Big Data: Related Technologies, Security Challenges, and Research Opportunities

Azman Samsudin
School of Computer Sciences, Universiti Sains Malaysia, 11800 Penang, Malaysia
azman.samsudin@usm.my

Abstract
Big Data is generally defined in association with the “Four V’s”: high volume, high velocity, high variety and high veracity. Millions of organizations worldwide are embarking into Big Data projects recently, in order to better support their R&D activities, decision making process, creating marketing strategies and customer relationship management. However, Big Data comes with challenges. This paper reviews the security challenges of Big Data and investigates whether existing data protection mechanisms can be further adapted to protect Big Data processing. This paper also looks into related technologies such as Hadoop, Cloud Computing and Internet of Things (IoTs), as well explore research opportunities in securing Big Data processing.

1.0 Introduction
With the advent of computing technologies, huge amount of data is being collected from diverse end-point, transferred through distributed networks, processed and analysed in an unprecedented rate, in which, such phenomenon is now known as “Big Data”. The Big Data paradigm was initially introduced to handle extensive scale of datasets during 1980s. Accordingly, in the early stage of Big Data development, both academia and industry had tried to define and characterize the explosive nature of the generated data [1], [2]. Some of the definition are listed below.

Oracle Definition: "Big Data typically refers to the following types of data: (i) Traditional enterprise data – includes customer information from CRM systems, transactional ERP data, web store transactions, and general ledger data; (ii) Machine-generated /sensor data – includes Call Detail Records (CDR), weblogs, smart meters, manufacturing sensors, equipment logs (often referred to as digital exhaust), trading systems data; (iii) Social data – includes customer feedback streams, micro-blogging sites like Twitter, social media platforms like Facebook." [3]

Forrester Definition: "Measured in terms of volume, velocity, and variety, Big Data represents a major disruption in the business intelligence and data management landscape, upending fundamental notions about governance and IT delivery. With traditional solutions becoming too expensive to scale or adapt to rapidly evolving conditions, companies are scrambling to find affordable technologies that will help them store, process, and query all of their data. Innovative solutions will enable companies to extract maximum value from Big Data and create differentiated, more personal customer experiences." [4]
EMC/IDC Definition: "Big Data technologies describes a new generation of technologies and architectures, designed to economically extract value from very large volume of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis." [5]

Microsoft Definition: “Big data is the term increasingly used to describe the process of applying serious computing power – the latest in machine learning and artificial intelligence – to seriously massive and often highly complex sets of information.” [2], [6]

Gartner Definition: “Big Data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.” [7]

NIST Big Data Public Working Group Definition: "Big Data refers to digital data volume, velocity and/or variety (veracity) that: i) Enable novel approaches to frontier questions previously inaccessible or impractical using current or conventional methods; and/or; ii) Exceed the capacity or capability of current or conventional methods and systems; iii) Differentiates by storing and analyzing population data and not sample sizes.” [8]

Clearly, these definitions were derived from different business interests and related supported technologies. For instance, Oracle’ definition was concerned on the data types, sources of Big Data (e.g. sensor, social data, etc.) and its underlying data processing infrastructure. Microsoft relates the Big Data with the machine learning and artificial intelligence technologies. Meanwhile, EMC/IDC, Forrester, Gartner and NIST Big Data Public Working Group have defined Big Data in the association with the three ‘V’ (3Vs) characteristics: volume, velocity and variety. IBM [9], Gantz and Reinsel [10] argued that the scope of Big Data is not limited only to 3Vs, but also encompasses the meaning of Big Data in which they refer it as the fourth ‘V’ – veracity. Subsequently, some researchers extended the definition of Big Data to include the 5th ‘V’ which is ‘value’.

Conclusively, Big Data consists of extensive datasets, primary associated with the 5Vs: high volume, velocity, variety, veracity and value, in which resulting traditional data processing mechanisms are unable to adapt to. The 5Vs characteristics are described as follows.

- **Volume.** High Volume of Big Data refers to a large scale of data, in terms of its amount or quantity. Generally, it involves processing a minimal terabytes of data weekly. For instance, sensors in Aviation industry that constantly capture engine and aircraft health information during a flight, a multiple gigabytes of data can be generated for a single flight [11].

- **Velocity.** While some researchers described velocity as the speed of data being generated, which sounds confusing with the term of volume. This paper refers velocity as the speed of data being processed, starting from data collecting, data pre-processing (e.g. cleaning, normalization), data processing (e.g. batching, streaming, real-time or near real-time analytical) and data storing.

- **Variety.** High variety refers to a high diversity of data format and structures. These includes structure data, semi-structured data, “quasi-structured” data and
unstructured data. Structure data is a data consists of a defined data type, format and structure that can be processed and stored in conventional database management system (e.g. MSSQL, DB2, Oracle), data warehouse and enterprises management systems. Semi-structured data is a data that always being associated with the discernible, pre-defined or standardized pattern in order to facilitate the message parsing and communication, such as XML and JSON files, and can be stored in NoSQL databases. “Quasi-structured” data are data with erratic data formats which most of the time contains inconsistency or redundancy, such as web clickstreams. Unstructured data are data that come without any inherent data structure and generally stored as a different data files such as video, pdf files, images, and social data (e.g. web chats, Facebook, etc.).

- **Veracity.** Veracity (also called as verification or variability) refers to uncertainty characteristic of the data such as data consistency, trustworthiness, availability, accountability, etc.

