

“TOURIST DATA ANALYSIS”

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

Computer Science and Engineering/Information Technology

By

Akriti Sood (161288)

Manika Kansal (161318)

Under the supervision of

Dr. Suman Saha

to

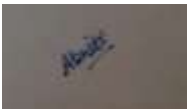


Department of Computer Science & Engineering and Information Technology

Candidate's Declaration

I hereby declare that the work presented in this report entitled “**Tourist Data Analysis**” in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology, Waknaghat is an authentic record of my own work carried out over a period from January 2020 to May 2020 under the supervision of **Dr. Suman Saha (Assistant Professor Sr. Grade Dept. of CSE)**.

The matter embodied in the report has not been submitted for the award of any other degree or diploma.



Akriti Sood (161288)



Manika Kansal (161318)

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)

Dr. Suman Saha
(Assistant Professor Sr. Grade)

. Dept. of CSE

Dated: 28/05/2020

Acknowledgement

It is our privilege to express our sincerest regards to our project supervisor **Dr. Suman Saha (Assistant Professor Sr. Grade)**, for their valuable inputs, able guidance, encouragement, wholehearted cooperation and direction throughout the duration of our project.

We deeply express our sincere thanks to our Head of Department for encouraging and allowing us to present the project on the topic “**Tourist Data Analysis**” at our department premises for the partial fulfillment of the requirements leading to the award of B-Tech degree.

At the end we would like to express our sincere thanks to all my friends and others who helped me directly or indirectly during this project work.

Date: 28-05-2020

Akriti Sood (161288)

Manika Kansal(161318)

TABLE OF CONTENTS

CONTENT	PAGES
DECLARATION	i
ACKNOWLEDGEMENT	ii
LIST OF FIGURES	vi
LIST OF SNIPPETS	vii
LIST OF TABLES	viii
ABSTRACT	ix
CHAPTER 1: INTRODUCTION	1-13
1.1 Introduction	
1.1.1. Java	1
1.1.2. MySQL	1
1.1.3. Netbeans	1
1.1.4. Swing	1
1.1.5. JFrameClass	1-5
1.1.6. Big Data	5-8
1.1.7. Use Of Big Data Analytics	9
1.2 Problem Statement	10
1.3 Objective	10
1.4 Methodology	11-13
1.4.1 Using MySQL	11-12
1.4.2 Hadoop Platform	13
CHAPTER 2: LITERATURE SURVEY	14-18
2.1 Hive a warehouse solution over a map reduce framework	14
2.2 Commercial product analysis using Hadoop map reduce	14
2.3 Hive- a petabyte scaledata warehouse using Hadoop	15
2.4 Review paper on mapreduce and Hadoop	15
2.5 Paper on Swings	16
2.6 Big Data : A Turnkey Solution	16

2.7 Study based on MySQL Storage Engine	16
2.8 Analysis of the perception of accommodation Consumers on the use of Online Travel Agencies	17
2.9 Scalability study of Hadoop map Reduce and Hive in Big Data Analytics	17
2.10 An Overview of the Map reduce/HBase /Hadoop framework And its current application in Biofield	18
2.11 Storage and processing speed for knowledge from enhanced Cloud framework	18
 CHAPTER 3: SYSTEM DEVELOPMENT	 19-43
3.1 Designing	19
3.2 Experimental Setup	
	3.2.1 For Java 19
3.2.2 For Hive And Pig Platform	20- 21
3.3 Analytical Model	
3.3.1 Traditonal Model-“Schema on write”	21
3.3.2 Big Data Model-“Schema on read”	22
3.3.3 Data Modelling in the Big Data Ecosystem	23
3.4 Analysis	
3.4.1 Data Analysis Using Hive	24-25
3.4.2 Design Analysis	26-27
3.5 Algorithms	
3.5.1 Using Java Applets	
3.5.1.1 To Import A CSV File Into A Database Using MySQL Server	28
3.5.1.2 Creating JDBC Application	28-29
3.5.2 MapReduce Algorithm	29-32
3.6 Switching To BigData Over SQL	32
3.6.1 Big Data Challenges	33
3.7 Test Plan	
3.6.2 DataSet	34-36

CHAPTER 4 : RESULT ANDPERFORMANCE ANALYSIS	37-52
4.1 Using SQL and Java Applets	37-41
4.2 Using Hive(HiveQL)	42-46
4.3 Using Pig	47-51
4.4 Comparison between existing and adopted solution	51-52
CHAPTER 5: CONCLUSION	53
5.1 Conclusion	
5.2 Future work	
References	54

List of Figures

Figure No.	Figure Name	Page No.
1	5Vs of BIG DATA	6
2	Values in 5Vs of Big Data	6
3	Big Data Uses	7
4	Summary Of Big Data	7
5	Significant Of Big Data	8
6	Data Models for Converting Data Into Tabular Format	22
7	Process for uploading data and creating table using the HDFS	23

8	Data Modeling for handling structured, unstructured and semi structured data	23
9	Flowchart of data analysis using Hive	24
10	Steps for Mapper & Reducer class	30
11	Mapper Stage	31
12	Reducer stage	32

List of Snippets

Figure No.	Figure Name	Page No.
1	Code Snippets #1	2
2	Code Snippets #2	2
3	Code Snippets #3	2
4	Code Snippets #4	3
5	Code Snippets #5	3
6	Code Snippets #6	4
7	Code Snippets #7	4
8	Snippets #8 for Use Case Diagram	26

9	Snippets #9 for Flow Chart	27
10	Code Snippets #10	28
11	Code Snippets #11	29

List Of Tables

Table No.	Name	Page No.
1	Dataset columns and its specifications	34
2	Sample record of Tourist Dataset	35

Abstract

Big data is the large amount of data that is generated every second. Such large data becomes very difficult to process using manual checking and old methods. Now, we have a number of tools and techniques to handle such data. With continually increasing customer use on online platforms, it is necessarily required to know customer's likes and interests. It is really necessary as through this, the preferences of the customers can be well known. The best way to handle this problem is by applying Big Data Analytics.

Tourism is considered to be the most favourite pass time . People travel with friends to have a good time or for business purposes, usually for a limited period of time. Tourism is usually associated with domestic or international travel. There are several travel organizations are available on the internet. Depending on their personal interest, people or tourists choose their cabs service with the favourable package. Travel companies rely on the enthusiasm associated with tourism to ensure that they increase their particular market value and provide massive package deals. Given the large amount of information available on the travel package, it is important to meet a tourist's personal needs and preferences to offer more attractive packages. In order to make their Travel Package more effective recommender system is becoming very popular and it attracts people, as it allows them in a short time to choose the best package cabs service.

Hobbies: Customers still have different opinions and understandings to choose a bundle due to variations in gender, profession, hobbies and knowledge. Customers have a lot of interest and taste so they can choose any kit based on their interest. Product descriptions are provided to the user by selecting the location where the products are displayed based on travel location and duration.

This is a project report on “**Tourist Data Analysis**”. During the exploration of this project, we develop new ideas and functionalities while using the Big Data technologies. So, this project is for recommending best cab service for a given location to a new customer based on rating of old customers

The project report covers the implementation of the project on Java and Big Data technologies and the reason why there is a shift from Java/SQL to Big Data using Hive and Pig over Hadoop with a concluding result at the end.

CHAPTER 1

INTRODUCTION

1.1 Introduction

1.1.1 JAVA:

Java is object oriented programming language, that was developed at Sun Microsystems by James Gosling. It was later on acquired by Oracle. Java was developed to counter some of the problems that incurred in C++ like portability, multithreading etc. The latest version of java that is released is Java 14 ,released on March 2020.

1.1.2 MySQL:

MySQL is the most famous open source relational database management System (RDBMS).It was initially released on 23 May,1995. Earlier it was owned by a Swedish company MySQL AB. But later on it was acquired by Sun Microsystems. And by 2010 it was taken by Oracle.

Right now, MySQL is used by many popular sites such as Facebook, twitter ,YouTube etc.

1.1.3 NetBeans:

Netbeans is an application of integrated development environment (IDE) for java . Basically it provides all the modules integrated at one place in order to build any java application. It has some features like: NetBeans Visual Library, User Interface management, Integrated development tools etc.

1.1.4 Swing:

Java Swing is built on the top of AWT(Abstract Window Toolkit).

Java Swing have lightweight components as compared to Abstract Window Toolkit.It is entirely written in java. Java Swing components are platform independent.Swing follow MVC(Model View Controller).

1.1.5 JFrame Class:

The JFrame Class need a container for your application whenever you create a graphical user interface with Java Swing functionality. This container is called a JFrame in Swing's case. A JFrame is required for all GUI applications. Some Applets actually even use a JFrame. What's the reason? Without a foundation, you can not build a house. The same is true in Java: you won't have a GUI application without a container to put all other elements in. In other words, for all other graphical components, the JFrame is required as the foundation or base container.

It is possible to run Java Swing applications on any system that supports Java. These are lightweight applications. This means not taking up a lot of space or using a lot of system resources. JFrame is a Java class with its own methods and builders. Methods are functions that affect the JFrame, such as size or visibility setting. When the instance is created, constructors are running:

one constructor can create a blank JFrame, while another can create it with a default name.

```
1 import javax.swing.*;
```

Code Snippet #1

Creating a JFrame

When you create a new JFrame, you actually create a JFrame class instance. You can create a title or an empty one. If you pass a string to the constructor, the following title will be created:

```
1 JFrame f = new JFrame();  
2 // Or overload the constructor and give it a title:  
3 JFrame f2 = new JFrame("The Twilight Zone");
```

Code Snippet #2

But, if you run this script, you won't see an application window! The explanation for this is that there is no concept of size or position. You have to give it a size and make it visible to the users.

Size and Location

After creating the JFrame, use the setSize and setLocation methods to set its size and location as follows:

```
1 //add the frame  
2 JFrame f = new JFrame("The Twilight Zone");  
3 //set size: width, height (in pixels)  
4 f.setSize(300, 325);  
5 //set the location (x,y)  
6 f.setLocation(150, 50);
```

Code Snippet #3

Setting the window size and location

Setting the Window Size and Location	
Method	Purpose
<code>void pack()</code> <i>(in Window)</i>	Size the window so that all its contents are at or above their preferred sizes.
<code>void setSize(int, int)</code> <code>void setSize(Dimension)</code> <code>Dimension getSize()</code> <i>(in Component)</i>	Set or get the total size of the window. The integer arguments to <code>setSize</code> specify the width and height, respectively.
<code>void setBounds(int, int, int, int)</code> <code>void setBounds(Rectangle)</code> <code>Rectangle getBounds()</code> <i>(in Component)</i>	Set or get the size and position of the window. For the integer version of <code>setBounds</code> , the window upper left corner is at the <i>x</i> , <i>y</i> location specified by the first two arguments, and has the width and height specified by the last two arguments.
<code>void setLocation(int, int)</code> <code>Point getLocation()</code> <i>(in Component)</i>	Set or get the location of the upper left corner of the window. The parameters are the <i>x</i> and <i>y</i> values, respectively.
<code>void</code> <code>setLocationRelativeTo(Component)</code> <i>(in Window)</i>	Position the window so that it is centered over the specified component. If the argument is <code>null</code> , the window is centered onscreen. To properly center the window, you should invoke this method after the window size has been set.

