International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 3, Issue 3, 2016, PP 1-11
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
DOI: http://dx.doi.org/10.20431/2349-4859.0303001
www.arcjournals.org

Big Data: Security Issues, Challenges and Future Scope

Kaustav Ghosh
Department of Computer Science
St. Xavier's College (Autonomous)
Kolkata, India
ghosh.kaustav88@yahoo.com

Asoke Nath
Department of Computer Science
St. Xavier's College (Autonomous)
Kolkata, India
asokejoy1@gmail.com

Abstract: *The work takes a look into the various properties that characterize big data, problems and issues related to big data handling and a broad view into the security and privacy issues related to big data analysis. The work ends with a glimpse into the future trends the world of big data is going to see.*

Keywords: *petabyte, zettabytes, veracity, valence, REST, HTTP basic access authentication, HTTP digest access authentication, injection attack, rollback attack, sybil attack*

1. INTRODUCTION

Big data is a collective term referring to data that is so **large** and **complex** that it exceeds the processing capability of conventional data management systems and software techniques. However with big data come big values. Data becomes **big data** when individual data stops mattering and only a large collection of it or analyses derived from it are of value. With many big data analyzing technologies, insights can be derived to enable better decision making for critical development areas such as health care, economic productivity, energy, and natural disaster prediction

The term **Big Data** appeared for the first time in **1998** in a **Silicon Graphics (SGI)** slide deck by John Mashey having the title **Big Data and the Next Wave of Infra Stress**. The first book mentioning **Big Data** is a data mining book that came to fore in **1998** too by **Weiss and Indrukya**. The **first academic paper** having the word **Big Data** in the title appeared in the year **2000** in a paper by **Diebold**.

The era of **Big Data** has bought with it a plethora of opportunities for the advancement of various branches of science, improvement of health care, promotion of economic growth, enhancement of education system and more ways of social interaction and entertainment. But as is said everything has its flip side as well, big data too has its issues. Security and privacy are great issues in big data due to its huge volume, high velocity, large variety like large scale cloud infrastructure, variety in data sources and formats, data acquisition of streaming data, inter cloud migration and others. The use of large scale cloud infrastructure having a varied number of software platforms across large networks of computers increases the region of attack to an all new level of the entire system.

The various challenges related to big data and cloud computing and its security and privacy issues and the reasons why they crop up are explained later in details.

Characteristics of Big Data:

Big Data possesses characteristics that can be defined by several V's

- **Volume:** The word **big** in big data is due to the sheer size of big data that it actually refers to. It refers to the vast amounts of data that is generated every second, minute, hour and day in our digitized world. It can come from large datasets being shared or many small data pieces collected over time. Every minute **204 emails** are sent, **200,000 photos** are uploaded **and 1.8 million likes** are generated on Facebook. On YouTube **1.3 million videos** are viewed and **72 hours of video** are uploaded. Its size is massive to the extent that they are measured by the likes of petabytes, exabytes and zettabytes. Some astounding examples of massive data generated (by machines) are:

- **CERN's large hadrons' collider** generates data of about **15 petabytes (2^50 bytes) a year**.
- Airbus **A380 engines.** Each has **4 engines** each and each generates **1 petabyte** of data on a flight from London to Singapore.
- **10,000 credit card transactions** are made per second.
- **1 million** customer transactions are made per second by Walmart.

According to predictions by an **IDC (International Data Corporation)** report sponsored by a big data company called EMC, digital data will grow by a **factor of 44** until the year **2020**, which is a growth of **0.8 zettabytes (2^80 bytes) [1]**. About 90% of world's data has been created in the last two years.

- **Variety:** Variety refers to the ever increasing different forms of data that can come in the form of texts, images, voices and geospatial data, computer generated simulations. The heterogeneity of data can be characterized along several dimensions. Some of these are:
  - **Structural variety**: It refers to the difference in the representation of the data. For example an EKG signal is very different from a newspaper article. Satellite images of wildfires from NASA are different from tweets sent out by people seeing the spread of fire.
  - **Media Variety**: Media variety refers to the medium in which the data gets delivered. For example: The audio of a speech and the transcript of a speech represent the same information in two different media.
  - **Semantic variety**: It comes from different assumptions of conditions on the data. Like conducting two income surveys on two different groups of people and not being able to compare or combine them without knowing more about the populations themselves.

In another way data can be **real time** like sensor data or **stored** like patient records.

