RESEARCH REVIEW JOURNALS
Volume-04
Issue-02
February -2019

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com [UGC Listed Journal]

Analysis of Climate Temperature Prediction Using Map reduce Hadoop and Hive

[1]ManpreetKaur and [2]Dr. Vijay Dhir

[1]Research Scholar MTech , Department of CSE Sant Baba Bhag Singh University, Jalandhar, India
[2]Professor of CSE Department, Sant Baba Bhag Singh University, Jalandhar, India

ARTICLE DETAILS

Article History
Published Online: 20 February 2019

Keywords
Big Data, Hadoop, Map reduce,HDFS, Hive

ABSTRACT

Now a day's big data increasing very fastly from various sources. It is a hottest topic in present days. Big data is a term that refers to a collection of huge amount of data that is very difficult to analyze, capture, manage and store into the traditional database. Big data is generated from numerous sources like internal and external sources. The generated data can be in any form like in structured from and unstructured from. To capture these types of data easily, the big data technologies and tools are used. The big data tools are hadoop framework, pig, hive, map reduce, hdfs and cassandra which manages the data into structured form easily. In this paper the tools of big data are described. To perform the analysis technique on the big data the mainly hadoop framework is used which includes many tools such as hdfs (Hadoop Distributed File System), map reduce and hive etc. In this research we use map reduce and hive tools to analyze the weather related data. Map reduce is a main component of hadoop framework which is used to perform various operations simultaneously. Second tool is hive which is used to perform database operations. It is a data warehouse tool which uses hiveql queries to process the data which is similar to the sql queries.

1. Introduction

Big data is the new topic of today that is increasing with speed. The technologies and initiatives are referred by the bigdata which involves the data to change fastly into conventional technologies, skills and infrastructure which can be addressed efficiently. It is a combination of large amount of data sets. Observing the size of the increasing data, big data came into existence. Big Data does not have to be big (peta/ exabyte). Even we can say that the 50 GB data is a big data when the data is too complicated to store into a normal database. Big data provides a platform for efficient handling of huge amount of data which is complex for traditional database or conventional methods to manage. Big data can be collected from so many sources and it has many forms. It is very difficult to analyze, manage and store the big data in traditional database thus hadoop framework is used to store and manage the big data[25].

1.1 Hadoop Framework

Hadoop is a tool or technology which is used to manage, analyze and store the big data. Hadoop was developed by computer scientists Doug Cutting and Mike Cafarella in 2006. Apache hadoop is freely available on internet which is written in programming language known as java. It is used for distributed processing and distributed storage of large set of data. Hadoop permits to process and store huge data in distributed environment via group of computers which use single programming model. This framework is made to measure the each processing from single server machine to several machines. Every machine has own storage device[34]. Many years ago users store and manage the data into relational database with the help of SQL queries. But now a days, data is in large sets. These large sets are not stored in the relational database that is why the hadoop framework was introduced which is capable to store and process those large sets of data. Mapreduce and HDFS are main components of hadoop[21].

1.2 HDFS

Hadoop has fault-tolerant storage system which is called Hadoop Distributed File System, or HDFS. It is distributed, scalable, portable file system written in java language for hadoop framework. HDFS is capable for storing large amount of data into it, process continuously and after the failure of crucial parts of storage the hdfs is survived without any loss of data[30]. Hadoop creates many set of nodes or clusters that interlinks with each other to know their active status and performance. A hadoop cluster has single name node with a cluster of datanodes. In hdfshadoop whenever any node is out of work or fails then the all work is shifted to the other nodes. These rest nodes are able to take the all responsibility of failed node with their backup. These nodes are known as secondary name node. Secondary namenode interacts perpetually with the primary name node and tells about description of the directory information of primary name node. The incoming files or data is broken into pieces by the nodes or clusters are called blocks, which hdfs manages to store into it. These blocks are stored smoothly in hdfs storage and group of servers. Every block is 128 MB. In hadoop1 the size of each block is 64 MB but now a days hadoop2 gives more block space than hadoop1 i.e. 128 MB. HDFS reserves at least three copies of each block or file and after copying each copy is sent to three different servers to save it[25].

