

Population Diagnosis System

T Jashwanth Reddy¹, Voddi Vijay Kumar Reddy², T Akshay Kumar³

Business Intelligence Senior Associate, NTT Data, Hyderabad, India¹

Student, MS in Data Science with Business Analytics, Saint Peter's University, New Jersey, USA²

Student, MS in Information Technology Professional Computing, Swinburne University, Hawthorn, Australia³

Abstract: Population diagnosis system in Hadoop is a project developed with Apache HIVE, an abstraction of Map reduces. Population diagnosis system provides an introduction to the key concepts and methods required for population analysis. The system will describe the vital of population change and enable people to learn basic methods for measuring population structure and the determinants of population size and change.

The system will also provide an introduction to population projections and describe and evaluate how demographic data are collected and used. Prominence is placed on the understanding and elucidation of statistic data, as well as methods of population analysis. The data what you are going to analyze is an semi-structured data. After uploading their data to cluster anyone can access them again provided they got to be in the cluster or can also use virtual machines that contain the right software to analyze them without any need for conversion.

Keywords: Apache Hadoop-1.2.1, Apache hive-0.12.0, Population Diagnosis System, My SQL.

I. INTRODUCTION

Big Data is a Broad Phrase and a new approach to analyze a composite and huge amount of data; there is no single accepted definition for Big Data. But many researchers working on Big Data have defined big data in distinct ways. One such is defining the 4V's of the big data The first "V" is Volume, from which the Big Data comes from. This is the data which is difficult to control in conventional data analytics. The 2nd "V" is velocity, the high speed at which the data is processed and analyzed. The 3rd "V" is variety which helps to analyze the data like face book data which contains all types of variety, like text messages, attachments, images, photos and so on; the forth "V" is Veracity, that is cleanliness and accuracy of the data with the available huge amount of data which is being used for processing. In Facebook posts or Snap chat. These types of data have different structures and configurations and are more arduous to store in a traditional business data base. Working with big data means handling a variety of data formats and structures. Big Data involve data from all fields such as Health data, flight data, financial data and population data such data brings as to another V, value which has been proposed by a number of researcher [3, 4 and 5] i.e, Veracity. Hadoop permits one to save and query Big Data in a single and multi cluster environment using simple programming models. It is intended to scale up starting with solitary machines and will be scaled to many machines. In this paper Hive tool is used. The aim of Hive is to provide results and evaluate system performance and check activity of users. For all these dynamically discard the data into MYSQL data, but now since large amount of data in Terabytes which is injected into Hadoop Distributed File System files and processed by Hive Tool.

II RELATED WORK

As far as data storage model considered by B-trees or distributed hash tables using key-value pair is too limited to manipulate large data sets. Many projects have attempted to give solutions for distributed storage at higher-level services over wide area networks, often at Internet scale which include take a shot at distribute hash tables that initiated with enterprise, for example, Chord [16], Tapestry [18], CAN [14] and Pastry [15]. These frameworks address worries that don't emerge for Bigtable, for example, profoundly variable data transfer capacity, untrusted members, decentralized control and Byzantine adaptation to internal failure are not Bigtable objectives. Several database developers have created parallel databases that can store huge volumes of information. Oracle's Real Application Cluster database [13] utilizes shared disks to store information (Bigtable uses GFS) and an appropriated lock director (Bigtable uses Chubby). IBM's DB2 Parallel Edition

III. PROBLEM DEFINITION

Big Data has come up because we are living in society that uses the Full-scaled use of increasing data technology. As there exist large amount of data, the various challenges are faced about the management of such extensive data .The

challenges include the unstructured data, real time analytics, fault tolerance, processing and storage of the data and many more.

The size of the data is growing day by day with the exponential growth of the enterprises. For the purpose of decision making in an organizations, the need of processing and analyses of large volume of data is increases.. Data is generated from the many sources in the form of structured as well as unstructured form. Big data amount can range from terabytes to petabytes . The processing and analysis of large amount of data or producing the valuable information is the challenging task. As the Big data is the latest technology that can be beneficial for the business organizations, so it is necessary that various issues and challenges associated with this technology should bring out into light. The two main problems regarding big data are the storage capacity and the processing of the data

IV.POPULATION ANALYSIS

The proposed technique is made by considering following scenario under consideration. An Country has huge amount of data related to number of states, district, cities and list of population in each country. The issue they faced untill now it's, they have ability to analyze limited data from databases. The Proposed model intension is to develop a model for the population data to provide platform for new analytics based on the following queries.

Schema of the dataset: Longitude; Country; City; AccentCity; Population; Region; Latitude;

Sample data: ad; andorra la vella; Andorra la Vella; 07; 20,430; 42.5; 1.5166667

Work in this Dataset: Find the city which has maximum number of population. Find the total number of population that is coming under a particular country and find the country which has the highest population.