- **Value.** Value refers to an added-value that the collected Big Data can bring to the intended process, activity or predictive analysis.

Today, more and more organizations, either businesses, non-profit organizations or governments are stepping into Big Data Projects as a transformative force to maximize their business profits, optimize their business process efficiency and sustainability, such as predicting the market shares or sales demands, support decision making and R&D activities. According to IDC, the worldwide Big Data and its business analytics will increased more than 50% percent over the five-year period, grow from USD 122 billion in 2015 to more than USD 187 million in 2019 [12] in order to support a wide range of industry applications such as manufacturing, logistic and transportation, medical and health care, etc. The Big Data relationship and its major industries are further demonstrated in Table 1.

Table 1. Big Data Characteristic and Its Potential in Major Industries [13]

<table>
<thead>
<tr>
<th>Industry</th>
<th>Volume</th>
<th>Velocity</th>
<th>Variety</th>
<th>Veracity</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical and Health Care</td>
<td>Very High</td>
<td>Very High</td>
<td>High</td>
<td>Very High</td>
<td>High</td>
</tr>
<tr>
<td>Logistic and Transportation</td>
<td>Very High</td>
<td>Very High</td>
<td>Moderate</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Manufacturing and Natural Resources</td>
<td>Very High</td>
<td>Very High</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Banking and Securities</td>
<td>Very High</td>
<td>Very High</td>
<td>Low</td>
<td>Very High</td>
<td>High</td>
</tr>
<tr>
<td>Education</td>
<td>Very Low</td>
<td>Very Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Government</td>
<td>Very High</td>
<td>High</td>
<td>Very High</td>
<td>Very High</td>
<td>Very High</td>
</tr>
<tr>
<td>Retail</td>
<td>High</td>
<td>High</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td>Wholesale Trade</td>
<td>Low</td>
<td>Low</td>
<td>Very Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Insurance</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td>Communication, Media Service</td>
<td>Very High</td>
<td>High</td>
<td>Very High</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Retail</td>
<td>High</td>
<td>High</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

While Big Data promises a transformative future for individuals, business and governments, it also introduces a number of security challenges and privacy risks, especially when involves a collection of personal data. This paper provides a comprehensive study of Big
Data — Its definition, characteristic, related technologies, security challenges and future research opportunities. The rest of this paper is organized as follows. Section 2, discusses the Big Data and its related technologies which includes Hadoop, cloud computing and Internet of Things (IoT). Section 3, discusses the Big Data architecture and its related technologies such as cloud computing and Internet of Things (IoTs). Section 4, elaborates the security challenges and how Big Data different from Traditional Data Protection. The discussion on future trends and research opportunities can be found in Section 5. Lastly, Section 6 the conclusion.

2.0 Big Data and Its Related Technologies

Big Data depends on diverse technologies in order to support its data life cycle, which includes data collection, data curation, data processing and storage. The emergence of Big Data, does converge with other related technologies such as Hadoop, cloud computing and Internet of Things (IoT). This section examines and discusses such technologies in relation to Big Data.

i. **Big Data Vs Hadoop.** Hadoop was initially developed by Doug Cutting and Mile Cafarella [14], [15] for processing a massive amount of public web data. Thereafter, Hadoop become a popular platform for analysing and processing data in private networks. Generally, Hadoop is an open source framework that capable to process an extensive dataset distributary with its processing module which is known as MapReduce and storage module — Hadoop Distributed File System (HFDS). Since Hadoop can handle high volume of data in high velocity, Hadoop has become popular as a platform for Big Data processing. The flexibility of Hadoop platforms to work with other software tools such as Apache Pig, Hive, HBase, Phoneix, Spark, Zookeeper, Clouder Impala, Flume, Oozie, Storm, etc., enables it to serve as a comprehensive technology to support Big Data processing activities.

ii. **Cloud Computing.** The recent advancement of cloud technologies promises a cost-effective, scalable and easier maintenance data solution for individuals, businesses and governments. The cloud computing enables a ubiquitous and on-demand access to a shared processing resources and data storage with a diversity of service models, includes Software as a Service (SaaS), Platform as a Service (Paas) and Infrastructure as a Service (Iaas) model. The cloud computing can be served as a technology that converging with the realization of Big Data by facilitate the data collection process, in which cloud technologies (e.g. MyRobots, WoTkit, Xively, etc.) provides a universal platform to integrate the collection of data from heterogeneous networks and diverse devices. Alternatively, cloud computing also offers a data processing tools such cloud-based Hadoop and application API to facilitate the data management and application development process.

iii. **Internet of Things (IoT).** Internet of Things (IoT) is an emerging network paradigm, that consists of billions sensor nodes, diverse end-point devices, from the tiniest of ultra-efficient connected nodes to a high performance gateway, which are intelligently connected via a unique addressing scheme, collaborated to collect, and exchange data. In most cases, these “things” are heterogeneous entities that located in the distributed networks [16]. When there is a massive number of “things” such as sensors in Aviation industry, that constantly capture engine and aircraft health information during a flight, transfer via several communication networks; or diverse “things” in smart cities applications such as sensors, mobile devices, camera, smart grid etc., collaborate to
generate and exchange data continuously, these generated data are always comes with the Big Data characteristics, in which they are high volume, exists in different formats and structures and involves a high speed analytical or nearly real-time analytics. Therefore, Big Data and IoTs are converging and complement to each other. IoT technologies offers automated mechanisms such as M2M and underlying architecture to capture, collect and transfer the machine data into data warehouse. Meanwhile, Big Data technologies provides a data processing platform such as Hadoop in data warehouse to curate, process and analysis these data, and turn it into a valuable information.