Code Snippet #4

Attaching image in GUI

```
//Set the frame icon to an image loaded from a file.  
frame.setIconImage(new ImageIcon(imgURL).getImage());
```

Code Snippet #5

Example for JFrame

```
private void prepareGUI(){
    mainFrame = new JFrame("Java Swing Examples");
    mainFrame.setSize(400,400);
    mainFrame.setLayout(new GridLayout(3, 1));

    mainFrame.addWindowListener(new WindowAdapter() {
        public void windowClosing(WindowEvent windowEvent){
            System.exit(0);
        }
    });
    headerLabel = new JLabel("", JLabel.CENTER);
    statusLabel = new JLabel("",JLabel.CENTER);
    statusLabel.setSize(350,100);
    msgLabel = new JLabel("Welcome to TutorialsPoint SWING Tutorial.", JLabel.CENTER);

    controlPanel = new JPanel();
    controlPanel.setLayout(new FlowLayout());

    mainFrame.add(headerLabel);
    mainFrame.add(controlPanel);
    mainFrame.add(statusLabel);
    mainFrame.setVisible(true);
}
```

Code Snippet #6

Output:



Code Snippet #7

Buttons

Generic buttons are designed from the class JButton (in the package javax.swing). Like other controls, hundreds of methods are inherited. The standard builder defines a string to be used as a button tag (or we can later call setLabel with any string). We can also call inherited methods such as setFont, setBackground, and setForeground, although default values are used for most standard buttons. We can also call the setEnabled method (with a boolean parameter) to tell Java if a button can be pressed (if not, the label appears as a faded color).

```
JButton b = new JButton("Cold");  
b.setFont(new Font("Monospaced",Font.BOLD,12));  
b.setBackground(Color.yellow);  
b.setForeground(Color.blue);
```

1.1.6 Big Data

Analytics is a multidimensional and detailed field. With today's technology, your data can be analyzed and immediately answered. Big data analysis analyses vast quantities of data to reveal hidden trends, comparisons, and other insights. One can make informed decisions with the use of Big Data Analytics without relying blindly on guesses.

5 Vs of Big Data :

- Volume: This meant huge voluminous information; all the requests is of terabytes and petabytes.
- Velocity: This means the high speed information which are we creating on our process.
- Variety: This meant to the enormous variant in the huge processed information.
- Veracity: This meant alludes to vulnerabilities and improper value in huge information, for example, missing, copy and fragmented sections.

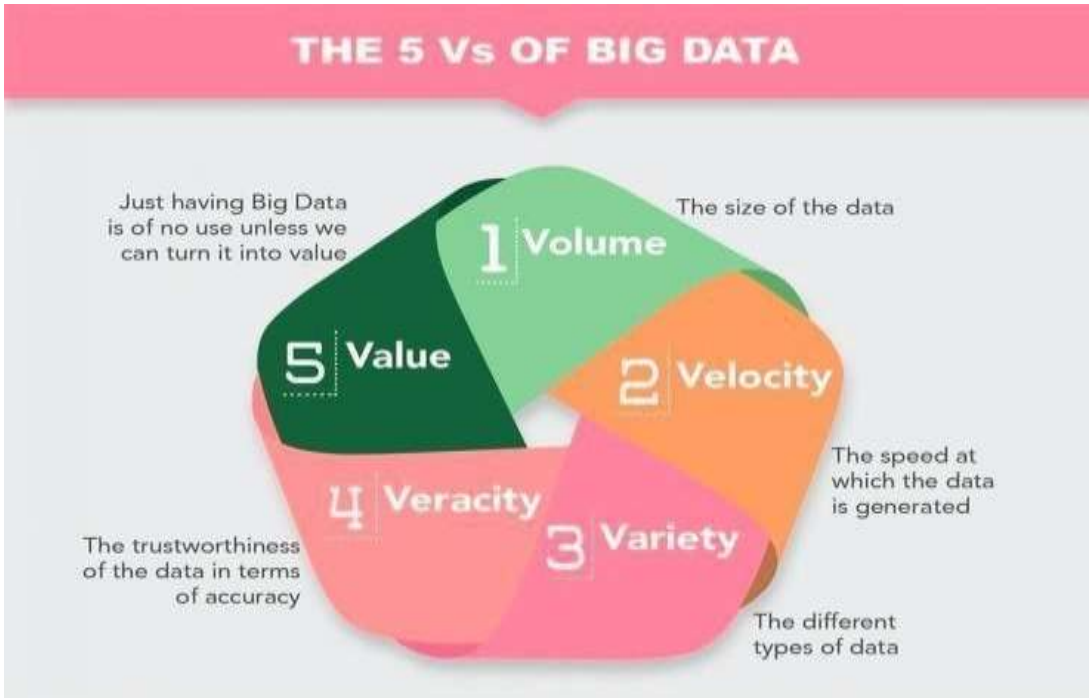


Fig. 1: 5 V's Of Big Data

- Value: This trademark alludes to the valuable information contained in huge information.

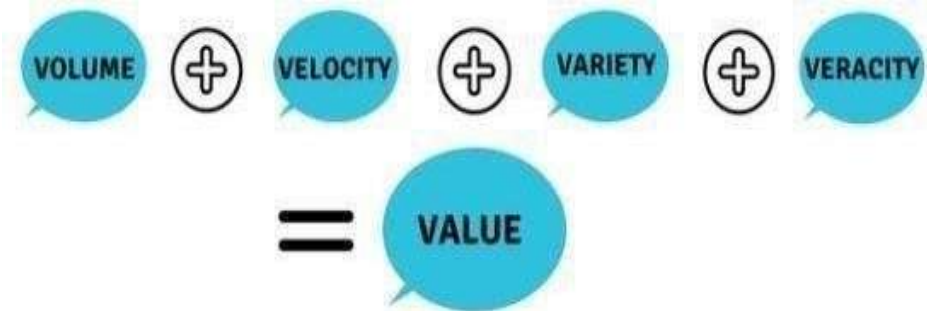


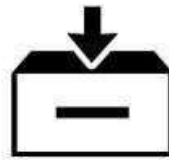
Fig. 2 Values In 5V's Of Big Data

BIG DATA

USED TO



Process



Storage



Analyze

Fig. 3 Big Data Uses

Big data analytics actually help companies to get accurate results which can be used in the favour of their company as well. Thus this leads to more productivity. Some of the important qualities of Big Data are:



Fig. 4 Summary Of Big Data



EFFICIENT COST REDUCTION

Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data - plus they can identify more efficient ways of doing business.



FASTER & BETTER DECISION MAKING

The speed of Hadoop and in-memory analytics, combined with the ability to analyze new sources of data, businesses are able to analyze information immediately and make decisions based on what they've learned.



NEW PRODUCTS & SERVICES

The ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want. Davenport points out that with big data analytics, more companies are creating new products to meet customers' needs.

Fig. 5 Significance of Big Data

1.1.7 Big Data Analytics uses:

Banking

Since Banking is a very big sector ,so it must contain very huge information which can easily be dealt by Big data. By analyzing the data ,it can protect people from frauds and risks.

Health care

The huge amount of information that a health care center can generate can easily be analyzed by big data analytics. Like the records of patients, information about prescriptions, records of medicines etc.

Manufacturing

Big data analysis can also help in manufacturing sector as well.As with the help of Big data analysis We can make more informed choices,can make our business strategies according to it and can remain ahead of our competition in the market.

Government

Nowadays governments are also using big data analysis for their use.By the analysis of the data they can easily decide what can be their next election strategy,how they can reduce crime etc.

Entertainment and Media

Big data even has a huge effect on TV and entertainment industry .As with the help of big data analysis one can show the right content at the right time to the right audiences. Simply they show different recommendations based on the shows, movies you have previously watched.Even the companies like Netflix ,Youtube are currently using this strategy to gain larger profit and provide more engaging experience for the user.

1.2 Problem Statement

There are several travel agencies available for a travel system. But currently there is no efficient recommendation system available , which actually help us to choose a particular travel agency based on people views. To overcome this problem, we are coming up with Travel Package Recommendation System using tourist data where you can select the best package.

There are two objectives for the development of this project.

Firstly, the manual dataset is observed and , with the requirement, various attributes are taken and shown together. The above dataset is executed on two different platform i.e. Java applets & on Hadoop. Also the data is, shown in tabular format. **Secondly**, we have analyze what sort of cabs services were taken by the customer in the preferred customer location and then we recommend them that particular cabs service according to their requirement .

Finally, we have examined the required problem through Big Data using HQL and PIG due to the constraint ,in the SQL. The reason of the l switch from SQL to Big Data platform is parallel and distributed processing in hadoop platform.

1.3 Objective

The objective is:

- To make the project using Java applets.
 - Displaying the GUI of cabs agencies recommendation.
 - Analyzing the reviews of a particular cabs service with the given location.
- To study the constraints of SQL & swittching to Big Data technology using Hadoop platform to get better results using Hive and Pig as both use map reduce framework which is done on parallel processing and thus comparing the Hadoop and sql platforms.

1.4 METHODOLOGY:

This part of the report ,covers the implementation of the above said problem using Java Swing and MySQL.

1.4.1 Using MySQL

MySQL is an open source database which is free to use .It is quite stable ,reliable and consists of some advanced features as: Security of data ,On demand Scalability and Complete Workflow Control etc.

The excel sheets comprises of dataset which includes the following attributes:

Person's Name(Name), Travel service used(Travel Agency), Experience With the service(Rating),Package opted(Package),Location of the places(Location).

The project is implemented via the Graphical User Interface which is made using java Swing and java abstract Window Toolkit. The GUI consists of:

Main JFrame form which asks the customer about the login details such as Name, Age , email , gender , package to be opted , no. of persons , no. of days and location.

Depending on the data provided, customer will be suggested best cabs service he/she should choose in order to meet their demands optimally.

Steps for implementing the project :

- To Select the dataset
- Then it is stored in Excel Sheets (csv format)
- Data is retrieved in database (Sql)
- GUI is made using Java Swing
- Following are the various queries that a user use to retrieve information:

Table name in database login values

List the various names of consumers who rated the service:

```
select name from login values
```

Listing all names of travel agencies:

```
select travelagencies from login values
```

List the variouspackage available: select

```
distinct(Package) from login values
```

List the various ratings available:

```
select distinct(rating)from login values
```

When the User enters the required details such as package information , location he wants to go,duration , no. of persons etc.,he will be provided with the choice of cabs service that suits him the best and for that result comes is done at the backend using sql .

