transformations and Actions on Streaming Data
- Integration with Flume and Kafka.
- Integration with Cassandra
- Monitoring streaming jobs.

SPARK SQL

- Introduction to Apache Spark SQL
- The SQL context
- Importing and saving data
- Processing the Text files,JSON and Parquet Files
- DataFrames
- user-defined functions
- Using Hive
- Local Hive Metastore server

SPARK MLIB.

- Introduction to Machine Learning
  Types of Machine Learning.
- Introduction to Apache Spark MLLib Algorithms.
- Machine Learning Data Types and working with MLLib.
- Regression and Classification Algorithms.

- Decision Trees in depth.
- Classification with SVM, Naive Bayes
- Clustering with K-Means
- Building the Spark server

What we are offering as part of this Course?

- 3 REAL TIME Hadoop Projects End-to-End Explanation with architecture.
- Mock Interviews will be conducted on a one-to-one basis after the course duration.
- Hard Copy & Soft Copy Materials for all the Components.
- Detailed Assistance in RESUME Preparation on a one-to-one basis with Real Time Projects based on your technical back ground.
- All the Real time interview questions and answers will be provided.
- Discussing the new happenings in Hadoop
- Discussing the Interview Questions on a daily basis
- Discussing Certification (CCA 175 – Spark and Hadoop Certification) Related topics on a daily basis.
- Proof Of Concept using complex architectures to give a real time idea