International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Issue 2, Volume 2 (February 2015) www.ijirae.com

Big Data Security Issues and Challenges

Raghav Toshniwal*
Department of Computer Science
St. Xavier's College (Autonomous)
Kolkata,India

Kanishka Ghosh Dastidar
Department of Computer Science
St. Xavier's College (Autonomous)
Kolkata, India

Asoke Nath
Department of Computer Science
St. Xavier's College (Autonomous)
Kolkata, India

Abstract— The amount of data in world is growing day by day. Data is growing because of use of internet, smart phone and social network. Big data is a collection of data sets which is very large in size as well as complex. Generally size of the data is Petabyte and Exabyte. Traditional database systems is not able to capture, store and analyze this large amount of data. As the internet is growing, amount of big data continue to grow. Big data analytics provide new ways for businesses and government to analyze unstructured data. Now a days, Big data is one of the most talked topic in IT industry. It is going to play important role in future. Big data changes the way that data is managed and used. Some of the applications are in areas such as healthcare, traffic management, banking, retail, education and so on. Organizations are becoming more flexible and more open. New types of data will give new challenges as well. The present paper highlights important concepts of Big Data. In this write up we discuss various aspects of big data. We define Big Data and discuss the parameters along which Big Data is defined. This includes the three V's of big data which are velocity, volume and variety. The authors also look at processes involved in data processing and review the security aspects of Big Data and propose a new system for Security of Big Data and finally present the future scope of Big Data.

Keywords— Big data, Petabyte, Exabyte, Database, velocity, volume, variety

I. INTRODUCTION

The term Big Data is now used almost everywhere in our daily life. The term Big Data came around 2005 which refers to a wide range of large data sets almost impossible to manage and process using traditional data management tools – due to their size, but also their complexity. Big Data can be seen in the finance and business where enormous amount of stock exchange, banking, online and onsite purchasing data flows through computerized systems every day and are then captured and stored for inventory monitoring, customer behaviour and market behaviour. It can also be seen in the life sciences where big sets of data such as genome sequencing, clinical data and patient data are analysed and used to advance breakthroughs in science in research. Other areas of research where Big Data is of central importance are astronomy, oceanography, and engineering among many others. The leap in computational and storage power enables the collection, storage and analysis of these Big Data sets and companies introducing innovative technological solutions to Big Data analytics are flourishing. In this article, we explore the term Big Data as it emerged from the peer reviewed literature. As opposed to news items and social media articles, peer reviewed articles offer a glimpse into Big Data as a topic of study and the scientific problems methodologies and solutions that researchers are focusing on in relation to it. The purpose of this article, therefore, is to sketch the emergence of Big Data as a research topic from several points: (1) timeline, (2) geographic output, (3) disciplinary output, (4) types of published papers, and (5) thematic and conceptual development. The amount of data available to us is increasing in manifold with each passing moment. Data is generated in huge amounts all around us. Every digital process and social media exchange produces it. Systems, sensors and mobile devices transmit it. [1] With the advancement in technology, this data is being recorded and meaningful value is being extracted from it. Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information.
The 3Vs that define Big Data are Variety, Velocity and Volume.

1) Volume: There has been an exponential growth in the volume of data that is being dealt with. Data is not just in the form of text data, but also in the form of videos, music and large image files. Data is now stored in terms of Terabytes and even Petabytes in different enterprises. With the growth of the database, we need to re-evaluate the architecture and applications built to handle the data.
2) Velocity: 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.
3) Variety: Today, data 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. We need to find ways of governing, merging and managing these diverse forms of data.
   There are two other metrics of defining Big Data
4) Variability: Variability. In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data involved.[2]

5) Complexity: 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. A data environment can lie along the extremes on any one of the following parameters, or a combination of them, or even all of them together.

II. BIG DATA TECHNOLOGY: OPERATIONS VS. ANALYTICAL

The Big Data landscape can be divided into two main categories: Systems which provide operational capabilities for real time, transactional/interactive situations where data is captured and stored. The other type is systems that provide analysis capabilities for retrospective and complex analysis of the data that has been stored. This document is a template. An electronic copy can be downloaded from the Journal website. For questions on paper guidelines, please contact the journal publications committee as indicated on the journal website. Information about final paper submission is available from the conference website. The following table is a comparison between Operation and Analytical Systems in the field of Big Data.

