Training for Bigdata and Hadoop

#I Background and Introduction

1. Introduction

Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

This course provides a quick introduction to BigData, Map Reduce algorithm, and Hadoop Distributed File System.

1.1 What is Big Data?

Big data means really a bigdata, it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not merely a data, rather it has become a complete subject, which involves various tools, techniques and frameworks.

1.2 What Comes Under Big Data?

Big data involves the data produced by different devices and applications. Given below are some of the fields that come under the umbrella of Big Data.

- **Black Box Data**: It is a component of helicopter, airplanes, and jets, etc.
- **Social Media Data**: Social media such as Facebook and Twitter hold information and the views posted by millions of people across the globe.
- **Stock Exchange Data** : The stock exchange data holds information about the 'buy' and 'sell' decisions made on a share of different companies made by the customers.
- **Power Grid Data** : The power grid data holds information consumed by a particular node with respect to a base station.
- **Transport Data** : Transport data includes model, capacity, distance and availability of a vehicle.

- **Search Engine Data:** Search engines retrieve lots of data from different databases.

2. Audience

This course has been prepared for TE students aspiring to learn the basics of Big Data Analytics using Hadoop Framework and become a Hadoop Developer.

3. Prerequisites

Before you start proceeding with this course, we assume that you have prior exposure to Core Java, database concepts, and any of the Linux operating system flavors.

This is the world of Internet of Things (IoT). Everything is available on internet, so handle the rise in data on internet and to manage it the various technologies have been introduced. Hadoop and Bigdata is one of them.

Apart from University requirement, In this era of advancement we should know the emerging technology hadoop and BigData. In order to prepare and make students ready for industry, Computer Engineering department has carved out a course that specifically aligns with industry requirements and conducted by industry experts.

Due to the advent of new technologies, devices, and communication means like social networking sites, the amount of data produced by mankind is growing rapidly every year. The amount of data produced by us from the beginning of time till 2003 was 5 billion gigabytes. If you pile up the data in the form of disks it may fill an entire football field. The same amount was created in every two days in 2011, and in every ten minutes in 2013. This rate is still growing enormously. Though all this information produced is meaningful and can be useful when processed, it is being neglected.

90% of the world's data was generated in the last few years.

#II Development of training program

The training program was identified as one of the add on courses at the end of the Semester I (2015-2016). The co-ordination of the course was handed over to Prof. Jyoti N. Gavhane. In the vacation, review was taken from TE students who are interested in developing the skills in the concern domain. The course structure was discussed with the HOD Prof. Dr. V. Y. Kulkarni.

- Coordinator visited different training institutes e.g. Edu Pristine, Blue Ocean Learning, to discuss and identify the course content with focus on topics to be covered in hands-on training .
- After discussing with the trainer, Blue ocean Learning institute was finalized to train the students.
- Both parties agreed to coordinate on timelines, financials as well as location for delivering the training program.

1. About the Course

The course 'Training of Bigdata and Hadoop' was designed as 24 hours of classroom and hands-on training, in which each session is of 3 Hrs and will be conducted per day. These sessions were conducted after college hours, so that it should not impact regular studies of the students. At a high level course covers a range of topics towards:

- Enabling students to understand basic Map Reduce programming concepts and programming for HDFS.
- To provide hands-on sessions to practice the concepts covered in the training.
- Enable students to develop applications in Java to implement MapReduce program.

2. About the trainer

Work history:

- php-postgresql-mongodb developer, db xento systems, Pune
- Hadoop trainer & technical content writer, technocrafty solutions, pune
- Hadoop mongodb trainer and consultant, freelancer, pune

Skills:

- Big data hadoop
  Hadoop 1.x, hadoop 2.x, hive, pig, flume, sqoop, spark, impala, mahout, kafka, hdp 2.0, cdh 5

- Data integration & visualization: informatica, tableau, pentaho
- Nosql & sql databases: mongodb, mysql, postgresql
- Programming: php, java, python

Recent projects and deployments:

- Has worked as a Consultant for couple of marketing companies to analyze the latest trends and help their clients to choose the right platform for targeted advertisements.
- Recent Training: MongoDB and Hadoop: Snapdeal ,manthan system, optra system jain university

#III Learning Product

1. Course Description

This course was designed for TE, BE and ME students of Computer Engineering department with an objective to make them aware with Bigdata and Hadoop technology and programming for map reduce.

The curriculum is divided into --- modules and is designed to be covered over a week period. The course was designed to ensure students get sufficient hands-on practice to master concepts.

2. Learning Objective

Upon completion of this course, participants will be able to:
Understand fundamentals of Concepts in Bigdata and hadoop etc
Understand fundamentals of Hadoop etc.
Be able to use the HDFS file system, debug and run simple Java programs for hdfs.
Be aware of the important topics and principles of software development and write better &more maintainable code
Be able to program using advanced Java topic like JDBC, Servlets and JSP .

Performance:

- Be able to write programs of simple to medium complexity.

