Data Analytics Using K-NN Techniques For Big Data

Tejashree S[1], Ravi Kumar V G[2], Shruthi K[3]
[1, 2, 3] Assistant Professor, Dept of Master of Computer Application
[1, 2, 3] JSSCACS Mysuru

Abstract- *Recently, there has been a significant increase in interest in the field of big data analysis. This data is fueled by a multitude of research challenges related to big and strong legitimate applications, such as modeling, processing, and distributing large amounts of repository data. A great deal of valuable data is lost as a result of incorrect data handling and storage. Effectively managing vast amounts of data is a challenging task. A challenge for academics is to extract the relevant data, analyze it, aggregate it, and store it in real time. For it, a better processing or management system is required. Here, the enormous volume was analyzed using K nearest neighbor approaches.*

Keywords- K Nearest Neighbor, Hadoop Distributed File System, Map Reduce, Data analysis

I. INTRODUCTION

The field of big data has seen a significant increase in interest recently, and this interest is driven by several research challenges that accompany big and strong legitimate applications, such as modeling, processing, and distributing large amounts of repository data.

Big data is a term used to describe specific types of data sets that include formless data that are extracted from technical computing application layers.

II. EXISITNG SYSTEM

The analysis in the current system is based on data from remote sensing applications that were taken by cameras or other sensors and used radiation to record various scenes. Following this, unique methods will be used to process the information gathered and analyzed in order to create various maps, including thermal and conventional maps, in addition to resource surveys.

2.1 Remote Sensing Big Data Acquisition (RSDU)

Remote sensing encourages the expansion of data collection and enables parallel data acquisition to meet requirements. Traditional methods lack the ability to supply the necessary power to process the data, necessitating parallel processing for the massive volume of data.

The RSDU collects data from several satellites located all over the world. The raw data is then transformed into an image format using the SPECAN algorithm, which results in lower storage costs, increased efficiency, and the removal of redundant data. The information is sent to the Earth Base Station via an antenna or a direct communication link, and it will be split into online and offline data and these will be processed further.

2.2 Data Processing unit (DPU)

Filtration and load balancing will be performed at the data processing unit, which also has processing capabilities. Through filtering, only the most pertinent data is used for analysis; the rest is blocked and released away. In response, the performance is improved.

Following filtering, the load balancing component offers the ability to access all of the valuable data that was made available by the filtration component and distributes it among the several processing servers. Every filtering and load balancing system has a distinctive algorithmic implementation and segment processing method.

2.3 Data Analysis and Decision Unit (DADU)

The main three parts of the data analysis and decision unit are the server result storage, data aggregation and compilation, and decision making. The data that has been compiled and organized is stored in the aggregation section. These copies are subsequently sent to the decision-making section. It facilitates the algorithm that examines various elements from the outcome and aids in the formulation of various decisions.

Upon completion of these tasks, the outcome will indicate whether the data used for the analysis is from the sea or the land.

III. PROPOSED SYSTEM

Both online and offline data can be processed by Hadoop; online data come directly from the HDFS, while offline data can be accessed from the HDFS's source.

We are attempting to process and analyze data—which may be offline or online—using Hadoop. The data is retrieved from the various effects of natural disasters depending on the various kinds of events.

The data is processed using the map reduce framework. Firstly, the data is split into multiple blocks. Next, the processing of the data begins. The data is processed in parallel in multiple blocks, shuffled and sorted to produce multiple <key, value> pairs. These blocks are then given to the reducer, which combines the data based on the <key, value> pairs and yields the consolidate output.

The Euclidean formula is used to calculate the values and classify the data using the KNN algorithm. It finds the closest value for the cluster by calculating the distance between two points. The data is then classified based on this value and it finds the nearest value for the cluster and finally it segregate the data based on the value it produce during map and reduce phase.

3.1 Hadoop Distributed File System (HDFS)

The data is evenly stored using the Hadoop Distributed File System. A file is internally divided into one or more blocks, which are then stored in Data Nodes. HDFS utilizes a master-slave architecture, exposes a file system namespace, and stores data in files. One Name Node (a master) and a number of Data Nodes (slaves) make up HDFS. Open, close, and rename operations for files and directories are handled by the Name Node. The mapping of blocks to Data Nodes is also defined by it. Serving read and write requests from file system clients is the responsibility of the data nodes. Additionally, Data Node creates, deletes, and replicates blocks in response to instructions from Name Node.

Massive volumes of data are stored using HDFS, which divides the data among several clusters. I'll give a brief example to clarify this. The HDFS will store all three files and reserve three disk spaces for each file if the three files are stored there under separate file names. In certain instances, a file may only have a few modified bytes, in which case HDFS is used to store the entire file. Thus, in order to offer effective storage facilities.

3.2 Tools, Data sets and Implementation Environment

The National Oceanic and Atmospheric Administration's (NOAA) storm data information serves as the foundation for the data sets' analysis. The primary goal is to identify which natural disasters have occurred in various nations, particularly in the United States of America, and to provide information about which is more dangerous.

The natural disaster under discussion is related to storm data as well as other phenomena with the potential to cause harm, property damage, fatalities, injuries, and other types of disruptions..

The existing data sets, known as the analytical dataset, are used to create a new dataset called summary_data, which is used storm data. This dataset's values are associated with a single natural disaster occurrence. In addition to totaling the crop damage, property damage, injuries, and fatalities—that is, the number of people who died during the disaster—this new dataset also summarizes the data into pseudo-categories. The following are the columns in the summary_data dataset:

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 training examples or datasets were categorized using the K Nearest Neighbor algorithm according to their values. Data categorization is made easier by the use of clustering techniques. Here, the K-nearest clustering algorithm is utilized for pattern recognition, and the k-Nearest Neighbors algorithm, or k-NN for short, is a technique for regression and classification.

