BIGDATA ANALYSIS USING HADOOP ECOSYSTEMS ON CLOUD PLATEFORM

Rajeshi Tanwar, Prof. Kailash Patidar[2] ,Mr. Rishi Kushwah[3]
Mr. Narendra Sharma[3]

[1]M. Tech Scholar, [2]HOD CSE, [3]Asst. Prof. CSE, SOE. SSSUTMS Sehore (M P) SOE. SSSUTMS Sehore (MP) SOE. SSSUTMS Sehore (M P) SOE. SSSUTMS Sehore (M P)
rashitanwar89@gmail.com.,hodcsit.sssist@yahoo.com, rishisinghkushwah@gmail.com narendra_sharma88@yahoo.com

Abstract— The traditional relational database systems cannot accommodate the need of analysing data with large volume and various formats, i.e., Big Data. Apache Hadoop as the first generation of open-source Big Data solution provided a stable distributed data storage and resource management system. However, as a MapReduce framework, the only channel of utilizing the parallel computing power of Hadoop is the API. Given a problem, one has to code a corresponding MapReduce program in Java, which is time consuming. Therefore, Hadoop can be a poor fit for interactive data processing.

The demand of interactive Big Data processing necessitated decoupling of data storage from analysis. The simple SQL queries of traditional relational database systems is still the most practical analysing tool that people without programming background can also benefit from. As a result, Big Data SQL engines have been spun off in the Hadoop Ecosystem. So for which we can using apache hive which is an hadoop ecosystem and works same same as SQL so easily work on bigdata by using SQL (hive) on Hadoop (Storing and processing bigdata). In this paper we can create a hadoop cluster on cloud environment and integrate hive on top of the hadoop and analyze the data stored in HDFS by using hive query. In this we can also compare the performance of hive by another hadoop ecosystem called apache pig.

***Keywords*-- Hadoop, bigdata, performance comparison, cloud, hive, pig.**

I.INTRODUCTION

Big data is a structured and unstructured data video, audio, pictures and information emails etc.it is very large amount of data provided by social site and daily activities of social media like news and news channels or new technology, television, mobile, and computers and industries all are big data [8]. That' s we can say it is more than thousands of information storage for the growth of the industries. Know if we have information or history of previous data than it's very easy for the next new changes for the industries or business. Today' s competitive world in this time industries and business are growing very fastly by the help of the storage of the previous data which is known as big data. Big data is very hard to process and analysis the data easily. But with the help of HADOOP [5] data is easily to process and analysis the data easily. Big data is a different-different collection of complex data sets. Big data [10] is produced by different kinds of sources like television, mobile and other sources like industries data records.

It is three characteristics of big data:
l. volume
2. velocity
3. verity.

A. Volume

Volume mainly defined is amount of data. Volume of data is growing exponentially megabytes, gigabytes, zeta bytes and petabytes. It' s very large amount of data and also hard to the process and storage. Like Some earlier estimates suggested by the websites that 20 petabytes of storage its very large space was used to store 260 billion Facebook photos and massages or tag. In 2010, it was some newly reported by one million photographs were processed by Facebook per second. Twitter is generate the data 12 terabytes of data per day its newly research. Now the Facebook in 2012 stated that 2.7 billion "likes" and "comments" and massages were registered per day by the peoples.

B. Velocity

Social media is one the major factor to provide the data exponentially. Social sites is continuous generated a complex data unstructured and semi-structured form of data. There are currently generated 90% data in last two years. Increase the velocity of big data with help of mobile, televisions more advance technology. Internet is main factor to collecting of huge data. According to the user requirement which is save somewhere. It is known as velocity.

C. *Verity*

It is different kind of data are structured and Unstructured format. Structured data is always fixed format everywhere there is no possibility of changes of this data like tabular data, ERP, backup storage for large volume of data. But Unstructured data always no fixed format like text, audio, video, images and many social sites' data like Facebook, twitter, LinkedIn, logs file web chats etc. All companies and industries are having up to 85% of the data semi-structured and unstructured and structured format of data.

HADOOP

In hadoop[6] developers can deploy programs written in any other languages or in java for the processing of data parallely across multiple commodity machines despite of the fact that hadoop framework is written in java. One of the key features of hadoop is that it partitions the computation and data across multiple nodes and then makes the application computation run in parallel on these nodes. Important features of hadoop are redundancy and reliability which means that if any of nodes fails due to technical fault or other failures, it automatically creates a backup for that node without any intervention of the operator.

