International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 01 | Jan-2016 www.irjet.net p-ISSN: 2395-0072

Clustering of Big Data Using Different Data-Mining Techniques

Manisha R. Thakare, Prof. S. W. Mohod, Prof. A. N. Thakare

[1] *M. tech, Computer science & engineering, B.D. College of Engineering Wardha, Maharashtra, India*
[2 3] *Assistant Professor, Computer science & engineering, B.D. College of Engineering Wardha, Maharashtra, India*

Abstract - ***There exist large amounts of heterogeneous digital data. The phenomenon of Big data which will be examined. The Big data analytics has been launched. Big data is large volume, heterogeneous, distributed data. Big data applications where data collection has grown continuously, it is expensive to manage, capture or extract and process data using existing software tools. Fast retrieval of the relevant information from databases has always been a significant issue. Clustering is a main task of exploratory data analysis and data mining applications. Clustering is one of the data mining techniques for dividing dataset into groups. Clustering is a kind of unsupervised data mining technique.***

Key Words: *Data Mining, Clustering, Classification, Clustering Algorithms, Big Data, Map-Reduce.*

1. INTRODUCTION

Big data is a largest buzz phrases in domain of IT, new technologies of personal communication driving the big data new trend and internet population grew day by day. The need of big data generated from the large companies like Facebook, yahoo, Google, YouTube etc for the purpose of analysis of enormous amount of data which is in unstructured form converted into structured form. The need of Big data analytics which is stored in relational database systems in terms of five parameters-variety, volume, value, veracity and velocity.

Volume: Data is ever-growing day by day of all types ever MB, PB, YB, ZB, KB, TB of information. The data results into large files. Excessive volume of data is main issue of storage. This main issue is resolved by reducing storage cost. Data volumes are expected to grow 50 times by 2020.

Variety: Data sources are extremely heterogeneous. The files comes in various formats and of any type, it may be structured or unstructured such as text, audio, videos, log files and more.

Velocity: The data comes at high speed. Sometimes 1 minute is too late so big data is time sensitive. Some organizations data velocity is main challenge.

Value: Value is main buzz for big data because it is important for business, IT infrastructure system to store large amount of values in database. It is a most important v in big data.

Veracity: The increase in the range of values typical of a large data set. When we dealing with high volume, velocity and variety of data, the all of data are not going 100% correct, there will be dirty data.

1.1 Data Mining Techniques

Data mining having many type of techniques like clustering, classification, neural network etc but in this paper we are consider only two techniques.

1.1.1 Clustering

Clustering is the most significant task of data mining. It is an unsupervised method of machine learning application. In clustering the classes are divided according to class variable. Two important topics are: (1) Different ways to group a set of objects into a set cluster. (2) Types of clusters. The result of the cluster analysis is a number of heterogeneous groups with homogeneous contents. The first document or object of a cluster is defined as the initiator of that cluster. The initiator is called the cluster seed.

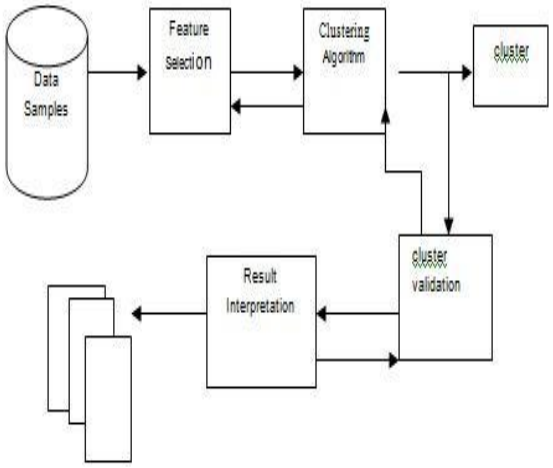


Fig1: Cluster Analysis

The procedures of the cluster analysis with four basic steps are as follows:

Feature Selection or extraction: Feature extraction utilizes some transformations to generate useful and novel features from the original ones. Feature selection chooses distinguishing features from a set of candidates. An elegant selection of features can greatly decrease the workload, and simplify the subsequent design process. Ideal features should be of use in distinguishing patterns belonging to different clusters, immune to noise, easy to extract and interpret.