### 3.0 Big Data Architecture and Its Related Technologies

Since Big Data is relatively a new phenomenon in computing, therefore there is virtually no properly defined architecture for Big Data. Existing architectures are either based on industry Big Data solutions, such as Oracle[17], Amazon AWS Big Data architecture [18] and IBM reference architecture[19] or alternatively an architecture that focuses on a specific Big Data application scenario. For instance, Zhang et al. [20] proposed Big Data architecture that is based on Product Lifecycle Management (PLM). Instead of focusing on specific scenario or supported technologies, Zhang proposed to develop the Big Data architecture and its security framework as an abstraction, with bottom-up approach as the core nature of Big Data that sits on top of diverse technologies and heterogeneous networks. Meanwhile, NIST Big Data Public Working Group (NBD-PWG) [21] has examined various publicly available Big Data architectures (e.g. IBM, Oracle, SAP, 9sight, LexisNexis, etc.) and subsequently proposed a NIST Big Data Reference Architecture (see Fig. 1) that is applicable to a diverse business environments, from tightly integrated enterprise systems to loosely coupled vertical industries.

![NIST Big Data Reference Architecture](image)

**Fig. 1. NIST Big Data Reference Architecture**
As illustrated in Fig. 1, the Big Data value chain is represented by the Information Value Chain (horizontal axis) and IT Value Chain (Vertical axis). The Information Value chain focuses on the values generated from Big Data processing processes, which includes data collection, curation, analysis and applications. The IT Value Chain concerns on the technical aspect to support Big Data processes and applications, which includes network infrastructure, platforms, application tools and other IT services (e.g. hosting). The NIST Big Data Reference Architecture consists of seven functional components as described in the following:

- **Data Provider.** Data Provider is responsible to generate, collect and transfer data or information into the Big Data system via diverse communication protocols or network technologies. Besides that, Data Provider also exposes a collection of services (interfaces) in order to support data discovery and data access model, either using a push from the data provider or a pull by the Big Data application model. The data providers can be tiniest sensor nodes, mobile devices, embedded systems to high performance computer systems.

- **Data Consumer.** Data consumers are end users or other systems that use the deliverable results from the data processing by submitting their desired tasks or operations to Big Data Application Provider, such as search and retrieve, visualization, download, analysis task, report or summarization, etc.

- **Big Data Application Provider.** Big Data Application Provider is responsible to execute a set of operation or tasks along with the data life cycle (e.g. data collection, preparation, analytics, visualization, access, etc.) which satisfy the System Orchestrator-defined requirements, as well as security and privacy requirements.

- **Big Data Framework Provider.** Big Data Framework Provider is responsible to establish a computing framework for executing certain transformation applications while providing data privacy and integrity protection. These computing framework includes infrastructure framework, data platform framework and process framework that underlying on various network, storage and data processing technologies. These technologies are designed to fulfil Big Data requirements while maintaining linear or near-linear performance.

- **System Orchestrator.** System Orchestrator is responsible to define, configure, integrate and manage the required data application activities into an operational vertical system. These includes monitoring and managing quality of service requirements, workflow requirements, business requirements, agreements, policies or constraints among the tightly coupled enterprises.

- **System Management.** System management is responsible to manage Big Data architecture, which includes provisioning, configuration, package management, software management, backup management, capability management, resources management, and performance management, in order to support Big Data environments.

- **Security and Privacy Management.** Security and Privacy Management is responsible to manage the security and privacy issues of data, systems, networks, technologies, frameworks, and applications, while maintaining a high level data quality and secure
accessibility. It encompasses all functional components, which includes Data Provider, Data Consumer, Big Data Application Provider, Big Data Framework Provider and System Orchestrator.

4.0 Security Challenges and Privacy Concerns

In the past, Big Data was limited to large-scale organizations such as government and large enterprises that could afford to establish their own infrastructure to process Big Data. These infrastructures and technologies were typically proprietary and totally isolated from public networks, therefore security and privacy issues is not a concerns. Today, more and more small to medium-sized organizations are investing on Big Data projects to support their decision making, research and development activities, in the hope to gain some insight views and subsequently achieving competitive advantages in market share. According to IDG Big Data report [22] a typical organization is processing an average of 164 terabytes data. Over the next 12 to 18 months, that number is expected to increase 76 percent, to 289 terabytes.

With a rapid growth of data, whether the traditional security and data protection still adequate to be adapted into Big Data architecture and environment is questionable. The security and privacy issues of Big Data significantly magnified by its 3Vs characteristic: Volume, Variety and Velocity. The differences between traditional data and Big Data is further summarized in Table 2.

Table 2: The differences between Big Data and traditional data [1]

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Traditional Data</th>
<th>Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume</td>
<td>Bytes - megabytes</td>
<td>High Volume (Terabytes- Zettabytes)</td>
</tr>
<tr>
<td>Velocity</td>
<td>Moderate Speed</td>
<td>High Speed</td>
</tr>
<tr>
<td>Variety</td>
<td>Structured Data</td>
<td>Structure, Semi-Structured, Unstructured, “Quasi” Structured</td>
</tr>
<tr>
<td>Environments</td>
<td>Homogenous</td>
<td>Heterogeneous</td>
</tr>
<tr>
<td>Data Storage</td>
<td>Data are collected and stored in application/service providers</td>
<td>High Scalability Data are stored at data owner/data producer, only the aggregated resulted will stored at application/service providers.</td>
</tr>
<tr>
<td>Data Redundancy</td>
<td>High Redundancy</td>
<td>Low Redundancy</td>
</tr>
<tr>
<td>Security and Privacy Protection</td>
<td>Data at Rest</td>
<td>Data at Rest</td>
</tr>
<tr>
<td></td>
<td>Data in Memory</td>
<td>Data in Memory</td>
</tr>
<tr>
<td></td>
<td>Data in Transit</td>
<td>Data in Transit</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data in Transform</td>
</tr>
</tbody>
</table>

Next, several security and privacy challenges of Big Data are further discussed as follows.

