International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Survey on Big Data-Next frontier for Innovation

Anupriya

Dept of CSE/IT, Bhagat Phool Singh Women University, Khanpur Kalan, Sonipat, Haryana, India

Abstract : ***The growing impact of Big Data deduces the importance of analyzing huge amount of data with a frequent and rapid rate of growth and change in databases and data warehouses This gigantic growth of data commonly known as Big Data has received a lot of attention from researches in the last decade and has accordingly gone through a thorough research. This literature study gives an overview of variety of data and analyzed the problem of integrating unstructured data with the traditional structured data. Moreover, currently available tools focus on either structured data or unstructured data.***

Keywords: Big data, structured data, unstructured data, Hadoop, MapReduce.

1. Introduction

Big Data has become one of the buzzwords in IT during the last couple of years. Initially it was shaped by organizations which had to handle fast growth rates of data like web data, data resulting from scientific or business simulations or other data sources. Some of those companies' business models are fundamentally based on indexing and using this large amount of data. The pressure to handle the growing data amount on the web e.g. leads Google to develop the Google File System and MapReduce. Efforts were made to rebuild those technologies as open source software. This resulted in Apache Hadoop and the Hadoop File System [1], and laid the foundation for technologies summarized today as 'big data'.

2. Literature Overview

With this groundwork traditional information management companies stepped in and invested to extend their software portfolios and build new solutions especially aimed at Big Data analysis. Among those companies were IBM [2], Oracle [3], HP [4], Microsoft [5] etc. Some of the 'big data' solutions are based on Hadoop contributions, others are self-developed and companies' 'big data' portfolios are often amalgated with existing technologies. This is e.g. the case when big data gets blended with existing data management solutions, but also for complex event processing solutions which are the basis (but got further developed) to handle stream processing of big data.

The effort taken by software companies to get part of the big data story is not surprising considering the trends analysts predict and the praise they sing on 'big data' and its impact onto business and even society as a whole. IDC predicts in its 'The Digital Universe' study that the digital data created and consumed per year will grow up to 40000 exabyte by 2020, from which a third will promise value to organizations if processed using big data technologies [6]. IDC also states that in 2012 only 0.5% of potentially valuable data were analyzed, calling this the 'Big Data Gap'. While the McKinsey Global Institute also predicts that the data globally generated is growing by around 40% per year, they furthermore describe big data trends in terms of monetary figures. They project the yearly value of big data analytics for the US health care sector to be around 300 billion $. They also predict a possible value of around 250 billion $ for the European public sector and a potential improvement of margins in the retail industry by 60% [7].

With this kind of promises the topic got picked up by business and management journals to emphasize and describe the impact of big data onto management practices. One of the terms coined in that context is 'data-guided management' [8]. In MIT Sloan Management Review Thomas H. Davenport discusses how organizations applying and mastering big data differ from organizations with a more traditional approach to data analysis and what they can gain from it [9]. Harvard Business Review published an article series about big data in which they call the topic a 'management revolution' and describe how 'big data' can change management, how an organizational culture needs to change to embrace big data and what other steps and measures are necessary to make it all work.

But the discussion did not stop with business and monetary gains. There are also several publications stressing the potential of big data to revolutionize science and even society as a whole. A community whitepaper written by several US data management researchers states, that a 'major investment in Big Data, properly done, can result not only in major scientific advances but also make the foundation for the next generation of advances in science, medicine, and business' [10]. The New York Times e.g. declared 'The Age of Big Data' [11]. There were also books published to describe how big data transforms the way 'we live, work and think' to a public audience and to present essays and examples how big data can influence mankind.

However the impact of 'big data' and where it is going is not without controversies. Chris Anderson, back then editor in chief of Wired magazine, started a discourse, when he announced 'the end of theory' and the obsolescence of the scientific method due to big data [12]. In his essay he claimed, that with massive data the scientific method - observe, develop a model and formulate hypothesis, test the hypothesis by conducting experiments and collecting data, analyze and interpret the data - would be outdated. He discussed that all models or theories are erroneous and the use of enough data allows skipping the modelling step and instead leveraging statistical methods to find patterns without creating hypothesis first.

