Competence and Challenges of Big-Data in Terms of Materializing Technologies

U. Ramya and Pradheeba Ulaganathan

Department of CSE, PKIET, India
Department of CSE, RMKCET, India

Abstract: Big Data is an arrangement of procedures and innovation that obliges new types of incorporation to reveal expansive concealed qualities from vast datasets that are different, complex and of an enormous scale. Hadoop gives an excellent framework for all the beginners to explore the competence of Big data. There is a perception about Map reduce framework that creates huge measure of moderate data. With expanding size of data in Data warehouse it is extravagant to perform data analysis thus for data handling, system transfer on bunch ideas and parallel execution is given by Map-Reduce. For every stage, all the above have appended an overview table and continued by examining the specialized difficulties. So this paper gives a glance over the competence of various Big – data and its related technologies for storage and application and thus by concluding the what has to be researched in the new era of Big-data.

Key words: Applications of Big data • Clustering • Hadoop • Map reduce • Storage

INTRODUCTION

Big data is a set of technique that requires new forms of integration to uncover large hidden values from large datasets that are diverse, complex and of a massive scale. Recently, industries become interested in the high potential of big data and many government agencies announced major plans to accelerate big data research and applications. In addition, issues on big data are often covered in public media, such as The Economist, New York Times and National Public Radio. Two premier scientific journals, Nature and Science, also opened special columns to discuss the challenges and impacts of big data. The era of big data has come beyond all doubt. Clustering is the method of collecting objects into a cluster which is none other than group in which every member in that cluster is unique in some manner. With the increasing size of data in data warehouse it is expensive to perform data analysis. For data processing framework relay on cluster computers and parallel execution framework provided by Map Reduce. It is an important task in data analysis and data mining applications. The main challenges in handling Big Data lie not only in the four V’s, namely, huge Volume in amount, high Variety in type, Velocity in terms of real-time requirements and Variability, i.e., constant changes in data structure and user interpretation, but also in the approach to understanding data. Therefore, Big Data calls for a revolutionary change in research methodology and in tools to be employed as well. As the development of cloud computing technologies is still at an early stage, hope that this paper will provide a better understanding, Hadoop is a java-based software framework that empowers data-intensive application in a distributed environment. Distributed data processing & storing technique has been used in this Hadoop. Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources. It is a scalable, open source, fault-tolerant Virtual Grid operating system architecture for data storage and processing. It runs on Commodity hardware, it uses HDFS which is fault-tolerant high-bandwidth clustered storage architecture. It runs Map-Reduce for distributed data processing and is works with structured and unstructured data.

Era of Big-Data: Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex and of a massive scale. Big data usually includes data sets

Corresponding Author: U. Ramya, Department of CSE, PKIET, India.
with sizes beyond the ability of commonly used software tools to capture, accurate, manage and process data within a tolerable elapsed time. Big data "size" is a constantly moving target, as of ranging from a few dozen terabytes to many petabytes of data. Big data can be described by the following characteristics,

a) **Velocity**: Velocity means the speed of generation of data and also how fast the data is generated and processed to meet the demand which will be on the path of growth and development.

b) **Veracity**: This will be useful since the mined data can vary greatly and using this the accuracy of the source data is observed.

c) **Variety**: The other aspect of big data is variety and it is useful in analyzing the data and so the data can be used efficiently by knowing the advantage.

d) **Variability**: It will be referring to the inconsistency and thus process of being handling and managing the data effectively. The following table gives a glance over the challenges and competence in Big-data Mining.

