AN OVERVIEW OF A BIG DATA MANAGEMENT AND THE PERSONAL PRIVACY CONSTRAINT

Oguntimilehin A.
Department of Computer Science
College of Sciences
Afe Babalola University
Ado-Ekiti, Nigeria
ebenabiodun2@yahoo.com

Ademola E.O
Department of Computer Science
College of Sciences
Afe Babalola University
Ado-Ekiti, Nigeria
ademolaeo@abuad.edu.ng

ABSTRACT

Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze. The accumulated huge amount of data that previously of no significant importance or value have been put into maximum use due to the availability of newly designed Big Data tools that surpass earlier available data mining tools. Big Data is now of tremendous importance to organizations and data mining researchers because better results are gotten from larger volume of data. Predictions and Analysis of business are becoming more accurate and interesting with the advent of Big Data Tools. The scale and scope of changes that Big Data are bringing about are at an inflection point, set to expand greatly, as a series of technology trends accelerate and courage. Data have always been part of information Technology (IT), and then the birth of Big Data is a plus to the IT profession. In this paper, we introduced readers to the concept of Big Data, the various sources of data for Big Data. Some of the advantages and applications that have been successfully implemented using Big Data tools. Some of the challenges of Big Data were also discussed with special reference to the most crucial of these challenges- the privacy or personal privacy issue which if not well managed could bring an individual or an entire organization using Big Data down. This paper is written to create awareness to researchers and to sensitize the existing and intending users of Big Data tools of the privacy issue and possible measures that can be of assistance.

Keywords: Data, Big Data, Big Data tools, Challenges, Security, Privacy.

1. INTRODUCTION

Data has become a torrent flowing into every area of the global economy. Companies churn out a burgeoning volume of transactional data, capturing trillions of bytes of data about their customers, supplies, and operations. Millions of networked sensors are being embedded into the physical world in devices such as mobile phones, smart energy meters, automobiles, and industrial machines that sense, create and communicate data in the age of the internet of things. Indeed, as companies and organizations go about their business and interact with individuals, they are generating a tremendous amount of digital "exhaust data" i.e, data that are created as a by-product of other activities. Social media sites, smart phones, and other consumer devices including PCs and laptops have allowed billions of individuals around the world to contribute to the amount of Big Data available and the growing volume of multimedia content has played a major role in the exponential growth in the amount of Big Data [6].

Organizations have inundated with data – terabytes and petabytes of it. To put it in context, 1 terabyte contains 2,000 hours of CD quality music, and 10 terabytes could store the entire US library of Congress print collection. Exabytes, zettabytes and yottabytes definitely are on the horizon. Data is pouring from every conceivable direction; from operational and transactional systems, from scanning and facilities management systems, from inbound and outbound customer contact prints, from mobile media and the web. 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 [8].

In today's world, we are surrounded by predictions. Problems with statistics and predictions are not limited to graphic representation and in fact, can be more complicated and challenging, especially with the advent of Big Data and its use in

making projections [3]. The source of data growth that are driving Big Data technology investments vary widely. Some represent entirely new data sources, while others are a change in the "resolution" of existing data generated [13].

2. WHAT IS BIG DATA?

Big Data is a term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis and visualization. The trend to larger data sets is due to the traditional information derivable from analysis of a single related data as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions [9]. Big Data is a term applied to data sets whose size is beyond the capability of commonly used software tools to capture, manage and process. The sheer size of data, combined with complexity of analysis and commercial imperative to create value from it has led to a new class of technologies and tools to tackle it. The term Big Data tends to be used in multiple ways, often referring to both the type of data being managed as well as the technology being used to manage it. In the most part, these technologies originated from companies such as Google, Amazon, face book, and linked-in, where they were developed for each company's own use in order to analyze the massive amounts of social media data they were dealing with [11].

As of 2012, limits on the size of data sets that were feasible to process in a reasonable amount of time were on the order of exabytes of data. Scientists regularly encounter limitations due to large data sets in many areas, including meteorology, genomics, connectomics, complex physics simulations, and biological and environmental research. Data sets grow in size in part because they are increasingly being gathered by ubiquitous information sending mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio- frequency identification readers, and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day, 2.5 exabytes (2.5×10^{18}) of data were created . The challenge for large enterprises is determining who should own Big Data initiatives that straddle the entire organization [7].

