Volume 5, Issue 3, March 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Big-Data Security

Kalyani Shirudkar, Dilip Motwani
Department of Computer Engineering
VIT, Mumbai, India

Abstract: Big data implies performing computation and database operations for massive amounts of data, remotely from the data owner's enterprise. Since a key value proposition of big data is access to data from multiple and diverse domains, security and privacy will play a very important role in big data research and technology. The limitations of standard IT security practices are well-known, making the ability of attackers to use software subversion to insert malicious software into applications and operating systems a serious and growing threat whose adverse impact is intensified by big data. So, a big question is what security and privacy technology is adequate for controlled assured sharing for efficient direct access to big data. Making effective use of big data requires access from any domain to data in that domain, or any other domain it is authorized to access. Several decades of trusted systems developments have produced a rich set of proven concepts for verifiable protection to substantially cope with determined adversaries, but this technology has largely been marginalized as "overkill" and vendors do not widely offer it. This talk will discuss pivotal choices for big data to leverage this mature security and privacy technology, while identifying remaining research challenges.

Keywords: Big Data, security, privacy, security Practices

I. INTRODUCTION

I. Big-Data

Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications. Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data.

Big data brings big value. With advanced big data analyzing technologies, insights can be acquired to enable better decision making for critical development areas such as health care, economic productivity, energy, and natural disaster prediction. The big data refers to massive amounts of digital information companies and government collect about us and our surroundings Voluminous data are generated from a variety of users and devices, and are to be stored and processed in powerful data centers. As such, there is a strong demand for building an unimpeded network infrastructure to gather geologically distributed and rapidly generated data, and move them to data centers for effective knowledge discovery**.**It's just standard data that's usually distributed across multiple locations, from a diverse array of sources, in different formats and often unstructured. The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations.

II. Why Larger Data ?

The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to spot business trends, prevent diseases, combat crime and so on.

III. Data Sets Grow In Size

In part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identifications (RFID) readers, and wireless sensor networks. The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization.

IV. How To Handle

Big data is difficult to work with using most relational database management system and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers"– What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain. For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration.

V. Characteristics

Big data can be described by the following characteristics:

1)Volume : The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered as Big Data or not. The name 'Big Data' itself contains a term which is related to size and hence the characteristic Many factors contribute to the increase in data volume. Financial Transaction-based data stored through the years. Unstructured data streaming in from social media , location-based data, customer interactions, the supply chain, as well as data produced by employees, contractors, partners and suppliers using social networking sites, intranets, extranets, and corporate wikis, in fact, sources such as mobile and online transactions, social media traffic and GPS coordinates now generate more. Increasing amounts of sensor and machine-to-machine data being collected. In the past, excessive data volume was a storage issue. But with decreasing storage costs, other issues emerge, including how to determine relevance within large data volumes and how to use analytics to create value from relevant data.

2)Variety : The next aspect of Big Data is its variety. This means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data analysts. This helps the people, who are closely analyzing the data and are associated with it, to effectively use the data to their advantage and thus upholding the importance of the Big Data. Data today comes in all types of formats. Structured, numeric data in traditional databases. Information created from line-of-business applications. Unstructured text documents, email, video, audio, stock ticker data and financial transactions. Managing, merging and governing different varieties of data is something many organizations still grapple with.

3)Velocity: The term 'velocity' in this context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development. Data is streaming in at unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations.

4)Variability : This is a factor which can be a problem for those who analyze the data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively. In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Is something trending in social media? Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data involved.

5)Complexity: Data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data. This situation, is therefore, termed as the 'complexity' of Big Data. Today's data comes from multiple sources. And it is still an undertaking to link, match, cleanse and transform data across systems. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control

VI. Why Big Data Should Matter To You

The real issue is not that you are acquiring large amounts of data. It's what you do with the data that counts. The hopeful vision is that organizations will be able to take data from any source, harness relevant data and analyze it to find answers that enable 1) cost reductions, 2) time reductions, 3) new product development and optimized offerings, and 4) smarter business decision making. For instance, by combining big data and high-powered analytics, it is possible to:

- Determine root causes of failures, issues and defects in near-real time, potentially saving billions of dollars annually.
- Optimize routes for many thousands of package delivery vehicles while they are on the road.
- Analyze millions of SKUs to determine prices that maximize profit and clear inventory.
- Generate retail coupons at the point of sale based on the customer's current and past purchases.
- Send tailored recommendations to mobile devices while customers are in the right area to take advantage of offers.
- Recalculate entire risk portfolios in minutes.
- Quickly identify customers who matter the most.
- Use click stream analysis and data mining to detect fraudulent behavior.

