IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 38-41
www.iosrjournals.org

Big Data with Knowledge Extraction

Chirag Thakkar[1], Dr.Kishore Dhole[2]
[1](Department, College/ University Name, Country Name)
[2](Department, College/ University Name, Country Name)

Abstract: *Nowadays Big Data has become one of the biggest concepts in the world of IT especially with the rapid development driving the increase of data. With today's information overload, it has become increasingly difficult to analyze the huge amount of data and to generate appropriate extraction decisions. Big data has emerged as an important area of study for both practitioners and researchers, reflecting the magnitude and impact of data related problems to be solved in the contemporary organizations. To a growing number of companies, Knowledge Extraction is more than just a concept or a sales pitch. Big Data can create efficient challenging solutions in health, Security, Government and more usher in a new era of analytics and decisions. Knowledge extraction comprises a set of strategies and practices used to identify, create, represent, distribute and enable creating experience creating that can constitute a real methodology. The scope of the conference on Big Data and is to bring together researchers, designers, developers and practitioners interested in the advances and applications in the field of communication technologies, sustainability, and technologies to realize smart Technologies of the future. This paper presents a state of issues in big data where we try to explore Big Data within the concept of Knowledge Extraction.*

Keywords: *Big Data, Knowledge Extraction.*

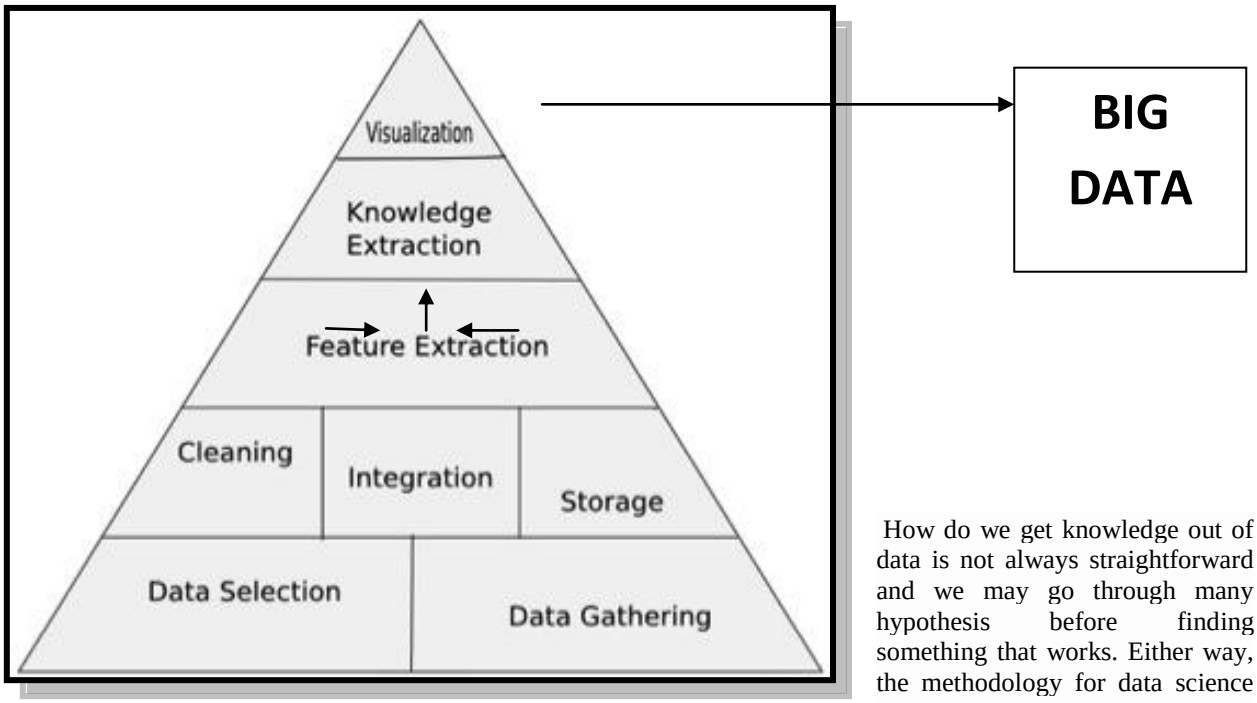
I. INTRODUCTION:

