# IMPLEMENTATION OF ATTENDANCE MANAGEMENT SYSTEM USING HADOOP

Project report submitted in partial fulfillment of the requirement for the degree of Bachelor of Technology

in

**Computer Science and Engineering/Information Technology**

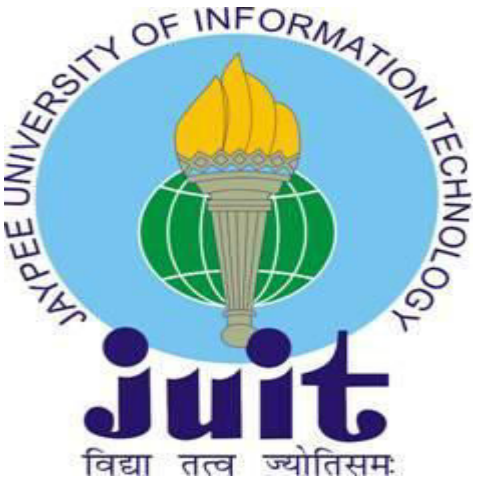
By

Rohan Gupta(151285)

Under the supervision of

Dr. Hemraj Saini(Associate Professor)

to




Department of Computer Science & Engineering and Information Technology

**Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh**

# Candidate's Declaration

We hereby declare that the work offered in this report entitled " Attendance Management System using Hadoop" in partial fulfillment of the necessities for the award of the diploma of Bachelor of Technology in Computer Science and Engineering/Information Technology submitted inside the branch of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology Waknaghat is an authentic document of our personal work achieved over a period from August 2018 to December 2018 below the supervision of(Dr. Hemraj Saini) (Associate Professor (CSE)).

The depend embodied in the report has now not been submitted for the award of every other diploma or diploma.

Rohan Gupta(151285)

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Dr. Hemraj Saini
Associate Professor
CSE
Dated:

# ACKNOWLEDGEMENT

We take this opportunity to express our profound gratitude and deep regards to our guide Dr. Hemraj Saini(Associate Professor , CSE) for her exemplary guidance, monitoring and constant encouragement throughout the course of this project. The blessing, help and guidance given by her time to time shall carry us a long way in the journey of life on which we are about to embark.

We are also obliged to staff members of JUIT University for the valuable information provided by them in their respective fields. We are grateful for their cooperation during the period of our assignment.

Lastly, we thank almighty, our parents and our classmates for their constant encouragement without which this assignment would not have been possible.

## ABSTRACT

Big Data has given crucial choice over the few years due to the quantity of records generated each year. The position of huge statistics comes into use when we need to store large amount of statistics which ordinary database aren't able to procedure and save.The undertaking in Big Data is not handiest storing of the records,however also accessing and reading the desired facts in specified amount of time. The problems of big statistics , they may be essentially solved by the usage of Hadoop. Hadoop is an open source platform and makes use of Map Reduce programming version for solving issues of Big Data. Hadoop provides the service to users to shop and method bulk amount of information which normal databases aren't capable of offer.

We are currently working on the implementation of Attendance Management System using Hadoop.It should provide real time functionality to the user.Our system is designed to maintain the attendance of the different colleges. This project provides a way of analyzing big data using Apache Hadoop which will process and analyze data on Hadoop cluster

# Contents

## Introduction



## 2 Literature



## 3 System Development

# List of Figures

CHAPTER**1**

# PROJECT OBJECTIVES

## 1.1 Introduction

The task is based totally on the implementation of Attendance Management System by the use of Apache Hadoop.It should provide the real time functionality to the user. Random Generator is used which is given the values of absent and present. Streams are Generated through Wamp server and then after this preprocessing of data occurs in which raw data is converted into some structured data. Aggregation and Filteration of data happens here. After this Output report is displayed in which we enter any name and University and we get the required attendance of the student.

## HADOOP

Hadoop is used for distributed storage and processing of every huge data units on laptop clusters constructed from the commodity hardware. It do all the processing in the parallel manner. Hadoop is based on Map Reduce algorithm. The Core of hadoop consists of a storage part known as HDFS and other part is map reduce model in which programming is done.It is built from the commodity hardware. Modules are designed in such a way assuming that hardware failures are common and should be handled and solved by the hadoop framework.

Major functions of Hadoop:

- Viewing all data
- Keep backup copies of data
- Distributive Processing

It consists of :

- **HDFS-**A distributed file system providing high bandwidth
- **Yarn-**Is a platform responsible for managing computing resources in clusters and using them for scheduling applications.
- **Map Reduce-**An implemention model for data processing.

## Why we use Hadoop?

- Disk Seek Time.
- It has a resistance to hardware failure.
- Processing Time is decreased when we are using Hadoop .It is fast.
- Horizontal scaling is linear.It can store and distribute very large data sets in parallel.
- Data is too big(terabytes per day)

## SQOOP

Sqoop is a device that is used to transfer the statistics between relational databases and Hadoop.

It kicks off the Map Reduce jobs to handle importing or exporting your data. Its role is to populate table on the HDFS, then you can use the Hive for querying

## How SQOP Works :

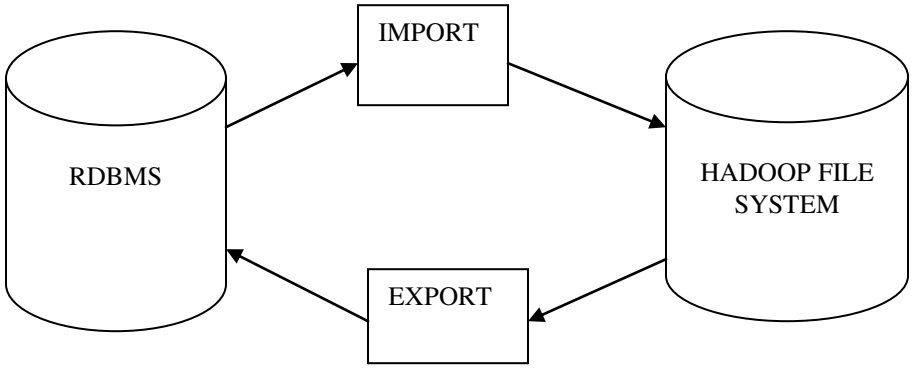



**Figure 1:Sqoop Architecture**

## SQOOP IMPORT:

It imports individual tables from RDBMS to HDFS. All record are stored as a text data in the text files.

Syntax: sqoop import (generic arguments) (export arguments).The import tool imports tables from RDBMS to HDFS. Each row in a desk is dealt with as a record in HDFS. All data are stored as text information in textual content files or as binary facts in Avro and Sequence documents.

## EXPORT:

It exports a file from HDFS to RDBMS. The export device exports a set of documents from HDFS again to an RDBMS. The files given as enter to Sqoop incorporate statistics, which might be referred to as as rows in desk. Those are study and parsed into a set of records and delimited with consumer-special delimiter.

Syntax: .sqoop export (generic arguments) (export arguments).

## PIG IN HADOOP

Pig is used to research massive statistics sets. The statistics sets are represented into records flows. It is used normally with the Hadoop. It can be utilised to perform all the records manipulation us. Pig falls in the category of HLL. Join can be done without problems done in Pig. Pig is used to carry out many operations like join, sorting, filter etc.