VII. Challenges To Consider

Many organizations are concerned that the amount of amassed data is becoming so large that it is difficult to find the most valuable pieces of information.

- What if your data volume gets so large and varied you don't know how to deal with it?
- Do you store all your data?
- Do you analyze it all?
- How can you find out which data points are really important?
- How can you use it to your best advantage?

Until recently, organizations have been limited to using subsets of their data, or they were constrained to simplistic analyses because the sheer volumes of data overwhelmed their processing platforms. But, what is the point of collecting

and storing terabytes of data if you can't analyze it in full context, or if you have to wait hours or days to get results? On the other hand, not all business questions are better answered by bigger data. You now have two choices:

Incorporate massive data volumes in analysis. If the answers you're seeking will be better provided by analyzing all of your data, go for it. High-performance technologies that extract value from massive amounts of data are here today. One approach is to apply high-performance analytics to analyze the massive amounts of data using technologies such as grid computing, in-database processing and in-memory analytics.

Determine upfront which data is relevant. Traditionally, the trend has been to store everything (some call it data hoarding) and only when you query the data do you discover what is relevant. We now have the ability to apply analytics on the front end to determine relevance based on context. This type of analysis determines which data should be included in analytical processes and what can be placed in low-cost storage for later use if needed.

VIII. Application

- Engage with your customers effectively – and individually – using insights from Big Data
- Identify and address potential fraud before it happens by uncovering patterns in Big Data
- Monitor supply and demand in real time to respond faster than ever before
- Understand the affinity between products and use Big Data to improve price promotions
- Create innovative business applications to deliver competitive advantage.

II. WHAT IS BIG DATA SECURITY?

Security and privacy issues are magnified by velocity, volume and variety of big data, such as large scale cloud infrastructures, diversity of data sources and formats, streaming nature of data acquisition ,and high volume inter-cloud migration. The use of large scale cloud infrastructure with diversity of software platforms, spread across large networks of computers, also increases the attack surface of entire system. therefore traditional security mechanisms, which are tailored to securing small scale static(as opposed to streaming)data, are inadequate. Ex. analytics for anomaly detection would generate too many outliers. similarly ,it is not clear how to retrofit provenance in exiting cloud infrastructure. streaming data demands ultra –fast response times from security and privacy solutions.

I. Privacy and Security

With a variety of personal data such as buying preference healthcare records, and location-based information being collected by big data applications and transferred over networks, the public's concerns about data privacy and security naturally arise. While there have been significant studies on protecting data centers from being attacked, the privacy and security loopholes when moving crow sourced data to data centers remain to be addressed. There is an urgent demand on technologies that endeavor to enforce privacy and security in data transmission. Given the huge data volume and number of sources, this requires a new generation of encryption solutions (e.g., homomorphic encryption). On the other hand, big data techniques can also be used to address the security challenges in networked systems. Network attacks and intrusions usually generate data traffic of specific patterns in networks. By analyzing the big data gathered by a network monitoring system, those misbehaviors can be identified proactively, thus greatly reducing the potential loss.

II. Security a Big Question of Big Data

Big data implies performing computation and database operations for massive amounts of data, remotely from the data owner's enterprise. Since a key value proposition of big data is access to data from multiple and diverse domains, security and privacy will play a very important role in big data research and technology. The limitations of standard IT security practices are well-known, making the ability of attackers to use software subversion to insert malicious software into applications and operating systems a serious and growing threat whose adverse impact is intensified by big data. So, **a big question is what security and privacy technology is adequate** for controlled assured sharing for efficient direct access to big data. Making effective use of big data requires access from any domain to data in that domain, or any other domain it is authorized to access. Several decades of trusted systems developments have produced a rich set of proven concepts for verifiable protection to substantially cope with determined adversaries, but this technology has largely been marginalized as "overkill" and vendors do not widely offer it.

