“Big Data Security Analysis”

Guan Wang, Pingram She, Fariah Mahzabeen

1. Executive Summary

The term *big data* refers to ever-increasing amount of information that organisations are storing, processing and analysing with the growing number of information sources in use. However, there are many potential security issues with big data. For example, organisations collect and process huge sensitive information regarding customers, employees, IPs (intellectual property), and financial information. Then, such confidential information are aggregated and centralized in one place for analysis in order to increase their value. Centralizing data in one place is risk per se - it is a valuable target for attackers and those confidential information might be exposed. Hence, we will first investigate potential security issues with big data, and then study corresponding state-of-art security solutions. For example, A typical security solution for centralized data security is called “Big Data Security Analytics”, which means doing big data analysis through secure analysis tools. *IBM Big Data Security Analytics Continuum* and *compliance-centric SIEM tools* are two examples of such tools. Thorugh investigation, there are four main categories of big data security solutons in prevalence: threat-collection based solution, AI-learning based solution, pipeline-pretection solution, and encryption-based solution. Any category of solution can be encapsulated as a security protection platform but each has different focus. In the solution section, we would also like to present some case studies of data breaches and how could they have been avoided with big data security solution. Finally, we will propose suggestions for orgnization that needs big data security protection based on our study in big data security issues as well as correpsonding solutions.

2. Problem Statement - Potential Security Risks

2.1 Sensitive Information

2.1.1 Customers' personal data: China's Central Television (CCTV)'s investigation into iPhones' Location Service

Recently, a new class action has been filed against Apple Inc. for its violation of iPhone user's privacy. This case relates to China's Central Television (CCTV) report on July 11 about Apple's location services. Though clearly presented on Apple's webpage about how to handle privacy on an iPhone, and can easily turned off in iPhone's settings, the location services actually are not stopping tracking users' activities. The truth is that Apply indeed transmits these highly sensitive and private consumers data to its own database for future reference, but will never share the data about consumers' daily whereabouts to any third party.

Similar cases happen ubiquitously in this big data era. Big data refers to not only the databases on-premise, but also those data in the cloud. The data includes web data, industry specific big transaction data, machine generated/sensor data and plain text. Enterprises continue collecting terabytes of sensitive data from their customers, via different forms on various platforms. These data may be captured for regulatory compliance and post hoc forensic analysis. The numbers will climb up if enterprises log more events in more sources except for the existing network events, software application events, and people's action events. The increase of staff and devices will also stimulate the growth of the number.

2.1.2 Employees' personal data

There are general laws which the tenant or cloud customer operating in the US, Canada or European Union are subject to. Some of these laws have specified markets, such as HIPAA for the health-care industry. However, such sensitive data are not only collected by health-care industry. Companies often store health-related information about their employees.

From companies' side, failure to provide adequate protection to those sensitive personal data can lead to severe consequences, including fines by governments or industry regulatory bodies. On the other side of employees, since no system is

absolutely secure, their personal data faces potential exposure when on their employers' hands.

2.2 Data Storage

2.2.1 Distributed Programming Frameworks

As Big Data has turned the centralized systems into distributed systems, or self-governance systems, the data become distributed simultaneously. What's more, the data doesn't stay in one place, they are flowing from one system to another, which increases the amount of data needed to be stored. One example for the parallelism in distributed programming frameworks is MapReduce framework, which splits one input into multiple groups.[7] However, untrusted mappers can return wrong results, which will late generate wrong aggregate results.

Retailer consumer data is often analyzed for targeted advertising or customer segmentation. Highly parallelized computations are executed over large data sets. However, leakages, whether intentional or unintentional, may happen in these mappers.

2.2.3 Security issues for Non-Relational Data Stores

Despite its advantages of scalability, flexibility, and administrator-friendliness compared to relational databases, NoSQL databases are weak in security infrastructure. Security was not considered at its design stages. NoSQL databases do not support for enforcing it within the database. The only way to secure the databases is to embed security in the middleware.

2.2.4 Secure Data Storage and Transactions Logs

The increasing amount of data has forced the IT managers to give up their direct control over the data to auto-tiering. Data and transaction logs now processed and stored in auto-tiering storage systems have to fit in a hierarchal tier for saving storage and money. Data with lower utility is moved to a lower and cheaper tier. However, IT managers can not have a clear image of where and when the data is moved, and the power of access authorization is weakened at the same time.

2.3 Data Processing

2.3.1 Real-time Security Monitoring

GIven the number of security alerts generated by devices, real-time monitoring can not always perform perfectly. These alerts are often simply ignored or clicked away. This problem severes when the volume and velocity of data streams increase. It calls on huge funds and storage in order to make faster and better decisions.

2.3.2 Scalable and Composable Privacy-Preserving Data Mining and Analytics

Enterprises and government agencies continually collect, and analyze user data for business or non-profit purposes. This involves inside analysts and potentially outside contractors or business partners. The insecure risks embedded here could be a malicious insider or untrusted partner abusing these datasets and disclosing private information from customers.