Depending on the process complexity the time of execution may vary from minutes to hours.[8] Hadoop has emerged out to be a potential solution for number of applications in web log analysis, visitor behaviour, search indexes, indexing and analysis of text content, applications in biology, genomics and physics, machine learning researches and natural language processing researches and in all sort of data mining.

Apache Hive

Facebook created Hive for analyzing large datasets. It is a most widely adopted data warehousing application which can provide the Relational model and SQL interface.

Hive infrastructure runs on the top of Hadoop. It mainly helps in providing summary of the data, query and analysis of the unstructured data. Since its incubation in 2008, Apache Hive is considered as standard for Batch and Interactive SQL workloads on data in Hadoop. The Hive tables are similar to relational databases, but the tables in Hive are made up of partitions.

Apache Pig

Yahoo started Pig as a research project to focus on analysis of large datasets. It was designed in the style of SQL and also MapReduce. Pig is used with Hadoop in general. Pig Latin is a procedural language used by Apache Pig. The programmers use Pig script and execute the command in the grunt shell. It runs MapReduce programs when running the pig script in grunt shell. The major components of Apache pig are, Parser: Checks the syntax of the script. Optimizer: Carries out the plan of the script as push down. Compiler: Compiles the plan into MapReduce job. Execution engine: Execute the MapReduce jobs and finally Hadoop produce the results.

II. LITERATURE REVIEW

In [1] the research work is carried out using Apache pig and hadoop on a crime dataset. It describes the large volume of data yielded from multiple sources and termed it as voluminous data. Crime and crime related datasets with ever growing population has risen to a higher extent and is a attention seeking subject to government for taking strict measures by prevailing law and procedure. Bigdata analytics using pig and hadoop has been applied on this crime dataset with the idea behind it as the optimal improvement for analysing some trends that needs to figure out, so among the citizens of the country there could be a feel of security and safety. Also it could help the government to furnish law and procedure and welfare among the people of the country. Analysis results shows the total number of crimes occurred in every state, crimes that took place against women, type of crime and from year 2000 to 2014 the total number of crimes that took place. Experimental setup was pseudo distributed mode of hadoop and it was concluded that scripting language Pig Latin has fewer lines of code as compared to mapreduce program but the execution time increases in pig as compared to mapreduce.

In [4] With the ever increasing man-machine interaction, automation of technique and decline in hardware and software package package value, the number of digital data generated and used is increasing day by day. the large data referred here is that the large amount of digital data generated in each and every second in structured, semi-structured and unstructured format throughout the world. This rising field of giant data analytic has driven the investigator worldwide toward vogue, development and implementation of assorted tools, technologies, style and platforms for analyzing the large volume of information generated day to day. Immense data embrace data sets that are difficult for inheritance management system to analysis. This paper details some analysis like feedback analysis, sentiment analysis and word-count. Feedback is vital for the system improvement, finding loop holes and still as for correct work distribution. Feedback is efficacious data that may be accustomed observe call. Feedback is vital not only has it highlighted weaknesses however additionally for strengths. If analysis of feedback is completed in wrong means then the results of analysis will be wrong. As a result, the pattern known will be incorrect therefore creating the entire system incorrect as a full. we are going to be implementing this projected system for feedback analysis victimization Map-Reduce framework for process massive knowledge set and for storage we are going to use Hadoop.

In [2], the author describes that huge knowledge analytics has attracted intense interest from all academe and trade recently for it's decide to extract data, data and knowledge type huge knowledge. huge knowledge and cloud computing, 2 of the foremost vital trends that ar shaping the new rising analytical tools. huge knowledge analytical capabilities victimization cloud delivery models might ease adoption for several trade, and most vital thinking to price saving, it might modify helpful insights that would providing them with totally different types of competitive advantage. several firms to produce on-line huge knowledge analytical tools a number of the highest most firms like Amazon huge knowledge Analytics Platform ,HIVE net primarily based Interface, SAP huge knowledge Analytics, IBM InfoSphere BigInsights, TERADATA huge knowledge Analytics, 1010data huge knowledge Platform, Cloudera huge knowledge answer etc. Those firms analyze huge amount of knowledge with facilitate of various sort of tools and additionally offer simple or easy program for analyzing data.