The algorithm provides details on the values and parameters that are employed in the suggested system.

The Euclidean formula $d(xi,xj) = \sqrt{(}$For all attributes a $\sum$ (xi,a – xj,a)2) is used to calculate the distance where x is a dataset with a range of (x1...xn) and the value of xj is

another point in the dataset. The distance between the points can be computed using this, and if there are more points than there are, the majority value will be taken.

The value of k varies according to the input and output values; if k=1, only the single nearest point among several points is taken into account. In order to minimize confusion when determining the centroid value of the point that will be used for analysis, the value of k will be selected as an odd number. Since k in the suggested system has a value of 3, or k=3, it updates its value by taking into account the three closest points.

IV. RESULT AND IMPLEMENTAION

The suggested algorithm is implemented using Apache Hadoop, a software library framework that enables the Map and Reduce program with only one node setup for sophisticated analysis. Hadoop offers the capability of parallel processing, high-performance computation of the data using a large number of servers. As a result, it works well for processing a sizable volume of distant sensory image data. Because Hadoop is preferred for analysis, algorithm development, and testing, it is because the architecture of the suggested system employs a similar mechanism for balancing the load. Data categorization is made easier by the use of clustering techniques. In this case, the K-nearest clustering algorithm is utilized for recognition analysis. 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.

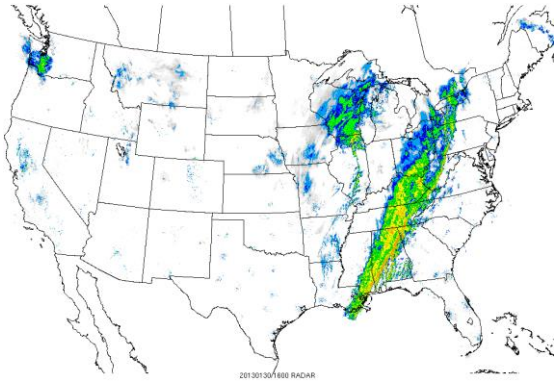

Fig. 1. Strom occurred region

4.1 Hadoop

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

The following procedures provide an overview of how to run Hadoop commands in the command prompt.

1) Using the Hadoop commands, create the directory first. Then, use the same command to create multiple directories.
2) After copying the file from the local drive, the KNN algorithm analyzes the data and determines the event type based on the data supplied.

4.2 Results

Overview of the Hadoop.

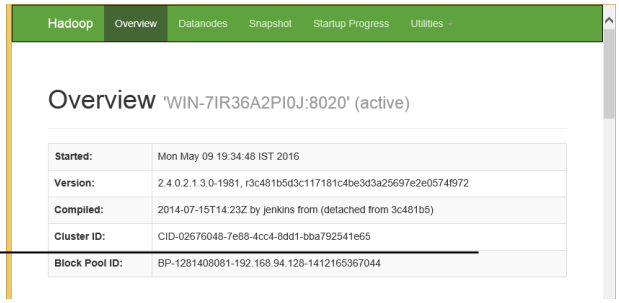

Fig. 2.Shows the directory which is present in the Hadoop

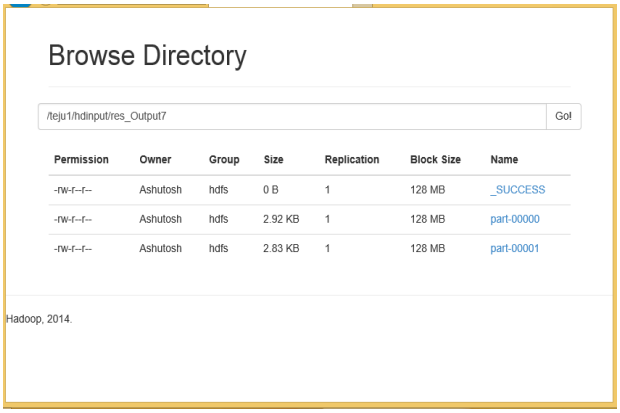

Fig. 3. The event type category

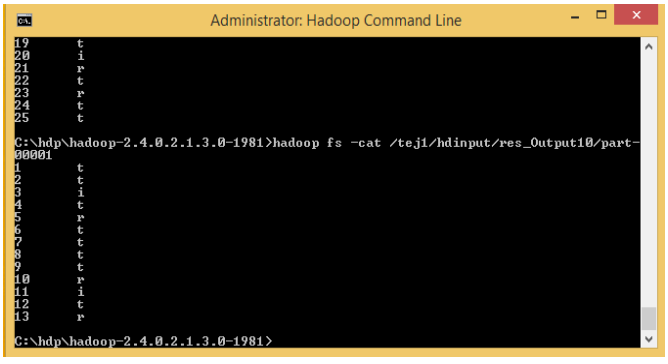

Fig.4. Output screen

V. CONCLUSION

The National Oceanic and Atmospheric Administration (NOAA) generates data sets that are analyzed using the Apache Hadoop software. Storm data is taken for analysis, and the type of event will be analyzed based on that. The data will be broken up into blocks using HDFS, which will then process them in parallel using the map and reduce parts before giving the output of the consolidate stage. After processing, the algorithm produces the result. Based on the data collected for the analysis, the outcome will be the various event types. The disaster names that transpire in the various regions will be the event type. In accordance with the cluster value, the processing will be done in an 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.
[9] Karthikeya, H. K., K. Sudarshan, and Disha S. Shetty. "Prediction of agricultural crops using KNN algorithm." Int. J. Innov. Sci. Res. Technol 5, no. 5 (2020): 1422-1424.