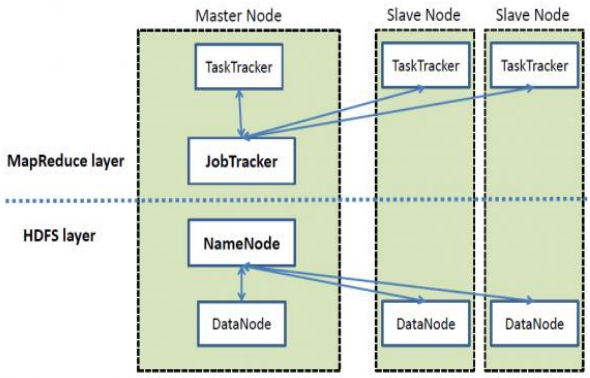

Fig. 1 Working of Hadoop

1.3 Map Reduce

In the Hadoop ecosystem the MapReduce framework is a backbone of parallel processing. This platform permits to operate a huge set of data. Firstly, the data and given problem is divided into various parts and combine the result of each part then run parallelly. It is more important tool of hadoop framework which is used to parallel processing. It works like a master slave architecture, in which one node works as a master node and other node works as slave node. For example, a vast sets of data are able to be divided into a smaller subgroups where process of analysis may be applied. In a concept of relational data warehouse the ETL process is applied on the transaction data to collect anything meaningful through the analysis process. These types of processes are written in hadoop by MapReduce[35].

2. Review of literature

More Priyanka et al.[1] Illustrated that the big data is a collection of large amount of sets of data which collects the meaningful data and valuable data from it to analyze, manage and store the data. This paper explains, the large set of weather data is extracted from centeralized storage and stored on Hadoop Distributed File System (HDFS) and then Map Reduce algorithm is applied on that data to analyze the minimum and maximum temperature. The weather data is load onto the hdfs (hadoop distributed file system) and mapreduce operations are applied on it. Reducer, reduces the result of given data and gives the output.

HassaniHossein et al.[3] Described that the big data is greatest weapon of industries. The main success of big data is analytical process which predicts the future outcome before any incident happen. This paper studies on climate change, what are the reasons for changing the climate and generates the result what is going tomorrow. To know the applications of weather and questions of weather related, firstly the previous years data is analyzed and then give the result why it happen. They covered almost 100 research papers and conclude that whenever big data predicts any future outcome for natural disaster and about agriculture then people aware for any incident before it happen.

Roja [4] Presents with the help of big data the changing in climate are stated that how the changing climate effects on agriculture. In this paper the reason of change in the climate and is predicted that how to get prepare from humidity, cold and hot weather. With the help of analysis of given data it can be decided whether the global warming is increasing or decreasing. With the use of map reduce data is examined which is stored in NCDC. Last few years data collected from the satellite and send to the base station and got the results that which is the hottest and coldest day in which year or in which month. This all processing is done by the big data technology.

ChoukseyPriyanka et al.[5] Introduced that every year many countries generate the data regarding weather. To store and manage this data is very challengable itself. In this paper the parameters of weather like wind, pressure, hot, cold and maximum are calculated and average is determined. Map reduce and Spark tools are used to analyze the weather data. The results of spark tool are efficient and better than mapreduce. In this research the performance of spark is better and on the basis of their result the weather data is predicted.

KaurSimranjot et al.[7] Quoted that entire world is upset from the changing of the weather. Data mining tools are used in this paper which manage the weather related data. As in agriculture field farmers have to face a lot of problems regarding unpredicted weather with it crops are damaged. With the help of this technique they know about the weather and aware for their crop germination. Weather forecasting is dependent on the molecules which are present in the air like carbon dioxide, nitrogen dioxide and ozone etc. On this collected data incremental K means clustering algorithm is used in which new data values are examined that what are the main causes of changing the values. She introduced an algorithm in which data is collected after every hour and saved in original database. The previous data is converted in structured data by using R tool after every two hours and stored in “Structural air pollution database” (SAPD).