Boyd and Crawford, while not denying its possible value, published an article to provoke an overly positive and

simplified point of view of 'big data' [13]. One point they raise is, that there are always connections and patterns in huge data sets, but not all of them are valid, some are just coincidental or biased. Therefore it is necessary to place data analysis within a methodological framework and to question the framework's assumptions and the possible biases in the data sets to identify the patterns that are valid and reasonable. It is a process of individual choices and interpretation. It starts with data creation and with deciding what to measure and how to measure it. It goes on with making observations within the data, finding patterns, creating a model and understanding what this model actually means [13]. It further goes on with drawing hypotheses from the model and testing them to finally prove the model or at least give strong indication for its validity. To draw valid conclusions from data it is also necessary to identify and account for flaws and biases in the underlying data sets and to determine which questions can be answered and which conclusions can be validly drawn from certain data. This is as true for large sets of data as it is for smaller samples. For one, having a massive set of data does not mean that it is a full set of the entire population or that it is statistically random and representative [13]. Different social media sites are an often used data source for researching social networks and social behaviour. However they are not representative for the entire human population. They might be biased towards certain countries, a certain age group or generally more tech-savvy people. Furthermore researchers might not even have access to the entire population of a social network. Twitter's standard APIs e.g. do not retrieve all but only a collection of tweets, they obviously only retrieve public tweets and the Search API only searches through recent tweets [13].

While all these discussions talk about 'big data', this term can be very misleading as it puts the focus only onto data volume. Data volume, however, is not a new problem. Wal-Mart's corporate data warehouse had a size of around 300 terabyte in 2003 and 480 terabyte in 2004. Data warehouses of that size were considered really big in that time and techniques existed to handle it. The problem of handling large data is therefore not new in itself and what 'large' means is actually scaling as performance of modern hardware improves. To tackle the 'Big Data Gap' handling volume is not enough, though. What is new, is what kind of data is analyzed. While traditional data warehousing is very much focused onto analyzing structured data modeled within the relational schema, 'big data' is also about recognizing value in unstructured sources. These sources are largely uncovered, yet. Furthermore, data gets created faster and faster and it is often necessary to process the data in almost real-time to maintain agility and competitive advantage e.g. due to noise note that this is often outside the influence of researchers using 'big data' from these sources e.g. the use of distributed databases e.g. text, image or video sources. Therefore big data technologies need not only to handle the volume of data but also its velocity and its variety. Gartner comprised those three criteria of Big Data in the 3Vs model [14]. Coming together the 3Vs pose a challenge to data analysis, which made it hard to handle respective data sets with traditional data management and analysis tools: processing large volumes of heterogeneous, structured and especially unstructured data in a reasonable amount of time to allow fast reaction to trends and events.

These different requirements, as well as the amount of companies pushing into the field, lead to a variety of technologies and products labeled as 'big data'.

3. Challenges presented by big data

a. Designing of filters

All the data generated from various sources like social media (Facebook, Twitter, and Google Plus), sensor networks etc are not important for storage. So to find the useful data from the unwanted data is a big challenge. Although, the generated data needs to be filtered .The separation of useful and needed information is also a time consuming process. The filtering of this redundant data may result in the compression of data by orders of magnitude. Thus filters have to be designed that may filter the data. It presents a great challenge to the people dealing with big data.

b. Ownership as a challenge

Ownership of data is also a challenge. Data is used by multiple people at a time unlike a physical entity. This also poses the question of who owns the data. Are the data collected by one is the personal property of him? Can the said personnel do anything with the collected data or is the personnel permitted to use the data the way he likes. This also marks a question mark on big data.

