A Model for Empirical Earthquake Prediction and Analysis in a Data Intensive Environment

Faraz Ahmad, Om Ashish Mishra, Shivam Bhagwani, Jabanjalin Hilda J

Abstract*—Characteristic perils like earthquakes are for the most part the consequence of spreading seismic waves underneath the surface of the earth. Tremors are dangerous absolutely in light of the fact that they're erratic, striking without warning, triggering fires and tsunamis and leading to deaths of countless individuals. If researchers could caution people in weeks or months ahead of time about seismic disturbances, clearing and different arrangements could be made to spare incalculable lives. An early identification and future earthquake prediction can be achieved using machine learning models. Seismic stations continuously gather data without the necessity of the occurrence of an event. The gathered data can be used to distinguish earthquake and non-earthquake prone regions. Machine learning methods can be used for analyzing continuous time series data in order to detect earthquakes effectively. The pre-existing linear models applied to earthquake problems have failed to achieve significant amount of efficiency and generate overheads with respect to pre-processing. The proposed work exploits parallel processing in Hadoop by using the various frameworks like Pig-Hive optimization, Map Reduce and Impala, in order to mine and analyze earthquake data to propose a model for predicting future earthquakes.

***Keywords*—earthquake, pig-hive, prediction, classification, machine learning, impala**

I. INTRODUCTION

Movement of seismic plates under the surface of the earth which supports life in many form, causes earthquakes which is a natural hazard. Seismometers, which are used to record motion of these plates, are installed at various locations on the planet. These instruments detect vertical motion of the plates to record it on the scale. The earth surface formally called the crust is divided into seven large tectonic plates. These larger plates are further divided into several small sub-plates which are being observed and are noticed to move apart continuously. There are variances of seismic types. These can be stated as divergence, convergence which lead to transformation of plate boundaries. When the plates distance themselves from each other, new boundaries are introduced. In the phenomenon of convergence, plates of different densities tend to approach nearer giving rise to new geographical structures. When these plates slide apart from each other, this type of motion is called transformation. Divergence, convergence and transformations are all together known as faults. A fault in any geological region causes stress. When the stress quantity is large, it is released by earth in the form of earthquakes and sometimes volcanic eruption (stress along with heat). Apart from faults, some other reasons leading to earthquakes include volcanic eruptions, nuclear activities, mine blasts. The point of origin of the earthquake is known as the focus point. Earthquakes are recorded by a modern form of geophones called seismometers. These geophones are very sensitive to even small energy patterns that they can record. They work in efficient way when they are installed in groups and work in a cluster. The cluster of geophones can be deployed to increase the accuracy in measurement of seismic values. Geophones are mainly used for two purposes. Firstly, they increase the accuracy by reducing noise results; and secondly, they record vertical displacements and ignore any kind of horizontal seismic vibrations. Horizontally moving seismic waves are also called ground rolls. They are considered as noise which is caused as a side effect of seismic energy patterns. Vertically propagating waves almost simultaneously strike the seismometers installed in a group and are recorded. All the vertical waves that hit the seismometers at the same time are recorded by the cluster and all others which hit with some delay are ignored. The sum of the propagating waves vertically can be calculated and in the end it can generate time series data for recording. Four stages that are included in the prediction of earthquake in the proposed work are:-

i) Pig hive optimisation technique is used in the process as it is faster than Hadoop tools like Zookeeper. Thus the processing of the data becomes faster. Apart from Hive, impala is also used for queries.

ii) Pre-processing phase would include elaborating the different metrics over which the study would be conducted.

iii) Feature extraction is performed on Hadoop and plots are generated for analysis.

iv) Prediction using ensemble algorithms as well as comparison of different clustering techniques.