Apache Pig is used to carry out massive records assets. The additives utilized in Pig are Optimizer, Parser,
Execution Engine and Compiler .The role of Pig in Map Reduce mode is to load records from HDFS and stores the results lower back in to HDFS. Pig is a excessive level scripting language this is used with Apache Hadoop. Pig permits information people to jot down complex information modifications without knowing Java. Pig's simple SQL-like scripting language is called Pig Latin, and appeals to builders already familiar with scripting languages and SQL.

Unstructured ,semistructured ,and polystructured are all a term for the records that doesn't in shape well in to the relational model. The records is like JSON,XML,RDF or different sources of data with a schema which can

range from file to record .Hive is extraordinarily powerful for dealing with data that doesn't quite suit into complex changes that is probably otherwise essential to address .This type of information in conventional

a relational gadget. Hive can also gracefully cope with the statistics that do not strictly comply with a table's schema for e.G If a few columns are lacking from the specific facts hive can address document via treating lacking columns as nulls.

## Tkinter

Python gives a couple of alternatives for growing GUI (Graphical User Interface). Out of all the GUI techniques, tkinter is most typically used method. It is a standard Python interface to the Tk GUI toolkit shipped with Python. Python with tkinter outputs the fastest and simplest way to create the GUI programs. Tkinter gives an expansion of integrated functions expand interactive and featured GUI (Graphical User Interface). After() function is likewise a Universal function which may be used without delay on the root in addition to with other widgets.

Following steps are taken to create Tkinter-

1.Import module.

2.Create the Window

3.Add Widgets

4.Apply the event Trigger

## Streaming

Hadoop streaming is a software that incorporates the Hadoop distribution. This application allows you to create and run Map/Reduce jobs with any executable or script because the mapper and/or the reducer. When a script is particular for mappers, each mapper assignment will release the script as a separate technique when the mapper is initialized. As the mapper mission runs, it converts its inputs into strains and feed the traces to the standard enter (STDIN) of the system. In the period in-between, the mapper collects the road-oriented outputs from the usual output (STDOUT) of the process and converts each line into a key/fee pair, that is accumulated as the output of the mapper. By default, the prefix of a line as much

as the primary tab individual is the key and the rest of the line (with the exception of the tab character) may be the cost.

## Preprocessing

The data generally comes in raw form and it is not structured and it is in the abrupt form. So here we convert the

Data into a structured form like Tree or in an easily recognizable pattern. From this data goes into a Computing

Server and then finally a report is generated in which there is an output showing the attendance record of any

Student studying in any university .We get our required output and Real-world information is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to include many mistakes. Steps taken

in data preprocessing are-

Step i : Import the libraries

Step ii : Import the statistics-set

Step iii : Check out the lacking values

Step iv: See the Categorical Values

Step v : Splitting the facts-set into Training and Test Set

Step vi : Feature Scaling

## Random Generator

A random wide variety generator  is a mathematical assemble, both computational or as a hardware device, this is designed to generate a random set of numbers that should no longer show any distinguishable styles of their appearance or era, hence the phrase random.

A real random number generator cannot rely on mathematical equations and computational algorithms to get a random number because if there's an equation involved, then it isn't always random. Applications that gain from genuine randomness are video games which include those associated with gambling like bingo, card games, the lottery and comparable video games. Video games that emphasize random loot series additionally benefit from genuine randomness, as pseudorandom variety era can cause frustration due to the fact it is able to move a long time without the goal wide variety being hit or the same variety may be obtained again and again

## Computing Server

Nearly all personal computers are capable of serving as network servers. However, commonly software program/hardware device devoted computer systems have functions and configurations just for this project. For example, committed servers may have excessive-overall performance RAM, a quicker processor and several high-ability hard drives. In addition, dedicated servers may be related to redundant power resources, several networks and different servers. Such connection features and configurations are essential as many client machines and purchaser packages may additionally depend on them to feature correctly, efficaciously and reliably.

## 1.2Problem Statement

It becomes very difficult to manage and organize the data of attendance of different colleges. Since our system will provide real time functionality it will have a large set of data(in

terabytes) in terms of variety, velocity and volume..We need to design a system which is capable of handling and analyzing the attendance of students studying in different colleges. We use Big Data to monitor and track the attendance of the students. It becomes very difficult for traditional methods for providing Real Time Functionality, therefore we are using Streaming .

Hadoop Streaming is done using Wamp Server which provides an efficient way for calculating the data. Also

We need a system which is capable of processing a large amount of data. We use SQOOP in our project to import the file from relational databases and Hadoop. It becomes very difficult and time consuming for us to sorting and processing he large data sets in our Attendance Management System .

Therefore the problem is solved by using the Map Reduce Algorithm. Map Reduce is a programming model that allows you to process your data in different clusters. We need a system which should be capable of providing or displaying the record of the students whose attendance is below a certain range. To get the desired result we have used HQL for querying the database to get the desired result.HQL is an object oriented query language which has a syntax similar to SQL. The difference between both of them is that HQL works with persistent objects and their properties whereas SQL is based on relational database model. We have used Hive that takes queries , make distributed data to set on the database.

Random Generator is used for generating the large amount of data for attendance .Then after this data is preprocessed which means that firstly data is not in a structured form and we are not able to recognize any pattern required for simplification. So we convert the data into a structured form and there we send the structured form of data to the computing server which generates an output report for the student belonging to any university and we Can check his attendance.

We can manage the attendance for any number of universities and any student at one place easily. With the help of Big Data Analytics we can easily manage and through hadoop streaming we can easily generate our output report which would have been otherwise very much difficult to manage. We can see the past record very easily. Since the data is kept confidential for the universities so we use Random Generator for generating large amounts of data and solve the problem quite easily.

## 1.3Objective

Our objective is to implement the "Attendance Management System" which should provide real time functionality to the user. In first step we need to gather and generate the information through various Random Generator programs etc. In second step the data is preprocessed converting unorganized unstructured data into the structured form. Aggregation and filteration occurs here which are very much important to eliminate the unwanted data. Data is stored in the HDFS and we have used the Map Reduce Algorithm where the data is mapped to the data sets and reduced to the smaller size. We have used hadoop streaming that generate streams of
Data and make distributed data to set on your database. And finally we have analysed the attendance of different students studying in different colleges with the help of computing server which generates the output report required .

Map Reduce is well known as the heart of Hadoop Ecosystem as each task of the Big Data is computed using this technique .Map Reduce distributes the processing of your data on your cluster. It divides the data in to partitions that are mapped(transformations) and reduced

(aggregated by mapper an reducer functions). Reduce is the technique that can be used to provide huge scalability across hundreds or thousands of clusters in Hadoop cluster.

A person can easily check the attendance belonging to any university for any student and can manage and update it at any time very easily. A GUI is made using Tkinter which consists of following steps-

i) importing the data
ii) creating main Window
iii) adding the widgets
iv) apply the event trigger

We can login as an admin to update and also login as to check the attendance for any student studying in any university and can see his records for any number of days. it can be easily managed through Hadoop Technologies. Attendance is also shown through local FS and hadoop streaming which solves the problem of handling and managing large amount of data which would otherwise have been difficult to organize and manage.
So This project serves all the objectives of developing an efficient attendance management system in the field of education and computing server generated the report which is very much easily calculated and displayed in less time which would have not been possible otherwise.