2.4 Data Accessing

2.4.1 Cryptographically Enforced Access Control and Secure Communication

Sensitive data stored in the cloud are not encrypted for easily performing or sharing records. This can not guarantee the data transferred from one system to the other are protected under one or multiple encryption methods. While, those less sensitive data, such as data useful for analysis, may also become targets of hackers or malicious person. They are unencrypted as well, which required secure communication framework.

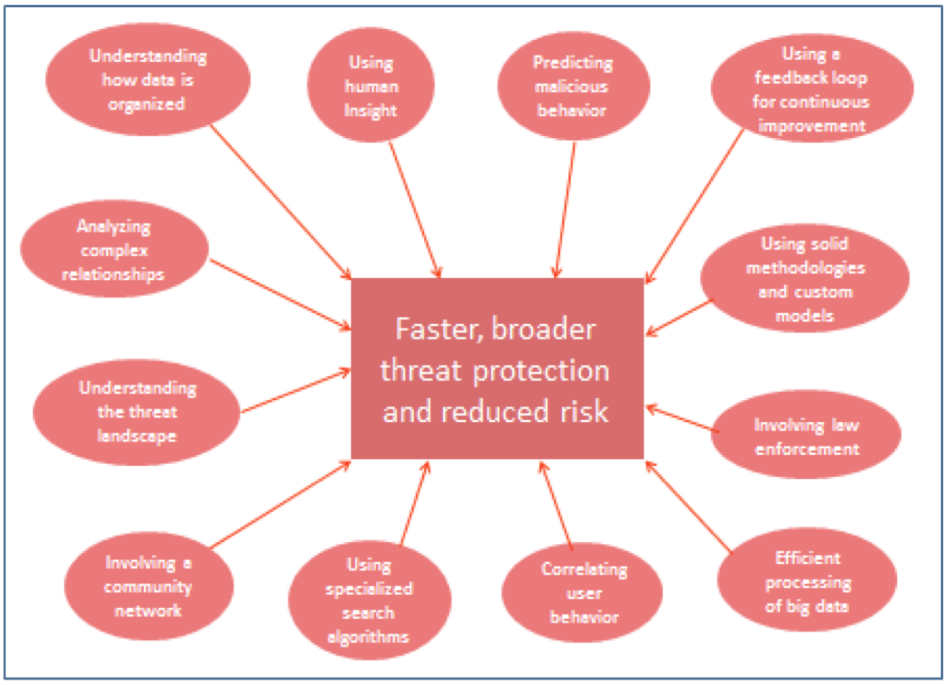
2.4.2 End-Point Input Validation/Filtering

For the purposes of data mining and data analysis, enterprises acquire data from many sources, such as end-point devices, such as hardware devices and software applications within enterprise networks. But the validation and filtering of those data become huge problems with the sizable data chunks and enterprises' scales. Especially, with the "Bring Your Own Device" (BYOD) model, employees are bringing their own laptops or smart phones to the company and connect them to the company's network often without company's knowledge or authorization. These untrusted input sources pose security challenges to big data processing for enterprises.

3. Analyse Results - Current Solutions

3.1 Treat-Based Solution: Trend Micro Smart Protection Solution [4]

Just like protection from virus and malware for software, Trend Micro provide security solutions for big data using collected big data threats. That is, Trend solution tries to collect as many potential threats as possible (either in its user feedback or from reported threats) and avoid any similar threats from the protected big data platform. Hence, the core technology of Trend Micro Smart Protection (TMSP) is to identify and collect any potential big data security threat, and use those threats to get rid of any potential similar threats. Following figure illustrates a numerous service provided by Trend using the collected threats, including customer feedback, malicious behaviour prediction, law enforcement, and correlate user behaviour.



3.1.1 Threat Definition & Identification

In order to collect as many useful threats as possible, it is crucial for Trend to first set the definition of “useful threats”. The threat landscape has evolved simultaneously with the number of threats increasing by orders of magnitude in short periods. As defined by Trend, today’s threat environment imposes the “three Vs of big data”: *volume, variety, and velocity,* and each of these is increasing at an astounding rate

and has required a shift in how security vendors manage threats. The definition of the three Vs is as following [4]:

- For ***volume***, “the threat landscape is evolving in various ways, including growth in the sheer volume of threats”. A typical example is the huge increase of receipt of spam messages from 1 or 2 spam messages per day in 1990s to 200 billion spam messages sent per day in 2010 [4].

- For ***variety***: “today, cyber-criminals are sophisticated, evolving their craft and tools in real time”. For example, malware created today often undergoes quality control procedures, and cyber-criminals are much more mature than ten years ago.

- For ***velocity***, “the need to manage, maintain and process this huge volume and variety of data on a regular basis presents security vendors with an unprecedented velocity challenge”. For example, in early 2012, cybercriminals installed an iFrame redirection on a popular news site in the Netherland, which made a legitimate website infect thousands of people during their lunch hour [4].