1.4.2 Hadoop platform

1.4.2.1 Big Data and Hadoop ecosystem

Hadoop capabilities

Apache Hadoop is a well-known Big Data software .It was developed with all the modules integrated at one place.Actually it is allowing processing which is distributed of larger datasets across fragments of clusters of computers. Storage part which is the core of the Hadoop is HDFS(Hadoop Distributed File System) whereas its processing model is Map Reduce .It actually split the files to larger blocks which then can be distributed over a node in the cluster Then the data is processed in parallel. Therefore in this the processing of data takes place much faster as compared to other architectures as here there is parallel computing. Eg: it can process terabytes of data in seconds whereas earlier it used to take hours and hours to analyze that data.

In addition, with several open source modules, the Hadoop community has helped to expand the ecosystem. In parallel, IT vendors have unique Hadoop distribution company hardening features.

1.4.2.2 Hadoop Distributed File System (HDFS)

HDFS is a storing unit in Hadoop. It can actually supports structured, semi-structured as well as unstructured data. It can handle huge volume of data(it can be even Zettabytes).HDFS has a master slave architecture. It actually distribute larger data into cluster .Each cluster have one master ie Namenode and several slave nodes ie Datanodes. Namenodes actually give a particular Task to a particular datanode. But since its based on fault tolerance so it keeps a copy in case of any failure. Usually it keeps 3 copies.

1.4.2.3 Data Querying Layer: Pig and Hive

Pig and hive are data querying languages in Hadoop .

Hive is a datawarehouse software built on the high level of apache Hadoop for providing data queries and analysis. It is written in Java.Hive is more like SQL since it has somewhat similar syntax as compared to SQL. Also it is known as HIVEQL.

Whereas pig is also one of ths data querying language built on the top of the apache Hadoop.The language for this platform is Pig Latin.It is basically a hhigh level programming language that is used to analyze larger datasets.It was developed at yahoo.

CHAPTER 2

LITERATURE SURVEY

21 “Hive , Warehouse Solution Over A Mapreduce Framework [1]

Paper actually talks about hive and its abilities that can be used as data warehouse.

In this paper how hive functions in the field of warehousing is appeared:

1. Usefulness — It indicates that HiveQL receives developed I with the aid of potential of the StatusMeme software .
2. 2 Tuning — It show off the inquiry design watcher o which demonstrates how HiveQL questions are converted into bodily plans of guide lessen employments.
3. UI — It demonstrates the graphical UI which permits customers to l investigate a Hive database, creator HiveQL inquiries, and display screen inquiry execution.
4. Versatility — It delineates the adaptability & of the framework with the aid of& increasing the sizes of the data data and the multifaceted mnature of the questions.

22 “Commercial Product Analysis Using Hadoop MapReduce” [2]

This paper discusses how an organisation can discover genuine lpen doorways in consolidating disconnected and on line informationl to give astuteness on how combining ldisconnected and on-line facts can be useful. Organizations makes use of inspiration calculations which have the above advantages. Proposal calculations are great perceived for their , utilization on on-line business Web sites. Here they utilize client's pursuits as a contribution to create an index of prescribed things.

These calculations are partitioned into two types :

The preliminary ones are referred lto as content based totally filtering. Content based filtering can likewise be known as as cognitive filtering, which prescribes items based on an examination between the lsubstance of the objects and a purchaser profile.

Also, tlhe 2nd one is collaborative filtering. It depends uponl no longer clearly the traits of the items but alternatively how men and women i.e distinct clientsf react to comparable articles. Associations need to get every one of the factss traits, disconnected and on the web, into a solitary database, which would be

additionally refined via cutting area examination strategies, and utilize the consolidated information for accuracy focusing on specific technologies.

23 “Hive – Petabyte Scale Warehouse Data using Hadoop” [3]

Hadoop is an open-source application framework that is used in multinational companies such as Yahoo, Twitter, to store and process significant data indexes on warehouse equipment to a large extent. Notwithstanding this, the guide diminishes the template of programming and encourages programmers to write personalized projects that are difficult to maintain and recycle. Hive, a warehousing information arrangement based on Hadoop, is used in this paper. HiveQL also enables customers to connect guide content reduction to inquiries.. The Hive stock room on Facebook contains a large number of tables and stores with more than 700 TB of information and is widely used by over 200 customers per month both announcing and specially appointing examinations.

24 “Review Paper on Map Reduce and Hadoop” [4]

With limitless Data is an data whose scale, better than common assortment, and multifaceted nature require new designing, strategies, figurings, and examination to direct it and focus regard and disguised gaining from it.

This paper actually discusses that why Hadoop is better. Following focuses center around the benefits of hadoop:

Adaptability, exceeds expectations at dealing with statistics of complicated nature and its open-source nature make it a great deal nicely known. In this paper the format of hadoop and mapreduce is clarified. MapReduce has been shown as a free stage as the advantage layer perfect for extraordinary want with the aid of cloud providers.

It in addition allows clients to respect the facts dealing with and exploring.

25 PAPER ON SWINGS [5]

As the components look and feel is decided by platform and not by java.

So this paper talks about swing. Swing was basically developed after Abstract Window Toolkit(AWT) because of some limitation of AWT. As swing components are light weight as compared to AWT. Swing supports platform independence and MVC. It supports platform

independence because of the fact that swing is written in Java. Also it provide some of the extra features as compared to AWT ie scroll panels, trees, lists etc

26 BIG DATA :A TURNKEY SOLUTION [6]

Big data is an ocean of datasets.It has changed various spheres of our life .It has helped in analyzing the data easily and effectively which in turn has helped us making decisions effectively.

Big data has contributed in various fields such as:

Medicines

Business sites

Science and Engineering

Government

This paper talks about,how big data has revolutioned our lives from science to government and from enterprises to customers.Its a turnkey solution because it has changed the scenario of business.as we these the e-commerce sites are on tremendous increase.With lots of advantages ,some challenges still remain like security and manymore.

27 STUDY BASED ON MY SOL STORING MACHINE [7]

MySQL provides different types of storage engines. Various engines uses various datastorage mechanism, techniques used for indexing ,lock level so as to provide distinct features and capabalities.

This paper basically talks about various storage engines of MySQL.

MY ISAM ENGINE:

It is a default storage engine in mysql 5.1 .

Foreign key and transaction are not supported in this. It has higher access speeds.

INNO DB ENGINE:

InnoDB provides MySQL trnsaction-safe tables .

They consists of the capabilities like trans action rollback and crash recovery.

These features increase multipl user.

Thus,this paper total up the capabilities of the two engines and introduces the optimization concepts by interrogating the two main engines.

28 Analysis of the Use of Online Travel Agencies (OTAs) based on the perception of Consumers[8]

In order to develop effective work in tourism companies, it is necessary to know the opinion of consumers about their services. Considering that the travel agency sector is constantly changing, because of the easy access of consumers to information technology, these companies must focus on service quality.

As for hotel managers and OTAs, they should not focus solely on direct sales, but should also be concerned about brand reputation, which is projected in UGC on the Internet, as highlighted throughout the paper.

Regarding traditional travel agencies, the conclusions obtained also suggest that an online presence should be created and maintained in order to survive and recover the competitiveness of the sector, since Internet use by young consumers is increasing. Soon, in the very near future, even older consumers will book accommodation through online platforms.

29 “Scalability Study Of Hive and Hadoop Mapreduce In Big Data Analytics” [9]

First inquiry emerges that how would you cross a contemporary facts framework to Hadoop, when that basis depends on conventional social databases and the Structured Query Language (SQL)?

This is the area Hive appears. Facebook created Hive which relies upon on herbal thoughts of tables, sections and segments, giving an ordinary state inquiry equipment for getting to information fromk their contemporary Hadoop distribution center.

In this paper this is a correlation between Hive versus Mapreduce :

All the work performed is even though j hive however the mapreduce is the one which I surely work inside. An examination was directed on phrase test in which hive performed out the exercise inside 35 seconds and well-known information reduce took around 1 min 10 seconds.

Along these lines this examination paper presumeksvthat hive is unmistakably greater finest to regular mapreduce & outperforms mapreduce execution.

210 “An Overview of the Mapreduce/Hbase /Hadoop framework and its present uses in Biofield” [10]

Hadoop is a product structure which is added on a Linux background to expand splendid records examinations . The Hadoop Distributed File System (HDFS) is a sturdy gadget given with the aid of hadoop and in addition it is a Java-based API that lets in parallel handling on the clusters of the group. Projects use a Map/Reduce , I execution which works as a incredible appropriated processing framework over usual informational indexes - a strategy given through Google.

There are discrete Map and Reduce steps, the place each 1 of the ranges are accomplished in parallel, working one through one on units of key-value sets. Process is parallelized greater than lots of clusters chipping away at terabyte or better measured informational 1 indexes. The Hadoop structure as a result plans t define close to the statistics 1 on which they will work, with "close" which meanss a comparable hub or, at any rate, ss the identical rack. Hub disappointments are moreover taken care of 1 consequently. Notwithstanding Hadoop 1 itself, which is a fantastic dimension Apache venture.

211 “Storage and Processing Speed For Knowledge from Enhanced Cloud Framework”[11]

Cloud is the pool of servers, all of the servers are inter - connecteds through web, The precept factor in cloud is improving data (learning) and operationvthat assortment of information and other factor is security for that information. Basically in todays time extraordinary kinds of capability assortment of facts (Structured, semi-organized and Unstructured information) is managed in the various social operational platform.

So,other trouble is verifiable statistics recovering. These types of troubles are settled with resource of hadoop and Sqoop and flume devices. Sqoop is stack records beginning database to Hadoop (HDFS), and flume stacks statistics from server documents to hadoop appropriated report frameworks. Last point is settling with assist of clusters in hadoop dispersed record machine with taking care of settling in information diminishin and pig and hive and begin, etc.

This paper here condenses capability & coping with speed in upgraded cloud with hadoop system.

CHAPTER 3

SYSTEM DEVELOPMENT

3.1 Designing

This topic completes the making of projects by Java and big data models using HQL and PIG.

The GUI of the project was developed in NetBeans in Windows 10. In this user has to register in the website, in order to execute the application. Once the user is registered in the travel recommendation system, the user is directed to login page of the application then they can access to recommendation system website for having best recommended travel agency.

3.2 Experimental setup

This area covers the hardware and software used while applying our problem statement using Hive and Pig and Sql.

3.2.1 Java

Software Requirements

- Language : Java
- Version : JDK 1.5
- IDE : Net Beans
- Back-end : Sql

3.2.2 For Hive and Pig platform

For Single Node , hardware requirements are described by the following details :

Type of hardware	Hardware (used)
Hardware RAM	- Quad core Intel i5 processor with 8 GB
- multiprocessor-based computer with a 2.00 Ghz processor	
- 64-bit operating system	(minimum 256 MB of RAM is essential)
Disk space	Atleast 4GB disk space is required
Memory	4 GB (2.5 GB on virtual machine)

Below details describes the software requirements that we used :

Softwares	Versions
Operating System	Linux (Ubuntu 12.04 LTS)
VMware workstation	15.1.2
Hadoop	2.7.1
Hive	1.2.1
Pig	
Netbeans IDE	8.0.2
Web browser	Google Chrome
Text Editor	notepad

Advantages of VMware Workstation are :

- Cloning potential of digital machines.
- Software development underneath a number working frameworks(OS).
- Turning of bodily PC into digital virtualized machine.
- Import the laptop and regulate virtual machines.
- Test programming below extraordinary working frameworks.