Doreswamy et al.[8] Stated that now a days store and manage the huge amount of data is very big challenge. Data mining is the useful technique to get the big data from various sources. Weather effects on many sectors like air traffic, agriculture and tourism. This paper explains a system or model that uses the previous or historical data to analyze the large amount of weather data with the help of hadoop framework and map reduce so the weather prediction has been done. Weather forecasting has various areas like temperature, cold, wind, hot, rain and thunder storm. This paper also explains big data characteristics, hadoop working and overall map reduce working. In future model uses the neural network for advancement.

3. Proposed Methodology

The Proposed System which uses data set of National Climatic Data Centers(NCDC) which provides the parameters as maximum temperature, minimum temperature and hot or cold months of the year. National Climatic Data Centers which contains huge historical weather datasets. Firstly, the files of weather data is downloaded from NCDC. These files are loaded into Hadoop Distributed File System that is HDFS with the help of hadoop commands. After loading the data files on the hdfs, the files are divided into equal parts or groups to analyze. To analyze these data files the map reduce commands are applied on it. Map reduce consists of

two modules map and reduce. First of all the input is sent to the map phase which contains their key and value. Map is a input phase of map reduce that is used for mapping. The files are mapped with their related groups which have similar keys and values. The output of mapper phase is a collection of key and value where key contains the date of month and year. Mapper is used divide and conquer method to generate the output. mapper's output is integrated and arranged by the key and value. After done this processing the results are sent to the reducers. Second module is a reducer which is known as the output phase of the map reduce. Reducer reduces the valuable results which are generated from the mapper. The working of every reducer is to calculate the maximum and minimum temperature on the basis of monthly and yearly. Then reducers store the final results in HDFS.

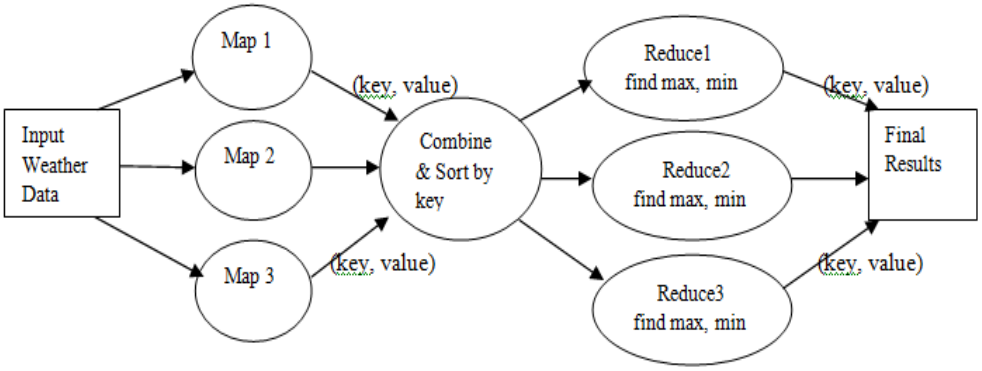

Fig 2. Map-Reduce Model

In this research work we use other tool is hive which uses hiveql queries to analyze the data. It stores the data into data warehouse. Hive tool is faster than the map reduce framework. Map reduce is used to perform the operations simultaneously so that it may take time less. In relational database the sql queries are used to analyze the data but in hadoop we use hive tool to perform transactional operations because relational database is not capable to store these large files into it. Hiveql queries are similar to sql queries that extracted the valuable data from the tables that are stored into a database.

4. Experimental Results

Previous 10 year weather related data is collected from the National Climatic Data Center. The data is loaded on hdfs and after that the data is extracted from that storage to analysis. We apply map reduce commands on that data to get the analysis result from previous years. We perform both map reduce and hive tools on that data to check their efficiency, performance and time. In below figure the y axis represents the temperature and x axis represents the years in which the temperature was read. Variation in this graph is according to previous year temperature. We analyze the weather data on the basis of month wise and year wise that which day has maximum temperature and minimum temperature. This analysis framework also provides hot and cold days of the year.