Big Data is relatively a new concept and a lot of definitions have been given to it by researchers, organizations and individuals. As far back as 2001, industry analyst Doung Laney (currently with Gartener), articulated the mainstream of definition of Big Data as the three Vs; Volume, Velocity and Variety. At SAS, SAS considered two additional dimensions when thinking about Big Data: the Variability and Complexity [8]. Our studies on Big Data show that Oracle defined Big Data in terms of four Vs – Volume, Velocity, Variety and Value [11]. Having gone through the literature of Big Data, in this paper, we will like to bring the definition of Big Data to a new state based on its genesis, bogusness and values. We define Big Data in terms of five Vs and a C.

These form a reasonable test as to determine if a Big Data approach is the right one to adopt for a new era of analysis. The five Vs are;

- Volume: the size of data. With technology, it is often very limiting to talk about data volume in any absolute sense. As technology marches forward, numbers get quickly outdated, so it is better to think about volume in a relative sense instead. If the volume of data you are looking at is an order of magnitude or larger than anything previously encountered in your industry, then you are probably dealing with Big Data. For some companies, this might be 10s of terabytes, for others, it might be 10s of petabytes [11].
- Velocity: data is streaming at unprecedented speeds and must be dealt with in a timely manner [8]. The rate at which data is being received and has to be acted upon is becoming much more real-time. While it is unlikely that any real analysis will have to be completed in the same time period, delays in execution will inevitably limit the effectiveness of campaigns, limit interventions or lead to sub-optimal processes [11].
- 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 [8].
- Variability: In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Is something trending on social media? Daily, seasonal and event-triggered peak data loads can be challenging to manage, even more so with unstructured data involved.
- Value: We need to consider what commercial value any new sources and forms of data can add to the business or scientific research. Are the existing problems that have defiled solutions due to unavailability of data now being solved?
- 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.

3. OPPORTUNITIES WITH BIG DATA

We are awash in a flood of data today. In a broad range of application areas, data is being collected at unprecedented scales. Decisions that previously were based on guesswork or on painstakingly constructed models of reality can now be made based on the data itself. Such Big Data analysis now drives nearly every aspect of our modern society, including mobile services, retail manufacturing, financial services, life sciences and physical sciences [4]. Scientific research has been revolutionized by Big Data. The Sloan Digital Sky Survey has become a central resource for astronomers the world over. The field of astronomy is being transformed from one where taking pictures of the sky was a large part of an astronomer's job to one where the pictures are all in a database already and astronomers' task is to find interesting objects and phenomena in the database [5].

In the biological sciences, there is now a well established tradition for depositing scientific data into a public repository, and also of creating public databases for use by other scientists. In fact, there is a discipline of bioinformatics that is largely devoted to the curation and analysis of such data. As technology advances, particularly with the advent of next generation sequencing, the size and number of experimental data sets available is increasing exponentially [5]. Big Data has the potential to revolutionize not just research, but also education. A recent detailed quantitative comparison of different approaches taken by 35 Charter schools in NYC has found that one of the top five policies correlated with measurable academic effectiveness was the use of data to guide instruction. Imagine a world in which we have to access a huge database where we collect every detailed measure of every student's academic performance. This data could be used to design the most effective approaches to education, starting from reading, writing, and math, to advanced, college-level courses. We are far from having access to such data, but there are powerful trends in this direction. In particular, there is a strong trend for massive web deployment of educational activities, and this will generate an increasingly large amount of detailed data about students' performance. It is widely believed that the use of information technology can reduce the cost of healthcare, while improving its quality by making care more preventive and personalized, and basing it on more extensive (home-based) continuous monitoring. McKinsey estimates a savings of 300 billion dollars every year in the US alone [5].

The use of Big Data will become a key basis of competition and growth for individual firms. From the standpoint of competitiveness and the potential capture of value, all companies need to take Big Data serious. In most industries, established competitors and new entrants alike will leverage data-driven strategies to innovate, compete, and capture value from deep and up to real time information [6]. The use of Big Data will matter across sectors, some sectors are set for greater gains. The computer and electronic products and information sectors, as well as finance and insurance, and government are poised to gain substantially from the use of Big Data [6]. In a similar vein, there have been persuasive cases made for the value of Big Data for urban planning (through fusion of high-fidelity geographical data), intelligent transportation (through analysis and visualization of live and detailed road network data), environmental modeling (through sensor networks ubiquitously collecting data), energy saving (through unveiling patterns of use), smart materials (through the new materials genome initiative), computational social sciences (a new methodology fast growing in popularity because of the dramatically lowered cost of obtaining data), financial system risk analysis(through integrated analysis of a web of contracts to find dependencies between financial entities), homeland security (through analysis of social networks and financial transactions of possible terrorists), computer security (through analysis of logged information and other events, known as security information and event management (SIEM)), and so on [5].