<table>
<thead>
<tr>
<th>Algorithm Name</th>
<th>Description</th>
<th>Competence</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. C4.5 and beyond</td>
<td>It generates classifiers expressed as decision trees in a comprehensible rule set form.</td>
<td>The primary one is allows two or more outcomes</td>
<td>The amount of cpu time and memory they require.</td>
</tr>
<tr>
<td>2. k-means algorithm</td>
<td>It is a simple iterative method to partition a given dataset into user-specified no of clusters k.</td>
<td>It is effective for much larger class of datasets with increasing k</td>
<td>It is sensitive to the presence of outliers.</td>
</tr>
<tr>
<td>3. Support Vector Machines (SVM)</td>
<td>The main aim of SVM is to obtain an optimal separating hyper-plane with the maximum margin (w) and a real value for organization of data.</td>
<td>It offers best generalization ability.</td>
<td>It is difficult to interpret computational efficiency.</td>
</tr>
<tr>
<td>4. Aprior algorithm</td>
<td>It is a seminal algorithm for finding frequent item sets using candidate generation.</td>
<td>It is quite simple and easy to implement.</td>
<td>It has cost suffering.</td>
</tr>
<tr>
<td>5. EM algorithm</td>
<td>It focus on the use of normal mixture models and to estimate the underlying density function.</td>
<td>It does not depend on irrelevant factors.</td>
<td>Regularity conditions do not hold for the likelihood ratio test.</td>
</tr>
<tr>
<td>6. Page rank</td>
<td>It is a search ranking algorithm using hyperlinks on the web.</td>
<td>It is simple. Actual convergence may not be necessary.</td>
<td>Fewer iterations are needed.</td>
</tr>
<tr>
<td>7. Ada Boost</td>
<td>To deal with multi-class problem was proposed.</td>
<td>It is simple and easy for learning.</td>
<td>This algorithm does not over fit. The original data is usually lost.</td>
</tr>
<tr>
<td>8. KNN(k-nearest neighbor algorithm)</td>
<td>It is one of the simple algorithm which performs classification only if the attributes of the object match.</td>
<td>It is particularly well suited for multi-model class.</td>
<td>It has many tests records will not be classified because they do not exactly match any of the training records.</td>
</tr>
<tr>
<td>9. Naive Bayes</td>
<td>The aim is to construct a rule which will allow us to assign future object to class.</td>
<td>It is readily applied to huge data sets. It is easy to interpret.</td>
<td>It have a complicated non-linear shape.</td>
</tr>
<tr>
<td>10. CART</td>
<td>It is a binary recursive partitioning procedure capable of processing nominal attributes both as targets &amp; predictors.</td>
<td>Class balancing. Dynamic feature construction.</td>
<td>The estimation is based on training data in the terminal nodes which is to be biased.</td>
</tr>
</tbody>
</table>

**Challenges of Big-Data**

![Big Data Processing Framework](Image)

Fig. 1: Big – Data Processing Framework
The challenges at tier I focus on low-level data accessing and arithmetic computing procedures. Challenges on information sharing and privacy. Tier II concentrate on high-level semantics, application domain knowledge for different applications of big data and the user privacy issues. The Outmost tier is tier III which challenges the actual mining algorithms. Tier III contains three stages. In first stage sparse, heterogeneous, uncertain, incomplete and multisource data is preprocessed by data fusion technique. In second stage after preprocessing stage complex and dynamic data are mined. Third stage is for local learning and model fusion, where the global knowledge is obtained. The other aspects of big-data are as discussed in the following sections with their competence and challenges (CC) in each section [2][1].

Big Data and Clustering – CC: Big data is a broad and a large term that is used for a massive volume of structured and unstructured data and because of this it is so complex to solve and process using usual databases and software. Big data can be explained by the characteristics such as volume, velocity, variety, variability, veracity and complexity. So clustering helps in storing large amount of data which is highly related with clustering mainly named as hadoop cluster[7][1].

Big Data and Power Iteration Clustering: Parallel Power Iteration Clustering is execute for minimizing correspondence cost, Due to its parallel approach distinctive hubs are likewise their so for make our framework more hearty and evade hub disappointment here we utilize Hadoop stage utilizing Map Reduce. Hadoop give nature to execute application in appropriated environment and it can equipped for taking care of hub disappointment. Because of making reproduction we fathom the issue of hub disappointment of existing framework. In future we can utilize another approach for bunching the information and check the execution by actualizing it in Hadoop furthermore we can chip away at. Hadoop framework in which (1) If Name Node will fall flat than how to deal with data (2) Try to create annex compose display in Hadoop (3) Extend the limit of NameNode in light of the fact that if A namespace with a to a great degree vast number of records surpasses NameNode ability to keep up.

Step 1: Choose \( k \) and similarity function \( s \)
Step 2: Derive \( A \) from \( s \), let \( W=I-D^{-1}A \), where \( I \) is the identity matrix and \( D \) is a diagonal square matrix \( D_{ii} = \sum_j A_{ij} \)

Step 3: Find eigenvectors and corresponding eigen values of \( W \)
Step 4: Pick the \( k \) eigenvectors of \( W \) with the \( 2^{nd} \) to \( k^{th} \) smallest corresponding eigen values as “significant” Project Eigen vectors the data points onto the space spanned by these vectors
Step 5: Run \( k \)-means on the projected data points

CC: Power iteration clustering is one of the simplest and scalable method, when compared to spectral type of clustering method the result which was produced by power iteration method is better at a very low cost. The main advantage of power iteration clustering is that of low cost when compared to spectral clustering and there is no need for predefine the number of clusters [12].