A **single data object** or a **collection of similar data objects** may not be uniform in themselves. For example: An **email** is an **hybrid entity** where some information can be in the form of **tables** and the body may have **texts** in it with the text being itself **designed** or **decorated** around it. The email may contain **files** which may in turn be images, files and other multimedia objects **[1]**.

Data can be **structured**, **unstructured** or **semi structured**. **Structured data** has semantic meaning attached to it like data stored in database SQL. **Unstructured data** has no latent meaning. It includes calls, texts, tweets, browsing through various websites, and exchanging messages by every means possible, transaction made through cards for various payment issues. **Semi structured** data include **XML** and other **markup languages**, **email**,

- **Velocity:** Velocity refers to the **speed** at which big data is created or **moves** from **one point to another** and the increasing pace at which it needs to be stored and analyzed. The processing of data in **real time** to match its **production rate** as it gets generated is the main goal of big data analytics. It allows personalization of advertisement on web pages one visits based on recent search, viewing and purchase history. Thus we can put it this way, if a business cannot take advantage of the data as it gets generated and analyze it at speed, it is missing opportunities.

Accurate yet old information is useless. Taking an example of real life say we are on a road trip and need information about weather conditions to start packing. In this case newer the information the higher is the relevance in deciding what to pack. As weather conditions keep on changing so looking at last month's information or last year's won't help us much rather information from the current week or rather the present day will help us a great deal. Obtaining latest information about weather, processing it and letting it reach us helps us in our decision making.

Sensors and smart devices monitoring human body helps detect abnormalities in real time and aid us in taking action, saving our lives.

New information that is streaming often is needed to be integrated with existing data to produce decisions in case of emergencies like in case of a tornado. **[1]**

These **three characteristics** volume, variety and velocity are the **three main dimensions** that characterize big data and describe its challenges **[1]**. More V's are included in the big data community as new challenges are discovered and newer ways to define big data are obtained.

- **Veracity:** Veracity refers to the quality of big data. It refers to the biases, noise and abnormality of data. It also often refers to the immeasurable uncertainties and truthfulness and trustworthiness of data. It is very important for making big data operational. Data is useless if it is not accurate. The results of big data analysis are only as good as the data being analyzed. Data that are erroneous, duplicate and incomplete or outdated, as a whole are referred to as **dirty data. [1]**
- **Valence:** Valence refers to the **connectedness** of big data in the form of graphs just like atoms. Data items are often **directly connected** to one another like a city is connected to its country. Two Facebook users are connected as they are friends. An employee is connected to his workplace. **Indirect connection** of data items include two scientists being connected as they are computer scientists. Valence measures the **ratio** of actually connected data items to the possible number of connections that could occur within the collection. Data connectivity **increases over time** like in a conference some attending scientists meet other scientists from around the globe whom they did not know beforehand. A high valence data is denser **[1]**.

The last and final V of big data is **Value**.

- **Value:** Value refers to the fact how big data is going to benefit us and our organization. Data **value** helps in measuring the usefulness of data in decision making. Queries can be run on the stored data so as to deduce important results and gain insights from the filtered data so obtained so as to solve most analytically complex business problems**[1] [2]**.

Hadoop

Hadoop is a **free, Java-based** programming frame work that aids in the processing of large sets of data in a **distributed computing environment**. It is a part of the **Apache project** sponsored by the Apache Software Foundation. Hadoop cluster uses a **Master/Slave structure**. Using Hadoop, large data sets can be processed across a cluster of servers and applications can be run on systems with thousands of nodes involving thousands of terabytes. Distributed file system in Hadoop helps in rapid data transfer rates and allows the system to continue its normal operation even in the case of some node failures. This approach reduces the risk of an entire system failure, even in the case of a significant number of node failures. Hadoop enables a computing solution that is **scalable**, **cost effective**, **fault tolerant** and **flexible**. Hadoop Framework is used by popular companies like Google, Yahoo, Amazon and IBM etc., to support their applications involving huge amounts of data. Hadoop has two main sub projects namely MapReduce and **Hadoop Distributed File System** (HDFS) **[10]**.

MapReduce

Hadoop MapReduce is a framework used to write applications that process large amounts of data in parallel on clusters of commodity hardware resources in a **reliable**, **fault-tolerant** manner. A MapReduce first divides the data into individual chunks which in turn are processed by Map jobs in parallel. The outputs of the maps sorted by the framework are then input to the reduce tasks. Usually the input and the output of the job are both stored in a file-system. Scheduling, Monitoring and re-executing failed tasks are taken care of by the framework **[10]**.

