Analysis, Searching & Sorting in Big data using Apache Hadoop

Chauhan Gaurav Kumar M.
B.E Student
Department of Computer Science & Engineering
DYPCET, Kolhapur, Maharashtra

Dhanawade Vaibhav K.
B.E Student
Department of Computer Science & Engineering
DYPCET, Kolhapur, Maharashtra

Aswale Shweta S.
B.E Student
Department of Computer Science & Engineering
DYPCET, Kolhapur, Maharashtra

Bhosale Rasika R.
B.E Student
Department of Computer Science & Engineering
DYPCET, Kolhapur, Maharashtra

Prof. Kekade Mandar
Assistant Professor
Department of Computer Science & Engineering
DYPCET, Kolhapur, Maharashtra

Abstract

Big Data is generally considered as a large collection of data sets (In PetaBytes) which are difficult to analyze, share, store, search, transference through traditional data processing applications. Apache Hadoop is a Java written application for Storing, Processing and Analyzing large datasets (i.e Big Data) with the help of in-built Hadoop Distributed File System (HDFS), Yet Another Resource Negotiator (YARN), Map-Reduce Algorithm type algorithms and additional software packages like Apache Pig, Apache Hive, Apache HBase, Apache Spark and others. The aim of this project is to minimize the response time for the Searching, Sorting and Analyzing of the Big Data using efficient algorithms and databases.

Keywords: Big Data, Apache Hadoop, Apache Hive

I. INTRODUCTION

It is not easy to measure the volume of data which is being generated and stored, which is approximately to about minimum of 5PetaBytes per day as per the analysis. Thus, maintaining such a huge amount of data is a tedious and challenging task. Consider, the example of Facebook the leading social networking site, which generates about 1Pb of data every day. Comparing this with others are Stock Exchange Offices which also generates 1Pb of data. There is another ancestry.com the genealogy site stores about 2.5Pb of data. Thus, there was advent of the term Big Data. The problem with Big Data is that of Storage and Analysis. It is because the storage media is growing in size, but the access speed such as read and write are gradually decreasing with respect to increase in size of storage. Such accessing speed of the data can be improved if we access multiple disks in parallel i.e in short usage of cluster. But, if we use multiple hardware there are more chances of failures which is a negative factor for such an application. Thus, to overcome such problems, Doug Cutting created Hadoop application which was responsible for managing Big Data. Deployed in 2006, major companies besides Yahoo!, Last.fm, Facebook, started using Apache Hadoop as it core Big Data processing tool by Jan 2008.

The main components used in the applications are:

1) Common: A set of components and interface for distributed file system and general I/O (serialization, Java RPC, persistent data structures)
2) MapReduce: A distributed data processing model and execution environment that runs on large clusters of commodity machines.
3) HDFS: A distributed file system that runs on large clusters of commodity machines.
4) Hive: A distributed data warehouse. Hive manages data stored in HDFS and provides a query language based on SQL (and which is translated by the runtime engine to MapReduce jobs) for querying the data.

II. OBJECTIVES

The objective of this work is to analyze the data obtained from phone companies. The data that is obtained is in simple text format and do not follow any schema. The text files also content the collective in formatted data which is difficult for simple program to decipher.

The following objectives are to be fulfilled during the completion of the work.

A. Collection of Unformatted Data

The data is obtained from phone company and imported as text format files.

B. Installation of Hadoop And Hive Configuration

the installation of hadoop is carried out on the ubuntu operating system. The hive server is configured.

C. Loading Text Files Into Hive Database

The data imported from the text files to the hive thrift database from hadoop distributed file system.

D. Analysis of Data

The data of phone calls is analyzed according to the requirement of the operator.