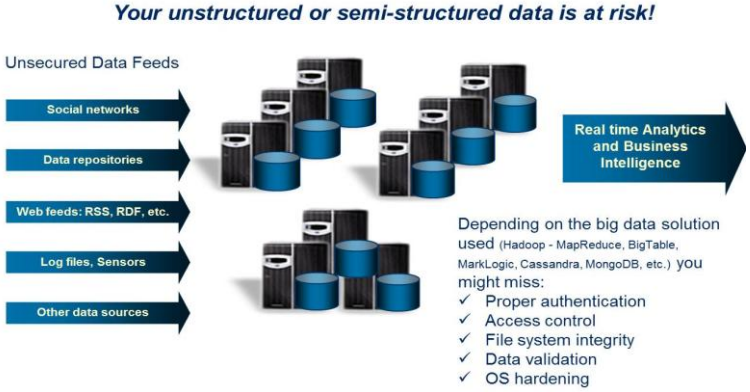

Fig 2.1 Your unstructured or semistructured data is at risk

III.Is Your Unstructured, Semi- or Structured Data at Risk?

With great power of data comes great responsibility! A big data initiative should not only focus on the volume, velocity or variety of the data, but also on the best way to protect it. Security is usually an afterthought, but Elemental provides the right technology framework to get you the deep visibility and multilayer security any big data project requires. Multilevel protection of your data processing nodes means implementing security controls at the application, **operating system** and network level while keeping a bird's eye on the entire system using actionable intelligence to deter any malicious activity, emerging threats and vulnerabilities.

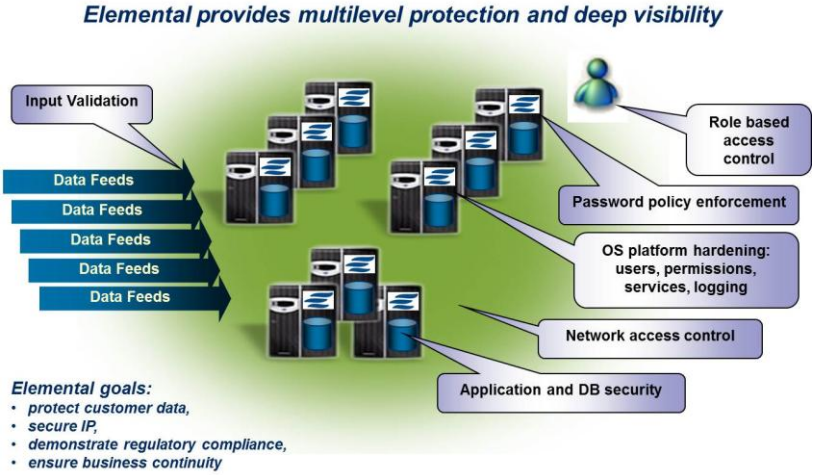

Fig 2.2 Elemental Multilevel Security

Harden Big Data Infrastructure with Elemental

Through specific big data **security policy** (controls) deployment, monitoring and enforcement, Elemental provides an integrated, cross-platform, comprehensive way to protect resources in a big data production environment.

ESP (Elemental Security Platform) is your best ally to ensure protection and compliance of your big data processing nodes:

1)Passwords : most NoSQL Big Data systems don't have any PW or use the default system PW, so anybody could easily access them. ESP can check on passwords and enforce them.

2)Input Validation : NoSQL systems aren't normally exposed to SQL injection problems, but they can still be injected using **JavaScript** or concatenation of strings. ESP could help check and not allow JavaScript.

3)Role-based Access Control : define and enforce who has access to what in the **data repository**. ESP can help check and enforce this.

4)OS Hardening : the operating system on which the data is processed should be hardened and locked down. The four main protection focus areas should be: users, permissions, services, logging. ESP provides with numerous policies for all including recommendations from CIS, Defense Info Systems Agency and **more**.

5)Persistent Control : constant monitoring and continuous enforcement of host-level security policies is provided by the Elemental system.

6)Responsive to Change : access controls automatically adapt to changes in roles and security posture.

7)In-Line Remediation : update configuration, restrict applications and devices, restrict network access in response to non-compliance.