3.3 Analytical Model

This part discusses about the analytical models in the field of Big Data.

3.3.1 Traditionals Model – “Schema on Write”

The layout of the table is forced in traditional databases in the middle of the information stack cycle, on the of chance that the stacked information does not conform to the construction, the information stack is rejected, this practice is known as Schema-on-Write . texture on-Write allows to execute the inquiry quicker, as the data is now arranged in a specific arrangement and locating the section folder or labeling the information is anything but difficult.

The main points of interest in composite mapping are precision and pace of questioning. In the normalmanner(RDBMS). Therefore, we assume that we construct a schema consisting of 9 columns and we try to lineup information that can satisfy only 8 segments with information would be reje cted so outcome of the composition schema, here the information is perused against blueprint before it is0 kept in contact with the data base.

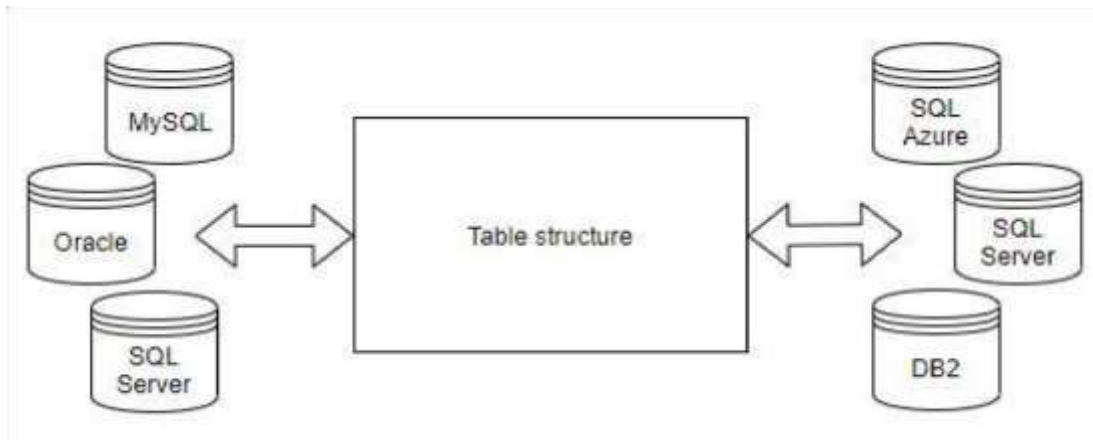


Fig. 6 Data Models for shaping data into tabular structure

3.3.2 Big Data Model – “Schema on Read”

In HIVE, the facts composition isn't always checked amid the heap time, alternatively it is proven whilst handling the question. Henceforth this technique in HIVE known as Schema-on-Read. Construction on- Read helps in speedy beginning data stack, considering that the records does now not want to pursue any inward schema(internal database design) to peruse or parse or serialize, as it is only a duplicate/move of a record.Organized is linked to the data just when it's perused, this enables unstructured data to be put away in the database. Since it is no longer important to represent the sample before putting away the statistics it makes it much less traumatic to get new facts sources.

With the Big Data and NoSQL worldview , "Schema-on-Read" implies we don't have to be aware of how we will utilize our information when we are inserting away it.

We do want to be aware of how we will utilize our statistics when we are using it and model in like manner.

Model: We might also at first put the information on HDFS in records , then practice a desk structure in Hive.

HDFS	File System
Exploration	File System Analyze and understand the data.
Hive	File System HDFS:

Fig. 7 Process for creating table and uploading data using the HDFS

3.3.3 Data Modeling in Big Data Process

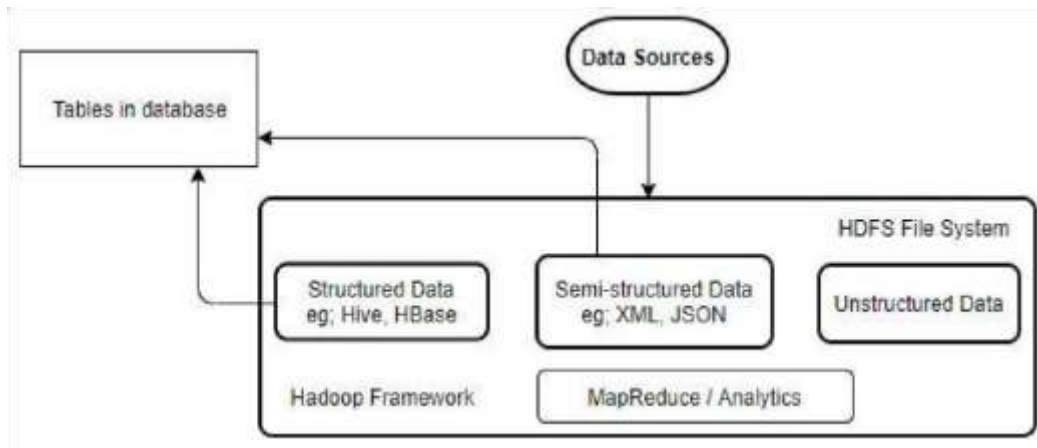


Fig. 8 Data Modeling used for handing all types of data

In the large records operational device , there are three sorts of datas i.e. structured information , semi structured and unstructured records . There are a few techniques through which we ought to execute these sorts of information.

For instance , let us think about the instance of organized data for that we have to make use of HIVE with HQL.

For unstructured statistics we have to at the beginning stack the report into HDFS record framework and after that convert it into an unthinkable employer using individual mapreduce strategies. For semi organized data we have to utilize JSON/XML records to change over it into an unthinkable configuration.

3.4 Analysis

This topic explains the exploring of data according to the required problem statement using Hive.

3.4.1 Data analysis using process of Hive

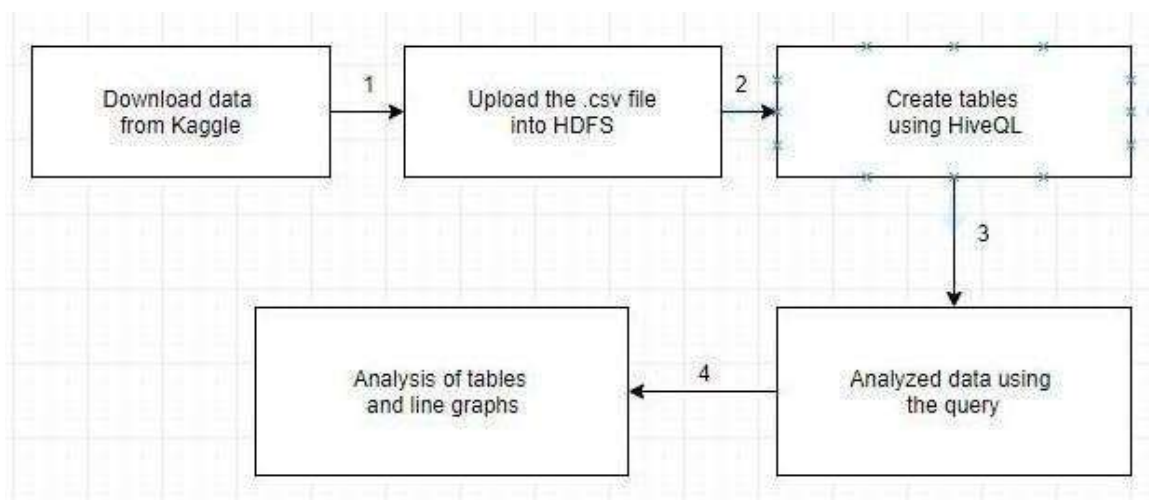


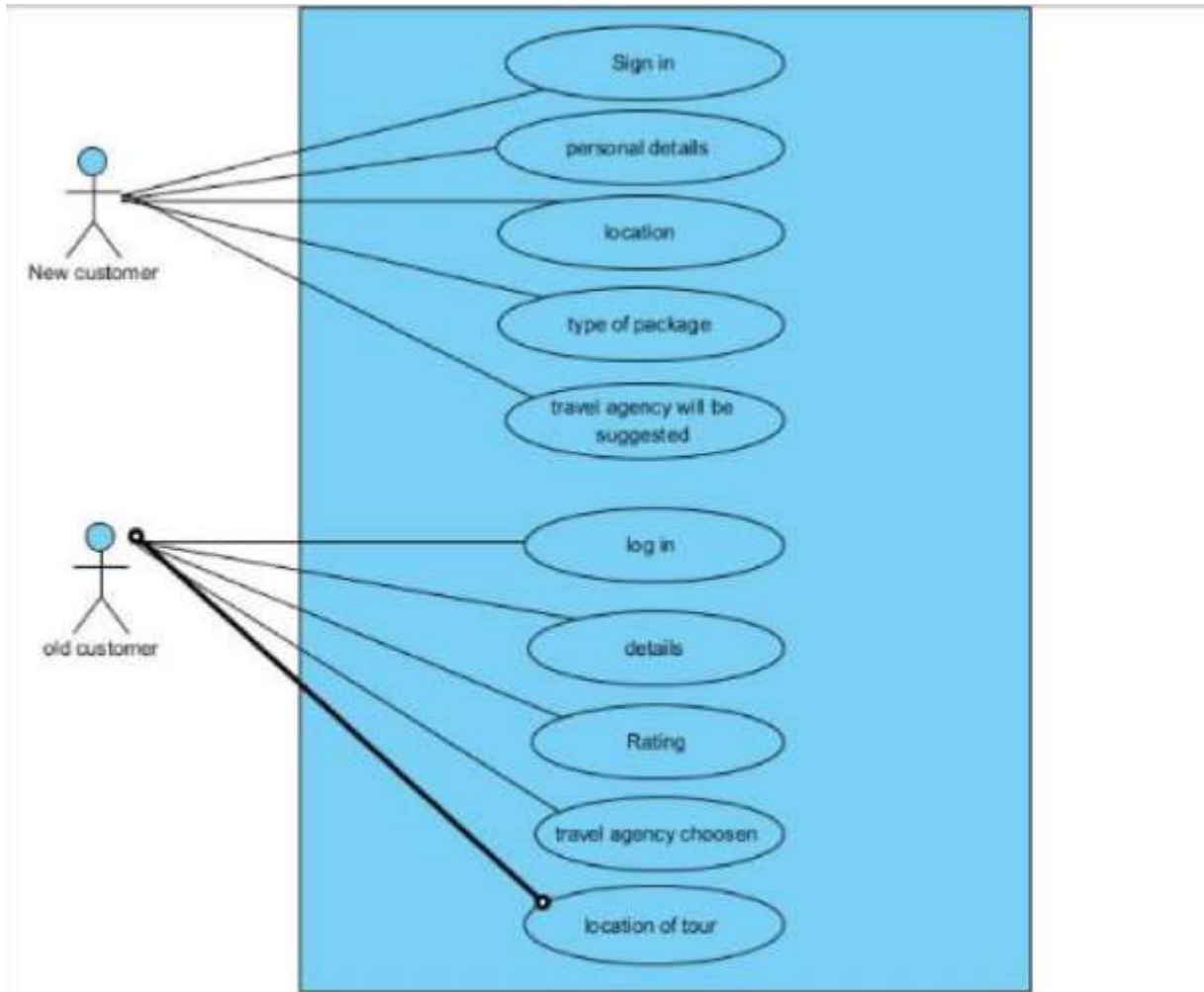
Fig. 9 Flowchart of process data analysis

In the Data Analysis, following steps takes place:

1. The downloaded .csv file (tourist.csv) is attached into the HDFS using:
load data local inpath 'login values.csv' into table rating;
2. After the file is attached to the HDFS, database (name - tourist) is made.
3. Queries are written to explore the data.
4. Data is showed using tables.

