Enhanced Approach to Attain Competent Big Data Pre-Processing

Prity Vijay, Bright Keswani

[a]*Research Scholar, Department of Computer Science, Suresh Gyan Vihar University, Jaipur, India*
[b]*Associate Professor, Department of Computer Science, Suresh Gyan Vihar University, Jaipur, India*

Abstract: - We are living in an era of Big Data. Mountain of data is generated every minute from everywhere and in a different format. These data can do wonder if utilize properly In order to extract hidden information from that data; proper machine learning algorithm should be used. But before feeding the data into an algorithm it has to wrangle. Lots of time and effort is needed in pre-process the data. In this paper, we discussed the importance/need of big data and steps involved in pre-processing of big data. This study presents the problem/challenge and available technologies with respect to big data pre-processing and discussed what else can be done to achieve an enhanced approach in order to transform complex data into momentous one.

Keywords- Big Data, Machine Learning, Data Preprocessing, Data Cleaning, Wrangle

1. Introduction

Computing has become global. Devices like Cell Phone and Smart Phones are creating lots of data on daily basis. A few years ago data was, just in terms of megabytes and gigabytes. But today data is counted in terms of terabytes and petabytes. Daily nearly 500 billion gigabytes of data originate through internet [1]. Now, the data is not human generated indeed lots of machines generating data [2]. This data blast is altering our world. These data are dollars but only, if handled properly. But, the question is how to get benefit from this data? Many challenges occur while storing and processing of Big Data. Therefore Big Data requires a new set of tools, applications, and frameworks. Big Data does not have a fixed measurement, for some group some GB or 1TB can be big, for others 10TB may be big. Doug Laney [3], characterizes Big Data in terms of three V 's. Volume refers to the size of data, with the growth of social media, the amount of data is growing very speedily. Velocity refers to the speed at which the data is being generated. In simple words, we can say, "Big Data is a circumstance where the Volume, Velocity, and Variety of data go beyond an organization's storage or computation capacity for precise and well-timed decision making". Different applications have different latency requirements, and in today's competitive world, decision makers want the important data information within a fraction of second, if possible. A few examples include stock exchange data, tweets on Twitter, status updates/likes/shares on Facebook, etc. This speed aspect of data generation is referred to as *Velocity* in the Big Data world. Variety refers to the dissimilar formats in which the data is being generated. Today, 70% of generated data are in unstructured form thus creating a challenge. RDBMS leader of IT world since last 30 years doesn't fit for Big Data because of the fact that it cannot handle unstructured data efficiently. RDBMS are strict towards schema having lots of constraints which goes well to a homogeneous dataset. In Big Data, maintaining the relationship between unstructured data (images, videos, Mobile generated information, RFID etc) is very tough. Apart from above Big Data Analytics should possess very fast processing speed like real-time or near to real time, which RDBMS doesn't guarantee. Therefore, NoSql with distributed file system can be a better approach for analyzing Big Data [4]. Maintaining and integrating Big Data is a very costly solution when we go with traditional approaches. All the above-mentioned reason together created very severe need for new approaches for Big Data analytics [5]. Storage problem of Big Data is only part of the game [6]. To cope up with, it astonishing techniques are required. Many approaches are available in the market; Hadoop along with its sub-projects (Hive, Pig, HBase, Zookepear, Flume, Oozie, Sqoop etc) could be a great option for the Big Data analytics [7]. In the last six years, Hadoop becomes the standard of many organizations [8]. Apache Hadoop, an open source framework designed to support distributed parallel processing of a large number of data sets across clusters of computers using simple programming models. It is written in Java. It can run on commodity hardware, scaling up from single nodes to thousands of computer, thus forming a cluster. Each node in the cluster offers local computation and storage. Hadoop promises high availability to end user. Instead of depending on hardware to deliver high-availability, its framework itself can identify and figure out a single point of failures at the application layer, thus providing a high availability service to its user[8]. It has a number of commercially supported distributions from