## 1.4Methodology

The following techniques we're going to comply with to investigate the attendance records:

- Designing Attendance Application
- Random Generator for generating lots of data.
- Hadoop Streaming
- Preprocessing of Data

- Computing server
- Output report

### 1.4.1 Designing Attendance Application

First we will create Graphical User Interface with the use of tkinter. Tk() is same antique interface library for Py . Py whilst mixed with tk() offers a fast and clean manner to create GUI programs. tkinter gives an effective object-orientated interface to the tk GUI toolkit. Creating a GUI application the use of tkinter is an easy assignment. Steps–

- A module (tk) is imported.
- Principal window is created. Include widgets in it.

Example

```
import.tkinter
a= tkinter.Tk ()
a.Mainloop()
```

### 1.4.2 Random Generator

We have taken a random generator which is taking the values of absent and present and is denoted by a Box symbol.. It is regularly in the shape of a function or blocks of code utilized in software program packages together with games in which an element of danger is needed. Random number mills are just the contemporary utility of randomness gadgets which have existed since historical times consisting of cube, shuffled playing cards, flipping coins and even drawing straws. In contemporary computing, random wide variety mills are carried out

thru programming based totally on deterministic computation, however this isn't always absolutely considered as proper random because the output can truly be anticipated if all seed values are regarded, so that is referred to as pseudorandom quantity era. From Random generator Streaming of data occurs and from there it is preprocessed
easily and convert into a structured form.

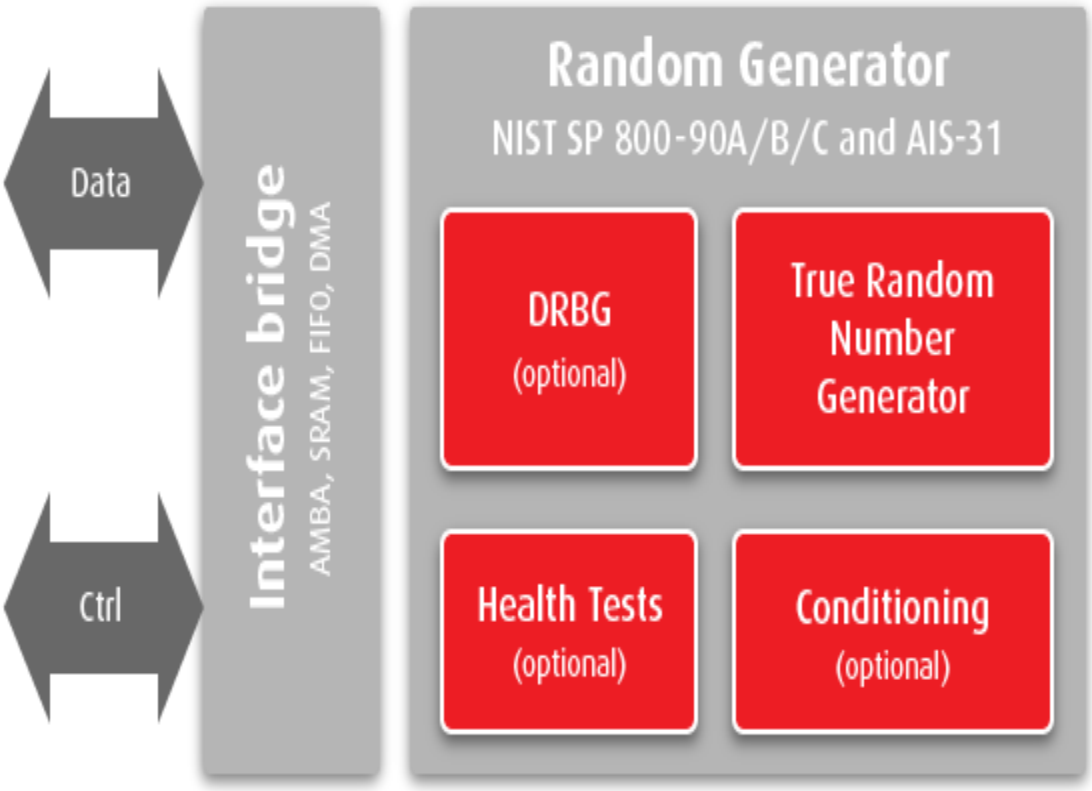



**Figure 2: Random generator Architecture [Reference no.(i)]**

### 1.4.3 Hadoop Streaming

Streams of data is there which is preprocessed later in which the unstructured data is converted into a more structured form . Wamp server is user in hadoop streaming. Hadoop streaming is a software that incorporates the Hadoop distribution. This application allows

you to create and run Map/Reduce jobs with any executable or script because the mapper and/or the reducer.

### 1.4.4 Preprocessing of Data

Data comes in raw and unstructured form. We need to preprocess it so that we can gather the required information and eliminated unwanted results or information. From this data goes into a Computing Server and then finally a report is generated in which there is an output showing the attendance record of any

Student studying in any university. We get our required output and Real-world information is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to include many mistakes. Steps taken

in data preprocessing are-

Step i : Import the libraries

Step ii : Import the statistics-set

Step iii : Check out the lacking values

Step iv: See the Categorical Values

Step v : Splitting the facts-set into Training and Test Set

Step vi : Feature Scaling

### 1.4.5 Computing Server and Report

Finally an output report is generated which displays the attendance record of any student studying in any university .We can see the past record very easily. We can see the past record

very easily. Since the Unstructured ,semistructured, andpolystructured are all a term for the records that doesn't in shape well in to the relational model. The records is like JSON,XML,RDF or different sources of data with a schema which can

range from file to record. Hive is extraordinarily powerful for dealing with data that doesn't quite suit into complex changes that is probably otherwise essential to address .This type of information in conventional

a relational gadget. Hive can also gracefully cope with the statistics that do not strictly comply with a table's schema for e.G If a few columns are lacking from the specific facts hive can address document via treating lacking columns as nulls.

data is kept confidential for the universities so we use Random Generator for generating large amounts of data
and solve the problem quite easily.

# CHAPTER 2

# LITERATURE SURVEY

Over beyond 10 years, industries and corporations does not want to save and perform a whole lot operations and analytics on information of the clients but around from 2005 the need to transform the whole lot in to statistics is tons entertained to meet the requirement of people so Big Data came in to image in the actual time commercial enterprise analysis of processing facts. Big Data means the information that is producing round us in regular lifestyles. It usually exceeds the capability of ordinary traditional conventional databases. For instance via combining a huge wide variety of alerts from the consumer movements and people in their buddies, Facebook evolved the large community region to the users to percentage their perspectives, ideas and a lot many stuff.

.The cost of Big Data to an organization falls in to two classes analytical use and permitting new products based totally on the existing once. Big Data can display the issues hidden through statistics that is too pricey to manner and carry out the analytics together with user transactions, social and geographical records issues faced by the enterprise.

The essential characteristics and demanding situations of Big Data are Volume, Velocity and Variety. These are referred to as 3V's of Big Data which might be used to characterise specific elements of Big Data.