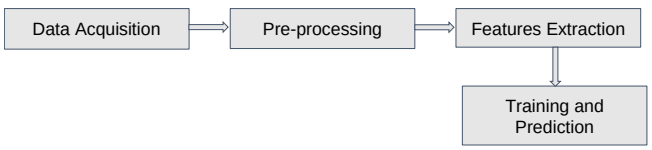

Figure 1: The stages involved in earthquake analysis

Revised Manuscript Received on April 12, 2019

Faraz Ahmad, Computer Science Engineering, Vellore Institute of Technology, Vellore, India.
Om Ashish Mishra, Computer Science Engineering, Vellore Institute of Technology, Vellore, India.
Shivam Bhagwani, Computer Science Engineering, Vellore Institute of Technology, Vellore, India.
Jabanjalin Hilda J, Vellore Institute of Technology, Vellore, India.

It is known that earthquakes are a major problem, even major countries like Japan are facing a lot of problems in detecting them in due course of time the main reason being that just on the basis of a simulated model we cannot say that a earthquake will take place at what time and at what location along with right or to some precision the magnitude. The proposed work uses Hadoop for performing the analysis in which there are two parts one is mapper and the other is reducer the mapper is used to form parallel segments or blocks (threads) and the reducer is used to extract the data from the mapper and store them into different directories after doing computation s these will be the two parts running in parallel and thereon check out what are the major earthquakes that have occurred in the past years so after scanning and deploying the charts are visualised in python/R and after that the locations where the most earthquakes took place is obtained. Earthquakes can take many forms in terms of destruction that they cause. They can be so weak that sometimes they cannot be felt and at the other times, so strong that they cause destruction at high scale and result in loss of life and property. They can make a whole area financially weak and needy. Term seismicity or seismic activity of a particular area is defined with the help of frequency, type and size of earthquake that are experienced over a period of time. The disturbance of seismic plates are the also the main reason of tsunamis. A tsunami is caused due to the epicentre of a large earthquake resulting somewhere in the ocean leading to displacement of seabed. Earthquakes can also trigger landslides, and sometimes volcanic activity. The best available example of such tsunami can be The Ring of Fire which is in the basin of Pacific Ocean. It has 452 volcanoes (more than 75% of the world's active and dormant volcanoes).

II. LITERATURE SURVEY

A lot of research has gone into the analysis of earthquakes. Mostly, due to the fact that earthquakes cause a lot of havoc and proper methods of predicting them are still not available. Earthquake being a very serious natural calamity needs to be predicted in order to prevent any kind of heavy loss to mankind. But, researchers and scientists are unable to do that due to many factors. This paper has listed some problems and issues which act as barriers to the prediction. As physical phenomena, earthquakes should be predictable up to a certain degree. But the prediction is difficult because the volume of source inside earth is inaccessible. The stress level of inner earth is one thing which can't be measured directly. Japan has come up with a technique which can analyze crystal deformations with the help of GPS and InSAR. But they have not claimed successful predictions till now [1].

BIRCH, a useful unsupervised learning technique of data clustering when the datasets are very large has been proposed [2]. Finding a useful pattern in a large dataset has already gained lots of interest recently, and identifying clusters is one of the most studied upon topic in this area. BIRCH incrementally and dynamically identifies multi-dimensional data and clusters them trying to produce best quality clustering with the available resources. Hadoop is a very honored platform to process big data volumes efficiently. Web analytics applications, social networks, scientific applications are some of the fields where Hadoop is used. MapReduce is a popular processing engine in it. It basically orders the processes in order of magnitude to boost performance. This particular journal has highlighted the similarities and differences between Hadoop MapReduce and Parallel DBMS [3].

