Map Reduce, Pig and Hive on Climatic Condition

K. Govinda, A. Shubham Nair, Somula Ramasubbareddy

Abstract: ***This paper addresses recent growth within thes caleand kind of Earth science knowledge has provided new opportunities to massive knowledge analytics analysis for understanding the Earth's physical processes. There has been associate upsurge of natural science datasets within the past few decades that are being frequently collected mistreatment various modes of acquisition, at deferent scales of observation, and in numerous knowledge sorts and formats. Earth science knowledge sets but exhibit some distinctive characteristics (e.g. adherence to physical properties and spatiotemporal constraints), that gift challengestoancientdata-centric approaches.In this paper the comparative study of Hadoop's programming paradigm (Map reduce)andHadoop's ecosystems Hive and Pig.The processing time of map reduce, hive and pig is implement edona data set with simple queries. It is observed that Map reduce processes the datain shorter time ascompared with Map reduceand Hive. It is not necessary that only Map Reduce is useful other techniques are also useful under differentconstraints.***

Keywords: Data Pre-processing, Hadoop, Sqoop, Pig, Hive, Map reduce.

I. INTRODUCTION

In the last few years 90% of the world's data was generated. The invention of new devices, technologies and communication means like social networking sites, the data produced every year is increasing day by day. The data produced till 2003 was 5 billion giga bytes.This would cover up entire foot ball field if we pile up the whole data. In 2011 the identical data was created in every two days, and in every ten minutes in 2013. This rate keeps growing enormously. Though all this information produced is meaningful and when processed can be useful, it is beingn eglected. Big Data is a collection of large data sets that cannot be processe dusing traditional computing techniques. It involves many areas of business and technology and is not restricted to single technique or a tool.

In this paper the work is to analyze Climate data by using Hadooptool along with some Hadoopeco systems like Hdfs, Mapreduce, Sqoop, Hive and Pig. By using these tools one can process large amount of data, no data lost problem, high through put can be achieved, maintence costals overy less and it is a open source software, it is compatible on all the platforms since it is Java based.

Revised Manuscript Received on July 22, 2019.

K. Govinda, SCOPE,VIT Univeristy, Vellore, Tamilnadu.
A. Shubham Nair, SCOPE,VIT Univeristy, Vellore, Tamilnadu.
Somula Ramasubbareddy, Information Technology, VNRVJIET, Hyderabad, Telangana.

II. MAP REDUCE

Approximately 2009, Hadoop was evidenced to be respectable method meant for Big Data across conventional ware housing of data. Nevertheless, writing Hadoop code was extremely problematic. Individuals started lack of ease of query language. Consequently, Facebook [1] originate with the key identified as Hive. Almost, in the same time Yahoo [2] introduced Pig. Goal of Hive and Pig intended to carry easiness to the intricate code of Map Reduce. Moreover, Hive and Pig are an open-source key made on uppermost of Hadoop. In that Hive is known as data ware house which provisions questions such as SQL named as Hive QL, that are assembled to works of map reduce and that are performed by means of Hadoop. Pig is referred as data flow scenario, creates the program concerning Pig Latin. Contrast to Pig, a schematic representation is mandatory in Hive.

Y Smart intentions to offer a generic outline to interpret SQL queries to enhanced Map Reduce works, and performing them competently on huge scale disseminated cluster schemes. Y Smart[3] can combined in Hive, to create Hive achieve well, and might also remain an self-determining decoder.

SCOPE [4] is seriously inclined by SQL. This mechanism is intended for simple and effectual dispensation of enormous volumes of information kept in circulated, consecutive records. It gives proficient inquiry preparing usefulness. Execution of Hive debased in map reduce on account of expense of saving transitional outcomes. Thusly network bandwidth cost ends up advanced and numerous imitations are generally kept. Automatic Query Analyzer) [5], an inquiry enhancement strategy intended for MapReduce built warehouse frameworks. Specified a SQL-like inquiry.

Water creates a succession of MapReduce occupations, which limits the rate inquiry preparing Objective was ponder and break down different scheduling methods, which are essential to expand execution in Hadoop. Quincy Scheduler [6], is improved regarding appropriating work into a distributed information hub..

Lucene [7], appropriate for application that needs complete content ordering and seeking capacity. Distributed Lucene depends on a couple of Apache open source schemes, Lucene and Hadoop besides it might be supposed as the initial of its kind that concentrates around content-based ordering. Lucene, a unrestricted open source information retrieval programming library, initially made in Hadoop [8], which shrouds the profoundly

specialized subtleties of Hadoop.