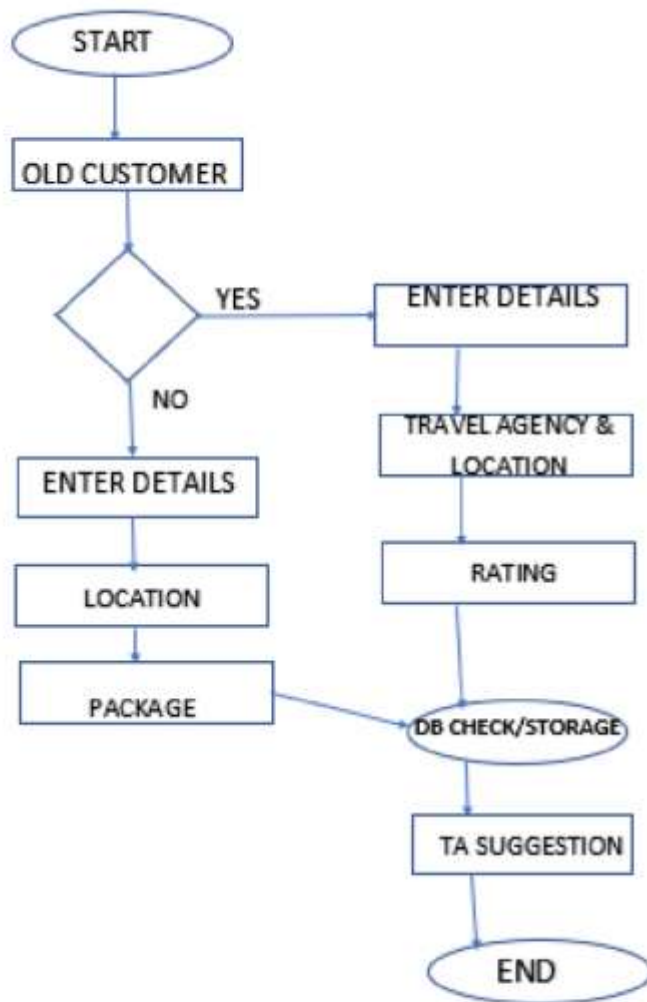
3.4.2 Design Analysis:

USE CASE DIAGRAM:



Snippet #8

FLOW CHART DIAGRAM:



Snippet #9

3.5 Algorithms

This topic details the algorithms we used in the concept of the problem topic of the project.

3.5.1 Using Java Applets

35.1.1 To import CSV file into a MySQL database using command:

CSV is dataset format used in ML and data science. MS Excel can be used in CSV format for basic data manipulation. Often, complex SQL queries need to be executed on CSV files, which is not possible with MS Excel.

Nonetheless, we need to convert CSV files to data tables before we can perform complex SQL queries on CSV files. There are many ways to transform CSV data into a database table format. One approach is to create a new table and copy all the information to the table from the CSV file. However, when the dataset is very large, copying and pasting data can be extremely cumbersome and time-consuming.

Another way is to write a script that reads the data from the CSV and inserts it into the data table. This method is faster than copy-pasting, but a manual script is still required.

```
mysql> load data local infile 'C:/Users/Akriti/Desktop/REALPROJECT.csv'
-> into table m4
-> fields terminated by ','
-> lines terminated by '\n';
Query OK, 49 rows affected (0.19 sec)
Records: 49 Deleted: 0 Skipped: 0 Warnings: 0
```

Code Snippet #10

35.1.2 Creating JDBC Application in Netbeans:

Below steps are involved in creating a JDBC application –

- **Import the packages:** Includes the packages containing the JDBC classes required for database programming. Largely, using *import java.sql.** would be sufficient.
- **To Register the JDBC driver:** Used to initialize a driver so as to open a communication channel with the database.

- **Establish a connection:** Using the *DriverManager.getConnection()* method to create a Connection object, shows a physical connection with the database.
- **To Execute a query:** Using an object of type Statement for framing and proposing an SQL statement to the database.
- **To abstract data from result set:** Using the appropriate *ResultSet.getXXX()* method to extract data from the result set.
- **Cleaning up the environment:** Explicitly closing all database resources versus depending on the JVM's garbage collection.

Connection of Sql with the GUI is done using:

```
private void jButton1ActionPerformed(java.awt.event.ActionEvent evt) { //GEN-FIRST:event_jButton1ActionPerformed
    try{
        String name=t1.getText();
        String password=t4.getText();
        Class.forName("com.mysql.jdbc.Driver");
        Connection con = DriverManager.getConnection("jdbc:mysql://localhost:3306/spc", "root", "manika");
        PreparedStatement pst=con.prepareStatement("Insert into login values(?,?)");
        pst.setString(1,name);
        pst.setString(2,password);
        int status=pst.executeUpdate();
        if(status>0)
        {
            System.out.println("\n\t Welcome to Travel Recommendation System \n\t");
            page4 p=new page4();
            p.setVisible(true);
            setVisible(false);
        }
        else{
            System.out.println("\n\t Data not Saved Successfully");
        }
    }
    catch(ClassNotFoundException | SQLException e){
        System.out.println(e.getMessage());
    }
} //GEN-LAST:event_jButton1ActionPerformed
```

Code Snippet #11

3.5.2 MapReduce Algorithm

The MapReduce calculation contains two important steps, in particular Map and Reduce.

The mapping is done by methods for Mapper Class

The reducing is done by methods for Reducer Class.

Mapper class catches information, fragments it, maps it and sorts it. The output of Mapper class is used as input by Reducers class, and thus seeks coordinating sets and lessens them.

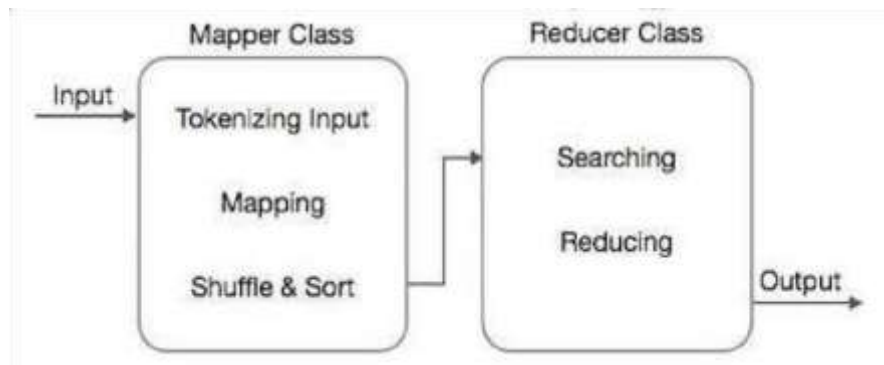


Fig. 10 Steps for Mapper & Reducer class

MapReduce actualizes different numerical calculations to separate a job into little fragments and appoint to different frameworks. In short, MapReduce calculation helps in direct the Map and Reduce undertakings for suitable servers bunch.

These above-said techniques incorporate the given accompanying :

- Sorting
- Searching

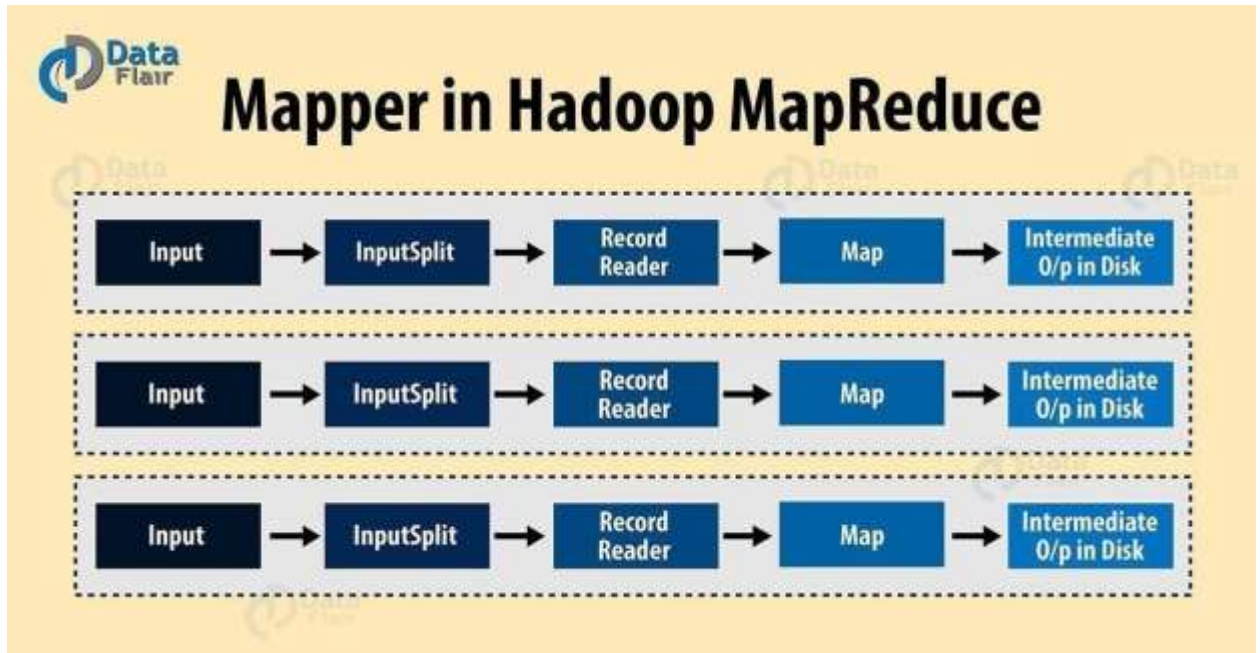


Fig. 11 Mapper Stage

- **Sorting**

Sorting is a basic MapReduce calculation to execute & fragment the given information. Map - Reduce executes arranging calculation naturally sortout the key-values sets from the mapperclass with the help of its keys.