- **Privacy Concerns on Big Data Collections and Usage.** Privacy threats generally refers to convey private, confidential or sensitive information through a public medium and in the process of disclosing it to an unwanted or unauthorized audience, without the data owner awareness or permission [23]. As the emergence of Big Data and its confluence technologies such as IoT, data can be collected and aggregated at anytime,
anywhere and anyplace, which poses a huge privacy threats to public. For instance, security surveillance camera technology can be used to determine and record person location with a time and geolocation. MAC address of computer or embedded systems can be used to locate and track person geolocation. Similarly, personal health data that collected from wearable devices can be further abuse for healthy product advertisement, etc. The privacy issue is becoming critically challenging when there is a cross-border data aggregation and transmission that involving different privacy requirements and legislation of countries.

- **Interoperability and Scalability Issues of Existing Security Mechanisms.** As Big Data security and privacy challenges are being magnified, so as its consequences. Traditional security mechanisms that tailored to secure a small scale and static data often fall short. The proposal for Big Data security framework should address interoperability and scalability issues. For instance, the security framework should support the communication and exchange of data among the different communication technologies that employ different communication protocols or data encryption algorithms, as well as enhancing the scalability of the Public Key Infrastructure (PKI) and the key management protocols.

- **Limitation of Conventional Point-to-Point Defences.** As conventional security mechanisms focus on end-to-end secure communication that operates on the basis of point-to-point defences (e.g. TLS/SSL, IPSec, etc.), these mechanisms are less efficient against the new cyber Advanced Persistent Threats (APT) attacks. The malicious intention can be targeted to any vulnerabilities or weak points of Big Data processing, networks or application system. For instance, multi-hop wireless broadcast communication is vulnerable to eavesdropping. These situations become worst with the Bring Your Own Device (BYOD) or Bring Your Own Technology (BYOT) environments. The hijacked or backdoor installed devices can be used to penetrate into Big Data networks.

- **Security of Big Data Processing Platform — Hadoop.** Hadoop was initially designed without any security countermeasure in mind as it was originally targeted for private usage such as government and large enterprises. As Hadoop became a popular platform for data analytics and processing for the masses, the increased security incidents (e.g. deleting massive amounts of data within seconds with a distributed delete, mischievous user could lower the priorities of other Hadoop jobs in order to make his job complete faster – or worse, kill the other jobs, etc.) has raised concerns on it security issues. Subsequently a team of Yahoo researchers had decided to integrate Kerberos authentication mechanism into Hadoop. However, Kerberos authentication mechanism only provides a limited access control and authorization capabilities (based on user and group permission Access Control Lists). Moreover, data in HDFS are stored without any encryption.

### 5.0 Future Research Trends

The inherent complexities of Big Data architecture and processing environment make its security and privacy issues becoming highly challenging. In 2013, Cloud Security Alliance has
identified the top ten security and privacy challenges of Big Data [24], which categorized into four categories as follow:

- **Infrastructure Security:**
  - Issue 1 – Secure computations in distributed programming framework
  - Issue 2 – Security best practices for non-relational data stores

- **Data Privacy**
  - Issue 3 – Privacy Preserving Data Mining and Analytics
  - Issue 4 – Cryptographically Enforced Data Centric Security
  - Issue 5 – Granular Access Control

- **Data Management:**
  - Issue 6 – Secure Data Storage and Transaction Logs
  - Issue 7 – Granular Audits
  - Issue 8 – Data Provenance

- **Integrity and Reactive security**
  - Issue 9 – End-point validation and filtering
  - Issue 10 – Real-time Security Monitoring

As Big Data engineering focuses on the “moving the processing to data, not the data to processing” [1], [11], the future trends of securing Big Data are directed toward a data centric approach, that is securing the data itself – wherever it goes (data-in-transit), whenever it exists (data-at-rest) and whatever it be (data-in-transform). These data centric approaches are focuses on providing end-to-end data protection, rather than end-to-end secure communication and point-to-point defences. Several data centric approaches that capable to ensure the security of data-in-transit, data-at-rest and data-in-transform are further discussed in the following.

- **Searchable Encryption.** Searchable Encryption (SE) scheme is a cryptographic primitive that offers a secure search functions (e.g. keyword search, ranking search, interval search, and subset search) over an encrypted data, thus preserving the data privacy [1], [25]. The recent research concerns on searchable encryption algorithm are concerns mainly on three issues: scope of query support (e.g. single query or multiple queries, adaptive query or non-adaptive query), data privacy (protection on query result or submitted query), application scenario (single user or multi-user environments with an access controlled by data owner). To be adapted into Big Data environments, the scalability of the algorithm to support multi-user environments (e.g. revocation of search privilege, lightweight and high efficient algorithm, or multi-owner data scenario) and the interoperable issues (e.g. able to search across the data that encrypted under different data encryption algorithm) are highly sought after.