IV. IBM Security Intelligence with Big Data

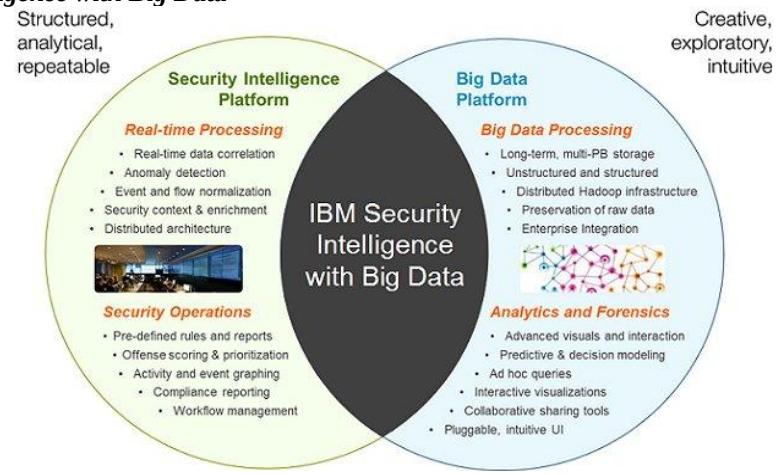

Fig.2.3 IBM Security Intelligence with Big Data

Key capabilities

- Real-time correlation and anomaly detection of diverse security data
- High-speed querying of security intelligence data
- Flexible big data analytics across structured and unstructured data – including security data; email, document and social media content; full packet capture data; business process data; and other information
- Graphical front-end tool for visualizing and exploring big data
- Forensics for deep visibility

III. BIG DATA SECURITY AND PRIVACY CHALLENGES

I. Secure Computations in Distributed Programming Framework

Distributed programming framework utilize parallelism in computations and storage to process massive amounts of the data .A popular example is map reduce framework, which splits an input file into multiple chunks in the first phase of map reduce, a mapper for each chunk reads the data, perform some computation ,and outputs a list of key/value pairs. In the next phase, a reducer combines the values belonging to each distinct key and outputs the result. There are two major attack prevention measure: securing the manners and securing the data in the presence of an untrusted manner.

II. Security Best Practices for Non Relational Data Stores

Non relational data stores popularized by NoSQL databases are still evolving with respect to security infrastructure. For instance, robust solutions to NoSQL injection are still not mature each NoSQL DBs were built to tackle different challenges posed by the analytics world and hence security was never part of the model at any point of its design stage. Developers using NoSQL databases usually embed security in the middleware .NoSQL databases do not provide any Support for Enforcing it explicitly in the database. However, clustering aspect of NoSQL databases poses additional challenge s to the robustness of such security practices.

III. Secure Data Storage and Transaction Logs

Data and transaction logs are stored in multi-tiered storage media manually moving data between tiers gives the it manager direct control over exactly what data is moved and when. However as the size of data set has been and continues to be, growing exponentially, scalability and availability necessited auto tiering for big data storage management. Auto tiering solutions do not keep track of where the data is stored ,which poses new challenges to secure data storage. new mechanisms are imperative to thwartun authorised access and maintain 24/7 availability.

IV. End Point Input Validation/Filtering

Many big data use cases in Enterprise settings require data collection from many sources, such as end point devices for example, a security information and event management system (SIEM) may collect event logs from millions of hardware devices and software application in an enterprise network . A key challenge in the data collection process is input validation :how can we trust the data? how can we validate that a source of input data is not malicious and how can we filter malicious input from our collection? input validation and filtering is a daunting challenge posed by untrusted input sources, especially with the bring your own device (BYOD) model.

V. Real –Time Security/Compliance Monitoring

Real time security monitoring has always been a challenge ,given the number of alerts generated by (security)devices. These alerts (correlated or not)lead to many false positive , which are mostly ignored or simply "clicked away", as humans cannot cope with the shear amount. This problem might even increase with the bid data given the volume and velocity of data streams however, big data technologies might also provide an opportunity, in the sense that these technologies do allow for fast processing and analytics of different types of data .Which in its turn can be used to provide, for instance, real time anomaly detection based on scalable security analytics.

VI. Scalable and Compos able Privacy-Preserving Data Mining And Analytics

Big data can be seen as a troubling manifestation of big brother by potentially enabling invasions of privacy ,invasive marketing, decreased civil freedoms ,and increase state and corporate control.