3.1.2 Threat Collection

The main resource of threat collection for Trend is from the feedback from Customers, where Trend establishes a synergistic relationship with customers and other third parties that are constantly exposed to ever-evolving malicious content. A key part of the partnership is to create “a licensing agreement that allows customers to anonymously donate suspicious data for analysis and reverse engineering can provide valuable access to real data on real machines operating in the real world. Based on data gathered from this community network, specialized search algorithms, machine learning, and analytics can then be brought to bear on this data to identify abnormal patterns that can signal a threat” [4]. A typical procedure for a computer user is to first follow a typical daily pattern which consist of “visiting a news site, encountering several ad servers, and logging on to Facebook” [4], and then an incident can be immediately prioritized for further analysis once that pattern suddenly changes (e.g. moving the user to a domain never previously visited). Hence, a feedback loop for process improvement is established between Trend and customers, and observation and curation of key data will be fed back into the process allowing for continual process improvement. Hence, the process can collect malicious behavior over the time [4].

Besides customers' feedback, human experts is another important resource for Trend to collect any potential complicated threats. Trained analysts constantly evolve the combination of methodologies, apply human intuition to complex problems, and identify new threats that computers miss from a diverse of resources like reports.

3.2 AI-Based Solution: Real-Time Intelligence Learning Solutions

Increasingly, organizations are gathering security event data and logs from systems and applications for a broad variety of purposes. Logs are an excellent starting point for effective security information and event management. The majority of responding organizations are leveraging security logs and event data for the following reasons [10] [12]:

- "Detect and track suspicious behavior"
- "Support forensic analysis and correlation"
- "Prevent incidents"
- "Achieve/prove compliance with regulatory requirements"

Hence, based on the security logs and event data, organizations could recognize patterns to detect and prevent potential security threats in big data. In this section, two learning method based solutions (*Supervised Learning Solution: SANS Real-time Security Intelligence & Unsupervised Learning Solution: Red Lambda's Big Data Security & Analytics Solutions*) are introduced with emphasis on the learning methods they are employing.

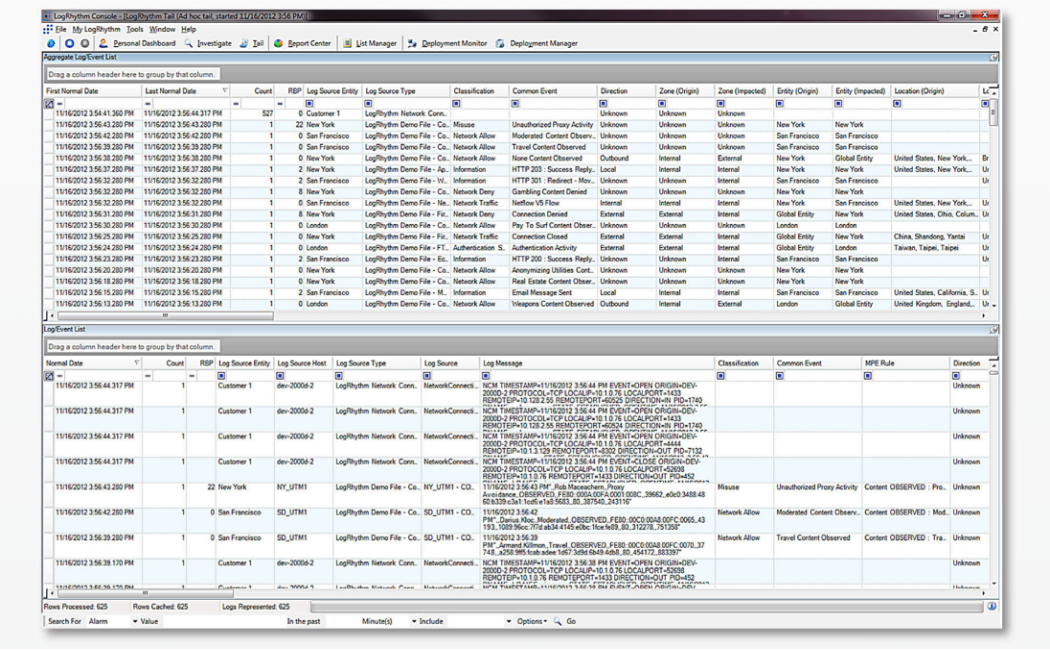
3.2.1 Supervised Learning Solution: SANS Real-time Security Intelligence

Supervised learning is the machine learning task of inferring a function from labeled training data, where the training data consist of a set of training examples. SANS employs its internal log system "LogRhythm" to manually generate threat examples used in the training examples to build a supervised learning model to predict new threats. Based on the trained models, SANS are able to do following real-time analysis first in order to predict and track suspicious patterns from the big data [10]:

- *General Analysis* - "Fundamental operations for a security analyst, including aggregation of logs and events, audit, operations, and security logs and events, are listed by classification categories, logs by type and logs by direction of traffic."