Some notable achievement involving Big Data

- The LAPD and university of California are using Big Data to predict crime before it even happens .
- Google Flu trends uses search terms to predict the spread of flu virus.
- Statisticians Nate Silver predicted the outcome of the US election down to each individual state in 2012.
- MIT is using mobile phone data to establish how people's locations and traffic patterns can be used for urban planning [1].

4. BIG DATA TOOLS

Big Data is a very complex subset of technology that can be difficult to implement with best practices still being defined. Many problems being solved with Big Data can be solved with the existing tools; they may just require a better implementation. That said, even though the problem may be solved with an existing tool, the cost of solving it may make a Big Data Solution a better option. We shall look at the categories of Big Data tools. First is Batch processing, Hadoop being the most notable tool. Hadoop is designed as a data storage and batch processing engine. It is very easy to load data into, but expect it to take minutes to hours to get the answer to your query. Some of the new SQL on Hadoop tools are enabling faster access to data in Hadoop, but as its core Hadoop is still a batch engine not suited for real time loading and processing of data [4].

Hadoop is designed for large volumes of data and is batch oriented in nature. Even a simple query may take minutes to come back. The dominant Big Data technologies in use today commercially are Apache's Hadoop and No-SQL databases [11]. Hadoop is appealing because it is open-source; therefore there is no software license fee [10]. Since Hadoop is (currently at least) batch oriented, other technologies or tools are required in order to support real-time interactions. The most common technologies currently in use within this area are Complex Event Processing (CEP), In-memory distributed data grids, In-memory database and traditional databases.

These may be supported by other related technologies such as No-SQL databases, either sitting on top of a Hadoop cluster or using a specific data storage layer [11]. No-SQL or New SQL tools are generally designed for fast ingestion and fast access to individual records. However, No-SQL databases usually are not built for aggregation or in-database processing of the data. It is possible to do aggregation and in-database processing of the data in these tools, but access to this aggregated data is not as fast as accessing individual records. These tools primary job is to ingest and make individual records available in real-time [4]. The final category is massively parallel processing (MPP) and Analytic databases. These are designed for every fast aggregation of data, but not fast loading of data. This makes them appropriate as a backend for a reporting and data science environment, but not as a transactional database for front end systems. There are also sub-category of SQL on Hadoop tools that are essentially MPP databases that use HDFS as their full system and sit on the same hardware as HDFS and Map Reduce processes. These databases usually require the data to be loaded into database proprietary file formats to achieve the speed advantage [4]. There are Big Data tools designed for batch processing of large amounts of data, designed for real time ingestion and access of data but not processing, and designed for speed of thought aggregation of data but not for fast loading. You need to determine what your business needs require [4].

5. CHALLENGES OF BIG DATA

On closer inspection, however, only two or three main issues appear capable of making or breaking the promise of Big Data, and these are related to: solution approach, personal privacy and intellectual priority (IP). The first issue deals with technology, deployment and the organizational context, whereas the latter two big ticket items raise concerns about the nature and applicable use of information or Big Data [2]. Other potential threats to the full utilization of Big Data are heterogeneity and incompleteness, scale, timeliness; another closely related concern is data security [5].

Heterogeneity: When humans consume 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 homogenous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step to (or prior to) data analysis. Computer systems work most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation, access and analysis of semi-structured data require further work [5].

Scale: As identified above, scale or volume is another major challenge of Big Data. The first thing anyone thinks of with Big Data is its size. After all, the word "Big" is there on the very name. Managing large and rapidly increasing volumes of data has a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore's law, to provide us with the resources needed to cope with increasing volume of data. But, there is a fundamental shift underway now: data volume is scaling faster that computer resources and CPU speeds are static. These unprecedented changes require us to rethink how we design, build and operate data processing components [5].

Timeliness: The flip side of size is speed. The larger the size of the data to be processed, the longer it will take to analyze it. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of velocity in the context of Big Data. Rather, there is an acquisition rate challenge and a timeliness challenge. There are many situations in which the result of the analysis is required immediately. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed-Potentially preventing the transaction from taking place at all. Obviously, a full analysis of a user's purchase history is not likely to be feasible in real-time. Rather, we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination [5].

Personal Privacy: Think of all the personal information that is stored and transmitted through ISPs, mobile network operators, supermarkets, local councils, medical and financial service organizations (e.g hospitals, banks, insurance and credit card agencies). Also, not forgetting information shared and stored on social networks, by religious organizations, educational institutions and or employers. Each organization has the headache of organizing, securing and exploiting their business, operational and customer data [2].