Big Data And K-Means Clustering Algorithm: K-means clustering algorithm is one of the most popular clustering algorithms which was used in metric spaces. At first \( k \) cluster centroids are selected at random. Then the k-means algorithm reassigns all the points to their nearest centroids and recomputed centroids of the newly assembled groups. The iterative relocation continues until the criterion function, e.g. square-error converges. Although it is ofwide popular it has some disadvantages which are \( k \)-means is very sensitive to noise and outliers since a small number of such data can substantially influence the centroids. Other are sensitivity to initialization, entrapments into local optima, poor cluster descriptors, inability to deal with clusters of arbitrary shape, size and density, reliance on user to specify the number of clusters[5].

The Algorithm steps are

Step 1: Choose the number of clusters, \( k \).
Step 2: Randomly generate \( k \) clusters and determine the cluster centers, or directly generate \( k \) random points as cluster centres.
Step 3: Assign each point to the nearest cluster center.
Step 4: Recomputed the new cluster centers.
Step 5: Repeat the two previous steps until some convergence criterion is met (usually that the assignment hasn't changed).

CC: K-means clustering algorithm has many competences over others like fast, robust and easier to understand. Its relatively efficient: \( O(tknd) \), where \( n \) is # objects, \( k \) is # clusters, \( d \) is # dimension of each object and \( t \) is # iterations. Normally, \( k, t, d \ll n. \) it gives best result when data set are distinct or well separated from each other.
The learning algorithm requires apriori specification of the number of cluster centers. The use of Exclusive Assignment - If there are two highly overlapping data then k-means will not be able to resolve that there are two clusters. Euclidean distance measures can unequally weight underlying factors. The learning algorithm provides the local optima of the squared error function. Randomly choosing of the cluster center cannot lead us to the fruitful result. Applicable only when mean is defined i.e. fails for categorical data. The algorithm unable to handle noisy data and outliers. Algorithm fails for non-linear data set.

Big Data And K-Medoids Clustering Algorithm: K-medoids or Partitioning Around Medoids (PAM) method a cluster is represented by its medoid that is the most centrally located object (pattern) in the cluster. Medoids are more resistant to outliers and noise compared to centroids. PAM begins by selecting randomly an object as medoid for each of the k clusters. Then, each of the non-selected objects is grouped with the medoid to which it is the most similar. PAM then iteratively replaces one of the medoids by one of the non-medoids objects yielding the greatest improvement in the cost function. Clearly, PAM is an expensive algorithm as regards finding the medoids, as it compares each medoid with the entire dataset at each iteration of the algorithm[12].

What to do with Hadoop and Big-Data?: In a matter of seconds, Hadoop is broadly utilized as a part of big data applications in the industry, e.g., spam sifting, system seeking, click stream analysis and social suggestion. In addition, considerable scholarly research is currently in the light of Hadoop. Some delegate cases are given beneath. As declared in June 2012, Yahoo runs Hadoop in 42,000 servers at four data focuses to bolster its items and administrations, e.g., searching and spam sifting and so on. At present, the biggest Hadoop bunch has 4,000 nodes, yet the quantity of nodes will be expanded to 10,000 with the arrival of Hadoop 2.0. In that month, Facebook reported that their Hadoop group can prepare 100 PB data, which became by 0.5 PB per day as in November 2012. Some surely understood organizations that use Hadoop to direct conveyed reckoning are listed in. Likewise, numerous organizations give Hadoop business execution and/or bolster, including Cloudera, IBM, MapR, EMC and Oracle. Among cutting edge mechanical apparatus and frameworks, sensors are broadly conveyed to gather data for environment checking and disappointment anticipating and so on.

Big-Data Storage: The data storage subsystem in a big data platform organizes the collected information in a convenient format for analysis and value extraction.

