Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.199

IJCSMC, Vol. 8, Issue. 5, May 2019, pg.252 – 258

Big Data Prediction Framework for Weather Temperature Based on Map-Reduce Algorithm

Abhishek Kumar; Guide Name: **Savita Shivani**
Jaipur National University, Jaipur, Rajasthan (India)
innovativeabhikumar@gmail.com
Address: Agra-Jaipur RD, Near New RTO Office, Jagatpura, Jaipur, Rajasthan (India)

Abstract- Weather is the most critical for human in many aspects of life. The study and knowledge of how weather Temperature evolves over time in some location or country in the world can be beneficial for several purposes. Processing, Collecting and storing of huge amounts of weather data is necessary for accurate prediction of weather. Meteorological departments use different types of sensors such as temperature, humidity etc. to get the data. The sensors volume and velocity of data in each of the sensor make the data processing time consuming and complex. This project aims to build analytical Big Data prediction framework for weather temperature based on MapReduce algorithm. Information Mining Package, can perform administered grouping methodology on immense measures of information, normally alluded as large information, on a conveyed framework utilizing Hadoop MapReduce. The instrument has arrangement calculations actualized, taken from Support Vector Machines (SVM). The aftereffects of an exploratory examination utilizing a SVM classifier on informational collections of various sizes for various bunch designs like K-Means shows the capability of the apparatus, just as perspectives that influence its execution.

Index Terms- Map Reduce algorithm, SVM, K-Means algorithm

INTRODUCTION:

Big Data is the procedure of look at extensive informational collections containing assortment of information types. The enormous information keeps up the colossal measure of information and procedure them. It is customary information examination; it can process the organized information, however not unstructured information. In huge information it can process both organized and unstructured information. Huge information generally incorporates informational collections with sizes past the capacity of regularly utilized programming devices to catch, clergyman, oversee and process the information. Enormous information estimate ranges from terabytes to numerous petabytes of information. Climate expectation is the use of innovation to anticipate the activity of the air for a given area. It is significant predominantly for business agriculturist, ranchers, fiascos the board and so forth climate expectation is one of the most intriguing and entrancing area and plays critical job in meteorology. There are a few confinements in better execution of climate estimating for instance in information mining systems; it can't anticipate climate present moment productively. They utilized little constrained territories for climate determining. It is troublesome undertaking to anticipate climate because of dynamic changes in the environment. Environmental change has been looking for a ton of consideration since long time. The adversarial impact of this atmosphere is being felt in all aspects of the earth. There are numerous precedents for these, for example, ocean levels are rising, less precipitation, increment in mugginess. The propose framework defeats the a few issues that happened by utilizing different strategies. In this task we utilize the idea of Big information Hadoop. In the proposed design we can process disconnected information, which is put away in the National Climatic Data Center (NCDC). Through this we can discover the greatest temperature and least temperature of year, what's more, ready to anticipate the future climate figure. At last, we plot the diagram for the got MAX and MIN temperature for every moth of the specific year to envision the temperature. In view of the earlier year information climate information of coming year is anticipated.

Related Work:-

[1] G. Nancy, G. David, A. Troy, "Big Data Effects on Weather and Climate", Informal Discussions on The New Economics. s.l. : SAIC, 2014.

Weather forecasting plays a vital role in daily routine, businesses and their decisions. The process of weather forecasting is developing as the effect of advancement in technology right from the realization of increasing size of data, Weather forecasting was found to be based on big data. The researchers have taken review with the objective to study the current forecasting process and methods, and the need of a data structure is recognized for handling the weather data, which is bigger in size, used for the process of weather forecasting. This paper presents a big data analysis framework for weather dataset based on MapReduce Algorithm, and offers not only weather dataset analysis, but also various analytic capabilities on huge amounts of data. However, this work establishes a guideline for researchers and industrial practitioners on how to analysis big data.

[2] V. Dagade, M. Lagali, S. Avadhani, P. Kalekar, "Big Data Weather Analytics Using Hadoop", International Journal of Emerging.

We want to build a platform that is extremely flexible and scalable to be able to analyze penta bytes of data across an extremely wide increasing wealth of weather variables. Here 8in this paper we are working on data analysis using Apache Hadoop and Apache Spark . We are performing experiments to decide the best tools among Hadoop using Pig and Hive Queries .And also we are comparing their performance based on pseudo node and Hadoop Distributed Multi node cluster.