- Time Analysis - “Analysts can easily view logs and events broken down by time of occurrence. These items are further broken down by type, direction of traffic and variably granular time options (e.g. weeks versus days)”
- *Statistical Analysis* - “This category allows security analysts to view granular data and graphs about logs and events from the environment. Analysis categories include log source statistics, origin login and host statistics (i.e., system and user information), impacted hosts and applications, and even vendor message.”
- *Topx Analysis* - “The TopX categories are open, flexible containers that analysts can populate with the top the Top 10 user accounts appearing in log events or the Top 20 domain names present in events. These adaptive monitoring options.”

Following is an example of the “Tail” analysis feature in SANS solution. In the figure below, it shows how SANS tracks a series of network connectivity events. It is easy to use the Tail feature to quickly get a “big picture” view of all events happening in this environment, at both the log layer and the event layer, across multiple log sources.



After those raw analysis features, SANS can apply its AI engine for advanced analysis. The AI Engine focuses on behavioral profiles and advanced correlation

between widely disparate data, which offers features and capabilities for “real-time analysis, detection and response, and support for rapid forensic drill down through the full universe of log, flow and event data recognized by the LogRhythm platform” [10].

3.2.2 Unsupervised Learning Solution: Red Lambda’s Big Data Security & Analytics Solutions [12]

Red Lambda’s Big Data Security & Analytics Solutions (RLBDSAS) Suite employs unsupervised learning methods to “immediately” detect new threats and prevent big data from them. The core technologies of RLBDSAS are two main points: 1) real time analysis; 2) unsupervised clustering of new events.

Real-time is defined by Red Lambda as “what’s happening right now”[12]. It is accomplished through its stream processing technology and real-time, incremental analytics, which empowers the organization to act on each new event and data point “the moment the data hits the edge of the network”[12].

Red Lambda holds *patent-pending Neural Foam*™ with artificial intelligence (AI) to “cluster massive amounts of data into its simplest, natural structure, without a single rule, signature or line of code” [12]. By grouping similar data, trillions of events can be reduced to just a few simple clusters for “immediate visibility and exploration”[12]. Finally, independent streams of events are “automatically correlated, classified and analyzed for novelties”[12].

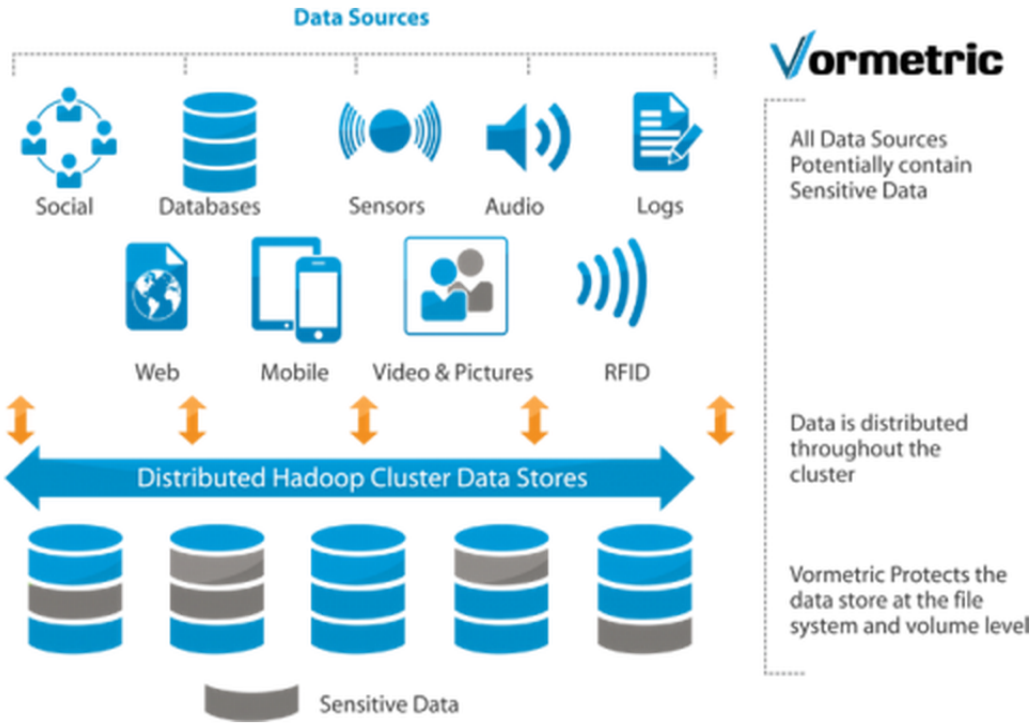
3.3 Encryption-Based Solution: Vormetric Data Security Solutions

Given the very large data sets that contribute to a Big Data implementations, either protected information or critical Intellectual Property (IP) may be in risk[9]. The key idea behind Vormetric Data Security Solutions is to first detect sensitive data and then encrypt them and manage encryption keys.

3.3.1 Protect large range of data source [9]

Vormetric protects various types of data including social network data, local database data, sensor data, audio data, logs data, web data, mobile data, video & picture data, and RFID data. Sensitive data can be potentially contained in all of those data source above. Also, those data could be distributed through Hadoop

clusters data stores. Hence, Vormetric tries to protect sensitive data in those data sources at the file system and volume level.