**Figure 3: Big Data Challenges**
**[Reference No.(iii)]**

## Characteristics of Big Data

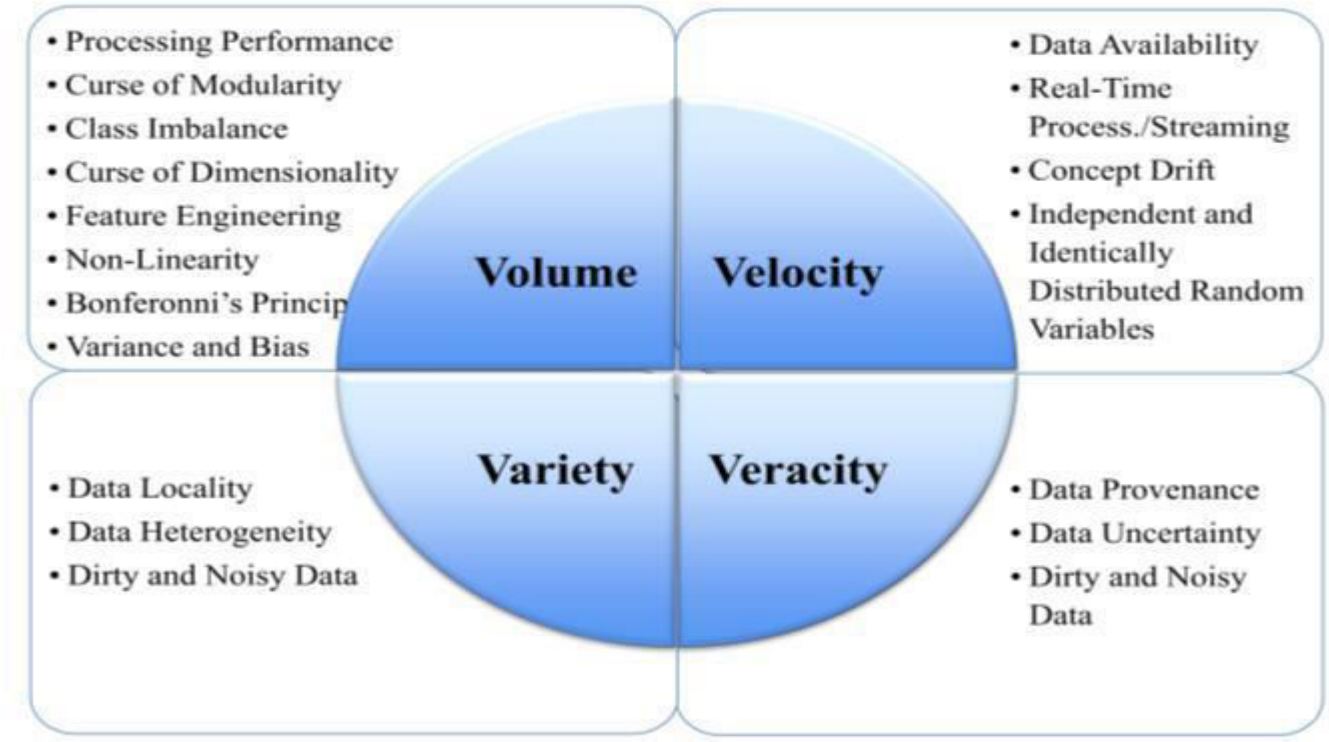



**Figure4:Big Data Characteristics [Reference no(iii)]**

## Benefits

**• Businesses can make use of outside intelligence even as taking choices**

Access to social records from search engines like Google and websites like Facebook, Twitter are allowing businesses to fine tune their business techniques.

**• Improved customer service**

Traditional patron remarks structures are becoming changed with the aid of new systems designed with 'Big Data' technologies. In those new structures, Big Data and natural language processing technologies are getting used to study and compare patron responses.

**• Early identity of threat to the product/offerings, if any**

**• Operational Performance**

'Big Data' technologies may be used for developing staging location or touchdown quarter for brand new statistics earlier than figuring out what information need to be moved to the information warehouse. In addition, such integration of 'Big Data' technologies and data warehouse helps employer to dump now and again accessed statistics.

### 2.2.5Big Data Analytics: A Literature Review Paper

**Authors:**
Ahmad Elregal, Lulea University of Technology
Nada Elgendy, German University in Cairo
**Year:** August 2014
**Abstract:**
The information generation, huge portions of statistics have become to be had accessible to selection makers. Big statistics refers to datasets that aren't best big, but moreover excessive in variety and speed, which makes them hard to handle using traditional device and techniques. Due to the speedy increase of such facts, answers want to be studied and supplied for you to deal with and extract rate and understanding from those datasets. Furthermore, desire makers need with a view to gain treasured insights from such varied and hastily changing statistics, ranging from every day transactions to customer interactions and social community records. Such cost can be supplied the usage of big records analytics, this is the software program of superior analytics strategies on massive records. This paper goals to investigate a number of the incredible analytics strategies and tools which can be executed to

big statistics, similarly to the opportunities provided by way of the software of huge statistics analytics in various decision domains.

**Big Data and Quality: A Literature Review Paper**

**Authors:**

- GunaAbdulkhadarLakshan
  College of Electrical Engineering
- SanzaBranes
  Mihazlo Pupin Institute

**Publication Year:** 2016 24th Telecommunications Forum (TELFOR)

**Date :** 22nd November2016

**Abstract:**

Big Data means information volumes in the variety of Exabyte (10^ 18) and past. Such volumes exceed the potential of cutting-edge on-line garage and processing systems. With traits like volume, speed and range large information throws challenges to the traditional IT institutions. Computer assisted innovation, actual time statistics analytics, patron-centric commercial enterprise intelligence, enterprise wide decision making and transparency are viable blessings, to mention few, of Big Data. There are many problems with Big Data that warrant nice evaluation methods. The problems are referring to storage and shipping, control, and processing. This paper throws mild into the existing country of great troubles related to Big Data. It presents valuable insights that can be used to leverage Big Data science activities.

**HADOOP- MAPREDUCE**

The records is to be had to customers and the assessment is accomplished from the databases in a certain quantity of time. By regular method, it's hard to gain massive data challenges. Hadoop is taken into consideration simplest of the answer to resolve massive facts problems. Hadoop is open source and is used for massive scale computation. It is used to way the facts on a community commodity of hardware gadget. The crucial additives in Hadoop are commodity hardware and Map Reduce.

Map Reduce called the heart of hadoop and every assignment is accomplished via this technique simplest .The phrases Map Reduce refers to the 2 obligations within the Hadoop this is the Map and Reduce .In step one it takes the records and converts it in to a few other set of facts.Here every word is referred as a key and the variety of occurences is treated as price .Therefore map reduce tuple includes of key price pairs.The 2nd step includes reduce operation wherein it takes the output from map as an input and combine those statistics tuples in to smaller set of tuples

The primary troubles in big statistics are storing and accessing the facts from large amount of records from the clusters. There is a platform needed to perform widespread computing since the facts is growing, and the statistics is saved into unique garage locations. Another problem is fetching the facts using a huge social media community. It will be very tough to layout the algorithm for dealing with the problems due to massive statistics.

