DVS Hadoop Development Content

MODULE 1- INTRODUCTION TO BIG DATA

- What is Big Data?
- Examples of Big Data
- Reasons of Big Data Generation
- Why Big Data deserves your attention
- Use cases of Big Data
- Different options of analyzing Big Data

MODULE 2- INTRODUCTION TO HADOOP

- What is Hadoop
- History of Hadoop
- How Hadoop name was given
- Problems with Traditional Large-Scale Systems and Need for Hadoop
- Understanding Hadoop Architecture
- Fundamental of HDFS (Blocks, Name Node, Data Node, Secondary Name Node)
- Rack Awareness
- Read/Write from HDFS
- HDFS Federation and High Availability

MODULE 3- STARTING HADOOP

- Setting up single node Hadoop cluster(Pseudo mode)
- Understanding Hadoop configuration files
- Hadoop Components- HDFS, MapReduce
- Overview Of Hadoop Processes
- Overview Of Hadoop Distributed File System
- The building blocks of Hadoop
- Hands-On Exercise: Using HDFS commands

MODULE 4- MAPREDUCE-1(MR V1)

- Understanding Map Reduce
- Job Tracker and Task Tracker
- Architecture of Map Reduce
- Data Flow of Map Reduce
- Hadoop Writable, Comparable & comparison with Java data types
- Creation of local files and directories with Hadoop API
- Creation of HDFS files and directories with Hadoop API
- Map Function & Reduce Function
- How Map Reduce Works
- Anatomy of Map Reduce Job
- Submission & Initialization of Map Reduce Job

- Monitoring & Progress of Map Reduce Job
- Understand Difference Between Block and Input Split
- Role of Record Reader, Shuffler and Sorter
- File Input Formats
- Getting Started With Eclipse IDE
- Setting up Eclipse Development Environment
- Creating Map Reduce Projects
- Configuring Hadoop API on Eclipse IDE
- Differences between the Hadoop Old and New APIs
- Life cycle of the Job
- Identity Reducer
- Map Reduce program flow with word count
- Combiner & Partitioner, Custom Partitioner with example
- Joining Multiple datasets in MapReduce
- Map Side, Reduce Side joins with examples
- Distributed Cache with practical example
- Stragglers & Speculative execution
- Schedulers(FIFO Scheduler, FAIR Scheduler, CAPACITY Scheduler)

MODULE 5- MAPREDUCE-2(YARN)

- Limitations of Current Architecture
- YARN Architecture
- Application Master, Node Manager & Resource Manager
- Writing a Map Reduce using YARN

MODULE 6- HIVE

- Introduction to Apache Hive
- Architecture of Hive
- Installing Hive
- Hive data types
- Exploring hive metastore tables
- Types of Tables in Hive
- Partitions(Static & Dynamic)
- Buckets & Sampling
- Indexes & Views
- Developing hive scripts
- Parameter Substitution
- Difference between order by & sort by
- Difference between Cluster by & distribute by
- File Input formats(Text file, RC, ORC, Sequence, Parquet)
- Optimization Techniques in HIVE
- Creating UDFs
- Hands-On Exercise
- Assignment on HIVE

MODULE 7- PIG

- Introduction to Apache Pig
- Building Blocks (Bag, Tuple & Field)
- Installing Pig
- Data types
- Different modes of execution of PIG
- Working with various PIG Commands covering all the functions in PIG
- Developing PIG scripts
- Parameter Substitution

  - ✓ Command line arguments
  - ✓ Passing parameters though a param file

- Joins (Left Outer, Right Outer, Full Outer)
- Nested queries
- Specialized joins in PIG (Replicated, Skewed & Merge Join)
- HCatalog(Getting data from hive to pig & vice versa)
- Working with un-structured data
- Working with Semi-structured data like XML, JSON
- Optimization techniques
- Creating UDFs
- Hands-On Exercise
- Assignment on PIG

MODULE 8- SQOOP

- Introduction to SQOOP & Architecture
- Import data from RDBMS to HDFS
- Importing Data from RDBMS to HIVE
- Exporting data from HIVE to RDBMS
- Handling incremental loads using sqoop
- Hands on exercise

MODULE 9- HBASE

- Introduction to HBASE
- Installation of HBASE
- Exploring HBASE Master & Region server
- Exploring Zookeeper
- CRUD Operation of HBase with Examples
- HIVE integration with HBASE(HBASE-Managed hive tables)
- Hands on exercise on HBASE

MODULE 10- OOZIE

- What is Oozie & Why Oozie
- Features of Oozie
- Job Types in Oozie
- Control Nodes & Action Nodes
- Oozie Workflow Process flow
- Oozie Parameterization
- Oozie Command Line examples – Developer
- Oozie Web Console
- Hands on exercise on OOZIE

MODULE 11- REAL TIME PROJECT

We will be providing raw data & requirements for the project & you will have to work. Finally we will have one Project execution session where we will be explaining the steps for project execution.

MODULE 12- FAQS, REAL TIME SCENARIOS AND REAL TIME INTERVIEW QUESTIONS

MODULE 13- OVERVIEW SESSIONS ON NEW MODULES

Workshops on Spark & Scala once in 2 months. Interested people can attend.

PRE-REQUISITES LIKE CORE JAVA, BASICS OF LINUX & BASICS OF SQL ARE ALSO COVERED AS PART OF THE TRAINING FOR FREE OF COST.

COURSE HIGHLIGHTS:

- ✓ Trainers are certified Real Time working professionals
- ✓ Main focus on Hands-on sessions
- ✓ Affordable course fee
- ✓ Course aligned to Cloudera Certification
- ✓ Flexible timings for working people
- ✓ Real-time projects on Hadoop
- ✓ Guidance in Resume Preparation
- ✓ 100% placement assistance
- ✓ Post training support
- ✓ Lifetime validity for re-attending classes