- **Homomorphic Encryption.** Homomorphic Encryption (HE) scheme is a cryptographic primitive that works similar as conventional data encryption algorithm, however, equipped with the additional capability to conduct computation (e.g. addition, multiplication, predictive analysis, etc.) over encrypted data. However, the recent work of HE scheme still suffers from the performance issues, such as slow speed and huge generated ciphertext and public key in Fully Homomorphic Encryption (FHE), while practical Partial Homomorphic Encryption (PHE) only capable to handle either addition or multiplication operations (e.g. RSA for multiplication and Pailler scheme for
addition operations). A highly efficient FHE scheme therefore has becoming a holy grail in cryptography research recently.

- **Attribute-Based Encryption.** Attribute Based Encryption (ABE) scheme is a cryptographic primitive that works similarly as the conventional data encryption algorithm, however, with the ability to have a fine-grained control access over the encrypted data. The access control policies are embedded directly into user’s private key (so called as Key Policy-Attributed Based Encryption, KP-ABE) or encrypted data (so called as Ciphertext Policy-Attribute Based Encryption, CP-ABE), therefore serves a promise to solve the long standing open problem of Public Key Infrastructure (PKI) inefficiency [26]. With its advantages over the conventional PKI, which includes the decentralized fine-grain access control in large dynamic networks and ubiquitous computing environments, the scalable key management and flexible data distribution, ABE scheme can be a better solution to fix the scalability and interoperability issues of Big Data environments. While existing ABE scheme are mainly constructed based on the bilinear map over elliptic curves approach that involves expensive bilinear pairing operations, it is very difficult to implement it into the Big Data processing platform. Therefore, a more efficient ABE scheme with low computation costs is an interesting topic to be explored.

### 6.0 Conclusions

The security and privacy concerns on Big Data is significantly becoming more critical as its inherent complexities of architecture and technologies. This paper provides a tour on Big Data and its related technologies, security challenges and future research trends. Firstly, this paper discussed the Big Data definitions and its associated 5Vs characteristics: Volume, Velocity, Variety, Veracity and Value. Subsequently, the Big Data relationship and its major industries are presented. The emergence of Big Data converging with other related technologies such as Hadoop, cloud computing and Internet of Things (IoT) are further discussed in this paper. Next, this paper discussed the Big Data architecture issues. Subsequently, this paper highlighted the Big Data security concerns and privacy challenges. While cloud security alliance identified ten security and privacy challenges of Big Data, this paper further emphasized the future trends towards a data centric approaches such as searchable encryption, homomorphic encryption and attribute-based encryption.

### References


IBM, “The Four V’s of Big Data.”


Big Data: Related Technologies, Security Challenges, and Research Opportunities

Presented by: Prof. Dr. Azman Samsudin
Universiti Sains Malaysia
1.0 Introduction: Big Data

2.0 Big Data and Related Technologies

3.0 Big Data Architecture

4.0 Security Challenges and Privacy Concerns

5.0 Future and Research Trends

6.0 Conclusion
1.0 Introduction

Rapid Growth of Global Data

The amount of data being generated is DOUBLING EVERY 18 MONTHS, which growing faster than Moore's Law (which stated that every two years, computer capacity: memory, speed, etc. increases by a factor 2) (Moore & Fellow, 1998).

(Krishnan, 2013; Labrinidis & Jagadish, 2012; O'Driscoll, Daugelaite, & Sleator, 2013)

Data Production will be growth 
10 TIMES GREATER in 2020 than it was in 2013

Data managed by Enterprises will grows by 50 TIMES

(Source: The Digital Universal of Opportunities: Rich Data and the increasing value of the Internet of Things, EMC Digital Universe with Research and Analysis by IDC, April 2014)
1.0 Introduction

What is Big Data?

The term “Big Data” can be traced back to the discussions of handling large groups of datasets in both academia and industry during the 1980s (Loukides et al., 2011; Yan, 2013).

"Big data typically refers to the following types of data: i) Traditional enterprise data – includes customer information from CRM systems, transactional ERP data, web store transactions, and general ledger data. ii) Machine-generated / sensor data – includes Call Detail Records (“CDR”), weblogs, smart meters, manufacturing sensors, equipment logs (often referred to as digital exhaust), trading systems data. iii) Social data – includes customer feedback streams, micro-blogging sites like Twitter, social media platforms like Facebook." (Dijcks, 2013)

"Big data is the term increasingly used to describe the process of applying serious computing power – the latest in machine learning and artificial intelligence – to seriously massive and often highly complex sets of information.” (Microsoft, 2013)
1.0 Introduction

What is Big Data?

"Big Data technologies describes a new generation of technologies and architectures, designed to economically extract value from very large volume of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis (Carter, 2011)

"Measured in terms of volume, velocity, and variety, Big Data represents a major disruption in the business intelligence and data management landscape, upending fundamental notions about governance and IT delivery. With traditional solutions becoming too expensive to scale or adapt to rapidly evolving conditions, companies are scrambling to find affordable technologies that will help them store, process, and query all of their data. Innovative solutions will enable companies to extract maximum value from Big Data and create differentiated, more personal customer experiences." (Galtieri, 2013)
1.0 Introduction

What is Big Data?