Hadoop is the era that is used for evaluation of attendance and works for big statistics. Hadoop runs the usage of Map Reduce Technique, wherein the records is processed in parallel. Hadoop works on diverse computer systems and plays computational analysis on large quantities of records. Hadoop which performs Map Reduce approach approaches huge quantities of data on heaps of nodes or clusters. Hadoop involves the use of HDFS File System. The HDFS is based totally on the Google Record Service and affords distributed document gadget this is designed to run on a machine

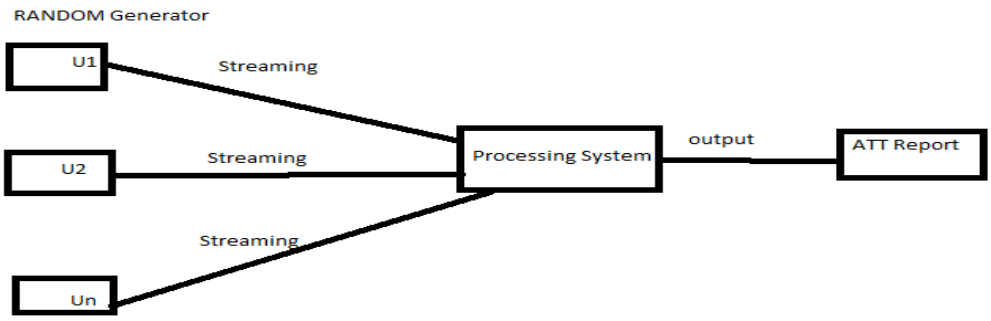
# Chapter 3

# System Development

. In first step we need to gather and generate the information through various Random Generator programs etc. In second step the data is preprocessed converting unorganized unstructured data into the structured form. Aggregation and filteration occurs here which are very much important to eliminate the unwanted data. Data is stored in the HDFS and we have used the Map Reduce Algorithm where the data is mapped to the data sets and reduced to the smaller size. We have used hadoop streaming that generate streams of

Data and make distributed data to set on your database. And finally we have analysed the attendance of different students studying in different colleges with the help of computing server which generates the output report required .

**Figure 4: Project Architecture**




## 3.1 Creating Attendance Application

First we will create Graphical User Interface with the use of tkinter. Tk() is same antique interface library for Py . Py whilst mixed with tk() offers a fast and clean manner to create GUI programs.tkinter gives an effective object-orientated interface to the tk GUI toolkit.Creating a GUI application the use of tkinter is an easy assignment.. It is a standard Python interface to the Tk GUI toolkit shipped with Python. Python with tkinter outputs the

fastest and simplest way to create the GUI programs. Tkinter gives an expansion of integrated functions expand interactive and featured GUI (Graphical User Interface). After() function is likewise a Universal function which may be used without delay on the root in addition to with other widgets.

Steps

- A module(tk) is imported.
- Principal window is created. Include widgets in it.
- Apply the event Trigger
- Finally a Gui is created

Example

```
Import .tkinter
a= tkinter.Tk()
a.Mainloop()
```

## 3.2 Getting Data Using Random Generator

It is regularly in the shape of a function or blocks of code utilized in software program packages together with games in which an element of danger is needed. Random number mills are just the contemporary utility of randomness gadgets which have existed since historical times consisting of cube, shuffled playing cards, flipping coins and even drawing straws.

.

## 3.3 Hadoop Streaming

Streams of data is there which is preprocessed later in which the unstructured data is converted into a more structured form .Wamp server is user in hadoop streaming. Hadoop streaming is a software that incorporates the Hadoop distribution. This application allows you to create and run Map/Reduce jobs with any executable or script because the mapper and/or the reducer.

An Oozie workflow is a collection of movements organized in adirected acyclic graph (DAG). This graph can comprise two varieties of nodes: manipulate nodes and action nodes. Control nodes, which can be used to outline mission chronology, provide the policies for beginning and ending a workflow and manage the workflow execution direction with viable selection points known as fork and be a part of nodes. Action nodes are use

cause the execution. **[Reference no(ii)]**

## 3.4 Preprocessing of data:

Data comes in raw and unstructured form. We need to preprocess it so that we can gather the required information and eliminated unwanted results or information. From this data goes into a Computing

Server and then finally a report is generated in which there is an output showing the attendance record of any

Student studying in any university. We get our required output and Real-world information is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to include many mistakes. Steps taken

in data preprocessing are-

Step i : Import the libraries

Step ii : Import the statistics-set

Step iii : Check out the lacking values

Step iv: See the Categorical Values

Step v : Splitting the facts-set into Training and Test Set

Step vi : Feature Scaling

Query for inserting student data  using HQL:

INSERT INTO project(name,attendace) VALUES ( '${hiveconf:name}',${hiveconf:attendance},${hiveconf:roll},'${hiveconf:class}','${hiveconf:email}',${hiveconf:phone})

Query for searching student record using HQL:

SELECT * FROM project WHERE class='${hiveconf:class}';

Query for searching the entire  record using HQL:

select * from project;

Each row in a desk is handled as a file in HDFS.

When we post Sqoop command, our principal mission gets divided into sub obligations that's treated by using person Map Task internally. Map Task is the sub project, which imports part of facts to the Hadoop Ecosystem. Collectively, all Map responsibilities imports the entire information.

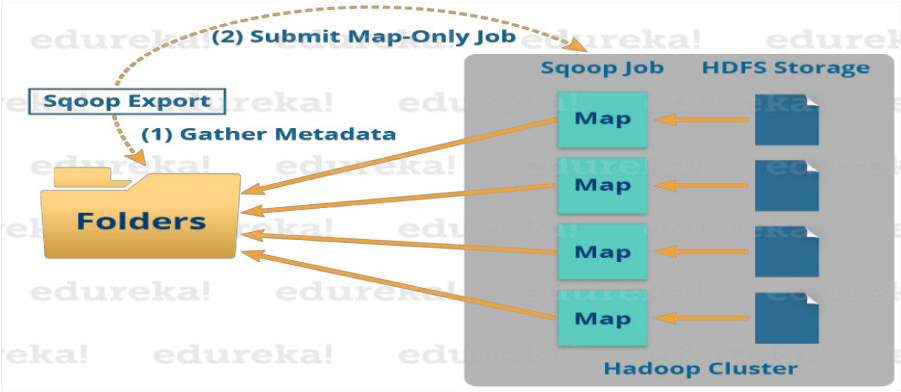


**Figure 5: Import And Export of data[Reference No(ii)]**

Export additionally works in a comparable manner The export device exports a fixed of documents from HDFS lower back to an RDBMS. The documents given as enter to Sqoop include statistics, which can be referred to as as rows in table. When we put up our Job, it's miles mapped into Map Tasks which brings the chew of statistics from HDFS. These chunks are exported to a based totally statistics destination. Combining a majority of these exported chunks of information, we get keep of the complete statistics at the vacation spot, which in maximum of the instances is an RDBMS.

mysql -root -cloudera</home/cloudera/Desktop/EP/sql

sqoop.import–connect.–username.root–password.cloudera --table.project–hive.-import

## Characterizing Data

Unstructured ,semistructured ,and polystructured are all a term for the records that doesn't in shape well in to the relational model. The records is like JSON,XML,RDF or different sources of data with a schema which can

range from file to record. Hive is extraordinarily powerful for dealing with data that doesn't quite suit into complex changes that is probably otherwise essential to address .This type of information in conventional

a relational gadget. Hive can also gracefully cope with the statistics that do not strictly comply with a table's schema for e.G If a few columns are lacking from the specific facts hive can address document via treating lacking columns as nulls.

It is pretty easy to look that there is a good bit of acomplexity to this statistics systems.Since aJSON can contain nested information structures it becomes very hard to force JSON information in to a general relational schema.Processing JSON information in a relational database might in all likelihood require a significant transformation making the task an awful lot extra cumbersome.. Looking at this unique bit of JSON there are some very interesting fields at the very top there may be a item whose existence suggests that those attendance become checked by means of another person.