- **Searching**

Searching plays a very important role in Map-Reduce calculations. Firstly, it is used in the combiner stage then in the Reducers stage.

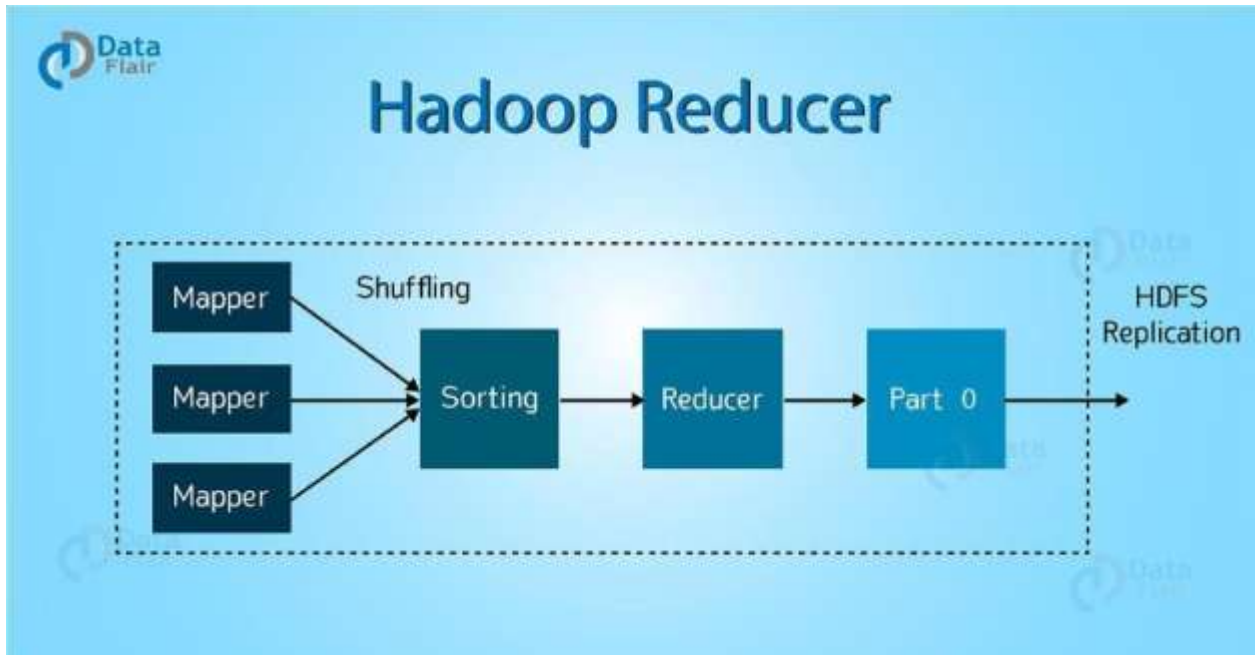


Fig. 12 Reducer Stage

3.6 Switching to Big Data over Sql

Big data applications has become increavsingly important over s past years. As the information from large volumes of data is increasingly dependent on many organizations from different sectors. Current data approaches and frameworks in the sense of big data are less well organized. Traditional approaches gave a slow response including lack of scalabilitys, reliability d & precision. A lot of work has been done to face the complex challenges of Big Data. This has resulted in j the

development of innumerable types of applications and technologies. With intention to help define and incorporate the best fusion of different Big Data technologies based on their technological requirements and determined applications. Not only offers a global impact of major Big Data technologies, but also correlations across various layers of structures like data storage layers, information processing layers, request layers, access layers, and suitable management layers.

Differences between SQL and HQL:

SQL is based on a relational database model whereas HQL is a combination of object-oriented programming with relational database concepts.

SQL manipulates data stored in tables and modifies its rows and columns. HQL is concerned about objects and its properties.

SQL is concerned about the relationship that exists between two tables while HQL considers the relation between two objects.

3.6.1 Big Data challenges

Big Data mining offers a lot of desirable prospects. Nonetheless, when research done on Big Data sets and abstract values and expertise from data mines, Researchers face many challenges. The difficulties includes : data captures, processing, search, evaluation, managing and visualizing at individual levels. With addition to , security and personal issues comes, specially in applications done by distributed data. Deluge of data and distributed flows sometimes surpass our abilities to harm. In reality, although Big Data's scale continues to grow exponentially, with the current technology ability to operate and inspect Big Data sets is relatively lower rates of data petabytes, exaabytes, and zettabytes

3.7 Test Plan

This area details the test plan for the executing of the data.

3.7.1 Dataset

A dataset is collection of information. An informational collection which relates to database table, where each column of the table called as an explicit variable with each line looks at to an individual from the educational file alluded to. Dataset taken is of travel agency rating and is generated randomly as it was not available on the internet.

The dataset has 5 columns and is of the review ratings of man travel agencies in distinct location.

In Table , the column name calls as the name of experienced customer who have reviewed the travel agencies .The field Travel Agency stands for various travel agencies which are available in the area. Rating field defines the rating of the travel agency by experienced customer. Package field represent the package that customer had opted for. Location field represents the location of the tour.

Field Name	Description
Name	Name Of customers
Travelagency	Name of travel agencies
Rating	Rating provided by customers
Package	Package opted by the customers
Location	Location of the tour

Table 1. Field names and its descriptions

One record from the dataset (login values.csv) is shown in the given figure:

Name	Travelage	Rating	Package	Location
akansha	cityland	3	A	Himachal
akshay	travelindi	4	A	Himachal
adnan	happy trav	2	B	Himachal
aatish	cityland	5	A	Himachal
sukanya	safari	1	B	Himachal
priyanka	skyblu	5	C	Himachal
suchetan	travelindi	3.5	D	Himachal
adnan k.	skyblu	4	A	Rishikesh
aashina	skyblu	2.5	D	Rishikesh
tanvi	cityland	1.5	A	Rishikesh
shristi	happy trav	5	B	Rishikesh
akriti	cityland	4.5	C	Rishikesh
manika	cityland	3	D	Rishikesh

Table. 2 Sample record of the tourist dataset

With this scenerio , the actual dataset is analyzed contains 100 rows.

Description of the dataset

- This dataset is for the Cabs agency rating analysis.
- This is a manual dataset which is used by a user to analyze the ratings of various cabs agencies of distinct locations.
- With the help of this, user can analyse whether available travel agencies could be choosen or not .
- Also, the manual dataset is choosen so as to analyse the data correctly and efficiently as online data gets updated very frequently.
- Also, to keep a track of its old customers to understand their preferences better.

This dataset is taken to perform the following tasks:

- to analyze the data records of a particular column field.

- to analyze the data records of various column fields collectively.

CHAPTER 4

RESULTS AND PERFORMANCE ANALYSIS

4.1 Using SQL and Java Applets

- **STEP 1: Collecting data regarding tourists and creation of a .csv file**

23	sunita	Rishikesh	cityland	3	Novembe	A	23		
24	gaurav	Rishikesh	cityland	3	Septembe	A	24		
25	shubham	Rishikesh	cityland	3.5	Septembe	C	25		
26	sunaina	Rishikesh	cityland	4	October	C	26		
27	shikha	Rishikesh	cityland	4	Novembe	C	27		
28	amit	Rishikesh	cityland	3	Novembe	A	28		
29	akash	Rishikesh	cityland	2	Septembe	A	29		
30	daman	Rishikesh	Happy Tra	2	October	A	30		
31	suman	Rishikesh	Happy Tra	2	Novembe	B	31		
32	pradyut	Rishikesh	Happy Tra	2	October	C	32		
33	prajwal	Rishikesh	Happy Tra	2.5	Novembe	C	33		
34	shikhar	Rishikesh	Happy Tra	2.5	Septembe	C	34		
35	samita	Rishikesh	Happy Tra	2.5	October	D	35		
36	balram	Rishikesh	Happy Tra	2.5	Novembe	D	36		
37	jai	Rishikesh	Happy Tra	2.5	Septembe	A	37		
38	shilpa	Rishikesh	Destinatic	3	October	A	38		
39	ankita	Rishikesh	Destinatic	3	Novembe	A	39		
40	aaradhna	Rishikesh	Destinatic	3	Septembe	A	40		
41	bala	Rishikesh	Destinatic	3.5	October	B	41		
42	alex	Rishikesh	Destinatic	3.5	October	B	42		
43	loka	Rishikesh	Destinatic	3.5	Septembe	B	43		
44	aradhna	Rishikesh	Destinatic	3.5	Novembe	C	44		
45	naina	Rishikesh	Destinatic	4	Novembe	C	45		
46	divyanshu	Rishikesh	Destinatic	4	Septembe	C	46		

REALPROJECT (+)

- **STEP 2: Importing The CSV file to MySQL 5.7**

```
mysql> load data local infile'C:/Users/Akriti/Desktop/REALPROJECT.csv'  
-> into table m4  
-> fields terminated by ','  
-> lines terminated by '\n';  
Query OK, 49 rows affected (0.19 sec)  
Records: 49 Deleted: 0 Skipped: 0 Warnings: 0
```

- **STEP 3: Making of GUI of the project .**

First page



Sign up page for new user



The sign-up page features a background image of a KLM airplane engine and wing against a sunset sky. The form is divided into two sections: **PERSONAL DETAILS** and **CONTACT DETAILS**. The **PERSONAL DETAILS** section includes input fields for NAME, AGE, GENDER (with a dropdown menu set to MALE), and PHONENO. The **CONTACT DETAILS** section includes input fields for HOUSE, CITY, STATE, COUNTRY, and PIN CODE. At the bottom of the form, there are two buttons: **SUBMIT** and **LOGIN PAGE**.

Log in page for registered user wanting recommendation on best cabs service for a particular asked location.



The travel recommendation form is set against the same KLM airplane background. It is titled **TRAVEL RECOMMENDATION** and contains the following fields: NAME, EMAIL ID., PHONE NO., DESTINATION (with a dropdown menu showing Rishikesh), PACKAGE (with a dropdown menu showing A), and DURATION (with a dropdown menu showing September). A **Click** button is located at the bottom of the form.

Log in page for user who wants to give rating for the cabs agency they have used.