The objective of [8] is to make an instrument that will make advancement of substantial scale processing of image and give understudy and analysts to make huge application

Figure 1:Proposed modular platform architecture generation effectively.

The mechanism used in [8] separating is stage which is earlier stage of mapping. lastly create image encoder and decoders that keep running behind. Now and again comparative numerous questions, normal tables, and join errands arrive at the same time, emerging numerous opportunities for calculation sharing regular occupations. Executing basic responsibilities only once can astoundingly diminish the absolute execution time of a cluster of inquiries. Likewise, Shared Hive[9] introduced to improve the general execution of Hadoop. SharedHive changes a lot of related HiveQL questions into another set of supplement inquiries that will create productions in a minimal execution time

III. PROPOSED METHOD

3.1 Data Preprocessing Module

In this module we have to create Data set for Weather dataset it contains a table with twenty cities each day temperatures subtleties for most recent 10 years and this information initially give in MySQL database help of this dataset we examination this task.

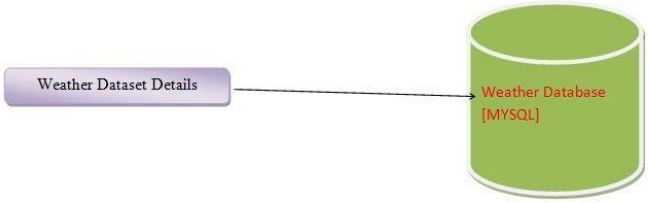

Figure 2:Working of Preprocessing Module

3.3 Data Analytic Module withHive

Hive is an information product house framework for Hadoop. Hive Query Language runs like SQL language and gets changed over into guide decrease programs. Information definition Language (DDL), Data Manipulation Language (DML) and client characterized capacities are bolstered by Hive. Facebook created Hive. In this module we need to examination the dataset utilizing HIVE apparatus which will be put away in Hadoop (HDFS).ForanalysisdatasetHIVEusingHQLLanguage.Using hiveweperformTablescreations, joins, Partition, Bucketing idea. Hive examination the main Structure Language. as Figure.4.