3.3.2 Detect Sensitive Data [9]

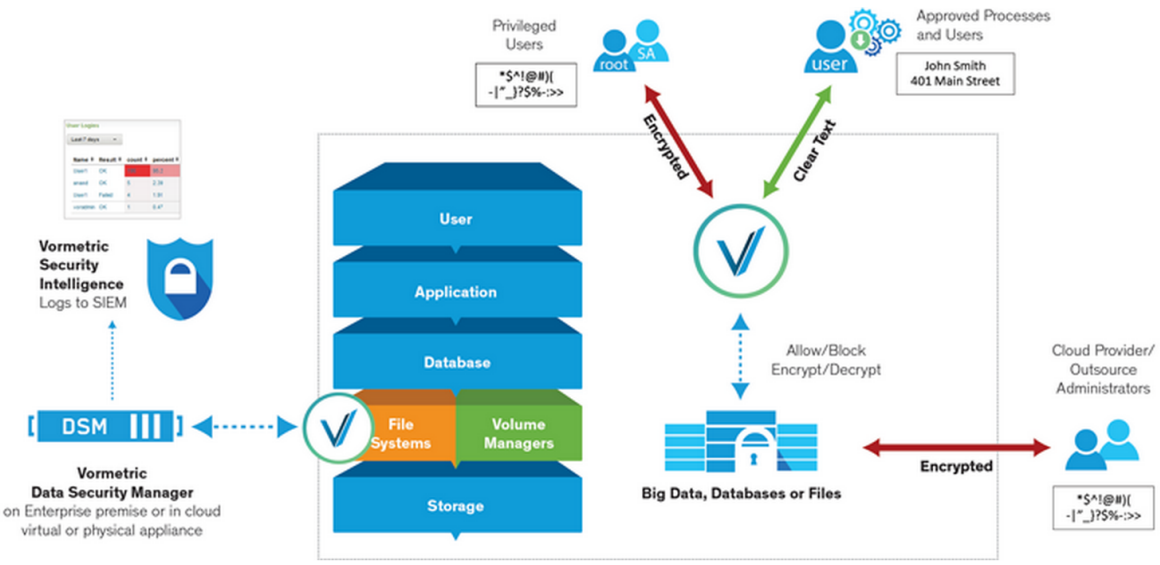
Vormetric defined two types of sensitive data: 1) data that a malicious insider is looking for; 2) data that Advanced Persistent Threat (APT) attacker is looking for. Those sensitive data are worth stealing due to of the high value information. For example, sensitive data in a report is much easier to find and extract than mining the information, and those information may be considered as sensitive data.

3.3.3 Seamlessly Encryption and Key Management [9]

Vormetric Encryption seamlessly protects Big Data environments at the filesystem and volume level. Data breach mitigation and compliance regimes require encryption to safeguard data. Vormetric provides the strong, centrally managed, encryption and key management that enables compliance and is transparent to processes, applications and users. In this way, users and applications have no sense about which data is actually sensitive, which is more secure.

3.3.4 All-in-one solution as Data Security Platform [9]

Besides encryption and key management features, Vormetric provides other advanced features for big data security protection and encapsulate them as one uniform platform. This Big Data analytics security solution allows organizations to gain the benefits of the intelligence gleaned from Big Data analytics while maintaining the security of their data. The data security platform includes "strong encryption, key management, fine-grained access controls and the security intelligence information"[9].



With those advanced features, it is easier to identify the latest in advanced persistent threats (APTs) and other security attacks on the big data. Other advanced features are as described as following:

- *Fine-grained Access Controls* - "Vormetric first set the fine-grained, policy-based access controls to the restrict access to data that has been encrypted. For processes and users, It only allows approved access to data and enforce any requirements to meet strict compliance requirements. Privileged users of all types are allowed to access plaintext information only if specifically enabled. System update and administrative processes continue to work freely – but see only encrypted data, not the plaintext source. "[9]