Clearly, privacy is an issue whose importance, particularly to customers, is growing as the value of Big Data becomes more apparent. However, people do still care about what and how their personal information is used, especially if it could become disadvantageous or harmful to them. There is a certain class of data which can easily become 'toxic' should a company suffers any loss of control, and it includes: personal information, strategic IP information, corporate sensitive data (e.g KPIs and results). The situation is further complicated by differing world of views on personal privacy as a constitutional and fundamental human right. The UK's Data protection Act is not applicable to personal information stored outside the UK, yet we deal daily with organizations, processes and technologies that are global in scale and reach. On the other hand, some users are happy to share personal data in exchange for financial gain [2]. The privacy of data is a huge concern, and one that increases in the context of Big Data. For electronic health records, there are strict laws governing what we can and cannot be done. For other data, regulations, particularly in the US are less forceful. However, there is a great fear regarding the inappropriate use of personal data, particularly through linking of data. For Nigeria, I do not think there is any and if there is, not fully emphasized, disseminated and strictly adhere to.

From multiple sources, consider the following example, data gleaned from location based services. These new architectures require a user to share his/her location with the service provider, resulting in obvious privacy concerns. Note that hiding the user's identity alone without hiding her location would not properly address this privacy concern. An attacker or a (potentially malicious) location based server can infer the identity of the query source from its (subsequent) location information. For example, a user's location information can be tracked through several stationary connection points (e.g., cell towers). After a while, the user leaves "a trail of packet crumbs" which could be associated to a certain residence or office location and thereby used to determine the user's identity. Several other types of surprisingly private information such as health issues (e.g., presence in a cancer treatment center) or religious preferences (e.g., presence in church) can also be revealed by just observing anonymous users' movement and usage overtime. Many online services require us to share private information, but beyond record-level access control we do not understand what it means to share data, how the shared data can be linked, and how to give users fine-grained control over this sharing [5].

Out of the numerous challenges facing Big Data today, privacy or personal privacy is the most important issue that needs urgent attention. For instance with availability of one's phone number or date of birth, attackers may gain a lot of ground to take over one's personal privacy. To protect competitively sensitive data or other data that should be kept private, addressing data security through technological and policy tools is essential. Questions about the intellectual property rights attached to data will have to be answered. Who "owns" a piece of data and what rights come attached with a dataset? What defines "fair use" of data? There are also questions related to liability, who is responsible when an inaccurate piece of data lead to negative consequences? As well as data repudiation i.e. the legitimate owner of the data cannot disown it legitimately. All these types of legal issues will need clarification, probably overtime, to capture the full potential of Big Data. When confidential information has been used or published by someone who has no right to do so, in certain circumstances, the law will provide some redress for its owner. But it is vital to use secure technical and practical means effectively to keep the information safe in the first place [12].

Portable devices containing personal data should be encrypted. On the social media, there should be restriction on exposing sensitive personal data that can be of assistance to the attackers. Individuals should be sensitized on the dangers associated with the exposure of personal data. Each country of the world should have legislative laws to guide the use of personal data. Organizations should encrypt personal data and only grant access to a legitimate user according to predefined policy. Data protection policy and enforcement should be a matter of priority. Companies need to present users with workable security guidelines. For the IT department, protecting personal data should be an on-going concern requiring constant review. Employers are faced not only with the threat of financial penalties but with immeasurable consequences of loss of reputation. Only recently it was reported that the personal details of about six million people have been in advently exposed by a bug in Face book's data archive.

6. CONCLUSION

Big Data is of economic and scientific importance. It is a scientific belief that the bigger the data used in a research, the better the accuracy. Data are created every second in real life which means the volume of data available can never reduce but increase. In fact, IDC's Digital Universe study predicts that between 2009 and 2020 digital data will grow 44 folds to 35ZB per year. It is also important to recognize that much of this data explosion is the result of an explosion in devices located at the periphery of the network, including embedded sensors, smart phones and tablet computers. All of these data create new opportunities for data analysts in human genomics, healthcare, Oil and gas, search, surveillance, finance and many other areas. There are uncountable applications and advantages of Big Data as some of them had been identified in this paper. It is to be noted that there are a lot of challenges facing the Big Data and in order to make optimal use of this modern discoveries, users must be quite aware of these challenges so as to providing a measurable adjustment or solutions to them as quick as possible.