companies such as MapR Technologies and Cloudera[9]. Doug Cutting, the mastermind of Hadoop[10], share the journey of Hadoop in an interview. Hadoop gets its name from the toy elephant of creator son[11]. Hadoop was started by two Yahoo employees, Doug Cutting and Mike Cafarella, in 2006. It was initially created to support Nutch[12], an open source web crawler. Hadoop was motivated by Google MapReduce and Google File System, launched by Google in 2003 to handle billions of data [13]. In order to share it, Google released white papers explaining GFS and MapReduce in 2004. After a year, Yahoo started to use Hadoop and then in 2008, it was taken over by Apache, thus, presently known as Apache Hadoop. It is an open source Apache framework, is mainly divided into two parts [14] – HDFS: Hadoop Distributed File System for storage and processing and MapReduce, programming language, in JAVA to look after all the programming task. Hadoop works on master and slave architecture consisting one Name node (Master) and various Data node (Slave). HDFS is cost-effective; highly fault tolerance, reliable, data processing system which is designed to run on cheap commodity hardware and to store terabytes (TB) and petabytes (PB) of distributed unstructured data very easily. Map Reduce, part of Hadoop and act as software for processing large datasets. It has basically two main functions Map and Reduce [15]. Map split the data into <key, value> pair and generate intermediate value, reduce conclude final output, out of intermediate value produced by map function. Workflow of MapReduce consists of mapping, sharing, shuffling and reducing. Jene and Dittrich, concluded that Hadoop, become standard for many organization various techniques come to boost up the performance of Hadoop.

The term "Data Warehouse" was first coined by Bill Inmon in 1990. According to Inmon [16], a data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data. This data helps analysts to make informed decisions in an organization. Data mining [17] deals with discovering important information from huge datasets. It possesses various methods with the combination of machine learning, statistics, and database systems. The main motive of data mining is to extract a significant pattern from big data ocean.

These patterns can further use in many areas to make decisions. According to Fayyad Usama and et-al," Data mining is the analysis step of the "knowledge discovery in databases" process or KDD". The greatness of data mining is because, it can be applied in any field of computer science including artificial intelligence, business intelligence, machine learning, etc. Machine Learning is a field of computer science which provides the computer an ability to learn without any outside interruption. It is an approach where applications are trained to build the model. The trained model, in turn, acts smartly. It was first discovered by Arthur Samuel in 1959. Machine learning is consist of huge sets of an algorithm which overcome the limitation of old static approach where in order to perform any task instruction should be fed into a system manually. In reverse machine learning algorithm is constructed in such a way that they can learn from the data itself and after learning it can predict on the data similar to it. Machine learning algorithm is mainly used in the department where an explicit algorithm could not work properly. Gmail is a smart engine where spamming of the mail is done automatically and smart phone application which can recognize gender, age and facial expression of a person is some of the use cases of machine learning algorithms.

2. Importance Of Data Pre-Processing For Data Scientist

Data is growing day-by-day. Extracting knowledge from it is a real threat otherwise it is nothing but garbage. Data, when loaded into the database from various source, emerge as a messy dataset. These datasets are of no use because extracting valuable information from it is very tough. Therefore data pre-processing is the first and mandatory step for any data scientist before mining. According to Lour [19], data scientists spend 50% - 80% of their valuable time and effort in data collection and preparing disorderly digital data, before it can be explored for useful nuggets.

