ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

# Improving the Quality and Accuracy of Clustering Algorithms on Big Data Sets Using Hadoop

Mr. K. Naveen Kumar
Student at Software Engineering,
Dept of IT,
Sree Vidyanikethan Engineering College.

Mr. A. Srinivasulu, M.Tech, (Ph.D)
Associate Professor,
Dept of IT,
Sree Vidyanikethan Engineering College.

## ABSTRACT:

Big Data is very drastically every day growing the firms and other data pools. Importance of Big data is enhancing the power of productivity, technology, analysis, design, business intelligence and data mining. In big data having drawbacks are capture, storage, search, sharing, analysis, visualization, less accurate patterns, less resolution, less performance and less quality of the clustering results. The proposed system uses this emerged the necessity of Big data techniques as Hadoop. The cluster problems such as cluster characteristic or behavior, divide into parts, cluster validity or rationality, and cluster performance or presentation, scalability. Finally to evaluating the accuracy and quality of Hadoop clustering and Map reduce with Data sets such as Face book, Twitter, Temperature and Geological data.

## Keywords:

Clustering, preprocessing, Map-Reduce, Big data techniques, Hadoop, Data sets.

## 1. INTRODUCTION:

Data generated in different organization, industries and other data pools. It is all about better Analytic on a broader spectrum classify something in terms of data, and therefore capability and possible to create even more differentiation among industry peers. Big Data-Facts are every day creates 2.5 quintillion bytes of data. Big Data Similar to 'Small-data' but bigger data requires Techniques, Tools, Architecture, Big Data Big Data Is more than just a DW that requires to store/analysis large volume of data.

The volume big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture or gathering, curation or activity or process, data base maintained, search, sharing, transfer, analysis, and visualization. The trend methods and technology to large amount of datasets is due to the additional information derivable from analysis of a single large set of relevant, same data is five myths 1. Big Data is about Massive Data Volume 2. Big Data run on

Hadoop3. Big Data Means Unstructured and structured data 4. Big Data is for Social Media Feeds and Sentiment Analysis 5. NoSQL means No SQL. Big data is the opportunity to extract insight from an immense volume, variety and velocity of data.The clusters canter residing in one map may have points within its purview that is part of the input data to another map function. The outputs from the reducer are then fed into the appropriate mapper and reducer to begin the next round of processing. Hadoop Mahout is that implements developments some of the clustering and classification algorithms which have been modified to fit the Map-Reduce model.

The Mahout implementations have been deployed within Apache Hadoop a Map Reduce based cloud runtime. Mahout has been designed to work specifically with Hadoop, there is nothing to preclude using the Mahout library within data pre-processing database system to supports the MapReduce paradigm and Clustering algorithms are an unsupervised machine learning technique that facilitates the creation of clusters, which allow us to group similar and relevant items (also called observations) together so that these clusters are similar in some definition of big data.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Clustering has more broad large applications in part of big data areas and industries such as data mining, large pre pre-exitsting database system, guidance of recommendation systems, design pattern recognition, identification of understanding Clustering algorithms have certain unique characteristics Hadoop environment.

## 2. Related work:

Big data is not equal to Hadoop. Hadoop is just its subset of working of process of big data. There are much more areas and particular items being part of area. Many people think big data is about Hadoop technology. The key requirements are to understand and navigate identity sources of big data to discover data base in that. New technologies have emerged on design of Hadoop big data to creation of catalogs, research and handle diverse sources of big data. Hadoop is a collection of open source capabilities. Two of the most outstanding on Hadoop File System for storing a variety of information and map reduce a parallel processing engine. Data warehouses to maintained big data- the volume of un structured to change the structured data is growing quickly implanted some of new application. The ability to run deep analytic and difficulty queries on huge volumes of structured data is a big data problem is converting to unstructured to structure. It requires massive parallel processing data warehouses and purpose-built appliances for deep analytics. Big data is just at resets also in motioned data to up loading, updated Streaming data represents query on to retrieve the data an entirely different big data problem the enable to quickly analyze and act upon data while it's still moving and respond the database. This new technology opens a world of possibilities from processing volumes of data that were just not practical to store on big data analytics' to solve and detecting the problems insight and responding quickly of database on user . As much of the worlds big data is unstructured and in data content, data analytics is a difficult component to analyze and derive meaning from text. Integration and governance of technology implements establishes the veracity of big data, and is critical in reminder whether information on big data is trusted. The reason that big data is currently a hot topic is partly due to this technology.