c. The big data talent shortage

This is the most prominent challenge that comes in the way of big data. Talent is an important asset for the growth of an emerging technology but unfortunately this is not to its fullest for big days as of now. For example, there will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone would face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the knowledge of-how to use the analysis of big data to make effective decisions. [15] Data Scientists are unavailable the days with the skills as defined by Granter in [16]. Professionals having this wide range of skills are rare and this explains why data scientists are currently in short supply.

d. Big data complexity

A number of challenges pertaining to the complexity are presented by big data. Since huge amount of data comes in the unstructured form like video, comments, images, other non-numeric data etc, so the challenges are like how we can capture, storage, search, analyze, curate, visualize, use etc this data. It is often unclear how the data should be interpreted. The main complexity is how we can analyze and understand this data given its volume and our computational capacity. Since big data must support present and future datasets. The algorithms that deal with such data should have provisions for expendability and scalability (algorithms. must be able to process increasingly expanding and more complex datasets).

e. Security as a challenge with big data

Once the data has been collected the question is how to protect and control this data i.e. how to secure this data. Because the data contains sensitive information. There is always an argument that what data must be collected and how to use that data or the data is shared without violating people's privacy. The huge datasets could not be effectively maintained and analyzed at present by the data service providers or owners. Their dependency on third party increases the potential risks as the data contains critical details like financial details. It is the responsibility of analysts to device methods for keeping the data secure.

f. The database as a challenge

The traditional data management and analysis are based on Relational Database Management Systems (RDBMS),that can only handle structured data but not semi-structured or unstructured. The emerging data brings huge challenges on data acquisition, analysis, storage and management. Traditional data management cannot handle this huge volume and heterogeneity of data. According to O'Reilly, ―Big data is data that exceeds the processing capacity of conventional database systems. To take advantage of these data, there must be an alternative way to process it [17] Distributed file systems [18] and NoSQL [19] may be the proposed solutions for the permanent storage and management of large –scale disordered datasets respectively. Above all there is funding challenge as well for the big data .Event though funds are released for the betterment of this technology yet they prove to be a minimum.

4. Techniques already used

a) Hadoop

Hadoop is an open source framework for writing and running distributed applications which are capable of batch processing large sets of data. Hadoop framework is mainly known for MapReduce and its distributed file system (HDFS). The MapReduce algorithm, that consists of two basics operations: "map" and "reduce" is a distributed data processing model that runs on large cluster of machines. Hadoop started out as a subproject of Nutch created by Doug Cutting. Nutch is itself an extension of Lucene, a full-featured text indexing and searching library. Nutch tried to build a complete web search engine using Lucene as its core component. Around 2004, Google published two papers describing the Google File System (GFS) and the MapReduce framework. Google affirmed to use these technologies for scaling its own search systems. Doug Cutting immediately saw the applicability of these technologies to Nutch and started to implement the new framework and ported Nutch to it. The new version of Nutch boosted its scalability. Then Doug Cutting created a dedicated project that brought together an implementation of MapReduce and a distributed file system. Hadoop was born. In 2006, Cutting was hired at Yahoo! to work on improving Hadoop as an open source project.

b) MapReduce

MapReduce is a data processing model. Its biggest advantage is the easy scaling of data processing over many computing nodes. The primitives to achieve data processing with MapReduce are called mappers and reducers. The data model is quite strict and, sometimes, decomposing a data processing application into mappers and reducers is not so easy. But once the application is written in the MapReduce form, scaling it to run over thousands of machines in a cluster is trivial. A MapReduce job takes as input files in HDFS. The mapper is responsible for splitting lines in the files word by word. Then the data model in MapReduce is based on key -value pairs. The word (the key) is mapped with a value the last step in the mapper is done by the partitioner. The partitioner decides the target reducer for the key value pairs. Then, in the mapper the key value pairs are sorted and grouped by key. Finally, the final result is composed of files in HDFS.

c) Hadoop limitations

Actually, Hadoop does its job very well; processing huge set of data. But the model has some limitations. Firstly, the data model is quite restrictive; it is not always easy to transform our problems into key value pairs. Moreover, Hadoop can only process a finished set of data; it means that MapReduce can only process data by batch. Therefore, when a batch is finished, data is already aged by, at least, the time required by the batch. Hadoop is definitely not a good fit for processing the latest version of the data.