Hadoop Distributed File System (HDFS)

HDFS is a file system that stretches over all the nodes in a Hadoop cluster for data storage. It links together file systems on local nodes to make it into one large file system. HDFS improves reliability by replicating data across multiple sources to overcome node failures **[10]**.

2. Issues and Challenges in Big Data

2.1. Big Data Issues and Challenges Related to Characteristics of Big Data:

- **Data volume:** When data volume is thought of the very first issue that occurs is **storage**. As data volume increases so the amount of space required to store data efficiently also increases. Not only has that but the huge volumes of data needs to be retrieved at a fast speed to extract results from them. Networking, bandwidth, cost of storing like in-house versus cloud storing are also other areas of concern **[1]**.

With the increase in volume of data the value of data records tend to decrease in proportion to age, type, richness and quality **[2]**. The advent of social networking sites have led to production of data of the order of terabytes every day. Such volumes of data are difficult to be handled using existing traditional databases **[2]**.

- **Data velocity:** Computer systems are creating more and more data, both **operational** and **analytical**;at increasing speeds and the number of consumers of that data are growing. People want all of the data and they want it as soon as possible leading to what is trending as **high-velocity data**. High velocity data can mean millions of rows of data per second.

  Traditional database systems are not capable enough of performing analytics on such volumes of data and that is constantly in motion. Data generated by both devices and actions of human beings like log files, website click-stream data like in E-commerce, twitter feeds can't be collected because the state of the art technology can't handle that data **[2]**.

- **Data variety:** Big data comes in many a form like messages, updates and images in social media sites, GPS signals from sensors and cell phones and a whole lot more. Many of these sources of big data are virtually new or rather as old as the networking sites themselves, like the information from social networks, Facebook, launched in 2004 and Twitter in 2006. Smart phones and other mobile devices can be bracketed in the same category. As these devices are ubiquitous, the traditional databases that store most corporate information until recently are found to be ill-suited to these data. Much of these data are unstructured and unwieldy and noisy which requires rigorous technique for decision making based on the data. Better algorithms to analyze them are an issue too **[5]**.

- **Data value:** Data are stored by different organizations to gain insights from them and use them for analytics for business intelligence. This storing produces a gap between the business leaders and the IT professionals. The business leaders are concerned with adding value to their business and obtaining profits from it. More the data more are the insights. This however doesn't go well with the IT professionals as they have to deal with the technicalities related to storing and processing the huge amounts of data **[2]**.

2.2. Big Data Management, Human Resource and Man Power Issues and Challenges:

Big data management deals with organization, administration and governance of large volumes of structured and unstructured data. It aims to ensure a high level of data quality and accessibility for business intelligence and big data analytics applications. Efficient data management helps companies, agencies and organizations in locating valuable information from large sets of data, of the order of terabytes and petabytes of unstructured or semi structured data. Sources may range from social media sites, system logs, call details and messages. There are however some challenges with big data and its management:

- Being new to big data and its management is the biggest challenge users of big data face. As organizations are new to big data it typically has **inadequate data analysts and IT professionals having the skills** to help interpret digital marketing data **[6]**.
- The sources of big data are varied with respect to size, format and method of collection. Digital data comes from many medium as comfortable to humans, like documents, drawings, pictures, sounds, video recordings, models and user interface designs, **with or without metadata** describing what the data is and its origin and how it was collected. Immaturity with these new data types and sources and inadequate data management infrastructure are a big problem. Hiring and training new consultants and progressing by virtue of learning are the only way out.
- The skill of a data analyst must not be limited to the **technical field**. It should be expanded to **research**, **analytical**, **interpretive** and **creative skills**. Along with the organizations that train for data scientist the universities too must include education about big data and data analysis to produce skilled and expert employees **[2]**.
- **IT investments are also lacking** like purchasing modern analytical tools to manage bigger data and analyze with better efficiency more complex data **[6]**.
- Due to **lack of governance or stewardship**, business sponsors and a compelling business case it is difficult for new projects to start **[7]**.