## 3. Background &Literature review:

Bigdata sets using clustering on Hadoop mahout, the technical Support call centers frequently receive several thousand customer queries on a daily basis.

Traditionally, such organizations discard data related to customer enquiries within a relatively short period of time due to limited storage capacity. The value of retaining and analyzing, information has become clear, enabling call centers' to identify customer patterns, improve first call resolution and maximize daily closure rates. Hadoop programming model, extended ecosystem of Mahout Big Data Analytics library for categorizing similar support calls for large technical support data sets.The clustering and classification on large data sets on MapReduce to work of Big data processing is currently becoming increasingly important in modern due to the continuous growth of data generated by various fields such as particle physics, human genomics, earth observations etc. However the efficiency of processing large-scale data on modern infrastructure is not clear of large set of data has to be taken for analysis as size of data grows in exponentially. Hadoop is one such framework that offers distributed storage and parallel data processing to build a combined clustering and classification model that run on Hadoop to process Bigdata. To optimize the performance of Bigdata analysis by integrating clustering and classification concepts with map reduce paradigm of Hadoop architecture.

The parallel power iteration clustering Distributed Data Mining is most popular topic in research area because as data are increasing in day to day life there are so many problems occurs to handle them and there are also a solutions for that but still they are not as per expectation, still there are some issue already there in the Distributed Data Mining, among them mainly we are focus in this papers that about reducing computational. Bigdata in hierarchical clustering MapReduce is a software framework that allows certain kinds of parallelizable or distributable problems involving large data sets to be solved using computing clusters in internet users by mining a huge volume of web access log of up to 500 gigabytes.The hierarchical clustering algorithm used MapReduce, a parallel processing framework over clusters on dataset. Implementation of the algorithms suffers from efficiency problem memory and higher execution time large datasets with Map-Reduce dimension of feature vectors and eliminate noise features. MapReduce is the most popular cloud computing programming model nowadays. It provides programmers with powerful APIs by encapsulating the details of parallelism, fault tolerance and load balancing. So cluster to improve the efficiency of time-consuming applications is an economical and efficient solution.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

## 4. Implementation:

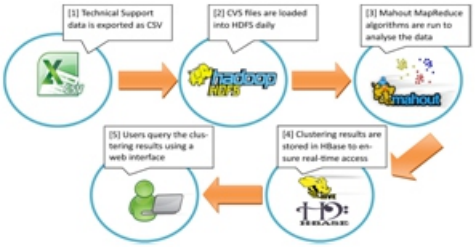


Fig 1: process of implementation

## Apache Hadoop:

Open-source software for dependable scalable, distributed computing Allows for the distributed processing of large data sets across clusters of computers Designed to scale up from single servers to thousands of machines offering local computation and storage High-availability on top of a cluster of computers. Apache Hadoop Modules Hadoop Common: The most common utilities that support the other Hadoop modules using to doing the MapReduce and clustering datasets.

## HBase:

Table storage for structured data Modeled on Google's Bitable Row/column store Billions of rows/millions on columns Column-oriented - nulls are free. To need random write, random read or both (but not neither) need to number of of operations per sec on multiple TB of data, to access patterns are simple Column – oriented database (column families)

» Table consists of Rows, each which has a primary Key (row key)
» Each Row may have any number of columns
» Table schema only defines Column familes
» Each cell value has a timestamp
» Get (row)
» Put (row, Map<column, value>)
» Scan (key range, filter)
» increment (row, columns)
» Check and Put,

### Read one column value from a row

Cell = table. get ("test_row1", "columnfamily1:column1");
To read one row with any columns, use HTable#getRow ()method. RowResultsingleRow=table.getRow(Bytes. toBytes ("test_row1"));