A recent analysis of how companies are leveraging data analytics for marketing purpose identified an example of how a retailer was able to identify that teenager was pregnant before her father knew. Similarlly anonym zing data for analytics is not enough to maintain user privacy. For example AOL released anonymized search logs for academic purposes ,but users were easily identified by their searchers .Netflix faced a similar problem when users of their anonymized data set were identified by correlating their Netflix movie scores with IMDB scores.

Therefore ,it is important to establish guidelines' and recommendations for preventating inadvertent privacy disclosures.

VII. Cryptographically Enforced Access Control And Secure Communication

To ensure that the most sensitive private data is end to end secure and only accessible to the authorized entities, data has to be encrypted based on access control policies. Specific research in this area such as attribute-based encryption (ABE)has to be made richer, more efficient, and scalable. To ensure authentication, agreement and fairness among the distributed entities, a cryptographically secure communication framework has to be implemented.

VII. Granular Access Control

The security Property that matters from the perspective of access control is secrecy-preventing access to data by people that should not have access .The problem with course-grained access mechanisms is that data that could otherwise be shared is often swept into a more restrictive category to guarantee sound security granular access control gives data managers a scalpel instead of a sword to share data as much as possible without compromising secrecy.

IX. Granular Audits

With real time security monitoring ,we try to be notified at the moment an attack takes place. in reality, this will not always be the case(e.g, new attacks, missed true positives). In order to get to the bottom of the missed attack ,we need audit information. This is not only relevant because we want to understand what happened and what went wrong ,but also because compliance, regulation and forensics reasons .in that regard ,auditing is not something new, but the scope and granularity might be different. For example, we have to deal with more data objects, which probably are (but not necessarily)distributed.

X. Data Provenance

Provenance metadata will grow in complexity due to large provenance graphs generated from provenance- enabled programming environments in big data applications. Analysis of such large provenance graphs to detect metadata dependencies for security/confidentiality applications is computationally intensive.

IV. SECURITY METHODS FOR BIG DATA

I. Type Based Keyword Search for Security Of Big Data.

1)Introduction: Big data provide many business opportunities to the information technology industry. Large scale applications of sensor networks, electronic health record systems, emails as well as social networks generate massive data each day. The volume of information collected and stored has exploded. Cloud computing system process extraordinary storage capacity and computation power and is promising to handle the big data processing system with its features. However, since the cloud data service provider system are distributed to share and process sensitive information assigned to them, a malicious data stealer might probably bring about serious privacy problems .

Data encryption technology is used for boost information privacy protection. However, traditional encryption primitives (such as symmetric key encryption and public key encryption) are not capable to ensure the usability and hinder even authorized users from searching several keywords of encrypted files **,** it is difficult for the users to retrieve desired information from encrypted big data. So It is necessary to explore new cryptographic primitives to provide data encryption and searchability for big data era. Searchable encryption technology could fulfill the requirements to realize operability and data confidentiality, simultaneously. In this method, we provide a novel keyword search method to enable customers easily searching keywords from encryption-protection data. Moreover, the encrypted big data could be managed by different type that was assigned by data owner. Moreover, the access right can be given to others according to the user's willingness

Researchers also explore new search patterns for searchable encryption, such as fuzzy search, subset search, rank search. The public key encryption with keyword search (PEKS) scheme was proposed in order to offers the user to retrieve files through keyword searching. Consider an electronic health record system. A user sends an encrypted file *m* appended with some encrypted keywords $w_1, w_2, \ldots. w_n$ that are extracted from the message to the data service provider.

The data are organized in the format

$PKE(pk_A, m) \,||\, PEKS (pk_A, w_1) || \ldots. \,||\, PEKS (pk_A, w_n)$,

in which pk_A is the public key of user. The user could generate a trapdoor that contains certain keyword W_i. After receiving the trapdoor, data service provider search the encrypted files and returns all files that contain W_i. Other researchers also try to extend searchable encryption scheme to multiple users.

*2)System model:*We will design a secure big data storage system that supports multiple users. In this system, authorized users are able to store encrypted data and carry out keyword queries on encrypted data without decrypting all the files. Moreover, data owners could delegate certain type of files to other users.

Data Service Provider: Data service provider is responsible to generate global parameter for the whole system. Its main responsibility is to store user's encrypted data, respond to user's retrieve request and return corresponding files. Moreover, a new functionality is provided: re-encrypt second level ciphertext to first level ciphertext on behalf of delegatee. We should point out that our scheme provides finegrained delegation authority management. In other reencryption based searchable encryption schemes, delegatee is capable to decrypt all files that belong to the data owner when delegation right is given. However, in this system, delegator could delegate a designated type of files to delegatee for decryption so that delegatee is only able to recover part of ciphertext of data owner.