Data is easier to capture and access through third parties such as Facebook , D & B, and others. Its not surprising that developers find increasing value in leveraging this data to enrich existing applications and create new one made possible by it. The use of the data is rapidly changing the nature of communication shopping, advertising, entertainment, and relationship Management..With the increasing data globally, the term big Data is mainly used to describe large datasets or complex that traditional data processing applications are inadequate. Compared with other traditional databases , Big Data includes a large amount of unstructured data that must be analyzed in real time. Big Data also brings new opportunities for the discovery of new values that are temporarily hidden. Big Data is a broad and abstract concept that is receiving great recognition and is being highlighted both in the field of Communication Technology and business. Accuracy in big data may lead to more confident decision making. And better decisions can mean greater operational efficiency, cost reduction and reduced risk.

Data sets grow in size in parts because they increasingly gathered by information sensing mobile, remote sensing, software logs, cameras ,microphones, radio frequency identification readers and wireless sensor networks. The worlds technological per-capita capacity to store information has roughly doubled every 40 months. The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization.

Analysis of data sets finds new correlations, to "spot business trends, prevent diseases, and so on.Scientists,business executives ,practioners of media and advertising and Governments alikeregularly met difficulties with large data sets in areas including Internet search,Finance and business informatics.Scientists encounter limitations in e-Science work, including meterology,genomics, complex physics simulations, and biological and environmental research.."For some organizations ,Facing hundred s of gigabytes of data for the first time may trigger a need to reconsider data management alternatives. Big Data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate ,manage, and process the data within a tolerable elapsed time. Big data sizea are a constantly moving target , from terabytes to many petabytes of data in a single data set. With this difficulty new platform of : Big data : tools has arisen to handle sense making over large quantities of data , as in the Apache Hadoop Big Data Platform.

Knowledge extraction is the creation of Knowledge from structured and unstructured sources. The resulting knowledge needs to be in a machine-readable and machine-interpretable format and must represent knowledge in a manner that facilitates inferencing. Although it is methodically similar to information extraction and data warehouse, the main criteria is that the extraction result goes beyond the creation of structured information or the transformation into a relational schema. It requires either the reuse of existing formal knowledge or the generation of a schema based on the source data.



How do we get knowledge out of data is not always straightforward and we may go through many hypothesis before finding something that works. Either way, the methodology for data science should be pretty general. Here I just want to outline what I have come to call the pyramid of data science, which at its core is based on the scientific method and simply outlines the steps I have come up with to execute a data science project. You may also find it resembles the DIKW (Data, Information, Knowledge, Wisdom) pyramid in some ways. The base upon which the pyramid sits is the idea, data selection and gathering are on the first tier, followed by data cleaning/integration and storage, then feature extraction, knowledge extraction, and finally visualization.

1. Data Selection and Gathering

At this tier we must think about what data we are going to gather for our question at hand. This is a most crucial step because whatever data parameters we choose to collect we are stuck with till the end (unless we come back to this step). It goes without saying that the final results will be entirely dependent on only these data parameters and nothing else. Here we might be gathering data from many sources, all of which will probably be in different data formats and contain data organized in many different ways.

2. Data Cleaning/Integration, and Storage

These first two tiers are generally going to take a substantial time of the project and could possibly be done simultaneously. Based on my own experience, I would argue that this step may take as much of 80% or more of the time involved in the project. Here we must design a database into which our data is to be stored. We must select a database management system that best fits our data. The data gathered must becleaned for such things as corrupt, missing, or inaccurate entries. Then, data from different sources needs to be integrated with each other into a cohesive dataset. This is where the storage comes in because a schema must be chosen so that all of the gathered data from the different sources can readily be reformatted into the database.

3. Feature Extraction

This tier is, in other words, dimensionality reduction of the data; feature extraction usually refers to image processing but I like to generalize it to all kinds of multidimensional data. We want to make the data easier for us to handle. The main goal here is to consolidate/combine our variables into more easily digestible chunks if at all possible. In this step we could use such techniques as principle component analysis or other transforms such as Fourier or wavelet transforms. In some cases it may not be possible to reduce the dimensionality of the data and we will just have to use all of the variables as they are. Another example could be the use of the rates at which the variables change in time series.

