ISSN No. 0976-5697

Volume 7, No. 6(Special Issue), November 2016

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Big Data Barriers and Opportunities

Kirpawant Kaur
Dept. of Computer Applications
Khalsa College For Women, Sidhwan Khurd,
Ludhiana
kirpawant@yahoo.com

Abstract*-The most important aspect that comes to mind with mention of big data is the enormity of data .This paper attempts to reveal the need of big data. It also offers a broader definition of big data that captures its other unique characteristics. Types of various technologies used in big data that makes it more reliable and useful. It also provides the applications of big data in various fields with respect to structured as well as unstructured data Manufacturing, Healthcare, Education, Media and many more. The future scope of Big Data is also described. The paper's primary focus is on Challenges and Opportunities with big data .There are various challenges involved while using big data. To take full advantage of big data we need to address these various challenges those are involved in big data. As there is a huge gap between the use of Big Data and its prospective. The main objective of this article is to discuss about the various flaws in big data. Some solutions have been provided for the same in this paper like Hadoop and some others could be introduced through further research. Enormous opportunities have also been provided in various fields by Big Data that are also considered in this article.*

1. INTRODUCTION

Big data refers to large data sets for which traditional data processing applications are inadequate. It is analyzed using advanced data analytics methods that extract value from data. Examples include web logs, call records, medical records, military surveillance, photography archives, video archives and large-scale e-commerce.

Big data is characterized by **3Vs**: the extreme volume of data, the wide variety of types of data and the velocity at which the data must be must processed.3Vs of big data-volume, velocity and variety of data exceed an organization's storage or compute capacity for accurate and timely decision making.

Volume: The quantity of generated and stored data. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not.

Velocity: For many applications, the speed of data creation is even more important than the volume. Real-time or nearly real-time information makes it possible for a company to be much more agile than its competitors.

Variety: Big data takes the form of messages, social networks data; readings from sensors, consumer browsing preferences for e-commerce,GPS signals from cell phones and more.

Application areas of Big Data:

Big data affects organizations across practically every industry.

Banking:

With large amounts of information streaming in from countless sources, banks are faced with finding new and innovative ways to manage big data. While it's important to understand customers and boost their satisfaction, it's equally important to minimize risk and fraud while maintaining regulatory compliance. Big data brings big insights, but it also requires financial institutions to stay one step ahead of the game with advanced analytics.

Education:

Educators armed with data-driven insight can make a significant impact on school systems, students and curriculums. By analyzing big data, they can identify at-risk students, make sure students are making adequate progress, and can implement a better system for evaluation and support of teachers and principals.

Government:

When government agencies are able to harness and apply analytics to their big data, they gain significant ground when it comes to managing utilities, running agencies, dealing with traffic congestion or preventing crime. But while there are many advantages to big data, governments must also address issues of transparency and privacy.

Health Care:

Patient records. Treatment plans. Prescription information. When it comes to health care, everything needs to be done quickly, accurately – and, in some cases, with enough transparency to satisfy stringent industry regulations. When big data is managed

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India

effectively, health care providers can uncover hidden insights that improve patient care.

Manufacturing:

Armed with insight that big data can provide, manufacturers can boost quality and output while minimizing waste – processes that are key in today's highly competitive market. More and more manufacturers are working in an analytics-based culture, which means they can solve problems faster and make more agile business decisions.

Big Data Tools

There are thousands of Big Data tools out there. All of them promising to save you time, money and help you uncover never-before-seen business insights. And while all that may be true, navigating this world of possible tools can be tricky when there are so many options.

Data Storage and Management

A good data storage provider should offer you an infrastructure on which to run all your other analytics tools as well as a place to store and query your data.

Hadoop

The name Hadoop has become synonymous with big data. It's an open-source software framework for distributed storage of very large datasets on computer clusters. All that means you can scale your data up and down without having to worry about hardware failures. Hadoop provides massive amounts of storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs. To truly harness its power, you really need to know Java.

MongoDB

MongoDB is the modern, start-up approach to databases. It's good for managing data that changes frequently or data that is unstructured or semi-structured.Common use cases include storing data for mobile apps, product catalogs, real-time personalization, content management and applications delivering a single view across multiple systems.