2.3. Big Data Technical Issues and Challenges:

- **Fault Tolerance**: With the advent of technologies like cloud computing the aim must remain such that whenever failure occurs the damage done must occur **within acceptable threshold** rather than the entire work requiring to be redone. **Fault-tolerant** computing is **tedious** and requires **extremely complex algorithms**. A foolproof, cent percent reliable fault tolerant machine or software is simply a far-fetched idea. To reduce the probability of failure to an acceptable level we can do:

  - **Divide the entire computation to be done into tasks** and assign these tasks to different nodes for computation.
  - **Keep a node as a supervising node** and look over all the other assigned nodes as to whether they are working properly or not. If a glitch occurs the particular task is restarted.

  There are however certain scenario where the entire computation can’t be divided into separate tasks as a **task can be recursive in nature** and requires the output of the previous computation to find the present result. These tasks can’t be restarted in case of an error. Here **checkpoints** are applied to keep the state of the system at certain intervals of time so that computation can restart from the last checkpoint so recorded **[2]**.

- **Data Heterogeneity**: 80% of data in today’s world are unstructured data. It encompass almost every kind of data we produce on a daily basis like social media interaction, document sharing, fax transfers, emails, messages and a lot more. Working with unstructured data is **inconvenient** and **expensive** too. **Converting** these to structured data is **unfeasible** as well **[2]**.

- **Data Quality**: As has been mentioned earlier, storage of big data is very expensive and there is always a tiff between business leaders and IT professionals regarding the amount of data the company or the organization is storing. The **quality of data** is an important factor to be looked into here. There is no point in storing very large data sets that are irrelevant as better result and conclusions can’t be drawn from them. Ensuring whether the amount of data is **enough** for a particular conclusion to be drawn or whether the data is **relevant** at all are further queries **[2]**.

- **Scalability**: The challenge in scalability of big data has led to **cloud computing**. It is capable of aggregating multiple different workloads with different performance goals into very large clusters. This needs high level of sharing of resources that is quite expensive and brings along with it various challenges like executing various jobs so that the goal of every workload is met successfully. It also has to deal with system failures in an efficient manner as it is quite common when working with large clusters.

  Hard disk drives being replaced by solid state drives and phase change technology do not have the same performance between sequential and random data transfer. The kind of storage device to be used is thus a large question looming around big data storage issue **[2]**.

2.4. Big Data Storage and Transport Issues and Challenges:

Big data processing issue has been well explained by the author of **[4]** by a very good example. Each time a new storage medium is invented the quantity of data becomes more and more. The capacity of current disks are about 4 terabytes per disk so 1 exabyte requires 25000 disks. Even if a single computer system is capable enough of processing 1 exabyte, to directly work with that many number of disks is well beyond its capacity.

Accessing this surge of data overwhelms current communication networks. If 1 gigabyte per second network has an effective sustainable transfer rate of 80%, its sustainable bandwidth is about 100 megabytes. This boils down to transferring 1 exabyte for 2800 hours, provided the sustainable transfer rate is maintained. This is actually transferring from the storage point to the processing point for a longer duration than actually processing it **[4]**.

2.5. Big Data Processing Issues and Challenges:

Effective processing of big data requires immense parallel processing and new analytics algorithms so as to provide rapid information. Often it may be unknown how to deal with a very large and varied volume of data and whether all of it needs to be analyzed. Challenges also include finding out data points that are really of importance and how to utilize the data to extract maximum benefit from it **[2]**.

2.6. Big Data Privacy and Security Issues and Challenges:

Often in big data analysis, the **personal information** of people from a database or from social networking sites need to be combined with external large data sets. Thus facts about anyone which might have been **confidential** become open to the world. Often it leads to taking insights in people's lives of which they are unaware of. Often it happens that a more educated person having better knowledge and concepts about big data analysis **takes advantage** of predictive analysis over a person who is less educated than him **[2]**.

3. Reasons for Security and Privacy Issues and Challenges in Big Data

Security and privacy are big concerns as far as big data are concerned and as big data grows by volume every day, every minute, every second, so are these concerns on the rise **[3]**.

- A prime reason for security and privacy concern in big data is because big data is now **widely accessible**. Data are shared on a large scale by scientists, doctors, business officials, government agencies and normal people. However the tools and technologies that have been developed till date to handle these huge volumes of data are not efficient enough to provide adequate security and privacy to data **[3]**.
- The technologies **lack enough security and privacy maintenance features** and the reason for this is because there is a **lack of basic understanding** about how to provide security to these huge volumes of data and sufficient **training is not provided** regarding how to provide security and privacy to these large scale data **[3]**.
- The data security and privacy maintenance regarding big data **lacks adequate policies** that ensure agreement with current approaches to security and privacy **[3]**.
- The present technologies have **weak security and privacy maintenance capability** so they are continuously being breached both accidentally and intentionally. Thus reassessing and updating current approaches to prevent data leakage has to be done on a continuous basis **[3]**.
- There is **lack of spending** on IT security to protect big data by the companies. About **10%** of a company's IT budget should be spent on security but **below 9%** is spent on an average thus making it tougher for themselves to protect their data.