MapReduce has been widely in use for large-scale data processing and analysis. MapReduce works fine when the hardware of a cluster is well configured to handle large data. However, the authors of this paper have conducted a survey which shows that most of the systems are memory constrained. Mammoth is a new MapReduce system which aims to improve the performance by using global memory management. Mammoth systems can decrease the execution time by more than 40% in most of the cases. The paper proves that the Mammoth system can have a promising potential and impact given the fact that MapReduce platform gained a great success [4]. A very impressive approach and probably the most defining move towards earthquake prediction was the use of neural networks. In [5] a method for the proposal of detecting earthquakes based on a number of geographical factors such as geometric field declination, humidity, rain hours etc were studied in order to predict the occurrence of earthquakes. As a case study, earth geomagnetic field measured data is used. Earthquake prediction methods have vastly improved, however there is no method of maximized accuracy using big data intensive approaches. It was observed that a neural network model was successfully able to predict the magnitude of earthquake along with an induced time lag of a seismic occurrence. This fact proves that neural networks require training by using the appropriate data in order to generalize and predict unknown seismic events accurately. The accuracy rates presented in the current paper are all based on the out-of-sample performance for each model. In other words, the data used for testing the networks are different from data used for training. For a good evaluation of a forecasting study, the method should be assessed based on the out-of-sample performance. In this way the predictive capability of the model will match the conditions of the real-world. Another approach towards the detection of earthquakes uses regression and ensemble learning. The paper combines several linear regression with ensemble learning in the context of big data and analyses the earthquakes in the regions of California. These include generalized linear models, gradient boosting machine, deep learning, random forests and stacking ensembles. A total of 1GB of data was analysed divided into 27 datasets and processed by means of cloud based infrastructure. The stacking-based ensemble learning had been applied, reporting relative errors verging on 10%. It was concluded that the methods based on trees yielded better results and lower regression errors [6].

With more research towards deep learning and soft computing, newer methods have been since taken into account for development of algorithms for predicting and observing damages caused by earthquakes. Another approach uses Hybrid neuro-fuzzy systems for estimating a

function, a Sugeno-type fuzzy system a special five-layer network. This algorithm can predict the seismic time between tremors thus acting as a benchmark for preparation and responsiveness of earthquakes. Lack of uniformity in seismic catalogue made the authors use available regression equations. These equations which are used to determine moment magnitude of earthquake records followed by calculating seismic moment lead to error [7]. Whenever we deal with natural phenomenon, we deal with a large amount of data. Thus processing and mining the right data becomes a very important task. Another approach to earthquake detection was the use of Spatio-Temporal Data Mining, Long Short Term Memory Neural Network Units. The algorithm can be used to make predictions and analyze earthquake and other natural disturbances occurring. The algorithm can be used to make accurate predictions with different temporal and spatial prediction granularities. The accuracy was improved by using more hidden layers. However a generic overhead was observed compared to other algorithms in obtaining and processing the input data [8].

III. PROPOSED ARCHITECTURE

The proposed work aims at using parallel processing in order to reduce the overheads that are generated when dealing with the processing and computation of earthquake data. Impala, a massively powerful parallel processing engine is used to perform the computations. The proposed work also analyses the earthquake dataset and performs analysis of the various clustering techniques like hierarchical and k-means clustering. The different problems solved by the proposed work includes the following:-

1. Predicting the most likely location of the future earthquake using past seismic data using the United States Geological Survey Dataset.
2. Classification of earthquakes based on types, magnitudes, location of occurrences,
3. Data analysis of past earthquakes and visualizations to better understand the factors behind occurrences of them.
4. Finding which regions are affected by earthquakes the most and successfully applying prediction and optimization algorithms on them.

The problems with Big Data are not only the size. There are three V's associated with Big Data that makes it difficult to process on a single machine.

1. Volume: The size of the data being generated.
2. Velocity : The rate at which the data is being generated.
3. Variety : the different sources and formats in which the data is coming.

The Hadoop map reduce algorithm is used to perform the processing. A number of other big data processing techniques have been used. Impala, a query processing variant in the big data environment is used. The proposed work includes a performance comparison of Hive and Impala. [9][10][11]

Data Collection

The proposed work uses the earthquake dataset of the United States Geological Survey.