[3] Madden, "From Databases to Big Data", Internet Computing, IEEE, 16(3), pp. 4 - 6. 2012. While technologies to build and run big data projects have started to mature and proliferate over the last couple of years, exploiting all potentials of big data is still at a relatively early stage. In fact, Big data is term refer to huge data sets, have high Velocity , high Volume and high Variety and complex structure with the difficulties of management , analyzing, storing and processing .Due to characteristic of big data it becomes very difficult to Management, analysis, Storage, Transport and processing the data using the existing traditional techniques. This paper introduces Big Data Analysis and storage. First we presents the Big data technology alongside it's the significance of big data in the modern world and venture existing which are successful and essential in changing the idea of science into huge science and society as well. Following that, we present How Fast Data is Increasing and The Importance of Big Data. In addition, we discuss Big Data Technologies include (Big Data Frameworks and Platforms and Databases for Big Data). Moreover, we discuss Data Storage and Big Data Management and Storage. Then, we present Big Data Analysis and Management include (Big Data with Data Mining, Big Data over Cloud Computing and Hadoop Distributed File System (HDFS) and MapReduce). Furthermore, we also discuss big data modeling and big data security issues. Finally Conclusion and Future work.

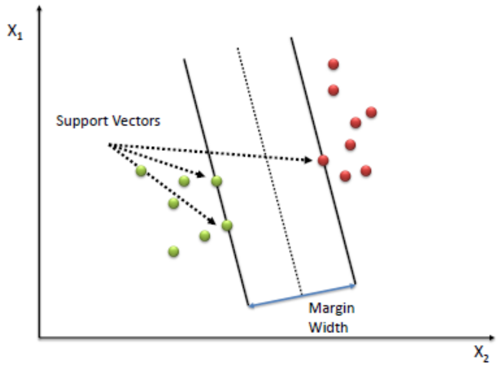
[4] Bernice Purcell The emergence of "big data" technology and analytics Holy Family University 12129pdf.

The Internet has made new sources of vast amount of data available to business executives. Big data is comprised of datasets too large to be handled by traditional database systems. To remain competitive business executives need to adopt the new technologies and techniques emerging due to big data. Big data includes structured data, semi-structured and unstructured data. Structured data are those data formatted for use in a database management system. Semi-structured and unstructured data include all types of unformatted data including multimedia and social media content. Big data are also provided by myriad hardware objects, including sensors and actuators embedded in physical objects, which are termed the Internet of Things. Data storage techniques used for big data include multiple clustered network-attached storage (NAS) and object-based storage. Clustered NAS employs storage devices attached to a network. Groups of storage devices attached to different networks are then clustered together. Object-based storage systems distribute sets of objects over a distributed storage system. Hadoop, used to process unstructured and semi-structured big data, uses the map-reduce paradigm to locate all relevant data then select only the data directly answering the query. No-SQL, MongoDB, and Terra-Store process structured big data. No-SQL data is characterized by being basically available, soft state (changeable), and eventually consistent. Mongo-DB and Terra-Store are both NoSQL-related products used for document-oriented applications. The advent of the age of big data poses opportunities and challenges for businesses. Previously unavailable forms of data can now be saved, retrieved, and processed. However, changes to hardware, software, and data processing techniques are necessary to employ this new paradigm.

Proposed methodology - In this proposed system we use Data Mining techniques as important methods which could be used for weather forecasting with Big Data. However, these techniques have been designed to handle data of comparatively smaller sizes as opposed to the size of Big Data. The architecture of the analytics needs to be redesigned so that it could handle historical data to forecasting. A detailed evaluation of challenges associated with the application of Data Mining techniques to Big Data. For data clustering we will use K-means and for data prediction using Support vector machine.

Support Vector Machine - Classification (SVM)

A Support Vector Machine (SVM) performs classification by finding the hyperplane that maximizes the margin between the two classes. The vectors (cases) that define the hyperplane are the support vectors.