"Big Data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.” (Gartner, 2016)

"Big Data refers to digital data volume, velocity and/or variety (veracity) that: i) Enable novel approaches to frontier questions previously inaccessible or impractical using current or conventional methods; and/or; ii) Exceed the capacity or capability of current or conventional methods and systems; iii) Differentiates by storing and analyzing population data and not sample sizes."
3Vs Characteristic of Big Data

High Volume
- Extensive Scale of Data
  - Tetrabyte, Petabyte, exabyte, zettabyte, yottabyte

High Velocity
- Speed of Data Processing
  - Batching, Streaming, Real-time, near-time

High Variety
- Diversity of Data Formats and Structures
  - Structured, Semi-structured, Quasi-structured, Unstructured

1.0 Introduction
1.0 Introduction

4Vs Characteristic of Big Data

- **High Volume**
  - Extensive Scale of Data
    - Tetrabyte, Petabyte, exabyte, zettabyte, yottabyte

- **High Veracity**
  - Uncertainty of Data
    - Data consistency, Trustworthiness, Availability, Accountability, Authenticity

- **High Velocity**
  - Speed of Data Processing
    - Batching, Streaming, Real-time, near-time

- **High Variety**
  - Diversity of Data Formats and Structures
    - Structured, Semi-structured, Quasi-Structured, Unstructured
1.0 Introduction

5Vs Characteristic of Big Data

- **High Volume**: Extensive Scale of Data
  - Tetrabyte, Petabyte, Exabyte, Zettabyte, Yottabyte

- **High Velocity**: Speed of Data Processing
  - Batching, Streaming, Real-time, near-time

- **High Value**: Added Value
  - Statistical, Events, Correlations, Hypothetical

- **High Variety**: Diversity of Data Formats and Structures
  - Structured, Semi-structured, Quasi-Structured, Unstructured

- **High Veracity**: Uncertainty of Data
  - Data consistency, Trustworthiness, Availability, Accountability, Authenticity
What is Big Data?

"Big Data consists of extensive datasets, primary associated with the 5Vs characteristic: high volume, velocity, variety, veracity and value, which resulting conventional data processing mechanisms are too expensive to scale or being adapted to store, manipulate, and analyze them efficiently."
1.0 Introduction

Growth of Big Data Market Size

![Bar chart showing the growth of Big Data market size revenue forecast worldwide from 2011 to 2026 in billion U.S. dollars.](chart.png)
1.0 Introduction

• Investment in Big Data Analytics

1. What portion of your overall technology expenditure is made in Big Data analytics?

- More than 30%: 3%
- 21%-30%: 22%
- 10%-20%: 51%
- Less than 10%: 24%

2. Do you expect this to increase, stay the same or decrease over the next year?

- Increase: 24%
- Stay the same: 76%

(Source: Accenture, 2014: Industrial Internet Insights Report For 2015)
1.0 Introduction

• Why Big Data?

(Source: Monetate, 2012)
http://content.monetate.com/h/i/12311844-the-retailer-s-guide-to-big-data
1.0 Introduction

Big Data Characteristic and Its Major Industries

<table>
<thead>
<tr>
<th>Industry</th>
<th>Volume</th>
<th>Velocity</th>
<th>Variety</th>
<th>Veracity</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical and Health Care</td>
<td>Very High</td>
<td>Very High</td>
<td>High</td>
<td>Very High</td>
<td>High</td>
</tr>
<tr>
<td>Logistic and Transportation</td>
<td>Very High</td>
<td>Very High</td>
<td>Moderate</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Manufacturing and Natural Resources</td>
<td>Very High</td>
<td>Very High</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Banking and Securities</td>
<td>Very High</td>
<td>Very High</td>
<td>Low</td>
<td>Very High</td>
<td>High</td>
</tr>
<tr>
<td>Education</td>
<td>Very Low</td>
<td>Very Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Government</td>
<td>Very High</td>
<td>High</td>
<td>Very High</td>
<td>Very High</td>
<td>Very High</td>
</tr>
<tr>
<td>Retail</td>
<td>High</td>
<td>High</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td>Wholesale Trade</td>
<td>Low</td>
<td>Low</td>
<td>Very Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Insurance</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td>Communication, Media Service</td>
<td>Very High</td>
<td>High</td>
<td>Very High</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Retail</td>
<td>High</td>
<td>High</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
</tbody>
</table>
• Confusing World of Big Data
• Big Data Vs Cloud Computing

Big Data

Hadoop

VS
• **Big Data Vs Cloud Computing**
• Big Data Vs Internet of Things (IoT)
• Oracle Big Data Reference Architecture
• **IBM Big Data Reference Architecture**
• EMC Big Data Reference Architecture

*Hadoop Ecosystem includes: Hive, Pig, Mahout, HBase, ZooKeeper, Oozie, Sqoop, Avro
3.0 Big Data Architecture

- SAP Big Data Reference Architecture
• NIST Big Data Reference Architecture

NIST Big Data Architecture Reference, 2013
### How it different from Traditional Data?