Final query: open hive terminal and perform

create table population1(Country STRING, City STRING, AccentCity STRING,Region INT,Population INT, Latitude FLOAT,Longitude FLOAT) row format delimited fields terminated by '|' stored as textfile;

loading data to hive: load data local inpath '/home/subbareddy/PopulationDataset' into table population;

querying the result: create table totalpopulation(City STRING, Population INT) row format delimited fields terminated by '|' stored as textfile;

```
INSERT OVERWRITE TABLE totalpopulation select City, sum(Population) as maxpop from population group by City sort by maxpop desc;
```

```
select City from totalpopulation LIMIT 1;
```

V.RESULTS

By using the above queries ,we are able to fetch the highest populated city from large set of data . Below are the screen shots of whole execution process and their related result after each step

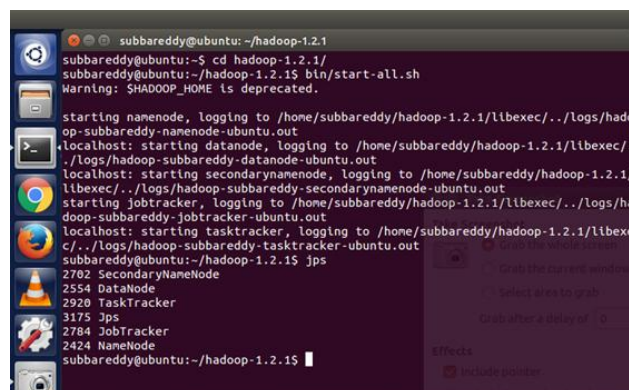
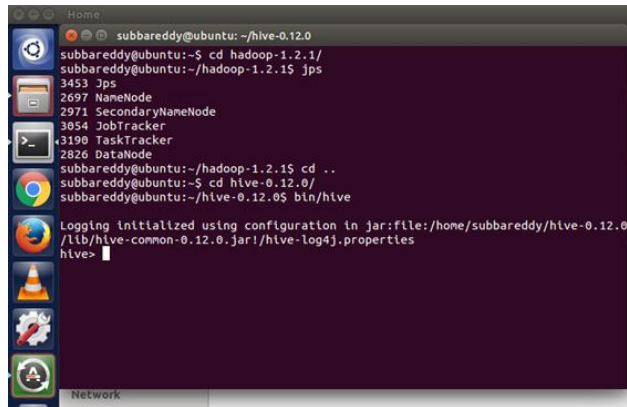
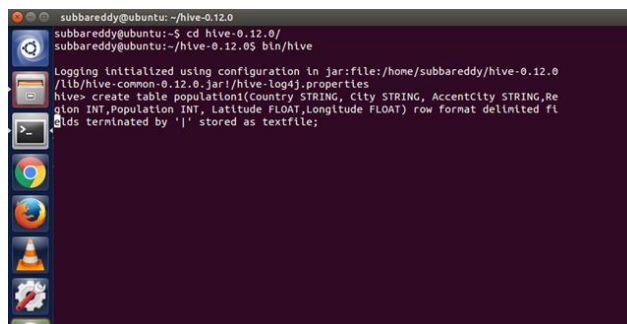


Fig1 : Starting all hadoop deamons and checking whether they are active or not .It also depicts the hadoop terminal in the unix environment



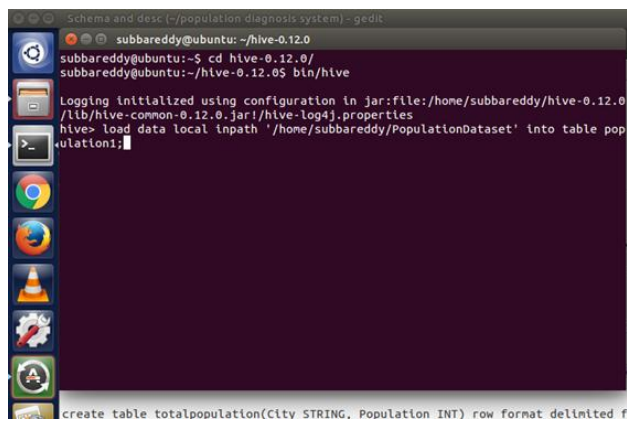
```
subbareddy@ubuntu: ~/hive-0.12.0
subbareddy@ubuntu:~$ cd hadoop-1.2.1/
subbareddy@ubuntu:~/hadoop-1.2.1$ jps
3453 Jps
2697 NameNode
2971 SecondaryNameNode
3054 JobTracker
3190 TaskTracker
2826 DataNode
subbareddy@ubuntu:~/hadoop-1.2.1$ cd ..
subbareddy@ubuntu:~$ cd hive-0.12.0/
subbareddy@ubuntu:~/hive-0.12.0$ bin/hive
```

