A Review of Issues and Challenges with Big Data

Article · October 2014

CITATIONS
7

READS
1,802

3 authors, including:

Bharti Kalra
Institute of technology gopeshwar chamoli india
12 PUBLICATIONS **29** CITATIONS
SEE PROFILE

D K Chauhan
Noida International University
6 PUBLICATIONS **32** CITATIONS
SEE PROFILE

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 4, pp: (97-101), Month: October - December 2014, Available at: www.researchpublish.com

A Review of Issues and Challenges with Big Data

[1]Bharti kalra, [2]Suryakant Yadav, [3] Dr. D.K. Chauhan

[1]P.hD. scholar (CSE), Noida international university, Greater Noida-India
[2]Assistant Professor, Department of CSE, SOET, Noida international university, Greater Noida-India
[3]Professor, Director Technical SOET, Noida international university, Greater Noida-India

***Abstract*: In today era of world data is very important for every field, many organizations and researchers. Companies having overwhelming volume of data for transactional processing, storing analyzing and to manage. The management, analysis, prediction of big data are becoming more accurate with the big data tools. This paper started with the introduction and summarizes the different issues and challenges with big data when different companies try to tackle with big data**

***Keywords:* Big Data, Issues, Challenges, Privacy.**

I. INTRODUCTION

We are in the world of technology and huge volume of data flowing and collected from many sources such as from our PC's, Laptop, mobile phone, social media site, web site and many other devices and sources .this data is characterized as structured, semi-structured and unstructured. The Big data is actually comes in different unit according to the need of the organization.

The McKinsey Global Institute has estimated that enterprises globally stored more than 7 exabytes (7 x 260 bytes) of new data on disk drives in 2010, while consumers stored more than 6 exabytes of new data on devices such as PCs and notebooks [5]. According to IDC, "in 2011, the amount of information created and replicated will surpass 1.8 zetabytes (1.8 trillion gigabytes), growing by a factor of nine in just five years". That is nearly as many bits of information in the digital universe as stars in the physical universe. The explosion of data is not new. It continues a trend that started in the 1970s. What has changed is the velocity of growth, the diversity of the data and the imperative to make better use of information to transform businesses [6].EBay.com uses two data warehouses at 7.5 petabytes and 40PB as well as a 40PB Hadoop cluster for search, consumer recommendations, and merchandising. Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. The core technology that keeps Amazon running is Linux-based and as of 2005 they had the world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB. Walmart handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes (2560 terabytes) of data – the equivalent of 167 times the information contained in all the books in the US Library of Congress. Facebook handles 50 billion photos from its user base.. FICO Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide. The volume of business data worldwide, across all companies, doubles every 1.2 years, according to estimates. [9]

II. ABOUT BIG DATA

Big data as the name define collection of large set of data with characteristics volume, variety and velocity" Traditionally, the term 'big data' has been used to describe the massive volumes of data analyzed by huge organizations like Google or research science projects at NASA," says Merv Adrian, research vice president at Gartner.[1]

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 4, pp: (97-101), Month: October - December 2014, Available at: www.researchpublish.com

Big data is a term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Although big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data.[2]

As far back as 2001, industry analyst Doug Laney (currently with Gartner) articulated the now mainstream definition of big data as the three Vs of big data: volume, velocity and variety.

Fig 1: Parameters [15]

- **Volume: Big data employs enormous volume of data.** Data volume is increase by many of the factors. Data is generated by machines, networks and human interaction on systems like social media the volume of data to be analyzed is massive. Unstructured data streaming in from social media. In the past, excessive data volume was a storage issue. But with decreasing storage costs, other issues emerge, including how to determine relevance within large data volumes and how to use analytics to create value from relevant data.
- **Velocity.** One meaning of Velocity is to describe data-in-motion; Data is streaming in at unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations.
- **Variety.** Data today comes in all types of formats. Structured, numeric data in traditional databases. Information created from line-of-business applications. Unstructured text documents, email, video, audio, stock ticker data and financial transactions. Managing, merging and governing different varieties of data is something many organizations still grapple with.

At SAS, we consider two additional dimensions when thinking about big data:

- **Variability.** In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Is something trending in social media? Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data involved.
- **Complexity.** Today's data comes from multiple sources. And it is still an undertaking to link, match, cleanse and transform data across systems. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control.

III. CHALLENGES WITH BIG DATA

In a broad range of application areas, data is being collected at an unique scale. Decisions that previously were based on guesswork, or on the basis of reality, at present now decision to be made using data-driven mathematical models. Such Big Data analysis now drives nearly every aspect of society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences.