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Traditional Data</th>
<th>Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Volume</strong></td>
<td>Bytes - megabytes</td>
<td>High Volume (Terabytes- Zettabytes)</td>
</tr>
<tr>
<td><strong>Velocity</strong></td>
<td>Moderate Speed</td>
<td>High Speed</td>
</tr>
<tr>
<td><strong>Variety</strong></td>
<td>Structured Data</td>
<td>Structure , Semi-Structured, Unstructured, &quot;Quasi&quot; Structured</td>
</tr>
<tr>
<td><strong>Environments</strong></td>
<td>Homogenous</td>
<td>Heterogeneous</td>
</tr>
<tr>
<td><strong>Data Storage</strong></td>
<td>Data are collected and stored in application/ service providers</td>
<td>High Scalability</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data are stored at data owner/data producer, only the aggregated resulted will stored at application/service providers.</td>
</tr>
<tr>
<td><strong>Data Redundancy</strong></td>
<td>High Redundancy</td>
<td>Low Redundancy</td>
</tr>
<tr>
<td><strong>Security and Privacy Protection</strong></td>
<td>Data At Rest</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data In Memory</td>
<td>Data At Rest</td>
</tr>
<tr>
<td></td>
<td>Data In Transit</td>
<td>Data In Memory</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data In Transit</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data in Transform</td>
</tr>
</tbody>
</table>

Privacy Concerns on Big Data Collections and Usage
Underlying on diverse network communication Protocols for collect, exchange and transfer data

Interoperability issues

Scalability issues
Underlying on diverse network communication Protocols for collect, exchange and transfer data
Conventional Point-to-Point Defenses subjected to a Single Point of Failure
Security of Hadoop and Cloud Computing
• Hadoop Security Model

Very Limited Auditing and Authorization Controls (HDFS File Permissions)
4.0 Security Challenges and Privacy Concerns

- Big Data Security: Hadoop Security Model

Kerberos Authentication

Developed by Yahoo! team Owen O’Malley, Kan Zhang, Sanjay Radia, Ram Marti, and Christopher Harrell

Rogers, S., 2014, Understanding Hadoop security
http://blogs.sas.com/content/sgf/2014/09/24/understanding-hadoop-security/
4.0 Security Challenges and Privacy Concerns

Kerberos in Hadoop

Oozie High-level Dataflow

Source: Malley, O. et al., 2009, Hadoop Security Design
• Big Data Security : Hadoop Security Model

Kerberos Authentication

No “Data at Rest” Encryption

Limited Authorization Capabilities

Rogers, S., 2014, Understanding Hadoop security
http://blogs.sas.com/content/sgf/2014/09/24/understanding-hadoop-security/
## 4.0 Security Challenges and Privacy Concerns

<table>
<thead>
<tr>
<th>Vendor/ Security Products</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloudera Sentry</td>
<td>Fine-Grained Authorization and role-based access control</td>
</tr>
<tr>
<td>IBM InfoSphere Optim Data Masking</td>
<td>Confidential Data Encryption</td>
</tr>
<tr>
<td>Intel's secure Hadoop distribution</td>
<td>Confidential Data Encryption, Role-based access control</td>
</tr>
<tr>
<td>DataStax Enterprise</td>
<td>Confidential Data Encryption, Data auditing</td>
</tr>
<tr>
<td>Zettaset Secure Enterprise Hadoop</td>
<td>Confidential Data Encryption, Fine-Grained Authorization and role-based access control</td>
</tr>
<tr>
<td>DataGuise</td>
<td>Confidential Data Encryption, Fine-Grained Authorization and role-based access control</td>
</tr>
</tbody>
</table>
4.0 Security Challenges and Privacy Concerns

PKI, Kerberos, Data Encryption

Source: Big Data Analytics Services & Solutions
4.0 Security Challenges and Privacy Concerns

• Example: DataGuise: DgSecure for Hadoop
Possible Solutions

• Install Tamper Resistance Hardware
5.0 Future and Research Trends

Top Ten Big Data Security And Privacy Challenges

Source: Expanded Top Ten Big Data Security And Privacy Challenges, CSA, 2013
5.0 Future and Research Trends

- Towards a Data Centric Approach

Searchable Encryption

Public Key Infrastructure (PKI) & Tamper Resistance Hardware
• Searchable Encryption
• Research Concerns on Searchable Encryption
• Towards a Data Centric Approach

Homomorphic Encryption

Public Key Infrastructure (PKI) & Tamper Resistance Hardware
Enterprise is able to protect confidentiality and privacy of these outsourced data by encrypting these private data before storing them into an un-trusted third party, meanwhile, allow corresponding third-party to **perform some computation on these encrypted data without decrypting** and accessing these private data.
Homomorphic Encryption

Definition 2.1: A homomorphic encryption scheme is an encryption scheme, which has the following property for all $c_1, c_2 \in C$, $m_1, m_2 \in P$ and $K$, where $c_1 = Enc(m_1)$ and $c_2 = Enc(m_2)$ and $C$ and $P$ are groups:

$$Dec(c_1 \odot_C c_2) = m_1 \odot_P m_2$$

where $\odot_C$ and $\odot_P$ are the group operation in the ciphertext and plaintext space respectively.

If the plaintext space $P$ is an additive group, the homomorphic cryptosystem is known as additive homomorphism, likewise, if the plaintext space is a multiplicative group, that homomorphic cryptosystem is considered as multiplicative homomorphism (Kukucka, 2013).
5.0 Future and Research Trends

Privacy Homomorphism (Rivest, 1978)

Suffers from major security flaws and efficient practical issues

1st Fully Homomorphic Encryption (FHE) (Gentry, 2009)

• Theoretically demonstrated by Gentry in 2009.
• Based on Lattice Approach
• Considered as being infeasible for practical deployments due to their slow running speed (bootstrapping and squashing techniques). (Naehrig, Lauter, & Vaikuntanathan, 2011; Sen, 2013)
5.0 Future and Research Trends

Existing Works of Homomorphic Encryption
5.0 Future and Research Trends

Conventional Algebraic Cryptosystem

1) **Enjoy some sort of homomorphism** either additive or multiplicative homomorphism.