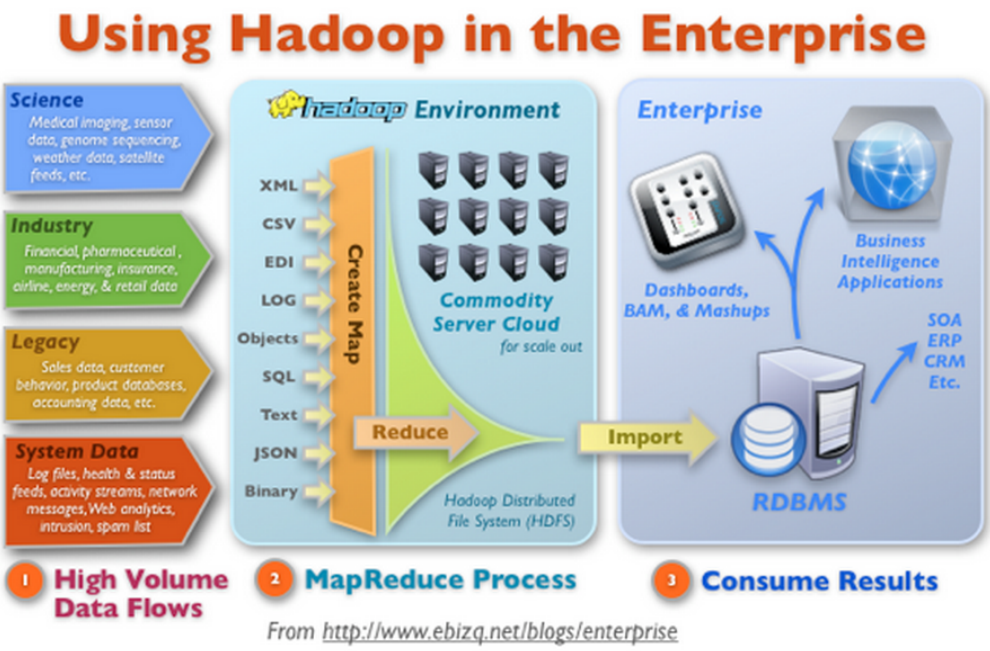
- *Security Intelligence* – “Vormetric logs capture all access attempts to protected data providing high value, security intelligence information that can be used with a Security Information and Event Management solution to identify compromised accounts and malicious insiders as well as finding access patterns by processes and users that may represent and APT attack in process. ”[9]

- *Automation* – “Use the Vormetric Toolkit to easily deploy, integrate and manage your Vormetric Data Security implementation with the rest of your big data implementation.” [9]

3.4 Framework Protection: Voltage Enterprise Security Solution for Big Data

3.4.1 Big Data Processing Architectures, Pipelines, and Frameworks

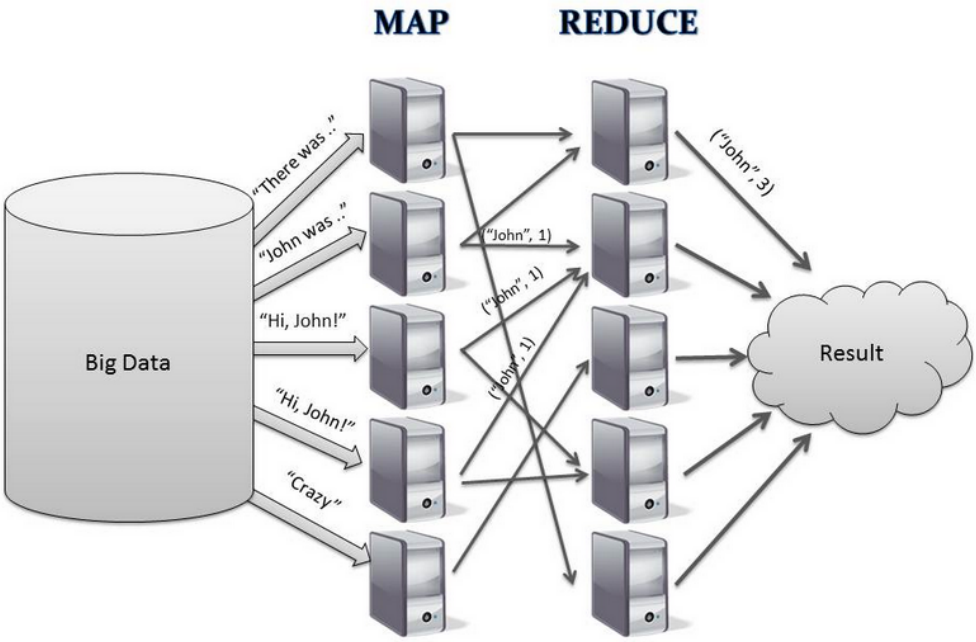
For most of the big data processing, it is dealing with a huge amount of data. Therefore, the traditional processing framework like sequential processing programs, and single threaded programs, is too time-consuming to process the big data. Hence, parallel computing architectures and frameworks are introduced to fasten the speed to process the big data. One of the most famous such architectures is Hadoop, which is able to “aggregate structured, semi-structured and unstructured data, allow parallel computations on large datasets, and continuously feed that store to enable data scientists to see patterns and trends”[11].



As shown above, Apache Hadoop framework is composed of the following modules[15]:

- Hadoop Common – “contains libraries and utilities needed by other Hadoop modules.”
- Hadoop Distributed File System (HDFS) – “a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster.”
- Hadoop YARN – “a resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications.”
- Hadoop MapReduce – “a programming model for large scale data processing.”

Among those modules, Hadoop MapReduce is the most widely used modulos. As shown in the figure below, a MapReduce program is composed of “a Map() procedure that performs filtering and sorting (such as sorting students by first name into queues, one queue for each name) and a Reduce() procedure that performs a summary operation (such as counting the number of students in each queue, yielding name frequencies)” [16]. The "MapReduce Framework" orchestrates “the processing by marshalling the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system, and providing for redundancy and fault tolerance” [16].