## Complex-Data Structures

Hive has native aid for the set of records systems that normally might either not exist in a relational database or could require definition of custom sorts. They are all the usual gamers integer, strings, floats and so on but thrilling one are the more extraordinary maps, arrays and structs. Maps and arrays works in a fairly intuitive manner just like how they paintings in many scripting languages. If the consumer mentions array is empty, Hives will just go back NULL for that document. Hive can take the queries, make dispensed data to set for your database.

## Serializers- & Deserializers-

If the consumer mentions array is empty, Hives will just go back NULL for that document. Hive can take the queries, make dispensed data to set for your database..The deserialiser

interface takes string or binary illustration of a file and translate into java item that hive can manage the serialiser however will take a java object that hive has been operating with ,and change into something that hive can write to HDFS .Commonly Deserialisers are used at query time to execute SELECT statements and serialisers are used while writing data .

## 3.5 Computing Server

Nearly all personal computers are capable of serving as network servers. However, commonly software program/hardware device devoted computer systems have functions and configurations just for this project. For example, committed servers may have excessive-overall performance RAM, a quicker processor and several high-ability hard drives. There are a number of categories of servers, which includes print servers, record servers, community servers and database servers. In concept, every time computers proportion sources with purchaser machines they're taken into consideration servers.

## 3.6 Output Report

The primary troubles in big statistics are storing and accessing the facts from large amount of records from the clusters. There is a platform needed to perform widespread computing since the facts is growing, and the statistics is saved into unique garage locations. Another problem is fetching the facts using a huge social media community. It will be very tough to layout the algorithm for dealing with the problems due to massive statistics.

Hadoop is the era that is used for evaluation of attendance and works for big statistics. Hadoop runs the usage of Map Reduce Technique, wherein the records is processed in parallel. Hadoop works on diverse computer systems and plays computational analysis on large quantities of records. Hadoop which performs Map Reduce approach approaches huge quantities of data on heaps of nodes or clusters..

Finally an output report is generated which displays the attendance record of any student studying in any university. We can see the past record very easily. We can see the past record very easily. Since the data is kept confidential for the universities so we use Random Generator for generating large amounts of data and solve the problem quite easily.

# CHAPTER 4

# ALGORITHMS

4.1 Map Reduce Algorithm:

- Map Reduce Algorithm distributes the processing of data on your cluster. It divides the data in to partitions that are mapped and reduced by the mapper and reducer functions. These are aggregated by mapper and reducer functions. Map reduce sorts and groups the Mapped Data.
- Mapper converts raw source data in to key/value pairs. Extract and organise the data in an organised fashion.

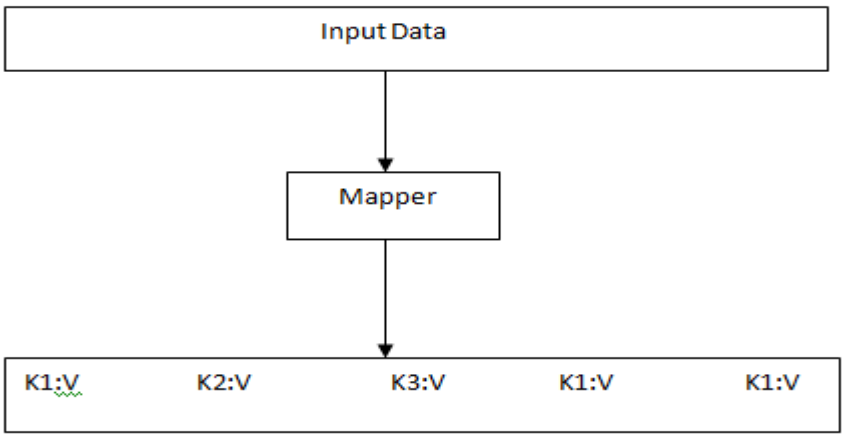



**Figure 6: Mapper Job**

i)

The Reducer processes each key values. Reducer is called once for each key.

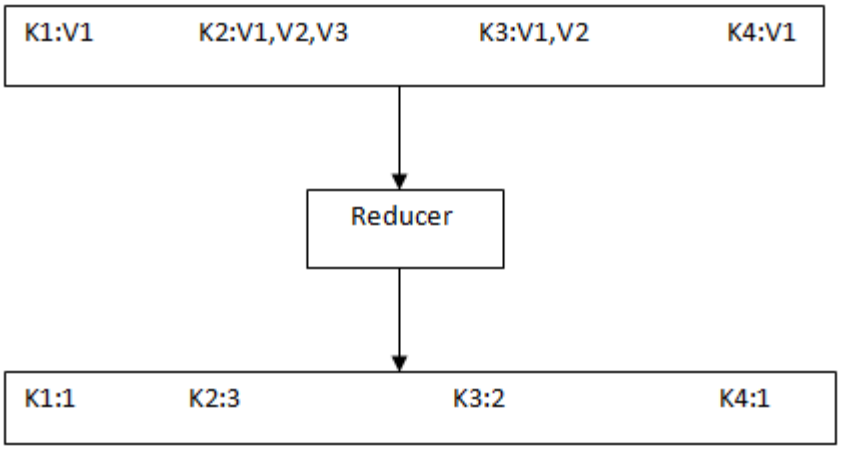



**Figure 7:Reducer Job**

- How are Mapper and Reducer written: Map Reduce is natively JAVA. Streaming allows interfacing to other languages.

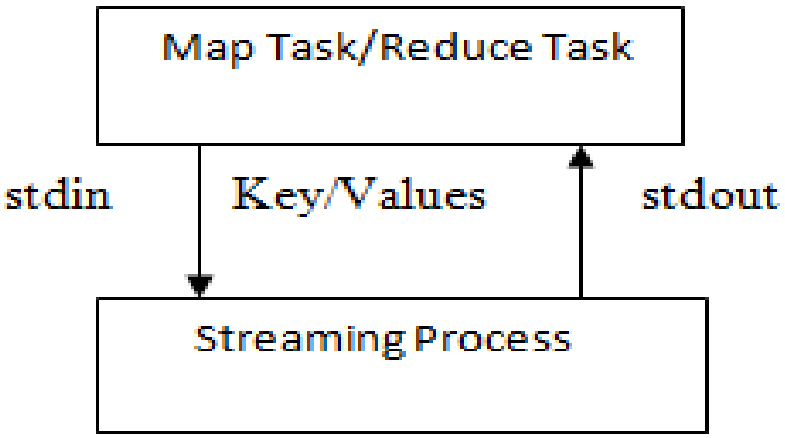


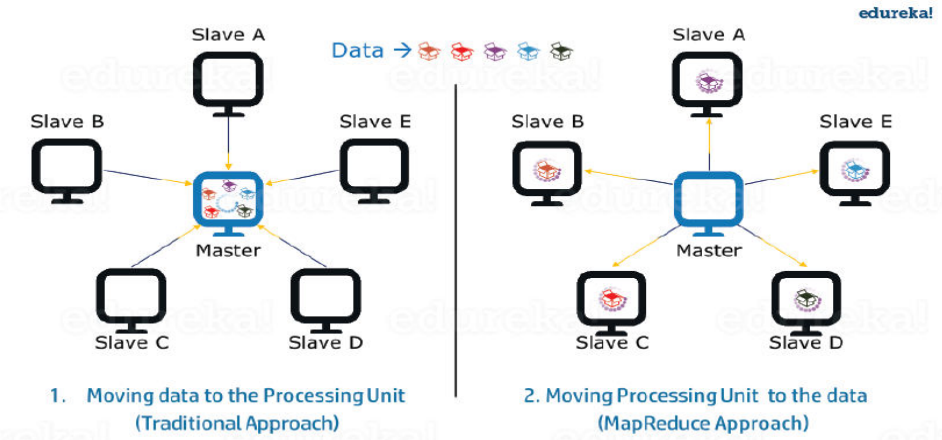

**Figure 8: Implementation of Map Reduce Algorithm**

## Advantages of Map Reduce:

### 1. Processing(Parallel)->

In Map -Reduce, we are partitioning the task among different nodes and every node works with a part of the task concurrently. So, Map Reduce is based totally on divide. Then conquer paradigm is applied which lets in us to gadget the information with the use of one of a kind machines. As the records is processed through the usage of a couple of system as opposed to a unmarried system in parallel, the time taken to approach the facts gets decreased through a terrific quantity as confirmed inside the decide underneath.