Data Attributes for earthquake prediction
Date: The day when it occurred
Time: The particular time it started
Latitude and Longitude: Location tracing
Type: Natural Earthquake / Nuclear Explosion
Depth: Epicentre Location
Magnitude: Intensity of Earthquake
Magnitude Type: MB, MW,ML etc.
Status: It is the place of origin or it received from some where

Table 1: The Dataset Features

Data Preprocessing and Techniques Used

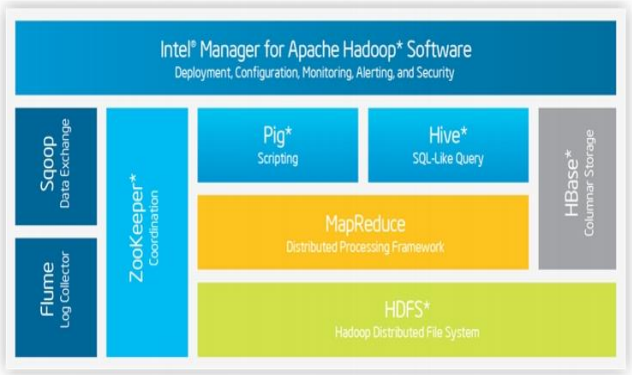

Figure 2: The Hadoop Architecture

The Pig Hive optimization technique is used. Pig is used here rather than other Hadoop tools like Hive or Zookeeper as it is fast and it processes the data faster. Just like the literal meaning of the pig, an animal who eats garbage, similarly here also the pig automatically processes the data(arranges) irrespective of the arranged or not. After the data is processed and cleaned, predictions on it is performed using K-Means. The K-Means algorithm is inefficient for large scale processing thus the proposed work also studied other areas such as evolutionary computing with emphasis on Swarm Particle Optimization to find a highly efficient algorithm that can be used. This would be based on the attributes defined in table 1. [12][13]

IV. IMPLEMENTATION

The proposed works compares the different Hadoop data processing technologies. The USGS dataset is used for data acquisition phase. The data is processed using Hadoop distributed computing algorithm. The techniques of Hive and Impala are used for the data pre-processing. The cleansed dataset is then fed to the different algorithms in order to observe insights and patterns. The first phase uses the K-means clustering algorithm to form the clusters based on the different locations of the earthquake. The hierarchical clustering and k-means clustering algorithm divide the earthquakes into clusters of Northern Hemisphere, Southern Hemisphere and the ones near the equator.

The Partition Cluster Algorithm

Step-0: Start

Step-1: Load the dataset into the environment

Step-2: Using Mapper, breakdown the tasks into individual clusters.

Step 3: Put the data in key and value pairs.

Step 4: Map the consequent values of a cluster through synchronised search into key pairs.

Step 5: Randomise the value to reduces through shuffle and sort

Step 6: Transfer the key pairs to reducer

Step 7 : Partition the data and store in HDFS

Step 8 : Compute the particle's closeness in free space and compute the clusters

Step 9: End

In stages of development of the algorithm in cloud-era environment, it was seen to take a lot of time in processing. Since a number of tasks were to be divided into mappers and reducers, the overhead and complexity increased enormously. Therefore, the data visualisation was carried out in a python environment. Figure 4 shows the values obtained from the k-means clustering algorithm.

Techniques	Description
K-Means	Simple clustering algorithm
Hierarchical Clustering	A more complex clustering algorithm.

Table 2: The clustering technique used

The figure 3(b) and 4(b) given below shows the results of the different clustering performed in the respective hemispheres. The elbow method is used to check the optimum number of clusters by checking the critical points on the graph. The most optimum number of clusters obtained here are 3.

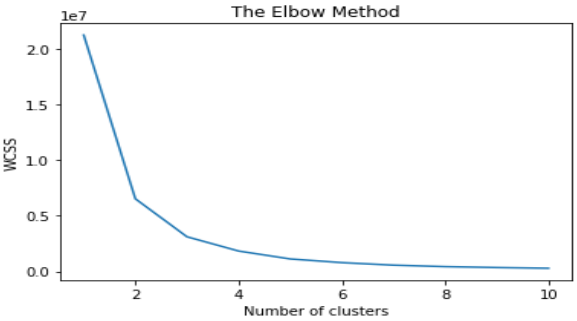

Figure 3(a): Elbow Method

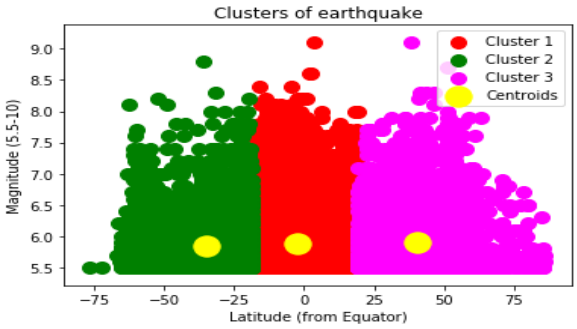

Figure 3(b): K-means Clustering

It was found that the distance between the clusters was more in K-means algorithm. Thus, hierarchical clustering was used as counter measure.