In [3], data technology provides utmost importance to process of knowledge. Some petabytes of knowledge isn't enough for storing great deal of knowledge. massive volume of unstructured and structured knowledge that gets created from varied sources like Emails, web logs, social media like Twitter, Facebook etc. the foremost obstacles with process huge knowledge embody capturing, storing, searching, sharing and analysis. Hadoop permits to explore complicated knowledge. it's an open supply framework written in Java that supports parallel and distributed processing and is employed for reliable storage of knowledge. With the assistance of huge knowledge analytics, several enterprises ar able to improve client retention, facilitate with development and gain competitive advantage, speed and scale back complexness. E-commerce firms study traffic on websites or navigation patterns to work out probable views, interests and dislikes of someone or a bunch as a full counting on the previous purchases. During this paper, they compare some generally used knowledge analytic tools.

III PROBLEM DEFINITION

The knowledge that we tend to we tend tore generating was growing in no time - as example we grew from a 15TB knowledge set in 2007 to a 700TB data set nowadays. The infrastructure at that point was therefore inadequate that some daily processing jobs were taking over daily to process and therefore the state of affairs was simply obtaining worse with each passing day. we tend to had urgent need to want for infrastructure that would scale in conjunction with our knowledge. As a result we tend to started exploring Hadoop as a technology to deal with our scaling wants. However writing a mapreduce program terribly troubles ome currently days for a fancy downside which require very high finish programming skills and additionally takes code maintenance time.

IV PROPOSED WORK

For analzing these large and complex data we use Hive which is a popular query languages like SQL and as a result users ended up spending hours (for writing mapreduce code) to write programs for even simple analysis.

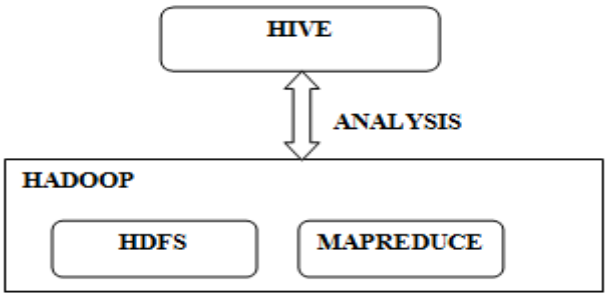

Figure.2. Workflow Diagram of SQL-on-Hadoop system

Our Steps or Algorithm Steps will follow:

Step 1: First we create a hadoop cluster on google cloud and loaded the dataset into HDFS.

Step 2: The datasets is pre-processed by mapreduce.

Step 3: Dataset in analyzed by huve and pig.

Step 4: Compare the performance of these bigdata analytical tools.

V. EXPERIMENTAL & RESULT ANALYSIS

All the experiments are performing on google cloud platform (GCP) on which we developed a heterogeneous clusters of five nodes and table-1 shows the hardware properties of each node. Cluster is implemented on ubuntu with hadoop is configure on it and top of the hadoop hive is working with default configuration, So to achieve this we are going to follow the following methods:

- Loading datasets into HDFS.
- Analyze the dataset by using Hive.
- Compare the performance of hive with pig.

Node	Hardware Property
Master	32-bit dual core processor with 20GB of HDD and 2GB of RAM.
Slaves (4 Datanodes)	32-bit dual core processor with 13GB of HDD and 1GB of RAM

Table-1 Hardware property of experimental environment

Loading Dataset into HDFS

For loading dataset into HDFS we first create an heterogenous cluster of hadoop , for which we can use a google cloud services to create a hadoop cluster. we can use Google Cloud Platform to create a cluster for five nodes which is shown in figure 2