4. Privacy and Security Issues and Challenges with Big Data

- **Secure Computations in Distributed Programming Frameworks:**

  Distributed programming frameworks use parallel computing and data storage for massive amounts of data. An example of this is MapReduce framework.

  As has been mentioned earlier MapReduce framework divides an input file into many chunks and then a **mapper** for each chunk reads the data, does computations and provides outputs in the form of **key/value pairs**. A reducer then combines the values belonging to each unique key and outputs the results. The main concerns here are: **securing the mappers** and **securing the data from a malicious mapper**.

  Mappers returning incorrect results are difficult to detect and it eventually results in incorrect aggregate outputs. With very large data sets malicious mappers are too hard to be detected as well and they eventually damage essential data. Mappers leaking, intentionally or unintentionally, private records are also an issue of concern.

  MapReduce computations are often subjected to **replay attack**, **man-in-the-middle attack** and **denial-of-service attack**. Rogue data nodes can be added to a cluster, and in turn receive replicated data or deliver altered MapReduce code. Creating snapshots of legitimate nodes and re-introducing altered copies is an easy attack in cloud and virtual environments and is difficult to detect **[8]**.

- **Security Best Practices for Non-Relational Data Stores:**

  Non relational databases used to store big data, mainly NoSQL databases, handle many challenges of big data analytics without concerning much over security issues. NoSQL databases consist of **security embedded in the middleware** and no explicit security enforcement is provided.

Transactional integrity maintenance is very lax in NoSQL databases. Complex integrity constrains can't be inculcated in NoSQL databases as it hampers with its functioning of providing better performance and scalability.

NoSQL databases have **weak authentication techniques** and **weak password storage mechanisms**. This exposes NoSQL to **replay attacks** and **password brute force attacks**, resulting in information leakage. NoSQL uses **HTTP Basic**- or **Digest**- based authentication that are prone to **replay** or **man-in-the-middle attack**. **REST** (Representational State Transfer) based on HTTP is prone to **cross-site scripting**, **cross-site request forgery** and **injection attacks**. Since NoSQL architecture employs **lightweight protocols** and **mechanisms** that are **loosely coupled**, it is susceptible to various injection attacks like: **JSON injection**, **array injection**, **view injection**, **REST injection**, **GQL** (Generalized Query Language) **injection**, **schema injection** and others. For example: An attacker can use **schema injection** to inject many columns in the database with data of the attacker's choice. This in turn may lead to database with corrupted data to Denial-of-Service attack resulting in total unavailability of the database. NoSQL is unsupportive of blocking with the help of **third party** as well. By maneuvering the REST-ful connection definition, it is possible to get access to the **handles** and **configuration parameters** of the underlying database, thereby gaining access to the file system. Some of the existing NoSQL databases offer authentication at the **local node level** however they fail to enforce authentication across all the cluster nodes

- ➢ **Cross-site scripting: Cross-site scripting** (**XSS**) is a type of computer security vulnerability typically found in web applications. XSS enables attackers to inject client-side scripts into web pages viewed by other users. A cross-site scripting vulnerability may be used by attackers to bypass access controls such as the same-origin policy **[12]**.
- ➢ **Cross-site request forgery: Cross-site request forgery**, also known as **one-click attack** or **session riding** (**CSRF** or **XSRF)** is a type of malicious exploit of a website where unauthorized commands are transmitted from a user that the website trusts. Unlike XSS, which exploits the trust a user has for a particular site, CSRF exploits the trust that a site has in a user's browser **[13]**.

Authorization techniques in one NoSQL solution differ from another. Most of the popular ones provide authorization at higher layers only without enforcing authorization on lower layers. It provides authorization on a per database level rather than at the level where the data are collected. There is no **role-based access control mechanism (RBAC)** built into the architecture because defining user roles and security groups with an RBAC mechanism is impossible.

NoSQL databases are subjected to **inside attacks** as well due to lenient security mechanisms. They may go unnoticed due to poor logging and log analysis methods along with other fundamental security mechanisms **[8]**.

