International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 5, Issue 3, May-June 2016 **ISSN 2278-6856**

Data Analytics Adopting K Nearest Neighbor Techniques For Big Data

Tejashree S[1], Swathi Y[2]

[1]Computer Science and Engineering, CMR Institute of Technology,
132, AECS Layout, IT Park Road, Kundalahalli, Bangalore 560037

[2]HOD & Associate Professor, Computer Science and Engineering, CMR Institute of Technology,
132, AECS Layout, IT Park Road, Kundalahalli, Bangalore 560037

Abstract

The great deal of interest has risen recently in the field of big data and the analysis has increases and this data is driven from the extensive number of the research challenges related to big and strong bonafide application such as modelling the data, processing the data and distributing large scale of the repository data. Lots of useful data is loss because of improper handle and storage of data. Handling of huge amount of data effectively is difficult tasks. Extracting the useful data, analysis of data, aggregate, and storage of these data in real time is a hurdle for researchers. It need a better management or processing system for that. Here the method used K nearest neighbor techniques for the analysis of the huge volume of data.

Keywords: KNN: K Nearest Neighbor, HDFS: Hadoop Distributed File System, Map Reduce

1. INTRODUCTION

The great deal of interest has risen recently in the field of big data and the analysis has increases and this data is driven from the extensive number of the research challenges related to big and strong bonafide application such as modelling the data, processing the data and distributing large scale of the repository data.

The term big data classifies the particular forms of data sets comprising formless data which will be taken from the technical computing application layers.

2. EXISITNG SYSTEM

In existing system, the analysis is based on the remote sensing application data, that was captured by cameras or sensors from which different sceneries are recorded using the radiations. After this special techniques will applied to process the data which is captured and interpreted for the purpose of producing the different kind of maps like conventional and thermal and also for resource surveys.

2.1 Remote Sensing Big Data Acquisition (RSDU)

The remote sensing promotes the expansion of the data which is captured and it allows provides parallel data acquisition to satisfies the requirements and using traditional way it's not able to provide the required amount of power to process the data so the parallel processing is required for the huge amount of data.

The RSDU gathers the data from different satellites around the globe , the raw data will be converted to image format using the algorithm called SPECAN, by doing so the storage cost will be less and it is more efficient and it also eliminate the redundant data . The data is transmitted to earth base station using the direct communication link or using the antenna and this data will be divided as online and offline data and these will be processed further.

2.2 Data Processing unit (DPU)

In data processing unit filtration and load balancing will be done and it also have the processing powers. Filtration uses the only useful amount of data for the analysis and the remaining data will be blocked and discarded. So it results in the enhancement of the performance.

After filtration the load balancing part provide the facility of diving the whole useful data which is available from the filtration part and it assign this data to the different processing servers. Each filtration and load balancing system have its own implementation part of the algorithm and the way of processing the segments.

2.3 Data Analysis and Decision Unit (DADU)

The data analysis and decision unit contain three major parts such as, aggregating and compilation of the data, storing the result of the server and finally the decision making part. The aggregation part stores the compiled and the data in the organized format. Later it send these copy to the decision making part. It supports the algorithm which analyze the different things form the result make helps to make different decisions.

After performing all these tasks, the result will provide whether the data which took for the analysis belongs to land or sea data.

3. PROPOSED SYSTEM

Hadoop is able to process both the online and offline data, online data are the ones which is directly taken form the HDFS and offline data can be access from the source of the HDFS

Using Hadoop we are trying to process and analyze the data which can be either offline or online, these data are

retrieved from the different consequences of natural disaster based on the different types of events.

The map reduce framework is used for processing the data, first the data will split into multiple blocks and the it starts the processing of the data and the data will be processed parallel in multiple blocks and these data will be shuffle and sorted and generate the multiple <key, value> pair and then these blocks will be given to the reducer and the reducer will combine the data based on the <key, value> pair and gives the consolidate output.

The KNN algorithm is used to classify the data and calculate the values using Euclidean formula, this formula will be used to calculate the distance between the two points and it finds the nearest value for the cluster and finally it classify the data based on the value it generate during map and reduce phase.

3.1 Hadoop Distributed File System (HDFS)

Hadoop Distributed file system is used to store the data in a distributed manner. HDFS has master-slaves architecture and it exposes a file system namespace and it stores the data in files and internally, a file is divided into one or more blocks and these blocks are stored in Data Nodes. HDFS has one Name Node (a master) and set of Data Nodes (slaves). The Name Node performs operations like open, close and rename files and directories. It also defines the mapping of blocks to Data Nodes. The Data Nodes are responsible for serving the requests like read and write from the file system's clients. And when Data Node gets instructions from Name Node, it performs block creation, deletion and replication.