Delegator: Delegator is usually the data owner and can issue the keyword search query. Only data owner has the right to update the encrypted file and the encrypted keyword index. The data file could be images, documents, videos, programs, etc. In addition, delegator is responsible to generate re-encryption key for delegatee.

Delegatee: Delegatee is responsible to generate its own private key and fulfill the delegation responsibility, i.e., to decrypt first level ciphertext on behalf of its delegator.

3) SECURITY ANALYSIS:

In this subsection, we discuss our type based keyword search for encrypted data from the following security requirements: data confidentiality, query privacy and query unforgeability. We assume that users' private keys are kept secret.

Data confidentiality: The meanings of information confidentiality in our scheme are three fold. Both the first level and second level cipher texts should be protected from both data service provider and malicious eavesdropper. Moreover, the curious data server and malicious adversary could not obtain any information about keyword from the encrypted index of keywords.

In our system, the second level ciphertext and index of keywords are enciphered before uploading to the data service provider through algorithm

Encrypt(m, pk_{Ri}, t, w) .

Since data owner's private key is kept secret, the data server could not get any information about the plaintext through illegal decryption operation without private key. The element r $\sum$ Z^*_p is chosen randomly to resist replay attack.

Query privacy: The meaning of query privacy here indicates that the protection of personal information of users and information which may be recovered by malicious party from the keyword retrieve phase. In the keyword retrieval process, the user firstly generates a trapdoor for the keyword and sends it to the data server. In the whole process, curious data server could not get any privacy information about keyword *w* .

Query unforgeability: In this system, an individual private key is utilized to encrypt keywords by each user. Various keyword trapdoor queries , generated by different users' secret keys *Risk* are distinctive. In multi-user big data system, no user can create a spurious trapdoor query on behalf of another illegal user. Thus, the query unforgeability is offered in this system.

In this paper, we construct a type based searchable encryption scheme to secure big data, which also allows reencryption function. The plaintexts are generated with respect to a certain type. The security analysis shows that our scheme could provide data confidentiality, query privacy as well as query unforgeability.

II. Achieving Big Data Privacy via Hybrid Cloud

1) Introduction: With the rapid development of electronic and communication technology, the amount of data produced by medical systems, surveillance systems or social networks has been grown exponentially, which makes it hard for many organizations to cost-effectively store and manage these big data. Cloud computing, a new business model, is attractive, provides the advantage of reduced cost through sharing of computing and storage resources. However, concerns in term of the privacy of data stored in public cloud have delayed the adoption of cloud computing for big data . On one hand, a large amount of image, such as medical systems or social networks, may contain sensitive information. On the other hand, Cloud Service Providers (CSPs), who own the infrastructures on which clients' data are stored, have full control of the stored data. Therefore, the data stored in public cloud may be scanned by CSPs for advertisement or other purposes. Furthermore, attackers may be able to access data stored in cloud if there is not sufficient secure mechanism provided by CSPs. Most existing solutions employ traditional cryptographic algorithms, such as AES, to encrypt data and then store encrypted data in public cloud. However, for image data, which have much larger size than text data, heavy computation overhead will be introduced by this approach. Meanwhile, for the mobile devices, which have been widely used, much battery energy will be consumed, and it will increase delay because of the limited computation resources. Therefore, the traditional cryptographic approaches are not suitable for big data privacy.