III. SYSTEM ARCHITECTURE

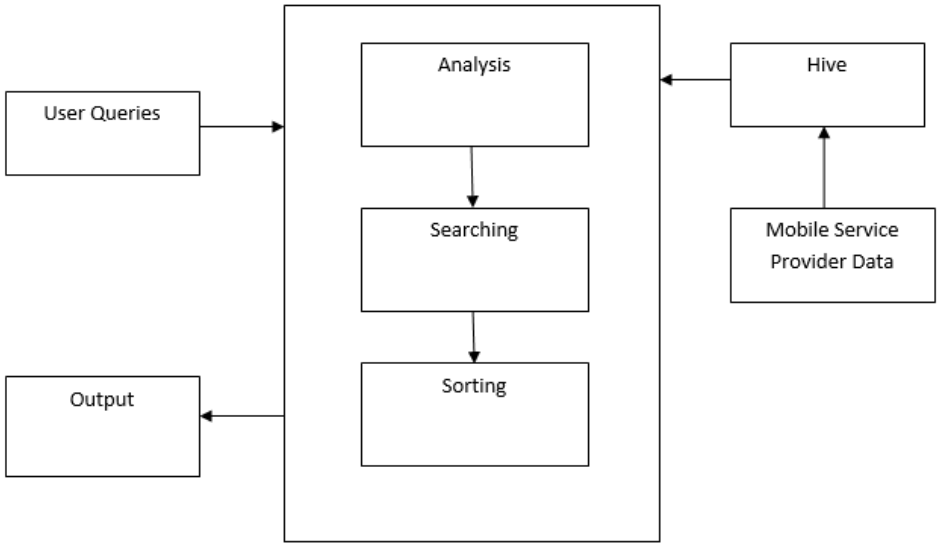

Fig. 1: System Architecture of Proposed System

The system consists of user queries where the user specifies the operation to be performed which includes analyzing of the data, searching and sorting of the data. We've provided the field to user for analyze, search and sort data. The Hive database include Mobile service provider data. The user provides queries to system and will get the required output.

IV. IMPLEMENTATION

A. Hadoop Implementation & Data Collection

1) Hadoop Implementation

Hadoop Installation is done on Ubuntu Operating System using the following procedure.

– In the Master server, download and install Hadoop using the following commands.

- # mkdir /opt/hadoop
- # cd /opt/hadoop/
- # wgethttp://apache.mesi.com.ar/hadoop/common/hadoop-1.2.1/hadoop-1.2.0.tar.gz
- # tar -xzf hadoop-1.2.0.tar.gz
- # mv hadoop-1.2.0 hadoop
- # chown -R hadoop /opt/hadoop
- # cd /opt/hadoop/hadoop/

– Install Hadoop on all the slave servers by following the given commands.

- # suhadoop
- $ cd /opt/hadoop

- $ scp -r hadoop hadoop-slave-1:/opt/hadoop
- $ scp -r hadoop hadoop-slave-2:/opt/Hadoop

2) Data Collection

– Initially the entire data shall be dumped in the hadoop from the Website Dumped Wikipedia.
– This will give a large number of Wikipedia files which are linked with each other.
– It is the data on which the Sorting and Searching operations will be performed.

B. Implementation of Map Reduce Algorithm

A Map-Reduce usually splits the input data set into independent chunks which are processed by the map tasks in a completely parallel manner. It works on master slave condition where Master is responsible for scheduling the job components task to the slaves, monitoring them and re-executing the tasks on failure and slaves on the other hand execute the tasks directed by the master

C. Analysis

– The data is collected or created for the various data mobile service providers.
– This data will be in the text format.
– These files contain the data regarding the calls logs, messages sent and the age group of the users and many more.
– This data will be analyzed according to the above categories by processing and cleaning the data files.

D. Implementation of Searching

Here the admin can search the information based on following parameters:
1) Search By Area: In this category admin can search information of user in the specific area(city)
2) Search By Number: Admin can get details of particular no by entering no in the given textbox
3) Search By Gender: Admin can check call details, message and net packs details of male and female.
4) Search By Age Group: Here can select age range for searching information about call and message details

E. Implementation of Sorting

Here the admin can perform sorting on the basis of the following parameters:
1) Sort by Age (low to high): Here user will be sorted on the basis of the age category.
2) Sort by City(Ascending): user will be sorted on the basis of where they are accessing the service.(city name in ascending)

V. RESULT

Below given are some of the result of system developed. These results show the analysis, searching and sorting of the data. The fig 2 shows login form. After login as a user, user will get analysis page as shown in fig 3. The fig 4 & 5shows the admin operation of searching.