- **Secure Data Storage And Transaction Logs:**

Data and transactions logs used to be kept in multi-tiered storage media. As data size grew scalability and accessibility became an issue hence **auto-tiering** for big data storage came to the fore. It doesn't keep track of where the data are stored unlike in previous multi-tiered storage media where IT managers knew which data resided where and when. This gave rise to many new challenges for data security storage.

Untrustworthy storage service providers often **search for clues**, from data transmission among tiers in a storage system, that help them correlate user activities and data sets and get to know certain properties, which can well prove vital to them. They however are not able to break into the data overcoming the encipherment but obtain certain useful properties of data. Due to the huge size of data it is quite infeasible to download the entire dataset to verify its availability and integrity

Auto-tiering places challenges on the service providers to guarantee **constant availability**. The weaker security at the lower tiers runs the risk of Denial-of-Service attacks. The **data owner** stores the cipher text in an auto-storage system and distributes the **public key** and **permission access** to each user; he gives the right to access data of certain portions to certain users. The **service provider** cannot interpret the data without the cipher key materials. However the service provider **may conspire** with users by exchanging the key and data hence he can obtain data to which he is not authorized to.

The service provider can instigate **roll-back attack** on users in case of a multi-user environment. He may serve outdated versions of data while the updated ones are already uploaded in the database.

Data tampering and data loss resulted by malicious users often **results in disputes** between the data storage provider or amongst users **[8]**.

- **End Point Input Validation/ Filtering:**

  Organizations collect data from a variety of sources including hardware devices, software applications and endpoint devices. As and when collecting these data, validation of the data as well as the source is a challenge.

  Often mischievous users tamper with the device from where the data are collected or tamper with the data collecting application installed in the device so that malicious data gets input into the central data collecting system.

  Fake IDs may be created by malicious users and provide malicious data as input into the central data collecting system. **ID cloning attacks** like **Sybil attacks** (**Sybil attack** in computer security is an attack wherein a reputation system is subverted by forging identities in peer-to-peer networks) are predominant in a **Bring Your Own Device (BYOD)** scenario where a malicious user brings his own device, faked as a trusted device and provides malicious input from there into the central data collecting system.

  Input sources of sensory data can be manipulated as well like artificially changing the temperature from a temperature sensor and inputting malicious input into the temperature collection process. GPS signals can be manipulated much the same way. The malicious user may change data while it is in transmission from a generous source to the central data collection system. It’s a **man-in-the middle attack** in a sense **[8]**.

- **Real-Time Security Monitoring:**

  Real-time security monitoring has been an ongoing challenge in the big data analysis scenario mainly due to the number of alerts generated by security devices. These alerts, may be co-related may be not, lead to many false positives and due to human being’s incapability to successfully deal with such a huge amount of them at such a speed, results in them being **clicked away** or ignored **[9]**.

  Security monitoring requires that the Big Data infrastructure or platform be inherently secure. Threats to a Big Data infrastructure include **rogue admin access** to applications or nodes, (**web**) **application threats**, and **eavesdropping** on the line. Infrastructure which is mostly an ecosystem of **different components**, the security of each component and the security integration of the components must be considered. In case of a **Hadoop cluster run in a public cloud** the s**ecurity of the public cloud**, itself being an ecosystem of components consisting of computing, storage and network components, needs to be considered. The security of the **Hadoop cluster**, the **security of the nodes**, the **interconnection among the nodes** and the **security of the data** stored in a node needs to be considered. The **security of the monitoring application** including applicable correlation rules that should follow secure coding principles, must be considered as well. The security of the **input source** (devices, sensors) from where the data comes from too must be taken into account **[8]**.

- **Scalable and Composable Privacy-Preserving Data Mining and Analytics:**

  Big data are subjected to **appropriation of privacy**, **invasive marketing**, **reduction of civil liberty** and **increase in state and corporate control**. User data collected by large organizations are constantly accessed by **inside analysts** and as well as **outside contractors** and **business partners**. A malicious insider or un-trusted partner can abuse these data sets and extract private information from customers.

  An insider of a company in charge of the big data store can **misuse his power** and **violate privacy policies**. For example: He can stalk people by monitoring through chats, if the company is a social networking one that facilitates chatting.

In case of a party owning the data outsources data analytics, an un-trusted partner can **infiltrate into private information** from users. This is also applicable in case of the usage of big data in the cloud as the cloud infrastructure, where the data are stored and processed, is usually not controlled by the owners of the data **[8]**.