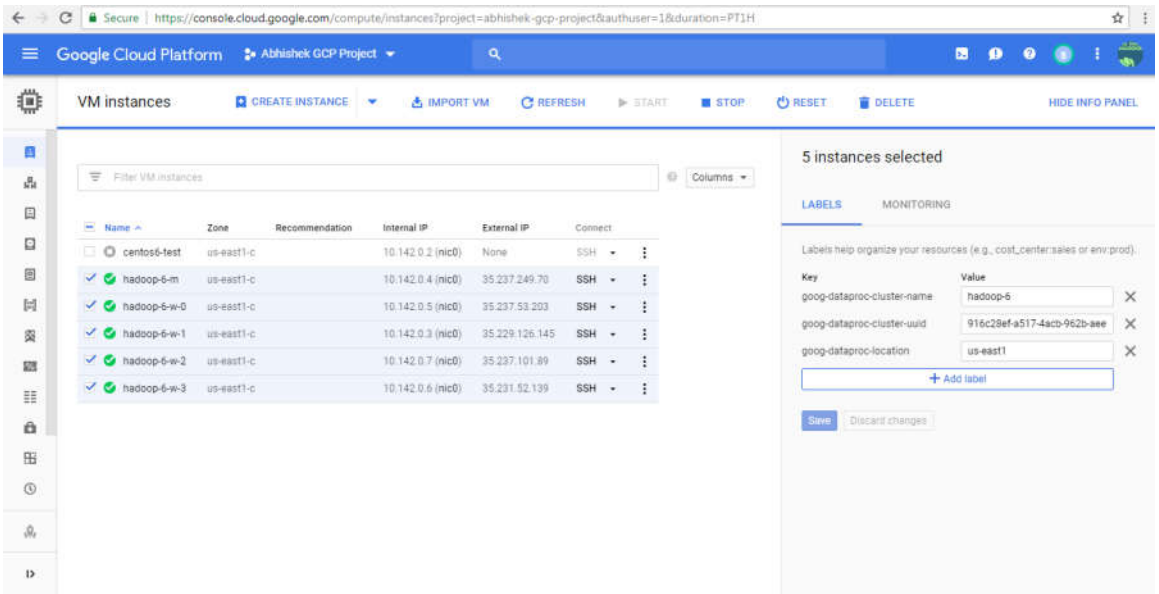

Figure 2. Heterogeneous cluster of five nodes

After that we can start the cluster and on master node we can operate all the functions so we just ssh to the master and the ssh create a connection between master terminal to browser. After connection to the VM, we can access the terminal of master node and just load the dataset (we take h1b visa application dataset) into HDFS by using hadoop put command and the data which is stored into the HDFS are shown in figure 3.

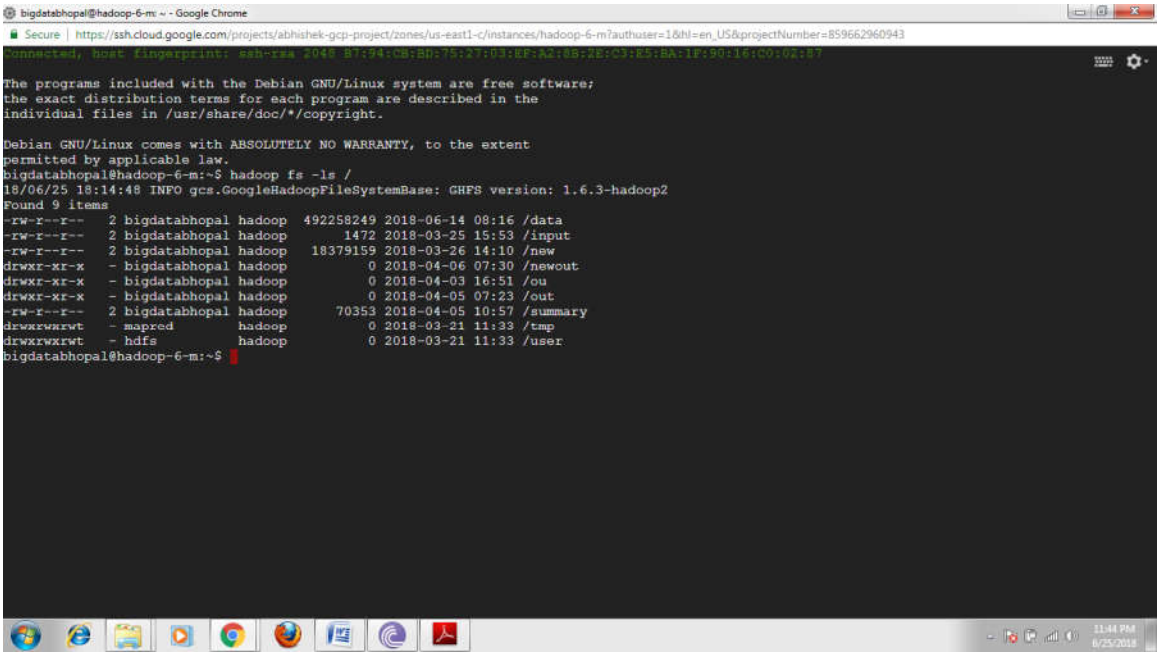

Figure 3. Dataset is loaded into HDFS

Analyzing the dataset by using Hive

After storing the h1b visa application dataset into HDFS, we are started hive over hadoop and for analyzing these data we first create a table for storing a dataset because hive works same as SQL so before analyzing we first create and stored the dataset into table. We can take h1b visa dataset for analyzing using hive. we can launch a SQL query on the table and the result and time taken by the query are shown in figure 2.