REFERENCES

[1] Big Data (2013), "What is Big Data", www.thegovlabacademy.org/.../govt, retrieved 07/02/14.
[2] Brian Runciman(2013), "IT NOW Big Data Focus, AUTUMN 2013", www.bcs.org
[3] Carlos Castillo (2014), "Predicting the future with Big Data", www.alrazeera.com/.../predicting-future..." , retrieved 10/02/14.
[4] Chris Deptula(2013), "With all of the Big Data Tools, what is the right one for me", www.openbi.com/blogs/chris%20Deptula, retrieved 08/02/14.
[5] Divyakant Agrawal, UC Santa Barba, Philip Bernstein, Microsoft Elisa Bertino,…(2012), " Challenges and Opportunities with Big Data", "www.cra.org/ccc/../BigdataWhitepaper.pd..., retrieved 15/03/14.
[6] James M, Michael C, Brad B, Jacques B, Richard D, Charles R, Angela H.B(2011), "Big Data: The next frontier for Innovation, Competition, and Productivity", www.McKinsey.com, retrieved 23/02/14.
[7] Martin Hilbert.net(2013), "Growth of and Digitization of Global Information Storage Capacity", www.martinhilber.net/worldinfocapacity.html, retrieved 19/02/14.
[8] Mark Troester(2013), "Big Data Meets Big Data Analytics", www.sas.com/resources/.../ WR46345.pdf, retrieved 10/02/14.
[9] META GROUP (2001), "Controlling Data Volume, Velocity and Variety" , blogs.gartner.com/…/ ad949-3D-…, retrieved 03/03/14.
[10] Neil Raden (2012), " Big Data Analytics Architecture", www.teradata.com/Big-Data-Analytics", retrieved 15/03/14.

[11] Oracle (2013), "Information Management and Big Data: A Reference Architecture", www.oracle.com/.../info-mgmt-big-data-r..., retrieved 20/03/14.
[12] Rachel burnett (2011), "Publishing of confidential information", IT NOW 2011, BCS, www.bcs.org.
[13] Richard L. Villars, Carl W. Olofson and Mathew Eastwood (2011), "Big Data: What is it and why you should care", www.idc.com, retrieved 24/03/14.

AUTHORS' BIOGRAPHY

Oguntimilehin Abiodun is a Lecturer in the Department of Computer Science, College of Sciences, Afe Babalola University, Ado-Ekiti, Nigeria. He obtained B.Sc(Hons) Computer Science from University of Ado-Ekiti, Ado-Ekiti, Nigeria and Master's of Technology (M.Tech) from Federal University of Technology, Akure, Nigeria. He is a chartered member of Computer Professionals Registration Council of Nigeria (CPN) and a member of Nigeria Computer Society (NCS). His research interests are Machine Learning, Medical Informatics and Soft Computing. He has a number of publications in both reputable local and international journals. He is at present a Ph.D student of the Department of Computer Science, School of Sciences, Federal University of Technology, Akure , Nigeria. He can be reached by phone on +2348038601431 and through E-Mail ebenabiodun2@yahoo.com

Prof. Emmanuel Ojo Ademola is a Professor of Information & Communication Technology (ICT) Management, Provost, College of Sciences, Afe Babalola University, Ado-Ekiti, (ABUAD), Nigeria, and Subject Matter Expert/Consultant and Managing Editor of Checklists. He joined the University after working as an Institute Professor in ICT Management, Information Systems and Services for ICT research, and so particularly keen on promoting interaction between research and practice. His interests are in the interactions between people and the whole Information and Communication Technology ecology'; ICT resources and systems of all kinds. He works in areas of digital literacy, knowledge management, information and communication behaviour of individuals, and the history and philosophy of the ICT Management.He has been examiner and assessor for a leading United Kingdom (UK) awarding body spanning over eight years. As a subject Matter Expert/Consultant and Principal Assessor on ICT Management and Citizenship Matters, he has developed different curricula for various institutions and awarding boards. He is a Fellow of various Chartered Professional bodies of international reputation in their subject expert such as Computer Professionals Registration Council of Nigeria (CPN), British Computer Society (BCS) UK, Chartered Management Institute (UK), Institute for Learning (UK), etc. He has authored over 70 Peer-reviewed journal articles, checklists, and books of multidisciplinary titles. His Research and Consultancy Interest are ICT Management, Managing Mobile Services (Technologies and Business practices), Information management and services, Change Management, Information, Communication, Education Technology Systems. He can be reached by phone on +234 (0) 810 628 4000 and by E-mail on ademolaeo@abuad.edu.ng