- **Cryptographically Enforced Data-Centric Security:**

  There exist two fundamental approaches of controlling visibility of data to individuals, organizations and systems, the first one being restricting access to **underlying systems like operating systems or hypervisor**. The second is **encapsulating the data** itself in a protective shell by virtue of cryptography. The first approach or the **system-based approach** provides a larger attacking surface. There are many attacks like **buffer overflow** (in computer security and programming, a **buffer overflow**, or **buffer overrun**, is an anomaly where a program, while writing data to a buffer, overruns the buffer's boundary and overwrites adjacent memory locations) and **privilege escalation attack** (privilege escalation attack is a type of network intrusion that takes advantage of programming errors or design flaws to grant the attacker elevated access to the network and its associated data and applications) that bypass access control implementations and access the data. Protecting data **end-to-end by encryption** provides a much smaller well-defined attacking surface. It is vulnerable to **covert side-channel attack** (a **side-channel attack** is any attack based on information gained from the physical implementation of a cryptosystem, rather than brute force or theoretical weaknesses in the algorithms. For example, timing information, power consumption, electromagnetic leaks or even sound can provide an extra source of information, which can be exploited to break the system) and can extract secret keys.

  Various threats associated with cryptographically enforced access control method using encryption are: It should not be identifiable by the adversary, the corresponding plaintext data looking at the cipher text even if he has to choose between a correct and an incorrect plain text. For a **cryptographic protocol** facilitating searching and filtering encrypted data the adversary should not be able to learn anything about the encrypted data beyond the corresponding predicate, whether satisfied or not. The cryptographic protocol must also ensure that adversary must not be able to forge data that came from the claimed source for this may well be false hence affecting integrity of data **[8]**.

- **Granular audits:**

  Real-time security monitoring notification at the very moment an attack takes place is a real challenge. There may often be new attacks or missed true positives. In order to discover a missed attack audit information is required. Audit information from any device must be **complete** or rather it must give us details about what exactly happened and what went wrong. It must give **timely access**, so that it serves the purpose of compliance, regulation and forensic investigation. It must **not be tampered** and must be **accessible only in authorized areas [8]**.

5. FUTURE SCOPE AND DEVELOPMENT

As far as the future of big data is concerned it is for certain that **data volumes will continue to grow** and the prime reason for that would be the drastic increment in the number of hand held devices and internet connected devices, which is expected to grow in an exponential order. **SQL** will remain as the standard for data analysis and **Spark**, which is emerging, will emerge as the complimentary tool for data analysis. **Tools for analysis without the presence of an analyst** are set to take over, with **Microsoft** and **Salesforce** both recently are announcing features letting non-coders to create apps for viewing business data. As per IDC half of all **business analytics software will include intelligence where it is needed** by 2020. In other words it can be said that prescriptive analytics will be built into business software. Programs like **Kafka** and **Spark** will enable users to make decisions in real time. **Machine learning** will have a far bigger role to play for data preparation and predictive analysis in businesses in the coming days. Privacy and security challenges related to big data will grow and by 2018, **50% of business ethics violations will be related to data**. **Chief Data Officer** will be a common sight in companies in the recent future though it is thought that it won’t last long. **Autonomous agents and things** like **robots**, **autonomous vehicles**, **virtual personal assistant** and **smart devices** will be a huge trend in the future. Big data talent crunch as is seen these days will reduce in the coming days. **The International Institute for Analytics** predicts that companies will use recruiting and internal training to budding data scientists to get their own problems done.

Businesses will soon be able to **buy algorithms** rather than program them by themselves and add their own data to it. Existing services like **Algorithmia**, **DataXu**, **Kaggle** will grow in a large scale, that is **algorithm markets will emerge**. More companies will try to derive their revenue from their data. The **gap between insight and action in big data is going to reduce** and more energy will be given to obtaining insights and execution rather than collecting big data. **Fast and actionable data** will replace big data. Companies are expected to ask the right questions and make better use of the data they have, much of the big data they have are unused these days **[11]**.

6. Conclusion

To handle big data and to work with it and obtaining benefits from it a branch of science has come up and is evolving, called **Data Science**. Data Science is the branch of science that deals with discovering knowledge from huge sets of data, mostly unstructured and semi structured, by virtue of data inference and exploration. It's a revolution that's changing the world and finds application across various industries like **finance**, **retail**, **healthcare**, **manufacturing**, **sports** and **communication**. Search engine and digital marketing companies like **Google**, **Yahoo and Bing**, social networking companies like **Facebook**, **Twitter** and finance and E-commerce companies like **Amazon** and **EBay** are requiring and will require a lot of data scientists.