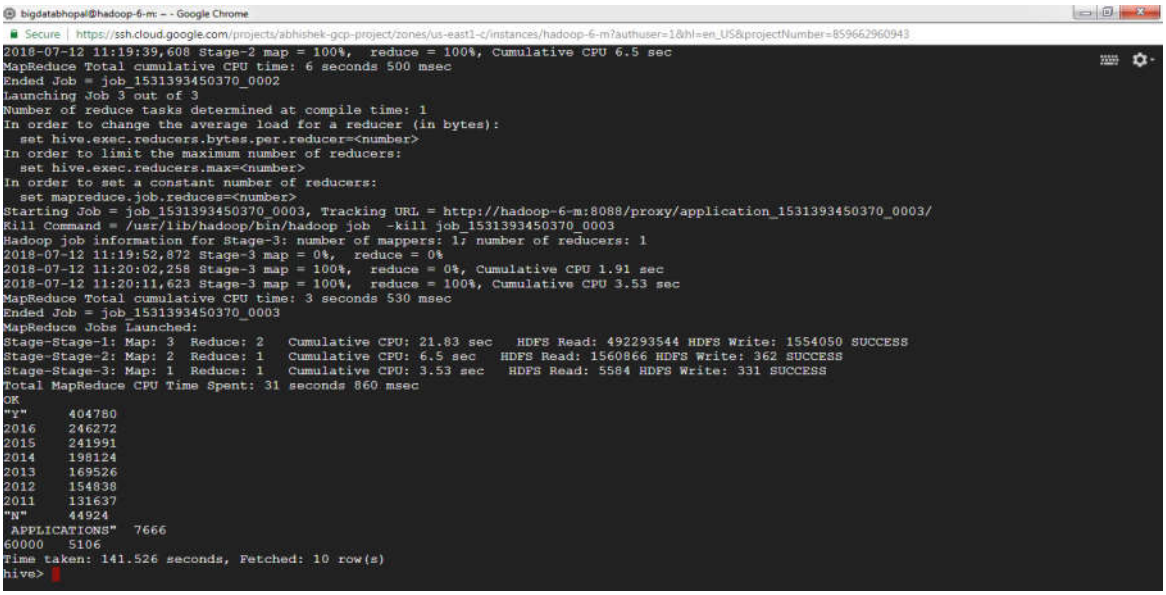

Figure 4. Time taken by Hive query

Cmparison with apache Pig

After analyzing the dataset by using hive we can also analyse the dataset using pig which is an another hadoop ecosystem used for analytics purpose. In this we can compare performance of hive and pig by analyzing data on hadoop cluster. Apache pig supports pig latin language which is a scripting language much similar like SQL, so we analyze the same dataset by using pig the output of pig script is shown in figure 5.