Clustering algorithm design or selection: Patterns are grouped according to whether they resemble one another. The construction of a clustering criterion function makes the partition of clusters an optimization problem. Clustering is ubiquitous, and a wealth of clustering algorithms has been developed to solve different problems in specific fields. Therefore, it is important to carefully investigate the characteristics of the problem on hand, in order to select or design an appropriate clustering strategy.

Cluster validation: Different approaches usually lead to different clusters and even for the same algorithm, parameter identification or the presentation order of the input patterns may affect the final result. Therefore, effective evaluation standards and criteria are important to provide the users with a degree of confidence, for the clustering results derived from the used algorithms. Generally, there are three categories of testing criteria: external indices, internal indices, and relative indices. These are defined on three types of clustering structures, known as partitional clustering, hierarchical clustering, and individual clusters.

Result interpretation: The ultimate goal of clustering is to provide users with meaningful insights into original data, so that they can effectively solve the problems encountered. Experts in the relevant fields interpret the data partition. It may be required to guarantee the reliability of the extracted knowledge.

1.1.2 Classification

Classification is a simple process to finding a model that describes and distinguishes data classes of test. It is both types supervised learning and unsupervised. It consists of two steps:

Model construction : It consists of set of predefined classes. The set of tuple used for model construction is known as training set. These models can be represented as classification rules, decision trees.

Model usage : This model is used for defining future or unknown objects. It is used unsupervised learning rule.

2. PROPOSED METHODOLOGY

2.1 Big Data Technologies

Big data Test infrastructure requirement assessment.
Big data Test infrastructure design.
Big data Test Infrastructure Implementation.

The processing of large amount of data. The various techniques and technologies have been introduced for manipulating, analyzing and visualizing the big data. There are many solutions to handle the Big Data but the Hadoop is one of the most widely used technologies. But in this paper we are consider only Map Reduce technique.

2.2 Map Reduce

Map Reduce is a programming framework for distributed computing which is created by the Google in which divide and conquer method is used to break the large complex data into small units and process them.
Map Reduce: The master node takes the input. It divide into smaller subparts and distribute into worker nodes. A worker node further leads to the multi-level tree structure. The worker node process the m=smaller problem and passes the answer back to the master node. The master node collects the answers from all the sub problems and combines together to form the output.

2.3 Clustering Algorithm:

2.3.1. K-means Algorithm:

This method is a type of hierarchical clustering method using K-means. The algorithm starts by putting all the documents in a single cluster. It partitions the original clusters into two clusters by using K-means i.e. K=2. It makes the cluster which has highest intra cluster similarity as permanent & recursively split the other cluster into two more clusters using K-means with K=2& continue this until the desired number of clusters are created.

2.3.2. Bisecting K-means Algorithm:

Bisecting k-means is most popular and reduced dimentionality. Bisecting k-means is a combination of k-means and hierarchical k-means algorithm. It starts with all objects in a single cluster. Bisecting K-means Algorithm for finding k-cluster.

Step1 : Pick a cluster to split.
Step2 : Find two sub-clusters using the basic K-mean algorithm
Step3 : Repeat Step2
The bisecting step for ITER times and takes the split that produces the clustering.
Step4 : Repeat Step1, 2, & 3 until the desired number of Clusters are reached.

