International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 3, Issue 4, 2016, PP 16-20
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
DOI: http://dx.doi.org/10.20431/2349-4859.0304003
www.arcjournals.org

Big Data and Challenges of the Century

Dr. Mozamel M. Saeed

Associate professor, Department of Computer Science,
Collage of Science Prince Sattam Bin Abdul-Aziz University, KSA

Abstract: *Research conducted to explore the importance and value of Big data In today's world. Its sources, features, and challenges. Big Data is used by many companies and organizations, for analysis purpose, utilizing their personal as well as organizational benefits, profits and achieving goals.*

It is not easy to process this data using data processing applications which are traditional. There are challenges like analysis, search, duration, exchange of data, storage of data, visualization, and privacy and security violations. The large database is in trend due to the large information derived by analysis of large co-relevant data set, as and when they compared with different smaller sets with the same total amount of data, allowing relationships to be found in "spotting business trends, to prevent diseases, controlling crime and so on." It is tough to process Big data using specifically the co-relational database sets and management systems as well as desktop statistics and visualization packages, instead of that requiring "parallel software executing on tens, few, or even thousands of servers". So Big data generally includes database with sizes which are beyond the ability of generally used software techniques to capture, cure, manage, and processing of data in the tolerable specified time. In Big data "size" is a continuously moving goal, as it is changing from a some dozen terabytes to many petabytes of data.

The paper concluded that the Hadoop can resolve the problem of Big Data and will be able to overcome the future challenges. The paper also concluded that Big Data should be handled carefully. Because of its scale, diversity, and complexity, and any misusing of such information (Big Data) will cause customers loss part of trust and faith in organizations.

Keyword: *Big data, HDFS, Hadoop, Map-Reduce, challenges of century.*

1. Introduction

Due to the abundance of information and its increased capacity, serious attention should be directed towards human frameworks capable of absorbing as much of the useful information. Coordinating, controlling, and exploiting the biggest payoff possible for development goals. This certainly has become the most dangerous and most prominent developmental community masimiz and civilization. The aim is how to create entities and technical methods that will exploit the most possible payoff information.

Now a days Data is considered as the heart of all organizations and companies. It can be used in myriad ways to run the business, market to customers, forecast sales, measure performance, gain competitive advantage, and discover new business opportunities. And lately, a convergence of new technologies and market dynamics has opened a new frontier for information management and analysis. This new generation of computing includes data with too greater volume, variety & velocity, than ever before.

Big Data is being used in smart ways to predict customer habits regarding shopping, detection of fraud, waste and analyze product, and quickly react to events and conditions of business changes. New business opportunities are derived by this force. Most companies are already using this analytics in the form of reports and dashboards to help run their business. This is largely based on well-structured data from operational systems that conform to pre-determined relationships. Although this model is not followed by Big Data in structural manner. All different streams are there and it is so difficult to establish the common relationships. But with its diversity and abundance come opportunities to learn and to develop new ideas, the ideas that can help change the business .To run this business, you should organize data to make it something specific; to change the business, you take data as is and determine what data can do for you. As we know power of unity this fact is true in this case also. So rather than using this approaches alone it is better to use the both together. Based on the survey it is observed that many innovative and latest solutions are combined approach of both techniques.

2. Definition of Big Data

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data duration, search, sharing, storage, transfer, visualization, querying and information privacy. The term often refers simply to the use of predictive analytics or certain other advanced methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk.

3. Sources of Big Data

Big data comes from different sources like sensors which are used to gather climate information from social media sites while posting digital pictures and videos from transaction records while purchasing goods, and from cell phone GPS signals etc and internet transactions. as shown in fig1.

