# Module 01 - Hadoop Introduction

Introduction to Big Data

Big Data and Its Importance

Simple Architecture of Big Data

Hadoop 1.0 Architecture

Hadoop 2.0 Architecture

Big Data Environments

Map Reduce Explanation with an example

YARN Architecture

Installation of Cloudera

Setting Up Cloudera environment

Sample Config files check in cloudera

Interview Process Discussion for Module 1

# Module 02 - Linux and Shell Scripting

What is Linux ?

Linux Basic commands sessions

Unix shell Scripting basics and handson

Hadoop Basic Commands

Hadoop commands Handson

Interview Process Discussion for Module 2

Assignment -2 ( Involves Liunux and hadoop based tasks)

# Module 03 - Data Ingestion Tool - SQOOP

Sqoop Introduction

Sqoop Internal Process

Sqoop Explanation with Example

Sqoop with Eval

Sqoop with Split by

Handson1 - Eval,SplitBy,Basic Import from MySQL

Sqoop Import Properties

Sqoop Incremental Import

Handson2 - Sqoop Incremental Import

Sqoop Incremental last_Modified

Handson-3 Sqoop Incremental Append

Sqoop Job Creation - Basic

Sqoop Job creating Password file

Direct Mode

Sqoop Import using Shell sripting

Sqoop Handson Session -2

Sqoop validate Command

Sqoop Import into Hive table

Sqoop Import All Tables

Sqoop Import All Tables Exclude command

Sqoop Export Introduction

Sqoop Export Internal Process

Sqoop Export Incremental load

Sqoop Export properties

Sqoop Export Transcationality

Assignment-3

Interview Process Discussion Sqoop

Project -1 : Sqoop Unix Shell Based Triggering Pipeline

# Module 04 – HIVE

Hive Introduction

Why HQL

HQL VS SQL

Hive Architecture

Different Types of Hive metastore

Different ways of Accessing Hive

Hive Beeline Explanation

Different types of Execution Engines in Hive

Hive - Hadoop Integration

Hive - Tables - Managed and External Tables

Hive Internal tables Explanation

How to create the Internal tables

Hive Internal Table creation on top of Directory

Loading Data from a File to Hive table

Hive External tables Explanation

How to external Tables on dirtectory

Difference between Internal and External table

Handson - Internal And External Tables

Partitions Introduction

Static Partition - Load and Insert

Dynamic Partitions Insert

Handson - Static and Dynamic Partitions

Hive Sub partitions Explanation

Handson - Sub partitions

Bucketing in Hive Explanation

Bucketing on INTEGER column

Bucketing on String Column

Bucketing in Date Column

Handson- Bucketing

Bucketing and Partition on Same Table

Hive Query Optimization

Hive built In Functions

Views in Hive

Hive Sub quires

SCD Types Explanation in Hive

Implementation of SCD Type 1 in Hive

How to remove the duplicates in Hive table

Hive Serde Properties Explanation

Hive table creation on parquet

Hive Table creation on Avro

Hive table creation on XML files

Different types of Joins in Hive

Map side Join In Hive

Bucket Map Join in Hive

Sort Merge Bucket Join in Hibe

Handson - Joins in Hive

Hive UDF creation

Handson - Handle Incremental Load in Hive through Views

Hive Ranking functions

Concept of Vectorization

Choosing File format in Hive - Industry based

Hive MSCK command Explanation

Hive Advanced commands

ACID Properties In Hive

Handson - DML operations in Hive

ORC vs Parquet Vs AVRO

Interview Session - 1

Interview Session - 2

Assignment - 3

Assignment - 3 Solution

Assignment -4

Assignment -4 Solution

Project 2 - Sqoop Hive Data Process Pipeline Creation

## Module 05 – Python

Python Introduction

Data types in Python

Collections in python

Python String Interpolation and data interpolation

Control statements (IF , While , For )

Python functions

python variables

Python Map , filter , Reduce

Python file handling , Read , Write and Append

Python classes and Objects

Inheritance and Multilevel Inheritance

How to write Wrapper Code in python

## Module 06 – Spark

Spark Introduction Why Spark?

Spark Ecosystem Components

Spark and mapReduce differences

Architecture of Spark

Different ways of process the data in Spark

Spark Core Introduction

What is SparkContext?

what is RDD and its importance? what is DAG? RDD Lineage

Concept of resilent

Lazy transformations

What is transformation in RDD Examples of Transformations in RDD

What is actions in RDD ?

