© 2016 IJSRST | Volume 2 | Issue 1 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Review on Big Data (Hadoop) processing model by implementing Data mining technique

Madhavi V. Shirbhate[1], Abhijit R. Itkikar[2]

ME Scholar, Dept of Computer Science and Engg, Sipna Collage of Engineering and Technology, Amravati, India[1]
Assistant Professor, Dept of Computer Science and Engg, Sipna Collage of Engineering and Technology, Amravati, India[2]

ABSTRACT

Big data is a term that describes the large volume of data –sensor data, tweets, photographs, raw data, and unstructured data. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves. Size of data has been exceeded Petabytes (1015 bytes) The size is not an issue but the processes are. Hadoop is a distributed computing open source framework for storing and processing huge unstructured datasets distributed across different clusters. The Business Intelligence in Hadoop retrieve the data from HDFS (Hadoop Data File System) and it locate that data in a database. The Database locate in a structured format. Due to this retrieving of data in cache duly consume the time and increase the factor of complexity. Here this paper present the data Mining algorithm to decrease the time and complexity factor for classification and clustering purpose. In this paper the identification of data present in data set is done using correlation and pattern. As the task of data mining is modelled ,a predictive or descriptive. A Predictive model makes a prediction about values of data using known results found from different data while the Descriptive model identifies patterns or relationships in data. Unlike the predictive model, a descriptive model serves as a way to explore the properties of the data examined, not to predict new properties. Predictive model data mining tasks include classification, prediction, regression and time series analysis. The Descriptive task encompasses methods such as Clustering, Summarizations, Association Rules, and Sequence analysis. So in this paper we will do classification and clustering of data on data set present in HDFS using the data mining algorithm. Like SOM (Self Organizing Maps), K-Means, Apriori.

Keywords: Big Data, Data Mining, Clustering, Classification, SOM (Self Organizing Maps), K-Means, Apriori.

I. INTRODUCTION

Anything that we requires in this hi-tech generation or you can say that is unaware to us then we go for Google and within few second we got several results according to entered queries. This may be a better example of big Data. Some of the day today example where the Big data used are like Our GPS is available because of big data. Thousands of reports and other maps are scanned in and used to make our GPS devices as accurate as possible. By bringing in data from incident reports, construction zone areas and individual data from apps, GPS devices are now more trustworthy than in the past. Other example of Medical Record, Medical records is now being put into computers to create electronic records for hospitals and doctors. This allows easier access to medical histories and helps doctors detect trends across all of the data. While this may feel like it jeopardizes patient privacy, take into account how doctors will be able to determine effectiveness of treatments better than ever before.

Big data is massive volume of both structured and unstructured data from various sources such as social data, machine generated data, traditional enterprise which is so large that it is difficult to process with traditional database and software techniques. Characteristics of Big Data include 4 Vs. They are Volume, Velocity, Variety and Veracity.

The manipulation on data set is done by taking all the required data in a relational database and performing the activity on data base instead of performing the activity on data set. Due to this the time and the complexity consumed increase and become difficult in handling the big data.

Data Mining is the knowledge discovery it is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data according to the user requirement. Data mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable ore. The objective of data mining is to identify valid novel, potentially useful, and understandable correlations and patterns in existing data . The main task of data mining is divided as either predictive or descriptive in nature. A ***Predictive*** model makes a prediction about values of data using known results found from different data while the ***Descriptive*** model identifies patterns or relationships in data. ***Predictive*** model data mining tasks include classification, prediction, regression and time series analysis. The ***Descriptive*** task encompases methods such as Clustering, Summarizations, Association Rules, and Sequence analysis. In this paper our aim is to focus on Classification and Clustering of data using SOM(Self Organizing Maps), K-Means, Apriori.

II. LITERATURE REVIEW

A variety of researches focusing on knowledge view, technique view, and application view can be found in the literature. However, no previous effort has been made to review the different views of data mining in a systematic way, especially in nowadays big data [1–2]; mobile internet and Internet of Things [3–4] grow rapidly and some data mining researchers shift their attention from data mining to big data. There are lots of data that can be mined, for example, database data (relational database, NoSQL database), data warehouse, data stream, spatiotemporal, time series, sequence, text and web, multimedia [5], graphs, the World Wide Web, Internet of Things data [6–7], and legacy system log. Motivated by this, in this paper, we attempt to make a comprehensive survey of the important recent developments of data mining research. This survey focuses on knowledge view, utilized techniques view, and application view of data mining. Our main contribution in this paper is that we selected some well-known algorithms and studied their strengths and limitations.