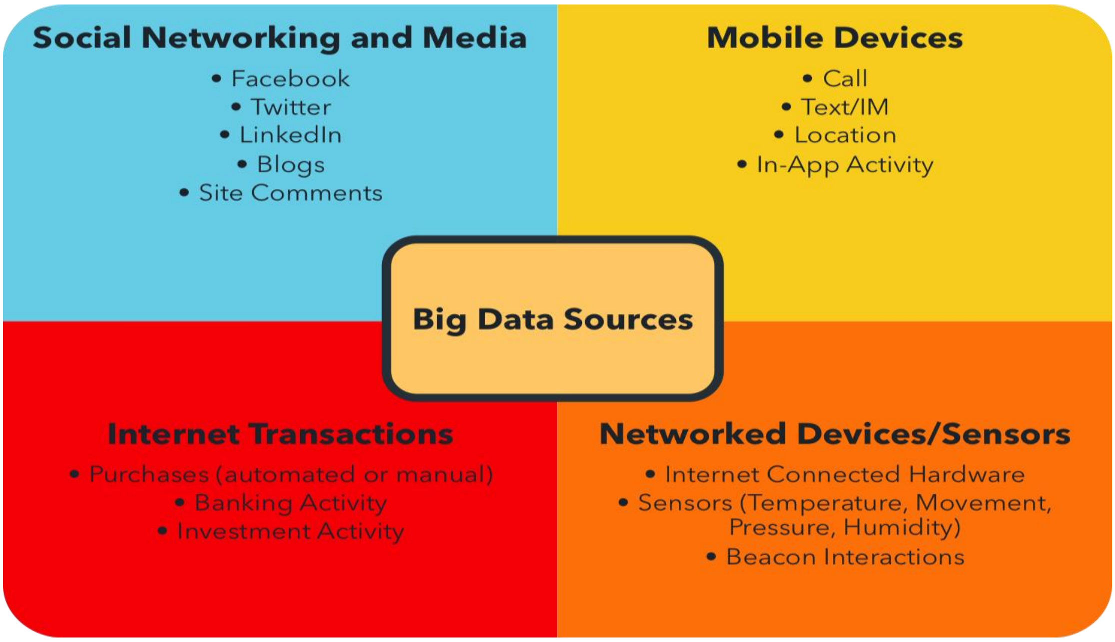

Fig1. *Sources of Big Data*

Using social media like Face book we produce lots of big data when we use to look it as when we look at another person's timeline, send or receive a message, when we post things like photos or videos on Face book etc. We receive data from or about the computer, mobile phone, or other devices you use to install Face book apps or to access Face book .We receive data whenever we visit a game, application, or website that uses Face book Platform. Sometimes we get data from some advertising partners, customers and other third parties that help us to deliver ads.

Different Physical Sensor data also produces high velocity, volume, variety of data. Sensors are used for geolocation, temperature, pressure, noise, humidity, biometrics purposes to collect data in a variety of ways. When we use this large data to user context and to predict behavior we have to process this huge data that is a big challenge.

The progress and innovation is no longer hindered by the ability to collect data. But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion is necessary.

4. Big Data Characteristics

Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization. Big data can be described by the following characteristics:

4.1. Volume

Big data is the term which indicates large volumes of data. It is used to create data or database of employees. Now that data is generated and obtained by machinaries, networks and human interaction with systems like social media. The volume of data to be analyzed is massive. Still, the volume of data is not as much the problem as other V‘s like veracity.

4.2. Variety

Variety tells us the phenomenon that many types and sources of data both which are unstructured and structured. The data is being stored with the help of various sources such as database as well as spreadsheets. Recently data is available in the several form like photographs, mails, as well as videos, various monitoring devices, PDF files, audio tracks or clips, etc. This so large and variety of unstructured data creates problems for mining, storage and analysis of data.

4.3. Velocity

Today Data is generated too fast and also need to be processed fast. Big Data Velocity deals with the pace at which data flows in from sources like business processes, machines, networks and human interaction with things like social media sites, mobile devices, etc. The flow of data is massive and continuous. This real-time data can help researchers only if you are able to handle the velocity. Velocity is applied to data in motion. There are various information streams and the increase in sensor network deployment has led to a constant flow of data at a pace that has made it impossible for traditional systems to handle. Initially analysis of data is done by using a batch process. With the new sources of data such as social and mobile applications, the batch process breaks down. The data is now streaming into the server in real time, in a continuous fashion and the result is only useful if the delay is very short.