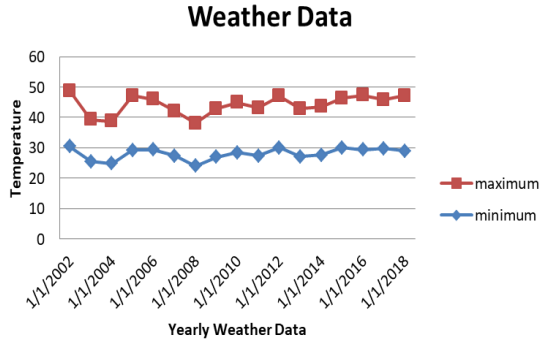

Fig 3. Result of yearly weather data

Fig 3.shows the output of yearly weather data. It is also generated from the hive technology. The input files of year wise weather data are stored into the hive data warehouse to analyze the result. X axis represents the date with their year and Y axis represents the temperature of that data. This is the hiveql query to perform the analysis on the yearly data. This gives result in the from of previous stored data from the month January. Because we enter the month is 1 from the month table. So the data will be analyzed of January month from previous year weather data which is stored in the data warehouse.

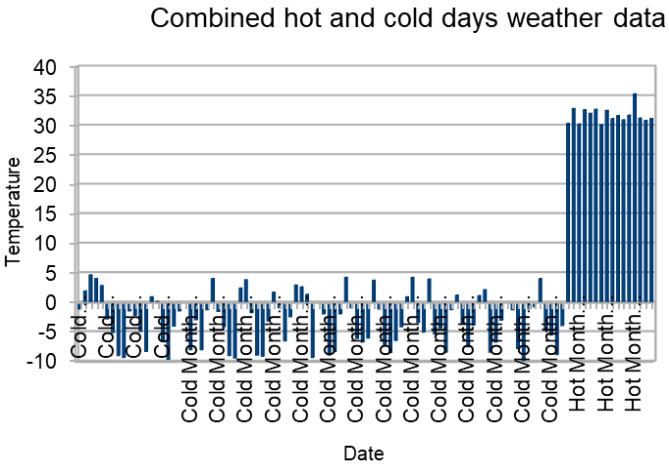

Fig 4. Result of hot and cold days of the year

Fig 4.shows the output of the hot and cold months of the previous years. This output is collect from the map reducer. This presents all hot and cold months in graph representation. X axis represents the dates with their year and Y axis represents the temperature of the year.

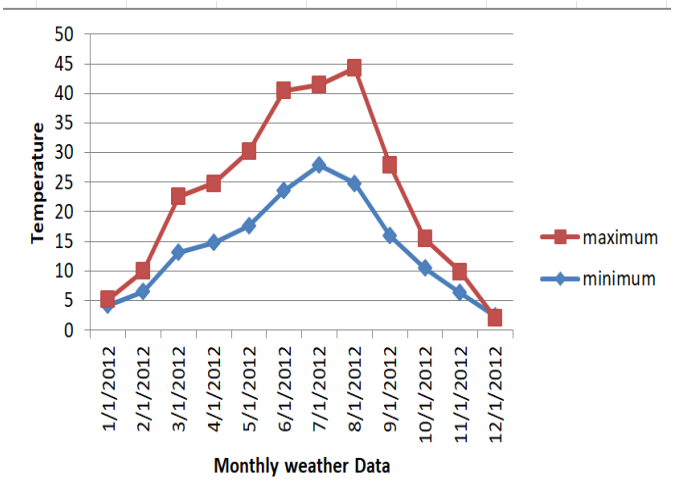

Fig 5. Result of month wise weather data

Fig 5.presents the output of the month wise weather data. This output is taken from the hive tool which is faster than the map reduce technology. This shows the maximum and minimum temperature of the monthly weather data from the 2012 year. The hive queries are performed on monthly data to collect the result from that data. Big data technology is very useful for every filed to analyze the data which we can't store and analyze in the traditional database.

5. Conclusion and Future Scope

We are in the era of big data where there is a need to introduce the new tools or technologies to process and analyze the data in efficient manner. Because the traditional database is not capable to manage and analyze the massive data which has many forms. In this paper we have discussed various tools of big data which are useful for analyzing the big data. New technologies are still under development, the current technologies are providing satisfactory result. We use two technologies to analyze the weather data, map reduce and hive technology in this research to compare their results on the basis of their efficiency, time and performance. We perform map reduce commands on collected data to generate the result about weather data. In this research we also use hive platform which contains hiveql queries to process the data. These results are collected from these two tools and compare them. With the help of this research you can analyze the maximum temperature, minimum temperature, hot days and cold days. This gives the result on monthly basis, yearly basis and daily basis.