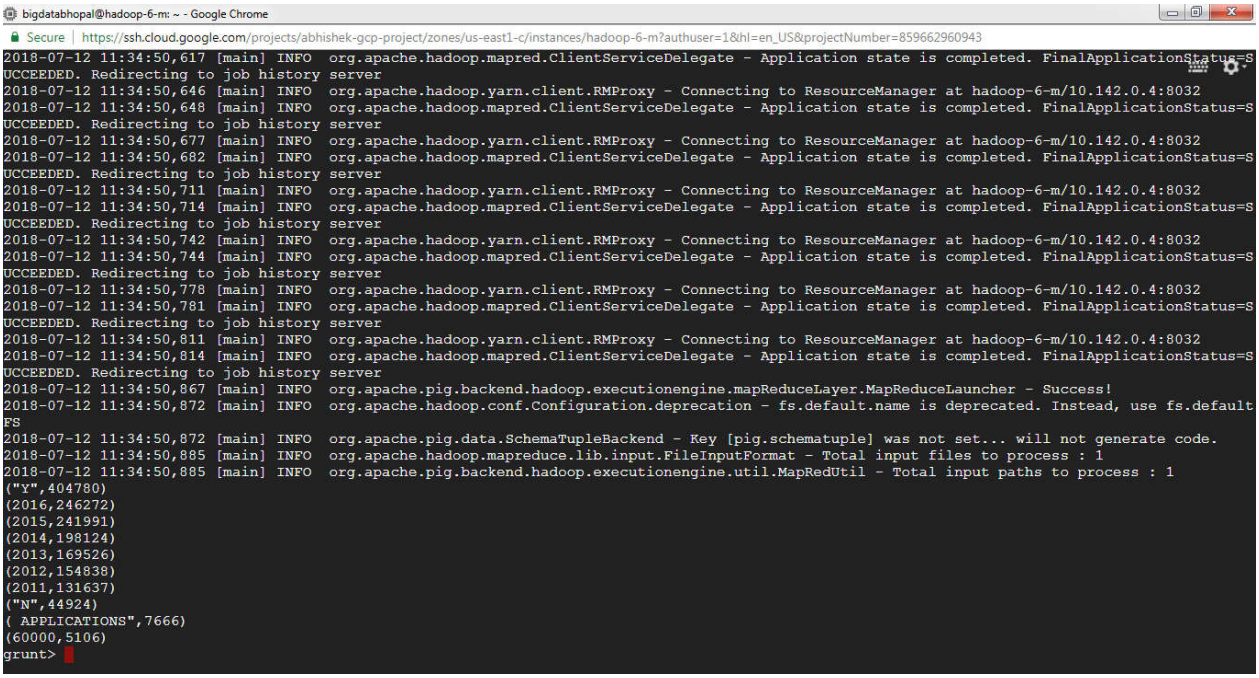

Figure 5. Output of Pig script

In these the pig latin script output is similar as hive query output so we can say that both the tools are very accurate and the time taken by pig is shown in figure 6.

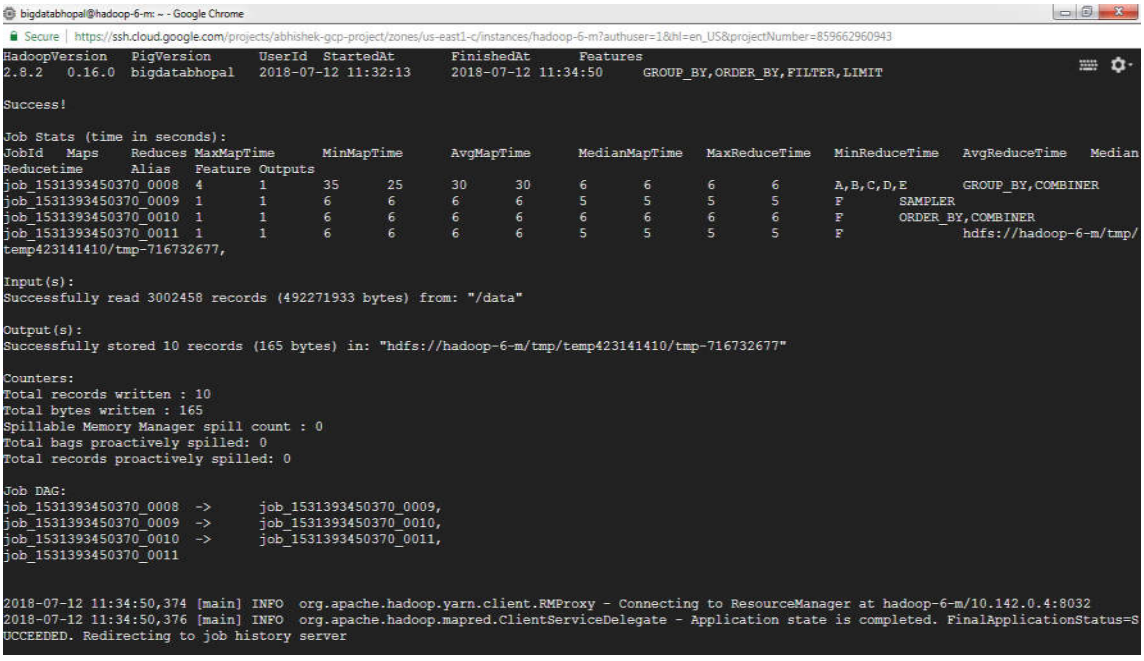

Figure 7. Time taken by pig

By comparing result of both the system , we can say that both the tools result are same means both are very accurate in terms of result but there is a difference between query execution time, the hive query takes less time as compared to pig script