As far as security is concerned, the existing technologies are promising to evolve as newer vulnerabilities to big data arise and the need for securing them increases.

Acknowledgement

The authors are very much grateful to the Department of Computer Science, St. Xavier's College Autonomous (Kolkata) for giving the opportunity to do the research work in big data and its security and privacy challenges.

References

[1] www.coursera.org, Introduction to Big Data, University of California, San Diego.
https://www.coursera.org/learn/big-data-introduction

[2] **http://www.slideshare.net/HarshMishra3/harsh-big-data-seminar-report**.
Published: 4th January 2014 in Technology, Education
Harsh Kishore Mishra. Center for Computer Science and Technology. School of Engineering and Technology,Central University of Punjab, Bhatinda

[3] Schmitt, C., Shoffner, M., Owen P., Wang, X., Lamm, B., Mostafa, J., Barker, M.,
Krishnamurthy, A., Wilhelmsen, K., Ahalt, S., & Fecho, K.
(2013): Security and Privacy in the Era of Big Data: The SMW, a Technological Solution to the Challenge of Data Leakage. RENCI, University of North Carolina at Chapel Hill.
Text: **http://dx.doi.org/10.7921/G0WD3XHT**
Vol. 1, No. 2 in the RENCI White Paper Series, November 2013.
Created in collaboration with the **National Consortium for Data Science.**
(**www.data2discovery.org**).

[4] **Big Data: Issues and Challenges Moving Forward.**
2013 46th Hawaii International Conference on System Sciences
Stephen Kaisler, i_SW Corporation. Frank Armour, American University. J. Alberto Espinosa, American University. William Money, George Washington University

[5] **Big Data: The Management Revolution.**
Andrew McAfee and Erik Brynjolfsson
October 2012. Harvard Business Review

[6] **http://www.dataversity.net/common-big-data-management-issues-solutions/**
The Most Common Big Data Management Issues (And Their Solutions). By: A.R. Guess. July 15 2014.
Last visited: 10th July 2016.

[7] **TDWI Research**.
TDWI Best Practices Report. **Managing Big Data**.
Fourth Quarter 2013. By Philip Russom

[8] **Expanded Top Ten Big Data Security and Privacy Challenges**
Big Data Working Group. April 2013.
© 2013 Cloud Security Alliance – All Rights Reserved

[9] **Challenges and Security Issues in Big Data Analysis.**
Reena Singh. Kunver Arif Ali.
IJIRSET. Volume: 5. Issue: 1. January 2016.

[10] **Security Issues Associated With Big Data in Cloud Computing**.
K.Iswarya Assistant Professor, Department of Computer Science. Idhaya College for Women, Kumbakonam, India.
SSRG International Journal of Computer Science and Engineering (SSRG - IJCSE) – volume1 issue 8 October 2014

[11] **http://www.forbes.com/sites/bernardmarr/2016/03/15/17-predictions-about-the-future-of-big-data- everyone-should-read/#2e45417d157c**
Mar 15, 2016 @ 04:04 AM. Contributor: Bernard Marr.

[12] **https://en.wikipedia.org/wiki/Cross-site_scripting**
Last accessed: 10th July 2016.

[13] **https://en.wikipedia.org/wiki/Cross-site_request_forgery**
Last accessed: 10th July 2016.

AUTHORS' BIOGRAPHY

Kaustav Ghosh, passed M.Sc. in Computer Science from St. Xavier's College (Autonomous), Kolkata in 2016. He has already published papers on Cognitive Radio in International Journals. Currently he is doing research work in Big Data Analytics, data science.

Dr. Asoke Nath, is an Associate Professor in the Department of Computer Science, St. Xavier's College (Autonomous) Kolkata. Apart from his teaching assignment he is involved in various research fields such as Cryptography and Network Security, Visual Cryptography, Steganography, Image Processing, Mathematical Modeling of Social Networks, Li Fi technology, Big Data analytics, Cognitive Radio, Data Science, e learning, MOOCs and so on. He has published more than 191 publications in Journals and conference proceedings. He is the life member of MIR Labs (USA) and CSI Kolkata Chapter.