3. IMPLEMENTED MODULES:

Data Mining Applications For Big Data

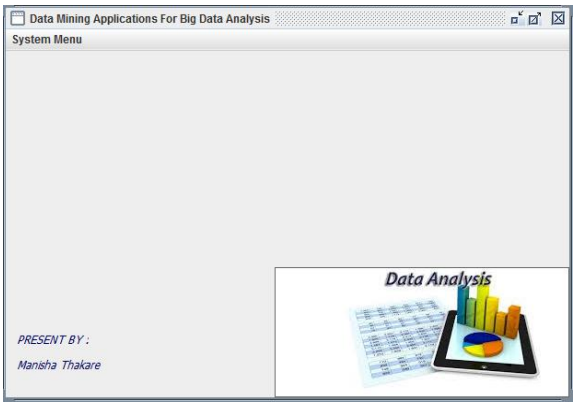


Fig 1 : Mining Applications For Big Data Analysis

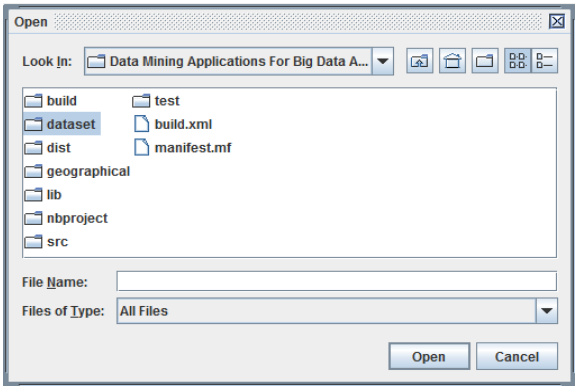


Fig 2 : Dataset

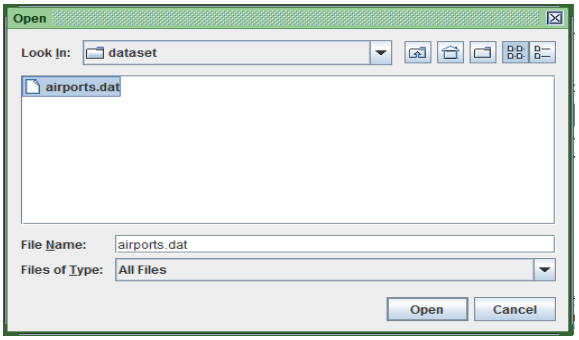


Fig 3 : Airport Dataset

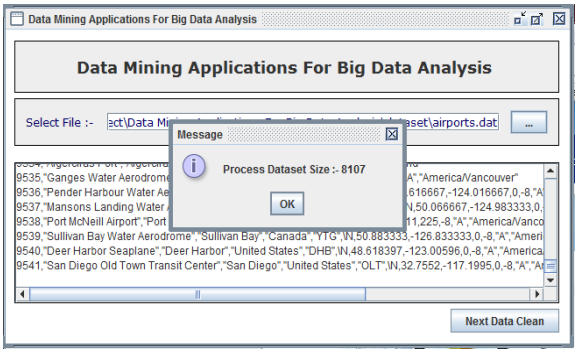


Fig 4 : Process Dataset Size

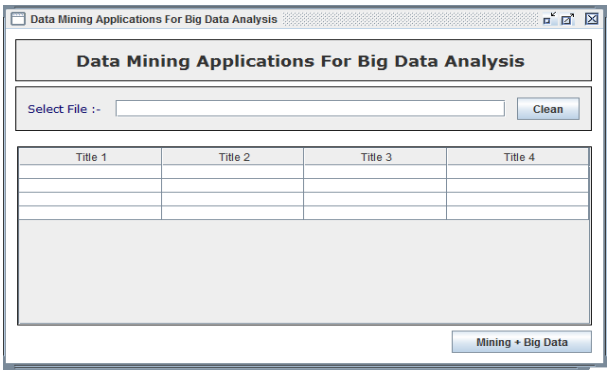


Fig 5 : Mining +Big Data

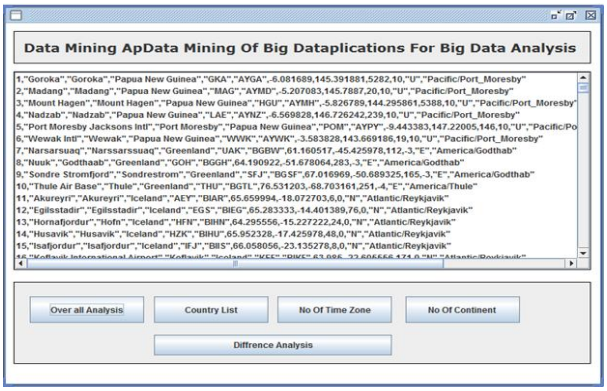


Fig 6 : Different Analysis

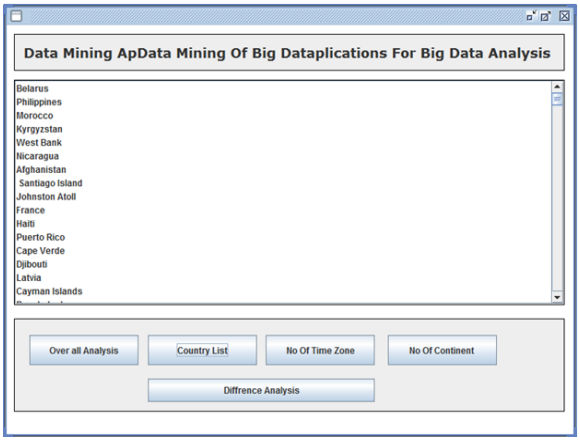


Fig 7 : Different Analysis : Country List

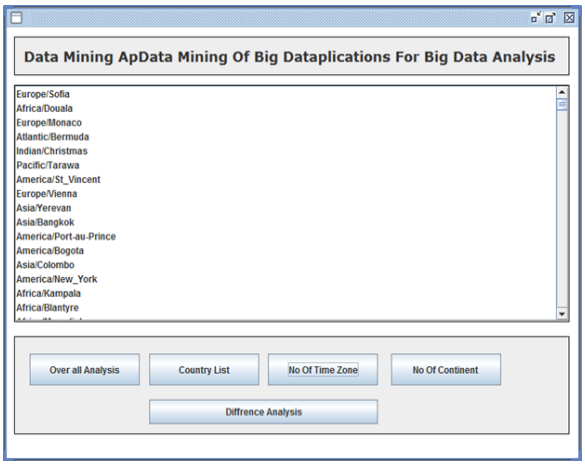