<table>
<thead>
<tr>
<th>PHE Scheme</th>
<th>Homomorphism</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSA (Rivest, 1978)</td>
<td>Multiplicative</td>
</tr>
<tr>
<td>Goldwasser-Micali (Goldwasser &amp; Micali, 1984)</td>
<td>Additive</td>
</tr>
<tr>
<td>ElGamal (Elgamal, 1985)</td>
<td>Multiplicative</td>
</tr>
<tr>
<td>Benaloh (Benaloh, 1987)</td>
<td>Additive</td>
</tr>
<tr>
<td>Domingo-Ferrer scheme (Ferrer, 1996)</td>
<td>Additive</td>
</tr>
<tr>
<td>Naccache-Stern (Naccache &amp; Stern, 1998)</td>
<td>Additive</td>
</tr>
<tr>
<td>Okamoto-Uchiyama (Okamoto &amp; Uchiyama, 1998)</td>
<td>Additive</td>
</tr>
<tr>
<td>Paillier (Paillier, 1999)</td>
<td>Additive</td>
</tr>
</tbody>
</table>

Majority of these scheme are **WELL-ESTABLISHED IN REAL-WORLD DEPLOYMENTS**

- **Speed**, **Security**

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Time for 1024 bits data block</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSA</td>
<td>took about 8.28 milliseconds</td>
</tr>
<tr>
<td>Paillier</td>
<td>only took about 2313 milliseconds</td>
</tr>
</tbody>
</table>

2) Their **LOW VERSATILITY** limiting their applicability in outsourced Big Data Computation
HE For Real-World Application: CryptDB

User Data → Song et al.'s encryption scheme → MySQL database

Keyword search query

Pai encryption scheme

Count query

Monomi: Support more complex query

5.0 Future and Research Trends

CryptDB and Its Variants

MrCrypt → MapReduce Operations

Crypsis → Pig Latin

Computing on Masked Data (CM) → NoSQL database

### Gentry’s original FHE scheme

| 1) Improves bootstrapping and squashing techniques in Gentry’s original scheme: |
| 2) Leverage the recent mathematical algorithmic advances: |
| 3) Combine Algebraic with lattice approach: |
| 4) Scheme conversion approach by convert the FHE into symmetric encryption scheme: |

#### 1) Improves bootstrapping and squashing techniques in Gentry’s original scheme:

1. Changes in Gentry’s original scheme: (Gentry & Halevi, 2011; Gentry & Smart, 2011; Smart & Vercauteren, 2012 Brakerski, 2012)

#### 2) Leverage the recent mathematical algorithmic advances:

1. Learning with the Error (LWE) & Ring LWE (RLWE) approaches: - Brakerski & Vaikuntanathan, 2011a, 2011b.

2. Number Theory Research Unit (NTRU) approach: (Stehle and Steinfeld, 2011; López-Alt et al., 2012)

#### 3) Combine Algebraic with lattice approach:

- (Dijk, Gentry, Halevi & Vaikuntanathan, 2010 (DGHV scheme); Coron et. al, 2011; Coron, Naccache & Tibouchi, 2012; Cheon, Coron, Kim & Lee, 2013; Kim, Lee, Yun & Cheon, 2013; Coron, Lepoint & Tibouchi, 2014)

#### 4) Scheme conversion approach by convert the FHE into symmetric encryption scheme:

- (Doroz et al., 2014a; Doröz et al., 2014; Gentry & Smart, 2012)

---

• Result in asymptotically **BETTER Fully Homomorphic Encryption** systems
Achieved efficiency of existing FHE schemes are **STILL FAR from being enough to support practical applications.**

---

**Gentry & Smart, 2012:**
- **LWE based FHE** (Brakerski et al., 2012)
- **AES symmetric scheme**
  - Generate single key: 30 minutes
  - Evaluate 1024 bits data: 36 hours

---

**Doroz et al (2014b):**
- **NTRU based FHE** (López-Alt et al., 2012)
- **Prince symmetric scheme**
  - Evaluate 1024 bits data: 79.5 minutes

---

**Lightweight, High Efficiency**
5.0 Future and Research Trends

• Towards a Data Centric Approach

Attribute Based Encryption

Public Key Infrastructure (PKI) & Tamper Resistance Hardware
5.0 Future and Research Trends

• Towards a Data Centric Approach

Attribute Based Encryption
• **What is Attribute Based Encryption (ABE) Scheme?**

  - Able to provide fine-grained access control on encrypted data with a data-centric approach.
  
  - Both user’s private keys and ciphertexts are associated with a set of descriptive attributes. Users are able to recover a message correctly if and only if his key’s attribute satisfies the requirement of attribute list on the ciphertexts.
• **What is Attribute Based Encryption (ABE) Scheme?**

- **Key-Policy ABE (KP-ABE)**
- **Ciphertext-Policy ABE (CP-ABE)**

**Advantages:**
- ✔ closer to support a real-world applications
- ✔ better in supporting the scalability of users.

---

**5.0 Future and Research Trends**

- **ABE for Securing Big Data Processing**

  **Two Main Streams**

  - **Bilinear map over elliptic curves approach**
  - **Lattice Based Approach**

  - Conventional Approach to construct ABE
    - Example:
      - KP-ABE schemes
      - CP-ABE schemes
  - Inspired by recent promising results on the hardness of lattice problems against the quantum attacks

Q & A Sessions
Presented by
School of Computer Sciences, Universiti Sains Malaysia