TABLE I
Overview of Operational vs. Analytical Systems

	Operational	**Analytical**
Latency	1 ms - 100 ms	1 min - 100 min
Concurrency	1000 - 100,000	1 – 10
Access Pattern	Writes and Reads	Reads
Queries	Selective	Unselective
Data Scope	Operational	Retrospective
End User	Customer	Data Scientist
Technology	NoSQL	MapReduce, MPP Database

III. BIG DATA ANALYTICS

Big data analytics refers to the process of collecting, organizing and analysing large sets of data ("big data") to discover patterns and other useful information. With the help of Big Data analytics, organizations use the large amounts of data made available to them to identify patterns and extract useful information. Big Data analysis not only helps us to understand the information contained in the data but also identify the information that is most important to the organization and future decisions. The most important goal of Big Data Analytics is to enable organizations to make better decisions. Data Scientists, predictive modellers and other analytics professionals deal with huge amounts of transactional data and use Big Data Analytics to tap this data that may be untapped by other, conventional Business Intelligence programs.Big data can be analysed with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics, data mining, text analytics and statistical analysis. Due to the Volume and Velocity of Big Data, data warehouses are unable to handle the processing demands posed by data sets that are being updated in real time and continually, such as the movements on social media websites. The newer technologies involved in Big Data Analytics involve Hadoop and related tools such as YARN, MapReduce, Spark, Hive and Pig as well as NoSQL databases.

IV. STAGES INVOLVED IN BIG DATA

1. **Data Acquisition:** The first step in Big Data is acquiring the data itself. With the growing medium the rate of data generation is rising exponentially. With the introduction of smart devices which are used with a wide array of sensors continuously generate data. The Large Haudron Collider in Switzerland produces petabytes of data. Most of this data is not useful and can be discarded, however due to its unstructured form; selectively discarding the data presents a challenge. This data becomes more potent in nature when it's merged with other valuable data and superimposed. Due to the interconnectedness of devices over the World Wide Web, data is increasingly being collated and stored in the cloud.
2. **Data Extraction:** All of the data generated and acquired is not of use. It contains a large amount of redundant or unimportant data. For instance, a simple CCTV camera, constantly polls sensor to gather information of the user's movements. However, when the user is in a state of inactivity, the data generated by the activity sensor is redundant and of no use. The challenges presented in data extraction are twofold: firstly, due to nature of data generated, deciding which data to keep and which to discard increasingly depends on the context in which the

data was initially generated. For instance, footage of a security camera with the same frames may be discarded however it is important not to discard similar data in a case where it is being generated by a heart-rate sensor. Secondly, a lack of a common platform presents its own set of challenges. Due to wide variety of data that exists, bringing them under a common platform to standardize data extraction is a major challenge.

3. **Data Collation:** Data from a singular source often is not enough for analysis or prediction. More than one data sources are often combined to give a bigger picture to analyze. For example a health monitor application often collects data from the heart-rate sensor, pedometer, etc. to summarize the health information of the user. Likewise, weather prediction software take in data from many sources which reveal the daily humidity, temperature, precipitation, etc. In the scheme of Big Data convergence of data to form a bigger picture is often considered a very important part of processing.
4. **Data Structuring:** Once all the data is aggregated, it is very important to present and store data for further use in a structured format. The structuring is important so queries can be made on the data. Data structuring employs methods of organizing the data in a particular schema. Various new platforms, such as NoSQL, can query even on unstructured data and are being increasingly used for Big Data Analysis. A major issue with big data is providing real time results and therefore structuring of aggregated data needs to be done at a rapid pace.
5. **Data Visualization:** Once the data is structured, queries are made on the data and the data is presented in a visual format. Data Analysis involves targeting areas of interest and providing results based on the data that has been structured. For instance, data containing average temperatures are shown alongside water consumption rates to calculate a relation in between them. This analysis and presentation of data makes it ready for consumption for users. Raw data cannot be used to gain insights or for judging patterns, therefore "humanizing" the data becomes all the more important.

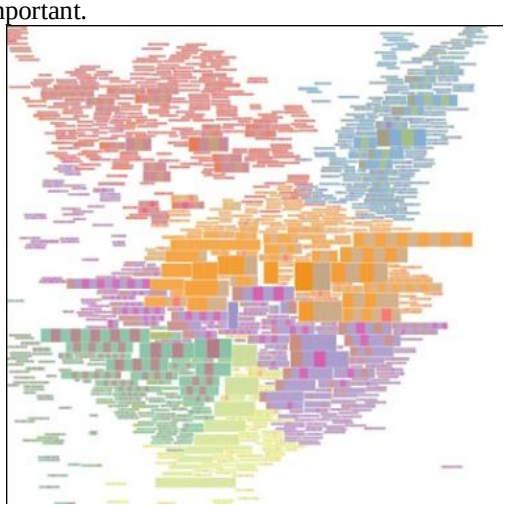

Fig 1: Big Data Visualization

6. **Data Interpretation:** The ultimate step in Big Data processing includes interpretation and gaining valuable information from the data that is processed. The information gained can be of two types: **Retrospective Analysis** includes gaining insights about events and actions that have already taken place. For instance, data about the television viewership for a show in different areas can help us judge the popularity of the show in those areas. **Prospective Analysis** includes judging patterns and discerning trends for future from data that is already been generated. Weather Prediction using big data analysis is an example of prospective analysis. Problems accruing from such interpretations pertain to fallacious and misleading trends being predicted. This is particularly dangerous due to an increasing reliance on data for key decisions. For example, if a particular symptom is plotted against the likelihood of being diagnosed with a particular disease, it might lead to misinformation about the symptom being caused due to the particular disease itself. Insights gained from data interpretation are therefore very important and the primary reason for processing big data as well.All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