Algorithm:

1. Define an optimal hyperplane: maximize margin
2. Extend the above definition for non-linearly separable problems: have a penalty term for misclassifications.
3. Map data to high dimensional space where it is easier to classify with linear decision.

surfaces: reformulate problem so that data is mapped implicitly to this space

Hadoop

Hadoop Is a stage that offers a productive and compelling technique for putting away and handling gigantic measures of information Unlike conventional contributions, Hadoop was structured and developed from the beginning location the prerequisites and difficulties of enormous information. Hadoop is incredible in its capacity to enable organizations to quit agonizing over structure enormous information competent framework and to concentrate on the main thing: removing business esteem from the information. Apache Hadoop use cases are many, and appear in numerous businesses, including: hazard, extortion and portfolio investigation in money related administrations; conduct examination and personalization in retail; interpersonal organization, relationship and conclusion examination for promoting; medicate communication demonstrating and genome information preparing in social insurance and life sciences, etc to give some examples. What's more, numerous organizations give Hadoop business execution and additionally support, including Cloudera, IBM, MapR, EMC, and Oracle. As per the Gartner Research, Big Data Analytics is a drifting subject in 2014. Hadoop is an open structure for the most part utilized for Big Data Analytics.

MapReduce is a programming worldview related with the Hadoop and have two isolated and unmistakable assignments. First is the guide work, which takes a lot of information and changes over it into another arrangement of information, where singular components are separated into (key/esteem) sets. The diminish work accepts the yield of guide as info and Shuffle/sort it and decrease the (key/esteem). As the succession of the name MapReduce the decrease work is constantly performed after the guide work. Putting the Map and Reduce capacities to work proficiently requires a calculation as well. The standard strides for a MapReduce work process. MapReduce utilizing to handling the information that putting away in Data Node. Hadoop disseminated record framework (HDFS) usage of an appropriated file system is intended to hold an immense measure of information, and give access to this information to numerous customers circulated over a system.

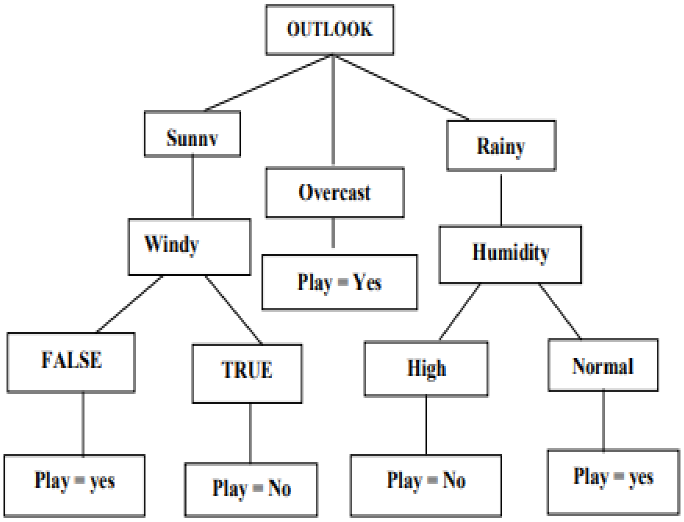
Clustering mining optimization algorithm based on MapReduce

The data set processed by MapReduce should have such characteristics: It can be broken down into many small data sets, and each small data set can be completely parallel processed. The process of K-means algorithm based on Hadoop mainly has two parts, the first part is to initial clustering centres, and divide the sample data set into a certain size of data blocks for parallel processing. The second part is to start the Map and Reduce tasks for parallel processing of algorithm in time, until process gets the clustering results.

The initial clustering centres of traditional algorithm selected randomly, will cause the instability of clustering results. This paper adopts a method of the initial clustering centre selection to improve the stability of the results. Optimized K-means clustering algorithm firstly choose k samples to initialize clustering centres according to certain algorithmic rules, then k clustering centres are stored in a file on the HDFS as a global variable.