Analytics of weather data can be considered for the future work. With the help of combination of big data and other technologies such as neural network, natural language processing, statistics and artificial intelligence the analytics framework may be developed. By using these technologies the weather may be predicted in advance and the people may be aware before happening any incident about the weather like heavy rain, thunder storm and snow. Analytics process may give advantages for future in various fields such as marketing, business, medical and agriculture.

References

[1] Priyanka More, Sunita Nandgave&MeghaKadam, "Climate Change Detection using Hadoop with MapReduce", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 7, Issue 3, Pages 1925-1932.

[2] Jesus Camacho, AshutoshChauhan& Alan Gates(2019)"Apache Hive: From MapReduce to Enterprise-grade Big Data Warehousing", International Conference on Management of Data.

[3] HosseinHassani, Xu Huang and Emmanuel Silva(2019), "A Review on Big Data and Climate Change", Big Data Cognitive Computing, Pages 1-17.

[4] Roja (2018), "Big-Data Analysis For Weather Forecasting and Predcition for Scalability Reports", International Journal of Engineering and Techniques, Vol.4, Issue 4.

[5] PriyankaChouksey&Chauhan (2017), "Weather Data Analytics using MapReduce and Spark", International Journal of Advanced Research in Computer and Communication Engineering, Vol.6, Issue 2.

[6] JyotiKumari&Surender (2016), "Statically Analysis on Big Data Using Hadoop", International Journal of Computer Science and Mobile Computing, Volume 6, Issue. 6, Pages 259 – 265.

[7] SimranjotKaur, Cheema (2017), "Big Data and Analysis Of Weather Forecasting System", International Journal of Advanced Research in Computer Science, Vol.8, No.7.

[8] DoreswamyAnd Ibrahim Gad (2016), "Big Data Techniques: Hadoop And Map Reduce For Weather Forecasting", International Journal Of Latest Trends In Engineering And Technology, Pages 194-199.

[9] SandeepIddalgave, M.Tejashree, G.Nikhil&DattuGoud, "Analyzing Natural Calamities Using Apache Hive", International Journal Of Research In Electronics And Computer Engineering, vol .6 issue 2, pages 820-823.

[10] Prakash Antony & Aloysius (2018)," Architecture Design for Hadoop No-SQL and Hive", International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Vol.3.

[11] Akshay More, Rathod, Patil&Sarode (2018), "Stock Market Prediction System using Hadoop",International Journal of Engineering Science and Computing, Vol. 8, No.3.

[12] PriyaParhate, GauravGhogle, JyotiBhange&Ashwini Ingle, "Review Paper on Big Data:Challenges And Applications", International Research Journal of Engineering And Technology, Volume 04 Issue 01, Pages 114-118.

[13] Riyaz P.A., Surekha Mariam Varghese (2015), "Leveraging Map Reduce With Hadoop for Weather Data Analytics" IOSR Journal of Computer Engineering, Volume 17 Issue 3, Pages 06-12.

[14] MeenaAgrawal, A.K.Pandey&C.PAgrawal (2017), "A Hadoop based Weather Prediction Model for Classification of Weather Data",IEEE.

[15] PrashantSahatiya(2018), "Big Data Analytics on Social Media Data: A Literature Review", International Research Journal of Engineering and Technology, Vol.5, Issue 2

[16] RaswithaBandi, T.Shravani Reddy, K.Nikhila&Mohd Abdul Javeed (2018), "Analyze Stock Data Using Apache Hive", International Journal Of Research In Electronics And Computer Engineering, Vol. 6 Issue 2, Pages 744-746.

[17] N. Deshai, S. Venkataramana& G. P.Varma, [2018] "Research Paper on Big Data Hadoop Map Reduce Job Scheduling", International Journal of Innovative Research in Computer and Communication Engineering, Volume 6, Special Issue 1, pages 103-114.