4. Knowledge Extraction

Finally, this is the tier that you have been waiting for! It is where you are at last going to be able to answer your question. With the data nicely formatted and the features selected it is time to apply whichever analysis is most appropriate for your data, be it a machine learning algorithm, statistical analysis, time series analysis, regression,

or what have you. Lets say you want to predict an outcome based on the features you have selected. You might what to use some machine learning algorithm, such a perceptron, naïve Bayes, or SVM; it is all up to you. Or, you might want to cluster your data to find hidden patterns. Many of these techniques are quite standard and implementations are generally straight forward, or you might use a machine learning tool such as WEKA, which is a tool that has many machine learning algorithms implemented in Java easily ready to be adapted to your project. There are many open source implementations out there, so just go out and look for them. For this reason, in my opinion, this step generally might take less time than the execution of the first 3 tiers. This leads to the final step of visualization, which is basically to explain your results with others.

5. Visualization

This, I leave up to the reader. This step takes some creativity, though you don't have to be a great graphic designer or anything, you just need a good visual way to get your point across. This also depends on who your audience is going to be. If it is to your colleagues that already are familiar with the data set and techniques you've used, then perhaps you can omit things that would be obvious to them. If it is to the CEO of your company and you have just discovered the greatest way to optimize profits then you will definitely have to get creative to tell a good story.

Issues in Big Data

Meeting the need for speed

In today's hypercompetitive business environment, companies not only have to find and analyze the relevant data they need, they must find it quickly. Visualization helps organizations perform analyses and make decisions much more rapidly, but the challenge is going through the sheer volumes of data and accessing the level of detail needed, all at a high speed. The challenge only grows as the degree of granularity increases. One possible solution is hardware. Some vendors are using increased memory and powerful parallel processing to crunch large volumes of data extremely quickly

Understanding the data

It takes a lot of understanding to get data in the right shape so that you can use visualization as part of data analysis. For example, if the data comes from social media content, you need to know who the user is in a general sense – such as a customer using a particular set of products – and understand what it is you're trying to visualize out of the data. Without some sort of context, visualization tools are likely to be of less value to the user. you're trying to visualize out of the data. Without some sort of context, visualization tools are likely to be of less value to the user.

Addressing data quality

Even if you can find and analyze data quickly and put it in the proper context for the audience that will be consuming the information, the value of data for decision-making purposes will be jeopardized if the data is not accurate or timely. This is a challenge with any data analysis, but when considering the volumes of information involved in big data projects, it becomes even more pronounced. Again, data visualization will only prove to be a valuable tool if the data quality is assured. To address this issue, companies need to have a data governance or information management process in place to ensure the data is clean. It's always best to have a proactive method to address data quality issues so problems won't arise later.

Displaying meaningful results

Plotting points on a graph for analysis becomes difficult when dealing with extremely large amounts of information or a variety of categories of information. For example, imagine you have 10 billion rows of retail SKU data that you're trying to compare. The user trying to view 10 billion plots on the screen will have a hard time seeing so many data points. One way to resolve this is to cluster data into a higher-level view where smaller groups of data become visible. By grouping the data together, or "binning," you can more effectively visualize the data.

Dealing with outliers

The graphical representations of data made possible by visualization can communicate trends and outliers much faster than tables containing numbers and text. Users can easily spot issues that need attention simply by glancing at a chart. Outliers typically represent about 1 to 5 percent of data, but when you're working with massive amounts of data, viewing 1 to 5 percent of the data is rather difficult.

II. CONCLUSION

Today Many technologies are emerging in the field of Big Data. It supports the running of applications on the hardware. Big data is directed to continue rising during the next year and every data scientist will have to handle a large amount of data every year. This data will be more miscellaneous, bigger and faster. In this paper the

issues with big data and how the knowledge extraction is done with the help of big data is explained. Big data is becoming the latest final border for precise data research and for business applications.

REFERCENCES

[1]. INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND MOBILE COMPUTING. Big Data Review Munesh Kataria, Ms. Pooja Mittal, *IJCSMC, Vol. 3,Issue 7 , July 2014 , pg. 106 – 110*
[2]. Papers refered from ieeexplore.ieee.org. , www.cmswire.com/social-buisness/knowledge.
[3]. www. Engpaper.net/big-data-research-papers-2013-2014.html.
[4]. The Roles of Big Data in the Decision-Support Process: An Empirical Investigation.
[5]. Thiago Poleto,Victor diogho Heuer de Carvalho, and Ana Paula Cabral Seixas Costa.
[6]. https://www.sas.com/resources/asset/five-big-data-challenges-article.pdf