Module -1 :Scala Programming

- Basics of Scala
- Why Scala is called Functional programming
- What is the use of Function programming
- Difference between VAL and VAR
- Data Types of Scala
- Use of UNIT Data Type
- Collection in Scala
- List Collection
- Set collection
- Tuple Collection
- Map Collection
- Range Collection
- Expressions in Scala
- Statements in Scala
- Scala Class Hierarchy
- For Loop
- If loop
- Match Expression
- Wild Card Pattern Matching
- String Interpolation
- Functions in Scala
- Methods and Operations
- Nested Functions
- Variable Args Functions
- Vector collection explanation
- Recursive functions
- Higher Order Functions Introduction
- What is the use of Higher Order functions
- Map Higher Order function
- Filter Higher Order function
- Foreach higher Order function
- Reduceby Higher Order function
- Currying In scala
- Singleton Object on Scala
- How to create a singeton object
- Classes in Scala
- Companion Object and Case Class
- Main method in Scala

- Factory Design Pattern
- Traits in Scala
- How to create traits in Projects
- Options in Scala
- Handling Nulls in Scala
- Assignment - 1
- Assignment - 1 Key

Module -2:SPARK PROCESSING FRAMEWORK

- 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
- Setting Up Eclipse IDE
- How to perform word count processing in Spark Core
- Jar creationJar deployment
- 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 - 7
- Assignment - 7 Solution
- Interview Preparation for Spark Core
- Real time Code preparation for Spark Core in Eclipse using Business Logic

Module 03 - 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 SparkSQL
- Complex Data processing - Struct Data processing (JSON) Complex Data processing - Array Data processing (JSON) How to create a Spark UDF ?
- Spark UDF in Dataframes
- Assignment - 8
- Assignment - 8 Solution
- Interview preparation for Spark sql

Module 04 - Spark Integrations

- Spark Hive Integration
- Spark Hive Hbase Integration
- Spark hbase Integration
- Spark Cassandra Integration
- Spark SQL PULL - RDBMS Spark SQL integration

Module 05 - Spark Use Case

- 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 sparkScala and Pyspark
- Pyspark deployments

Module 06-Hbase

- Introduction to Hbase
- Types of NOSQL Databases
- Characteristrics 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 ?
- Handson 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

- **Module 07 - Spark Streaming And Structured Streaming**

  What is Real Time Processing
- Why Real time processing is ruling the IT industry?
- Batch Vs Reak Time processing
- Streaming Context
- Spark Dstreams
- What is batch interval
- Transformed Dstreams What is sliding interval
- What is Window Size
- Window operations in Spark Streaming
- Stateless Transformations
- Use Case - Spark Streaming using Wordcount program
- What is Structured Streaming
- Advantages of Structured Streaming
- Data Stream Output mode
- What is checkpoint location
- Assignment 09
- Interview Session

- Module 11-KAFKA
- 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 copies in Kafka
- Consumer Groups
- Data Serialization in Kafka
- Assignment - 10
- Assignment - 10 Solution
- Project 4: Kafka Spark Streaming Use case

Module 08 - Cloud Computing

- What is cloud computing?
- Different types of Cloud
- Customer Definition for Cloud Computing
- Business Definition for Cloud Computing
- What is Public Cloud
- What is Private Cloud What is Hybrid Cloud What are cloud services?
- IaaS Service
- PaaS Service
- SaaS Service

Module 09 - AWS In Big Data

- Why we go for AWS?
- Why AWS is a world 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 requires for Big Data processing What is Athena Why we go for Athena?

Module 10 - Azure in Big Data

- What is Azure?
- Why Azure is world number one Cloud provider in terms of Security
- Services Offered by Azure
- How to create the Free Azure account
- Services Offered by Azure
- Data Storage Services in Azure - Azure BLOB storage
- HDInsight cluster - Hadoop service in Azure
- Creation of HDInsight cluster
- Performing Hive analytics in Azure HDInsight Cluster
- Upload the processed data into Azure Data lake Storage

NOTE:

- **Real Time batch Processing Project Explanation with Different layers**
- **Resume Preparation Session**
- **Mock Interview will be conducted for every Module**
- **Interview Questions**
- **How to handle the Design round for 10+ Experienced**
- **SQL interview Preparation Session**

PySpark Curicullam :

Python Programming:

- 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

Note:

- For Pyspark only python programming will change and spark concepts remains same, pyspark deployments will be taken care separately.