RATING SUBMISSION

ID

NAME

DESTINATION

PACKAGE

TRAVEL AGENCY

RATING

DURATION

ID1

- **STEP 4: Connection built – java applets to mySql 5.7 using mysql connector.**

```
Class.forName("com.mysql.jdbc.Driver");
Connection con;
con = DriverManager.getConnection("jdbc:mysql://localhost:3306/e", "root","rohit");
PreparedStatement pst=con.prepareStatement("Insert into t1 values(?,?,?,?,?,?,?,?)");
pst.setInt(1,id1);
pst.setString(2,namel);
pst.setString(3,destination1);
pst.setString(4,travelagency1);
pst.setString(5,rating1);
pst.setString(6,duration1);
pst.setString(7,pack1);
pst.setInt(8,id2);
```

- **STEP 5: Analysing the datasheet for recommendation on best cab agency for a given location using sql query.**

```

}
Statement stmt=con.createStatement();
ResultSet rs=stmt.executeQuery("select travelagency,avg(rating) from t1 where package in(select package from t3 where name='"+name1;
+"') group by travelagency order by rating desc");

```

- **STEP 6: Result**

```

mysql> select travelagency,avg(rating) from t1 where package in(select package from t3 where name = "aman")group by travelagency order by rating desc;
+-----+-----+
| travelagency | avg(rating) |
+-----+-----+
| Destination Go | 3.8333333333333335 |
| Cityland | 3.375 |
| TRavelIndia | 2 |
| SkyBlu | 3 |
| Safari | 3 |
| Happy Travellers | 2.3333333333333335 |
+-----+-----+
1 rows in set (0.00 sec)

```

```

mysql> select travelagency,avg(rating) from t1 where package in(select package from t3 where name = "aman")group by travelagency order by rating desc limit 1;
+-----+-----+
| travelagency | avg(rating) |
+-----+-----+
| Destination Go | 3.8333333333333335 |
+-----+-----+
1 row in set (0.00 sec)

```

The screenshot displays an IDE interface with a project explorer on the left, a code editor in the center, and an output window at the bottom.

Project Explorer: Lists several Java applications: JavaApplication18, JavaApplication24, JavaApplication25, JavaApplicationhh, JavaFXApplication1, JavaFXSwingApplication1, and Realproject1.

Code Editor: Shows Java code from lines 183 to 196. The code includes a conditional check for a status variable, a database query execution, and a message dialog display.

```
183     if(status>0)
184     {
185         System.out.println("\n\t Welcome to Travel Recommendation System
186     }
187     }
188     else{
189         System.out.println("\n\t Data not Saved Successfully");
190     }
191     Statement stmt=con.createStatement();
192     ResultSet rs=stmt.executeQuery("select travelagency,avg(rating) from tl wh
193     pst.setString(1, name1);
194     if(rs.next())
195     {
196         JOptionPane.showMessageDialog(null," Successful");
        String travelagency2 = rs.getString("travelagency");
```

4.2 Using HIVE(HiveOL)

- **STEP 1: Loading the data set from local file system to hadoop file system. Data set contains attributes i.e id(int) , name(String) , destination(String) , travelagency(String) , rating(float) , month(String) , package(String) ,id2(int).**

```
hive> create database new;
OK
Time taken: 0.044 seconds
hive> use new;
OK
Time taken: 0.012 seconds
hive> create table new(id1 int,name String,destination String,travelagency String,rating float,month String,package String,id2 int)row format delimited fields terminated by '\t' lines terminated by '\n' stored as textfile;
OK
Time taken: 0.073 seconds
```

- **STEP 2: Dataset contains customer information who had given ratings on travelagency used by them in different packages.**

```
hive> select * from new1;
OK
1   akansha Rishikesh      Cityland      3.0    September    A      1
2   akshay  Rishikesh      Cityland      3.5    October B    2
3   adnan   Rishikesh      Cityland      2.0    November    C      3
4   aatish  Rishikesh      Cityland      4.0    October D    4
5   sukanya Rishikesh      TRavelIndia   5.0    November    A      5
6   priyanka Rishikesh      TRavelIndia   3.0    September    B      6
7   suchetan Rishikesh      TRavelIndia   2.0    October C    7
8   adnan k. Rishikesh      TRavelIndia   2.5    November    D      8
9   aashina Rishikesh      SkyBlu        3.0    September    A      9
10  tanvi   Rishikesh      SkyBlu        3.5    October B    10
11  shrusti Rishikesh      SkyBlu        4.0    November    B      11
12  akriti  Rishikesh      SkyBlu        2.0    October C    12
13  manika  Rishikesh      SkyBlu        1.0    October D    13
14  anusheel Rishikesh      SkyBlu        4.0    November    C      14
15  srishti Rishikesh      Safari        5.0    October D    15
16  danish  Rishikesh      Safari        4.0    November    A      16
17  divyansh Rishikesh      Safari        3.0    September    B      17
18  anish   Rishikesh      Safari        2.0    October C    18
19  kapil   Rishikesh      Safari        4.0    November    C      19
20  abhishek Rishikesh      Safari        4.0    October D    20
21  aarav   Rishikesh      Safari        3.0    September    D      21
22  meena   Rishikesh      cityland      3.0    October A    22
23  sunita  Rishikesh      cityland      3.0    November    A      23
24  gaurav  Rishikesh      cityland      3.0    September    A      24
25  shubham Rishikesh      cityland      3.5    September    C      25
26  sunaina Rishikesh      cityland      4.0    October C    26
27  shikha  Rishikesh      cityland      4.0    November    C      27
```

- **STEP 3: Creating table new2 in hiveQL. This dataset contains customer information who had request of suggesting them travelagency as pertheir information provided.**

```
hive> create table new2(name String, emailid String, phnno bigint, duration String, destination String, package String) row format delimited fields terminated by '\t' lines
erminated by '\n' stored as textfile;
OK
Time taken: 0.146 seconds
```

- **STEP 4: Loading dataset to hdfs table names new2. Dataset contains name(String) , emailed(String) , phnno(int), duration(String) , destination(String) , package(String).**


```
hive> load data local inpath '/home/training/Desktop/bs.txt' into table new2;
Copying data from file:/home/training/Desktop/bs.txt
Copying file: file:/home/training/Desktop/bs.txt
Loading data to table new.new2
OK
Time taken: 0.783 seconds
hive> select * from new2;
OK
name      emailid NULL      duration      destination      package
Akriti    soodakriti05@gmail.com  9816787169      September      Rishikesh      A
Manika    kansalmk712@gmail.com  9805084468      September      Rishikesh      A
Tanvi     sehgal123@gmail.com    7018803501      October Rishikesh      C
Aashina   aashina67@yahoo.com    9876534236      November      Rishikesh      D
Shristi   sahasiri@gmail.com     9837117004      September      Rishikesh      D
rohit     rgggtb@outlook.com     9876334998      November      Rishikesh      D
aman      aman@hotmail.com        8097658845      October Rishikesh      C
Time taken: 0.141 seconds
```

- **STEP 5: Analyzing the table new1 and new2 in order to fulfill the customer request who wanted to have suggestion of travelagency in terms of package , duration and destination.**

```
hive> select n.travelagency,avg(n.rating) from new1 n join new2 o on(n.package=o.package)where o.name='aman' group by n.travelagency;
Total MapReduce jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_202005080006_0021, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_202005080006_0021
Kill Command = /usr/lib/hadoop/bin/hadoop job -Dmapred.job.tracker=localhost:8021 -kill job_202005080006_0021
2020-05-24 00:09:17,231 Stage-1 map = 0%, reduce = 0%
2020-05-24 00:09:21,246 Stage-1 map = 100%, reduce = 0%
2020-05-24 00:09:28,324 Stage-1 map = 100%, reduce = 33%
2020-05-24 00:09:29,329 Stage-1 map = 100%, reduce = 100%
Ended Job = job_202005080006_0021
Launching Job 2 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
```

```
Cityland      2.0
Destination Go 3.8333333333333335
Happy Travellers      2.3333333333333335
Safari 3.0
SkyBlu 3.0
TRavelIndia      2.0
cityland      3.8333333333333335
Time taken: 30.251 seconds
```

- **STEP 6: Using ORDER BY the analyzed result would be :**

```

UN
Cityland          2.0
TRavelIndia      2.0
Happy Travellers  2.3333333333333335
Safari           3.0
SkyBlu           3.0
Destination Go   3.8333333333333335
Time taken: 44.457 seconds

```

- **STEP 7: Recommended travelagency to customer according to the information provided by them in terms of duration , package and destination by using the rating of existing customer .**

```

hive> select n.travelagency,avg(rating) as rating from new1 n join new2 o on(n.package=o.package) where o.name='aman' group by n.travelagency order by rating desc limit 1;
Total MapReduce jobs = 3
Launching Job 1 out of 3
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_202005080006_0043, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_202005080006_0043
Kill Command = /usr/lib/hadoop/bin/hadoop job -Dmapred.job.tracker=localhost:8021 -kill job_202005080006_0043
2020-05-24 00:36:42,134 Stage-1 map = 0%,  reduce = 0%
2020-05-24 00:36:46,266 Stage-1 map = 100%,  reduce = 0%
2020-05-24 00:36:54,335 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_202005080006_0043
Launching Job 2 out of 3
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):

```

- **STEP 8: Final result : Destination Go would be the best travel agency for the customer named “Aman” requested for package “C “ for destination = “Rishikesh” in the duration = “October”.**

```

UN
Destination Go   3.8333333333333335
Time taken: 42.521 seconds

```

4.3 Using PIG

- **STEP 1: Both datasets are uploaded to pig interface.**

```
(1,akansha,Rishikesh,Cityland,3.0,September,A,1)
(2,akshay,Rishikesh,Cityland,3.5,October,B,2)
(3,adnan,Rishikesh,Cityland,2.0,November,C,3)
(4,aatish,Rishikesh,Cityland,4.0,October,D,4)
(5,sukanya,Rishikesh,TRavelIndia,5.0,November,A,5)
(6,priyanka,Rishikesh,TRavelIndia,3.0,September,B,6)
(7,suchetan,Rishikesh,TRavelIndia,2.0,October,C,7)
(8,adnan k.,Rishikesh,TRavelIndia,2.5,November,D,8)
(9,aashina,Rishikesh,SkyBlu,3.0,September,A,9)
(10,tanvi,Rishikesh,SkyBlu,3.5,October,B,10)
(11,shristi,Rishikesh,SkyBlu,4.0,November,B,11)
(12,akriti,Rishikesh,SkyBlu,2.0,October,C,12)
(13,manika,Rishikesh,SkyBlu,1.0,October,D,13)
(14,anusheel,Rishikesh,SkyBlu,4.0,November,C,14)
(15,srishti,Rishikesh,Safari,5.0,October,D,15)
(16,danish,Rishikesh,Safari,4.0,November,A,16)
(17,divyansh,Rishikesh,Safari,3.0,September,B,17)
```

```
(name,emailid,,duration,destination,package)
(Akriti,soodakriti05@gmail.com,9816787169,September,Rishikesh,A)
(Manika,kansalmk712@gmail.com,9805084468,September,Rishikesh,A)
(Tanvi,sehgal123@gmail.com,7018803501,October,Rishikesh,C)
(Aashina,aashina67@yahoo.com,9876534236,November,Rishikesh,D)
(Shristi,sahasiri@gmail.com,9837117004,September,Rishikesh,D)
(rohit,rggtb@outlook.com,9876334998,November,Rishikesh,D)
(aman,aman@hotmail.com,8097658845,October,Rishikesh,C)
```