[18] K.Tamilselvi, V.Sumithra&Dhanapriyadharsini (2018), "Big Data Analytics Using Hadoop Technology", International Research Journal of Engineering and Technology, Volume 05 Issue 01, Pages 1507-1511.

[19] JakubKudlacek (2015), "Big data analytics for mobile networks", International Journal Of Advanced Research In Computer And Communication Engineering.

[20] VenkateshNaganathan, "Comparative Analysis of Big Data, Big Data Analytics: Challenges and Trends", International Research Journal of Engineering and Technology, Volume: 05 Issue: 05, Pages 1948-1964.

[21] Mr. Sunil Navadia (2017), "Weather Prediction: A novel approach for measuring and analyzing weather data", IEEE International conference on I-SMAC-2017

[22] Annapoorani&Srinivasan (2018), "Improving Performance of Data in Hadoop Clusters using Dynamic Data Replica Placement: A Survey", International Journal Of Engineering Sciences & Research Technology, pp. 153-156.

[23] AthulaBalachandran (2014), "Large Scale Data Analytics of User Behavior for Improving Content Delivery, IEEE Conference On Open Systems.

[24] P. Chandrashaker Reddy & A. Suresh Babu (2017), "Survey on Weather Prediction using Big Data Analytics, International Conference on High Performance Computing Workshops IEEE.

[25] ManpreetKaur& Vijay Dhir (2018), "A Survey on Big Data and its Applications using Hadoop", Journal
of Emerging Technologies and Innovative Research, Vol.5, Issue 9.

[26] MashooqueMemon, Soomro, Jumani&Kartio(2017), "Big Data Analytics and Its Applications", International Journal of Innovative Research in Computer and Communication Engineering, Vol.6, Issue1.

[27] Shams Uddin Al Azad (2017), "Big Data Analytics: Performance Analysis of NoSQL Databases and Hadoop Ecosystem", International Journal Of Latest Trends In Engineering And Technology.

[28] Yulan Liang &Arpad Kelemen (2016),"Big Data Science and Its Applications in Health and Medical Research: Challenges and Opportunities", Journal of Biometrics & Biostatistics, Volume 7, Issue 3

[29] Himanshi Jain &Raksha Jain, "Big Data in Weather forecasting: Applications and Challenges", International Conference On Big Data Analytics and computational Intelligence, pages 138-142.

[30] K.Anusha, K.Usha Rani (2017), "Big Data Techniques for Efficient Storage and Processing of Weather Data", International Journal for Research in Applied Science &EngineeringTechnology(IJRASET), Volume 5 Issue 7.

[31] VeershettyDagade, Mahesh Lagali,SupriyaAvadhani and PriyaKalekar (2015), "Big Data Weather Analytics Using Hadoop", International Journal of Emerging Technology in Computer Science & Electronics, Volume 14, Issue 02.

[32] AleksandarVelinov&ZoranZdravev (2018), "Analysis of Apache Logs Using Hadoop and Hive", TEM Journal, Vol.7, Issue 3, Pages 645-650.

[33] M. R. Bendre& R.C. Thool (2015), "Big Data in Pricision Agriculture: Weather Forcasting for Future Farming", International Conference on Next Generation Computing Technologies, pages 744–750.

[34] Big Data Analytics, www.tutorialpoint.com.

[35] IqbaldeepKaur, NavneetKaur et al. (2016), "Research Paper on Big Data and Hadoop", International Journal of Computer Science And Technology, Vol. 7, Issue 4, pages 50-53.

[36] Kelvin KF Tsoi, Simon K Poon, Patrick Hung (2020), "Minitrack for Big-Data on Healthcare Application" , Proceedings of the 53rd Hawaii International Conference on System Sciences.

[37] N. Krishnaveni, A. Padma (2020), "Weather forecast prediction and analysis using sprint algorithm", Journal of Ambient Intelligence and Humanized Computing.

[38] Chong Chong Qi (2020), "Big Data Management in the Mining Industry", International Journal of Minerals, Metallurgy and Materials, Pg, 131- 139.

[39] Zhaohao Sun, Kenneth David Strang (2020), Francisca Pambel, "Privacy and security in the big data paradigm", Journal of Computer Information Systems, Volume 60 Issue 2, Pg. 146-155.