3.4 Data Analytic Module withPig

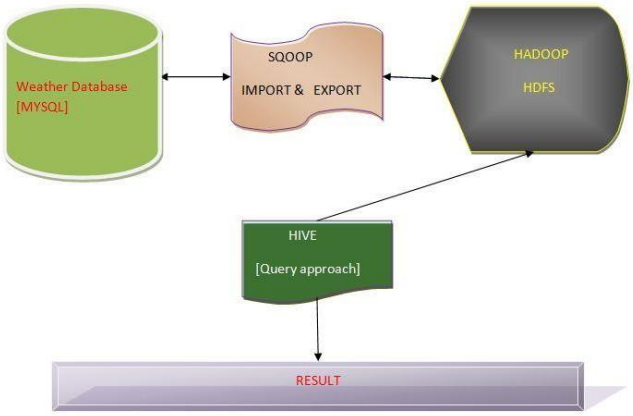

Figure 3:Working of Hive

Figure 4:Working of Hive

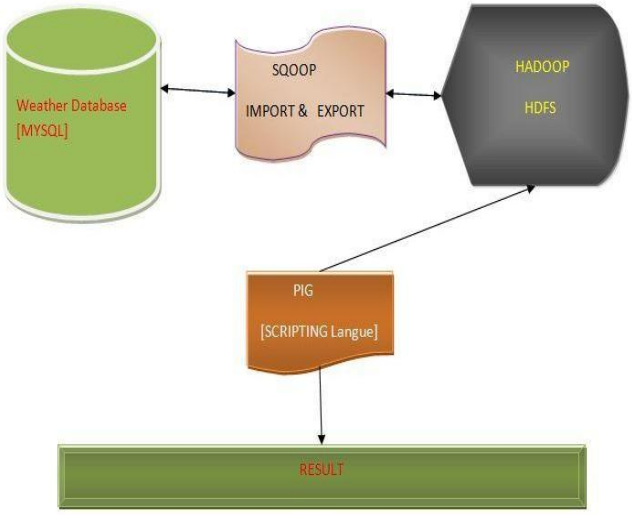

Figure 5:Working of Pig

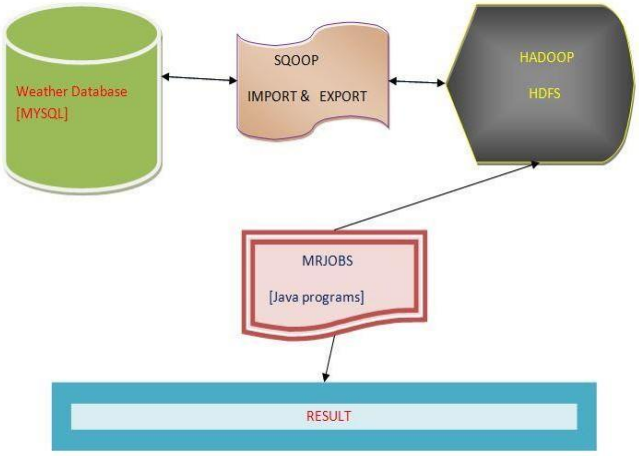

Figure 6:Working of Map Reduce

Implementation is achieved using map reduce program on the data. The map reduce program contains approximatelyaround300linesofcodetoperformthemappinga ndreducingtask.Thelanguageusedis Java. If the implementation is done without the use of Hadoop framework in a traditional manner, it becomes very difficult and the number of lines of code also increases apparently.

IV. IMPLEMENTATION

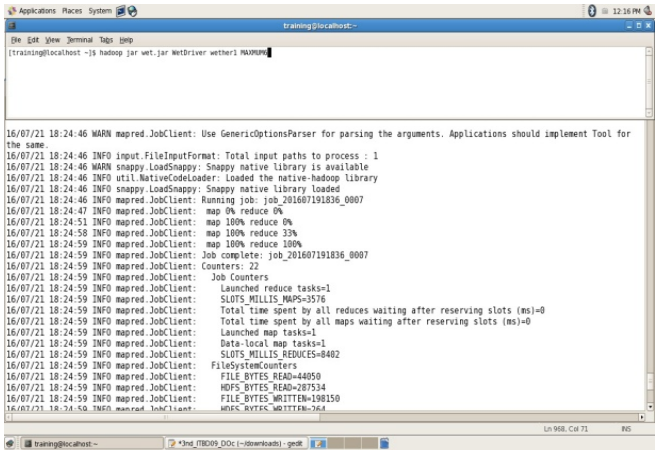

Figure 7(a):Output of Map reduce algorithm on weather datasets

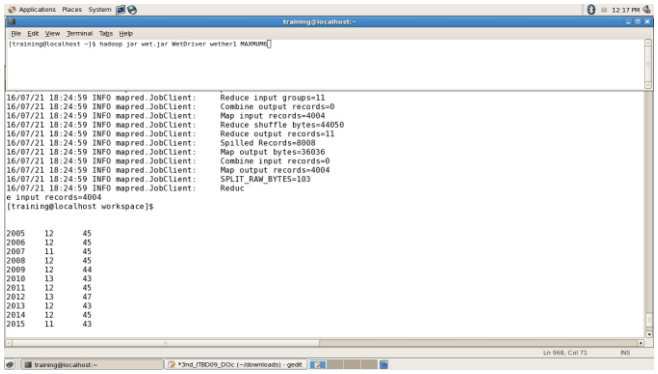

Figure 7(b):Output of Map reduce algorithm on weather datasets

The output of the map reduce program is shown in the above Figure 7(a) & 7(b). The time taken by the map reduce program using mapper and reducer is 14 secs.

4.1 Hive

The next tool that has been used to find the temperature of the location is Hive. The same concept is being implemented using Hadoop ecosystem known as Hive. It performs mapping and also runs ontop ofHadoopframework.Allthe300linesofcodeusedinmapreduce programcanbereducedusinghive to 10 lines of code resulting in the same output.

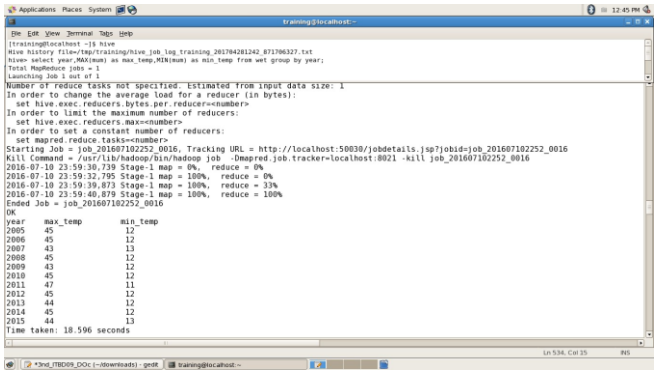

Figure 8:Output of Hive queries on weather data sets

The output of the program is shown in the above Figure.8. The time taken to execute the program in hive is 19 secs.