4.4. Veracity

Big Data Veracity refers to the biases, noise and abnormality in data. Is the data that is being stored, and mined meaningful to the problem being analyzed. Veracity in data analysis is the biggest challenge as compared to volume and velocity. In scoping your big data strategy ones need to have his own team and partners together to work to clean the data in the system to avoid ‘dirty data’ from accumulating in his systems.

4.5. Validity

The validity implies us that data is correct and accurate for intended use. Clearly to make right decisions the valid data is key. Also the tools they offer to help with data veracity and validity.

4.6. Volatility

Volatility is the concept in Big Data which gives the answers of two important questions. One of them is data validity i.e. how huge data is valid and second one is for how long time it should be stored. In this world of real time data you need to determine at what point is data no longer relevant to the current analysis. Big data management clearly deals with issues beyond volume, variety and velocity to other concerns like veracity, validity and volatility to hear about other big data trends and presentation and uses.

5. HADOOP : A BIG DATA MANAGEMENT PLATFORM

Hadoop is a processing engine that is designed to handle extremely high volumes of data in any structure. Hadoop provides distributed storage and distributed processing for very large data sets. Hadoop has two main components:

The Hadoop distributed file system (HDFS), which supports data in structured relational form, in unstructured form, and in any form in between. HDFS is a reliable distributed file system that provides high through put access to data.

The MapReduce programing paradigm which is meant for managing applications on multiple distributed servers. It is a framework for performing high performance distributed data processing and is based on divide and aggregate paradigm.

HDFS, or the Hadoop Distributed File System, gives the programmer unlimited storage. The advantages of HDFS are:

Horizontal scalability: Thousands of servers holding petabytes of data. When you need even more storage, we don't switch to more expensive solutions, but add servers instead.

Commodity hardware: HDFS is designed with relatively cheap commodity hardware in mind. HDFS is self-healing and replicating.

Fault tolerance: Hadoop also deal with hardware failures.

MapReduce takes care of distributed computing. It reads the data, usually from its storage, the Hadoop Distributed File System (HDFS), in an optimal way. However, it can read the data from other places too; including mounted local file systems, the web, and databases. It divides the computations between different computers (servers, or nodes). It is also fault-tolerant. If some of nodes fail, Hadoop knows how to continue with the computation, by re-assigning the incomplete work to another node and cleaning up after the node that could not complete its task. It also knows how to combine the results of the computation in one place

A set of machines running HDFS and MapReduce is known as a Hadoop Cluster. Individual machines are known as nodes. A cluster can have as few as one node, as many as several thousands. More nodes in the cluster mean better is the performance.

6. Challenges in Big Data

As we are aware now the exclusive importance of big data into the enterprise. The practitioners of information technology (IT) and business sponsors alike will have to face against a number of challenges that must be addressed before any big data program can be successful. Five of those challenges are:

6.1. Uncertainty of The Data Management Landscape

There are so many competitive techniques and technologies, and at the same time within each technical area there are numerous rivals also. Our first challenge is making the best choices while not introducing additional unknowns and risk to big data adoption.

6.2. The Big Data Talent Gap

The Talent gap is the second challenge faced by us. - The curiosity and excitement about big data and applications of big data shows to imply that there is a community of experts which is broad enough available to help us in implementation. But, this is not yet the case. That's why talent gap is our second challenge.

6.3. Getting Data into the Big Data Level

The volume and assortment of data to be grasped into a big data state of affairs can get over the unprepared data practitioner. Making data accessibility as well as integration of data is our third challenge.

6.4. Lineament relation Across the Data Sources

Data base from various resources are co-ordinated into an synthetical platform. The potential for time delays to impact data currency and similarity is nothing but our fourth challenges.