**Figure 9:Master and Slave [Reference no.(ii)]**




### 2. Locality->

Big Data provides data locality. We don't need to move to the processing units like the earlier traditional methods. In this we only need to move processing units to the facts within Hadoop framework. Facts grew massive with time so it handles them quite easily.

# CHAPTER 5

# RESULT AND PERFORMANCE ANALYSIS->

In first step we need to gather and generate the information through various Random Generator programs etc. In second step the data is preprocessed converting unorganized unstructured data into the structured form. Aggregation and filteration occurs here which are very much important to eliminate the unwanted data. Data is stored in the HDFS and we have used the Map Reduce Algorithm where the data is mapped to the data sets and reduced to the smaller size. We have used hadoop streaming that generate streams of
Data and make distributed data to set on your database. And finally we have analysed the attendance of different students studying in different colleges with the help of computing server which generates the output report required .

Query for inserting student data using HQL:

```
INSERT INTO project(name,attendace,roll_ no,class,email,phone) VALUES (
'${hiveconf:name}',${hiveconf:attendance},${hiveconf:roll},'${hiveconf:class}','${hiveconf:
email}',${hiveconf:phone})
```

Query for searching student record using HQL:

```
SELECT * FROM project WHERE class='${hiveconf:class}';
```

Query for searching the entire record using HQL:

```
select * from project;
```

From the results we see that we have easily maintained and managed the records for attendance using Big Data Technologies.

Big Data Applications (BDA) are a manner to system part of such big quantities of data by means of platforms, gear and mechanisms for parallel and dispensed processing. ISO subcommittee 32 mentions that BD Analytics has grow to be a main riding utility for information warehousing, with using Map Reduce outside and inside of database management systems, and the usage of self-service statistics marts [2]. Map Reduce is one of the programming fashions used to broaden BDA, which turned into advanced with the aid of Google for processing and generating massive datasets.

Unstructured ,semistructured, and polystructured are all a term for the records that doesn't in shape well in to the relational model. The records is like JSON,XML,RDF or different sources of data with a schema which can

range from file to record. Hive is extraordinarily powerful for dealing with data that doesn't quite suit into complex changes that is probably otherwise essential to address .This type of information in conventional

a relational gadget. Hive can also gracefully cope with the statistics that do not strictly comply with a table's schema for e.G If a few columns are lacking from the specific facts hive can address document via treating lacking columns as nulls.

Hadoop is the Apache Software Foundation's top stage project, and encompasses the various Hadoop sub tasks. The Hadoop project affords and helps the improvement of open supply

software program that components a framework for the improvement of notably scalable allotted computing packages designed to address processing info, leaving builders loose to consciousness on utility good judgment [12]. Hadoop is split into several sub tasks that fall under the umbrella of infrastructures for dispensed computing. One of those sub tasks is Map Reduce, that is a programming model with an associated implementation, both developed with the aid of Google for processing and generating massive datasets.

Performance of this project will be quite beneficial in the field of education. With the assist of large information, custom designed programs for every man or woman student can be created. Even if colleges and universities have lakhs of students, custom designed packages may be created for each of those college students.

There are a number of categories of servers, which includes print servers, record servers, community servers and database servers.

**[Reference no(v)]**

In concept, every time computers proportion sources with purchaser machines they're taken into consideration.

One crucial element of massive facts is the method which handles huge data, particularly referred to as massive data analytics manner. Studying this process and measuring it for better performance will have massive advantages for those who are inclined to reap the price of huge records. The focus on techniques is more commonplace in different disciplines inclusive of enterprise and software development but in big statistics analytic, it's miles quite new. The purpose of this paper is to research massive information analytics (BDA) procedure and become aware of suitable overall performance measures for it. Existing BDA techniques

and performance dimension strategies are studied and the outcomes are supplied on this paper

Hive has native aid for the set of records systems that normally might either not exist in a relational database or could require definition of custom sorts. They are all the usual gamers integer, strings, floats and so on but thrilling one are the more extraordinary maps, arrays and structs. Maps and arrays works in a fairly intuitive manner just like how they paintings in many scripting languages. If the consumer mentions array is empty, Hives will just go back NULL for that document. Hive can take the queries, make dispensed data to set for your database.

Finally an output report is generated which displays the attendance record of any student studying in any university. We can see the past record very easily. We can see the past record very easily. Since the data is kept confidential for the universities so we use Random Generator for generating large amounts of data
and solve the problem quite easily.

Unstructured , semistructured, and polystructured are all a term for the records that doesn't in shape well in to the relational model. The records is like JSON,XML,RDF or different sources of data with a schema which can

range from file to record. Hive is extraordinarily powerful for dealing with data that doesn't quite suit into complex changes that is probably otherwise essential to address .This type of information in conventional

a relational gadget. Hive can also gracefully cope with the statistics that do not strictly comply with a table’s schema for e.G If a few columns are lacking from the specific facts hive can address document via treating lacking columns as nulls.

It is  pretty easy to look that there is a good bit of acomplexity to this statistics systems.Since aJSON can contain nested information structures it becomes very hard to force JSON information in to a general relational schema.Processing JSON information in a relational database might in all likelihood require a significant transformation making the task an awful lot extra cumbersome.

# APPLICATIONS AND SCOPE

## Entertainment->

Big statistics inside the enjoyment enterprise holds the capacity to convert the way media houses are operating currently. As massive data analytics provides insights, government within the enjoyment enterprise can effectively have the availability of statistics which could assist them in making green decisions .Big records analytics is a technology that is experiencing first rate growth and recognition. As currently there are various assets of large information, corporations focus on gathering all the records viable. Post series of statistics, corporations try to leverage analytics to achieve intuitive and actionable insights. Big data in the amusement enterprise holds the capability to convert the way media houses are operating currently. As big information analytics offers insights, authorities in the amusement enterprise can efficiently have the supply of records that can help them in making green selections. The insights encompass patron feelings to a particular show or track or traits of how commercials are displayed.

Unstructured ,semistructured, and polystructured are all a term for the records that doesn't in shape well in to the relational model. The records is like JSON,XML,RDF or different sources of data with a schema which can fit easily.

## Manufacturing->

Advanced analytics refers back to the utility of information and other mathematical tools to enterprise records so one can verify and improve practices (show off). In manufacturing, operations managers can use superior analytics to take a deep dive into historical procedure information, discover styles and relationships amongst discrete

technique steps and inputs, after which optimize the elements that show to have the finest impact on yield. Many international manufacturers in quite a number industries and geographies now have an abundance of actual-time keep-floor facts and the capability to conduct such state-of-the-art statistical tests. They are taking previously isolated records sets, aggregating them, and reading them to show crucial insights.