5. Conclusion

This is a time of big data. So to analysis big volume and variety of data, a lot of research work has been done. Some survey has been presented in literature survey regarding this work. A main challenge in the field of big data is also being included. The clear understanding of these challenges will facilitate the researchers to better understand and design optimal ways of extracting useful information from large data source in a time efficient manner. Technologies for big data like Hadoop and MapReduce technology are also being discussed.

References

[1] Hadoop 1.2.1 Documentation. Online documentation, Apache Software Foundation, 2013. URL http://hadoop.apache.org/docs/r1.2.1/index.html.
[2] What is big data? Company website, IBM, 2014. URL http://www-01.ibm.com/software/data/bigdata/.
[3] Oracle and Big Data: Big Data for the Enterprise. Company website, Oracle, 2014. URL http://www.oracle.com/us/technologies/big-data/index.html.
[4] Big Data Solutions. Company website, Hewlett-Packard, 2014. URL: http://www8.hp.com/us/en/business-solutions/big-data-overview.html.
[5] Big Data. Company website, Microsoft, 2013. URL http://www.microsoft.com/enterprise/it-trends/big-data/.
[6] John Gantz and David Reinsel. THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Study

report, IDC, December 2012. URL www.emc.com/leadership/digital-universe/index.htm
[7] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh and Angela Hung Byers , Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. May 2011.
[8] S. Lohr, "The Age of Big Data" The New York times Publication, February 2012.
[9] Thomas H. Davenport, Paul Barth, and Randy Bean. How 'Big Data' Is Different. MIT Sloan Management Review, Fall 2012, July 2012. URL http://sloanreview.mit.edu/article/how-big-data-is-different/.
[10] Divyakant Agrawal, Philip Bernstein, Elisa Bertino, Susan Davidson, Umeshwar Dayal, Michael Franklin, Johannes Gehrke, Laura Haas, Alon Halevy, Jiawei Han, H. V. Jagadish, Alexandros Labrinidis, Sam Madden, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, Kenneth Ross, Cyrus Shahabi, Dan Suciu, Shiv Vaithyanathan, and Jennifer Widom. Challenges and Opportunities with Big Data: A community white paper developed by leading researchers across the United States. Whitepaper, Computing Community Consortium, March 2012. URL http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf.
[11] S. Lohr, "The Age of Big Data" The New York times Publication, February 2012.
[12] Chris Anderson. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired Magazine, 16.07, July 2008. URL http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.
[13] Danah Boyd and Kate Crawford. Six Provocations for Big Data. In A Decade in Internet Time:Symposium on the Dynamics of the Internet and Society, September 2011. doi: 10.2139/ssrn.1926431.
[14] Kaiser J. Giri ,Towseef A. Lone, Big Data - Overview and Challenges, International Journal of Advanced Research in Computer Science and Software Engineering 4(6), June - 2014, pp. 525-529
[15] Source: McKinsey Global Institute; Big data: The next frontier for innovation, competition, and productivity
[16] Douglas Laney, Lisa Kart. Emerging Role of the Data Scientist and the Art of Data Science. [Online] Available from: http://www.gartner.com/id=1955615
[17] Edd Dumbill. What is big data? [Online] Availablefrom:http://radar.oreilly.com/2012/01/what-is-big data.html
[18] Howard JH, Kazar ML, Menees SG, Nichols DA,Satyanarayanan M, Sidebotham RN, West MJ (1988) Scale and performance in a distributed filesystem. ACM Trans ComputSyst (TOCS) 6(1):51–81
[19] Cattell R (2011) Scalable sql and nosql data stores. ACM SIGMOD Record 39(4):12–27