- **Timeliness and heterogeneity**

When there is a lot of information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured prior to data analysis. Even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain. This incompleteness and these errors must be managed during data analysis. Doing this correctly is a challenge.[8]

- **Scalability**

Scalability is the major challenge with the big data. You want to be able to scale very rapidly and elastically, whenever and wherever you want. There is a need of effective solution to enable the cost-effective, feasible, scalable storage and processing of large volume of data. Most NoSQL solutions like MongoDB or HBase have their own scaling limitations.[10]

- **Performance**

In the world of internet big data must move at extremely high velocities no matter how much you scale or what workloads your database must perform. You need that big data analysis is performed within time constraints as require. The data handling hoops of RDBMS and most NoSQL solutions put a serious drag on performance.[10]

- **Continuous Availability**

When you rely on big data to feed your essential, revenue-generating 24/7 business applications, even high availability is not high enough. Your data can never go down. The capabilities of existing system to process streaming information and answer queries in real-time and for thousands of concurrent users are limited. people expect real-time or near real-time responses from the systems they interact with.[10]

- **Workload Diversity**

Big data comes in all shapes, colors and sizes. Rigid schemas have no place here; instead you need a more flexible design. You want your technology to fit your data, not the other way around. And you want to be able to do more with all of that data – perform transactions in real-time, run analytics just as fast and find anything you want in an instant from oceans of data, no matter what from that data may take.[10]

- **Data Security**

Security is the big concern with the big data. As larger amount of data is processed and transfer among the organizational boundaries and this Big data carries some big risks when it contains credit card data, personal ID information and other sensitive assets. Now the challenge is how to protect this sensitive data and how to keep private. Most NoSQL big data platforms have few if any security mechanisms in place to safeguard your big data. Security concerns about data protection are a major obstacle preventing companies from taking full advantage of their data.[10]

- **Identifying Right Data**

Identifying the right data from the vast amount of data is the big challenge. Since there are large number of sources such as social networking sites, blogs, different types of content such as articles, comments. companies have difficulty identifying the right data and determining how to best use it. Therefore, there is the need to find out the rules that will help in identifying the right data Building data-related business cases often means thinking outside of the box and looking for revenue models that are very different from the traditional business.

- **Identifying right talent**

Companies are struggling to find the right talent capable of both working with new technologies and of interpreting the data to find meaningful business insights.

- **Identifying right Platform**

  Data access and connectivity can be an obstacle. A majority of data points are not yet connected today, and companies often do not have the right platforms to aggregate and manage the data across the enterprise. In order to address the growing volume of data created as a part of power grid operation

- **Identify right architecture**

  The technology landscape in the data world is evolving extremely fast. Leveraging data means working with a strong and innovative technology partner that can help create the right IT architecture that can adapt to changes in the landscape in an efficient manner.

- **Collaborating across functions and businesses.**

  Leveraging big data often means working across functions like IT, engineering, finance and procurement and the ownership of data is fragmented across the organization. To address these organizational challenges means finding new ways of collaborating across functions and businesses.

According to SAS following challenges are outlined in terms of data visualization.

- **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. Visual-ization 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. Another method is putting data in-memory but using a grid computing approach, where many machines are used to solve a problem. Both approaches allow organizations to explore huge data volumes and gain business insights in near-real time.

- **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.

  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.

- **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 pro-active 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. How do you represent those points without getting into plotting issues? Possible solutions are to remove the outliers from the data (and therefore from the chart) or to create a separate chart for the outliers. You can also bin the results to both view the distribution of data and see the outliers. While outliers may not be representative of the data, they may also reveal previously unseen and potentially valuable insights.

IV. CONCLUSION

Big data has its own importance in the world of technology. Day by day many data is used and generated by many companies and researchers. So in this paper we start with the big data and review the possible challenges of big data, users must be aware of these challenges and will follow with appropriate solution.

REFERENCES

[1] For Big Data Analytics There's No Such Thing as Too Big The Compelling Economics and Technology of Big Data Computing March 2012 By: 4syth.com Emerging big data thought leaders

[2] http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data

[3] Big Data, Bigger Opportunities – Data it. gov's roles: Promote, lead, contribute, and collaborate in the era of big data ,Jean Yan ,April 9, 2013, www.meritalk.com/pdfs/bdx/bdx-whitepaper-090413.pdf

[4] http://www-01.ibm.com/software/data/infosphere/hadoop/what-is-big-data-analytics.html

[5] Understanding Big Data: A Management Study, September 22, 2011

[6] Mark Troester(2013), "Big Data Meets Big Data Analytics", www.sas.com/resources/.../ WR46345.pdf

[7] http://www.sas.com/en_us/insights/big-data/what-is-big-data.html

[8] Divyakant Agrawal, UC Santa Barba, Philip Bernstein, Microsoft Elisa Bertino,…(2012), " Challenges and Opportunities with Big Data",

[9] http://en.wikipedia.org/wiki/Big_data

[10] http://blogs.wsj.com/experts/2014/03/26/six-challenges-of-big-data/

[11] http://cacm.acm.org/magazines/2014/7/176204-big-data-and-its-technical-challenges/abstract

[12] http://www.sas.com/resources/asset/five-big-data-challenges-article.pdf

[13] http://www.datastax.com/big-data-challenges nessi white paper, December 2012,big data, new world of opportunities.

[14] Shilpa et al., Big data and Methodology: A review, International Journal of Advanced Research in Computer Science and Software Engineering 3(10), October - 2013, pp. 991-995