Han et al. provided a comprehensive survey, in database perspective, on the data mining techniques developed recently [8]. Several major kinds of data mining methods, including generalization, characterization, classification, clustering, association, evolution, pattern matching, data visualization, and meta-rule guided mining, was reviewed by them. Techniques for mining knowledge in different kinds of databases, included relational, transaction, object- oriented, spatial, and active databases, as well as global information systems, was examined by them [8].

Clustering is the most commonly used technique of data mining under which patterns are discovered in the underly-ing data.[9] Sidhu et al. presented that how clustering was carried out and the applications of clustering. They also provided us with a framework for the mixed attributes clus-tering problem and also showed us that how the customer data can be clustered identifying the high-profit, high-val-ue and low-risk customer [9].

KarthikeyanT. And Ravikumar N.(2014) In this paper they reviewed algorithmic approach of association rule mining and observe that a lot of attention was focused on performance and scalability of the algorithm, but the quality of rule generated are poor. So to enhance the quality of rule, to reduce the execution time, complexity and improve the accuracy they enhance the algorithm.
Sakthi et al. [10]. Discuss in this paper that due to the increment in the amount of data across the world, analysis of the data turns out to be very difficult task. To understand and learn the data classify those data into remarkable collection. So , there is a need of data mining techniques.

[11] Shafeeq et al present the modified K-means algorithams to improve the cluster qulity and to fixed the optimal number of cluster. As input number of cluster (K) given to the K-means algorithms by the users. But in the practical scenario, it is very difficult to fix the number cluster in advance. The method proposed in this paper works for both the cases i.e. for known number of clusters in advance as well as unknowm number of

clusters. The users has the flexibility either to fixed the number of cluster or input the minimum number of cluster required . the new cluster centers are computed by the algoritham by incremeny yhe cluster counter by one in each iteration until it satisfied the validity of cluster quality. This algoritham will over come this problems by finding the optimal number of cluster on the run. The proposed approach takes more computational time than the K-means for larger data sets. It is the major drawback of this approach.

[12] Libao ZHANG et al. Proposed simple and qualitative methodology using k-means clustering algoritham to classify NBA guards and used the euclidean distance as a measure of similarity distance. This work displya by using k-means clustering algorithm and 120 NBA guards data. Manual classification of traditional methods is improved using this model. According to the existict statical data, the NBA players are classifed to classsification and evolution objectively and scintifically. This work shows that this is very effective and reasonable methodology. Therefore based on classification results the guards type can be defined properly. Meanwhile, the guards function in the team can be evaluated in a fair and objective manner.

III. METHODOLOGY

All propose model explore on the big data issue which are created during the manipulation of data done in data base. Here in this paper the main goal is to focus on classification and clustering of data done on HDFS file system in Hadoop platform. The tradition method of manipulation done on Big data was done by retrieving the required data in database due to this time and complexity were increases.

In this paper we had acquired the wall mart dataset and copied the file into the HDFS. From that Datasets field are retrieved and various processes are applied on it like. Here we had focused on classification and clustering on data set. For this classification and clustering of data we had used the standard algorithm of Data mining like SOM (Self Organizing Maps), K-Means, Apriori.

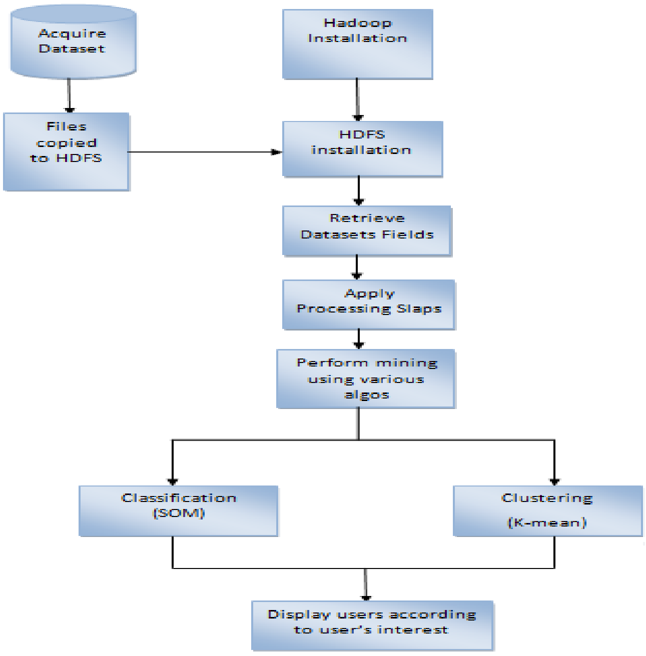

Figure 1. Flow for classification and clustering

Aim of the data mining technique In BIG DATA.