3.4.2 Risk from Architectures, Pipelines, and Frameworks [11]

Those pipelines, architectures and frameworks may introduce big potential risks. There might be an increased risk of a potential data breach with the massive amounts of data. In the way, sensitive data might be exposed, and compliance and data security will eventually be violated. Moreover, data residency laws might be broken. For example, data are aggregated across different borders. Hence, a solution to secure sensitive data while enable analytics for meaningful insights, is necessary for any Big Data frameworks, which is the key technology provided by Voltage.

3.4.3 Architectures, Pipelines, and Frameworks Protection

Voltage Security delivers protection strategy on each components of the the Architectures, Pipelines, and Frameworks in the following procedures [11]:

- “Secure sensitive data entering Hadoop, then control access.”
- “Protect data from any source, of any format, before it enters Hadoop.”
- “Set policies enabling which applications and which users get access to which original data, with protection of sensitive data that maintains usable, realistic values for accurate analytics and modeling on data in its encrypted form.”
- “Assure global regulatory compliance”

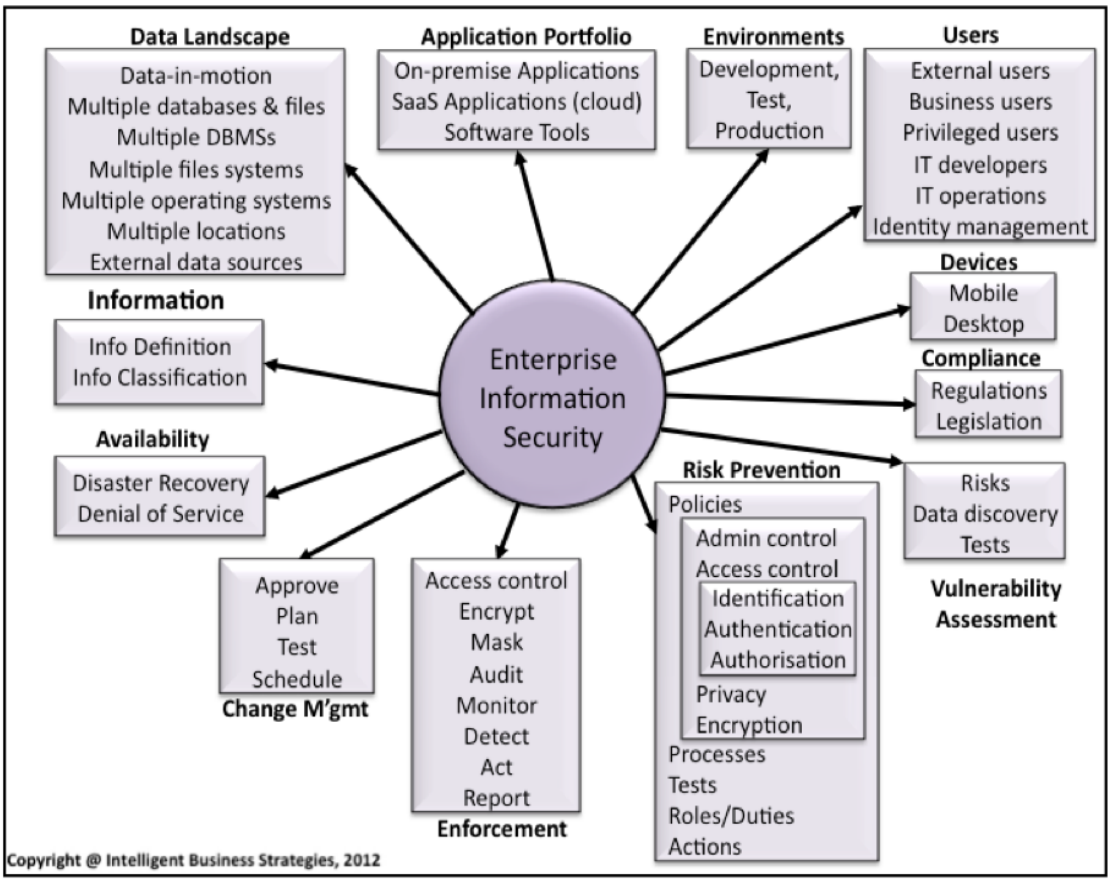
- “Securely capture, analyze and store data from global sources, and ensure compliance with international data security, residency and privacy regulations. Address compliance comprehensively, not system-by-system.”
- “Optimize performance and scalability”

In this way, Voltage will secure any components of the framework, architecture and pipeline. Eventually, Voltage is able to “Integrate data security fast, with quick implementation and an efficient, low-maintenance solution that won’t degrade performance and will scale up” [11].

4. Proposed Suggestions

Based on our study in big data security issues as well as correpsonding solutions, we would like to provide suggestions for orgnization that needs big data security

solutons for Enterprise Information Security. As shown in the figure below, we will introduce our suggestions in several aspects including Information, Data Landscape, User, Mobile, etc.