HDFS is used to store massive amount of data with the data stored in multiple clusters. To elaborate this, take a simple following example. If the three files are stored to the HDFS using different file names, then the HDFS will store all the three files and also for each file it reserves three disk spaces. In some cases only few bytes of file are changed and this leads to store the whole file in HDFS. So to provide the efficient storage facilities.

3.2 Tools, Data sets and Implementation Environment

The study of the data sets is based on the storm data information and which is published by National Oceanic and Atmospheric Administration (NOAA). The main objective is determination of which natural disaster has occur in the different country especially in the United States of America and it also provides the information regarding which is more harmful.

The natural disaster which are discussing here is about the storm data and also about the other phenomena having the intensity to cause the damages, injuries, loss of life, damage of the property and also some kind of disruption.

A new dataset which storm data is used called summary_data, is generated from the existing data sets called analytic dataset. Each value of this dataset corresponds to one natural disaster event. This new dataset summarizes the data into the pseudo-categories and also summing out the property damage during the disaster occur , the crop damage, the injuries and the fatalities i.e number of people dead during the disaster. The columns of the summary_data dataset are the following:

- ABRV_EVTYPE: gives the pseudo-category abbreviated name.
- COMP_EVTYPE: gives the complete category name. Note that this name only correspond to one of the raw categories (some pseudo-categories contain more raw categories).
- FATALITIES: it provides the sum of fatalities for this pseudo-category from the raw datasets.
- INJURIES: it gives the sum of injuries for this pseudo-category from the raw datasets.
- PROPDMG_DOLLARS: the total amount of property damage during the disaster for all the events of this pseudo-category.
- CROPDMG_DOLLARS: the total amount of crop damage and gives all the events of this pseudo-category.

3.3 The K Nearest Neighbor Algorithm (KNN)

The K Nearest Neighbor algorithm used to classify the datasets or the training examples based on their values. Clustering techniques is used for the ease of data categorization. Here the algorithm which used for the analysis is K-nearest clustering algorithm for the recognition of the pattern, the k-Nearest Neighbors algorithm (or k-NN for short) is a method used for classification and regression.

The algorithm gives the information about the parameters and values which is used in the proposed system.

The distance is calculated using the Euclidean formula

$d(xi,xj) = \sqrt{(\text{For all attributes a} \sum (xi,a - xj,a)2)}$

where x is a dataset and it ranges from (x1…..xn) and the value of xj is the another point in the datasets, by using this the distance between the points will be calculated and if the number of points will be more than it takes the majority value.

The value of k varies based on the input and the output value, if the value of k=1 then it considers only the single nearest point among multiple points. The value of k will be chosen as odd value because to avoid the confusion while calculating the centroid value of the point which taken for the analysis. In the proposed system , the value of k is 3 i.e k=3, so it considers the 3 nearest points and update its value.

4. RESULT AND IMPLEMENTAION

Apache Hadoop is a software library framework that allows the Map and Reduce program using only one node setup for sophisticated analysis is used for the implementation of the proposed algorithm, since Hadoop will provides the facility of parallel processing, high-performance computation of the data using a large number of servers. Therefore, it is suitable for analyzing a large amount of remote sensory image data. The architecture of the proposed system uses a similar mechanism for balancing the load and hence, the preference is given to Hadoop for the analysis, and also used for algorithm development, and finally for testing. Clustering techniques is used for the ease of data categorization. Here the algorithm which used for the analysis is K-nearest clustering algorithm for the recognition of the pattern, the k-Nearest Neighbors algorithm (or k-NN for short) is a method used for classification and regression.

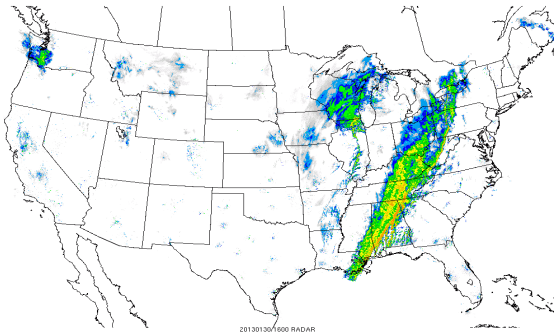


Fig. 1. Strom occurred region

4.1 Hadoop

The Hadoop is used to analyze the datasets and command prompt is used to run the command.