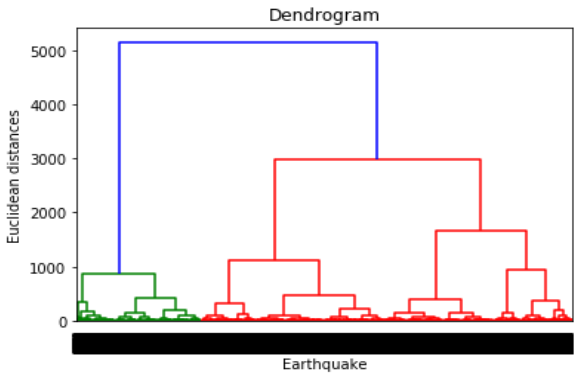

Figure 4(a): Dendrogram

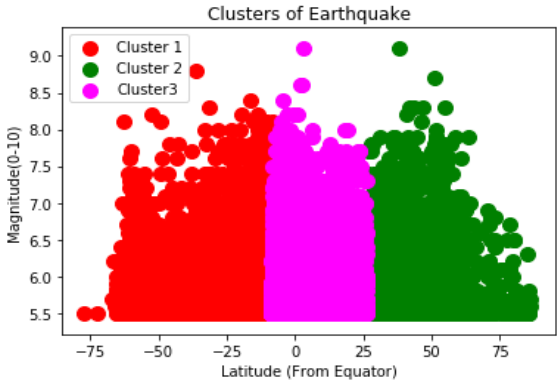

Figure 4(b): Hierarchical Clustering

The dendrogram in figure 4(a) shows the Euclidean distance between the clusters and groups them based on the least distance between the points. The clustering methods are used to obtain the centroid of the latitudes and longitudes and then the most prone locations to earthquakes are found, depending on the distance of the point, the clusters can be mapped as shown in figure 5(b).

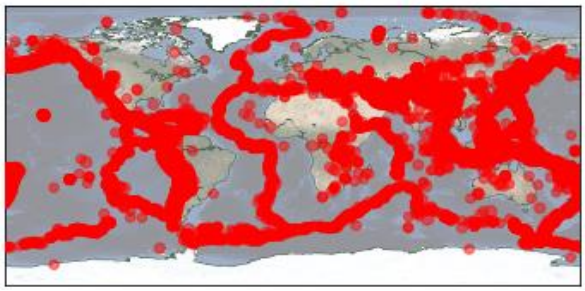

Figure 5(a): The plotting of earthquake dataset

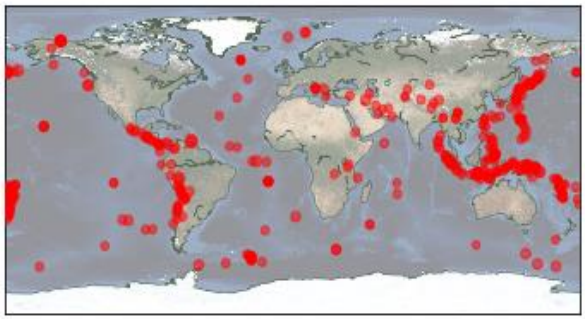

Figure 5(b): The predicted earthquake prone regions

V. RESULT AND PERFORMANCE EVALUATION

The different methods of Hadoop like Impala and Hive were compared. In phase 1, the data processing was done and insights on the different kinds of data was observed. The map reduce algorithm although versatile and universal, it fails to match efficiency and fast analytical parallel processing of Impala. Figure 6, shows a comparison for different number of tasks in Hadoop and Impala.