Fig 8 : Big Data Analysis : No. of Time Zone

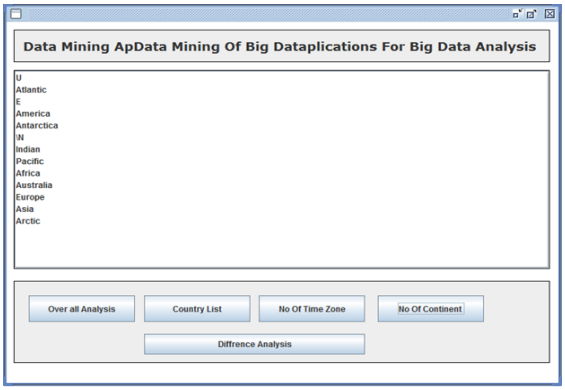


Fig 9 : Big Data Analysis : No. of Continent

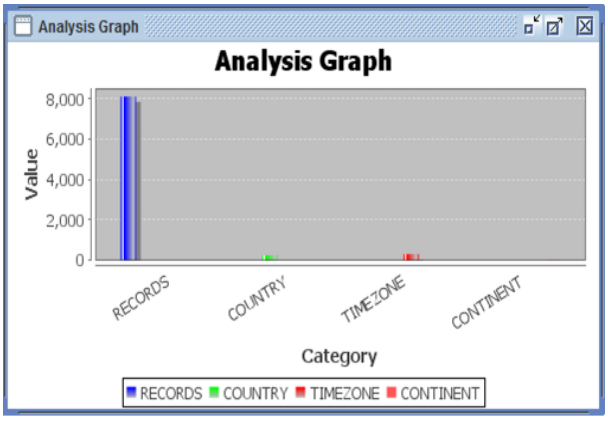


Fig 10 : Big Data Analysis Graph

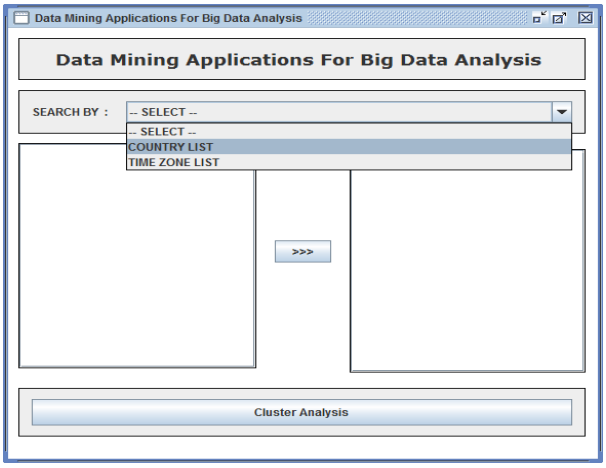


Fig 11 : Cluster Analysis : Country List

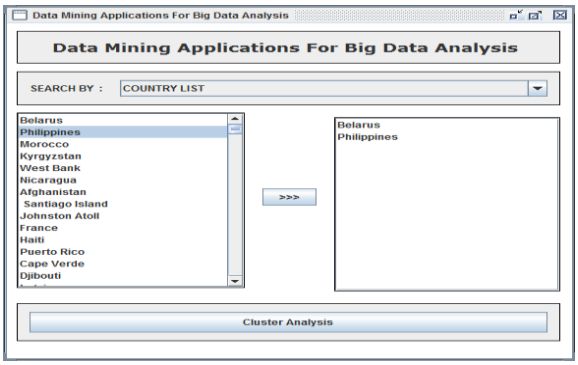


Fig 12 : Search by Country List

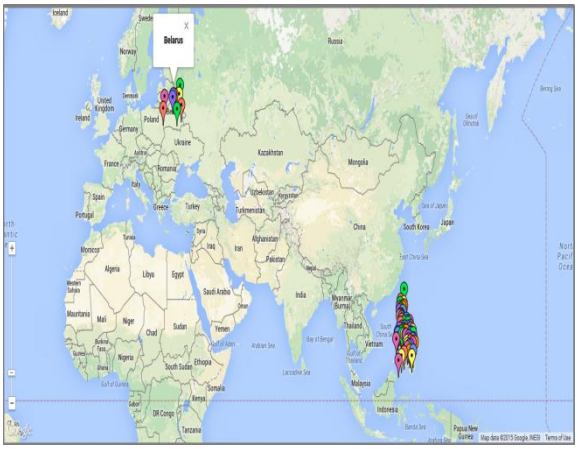


Fig 13 : Cluster Formation

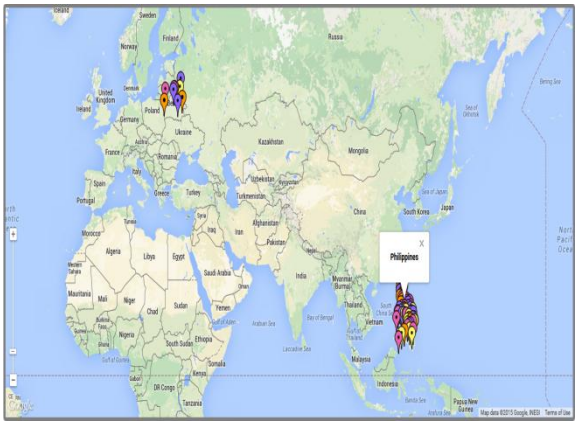


Fig 14: Cluster Result

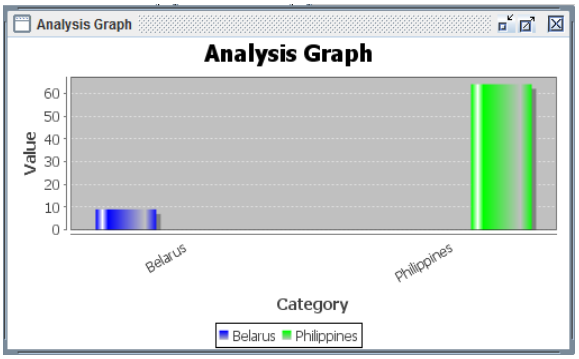


Fig 15: Analysis Graph : Country List

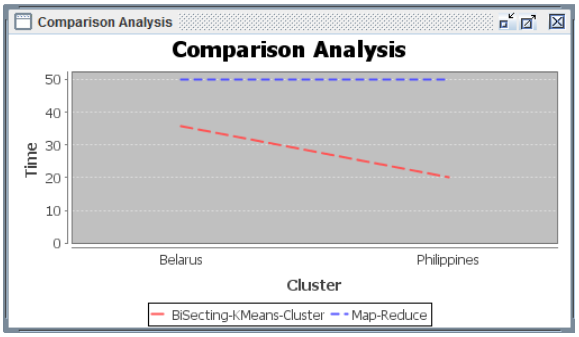