Time Taken by (in sec)	Hive	Pig
Query -1	141.5	157.0
Query-2	131.0	161.0

Table 2. Time Taken by Hive & Pig

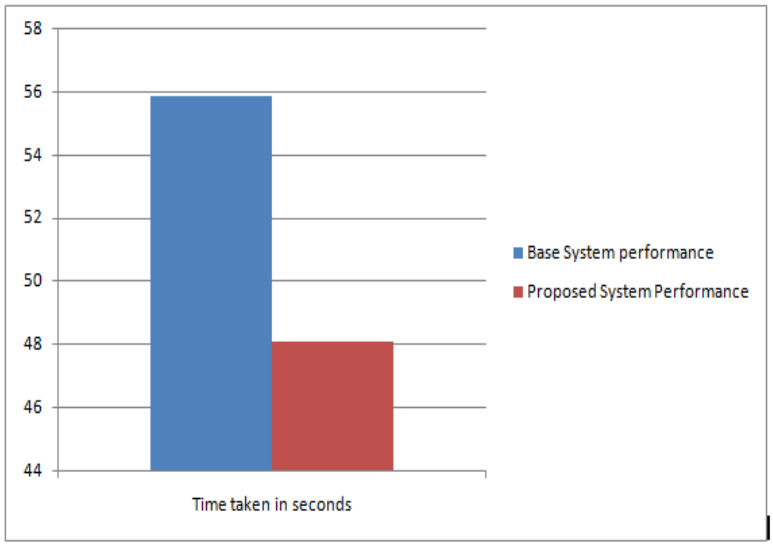

Figure 8. Time taken by Hive & Pig

VI CONCLUSION

The demand of interactive Big Data processing necessitated decoupling of data storage from analysis. The simple SQL queries of traditional relational database systems is still the most practical analyzing tool that people without programming background can also benefit from. As a result, Big Data SQL engines have been spun off in the Hadoop Ecosystem. So for which we can using apache hive which is an hadoop ecosystem and works same same as SQL so easily work on bigdata by using SQL (hive) on Hadoop (Storing and processing bigdata). In this we can create a hadoop cluster on cloud environment and integrate hive on top of the hadoop and analyze the data stored in HDFS by using hive query. In this we can also compare the performance of hive by another hadoop ecosystem called apache pig and we can say that hive query takes less execution time as compared to pig.

REFERENCES

[01]] Arushi Jain, Vishal Bhatnagar, "Crime Data Analysis Using Pig with Hadoop" in International Conference on Information Security & Privacy (ICISP2015), 11-12 December 2015, Nagpur, INDIA, in ELSEVIER 2015.

[02] Rahul Kumar Chawda, Dr. Ghanshyam Thakur, "Big Data and Advanced Analytics Tools", 2016 Symposium on Colossal Data Analysis and Networking (CDAN), IEEE 2016, ISSN: 978-1-5090-0669-4/16.

[03] Mrunal Sogodekar, Shikha Pandey, Isha Tupkari, Amit Manekar, "BIG DATA ANALYTICS: HADOOP AND TOOLS", in 978-1-5090-2730-9/16, 2016 IEEE

[04] Kusum Yadav, Manjusha Pandey, Siddharth Swarup Rautaray, "Feedback Analysis Using Big Data Tools" in IEEE 2016.

[05] Dave Jaffe "Three Approaches to Data Analysis with Hadoop"

[06] http://hadoop.apache.org/

[07] https://pig.apache.org/

[08] Aditya B. Patel, Manashvi Birla, Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce",6-8 Dec. 2012.

[09] Michael G. Noll, Applied Research, Big Data, Distributed Systems, Open Source, "Running Hadoop on Ubuntu Linux (Single-Node Cluster)", [online], available at http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/

[10] Sagiroglu, S., & Sinanc, D, "Big data: A review", IEEE International Conference on Collaboration Technologies and Systems (CTS), 2013, pp 42-47.

[11] https://hive.apache.org/

[12] Dave Jaffe "Three Approaches to Data Analysis with Hadoop"

[13] Jurmo Mehine, Satish Srirama, Pelle Jakovits "Large Scale Data Analysis Using Apache Pig"