4.2 Pig

The final tool used to implement the prediction of future climatic condition sis Pig. The10 lines of code use dinb hive to calculate them axan dmin temperature of a city can be further reduce din to 5 lines of code using pig tool.

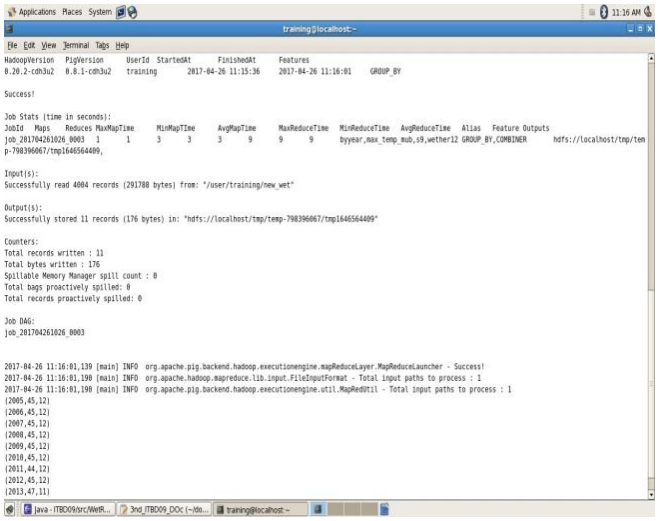

Figure 9:Output of Pig's grunt shell on weather data sets

The output of the program is shown in the above Figure.9. The time taken to execute the program in hive is 35 secs.

Hive and Pig are open source and were built as an alternative to Hadoop Map reduce so that the developers of Hadoop could do the same work in Java with lesser complexion by writing fewer lines of code. The code can be easily understood by the developers. For predicting the climatic conditions of a city, as we can see from the implementation portion, Map reduce gives faster and accurate results than Hive and Pig tool. But in various other situations, Pig or Hive may prove to be better than Map reduce program. There are various other reasons why it is useful to implement Map Reduce in predicting the weather conditions. They are described as under:

- Using virtualization and running two different operating systems in a single machine may need definite driver control which can be easily achieved using the Hadoop Mapreduce.
- .Sincethejobrequirestheuseofacustompartitionedtheref oredevelopersshouldutilizeHadoop Map Reduce.
- Due to the presence of pre-defined library of Java Mappers and Reducers, we should use Map Reduce.
- With large number of data sets there was a need for good amount of testability, so it is better to use Map Reduce than Hive or Pig.
- Since optimization is the requirement of the job, at a particular stage it can be done by using tricks same as that of mapper combining for which Hadoop Map Reduce is a goodoption.
- Although Pig and Hive have less lines of code compared to map reduce algorithm, but in terms of execution and producing the outputs faster(can be inferred from Figure.7(b),Figure.8 & Figure.9),map reduce algorithm is most suited for finding the weather condition of acity.

V. RESULTS

The proposed system uses the temperature datasets of 2005 to 2015 taken from UI repositories. The Process is known as data preprocessing. These records are stored in HDFS using Sqoop tool and Perform map reduce, hive and pig functions on it. The execution of these functions output are shown in fig respectively. The execution time taken by three of them proves that Map reduce algorithm produces results faster than Pig and Hive. The following is the graphical outcome result of weather from 2005-2015.

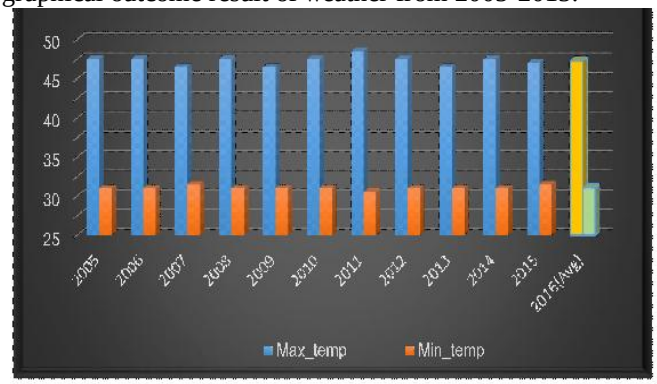

Figure 10: Graphical representation of Map reduce output (Mumbai Weather Forecast)

