IJCSN International Journal of Computer Science and Network, Volume 4, Issue 2, April 2015
ISSN (Online): 2277-5420 www.IJCSN.org
Impact Factor: 0.417

A Survey Paper on Big Data and Datamining Issues

Anuradha N. Nawathe

Amrutvahini College of Engineering, Sangamner. University of Pune, Ahemadnagar,
State Maharashtra, Country India.

Abstract - Big data is an issue that having large amount of data. To handle this data is a big challenge. In big data mining is collecting the needful information. Now a day's enterprise and its data is increasing rapidly also there is complexity of data is also increasing. In many areas such as biology, physics, medicine, astronomy, climate forecasting, etc. To find patterns and trends in the data is increasingly important and challenging for decision making. The analysis and management of Big Data is today a main concern of the computer science community, especially for complex data (e.g., images, graphs, audio and long texts). Big Data is a new term used to identify the datasets that due to their large size, we cannot manage them with the typical data mining software tools. Instead of defining "Big Data" as datasets of a concrete large size, for example in the order of magnitude of petabytes, the definition is related to the fact that the dataset is too big to be managed without using new algorithms or technologies.

Keywords - *Big data, Database, DataMining, Large Databse.*

1. Introduction

The databases used in many important and novel applications are often uncertain. For example, the locations of users obtained through RFID and GPS systems are not precise due to measurement errors. As another example, data collected from sensors in habitat monitoring systems (e.g., temperature and humidity) are noisy. Customer purchase behaviors, as captured in Supermarket basket databases contain statistical information for predicting what a customer will buy in the future. Integration and record linkage tools also associate confidence values to the output tuples according to the quality of matching. In structured information extractors, confidence values are appended to rules for extracting patterns from unstructured data. To meet the increasing application needs of handling a large amount of uncertain data, uncertain databases have been recently developed. A simple way of finding PFIs is to mine frequent patterns from every possible world, and then record the probabilities of the occurrences of these patterns. This is impractical, due to the exponential number of possible worlds. To remedy this, some algorithms have been recently developed to successfully retrieve PFIs without instantiating all possible worlds. Mining of uncertain big data is done by some algorithms. The Big Data is used in many applications such as Facebook, Ya- hoo!, Twitter, LinkedIn but the big data is also handle by Hadoop, Pig ,Hive and other software projects.

1.1 Comparison of Big data and Data Mining

DataMining: - Data mining is analyzing a data from different sources and produces it into meaningful information. Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. In DataMining there are following six activities:-

1. Classification
2. Estimation
3. Prediction
4. Association rules
5. Clustering
6. Description

A Classification is a process of generalizing the data according to different instances. Several major kinds of classification algorithms in data mining are Decision tree, k-nearest neighbor classifier, Naive Bayes, Apriori and AdaBoost. Classification consists of examining the features of a newly presented object and assigning to it a predefined class. The classification task is characterized by the well-defined classes, and a training set consisting of reclassified example. Estimation deals with continuously valued outcomes. Given some input data, we use estimation to come up with a value for some unknown continuous variables such as income, height or credit card balance.

A Prediction It's a statement about the way things will happen in the future, often but not always based on experience or knowledge. Prediction may be a statement in which some outcome is expected. Association Rules an association rule is a rule which implies certain association relationships among a set of objects (such as "occur together" or "one implies the other") in a database. Clustering can be considered the most important unsupervised learning problem; it deals with finding a structure in a collection of unlabeled data.

2. Big Data Mining and Analytics

[1] Big data is a broad term for datasets so large or complex that traditional data processing applications are inadequate. Big data is categories by three V's Variety, Volume and Velocity so mining and processing of this type of data requires some special kind of framework which is distributed and easily scalable. One such well known and popular framework is Hadoop. Hadoop is an open source distributed framework based on Google's Map Reduce paradigm. Hadoop utilizes the power of distributed systems to analyze big data. It is highly scalable and adding and administrating computing nodes in Hadoop is very simple and does not require any special knowledge. Data processing in Hadoop works in two phases The Map phase and The Reduce phase. Each phase will require the data in the form of key value and will produces the results in Key and Value. To handle the variety of data in Big Data Hadoop uses its own file system called as Hadoop Distributed File system (aka. HDFS). HDFS is distributed file system designed based on Google File System (GFS). Hadoop has been adopted by most of the well known enterprises like yahoo to process their data. Following diagram depicts the working of map reduce framework in Hadoop:

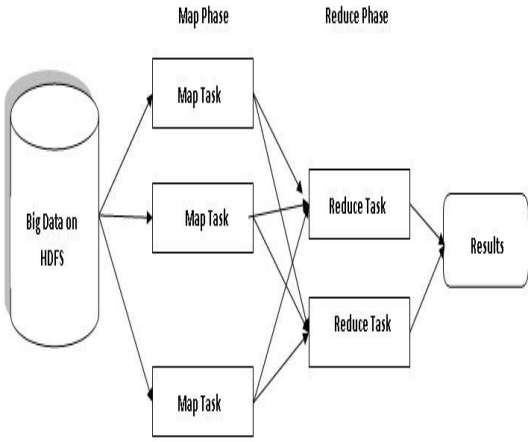

Fig. 1:-Working of map reduce framework

Different tools have been developed on top of Hadoop which provides abstract layers that minimize the effort of writing the map reduce programs. In next section we discuss such tools that are used for mining big data.