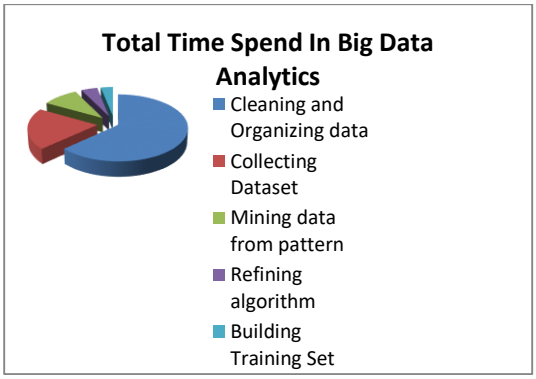

Figure: Total Time Spend In Big Data Analytics [22]

1. Steps Implicated In Data Preparation

Data is always messy, in order to convert messy data into a meaning dataset, data has to undergo different phases. These phases can be described in different steps:

1.1 Data Cleaning

The first and most important step in data pre-processing is data cleaning. In today's world, there are many possibilities where data is not consistent. Dirty data can give wrong output. Several reasons end up loading wrong data in a dataset. While cleaning the data different types of errors increases the complexity of the dataset. These errors are categorized into various group syntactical, semantic and coverage. Each group includes a set of error types. Those are Outliers, (an extreme or unexpected value which are different from the majority of values in the dataset), Duplicate or redundant (which occurs more than one in a dataset), incomplete (missing data) or noisy data (false data). Each of these errors should be treated in this phase.

1.2 Data Integration

Merging of data from various sources is termed data integration. One of the challenges here is data redundancy. Redundancy occurs if an attribute derived from another attribute or a set of the attribute.

1.3 Data Transformation

Here data is transformed into a format suitable for data mining. Data scientist follows some use normalization, aggregation and generalization for data transformation.

1.4 Data Reduction

It is not realistic to perform a complex data analytics on a massive dataset. Data reduction is a technique to achieve a smaller dataset which will generate the same result as the huge dataset.

In this step data, scientists attain a reduced representation of the data set, which is smaller in size but yields almost the same analysis outcomes.

1.5 Data Discretization

Dataset usually contains 3 types of attributes-continuous, nominal and ordinal. In this phase, data is converted into the type which is under stable by the algorithm applied on it.

2. Challenges And Problems Of Data Pre-Processing

It was concluded in one of the research report of Gartner that time and efforts for data preparation cost an average organization $13.5 million every year which is too high to bear. They have to remove incorrect values and the values which are duplicated from the dataset. In many cases, it happens that they don't have proper information about such abnormal data and they don't have a proper understanding of how to deal with this. In that case, deleting such entries may be the only option. This deletion can be costly if a large amount of data is deleted due to loss of information [20]. As we have seen earlier also data cleaning is a very much time-consuming process. Once the cleaning is completed and dataset become free from error repetition of this process again is a big headache. But data is changing every minute; it happens that new data is added into the existing dataset after the cleaning part was done. Therefore in order to avoid such a situation, the proper management tool is still in need. In lots of companies, data cleaning has to repeat every time whenever new data comes which increases response time and lower efficiency. In a lot of cases as the data is so big that we don't have a complete picture of error-prone data at the beginning of data cleaning. This again led to a big problem. In order to overcome this problem, we require tools which contain a collection of methods for detecting an error and correcting it automatically.

3. Tools And Technique Available For Data Pre- Processing

There are many tools and techniques which can deal with huge dataset preparation [21], some of which are Trifacta, OpenRefine, Python, Pandas. R contains many packages for data preparation. Optimus is an open source framework which can be used under Spark in a distributed environment for data cleaning. it contains several data wrangling tools. Although many tools have been inventing to solve the purpose of data preprocessing yet a lot more is needed because of its importance in machine learning and big data.

4. Efficient Approach To Attain Data Preprocessing

Errors in the data increase its complexity and present tools and techniques are either time consuming or costly or manual. As we have seen that almost 80% time is spent on data preparation and the process can be repeated also. This is the area where enhancement is strictly needed. One option to attain enhancement and to lower down the process time is automation. In our words, enhanced data preparation tools are in need which can automatically at the same time efficiently clean the data. Thus decreases complexity in a data and that too in a less time. However, it is known that data cleaning could not be fully automated in most cases because no two data will alike. Therefore smart and semi-automatic approaches are practical and good enough as a collection and methods are needed which can detect each kind of errors individually and then proper data cleaning can be applied on it automatically and all of these collections are combined as a new tool for data more effective data cleaning.