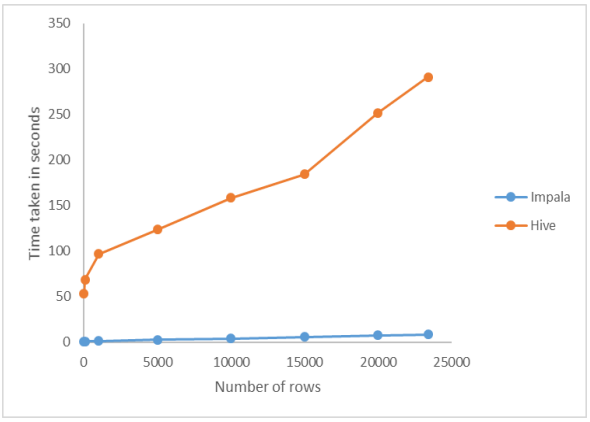

Figure 6: The Job Processing Speed Comparison

From Figure 3 it can be seen that the time taken for processing the same set of tasks in Impala was found to be far less compared to that of Hive. Impala demonstrates the brute force processing power to give lightning fast analytic results. Impala responds quickly through processing whereas Hive translates the given query into MapReduce jobs. Due to this, the overhead increases and thus leads to more processing time. Another advantage with impala is that it avoids start-up overhead as the processes are started at the boot itself, thus always being ready to process a query. Similarly, hive generates overhead during start.

The phase 2 of the proposed work aims at using the cleansed data in order to perform predictions. The train vs test ratio is 95:5. The different clustering techniques were compared and hierarchical clustering was found to give more condensed results. The table 3, compares the values for the different types of clustering techniques used.

No. of rows	K-means (s)	Hierarchical (s)
100	1.72	2.76
1000	2.65	4.11
5000	3.08	6.89
10000	6.11	12.19

Table 3: Comparison of time in different clustering techniques

The hierarchical clustering algorithm takes more processing speed however, further analysis stated that the entropy of hierarchical clustering is less. Entropy refers to the disorder with respect to a given clustering technique. The time overhead is used for making the clusters more precise. Table 4, draws a comparison on the given entropy for K-means and Hierarchical clustering. Similarly the figure 5 draws a line plot to further evaluate the performance between K-means and hierarchical clustering.

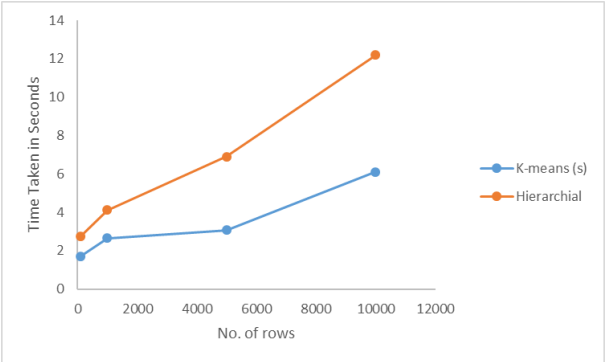

Figure 7: The time comparison between K-means and hierarchical clustering

No of Rows	K-means	Hierarchical
100	0.479	0.217
1000	0.585	0.312
5000	1.633	0.430
10000	2.172	0.768

Table 4: Comparison of K-means and Hierarchical Clustering in terms of entropy

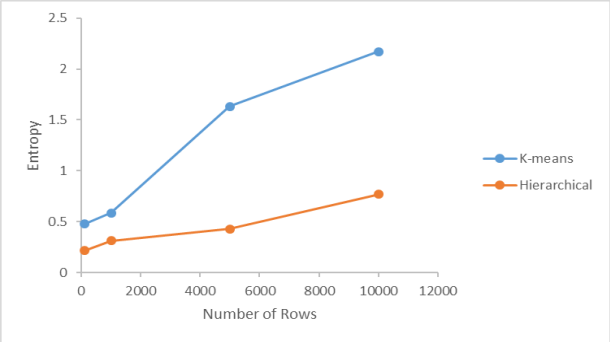

Figure 8: Comparison of Entropy between K-means and Hierarchical Clustering

The proposed work also compared other metrics for evaluating the performance of various clustering methods. F-measure and Coefficient of variance were taken into account and compared. The method of exhaustive enumeration was followed in order to obtain the values. The generic algorithm used for analysis was based on the performance of the earthquake dataset. The following algorithm can be used in on other datasets as well for computing the performance. F-measure is used for measuring the accuracy of clustering methods. This value is calculated by weighted average of recall and precision.