4.1 Information

In order to secure enterprise information security in Big Data environment, the first step is to classify data which is sensitive and asks for extra care. It involves what data needed to be protected and where the data is stored. The data are not only stored in on -premise databases, but also cloud-based platforms. The widely distributed sensitive data across the enterprise network increases the risk of security breaches.

For structured data, there are some common data definitions that can be used to classify data as sensitive. Policies also need to be attached to these sensitive data to control data privacy and access security. Some unstructed data needs protections, too, such as supplier constracts, marketing brochure, and so on.

4.2 Data Landscape

Then, as IT managers, a good understanding of existing data landscape help them to locate sensitive information in order to protect it. Some data streams may be sensitive becaused data protection should include also data in motion whether it has been stored or not.

4.3 User

Which type of users and devices are allowed to access sensitive data or applications also needs to be predefined to manage authentication and authorization. The identity of individual users or devices should be monitored to govern what data they can access.

4.4 Device

Enterprises should enforce device authentication to guarantee trusted access to sensitive data. Control over access to mobile applications and functionality includes mobile application authentication via user ID and password, role-based access allowing users to specific application functionality, and a set time period to specific applications from a mobile device.

5. Conclusion & Future Work

In this paper, we have discuessed modern big data security issues and corresponding state-of-art security solutuons provided by various technical companies. We found that there is no solution that is absolutely better than the other solutions. Instead, those solutions have different focuses, and choosing which type of solution heavily depends on the organizations' need - which part of big data do they actally want to protect or how to protect their big data. For example, whether they want to protect the big data processing pipeline (in this case, Voltage Solution is a better one) or only the sensitive data (in this case, Vormertic is a preferred one), whethter they trust previously-collected threat (just like what Trends has provided) or trust AI learning techniques to do real-time detection and prevention (where Red Lamba or SANS is an idea solution). Hence, different organizations have different requirements on their big data and may adopt different big data security solution sets. In our future work, we will keep an eye closely on the development of big data technology and correpsonding emerging security issues and solutions within it. Those updated could be collected through various resources like reports, blogs, news, etc. Also, we extend our study target to global organizations - whether they have ths similar big data security issues or similar security solutions. Finally, besides security issues, we will step into potential ethical issues resulted from big data for citizens.

6. References

[1] Colin Tankard, Big data security.
http://www.sciencedirect.com/science/article/pii/S1353485812700636#

[2] Using big data to reduce security risks.
http://www.sciencedirect.com/science/article/pii/S1361372312700805

[3] IBM- An Early Leader across the Big Data Security Analytics Continuum, EMC
http://www.ndm.net/siem/pdf/IBM%20An%20Early%20Leader%20across%20the%20Big%20Data%20Security%20Analytics%20Continuum.PDF

[4] Addressing Big Data Security Challenges: The Right Tools for Smart Protection, Trend
http://www.trendmicro.com/cloud-content/us/pdfs/about/wp_big-data-security-challenges.pdf

[5] Big security for big data, HP
https://ssl.www8.hp.com/ww/en/secure/pdf/4aa4-4051enw.pdf

[6] The big data security analysis era is here, EMC
http://www.emc.com/collateral/analyst-reports/security-analytics-esg-ar.pdf

[7] Big data top ten,
http://www.isaca.org/Groups/Professional-English/big-data/GroupDocuments/Big_Data_Top_Ten_v1.pdf

[8] IBM Security Intelligence with Big Data
http://www-03.ibm.com/security/solution/intelligence-big-data/

[9] Vormetric Data Security Solutions BIG DATA SECURITY
http://www.vormetric.com/data-security-solutions/applications/big-data-security

[10] SANS Security Intelligence in Action
http://ecrm.logrhythm.com/WSANSReviewBigDataSecurityanalytics1212.html?utm_medium=cpc&utm_campaign=DataSecurity&AdGroup=BigDataSecurity&gclid=Cj0KEQjw3cKeBRDG-KKqqIj4qJgBEiQAOamX_bCqzXXvwXN5mGGbpqFifeAvYonMOq07xhu-IrIAkQEaAtag8P8HAQ

[11] Voltage Enterprise Security for Big Data
http://www.voltage.com/solution/enterprise-security-for-big-data/

[12] Red Lambda's Big Data Security & Analytics Solutions
http://www.redlambda.com/why-red-lambda/real-time

[13] New Class Action Brewing Against Apple over Invasion of Privacy Brought on by a TV Report in China
http://www.patentlyapple.com/patently-apple/2014/07/new-class-action-brewing-against-apple-over-invasion-of-privacy-brought-on-by-a-tv-report-in-china.html

[14] Apple Responds to China's CCTV Report on Privacy and Location Services
https://www.apple.com/cn/your-location-privacy/#english

[15] Apache Hadoop,
http://en.wikipedia.org/wiki/Apache_Hadoop

[16] Mapreduce,
http://en.wikipedia.org/wiki/Mapreduce