Fig 16 : Comparison Analysis

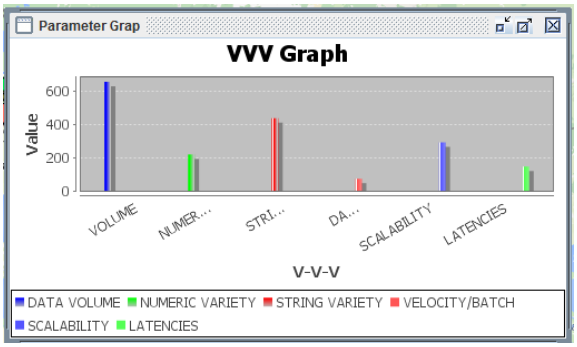


Fig 17 : Parameters graph for Big Data

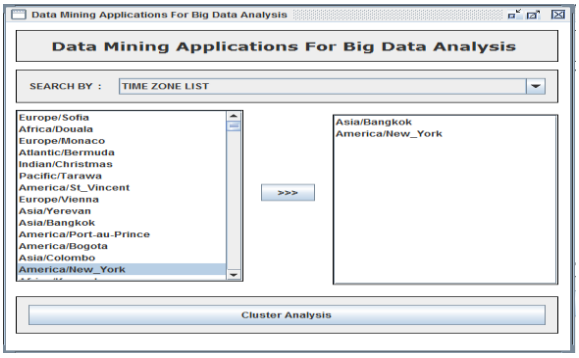
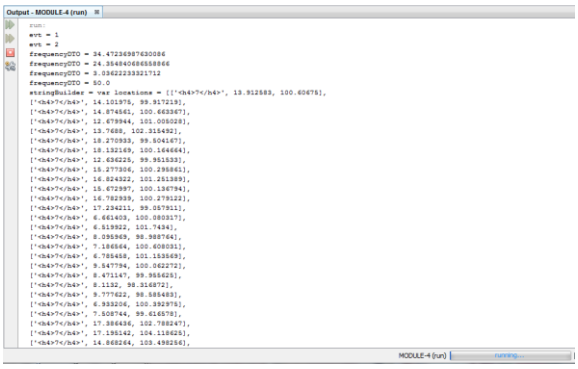
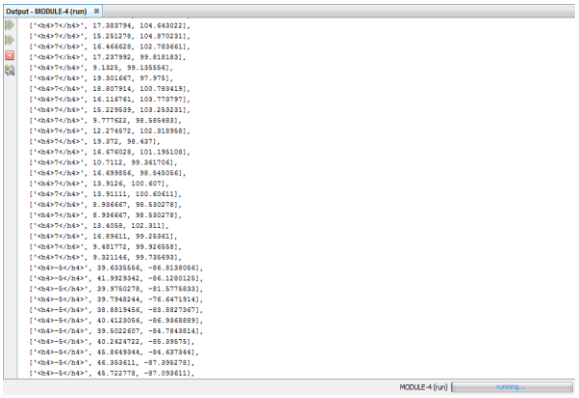


Fig 18 : Time Zone List

Fig 19 : Cluster Result

Fig 20: Cluster Formation





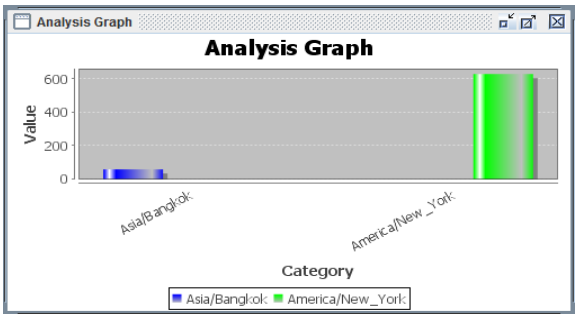


Fig 21 : Analysis Graph : Time Zone List

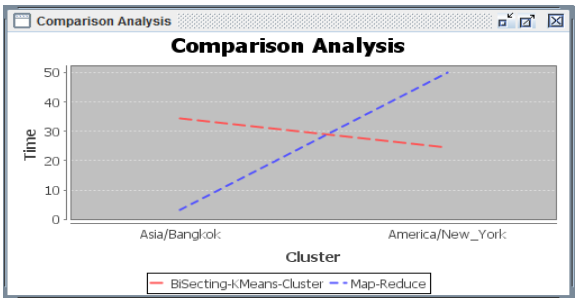


Fig 22 : Comparison Analysis : Time Zone List

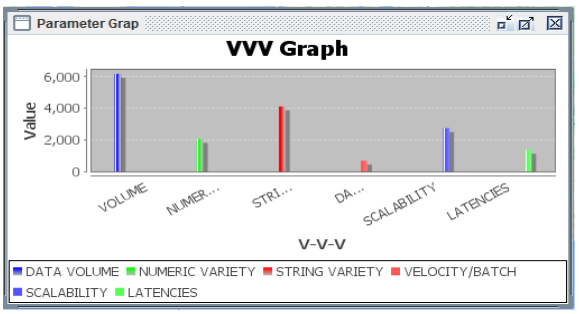


Fig 23 : Parameter Graph

4. CONCLUSIONS

Big data framework needs to consider complex relationships between samples, models and data sources. Big data mining high performance computing platforms are required. With Big data technologies we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at realtime. Map reduce mechanisms suitable for large scale data mining by testing series of standards data mining tasks on cluster. Map reduce implementation mechanism evaluated the algorithm. A system needs to be carefully designed so that unstructured data can be linked through their complex relationships to form useful patterns, the growth of data volumes and item relationships should help from legitimate patterns.