## 5. Results:

| Clustering Techniques | Mean | Min | Max | SD |
|---|---|---|---|---|
| k-Means | 5122.81 | 5002.786 | 5683.732 | 137.581 |
| Canopy | 8423.78 | 8414.22 | 8431.151 | 5.892 |
| Fuzzy k-Means | 2266.084 | 2212.842 | 2294.902 | 25.081 |
| Dirichlet Allocation | 31933.61 | 31826.97 | 32332.79 | 131.541 |

Fig 2: Cluster quality of Hadoop Clustering Algorithms

The output of the clustering algorithms is now examined as well as the accuracy of the clusters. Performance of evolution of quality improve of clustering results by using of Hadoop on big data sets, in that to any data sets to analyze and read the file to write the map reduce programming data sets after results to perform the Hadoop mahout clustering. In that results to store the data sets to push the data in HBase or Hive, to retrieve data them to write Nosql command in Hive or HBase see the comparison of clustering process and results all considered clustering algorithms for clusters on data sets related to evaluation of performance issues and clarity of results. The same of the results similar as k-means, fuzzy k-means with a pre-processing, to evaluate the clustering capturing results, evalution of mean,min,max,sd execution time of Hadoop clustering,mapreduce perfomnce results It can be fined out the results are k-means, fuzzy k-means with Hadoop using The number of data sets included in the clusters is comparable using all techniques. The output of the probability of clustering results more performance of evaluation, clarity of clustering results on data sets with use of Hadoop.

## 6. Conclusion

Clustering is the well designed as it may easily overcome many clustering issues such as cluster tendency, cluster partition and cluster validity. Although they are many clustering algorithms in Mahout and Weka are unwanted elements to effective ones. In the mahout it performs clustering algorithms and evaluated, and also Cluster Cohesiveness and Cluster Quality of all the clustering algorithms in Hadoop Mahout by giving the regular Bigdata sets.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

To evaluate the datasets in data pre-processing and Hadoop clusters and Hadoop map reducing on dataset to calculate the accurate the results and decries the evaluation time of data sets. Mahout is capable of managing large amount of data as it supports scalability, its clustering algorithms are well designed as it may easily overcome many clustering issues such as cluster tendency, cluster partition and cluster validity. Although, they are many clustering algorithms on Mahout, now they are refined to effectiveness evaluated the performance, Cluster cohesive and good Quality results of all the clustering algorithms in Hadoop by giving the Standard Bigdata datasets.

## References:

[1]Arantxa Duque Barrachina1 and Aisling O'Driscoll, "A big data methodology for categorizing technical support requests using Hadoop and Mahout", Springer Journal of Big Data, Vol 1:1, 2014.

[2]Somesh S Chavadi, Dr. Asha T, "Text Mining Approach for Big Data Analysis Using Clustering and Classification Methodologies", International Journal of Emerging Technology and Advanced Engineering , Volume 4, Issue 8, August 2014.

[3]Ankit Darji , Dinesh Waghela, "Parallel Power Iteration Clustering for Big Data using Map Reduce in Hadoop", International Journal of Advanced Research in Computer Science and Software Engineering , Volume 4, Issue 6, June 2014.

[4]Vadivel.M, Raghunath.V, "Enhancing Map-Reduce Framework for Big Data with Hierarchical Clustering", International Journal of Innovative Research in Computer and Communication Engineering, Vol.2, Special Issue 1, March 2014.

[5]Kathleen Ericson, Shirideep allickara, "On the performance of high dimensional data clustering and classification algorithms" Future Generation Computer Systems, Elsevier 2013 - doi: 10.1016.

[6] Amita Verma, Ashwani kumar, " Performance Enhancement of K-Means Clustering Algorithms for High Dimensional Data sets" International Journal of Advanced Research in Computer Science and Software Engineering Vol 4, Issue 1, January 2014.

[7]Krishnakumar.K, Ganesh Karthikeyan.V, "Analysis and Performance Evaluation of Large Data Sets Using Hadoop" IJournal of Research in Advent Technology, Vol.2, No.5, May 2014.