6.5. Getting Useful Information out of the Big Data Platform

This is last but not least. Using big data for various purposes starting from storage diminution to enabling high-performances analysis is obstructed .But the condition is if the information cannot be adequately conditioned back within the all other components of the enterprise information structure. That's why big data syndication is our fifth challenge.

7. Conclusion

We are leaving in a new age of Big Data technology which can lead to innovation, techniques, and automation. In today‘s world various organizations like Banking, Insurance companies, Government sectors, and Educational Sectors are using Big Data for analysis purpose. Many organizations are carefully utilizing Big data for their personal as well as organizational benefits, profit and also to achieve their goals.

This paper explored what Big Data actually means, its sources, characteristics, Features, full Applications in today‘s world, and the challenges it face. The paper concluded that the Hadoop can resolve the problem of Big Data and will be able to overcome the future challenges. The paper also concluded that Big Data should be handled carefully. Because of its scale, diversity, and complexity, and any misusing of such information (Big Data) will cause customers loss part of trust and faith in organizations.

REFERENCES

[1] Agrawal R.,Srikant R., ``Privacy Preserving Data Mining., "In the Proceedings of the ACM SIGMOD Conference.2000.

[2] Big Data Analytics‖ ericsson White paper,284 23-3211 Uen, August 2013.

[3] Big Data Analytics for Security Intelligence,, CLOUD SECURITY ALLIANCE , September 2013.

[4] Ben Spivey, Joey Echeverria, Hadoop Security: Protecting Your Big Data Platform Paperback – May, 2015, Edition: 1st

[5] Beyer, M.A, Laney, D.: The Importance of 'big data': a Definition. Gartner,.2012.

[6] Boyd, Danah and Kate Crawford, ―Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon.‖Information, Communication, & Society 15:5, p. 662-679(2012).

[7] C.L. Philip Chen, Chun-Yang Zhang.‖ Data-intensive applications, challenges, techniques and technologies A survey on Big Data‖ . Elsevier 2014.

[8] Garry Turkington, Gabriele Modena, Learning Hadoop – February,2015.

[9] Hilbert, Martin. "Big Data for Development: A Review of Promises and Challenges. Development Policy Review.". martinhilbert.net. Retrieved 2015-10-07.

[10] Ira S. Rubinstein, ‗ Big Data: The End of Privacy or a New Beginning, International Data Privacy Law Advance Access published January 25, 201

[11] Kyuseok Shim, MapReduce Algorithms for Big Data Analysis, DNIS 2013. [12]Michael Frampton, Big Data Made Easy: A Working Guide to the Complete Hadoop Toolset – December, 2014, Edition: 1st

[12] P.Kamakshi. ―SURVEY ON BIG DATA AND RELATED PRIVACY ISSUES ― IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308 DEC 2014.

[13] P. Russom, ― Big Data Analytics‖, Best Practices Report, Fourth Quarter, The DataWarehouse Institute, Renton, WA, September 18 2011 [11]. IDC, Digital data to double every 18 months, worldwide marketplace model and forecast, Framingham, MA. available at www.idc.com May 2009

[14] R. Agrawal, R. Srikant, ―Privacy-preserving data mining‖, In: Proceedings of the 2000ACM-SIGMOD on management of data, Dallas, TX, USA, May 15-18, 2000.

[15] Thomas M. Lenard and Paul H. Rubin, ―The Big Data Revolution: Privacy Considerations‖, December 2013.

[16] Vinayak Borkar, Michael J. Carey, Chen Li, Inside “Big Data Management”:Ogres, Onions, or Parfaits?, EDBT/ICDT 2012 Joint Conference Berlin, Germany,2012 ACM 2012,

[17] VINT research report on ― Privacy, technology and the law Big Data for everyone through good design.

AUTHOR’S BIOGRAPHY

Dr. Mozamel M. Saeed, is the head department of Computer Science at Faculty of Science, Sattam Bin Abdul Aziz University. I've published some books & papers internationally.