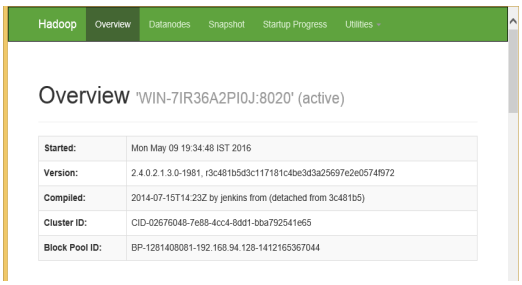
Hadoop commands are used to run in the command prompt the following steps gives the overview.

1) First create the directory using the Hadoop commands and then multiple directories will be created using the same command.

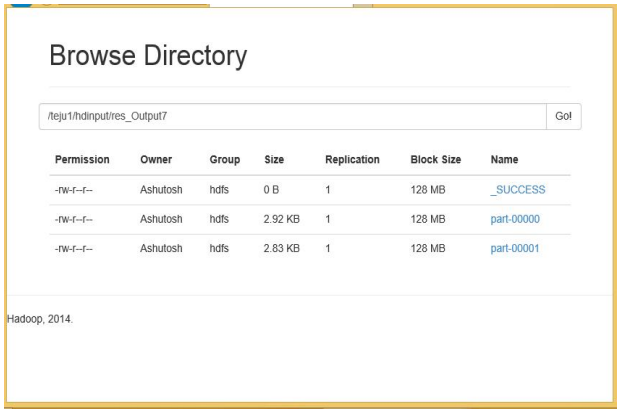
2) The file will be copied from the local to and the KNN algorithm is used to analyze the data and it returns the event type based on the data provided for the analysis.

4.2 Results

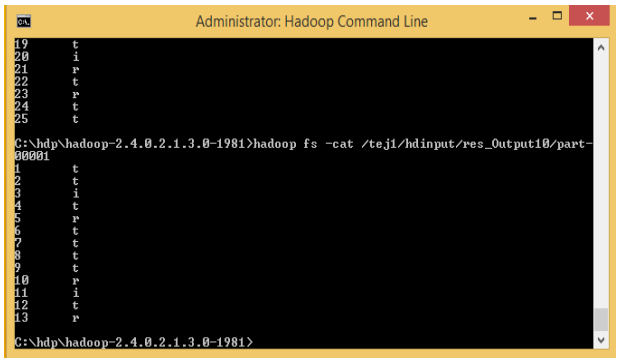
Overview of the Hadoop.



Shows the directory which is present in the Hadoop



The event type category



5.CONCLUSION

The Apache Hadoop software is used to for the analysis of the data sets which is generated from the National Oceanic and Atmospheric Administration (NOAA), the storm data is taken for the analysis and depending on that the event type will be analyzed. HDFS will be used to divide the data into blocks and it will process parallel using map part and after this the resulted blocks will be given to the reduce part, finally it gives the consolidate output. The algorithm process and gives the output. The result will be the different event types based on the data taken for the analysis. The event type will be the disaster names that occur in the different regions. Based on the cluster value the processing will be done. Furthermore, the algorithm have the capabilities to analyze and process the data in the efficient manner.

References

[1]. Improving Decision Making in the World of Big Data http://www.forbes.com/sites/christopherfrank/2012/03/25/improvingdecision- making-in-the-world-of-big-data/

[2]. Pakize, S., & Gandomi, A. (2014). Comparative Study of Classification Algorithms Based On MapReduce Model. International Journal of Innovative Research in Advanced Engineering (IJIRAE), 1(7), 251-254.

[3]. DeWitt D, Gray J (1992) Parallel database systems: the future of high performance database systems. Communication ACM35(6):85–98
[4]. A. Plaza et al., "Recent advances in techniques for hyperspectral image processing," Remote Sens. Environ., vol. 113, pp. 110–122,2009.
[5]. J. A. Richards and X. Jia, Remote Sensing Digital Image Analysis: An Introduction. New York, NY, USA: Springer, 2006
[6]. Bhagattjee, B. (2014). Emergence and Taxonomy of Big Data as a Service.
[7]. Anchalia, Prajesh, and Kaushik Roy. The K-Nearest Neighbor Algorithm Using MapReduce Paradigm. Fifth International Conference on Intelligent Systems, Modelling And Simulation. 2014. Web. 15 Oct. 2015.
[8]. K. Fukunaga and P. M. Narendra, "A branch and bound algorithm for computing k-nearest neighbors," IEEE Trans. Comput., vol. C-24, no. 7, pp. 750–753, Jul. 1975.