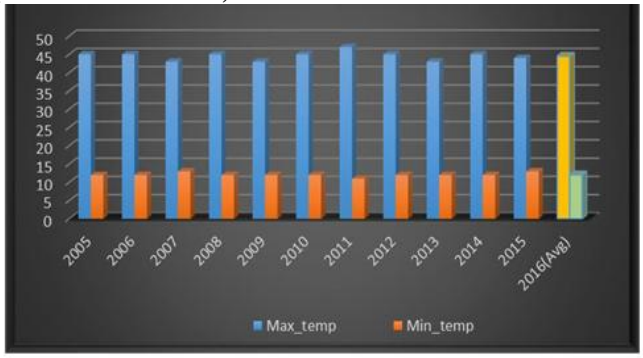

Figure 11: Graphical representation of Hive's output (Mumbai Weather Forecast)

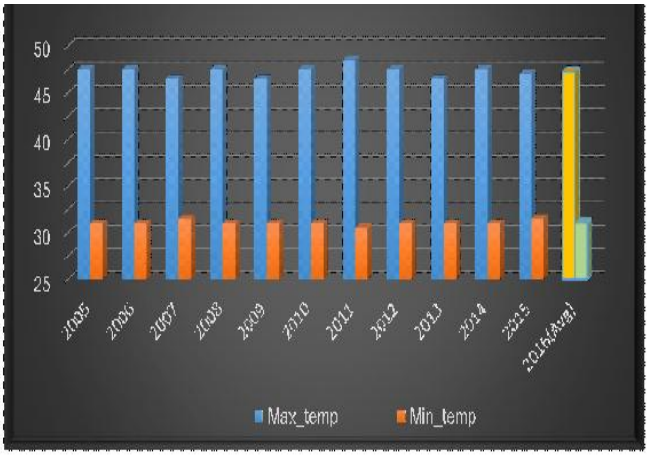

Figure 12: Graphical representation of Pig's output(Mumbai Weather Forecast)

The maximum and minimum temperature for the year 2016 was 44.5 & 12 (degree Celsius respectively). Map reduce algorithm's 2016 average maximum and minimum temperature came out to be 44.454 & 12(degree Celsius respectively, shown in Figure.10.).Hence we observe that map reduce algorithm produces accurate results as well.

VI. CONCLUSION

Using a large number of commodity computers, Map Reduce can be utilized as a framework for executinghighlyparallelizableanddistributablealgorithmsacr osshugedatasets.UsingMapreduce with Hadoop, the temperature can be analyses effectively. The scalability bottleneck is removed by using Hadoop with Map Reduce. Addition of more systems to the distributed network gives faster processing of the data. There are various ways for Hadoop to run the job. The three programming approaches that are Map Reduce, Hive and Pig are used. After performing practically we conclude that Map Reduce takes less time and produces accurate results as compared to both Pig and Hive. But all approaches have pros and cons so we have to choose according to our need and data.

REFERENCES

1) K.E. Taylor, R.J. Stouer, and G.A. Meehl, "An Over-perspective on CMIP5 and the Experiment Design," Bul-letin Am. Meteorological Soc., vol. 93, no. 4, 2012, pp.485–498.
2) Facebook Data Infrastructure Team(2010),"Hive–A Petabyte Scale Data Warehouse Using Hadoop", Facebook Data Infrastructure Team, http://infolab.stanford.edu/~ragho/hive-icde.pdf,2010. Yahoo Team,"The Pig Experience", VLDB,2009.
3) Facebook Data Infrastructure Team,"YSmart:Yet Another SQL-to-MapReduce Translat-or,"International Conference on Distributed Systems",2011.
4) Chaiken R,Jenkins B,Larson P,Ramsey B,Shakib D, Weaver S, Zhou J,"SCOPE: Easy And Efficient Parallel Processing of Massive Data Sets", research.microsoft.com/en-us/um/individuals/jrzhou/bar/Scope.pdf,2011
5) Wu S, Li F ,Mehrotra S,Chin Ooi,"Optimization for Massively Parallel Data Processing " ,2011.

6) Arora S and Goel M," Survey Paper on Scheduling in Hadoop, International Journal of Advanced Research in Computer Science and Software Engineering", Vol 4(5),2014.
7) Butler M and Rutherford J," Distributed Lucene: A dispersed free book list for Hadoop ", HP Laboratories,2008.
8) 7) Sweeney C, Liu L, Arietta S, Lawrence J,"HIPI: A Hadoop Image Processing Interfa-ce for Image-based mapreduce tasks",cs.ucsb.edu/~cmsweeney/papers/undergrad_thesis .pdf,2011.
8) Tansel Dokeroglu, Serkan Ozal1, Murat Ali Bayir, Muhammet Serkan Cinar and Ahmet Cosar (2014), "Improving the presentation of Hadoop Hive by sharing sweep and calculation undertakings", Accessed at: http://link.springer.com/article/10.1186%2Fs13677-014-0012-6#page-1