Hadoop (Big Data)

SKILLs GAINED

1. Hadoop
2. HDFS
3. MapRedcue
4. Hive
5. Pig
6. Sqoop
7. Spark and Kafka
8. HBase//Mongo
9. Oozie
10. Flume, Chukwa, Scribe
11. Projects

TRAINING METHODOLOGY

Hadoop Developer Training has a major focus on giving you the complete knowledge to build Big Data Analytics system using Hadoop and Hadoop Ecosystem. You will receive hands-on training on HDFS, MapReduce, Hive, Sqoop, Pig, HBase, Spark, Kafka and Oozie in an effective way. Our vast experienced trainer and tutors will cover all concepts with assignments at every session. You can evaluate the knowledge acquired at the end of the session. The complete agenda extensively covers all the topics required to gain expertise in Hadoop stack. Dedicated tutors will be available all the day on email and specific hours on call to clarify the questions and help you in completing your assignments on time. If you are not satisfied with trainer or tutor, we will reimburse your complete money before the first 7 sessions of the training without any questions asked. So, why are you waiting for? Enroll today and be ready to build Big data applications. If you have any questions please write to us.

OVERVIEW

This is an instructor lead Hadooptraining that delivers key concepts and expertise necessary to create robust data processing applications using Apache Hadoop. Through lecture and interactive hands-on exercises, attendees will learn Hadoop and its ecosystem components in a very effective manner.

AGENDA

Chapter 1: Introduction to Apache Hadoop

Learning Objectives:

In this chapter, we will provide overview of how Hadoop works. What is Cloud and how Hadoop is different from Cloud. We will also cover role of Hadoop in analytics and data science world. We will cover different Hadoop distributions available in market and their relative merits and demerits.

Topics Covered:

- Hadoop in cloud
- What is Big Data
- Introduction to Analytics and the need for big data analytics
- Hadoop Solutions - Big Picture
- Hadoop distributions
    a. Apache Hadoop
    b. Cloudera Hadoop
    c. Horton Works etc.
- Comparing Hadoop Vs. Traditional systems
- Data Retrieval - Radom Access Vs. Sequential Access
- NoSQL Databases
    a. HBase
    b. Cassandra
    c. MongoDB etc.

Chapter 2: HDFS(Hadoop Distributed File System)

Learning Objectives:
We will learn HDFS in more detailed and comprehensive manner in this chapter. You are going to setup your own Hadoop environment where we can practice and experience Hadoop cluster first hand. You are going to interact with Hadoop and its demons in various ways to learn practical understanding of HDFS and Hadoop.

Topics Covered:

- Blocks and Splits
- Input Splits
- HDFS Blocks
- Data Replication
- Hadoop Rack Aware
    - Data high availability
    - Data Integrity
    - Cluster architecture and block placement
    - Accessing HDFS
- JAVA Approach
- CLI Approach
- Master Daemons
    - Name node
    - Job Tracker
    - Secondary name node
- Slave Daemons
    - Data node
    - Task tracker

Chapter 3: Map Reduce

Learning Objectives:

In this chapter, we will understand what is Map Reduce and what is the necessity of Map reduce in big data world. We will learn how map reduce is different from traditional programming and map reduce framework as a whole. We are going to explore, learn and practice at least 15 different map reduce programs covering different business domains. We will use Eclipse as an IDE for doing development and debugging. We will practice on popular Hadoop vendor distributions namely Cloudera, Horton Works and Pivotal.

Topics Covered:

- Writing a MapReduce Program
  - Examining a Sample MapReduce Program
  - With several examples
  - Basic API Concepts
  - The Driver Code
  - The Mapper
  - The Reducer
  - Hadoop's Streaming API
- Performing several Hadoop jobs
  - The configure and close Methods
  - Sequence Files
  - ToolRunner
  - Using The Distributed Cache
- Monitoring and debugging on a Production Cluster
  - Counters
  - Skipping Bad Records
  - Rerunning failed tasks with Isolation Runner
- Tuning for Performance in MapReduce
  - Reducing network traffic with combiner
  - Partitioners
  - Reducing the amount of input data
  - Using Compression
  - Other Performance Aspects
- Programming Practices & Performance Tuning
  - Developing MapReduce Programs in
  - Local Mode
    - Running without HDFS and Mapreduce
  - Pseudo-distributed Mode
    - Running all daemons in a single node
  - Fully distributed mode
    - Running daemons on dedicated nodes
- YARN and MR2

- What is difference between YARN and MR2
- MR2 demons
- RM, NM
- Executing a program in YARN
- Name Node High – Availability
- Name Node federation

Chapter 4: Hive

Learning Objectives:

In this chapter, we will study Hive, a very easy and powerful ecosystem component of Hadoop. Good thing about Hive is that people who know SQL already know Hive. We will talk about what are the advantages of Hive over standard map reduce and we will also cover when to use map reduce over Hive. We will cover Hive concepts in a detailed manner post which we will switch to practical's; We will share a document containing good number of exercises where each and every participant will practice/perform these exercises. Trainer/tutor will be assisting participants for doing those exercises.We will practice on popular Hadoop vendor distributions namely Cloudera, Horton Works and Pivotal.