Storage System for Massive Data: Existing massive storage technologies can be classified as Direct Attached Storage (DAS) and Network storage. In Direct Attached Storage (Das), various hard disks are directly connected with servers and data management is server-centric, such that storage devices are peripheral equipment, each of which takes a certain amount of I/O resource and is managed by individual application software. Network Storage is to utilize network to provide users with a union interface for data access and sharing. It is characterized with strong expandability. Network Attached Storage (Nas) is actually an auxiliary storage equipment of a network and it is directly connected to a network through a hub or switch through TCP/IP protocols. In Storage Area Network (San), data storage management is relatively independent within a storage local area network, where multipath based data switching among any internal nodes is utilized to achieve a maximum degree of data sharing and data management.
**Storage Mechanism for Big-Data:** Existing storage mechanisms of big data may be classified into three bottom-up levels: File systems, Databases, Programming models. File Systems are the foundation of the applications at upper levels and Google’s GFS is an expandable distributed file system to support large-scale, distributed, data-intensive applications. Databases systems are developed to handle datasets at different scales and support various applications and MySQL databases are becoming the core technology for of big data. The following three main MySQL database are Key-value databases, column-oriented databases, document-oriented databases. (i) Programming Models: The programming model is critical to implementing the application logics and facilitating the data analysis applications. However, it is difficult for traditional parallel models to implement parallel programs on a big data scale.[11]. Many tools for big data mining and analysis are available, including professional and amateur software, expensive commercial software and open source software. Here we briefly review the top five most widely used software. R (30.7 %): R, an open source programming language and software environment, is designed for data mining/analysis and visualization. While computing intensive tasks are executed, code programmed with C, C++ and FORTRAN may be called in the R environment. Excel(29.8 %): A core component of Microsoft Office, provides powerful data processing and statistical analysis capabilities. Excel is also the only commercial software among the top five. Rapid-I Rapid Miner (26.7 %): Rapid miner is open source software used for data mining, machine learning and predictive analysis. Data mining and machine learning programs provided Rapid Miner include Extract, Transform and Load (ETL), data pre-processing and visualization, modeling, evaluation and deployment. Weka/Pentaho(14.8 %): Weka, abbreviated from Waikato Environment for Knowledge Analysis, is a free and open-source machine learning and data mining software written in Java. Weka provides such functions as data processing, feature selection, classification, regression, clustering, association rule and visualization, etc.[11]

**What has to be done?:** The above survey pointed out the various dimensions of Big-data. Each section describes the basic technology involved in storage, clustering etc. Look into on Big Data rose in the 1970s however has seen a blast of productions since 2008. In spite of the fact that the term is ordinarily connected with software engineering, the information demonstrates that it is connected to various controls including earth, wellbeing, building, expressions and humanities and ecological sciences. A more intensive take a gander at the ideas and subjects inside the edited compositions and titles after some time indicate how this region, which started as a PC and innovation center territory with some satellite applications, formed into a nearby and tight-weave train highlighting applications, techniques and imaginative arrangements running from could to disseminated figuring and concentrating on client encounter. The multi-dimensional normal for this subject is found in the writing and additionally in the web-based social networking and online productions. The idea of Big Data as an examination theme is by all accounts developing and it is likely that before the finish of 2012 the quantity of distributions will twofold, if not more and its investigation and applications will be seen in different orders. There is still much to do in building up an expert way to deal with information administration in science. The principle remarkable issues were tended to by the ICSU (International council for Science) Strategic Coordinating Committee on Information and Data noted above and incorporate the accompanying: better direction for best practice on information administration; enhanced meanings of the different terms utilized as a part of the expression "open get to"; more prominent acknowledgment of the distribution of information by researchers and additionally the production of diary articles and books; functional help in information administration for less monetarily created nations through association with individuals from the ICSU family and others; and collaboration with business organizations for common advantage. Huge Data presents science with many difficulties, however in the meantime displays numerous chances to impact how science develops and creates for the better, not slightest by adding information driven science to speculation driven science. Changes in expert information administration will bring about better science. The three aspects of computational science cyber infrastructure, interdisciplinary research and computational research have be considered important on the grounds at the most abnormal amounts and seen as a critical part of scholastic research. HPC facilities on grounds should be coordinated with national assets and give a pathway to grounds research to effortlessly associate with national and worldwide exercises. Instruction and preparing of understudies and personnel is pivotal; incomprehensible changes are required ever the little numbers at present came to through HPC focus instructional exercises; calculation and computational intuition should be a piece of new educational program over all orders.
CONCLUSION

Big Data is on the rise during the upcoming years and every data scientist has to accomplish much more range of data annually. This data is going to be more diverse, huge and faster. This paper has some insights about the topic and the main challenges for the future. Big Data is becoming the new Final Frontier for scientific data research and for business applications. In this paper, the background of big data mining and its related technologies such as clustering, hadoop, mapreduce, cloud computing are reviewed and also has compared various algorithm of big data clustering and map reduce by tabulating them.

REFERENCES

17. Halevi, G., Welcome to the 30th issue of Recent Trends in Big- Data.