REFERENCES

[1] BABU, G.P. and MARTY, M.N. 1994. Clustering with evolution strategies Pattern Recognition, 27, 2, 321-329.

[2] Fayyad, U. Data Mining and Knowledge Discovery: Making Sense Out of IEEE Expert, v. 11, no. 5, pp. 20-25, October 1996.

[3] Guo, G, Neagu, D. (2005) Similarity-based Classifier Combination for Decision Making . Proc. Of IEEE International Conference on Systems, Man and Cybernetics, pp. 176-181

[4] Jyothi Bellary, Bhargavi Peyakunta, Sekhar Konetigari "Hybrid Machine Learning Approach In Data Mining", 2010 Second International Conference on Machine Learning and computing.

[5] Oyelade, O. J, Oladipupo, O. O, Obagbuwa, I. C" Application of k- means Clustering algorithm for prediction of Students Academic Performance" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, 2010.

[6] Varun Kumar and Nisha Rathee, ITM University, "Knowledge discovery from database Using an integration of clustering and classification", International Journal of Advanced Computer Science and Applications, Vol. 2, No.3, March 2011.

[7] McKinsey Global Institute (2011) Big Data: The next frontier for innovation, competition and productivity.

[8] Chen, H., Chaing, R.H.L. and Storey, V.C. (2012) Business Intelligence and Analytics: From Big Data to Big Impact, MIS Quarterly, 36, 4, pp. 1165-1188.

[9] Patel, A.B., Birla, M. and Nair, U. (2012) Addressing Big Data Problem Using Hadoop and Map Reduce, NIRMA University Conference on Engineering, pp. 1-5.

[10] Wu Yuntian, Shaanxi University of Science and Technology, "Based on Machine Learning of Data Mining to Further Explore", 2012 International Conference on Machine Learning Banff, Canada.

[11] Neelamadhab Padhy, Dr. Pragnyaban Mishra and Rasmita Panigrahi, "The Survey of Data Mining

Applications And Feature Scope", International Journal of Computer Science and Information Processing(CSIP).

[12] Aditya B. Patel, Manashvi Birla, Ushma Nair, (6-8 Dec. 2012),"Addressing Big Data Problem Using Hadoop and Map Reduce".

[13] Shiv Pratap Singh Kushwah, Keshav Rawat, Pradeep Gupta" Analysis and Comparison of Efficient Techniques of Clustering Algorithms in Data Mining" International Journal of Innovative Technology and Exploring Enginneering (IIJITEE) ISSN: 2278-3075, Volume-1, Issue-3, August 2012.

[14] Tekiner F. and Keane J.A., Systems, Man and Cybernetics (SMC), "Big Data Framework" 2013 IEEE International Conference on 13-16 Oct. 2013, 1494-1499.

[15] Dong, X.L.; Srivastava, D. Data Engineering (ICDE),' Big data integration" IEEE International Conference on , 29(2013)1245-1248.

[16] Sagiroglu, S.; Sinanc, D.,"Big Data: A Review",2013,20-24.

[17] Kyuseok Shim, MapReduce algorithms for Big Data Analysis, DNIS 2013, LNCS 7813, pp. 44-48, 2013.

[18] Madhuri V. Joseph, Lipsa Sadath and Vanaja Rajan" Data Mining: A Comparative Study on various Techniques and Methods" International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 2, February 2013 ISSN: 2277.

[19] Aastha Joshi, Rajneet Kaur A Review: Comparative Study of Various Clustering Techniques in Data Mining" International Journal of Advanced Research in computer Science and Software Engineering, Volume 3, Issue 3, March 2013.

[20] Yaxiong Zhao; Jie Wu INFOCOM, "Dache: A Data Aware Caching for Big-Data Applications Using the MapReduce Framework" 2014 Proceedings IEEE 2014, 35 – 39 (Volume 19).

[21] Wu, X., Zhu, X., Wu, G., Ding, W. (2014) Data Mining with Big Data, Knowledge and Data Enginnering , IEEE Transactions.