One manufacturer is the use of big statistics to reduce danger in shipping of uncooked materials, no matter what takes place in the deliver chain.

## Security->

This proposed system maintains and organizes the data of attendance for various universities. Therewill be 24 hours of updation and checking of attendance .Your encryption equipment want to relaxed records in-transit and at-rest, and they want to do it throughout big statistics volumes. Encryption additionally wishes to perform on many exceptional sorts of facts, each person- and gadget-generated. Encryption equipment additionally need to paintings with extraordinary analytics toolsets and their output records, and on common large facts storage formats including relational database management structures (RDBMS), non-relational databases like NoSQL, and specialised filesystems which includes Hadoop Distributed File System (HDFS)

## Transportation->

Traffic is despised by everybody. Big information involves rescue, being used to scale down traffic and decorate visitors control by means of helping prediction and management of congestion. Using a combination of actual time records, ancient developments, and clever algorithms, large records is translating vehicle speeds, climate conditions, network events, and resources of acceleration and deceleration for road operators. Sensors built on delivery networks and fleet motors permit corporations to accumulate records streams from neighborhood delivery authorities. Besides, big facts additionally gives clever records to traffic authorities to control traffic.

## Insurance->

Extracting cost from uncertain data

“Big statistics” – which admittedly approach many things to many human beings – is now not limited to the world of era. Today, it is a business imperative. In addition to supplying solutions to coverage companies’ lengthy-standing commercial enterprise demanding situations, huge information solutions provide the electricity to convert procedures, corporations and entire industries.

Big data is especially promising and differentiating for coverage corporations. With no bodily products to manufacture, records is arguably certainly one of their maximum important property. Financial, actuarial, claims, danger, patron, manufacturer/wholesaler and plenty of other types of information form the basis for truly each selection an insurer makes. And while the enterprise has made progress in taking pictures and studying an awful lot of the dependent information related to their merchandise and policyholders, there's value in unstructured and semi-structured data that stays untapped.

Map Reduce is well known as the heart of Hadoop Ecosystem as each task of the Big Data is computed using this technique .Map Reduce distributes the processing of your data on your

cluster. It divides the data in to partitions that are mapped(transformations) and reduced(aggregated by mapper an reducer functions). Reduce is the technique that can be used to provide huge scalability across hundreds or thousands of clusters in Hadoop cluster.

## Energy and utilities->

Energy and application companies practice smart technology to their landscape, which includes sensors, cloud computing technologies, wireless, electricity planning, and community communication. These produce huge statistics sets on a continuous foundation which gets accumulated over a time period. For instance, a software enterprise, using clever meters and power, can collect round three petabytes of information every 15 minutes for a 12 months for about 1,000,000 households.

If we start increasing shrewd gadgets like sensors and thermostats, we are talking of huge-extent facts units being generated throughout power technology to transmission to distribution after which to customers via substations. Businesses across the utility industry are facing a variety of demanding situations to draw insights out of this valuable data and conduct strength planning.

## SCOPE->

These days large data has end up the buzzword in IT industry corporations .The need of reading and processing of information has grown lots .We have fulfilled the analysis of the attendance of various students in specific schools. Further analysis can be performed to photographs and all kinds of multimedia documents based totally on index guide. The result of textual content mining and facts analysis might help in signifying associated pages primarily based on extraordinary types of records. So that industries make the statistics effortlessly available humans who's the usage of and retrieving such form of data

We have used hadoop streaming that generate streams of Data and make distributed data to set on your database. And finally we have analysed the attendance of different students studying in different colleges with the help of computing server which generates the output report required .

With this gadget, you may construct a console to show the facts related to attendance for diverse universities in real time .But the problem arises while coping with large facts of unstructured type. It is solved by means of Hadoop and its packages. Using traditional methods recuperating abilities and processing time are very difficult but it's far made easy with the aid of the use of Hadoop and its various technologies.

# CHAPTER 6
# CONCLUSIONS

## 6.1 Conclusions

Outdated Data Warehouses do not have the functionality to keep up with hastily increasing information for attendance records for big range of universities. With this gadget, you may construct a console to show the facts related to attendance for diverse universities in real time .But the problem arises while coping with large facts of unstructured type. It is solved by means of Hadoop and its packages. Using traditional methods recuperating abilities and processing time are very difficult but it's far made easy with the aid of the use of Hadoop and its various technologies.

Data is stored in the HDFS and we have used the Map Reduce Algorithm where the data is mapped to the data sets and reduced to the smaller size. We have used hadoop streaming that generate streams of Data and make distributed data to set on your database. And finally we have analysed the attendance of different students studying in different colleges with the help of computing server which generates the output report required .

Map Reduce is well known as the heart of Hadoop Ecosystem as each task of the Big Data is computed using this technique .Map Reduce distributes the processing of your data on your cluster. It divides the data in to partitions that are mapped(transformations) and reduced (aggregated by mapper an reducer functions). Reduce is the technique that can be used to provide huge scalability across hundreds or thousands of clusters in Hadoop cluster.

.

The essential characteristics and demanding situations of Big Data are Volume, Velocity and Variety. These are referred to as 3V's of Big Data which might be used to characterise specific elements of Big Data.

## 6.2 Future Scope

These days large data has end up the buzzword in IT industry corporations .The need of reading and processing of information has grown lots .We have fulfilled the analysis of the attendance of various students in specific schools. Further analysis can be performed to photographs and all kinds of multimedia documents based totally on index guide. The result of textual content mining and facts analysis might help in signifying associated pages primarily based on extraordinary types of records. So that industries make the statistics effortlessly available humans who's the usage of and retrieving such form of data. Hadoop is the twiglet of Big Data. To be professional in Hadoop is a deciding aspect in getting a springboard for your career or getting left at the back of. If you are a brisker there is a massive scope if you are skilled in Hadoop. Amongst the open source framework, there's almost no other opportunity that may deal with petabytes of records as Hadoop can. In 2015 changed into it turned into anticipated that Indian Big Data Hadoop industry will grow five folds within the analytics centre.

With this gadget, you may construct a console to show the facts related to attendance for diverse universities in real time .But the problem arises while coping with large facts of unstructured type. It is solved by means of Hadoop and its packages. Using traditional methods recuperating abilities and processing time are very difficult but it's far made easy with the aid of the use of Hadoop and its various technologies.

The pool of educated professionals in facts analytics with Hadoop expertise is low compared to the modern and predicted call for. Hadoop market in India isn't always a frizz so as to dilute with time, at the contrary, it is extra special in call for, getting to know the talent guarantees better income and better job prospects for both skilled and fresher's alike. Currently every principal IT agency like, Facebook, Jabong, Snapdeal, Amazon and manyothers.., are the usage of Hadoop to convert zettabytes of data created through those portals hence if you are educated in Hadoop you may be the apple of any developer in India.

## References


i).Big Data Analytics: Research Paper,Scopus (10 January,2015)

ii) IEEE:Author,IEEESpectrum(15 May 2017)

iii) Big Data Course:Udemy,Mr.Frank(8 Feb 2018)

iv) Sqoop Tutorial:Edureka,Author(January 2016)

v) Hive Tutorials:Tutorial Point team(January)

vi) Big Data Black Book: Dreamtechpress(1 Jan 2016)