- **STEP 2: Filtering dataset2 by a name = “aman”.**

```
res1 = filter a2 by name2=='aman';
grunt> dump res1;
```

```
(aman,aman@hotmail.com,8097658845,October,Rishikesh,C)
```

- **STEP 3: Joining table1 and table2 on package = “C” to get all existing customer information who had given rating on same package.**

```
grunt> res4 = join res3 by package1, res2 by package2;
grunt> dump res4;
```

```
{Destination Go,3.5,C,aman,C)
{Destination Go,4.0,C,aman,C)
{Happy Travellers,2.0,C,aman,C)
{Happy Travellers,2.5,C,aman,C)
{Happy Travellers,2.5,C,aman,C)
{SkyBlu,2.0,C,aman,C)
{Cityland,2.0,C,aman,C)
{SkyBlu,2.0,C,aman,C)
{SkyBlu,4.0,C,aman,C)
{TRavelIndia,2.0,C,aman,C)
{Happy Travellers,2.0,C,aman,C)
{cityland,4.0,C,aman,C)
{Happy Travellers,2.5,C,aman,C)
{Happy Travellers,2.5,C,aman,C)
{Destination Go,4.0,C,aman,C)
{Safari,2.0,C,aman,C)
{Safari,4.0,C,aman,C)
{Destination Go,3.5,C,aman,C)
{Destination Go,4.0,C,aman,C)
{Destination Go,4.0,C,aman,C)
{Cityland,2.0,C,aman,C)
{TRavelIndia,2.0,C,aman,C)
{cityland,4.0,C,aman,C)
{Safari,4.0,C,aman,C)
{Safari,2.0,C,aman,C)
{cityland,3.5,C,aman,C)
{cityland,4.0,C,aman,C)
{cityland,4.0,C,aman,C)
{cityland,3.5,C,aman,C)
{SkyBlu,4.0,C,aman,C)
-----
```

- **STEP 4: Using FOR EACH and GENERATE to get travelagency and ratings for the package = “C”.**

```
grunt> res5 = foreach res4 generate travelagency,rating1;  
grunt> dump res5;
```

```
(Destination Go,3.5)  
(Destination Go,4.0)  
(Happy Travellers,2.0)  
(Happy Travellers,2.5)  
(Happy Travellers,2.5)  
(SkyBlu,2.0)  
(Cityland,2.0)  
(SkyBlu,2.0)  
(SkyBlu,4.0)  
(TRavelIndia,2.0)  
(Happy Travellers,2.0)  
(cityland,4.0)  
(Happy Travellers,2.5)  
(Happy Travellers,2.5)  
(Destination Go,4.0)  
(Safari,2.0)  
(Safari,4.0)  
(Destination Go,3.5)  
(Destination Go,4.0)  
(Destination Go,4.0)  
(Cityland,2.0)  
(TRavelIndia,2.0)  
(cityland,4.0)  
(Safari,4.0)  
(Safari,2.0)  
(cityland,3.5)  
(cityland,4.0)  
(cityland,4.0)  
(cityland,3.5)  
(SkyBlu,4.0)
```

- **STEP 5: GROUP** all travelagency for calculating the average of ratings in order to get the highest rating for recommending the best travelagency to the customer in terms of package , duration , rating .

```
grunt> res6 = group res5 by travelagency;  
grunt> dump res6;
```

```
(Safari,{{Safari,2.0},{Safari,2.0},{Safari,4.0},{Safari,4.0}})
(SkyBlu,{{SkyBlu,4.0},{SkyBlu,2.0},{SkyBlu,2.0},{SkyBlu,4.0}})
(Cityland,{{Cityland,2.0},{Cityland,2.0}})
(cityland,{{cityland,4.0},{cityland,3.5},{cityland,4.0},{cityland,4.0},{cityland,3.5},{cityland,4.0}})
(TRavelIndia,{{TRavelIndia,2.0},{TRavelIndia,2.0}})
(Destination Go,{{Destination Go,3.5},{Destination Go,4.0},{Destination Go,3.5},{Destination Go,4.0},{Destination Go,4.0},{Destination Go,4.0}})
(Happy Travellers,{{Happy Travellers,2.0},{Happy Travellers,2.5},{Happy Travellers,2.5},{Happy Travellers,2.5},{Happy Travellers,2.5},{Happy Travellers,2.0}})
grunt>
```

```
grunt> res7 = foreach res6 generate group ,AVG(res5.rating1) as rating;
grunt> dump res7;
```

```
(Safari,3.0)
(SkyBlu,3.0)
(Cityland,2.0)
(cityland,3.8333333333333335)
(TRavelIndia,2.0)
(Destination Go,3.8333333333333335)
(Happy Travellers,2.3333333333333335)
```

- **STEP 6: Using ORDER for obtaining the descending order of travel agency based on ratings.**

```
grunt> res8 = ORDER res7 by rating desc;
grunt> dump res8;
```

```
{Destination Go,3.8333333333333335)
{Safari,3.0)
{SkyBlu,3.0)
{Happy Travellers,2.3333333333333335)
{Cityland,2.0)
{TRavelIndia,2.0)
```

- **STEP 7: Using limit operator Destination Go would be the best travel agency by analyzing the table having existing customers rating . asked by the customer =”aman” for the destination =”Rishikesh” in duration =”October” for the package =”C”.**

```
grunt> res9 = limit res8 1;
grunt> dump res9;
```

```
Destination Go 3.8333333333333335
Time taken: 42.521 seconds
```

4.4 Comparison between existing system and adopted solution

Existing System:

1. Travel information are normally less.
2. Nearly every travel package contains numerous landscapes with lots of interest and attractions for people and therefore has intrinsic complex spatial-temporal relationships. A travel package includes, for example, locations that are geographically related.
3. Existing recommender systems typically rely on data collected and analyzed based on userspecified scores, but travel information is not conveniently available.

Disadvantages:

1. Travel information is much less sparsely compared to typical objects.
2. Modern recommendation items usually have a long stable value span, whereas travel package prices can easily depreciate over time.
3. The recommendation systems is for real world travel are typically very complex.
4. Travel package is made up of suitable landscape and therefore have an underlying dynamic temporared spatial relation.

Adopted New Solution:

Topic including unique features to distinct suggestions for suitable travel packages from standard recommender systems remains very open .Due to the disadvantages of above existing system i.e not able to handle very large data so we shifted to Big Data(Distributed Architecture).With coming practical and domain challenges in structuring and executing the suitable system of recommendations in customized travel recommendations. Plan would aid

visitors to have best travel agency with the selected package deal among all the Travel agencies. A customer will have a best travel agency with his desired package for a particular destination based on the ratings given by the existing customers who had experience with the packages. With this travel recommendation system aids in making the right choice with the best travel agency makes deal easy for the client.

Advantages:

1. Sharing of resources – Sharing of resources for hardware and software.
2. Openness – Flexibility to use various vendors ' hardware and software.
3. Competition – Concurrent performance enhancement processing.
4. Scalability – Improved throughput through the addition of new capital.
5. Malfunction tolerance – The ability to continue to work after a malfunction.
6. By using this recommendation approach the flaws of the existing system will be eliminated as it performs much better than traditional techniques.

CHAPTER 5

CONCLUSION

Conclusion

This project is done using Sql and Java applets but for the optimal efficiency when data set is too big to query with sql it is done using Pig and Hql over MapReduce framework .

Application helps to suggest best cabs agency with respect to factors like package, duration and destination etc among allother cabs services. Customer would choose a cabs agency for a destined place based on the recommendations provided by the existing customers who had experience with the same package in a travel agency. This makes simple and ease for the new customers to select the best travel agency deal.

New customers could select the best travel agency in short amount of time (instead of navigating to other websites).

Finally, the aim of project is to have an optimal system which is effecient in terms of cost , time and money plus with less hardwork. This can be seen in the results of sql , hive and pig. Computing time for analyzing the query by them is SQL > Hive >Pig.

Also, we conclude that existed system is not efficient for handling real time data which is very large so moving to distributed architecture and parallel computing using Hive & Pig over Map-Reduce framework have certain advantages over it .

REFERENCES

- [1] Ashish Thusoo, Joydeep Sen Sarma, Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff and Raghotham Murthy , ' Hive - A Warehousing Solution Over a Map-Reduce Framework'
- [2] Swapna Sahu , 'Pattern Finding In Log Data Using Hive on Hadoop', IJIRMPS | Volume 6, Issue 4, 2018
- [3] Scalability Study of Hadoop MapReduce and Hive in Big Data Analytics Jabeen1, Dr TSS Balaji2 1B , International Journal Of Engineering And Computer Science ISSN: 23197242 ,Volume 5 Issue 11 Nov. 2016, Page No. 18790-18792
- [4] Kshitij Jaju1, Abhishek Konduri3, ' Commercial Product Analysis Using Hadoop MapReduce', International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 04 | April-2016 , © 2016 IJSRSET | Volume 2 | Issue 2
- [5] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng , Prasad Chakka, Ning Zhang, Suresh Antony, Hao Liu and Raghotham Murthy , ' Hive – A Petabyte Scale Data Warehouse Using Hadoop '
- [6] SK. Jilani Basha, P. Anil Kumar, S. Giri Babu , 'Storage and Processing Speed for Knowledge from Enhanced Cloud Computing With Hadoop Frame Work : A Survey'
- [7] Ronald Taylor , ' An Overview of the Hadoop/Mapreduce/Hbase framework and its current applications in bioinformatics'
- [8] Namrata B Bothe , 'Migration of Hadoop To Android Platform Using 'Chroot', Volume 1 | Issue 5
- [9] Nishant Rajput , Nikhil Ganage ,and Jeet Bhavesh Thakur, ' REVIEW PAPER ONHADOOP AND MAP REDUCE', IJRET: International Journal of Research in Engineering and Technology, Volume: 06 Issue: 09 | Sep-2017
- [10] K. R. Srinath, ' Python – The Fastest Growing Programming Language' International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 12 | Dec-2017.
- [11] Wanliang Tan Xinyu Wang Xinyu Xu, ' Travel agencies OTA 2015
- [12] Callen Rain, ' sPerception of customers on online travel agency', 2016

ORIGINALITY REPORT

19%	2%	1%	19%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Jaypee University of Information Technology Student Paper	17%
2	Submitted to Higher Education Commission Pakistan Student Paper	1%
3	Submitted to CTI Education Group Student Paper	<1%
4	Submitted to Pathfinder Enterprises Student Paper	<1%
5	Submitted to Imperial College of Science, Technology and Medicine Student Paper	<1%
6	Submitted to Southern Cross University Student Paper	<1%
7	mafiadoc.com Internet Source	<1%
8	"Ubiquitous Computing and Ambient Intelligence", Springer Science and Business	<1%
		<1%
		<1%



www.iaeme.com
Internet Source

<1 %

Exclude quotes Off

Exclude matches Off

Exclude bibliography On