3. Big Data Mining Tools

For mining Big Data there are many open source tools. The most popular are as follows:

3.1 Apache Mahout

Apache Mahout is scalable machine learning and data mining open source software based mainly in Hadoop. It has implementations of a wide range of machine learning and data mining algorithms: clustering, classification, collaborative filtering and frequent pattern mining.

3.2 Apache Hive

Apache Hive data warehouse software facilitates querying and managing large datasets residing in distributed storage. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL. At the same time this language also allows traditional map/reduce programmers to plug in their custom mappers and reducers when it is inconvenient or inefficient to express this logic in HiveQL

3.3 Apache Spark

Apache Spark is an in-memory distributed framework which is used for fast mining and analyzing data.

3.4 Apache Giraph Pregel

Apache Giraph is used for mining the relationship data that is generally based on graph theory. It is an iterative graph processing system built for high scalability. For example, it is currently used at Facebook to analyze the social graph formed by users and their connections. Giraph originated as the open-source counterpart to Pregel, the graph processing architecture developed at Google.

4. Issues with Bigdata

4.1 Security

Security has always been an issue when data privacy is considered. Data integrity is one of the primary

components when preservation of data is considered. Access and sharing of Data which is not meant for public, has to be protected. For this type of security many researches have been done. Security has always been an issue when data are considered. Many of the Big Data tools lack security and many of them have been compromised. But these tools can be integrated with other tools like Kerberos to provide authentication and authorization.

4.2 Dependent on Statistical Analytics Tools

The second issue with big data is the Big Data alone is not sufficient for good analysis, it only works in adjunct with statistical algorithms.

4.3 Validating the Result Is Difficult

Tools require to analyze big data are more error prone due to the complex variety of data. Validating the results of such kind of processing is very difficult.

Finally big data is at its best when analyzing things that are extremely common, but often falls short when analyzing things that are less common. For instance, programs that use big data to deal with text, such as search engines and translation programs often rely heavily on something called trigrams: sequences of three words in a row (like “in a row”). Reliable statistical information can be compiled about common trigrams, precisely because they appear frequently. But no existing body of data will ever be large enough to include all the trigrams that people might use, because of the continuing inventiveness of language.

5. Conclusion

It’s no secret that big data has led to major changes within the business world. Companies are utilizing the benefits from data analytics to positively impact their bottom line, resulting in growing revenues and greater efficiency. The uses of big data are many and can apply to areas that many might not have thought of before. One area that sees a lot of potential in big data is the mining industry. For an industry that does trillions of dollars in business every year, big data is not seen as a luxury but as a necessity. Researchers are continuously working on the algorithms to mine big data efficiently and quickly.

References

[1] http://en.wikipedia.org/wiki/Big_data
[2] SAMOA, http://samoa-project.net, 2013.
[3] C. C. Aggarwal, editor. Managing and Mining Sensor Data. Advances in Database Systems. Springer, 2013.
[4] Apache Hadoop, http://hadoop.apache.org.
[5] Apache Mahout, http://mahout.apache.org.
[6] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. MOA: Massive Online Analysis http://moa. cms.waikato.ac.nz/. Journal of Machine Learning Research (JMLR), 2010.
[7] Jason Palmer (2013). "Google searches predict market moves". BBC Retrieved 2013
[8] Kalil, Tom. "Big Data is a Big Deal". White House Retrieved 2012
[9] Executive Office of the President (2012). "Big Data across the Federal Government" White House Retrieved 2012
[10] "How big data analysis helped President Obama defeat Romney in 2012 Elections". Bosmol Social Media News 2013 Retrieved 2013
[11] Hoover, J. Nicholas. "Government's 10 Most Powerful Supercomputers" Information Week UBM Retrieved 2012
[12] Bamford, James. "The NSA Is Building the Country‘s Biggest Spy Center (Watch What You Say)". Wired Magazine Retrieved 2013
[13] Groundbreaking Ceremony Held for $1.2 Billion Utah Data Center" National Security Agency Central Security Service. Retrieved 2013
[14] Layton, Julia. "Amazon Technology" Money.howstuffworks.com Retrieved 2013.

Authors Details

Prof. Anuradha Narendra Nawathe Assistant Professor in Amrutvahini college of Engineering Sangamner in Pune University. Having 16 years of teaching experience in the computer engineering field. Near about 20 international paper published in reputed gernals10 national level paper published .Life Member of ISTE,IE and CSI.I am very much positive related teaching and learning process. This year for research project 1,30,000 grant is received from university of Pune. Work in various committees positively. I have completed my B.E. From Government College of Engineering Amravati and M.E. from Prof .Ram Meghe Institute of Technology and Research, Badnera, Amravati University. I have 10 years Teaching Experience from government college of Engineering, Amravati. I have attended 30 workshops uptill now.