$$\text{Recall}(i,j) = \frac{\text{Number of elements of class i in j}}{\text{Number of elements of class i for cluster j}}$$

$$\text{Recall}(i,j) = \frac{\text{Number of elements of class i in j}}{\text{Number of elements of cluster j}}$$

F measure is a result of weighted average of Precision and recall for each class i, and |i| is the given size of the cluster class.

$$F(i,j) = \frac{\sum_i (|i| * F(i))}{\sum_i}$$

The table 4 applies the given formula for evaluating the F measure.

No. of clusters	K-Means	Hierarchical
1	0.008	0.01
2	0.0174	0.0312
3	0.0203	0.0389
4	0.0341	0.0412
5	0.0371	0.0562

Table 4: Comparison of F-measure

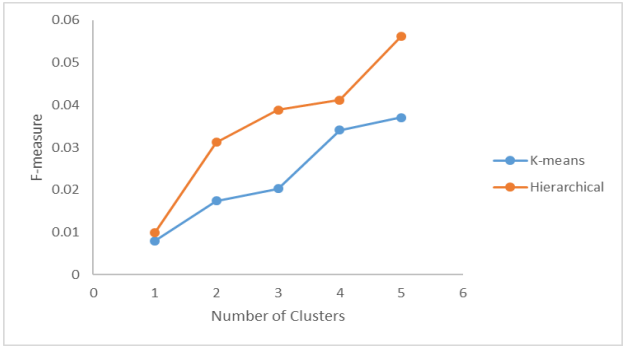

Figure 9: Comparison of F-measure between K-means and Hierarchical Clustering for different number of clusters

Similarly, the coefficient of variation was found out for the given clustering techniques. The coefficient of variation is obtained from the mean and standard deviation. Table 5 incorporates the coefficient of variance for the given techniques using the same method of exhaustive enumeration.

Number of clusters	K-means	Hierarchical
2	0.372	0.389
3	0.346	0.218
4	0.414	0.357
5	0.407	0.311

Table 5: Comparison of Coefficient of Variance

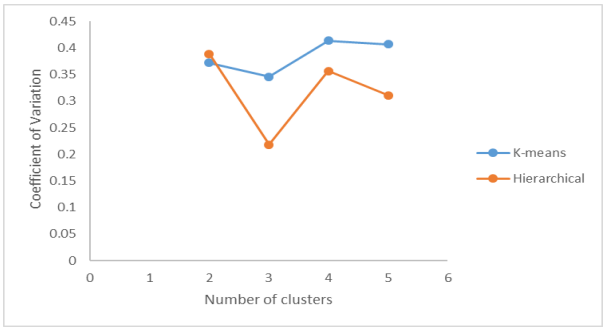

Figure 10: Comparison of Coefficient of variance between K-means and Hierarchical Clustering

VI. CONCLUSION

Thus it can be observed that by using the following algorithmic model for earthquake prediction, proper methods can be implemented for deploying warnings and preparing for earthquakes. The proposed algorithmic model efficiently performs data analysis using Hadoop and can be used for observing insights related to earthquakes. A deep study observed a number of areas that are more prone to earthquakes. Some of these regions include the pacific ring of fire, the Hindukush and the Himalayas, the Japanese coastal spread and the Philippines. It was observed that a number of reasons were responsible for earthquakes, the most dominant were tectonic disturbances followed by

nuclear activities. A number of clustering algorithms were used through the course of the research such as K-means and Hierarchical clustering. Hierarchical Clustering was found to be more efficient in terms of entropy but takes more processing time. Similarly, the coefficient of variance of hierarchical clustering was lower than K-means but it was found to have a higher F-measure. Similarly the data analysis was done in the Hadoop environment and techniques like Hive and Impala were compared. However there were a number of drawbacks related to the technique for prediction which can be improved upon.