3. Course Contents

1. What is Big Data

   - ✓ Introduction to Data Lake
   - ✓ Why do companies care for Big Data
   - ✓ What do we do out of Big Data
   - ✓ Where did this data come from
   - ✓ Social Media – A way to get the true customer insights
2. History of Hadoop

   - ✓ Hadoop Timeline
   - ✓ Why Hadoop
   - ✓ Hadoop 1.X Architecture
   - ✓ Hadoop 1.X Core Components
   - ✓ Hadoop 1.X Job Process

3. Importance of HDFS

   - ✓ HDFS Daemons
   - ✓ Name Node
   - ✓ Data Node
   - ✓ Secondary Name Node
   - ✓ Node Level Failure Handling in Hadoop 1.X
4. Different Phases in Map Reduce

   - ✓ Input – Output formats in each phase
   - ✓ Modeling Real World applications into Map Reduce
   - ✓ Understanding Map Reduce Program Execution
   - ✓ Problems in Map Reduce
5. Hadoop 1.X labs

   - ✓ Setting up a Pseudo Mode Hadoop Cluster
   - ✓ Executing a sample Map Reduce Program
   - ✓ Writing and Understanding Basic Map Reduce Program
6. Apache HIVE

   - ✓ Introduction to Hive Meta store
   - ✓ SQL vs. Hive
   - ✓ Hive Query language
   - ✓ Managed and External tables
   - ✓ Querying data
   - ✓ Hive thrift server
   - ✓ Working on HIVE Beeline
   - ✓ Joins, Sub Queries and other Aggregations
7. Apache PIG

   - ✓ Introduction to PIG

- ✓ Map Reduce vs. PIG
- ✓ PIG in local mode
- ✓ PIG in Map Reduce mode
- ✓ Local mode vs. Hadoop mode
- ✓ Execution mechanism and data processing
- ✓ Writing PIG scripts
- ✓ User defined functions in PIG

8. SQOOP

- ✓ Introduction to SQOOP framework
- ✓ SQOOP flavors of Import
- ✓ SQOOP flavors of Export
- ✓ SQOOP CLI Options

9. FLUME

- ✓ Introduction to Messaging Service
- ✓ Applications of a Messaging Service
- ✓ FLMUE Architecture Framework
- ✓ Working of a FLUME Agent
- ✓ Understanding FLUME Configurations
- ✓ Hadoop Ecosystem Labs
- ✓ Importing data from MYSQL and querying it using HIVE
- ✓ Configuring a FLUME agent to listen to local log files

#IV Outcome of Course

1. Exam and Test:

During the course the trainer has conducted 1 test after half syllabus completion and final exam at the end of the course. Depending on the marks scored by the participant, grades have been given.

2. Certification:

The training institute with collaboration of MIT Pune has distributed the certificate to each participant as Hadoop developer at basic level.

3. Participants:

Sr. No.	Roll No.	Name
1	302020	Kavale Arya Mahesh

2	302033	Agrawal Parag Sanjay
3	302035	Ankit Anil Sharma
4	302076	Kedare Rohit Gautam
5	302077	Kewal Upendra Shah
6	302080	Krishnamohan Manmohan
7	302086	Manas Pandey
8	302088	Naval Vaidya
9	302091	Rishav Dasgupta
10	303001	Bhalekar Neha
11	303003	Deshmukh Namrata Balasaheb
12	303004	Gawade Sayali Dattatraya
13	303008	Naik Neha Hemant
14	303009	Nikam Shraddha Kailas
15	303011	Pande Priyanka Deepak
16	303017	Raipurkar Rutuja Mukund
17	303021	Nakashe Sayali Sanjay
18	303029	Vanvari Shriya Nirmal
19	303030	Tibdewal Vinita Manoj
20	303031	Bhimte Ashish Dhanaraj
21	303033	Bhangare Kalpesh
22	303036	Mankoji Ashish Rajendra
23	303040	Palliyalil Suraj S.
24	303041	Panwala Aamir Juned
25	303042	Paramane Hrishikesh Ravindra
26	303049	Pawar Ashish Kalyan
27	303053	Rajmane Omkar Ramesh
28	303059	Joshi Rushikesh Vinayak

29	303061	Saha Ankan Ashish
30	303062	Sakargayen Kundan Vivek
31	303065	Shah Yash Jayant
32	303070	Sonawane Saurabh Vishwasrao
33	303072	Tayade Harshal Milind
34	303073	Tayade Shubham Gajanan
35	303069	Pratik Singhavi
36	303024	Snehal Shinde
37	303007	Varsha Mundra
38	303028	Sakshi Uplenchwar
39	303010	Sarika Padare
40	ME	Prajakta Joglekar
41	ME	Priti Patil
42	ME	Amit Wale
43	ME	Payal Kumbhalkar
44	ME	Mayri Dhamdhere
45	ME	Pooja Kavare

Prof. Jyoti N. Gavhane
Course Co-Ordinator
Department of computer Engineering,
MIT Pune.

Prof. Dr. V. Y. Kulkarni
HOD, Department of computer Engineering
MIT Pune