1. Classification: Classification is the process related to categorizing the process in which ideas and object are recognize. Classification is the processing of finding a set of models which describe and distinguish data classes or concept . Here we are doing classification with the help of SOM algorithm used in data mining process. This classification done on data set in Hadoop system help to find the set of model easily.

2. Clustering: Clustering is the process of partitioning or grouping a given set of patterns into disjoint *clusters*. This is done such that patterns in the same cluster are alike and patterns belonging to two different clusters are different. Clustering has been a widely studied problem in a variety of application domains including neural networks, AI, and statistics. The k-means method has been shown to be effective in producing good clustering results for many practical applications. However, a direct algorithm of k-means method requires time proportional to the product of number of patterns and number of clusters per iteration. So we r using the K mean algorithm form clustering of data.

IV. CONCLUSION

There is a frequent need of retrieving and manipulating a data according to the user convenience. Mostly performing a manipulation on data set in big data is done by retrieving the data in the database system .Due to retrieval of data in a database it consume time and complexity . So to reduce this factor we are using the Data mining technique . Here we mostly focused on the Classification and clustered of Data set which increased the robustness. The Classification and clustering of set in data sets is done by using The data mining algorithm like SOM used for classification, k-mean done for clustering and aproior done for association.

V. REFERENCES

[1] Zhang Y., Chen M., Mao S., Hu L., Leung V.CAP: crowd activity prediction based on big data analysisIEEE Network2014284525710.1109/mnet.2014.6863132 Google Scholar CrossRef

[2] Chen M., Mao S., Zhang Y., Leung V.Big Data: Related Technologies, Challenges and Future Prospects2014SpringerSpringerBriefs in Computer Science Google Scholar CrossRef

[3] Wan J., Zhang D., Sun Y., Lin K., Zou C., Cai H.VCMIA: a novel architecture for integrating vehicular cyber-physical systems and mobile cloud computingMobile Networks and Applications201419215316010.1007/s11036-014-0499-62-s2.0-84898828128 Google Scholar CrossRef

[4] Chen F., Rong X.-H., Deng P., Ma S.-L.A survey of device collaboration technology and system softwareActa Electronica Sinica20113924404472-s2.0-79955052781 Google Scholar

[5] Zhou L., Chen M., Zheng B., Cui J.Green multimedia communications over Internet of ThingsProceedings of the IEEE International Conference on Communications (ICC '12)June 2012Ottawa, Canada1948195210.1109/icc.2012.63639092-s2.0-84871967365 CrossRef

[6] Deng P., Zhang J. W., Rong X. H., Chen F.A model of large-scale Device Collaboration system based on PI-Calculus for green communicationTelecommunication Systems20135221313132610.1007/s11235-011-9643-92-s2.0-84879603230 Google Scholar CrossRef

[7] Zhang J., Deng P., Wan J., Yan B., Rong X., Chen F.A novel multimedia device ability matching technique for ubiquitous computing environmentsEURASIP Journal on Wireless Communications and Networking201320131, article 1811210.1186/1687-1499-2013-1812-s2.0-84894120909 Google Scholar CrossRef

[8] Han, Jiawei. "Data mining techniques." In ACM SIGMOD Record, vol. 25, no. 2, p. 545. ACM, 1996

[9] Sidhu, Nimrat Kaur, and Rajneet Kaur. "Clustering In Data Mining.

[10] Sakthi, M. Thanamani. A, " An Enhanced K Means Clustering using improved Hopfield artificial neural network and genetic algorithm:, international jouranal of recent technology and engineering (IJRTE) ISSN: 2277-3878, Vol-2, 2013

[11] Shafeeg a., Hareesha K., "Dynamic clustering of data with modified K-means algorithams" International conference on Information and Computer Networks, vol. 27, 2012

[12] Libao ZHANG, Faming LIU, Pingping GUO, Cong LIU," application of K-means clustereing algoritham fpr classification of NBA guards", international jouranal of science and engineering application volumn 5 issue1, 2016, ISSN-2319-7560(Online).