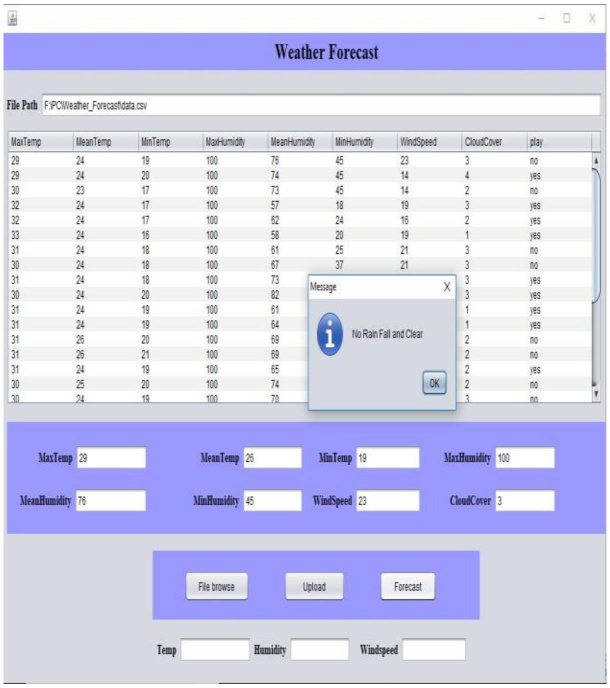
Simulation & Result

The following Figure 1 explains the overview of weather prediction process. Weather prediction mainly based on weather conditions. (Figure 1) outlook label have three seasons like Sunny, Overcast and Rainy. And also windy and humidity level also checked this process.



The following table explains the comparison of existing and proposed system. K-means, Hive and SVM algorithm produced a result 92.23%, KNN+ K- Means produced 89.1% K-Means+ Naive Bayes 88.32% result on existing system.

Algorithm	Accuracy
Hive+K-Means+SVM (Proposed)	92.23%
K- Means + KNN (Existing)	89.1%
K-Means+ Naive Bayes (Existing)	88.2%



File Path F:\PC\Weather_Forecast\data.csv

MaxTemp	MeanTemp	MinTemp	MaxHumidity	MeanHumidity	MinHumidity	WindSpeed	CloudCover	play
29	24	19	100	76	45	23	3	no
29	24	20	100	74	45	14	4	yes
30	23	17	100	73	45	14	2	no
32	24	17	100	57	18	19	3	yes
32	24	17	100	62	24	16	2	yes
33	24	16	100	58	20	19	1	yes
31	24	18	100	61	25	21	3	no
30	24	18	100	67	37	21	3	no
31	24	18	100	73			3	yes
30	24	20	100	82			3	yes
31	24	19	100	61			1	yes
31	24	19	100	64			1	yes
31	26	20	100	69			2	no
31	26	21	100	69			2	no
31	24	19	100	65			2	yes
30	25	20	100	74			2	no
30	24	19	100	70			3	no

Message

No Rain Fall and Clear

OK

MaxTemp 29 MeanTemp 26 MinTemp 19 MaxHumidity 100

MeanHumidity 76 MinHumidity 45 WindSpeed 23 CloudCover 3

Temp Humidity Windspeed

CONCLUSION:

In this paper have proposed climate forecast utilizing enormous information condition. The strategy utilized in our undertaking is Hadoop with guide lessens to break down the sensor information, which is put away in the National Climatic Data Center (NCDC) is a productive arrangement. Guide diminish is outline work for exceptionally parallel and conveyed frameworks crosswise over gigantic dataset. It is utilized to examine for the given information and anticipate expected yield to our venture. By utilizing map lessen with Hadoop and java helps in evacuating adaptability bottleneck. For information bunching we will utilize K-Means and for information expectation utilizing Support vector machine. This kind of innovation used to dissect huge informational indexes can possibly extraordinary upgrade to climate estimate. Subsequently we anticipate the future climate estimate, least and most extreme temperature, hot days and cold days dependent on the information got from the NCDC. This encourages for the general population to preplanning for outside occasions dependent on the climate conditions.

References

[1] G. Nancy, G. David, A. Troy, "Big Data Effects on Weather and Climate", Informal Discussions on The New Economics. s.l. : SAIC, 2014.

[2] V. Dagade, M. Lagali, S. Avadhani, P. Kalekar, "Big Data Weather Analytics Using Hadoop", International Journal of Emerging.

[3] Madden, "From Databases to Big Data", Internet Computing, IEEE, 16(3), pp. 4 - 6. 2012.

[4] Bernice Purcell The emergence of "big data" technology and analytics Holy Family University 12129pdf.

[5] V. Borkar, M. Carey, "Inside big data management: Ogres, onions, or parfaits?". in Proc. 15th Int. Conf. Extending Database Technol. pp. 3_14, 2012.