CONCLUSION

According to IDC by the end of 2020 time consumed by data preparation tools will grow 2.5 times faster than a regular IT-based tool. Many approaches are present for data preparation but still, it is a hot area of research. Work needs to be done in the pre-analysis stage so that the results which will be achieved are more perfect and less time-consuming. Right degree of automation should have to infuse into the techniques dealing with data preparation stage.

REFERENCES

Shilpa, M. Kaur, "Big Data and Methodology-A review", International Journal of Advanced Research in Computer Science and Software Engineering, Vol: 3, Issue: 01, 2013

SAS White Paper, "Big Data Meets Big Data Analytics". Accessed: http://docplayer.net/765971-Big-data-meets-big-data-analytics.html,2012

D Laney," 3d data management: Controlling data volume, velocity and variety", META Group Inc,2001

V. Shukla, P. K.Dubey," Big Data: Moving Forward with Emerging Technology and Challenges ", International Journal of Advanced Research in Computer Science and Management Studies, Vol.2, 2014

Lavastorm Analytics, Why Most Big Data Projects Fail", Accessed:http://www.lavastorm.com/assets/Why-Most-Big-Data-Projects-Fail-White-Paper.pdf, 2014

B. Purcell," The emergence of "big data" technology and analytics", Journal of Technology Research,pp.1-6

V. S. Patil, P. D. Soni, "HADOOP SKELETON & FAULT TOLERANCE IN HADOOP CLUSTERS ", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Vol: 2, Issue:2,2013

Hadoop Wiki, "Apache Hadoop", Accessed: http://wiki.apache.org/hadoop

Juniper Networks, "Introduction to Big Data: Infrastructure and Networking, Considerations", Accessed: http://www.juniper.net/us/en/local/pdf/whitepapers/2000488-en.pdf,2012

D. Harris"The history of Hadoop: From 4 nodes to the future of data".Accessed:https://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/,2013

A. Vance, Ashlee, "Hadoop, a Free Software Program, Finds Uses Beyond Search", The New York Times. Accessed at: Hadoop, a Free Software Program, Finds Uses Beyond Search",2016

R. Khare, D. Cutting, K. Sitaker, A. Rifkin," Nutch: A Flexible and Scalable Open-Source Web Search Engine", Accessed: http://www.master.netseven.it/files/255-Nutch.pdf,2005

Intellipaat, "Hadoop Creator goes to Cloudera". Accessed: https://intellipaat.com/blog/hadoop-creator-goes-to-cloudera/,2016

S. D. Vidyasagar, " A Study on "Role of Hadoop in Information Technology era" ", Global Research Analysis, Vol 2, Issue:2,2013

D. Maclean, "A Very Brief Introduction MapReduce", Accessed: http://hci.stanford.edu/courses/cs448g/a2/files/map_reduce_tutorial.pdf,2011

W.H. Inmon, "Building the Data Warehouse", John Wiley. pp. 33, 1996

M.S. Chen and etal, " Data Mining: An Overview from a Database Perspective", IEEE Transactions on Knowledge and Data Engineering, 8, Issue:6,1999

A.L. *Samuel*, "Some Studies in Machine Learning Using the Game of Checkers". *IBM Journal of Research and Development, 1959*

Dezyr, Why data preparation is an important part of data science? (2016)Accessed: https://www.dezyre.com/article/why-data-preparation-is-an-important-part-of-data-science/242,2016

Wikipedia,Accessed: : https://en.wikipedia.org/wiki/Data_cleansing

Prity Vijay, K.Bright Keshwani," A Study on Big Data Analytics through R", International Journal of Innovative Research in Computer and Communication Engineering,2016

Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says, Accessed: https://www.forbes.com/sites/unicefusa/2018/06/19/unicef-helps-rohingya-babies-born-into-a-legacy-of-sexual-violence/#2f95c2545318