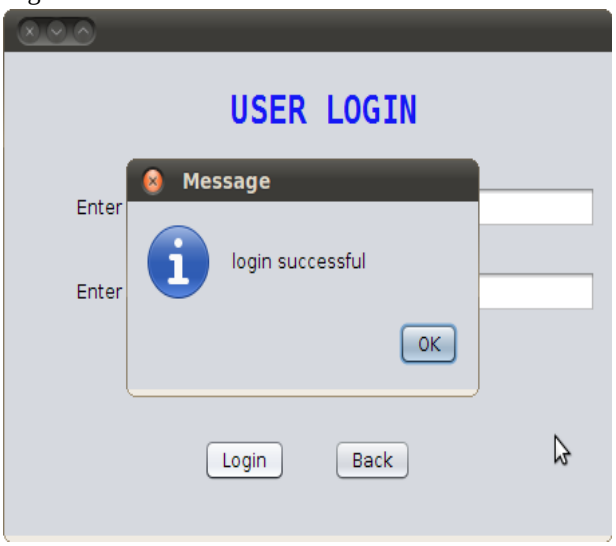

Fig. 2:

Fig. 3:

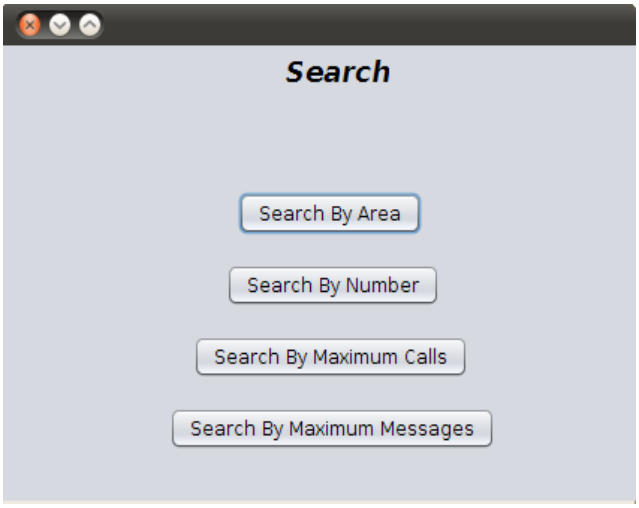

Fig. 4:

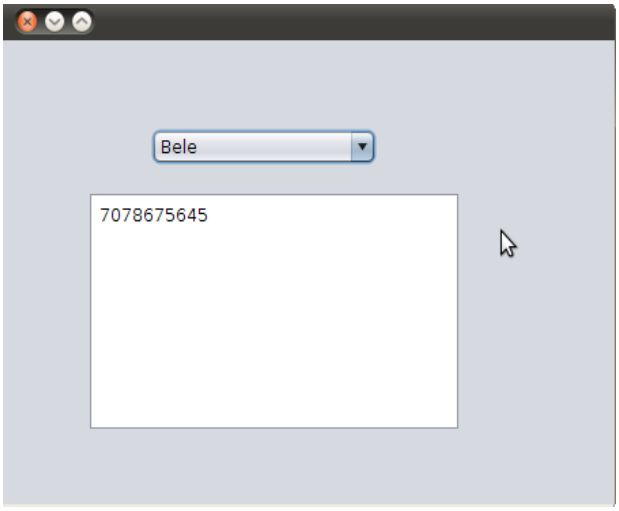

Fig. 5:

REFERENCES

[1] File and File Content Sorting using Hadoop framework paper by www.a4academics.com
[2] Approaches for keyword Query Routing from www.ijera.com
[3] Hadoop Tutorial by www.tutorialpoint.com
[4] Hadoop: The Definitive Guide 3rd Edition by Tom White `
[5] Hive Programming by Jason Rutberglen, Dean Wampler&Edward Capriolo
[6] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," in Proc. of OSDI '04, 2004.
[7] Hadoop in action by CHUCKLAM