Examples of RDD Actions

Narrow and Wide Transformation

How to perform word count processing in Spark Core

Spark Submit Introduction

Spark Submit Architecture explanation

Spark Submit - Stages in Spark

Different modes of Spark Submit

Spark Submit in Client mode

Spark Submit In cluster mode

Spark submit in Standalone mode

Spark Dynamic memory Allocation of resources

Difference between Group By Vs ReduceBy

Concept of Accumulators

Concepts of Broadcast varibales

How to Accumulators and broadcast variables acts as a Optimization techniques in Spark

Repartition

Coalesce

Difference between repartition and Coalesce - Real time scenerio

How to increase the parallelism in spark

Hands On Document for Spark Core

Spark Core HandsOn Session -1

Spark Core HandsOn Session -2

Concept of Map partition

Cache Concept In Detail

Units of Caching

Different memory Levels in Spark

Difference between cache vs persist

Concept of Serialization in Spark

Java serialization Kyro Serialization why Kyro Serialization is best for Spark?

Joins in Spark Core Benefits of Repartitions partitionBy vs bucketBy saving file in various file format

Assignment - 5

Assignment - 5 Solution

Interview Preparation for Spark Core

Real time Code preparation for Spark Core in Pycharm using Business Logic

## Module 07 – Spark SQL

Spark SQL Introduction Components of Spark SQL?

Data Source API explanation

Data Frame Explanation

Hive Thrift Service in Spark Explanation Tungsten Memory management in Spark SQL What is SparkSession?

Difference Between SparkSession and SparkContext What is Data set?

Advantages of Data set?

RDD Vs Dataframe Vs Data set

Dataframe creation from CSV file format

Dataframe creation from JSON file format

Dataframe creation from AVRO file format using External Jar

Dataframe creation from XML file format using External Jar

Dataframe creation from Parquet File format

Dataframe creation in spark shell for AVRO , XML using SparkConf property

Creating a Dataframe from a file (without schema)

Case class using toDF()

Create dataframe method with RowRDD and Struct variable

Create Dataframe using Schema - Seamless Dataframe

Write Modes in Dataframe

Dataframe using partitionBy

Joins in Spark SQL

Usage of BroadCast Join

Domain Specific language Operations on Dataframes

withColumn in Dataframe DSL operation - Session 1

DSL operation - Session 2

Aggregation in Spark SQL

Window Aggregations in Spark SQL

Complex Data processing - Struct Data processing (JSON) Complex Data processing - Array

Data processing (JSON) How to create a Spark UDF ?

Spark UDF in Data frames

Assignment - 6

Assignment - 6 Solution

Interview preparation for Spark sql

Project 3 – Spark Processing through Web URL and HDFS storage

# Module 08 - HBase

Introduction to Hbase

Types of NOSQL Databases

Characteristics of NOSQL

CAP THEOREM

Why Column Based Storage is highly preferred than Row Based

RDBMS vs Hbase

Storage Hierarchy in HBASE

Hbase Architecture

TABLE design HBASE

What is column family in Hbase ?

Hands on Session on HBASE commands

How to create the Hbase table

How to insert the data into Hbase Table

How to scan the data

How to enable the table

How to disable the table

Assignment-5

Assignment-5 Solution

Project 4 : Sqoop Hive Hbase  Spark Data processing Pipeline

## Module 9 – Spark Integrations & Use Cases

Spark Hive Integration

Spark Hive Hbase Integration

Spark hbase Integration

Spark Cassandra Integration

Spark SQL PULL - RDBMS Spark SQL integration

## UseCases:

How to handle Null values in Spark SQL

How to choose the number of executors for a given configuration

How to calculate the number of cores

How to mask the data for a given Dataframe

How to handle error records in Dataframe

How to do resource Level optimization

When to go for broadcast join and simple join How to handle memory out of

exceptions in Spark What is Data skew ?

How to resolve Data Skew using Salting technique?

Spark Speculative execution Mode

How to handle the Ambiquous column in Spark Dataframe

How to do the PIVOT in spark SQL

Difference between partition and partitioner

Hard Coding in Spark Projects

What is Pyspark

Difference between spark scala and Pyspark

Pyspark deployments

## Module 10 – Kafka & Spark Streaming

Introduction to KAFKA

Why Kafka?

Kafka explanation with real time scenario

Kafka Message Queue Components explanation

Topic,partition,Replication

What is Producer and Consumer?

Broker and its importance

Controller Broker explanation and its election Use of Zookeeper What is

Offset ?

what is BootStrap Servers?

Installing One Node Kafka cluster locally

Introduction to KAFKA

Data storage in Brokers

Leader Copy in Kafka

Follower copy in Kafka

Consumer Groups

Data Serialization in Kafka

## Module 11 – AWS in Big Data

Module 16 - AWS In Big Data

Why do we go for AWS?

Why AWS is the world’s largest cloud provider?

Storage services in AWS What is S3 Storage?

How to upload the data in S3 Storage?

How to process the data that is present in S3 Storage? EMR - Hadoop service in AWS

how to create EMR cluster

How to process the data in EMR through Hive? how to create hive tables in EMR

on S3 Storage

How to copy the data from S3 to local

How to create EC2 Instance

How to generate Key value pair

AWS basic commands are required for Big Data processing

What is Athena

when we go for Athena