V. SECURITY AND BIG DATA

The advent of Big Data has presented nee challenges in terms of Data Security. There is an increasing need of research in technologies that can handle the vast volume of Data and make it secure efficiently. Current Technologies for securing data are slow when applied to huge amounts of data.

TABLE II
ENCRYPTION RATES OF POPULAR ALGORITHMS

Algorithm	Megabytes(2^20 bytes) Processed	Time Taken	MB/Second
Blowfish	256	3.976	64.386
Rijndael (128-bit key)	256	4.196	61.010
Rijndael (192-bit key)	256	4.817	53.145
Rijndael (256-bit key)	256	5.308	48.229
Rijndael (128) CTR	256	4.436	57.710
Rijndael (128) OFB	256	4.837	52.925
Rijndael (128) CFB	256	5.378	47.601
Rijndael (128) CBC	256	4.617	55.447
DES	128	5.998	21.340
(3DES)DES-XEX3	128	6.159	20.783
(3DES)DES-EDE3	64	6.499	9.848

From the above table we can conclude that even the most efficient algorithms give an encryption rate of 64.3 MB/sec. However, in the light of Big Data where the amounts of data extend to a Gigabytes or even Petabytes, we can see a significant bottle neck for encrypting such large amounts of data. This is detrimental to the nature of Big Data which have real time processing and results. A need for a secure but faster encryption technique is increasingly required.

Another glaring challenge in Big Data is query processing on encrypted data. Currently, queries in both unstructured and structured encrypted data need decryption of the data first. Due to vast amounts of data this can take significant amounts of time and Query Processing can take significant time.
We now look into an alternative scheme of data encryption.

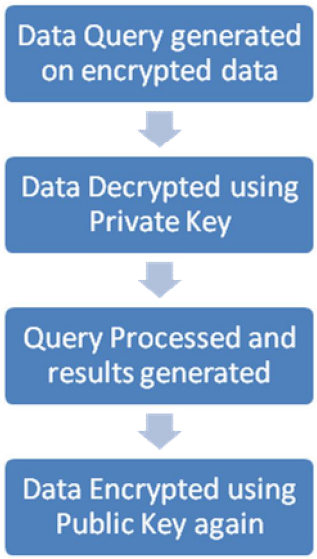

Fig I: CURRENT SYSTEM OF DATA ENCRYPTION
Figure I depicts how query on traditional encryption system allow Query Processing. The entire database needs to be encrypted to process any query. As discussed before, this takes significant amounts of time due to the slower rate of secure cryptographic techniques.

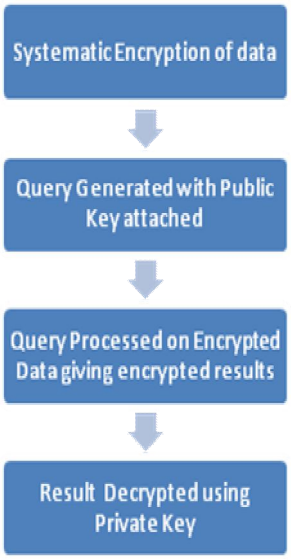

Fig II PROPOSED SYSTEM OF DATA ENCRYPTION

Figure II depicts the processed system of data encryption where the entire database is not encrypted as a whole but systematically encrypted to preserve some level of flexibility so as to allow query processing. This can be achieved by encrypting individual parts of data rather than the whole database.

Example:

TABLE III ORIGINAL DATABASE WITH TABLE 'ANIMAL'

Animal _ID	Animal_Name	Animal_Colour
0001	Goat	Grey
0002	Cat	Black

Each cell of the table is encrypted with the Public Key to result in the following table.

TABLE IV- ENCRYPTED DATABASE WITH TABLE 'ANIMAL'

Animal_ID	Animal_Name	Animal_Colour
#$@1	(@*4	&@($
@(#*	(*$	!&#

***Regular Query**: select * from Animal where Animal_Colour = 'Grey'*
'Grey' is encrypted using public key to '&@($'
***Processed Query**: select * from Animal where Animal_Colour = '&@($'*

Result:

Animal_ID	Animal_Name	Animal_Colour
#$@1	(@*4	&@($

Decryption using Private Key to result:

Animal_ID	Animal_Name	Animal_Colour
0001	Goat	Grey

One significant disadvantage of the proposed system is that security is compromised when individual elements are encrypted and the process takes longer to encrypt the first time than encrypting Data Base as a whole.