Data Cleaning

Before you can really mine your data for insights you need to clean it up. Even though it's always good practice to create a clean, well-structured data set, sometimes it's not always possible. Data sets can come in all shapes and sizes (some good, some not so good!), especially when you're getting it from the web. The companies below will help you refine and reshape the data into a useable data set.

OpenRefine

OpenRefine (formerly GoogleRefine) is an open source tool that is dedicated to cleaning messy data. You can explore huge data sets easily and quickly even if the data is a little unstructured. OpenRefine is that it has a huge community with lots of contributors meaning that the software is constantly getting better and better.

Big Data Techniques:

There are many techniques being used to analyze datasets. Researchers continue to develop new techniques and improve on existing ones, particularly in response to the need to analyze new combinations of data..

1. **A/B testing**: A technique in which a control group is compared with a variety of test groups in order to determine what treatments (i.e., changes) will improve a given objective variable, e.g., marketing response rate. This technique is also known as split testing or bucket testing. An example application is determining what copy text, layouts, images, or colors will improve conversion rates on an e-commerce Web site. Big data enables huge numbers of tests to be executed and analyzed, ensuring that groups are of sufficient size to detect meaningful (i.e., statistically significant) differences between the control and treatment groups. When more than one variable is simultaneously manipulated in the treatment, the multivariate generalization of this technique, which applies statistical modeling, is often called "A/B/N" testing

2. **Association rule learning**: A set of techniques for discovering interesting relationships, i.e., "association rules," among variables in large databases. These techniques consist of a variety of algorithms to generate and test possible rules. One application is market basket analysis, in which a retailer can determine which products are frequently bought together and use this information for marketing.

Challenges with Big Data:

There are many challenges involved with Big Data. To take the full advantage of Big Data we have to tackle all these challenges. Some of them are described below.

Data Acquisition:

Most of the most practical uses cases for big data involve data availability. It is recorded from some data generating source. Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude. One challenge is to define these filters in such a way that they do not discard useful

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India

information. 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. One solution to this challenge is to have the proper domain expertise in place. Make sure the people analyzing the data have a deep understanding of where the data comes from, what audience will be consuming the data and how that audience will interpret the information.

Information Extraction and Cleaning:

The second one big challenge is to generate the right metadata to describe what data is recorded and how it is recorded..For example we need to know the source for each report if we wish to examine duplicates. we require an information extraction process that pulls out the required information from the underlying sources and expresses it in a structured form suitable for analysis. Doing this correctly and completely is a continuing technical challenge.

Data Integration and Representation:

The next challenge is to pull out the required information from various underlying recourses and integrate it in meaningful information and to represent it in a structured form suitable for analysis.

Query Processing, Data Modeling, and Analysis:

Data analysis is considerably more challenging than simply locating, understanding and identifying data. Effective large scale analysis requires differences in data structure and semantics to be expressed in form that are computer understandable. Methods for querying and mining Big Data are fundamentally different from traditional statistical analysis on small samples. Big Data is often noisy, dynamic, heterogeneous, inter-related and untrustworthy. Nevertheless, even noisy Big Data could be more valuable than tiny samples because general statistics obtained from frequent patterns and correlation analysis usually overpower individual fluctuations and often disclose more reliable hidden patterns and knowledge.

CONCLUSION

We have entered an era of Big Data. Big Data is a sea change that, like nanotechnology and quantum computing, will shape the next century. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. All of our challenges reflect different facets of more fundamental issues. My hope is that this paper contributes to generating exchanges, debates and interest

among a wide range of readers to advance Big Data for Development and will help to overcome the challenges.

REFERENCES

[1] Big data: The next frontier for innovation, competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. May 2011.

[2] Advancing Discovery in Science and Engineering. Computing Community Consortium. Spring 2011. [CCC2011b] Advancing Personalized Education.

[3] "The World's Technological Capacity to Store, Communicate, and Compute Information". *MartinHilbert.net*. Retrieved 13 April 2016.

[4] "Data Integration, Information Extraction" <https://www.sas.com/resources/asset/five-big-data-challenges-article.pdf >

[5] "Big Data challenges" <https://www.progress.com/docs/default-source/default-document-library/Progress/Documents/Papers/Addressing-Five-Emerging-Challenges-of-Big-Data.pdf