Fig2 : Entering into hive terminal



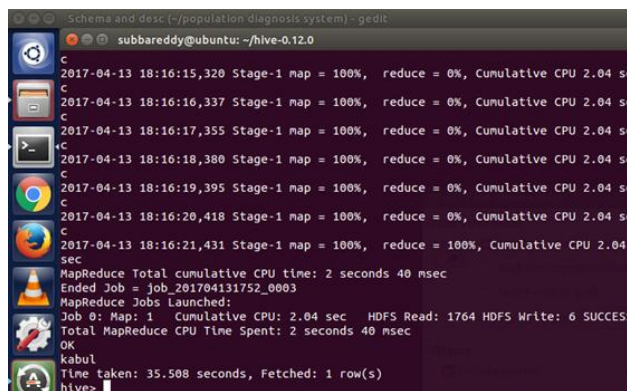
```
subbareddy@ubuntu:~/hive-0.12.0/
subbareddy@ubuntu:~/hive-0.12.0$ bin/hive
Logging initialized using configuration in jar:file:/home/subbareddy/hive-0.12.0/lib/hive-common-0.12.0.jar!/hive-log4j.properties
hive> create table population1(Country STRING, City STRING, AccentCity STRING, Region INT, Population INT, Latitude FLOAT, Longitude FLOAT) row format delimited fields terminated by '|' stored as textfile;
```

Fig3 : Creating a table in the Hive environment



```
subbareddy@ubuntu:~/hive-0.12.0/
subbareddy@ubuntu:~/hive-0.12.0$ bin/hive
Logging initialized using configuration in jar:file:/home/subbareddy/hive-0.12.0/lib/hive-common-0.12.0.jar!/hive-log4j.properties
hive> load data local inpath '/home/subbareddy/PopulationDataset' into table population1;
```

Fig4 : Loading data into the newly created table from local storage



```
2017-04-13 18:16:15,320 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.04 sec
2017-04-13 18:16:16,337 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.04 sec
2017-04-13 18:16:17,355 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.04 sec
2017-04-13 18:16:18,380 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.04 sec
2017-04-13 18:16:19,395 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.04 sec
2017-04-13 18:16:20,418 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.04 sec
2017-04-13 18:16:21,431 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 2.04 sec
MapReduce Total cumulative CPU time: 2 seconds 40 msec
Ended Job = job_201704131752_0003
MapReduce Jobs Launched:
Job 0: Map: 1 Cumulative CPU: 2.04 sec HDFS Read: 1764 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 40 msec
OK
kابل
Time taken: 35.508 seconds, Fetched: 1 row(s)
hive>
```

Fig5 : Final output (highest populated city)

VI. CONCLUSION

This paper points out the related work of huge data sets that were found in general scene, challenges and analysis on population that using Hive. we attempted to explore detailed analysis on population data sets such as listing countries cities and population in it. Here we focused on the processing the big data sets using hive component of Hadoop system in single cluster environment. This work will benefit the developers and analysts in accessing and processing their user queries.

REFERENCES

Web Resources

- 1) https://hadoop.apache.org/docs/r1.2.1/single_node_setup.html
- 2) http://www.tutorialspoint.com/hbase/hbase_overview.htm
- 3) <https://cwiki.apache.org/confluence/display/Hive/GettingStarted>
- 4) http://hbase.apache.org/0.94/book.html#getting_started
- 5) http://www.tutorialspoint.com/hadoop/hadoop_introduction.htm
- 6) http://www.tutorialspoint.com/hive/hive_introduction.htm

Bibliography

1. Haloi, S.; Jorapur, R.; Singhvi, R.; Mukherjee, A.; Akram, W.; Datta, J., (18-22 Dec., 2012) , “Shared disk big data analytics with Apache Hadoop”
2. Manashvi Birla, Ushma Nair, Aditya B Patel ,(6-8 Dec. 2012), “Addressing Big Data Problem Using Hadoop and Map Reduce”
3. Tien, J.M.(17-19 July, 2013),” Big Data: Unleashing information”
4. Yu Li ; Wenming Qiu ; Awada, U. ; Keqiu Li.,(Dec 2012),” Big Data Processing in Cloud Computing Environments”
5. Sagiroglu, S.; Sinanc, D. ,(20-24 May 2013),”Big Data: A Review”
6. Grosso, P. ; de Laat, C. ; Membrey, P.,(20-24 May 2013),” Addressing big data issues in Scientific Data Infrastructure”
7. Kogge, P.M.,(20-24 May, 2013), “Big data, deep data, and the effect of system architectures on performance”
8. Szczuka, Marcin,(24-28 June, 2013),” How deep data becomes big data”
9. Zhu, X. ; Wu, G. ; Ding, W.,(26 June, 2013),” Data Mining with Big Data”
10. Zhang, Du,(16-18 July, 2013),” Inconsistencies in big data”