REFERENCES

[1] Wyss, Max. "Why is earthquake prediction research not progressing faster?." Tectonophysics 338.3-4(2001):217-223.
[2] Zhang, Tian, Raghu Ramakrishnan, and Miron Livny. "BIRCH: an efficient data clustering method for very large databases." ACM Sigmod Record. Vol. 25. No. 2. ACM, 1996.
[3] Dittrich, Jens, and Jorge-Arnulfo Quiané-Ruiz. "Efficient big data processing in Hadoop MapReduce." Proceedings of the VLDB Endowment 5.12 (2012): 2014-2015.
[4] Shi, Xuanhua, et al. "Mammoth: Gearing hadoop towards memory-intensive mapreduce applications." IEEE Transactions on Parallel and Distributed Systems 26.8 (2015): 2300-2315.
[5] Moustra, Maria, Marios Avraamides, and Chris Christodoulou. "Artificial neural networks for earthquake prediction using time series magnitude data or Seismic Electric Signals." Expert systems with applications 38.12 (2011): 15032-15039.
[6] Asencio–Cortés, G., et al. "Earthquake prediction in California using regression algorithms and cloud-based big data infrastructure." Computers & Geosciences 115 (2018): 198-210.
[7] Mirrashid, M., et al. "Performance investigation of neuro-fuzzy system for earthquake prediction." (2016): 213-223.
[8] Wang, Qianlong, et al. "Earthquake prediction based on spatio-temporal data mining: an LSTM network approach." IEEE Transactions on Emerging Topics in Computing (2017).
[9] Bittorf, M. K. A. B. V., et al. "Impala: A modern, open-source SQL engine for Hadoop." Proceedings of the 7th Biennial Conference on Innovative Data Systems Research. 2015.
[10] Rodrigues, R. A., Lima Filho, L. A., Gonçalves, G. S., Mialaret, L. F., da Cunha, A. M., & Dias, L. A. V. (2018). Integrating NoSQL, Relational Database, and the Hadoop Ecosystem in an Interdisciplinary Project involving Big Data and Credit Card Transactions. In Information Technology-New Generations (pp. 443-451). Springer, Cham.Bittorf, M. K. A. B. V., et al. "Impala: A modern, open-source SQL engine for Hadoop." Proceedings of the 7th Biennial Conference on Innovative Data Systems Research. 2015.
[11] Bittorf, M. K. A. B. V., et al. "Impala: A modern, open-source SQL engine for Hadoop." Proceedings of the 7th Biennial Conference on Innovative Data Systems Research. 2015.
[12] Suneetha, N., et al. "Comprehensive Analysis of Hadoop Ecosystem Components: MapReduce, Pig and Hive." (2017).
[13] Hajikhodaverdikhan, P., Nazari, M., Mohsenizadeh, M., Shamshirband, S., & Chau, K. W. (2018). Earthquake prediction with meteorological data by particle filter-based support vector regression. Engineering Applications of Computational Fluid Mechanics, 12(1), 679-688.

AUTHORS PROFILE

Faraz Ahmad: Computer Science and Engineering student at Vellore Institute of Technology. Member of Association for Advancement of Artificial Intelligence, Computer Society of India and IEEE. Areas of research include Data Science, Cloud Computing, Automata Theory and Software Development Methodologies.

Om Mishra: Computer Science and Engineering student at Vellore Institute of Technology. Member of Microsoft Student Community. Areas of research include data mining, machine learning and big data technologies.

Shivam Bhagwani: Computer Science and Engineering student at Vellore Institute of Technology. Member of IEEE SSIT.

Jabanjalin Hilda J: Professor at Vellore Institute of Technology. Areas and publications of research include Cloud Computing, Data Mining, Data Visualization.