VI. CHALLENGES FACING BIG DATA

There are numerous challenges facing Big Data

1) The first challenge for organizations is to choose and select the relevant and important data. With such high volumes of data, it becomes important for organizations to able to separate the relevant data.

2) The second challenge is that even now, in organizations, many data points are not connected. This problem of connectivity is a severe hurdle. Big Data is all about collection of data from various transaction points. Organizations need to be able to manage data from across its enterprises. In order to address the growing volume of data created as a part of power grid operation, Siemens and Accenture recently formed a joint venture in the smart grid field to focus on solutions and services for system integration and data management. [3] These offerings will enable utilities to integrate operational technologies, such as real-time grid management, with information technologies like smart metering.
3) To leverage Big Data, one has to work across departments such as IT, Engineering and Finance. Thus the ownership and procurement of this data has to be a co-operative endeavour across these departments. This proves to be a significant organizational challenge.
4) There is a security angle related to Big Data collection. This is a major obstacle preventing companies from taking full advantage of Big Data Analysis.

Several issues will have to be addressed to capture the full potential of big data. Policies related to privacy, security, intellectual property, and even liability will need to be addressed in a big data world. Organizations need not only to put the right talent and technology in place but also structure workflows and incentives to optimize the use of big data. Access to data is critical—companies will increasingly need to integrate information from multiple data sources, often from third parties, and the incentives have to be in place to enable this.

VII. CONCLUSION AND FUTURE SCOPE

Big Data is changing the way we perceive our world. The impact big data has created and will continue to create can ripple through all facets of our life. Global Data is on the rise, by 2020, we would have quadrupled the data we generate every day. This data would be generated through a wide array of sensors we are continuously incorporating in our lives. Data collection would be aided by what is today dubbed as the "Internet of Things". Through the use of smart bulbs to smart cars, everyday devices are generating more data than ever before. These smart devices are incorporated not only with sensors to collect data all around them but they are also connected to the grid which contains other devices. A Smart Home today consists of an all encompassing architecture of devices that can interact with each other via the vast internet network. Bulbs that dim automatically aided by ambient light sensors and cars that can glide through heavy traffic using proximity sensors are examples of sensor technology advancements that we have seen over the years. Big Data is also changing things in the business world. Companies are using big data analysis to target marketing at very specific demographics. Focus Groups are becoming increasingly redundant as analytics firms such as McKinsey are using analysis on very large sample bases that have today been made possible due to advancements in Big Data. The potential value of global personal location data is estimated to be $700 billion to end users, and it can result in an up to 50% decrease in product development and assembly costs, according to a recent McKinsey report. Big Data does not arise out of a vacuum: it is recorded from some data generating source. For example, consider our ability to sense and observe the world around us, from the heart rate of an elderly citizen, and presence of toxins in the air we breathe, to the planned square kilometer array telescope, which will produce up to 1 million terabytes of raw data per day. Similarly, scientific experiments and simulations can easily produce petabytes of data today. Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude. There is immense scope in Big Data and a huge scope for research and Development.

ACKNOWLEDGMENT

The authors are very much grateful to Prof. Shalabh Agarwal, Head, Department of Computer Science, St. Xavier's College (Autonomous), Kolkata for his encouragement in research work in Big Data. One of the authors AN is also grateful to Fr. Dr. John Felix Raj, Principal of St. Xavier's College (Autonomous) for his constant support for doing research work with students and staff members.

REFERENCES

[1] Dona Sarkar, Asoke Nath, "Big Data – A Pilot Study on Scope and Challenges", International Journal of Advance Research in Computer Science and Management Studies (IJARCSMS, ISSN: 2371-7782), Volume 2, Issue 12, Dec 31, Page: 9-19(2014).
[2] http://www.cra.org/ccc/files/docs/init/bigdatawhitepaper.pdf
[3] http://www.nessi-europe.com/Files/Private/NESSI_WhitePaper_BigData.pdf
[4] http://sites.amd.com/sa/Documents/IDC_AMD_Big_Data_Whitepaper.pdf
[5] Sagiroglu, S.; Sinanc, D. ,"Big Data: A Review"
[6] Grosso, P. ; de Laat, C. ; Membrey, P.,(" Addressing big data issues in Scientific Data Infrastructure"
[7] Kogge, P.M.,(20-24 May,2013), "Big data, deep data, and the effect of system architectures on performance" Szczuka, Marcin,(24-28 June,2013)," How deep data becomes big data".
[8] META Group. "3D Data Management: Controlling Data Volume, Velocity, and Variety." February 2001. Performance Analysis of Data Encryption Algorithms: Abdel-Karim Al Tamimi