[8]Harish Chauhan, "Harnessing Hadoop: Understanding the Big Data Processing Options for Optimizing Analytical Workloads" cognizant 20-20 insights white paper -|November 2013.

[9] Yun Lu, Ming Zhao, Lixi Wang and Naphtali Rishe "v-TerraFly: large scale distributed spatial data visualization with autonomic resource Management", Springer Journal of Big Data, 1:4,2014

[10]Aloysius George, "Efficient High imension Data Clustering using Constraint-Partitioning K-Means Algorithm" The International Arab Journal of Information Technology, Vol. 10, No. 5, September 2013.

[11]Durai Raj, "Un-moderated real-time news trends extraction from World Wide Web using Apache Mahout", Project Report, San Jose University, USA,2011.

[12]Esteves R M and Chumming Rong, "k-means clustering for the cloud : A Mahout Test", Advanced Information Networking and Applications (WAINA), 2011 IEEE of International Conference , PAGE NO: 514 - 519 , 22-25 March 2011.

[13]Carlos E.Seminario and David C. Wilson, "Case Study Evaluation of Mahout as a Recommender Platform", Workshop on Recommendation, Utility Evaluation: Beyond RMSE (RUE 2012), held in conjunction with ACM RecSys 2012. September 9, 2012, Dublin, Ireland.

[14]Dr. Senthil kumar and John Paul, "Performance analysis of various recommendation algorithms using Apache Hadoop and Mahout", International Journal of Scientific & Engineering Research, Volume 4, Issue 12, ISSN 2229-5518, December-2013.

[15]J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," ACM Communication., vol. 51, pp. 107-113, Jan. 2008.

[16]Berkhin, "A Survey of Clustering Data Mining Techniques," in Grouping Multidimensional Data: Recent Advances in Clustering, J. Kogan and C. K. Nicholas, Eds., ed: Springer, pp. 25-83, 2006.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

## ABOUT RESOURCE PERSON:

**K.Naveen Kumar** received the B.Tech (CSE) Computer Science Engineering from JUNTU Anthapur, Andhra Pradesh in the year 2012 and Pursuing his M.Tech (SE) in Software engineering from Sree Vidyanikethan Engineering College (Autonomous)., Tirupati,Andra Pradesh, He is a professional member of Computer society of India . His research interests include area of interest includes Big Data Analytics, Data Mining and Data warehousing, Cloud Computing, Intelligent Systems.

**Asadi Srinivasulu** received the B.Tech (CSE) Computer Science Engineering from Sri Venkateswara University, Tirupati, Andhra Pradesh in the year 2000 and completed his M.Tech (IS) in Intelligent Systems from Indian Institute of Information Technology (IIIT), Allahabad, Uttar Pradesh, in the year 2004. Currently he is pursuing his Ph.D from J.N.T.U.A, Anantapur and He is submitted the Thesis in 2014. He has got 12 years of teaching and industrial experience. He served as the Head, Dept of Information Technology, S.V.College of Engineering, Karakambadi, Tirupati during 2007-2009. He has 07 IBM Professional Certifications to his credit such as RAD, RFT, DB2, RTC, TDS, LOTUS Domino and WID. He is a professional member of CSI, ACM, ISTE, IAENG, IACSIT, ICST, SCIEI, VSRDIJ, NASSCOM, ASDF and ICA (Indian Congress Association. He is a professional reviewer of CSI, Data Mining & Knowledge Management, IAENG and VSRD-CSIT. He got best Mentor award from IBM-TGMC-2011, 2012. He has published more than 48 papers in International Journals and Conferences. He has attended more than 70 Workshops, Symposiums and Seminars. He has conducted and acted as resource person more than 22 Conferences, Workshops, Symposiums and Seminars. Some of his publications are listed in digital libraries such as IEEE Xplore, Elsevier Xplore, IJCA, Bookman International Journal(Free), IJAIS, IJCSNS, JACRIJ, DOAJ, PDF DIGITAL LIBRARY, IJCSET, IJCSIT, IJARCET, IJARCSSE, Horizon Publication, CSI and COGNIZANCE Journals. He visited many countries like Malaysia, Thailand and Singapore.