Topics Covered:

- Hive concepts
- Hive architecture
- Install and configure hive on cluster
- Create database, access it from java client
- Buckets
- Partitions
- Joins in hive
- Inner joins
- Outer Joins
- Hive UDF
- Hive UDAF
- Hive UDTF
- Develop and run sample applications in Java/Python to access hive

Chapter 5: Pig

Learning Objectives:

In this chapter, we will study Pig, another very easy and powerful ecosystem component of Hadoop. Good thing about Pig is that people who know any scripting languagecan easily learn Pig. Doing programming is Pig is more of an algorithmic way rather than language dependent way. We will talk about what are the advantages of Pig over standard map reduce and we will also cover when to use map reduce over Pig. We will cover Pig concepts in a detailed manner post which we will switch to practical's; We will share a document containing good number of exercises for Pig where each and every participant will practice/perform these exercises.

Trainer/tutor will be assisting participants for doing those exercises.We will practice on popular Hadoop vendor distributions namely Cloudera, Horton Works and Pivotal.

Topics Covered:

- Pig basics
- Install and configure PIG on a cluster
- PIG Vs MapReduce and SQL
- Pig Vs Hive
- Write sample Pig Latin scripts
- Modes of running PIG
- Running in Grunt shell
- Programming in Eclipse
- Running as Java program
- PIG UDFs
- Pig Macros

Chapter 6: Sqoop

Learning Objectives:

In this chapter, we will study Sqoop, an easy and ubiquitous ecosystem component of Hadoop. It's a component using which we can import data from RDBMS systems like Oracle, Mysql into Hadoop/HDFS and similarly we can export data from Hadoop to MySQL. We will talk about what are the advantages of Sqoop over other traditional approaches. We will cover Sqoop concepts in a detailed manner post which we will switch to practical's; We will share a document containing good number of exercises for Pig where each and every participant will practice/perform these exercises. Trainer/tutor will be assisting participants for doing those exercises.We will practice on popular Hadoop vendor distributions namely Cloudera, Horton Works and Pivotal.

Topics Covered:

- Install and configure Sqoop on cluster
- Connecting to RDBMS
- Installing Mysql
- Import data from Oracle/Mysql to hive
- Export data to Oracle/Mysql
- Internal mechanism of import/export
- Imports using multi node cluster
-

Chapter 7 :NoSqls like Hbase , Cassandra and MongoDB

Learning Objectives:

In this chapter, we will study NoSqls like HBase, Mongo and Cassandra. Facebook and Twitter currently store their data into HBas. It's a Big Data based database which is originally originated from Google. We will cover HBase concepts in a detailed manner post which we will switch to practical's; We will share a document containing good number of exercises for

MongoDB where each and every participant will practice/perform these exercises. Trainer/tutor will be assisting participants for doing those exercises.
Topics Covered:

- NoSqls concepts
- HBase architecture
- Region server architecture
- File storage architecture
- HBase basics
- Column access
- Scans
- HBase   use cases
- Install and configure HBase on a multi node cluster
- Create database, Develop and run sample applications
- Access data stored in HBase using clients like Java, Python and Pearl
- Map Reduce client to access the HBase data

Chapter 8 :Oozie

Learning Objectives:
In this chapter, we will study Oozie, a workflow and scheduling component of Hadoop family. Facebook and Twitter and Yahoo runs thousands of Oozie workflows every day. We cover Oozie concepts in a detailed manner post which a tutorial document will be shared. The tutorial contains good number of exercises for Oozie where each and every participant will practice/perform these exercises. Trainer/tutor will be assisting participants for doing those exercises.

Topics Covered:

- Oozie architecture
- XML file specifications
- Install and configuring Oozie and Apache
- Specifying Work flow
  - Action nodes
  - Control nodes
- Oozie job coordinator
- Accessing Oozie jobs using command line and using web console
- Develop sample workflows in Oozie and run on Apache's Oozie and on other Hadoop distributions.

Chapter 9: Spark and Kafka

Learning Objectives:

In this chapter, we will study Spark and Kafka, very powerful ecosystem components of Hadoop designed for real time data processing. Spark started penetrating into Big Data world at an astronomical rate these days. It's a component using which we can process real time and perform streaming data analytics. Twitter trending topics is a classic example of streaming application and Spark is a perfect fit for it. We will talk about what are the alternatives of Spark and advantages of Spark over others.We will cover Spark concepts in a detailed manner post which we will switch to practical's; We will also cover ecosystem of Spark. We will cover Kafka, an open-source messaging system. We will talk about how Kafka is different from RabbiMQ, ActiveMQ etc. We will talk about when to use Kafka over traditional messaging systems. We will also talk about how to integrate Kafka with processing systems like Spark streaming, Strom and Samza.

Topics Covered:

- Kafka basics
- Kafka Vs RabbitMQ, ActiveMQ
- Install and configure Kafka
- Hands-on Kafka
- Spark basics
- Spark Vs MapReduce
- Install and configure Spark on a cluster
- Spark alternatives
- Spark Streaming use cases
- Running Spark programs written Scala and
- Analytics using Spark
- Spark streaming
- Running Spark in various models like local, stand alone and in YARN
- Integrate Kafka with Spark streaming
- Integrate Kafka with MongoDB
- Simulate a Kafka-Spark streaming system and post few million messages and get them processed and make the processing result in NoSql database

Chapter 10: Projects and Resume preparation

Learning Objectives:

- Every participant will get a chance to do POC of a production grade project with real time data.

- At least 6 real time projects that are being used at production in various companies will be explained at end of course.
- 800 interview questions will be provided.
- A 400 page material will be provided to each and every individual.

About Instructor

The trainer has overall experience of 11 years, out of which 5 years in Hadoop and Big Data Analytics and rest into Java. The trainer had handled 60+ batches already. The trainer had handled few corporate batches also at big companies like Infosys and Deloitte. Apart from doing full time job, he had done multiple consulting assignments overseas.