In recent years, various image encryption algorithms have been proposed to speed up the process, among which the chaos-based approach with a substitution-diffusion layout appears to be a promising direction. In the substitution stage, the positions of pixels of the image are shifted via some chaotic map, and then the pixel values of the shuffled image are changed by chaotic sequences in the diffusion stage. However, the chaos system itself causes large computation overhead. Another approach is to take advantage of hybrid cloud by separating sensitive data from non-sensitive data and storing them in trusted private cloud and un-trusted public cloud respectively. However, if we adopt this approach directly, all images containing sensitive data or the ones that would not like to be seen by others have to be stored in private cloud, which would require a lot of storage in private cloud. Most users want to minimize the storage and computation in private cloud, and let public cloud do most of the storage and computation. To address the above challenge, we need to answer an important problem: How to efficiently achieve big data privacy by using hybrid cloud? Compared to using public cloud only, using hybrid cloud would have communication overhead between private and public cloud. Besides achieving data privacy, we want to reduce storage and computation in private cloud, as well as communication overhead between private and public cloud. In addition, the delay introduced by communications between private and public cloud should be small. In this paper, we present a scheme that can efficiently achieve image data privacy in hybrid cloud. A novel random one-to-one mapping function is proposed for image encryption, which makes the pair wise affinity among jigsaws unreliable and at the same time significantly speeds up the process of substitution and diffusion. Only the random parameters of the mapping function are stored in private cloud. . In this paper, we propose an efficient scheme for image data, which has much more volume than text data. We evaluate our scheme in real networks (including Amazon EC2), and our experimental results on image show that:

1)our scheme achieves privacy but only use 1/585.8∼1/398.6 the time of the AES algorithm; (2) the delay of our hybrid-cloudbased scheme is only 3%∼5% more than that of the traditional public-cloud-only approach.

2)System and Threat Model:

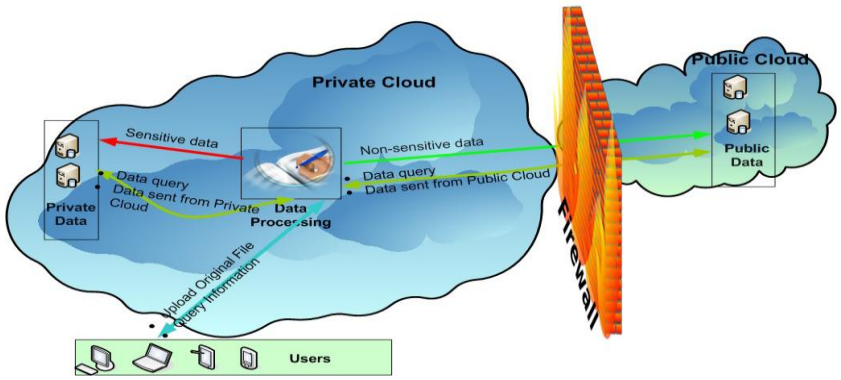

Fig. 4.2. The architecture of hybrid cloud

The architecture of a hybrid cloud is illustrated in Fig. 1. The original data come from private cloud, and are processed on servers within private cloud. If there are no sensitive data, the original data may be sent to public cloud directly. Otherwise, the original data will be processed to make no sensitive data leaked out. After being processed, most data are sent to public cloud, and a small amount of sensitive data are kept in private cloud. When a user queries the data, both private cloud and public cloud will be contacted to provide the complete query result. We consider an un-trusted public cloud who are curious and may intend to browse users' data. The public cloud has full control of its hardware, software, and network.

3)Design Goals

We want to protect image data privacy stored in public cloud via hybrid cloud. Specifically, we want to remove sensitive data and store them in trusted private cloud, and store the processed data (without sensitive information) in un-trusted public cloud. It would require too much storage in private cloud if we simply store the entire image with sensitive information in private cloud. Therefore, our design goal is to achieve image data privacy via hybrid cloud and at the same time reduce the following overheads: (1) the amount of data stored in private cloud, (2) the communication overhead between private and public cloud, and (3) the delay introduced by communications between private and public cloud.

To promote the cloud computing as a solution for big data, we proposed an efficient scheme to address the increasing concern of data privacy in cloud for image data. Our scheme divides an image into blocks and shuffles the blocks with random start position and random stride. Our scheme operates at the block level instead of the pixel level, which greatly speeds up the computation.

V. APPLICATION

Malicious URL filtering-Big Data Application

I. Introduction: Malicious

URLs have become a common channel to facilitate Internet criminal activities such as drive-by-download, spamming and phishing. Many attackers try to use these web sites for spreading malicious programs or stealing identities There is an urgent need to develop a mechanism to detect malicious URLs from the high volume, high velocity and high variety data. The blacklist is a general solution for protecting users from the malicious web sites, examples include PhishTank, SORBS and URIBL . These services provide a list of malicious web sites reported by volunteers or collected by web crawlers and verified by some reliable back-end system. The content-based analysis service BLADE is also a well known solution for detecting malicious web sites. They download the web page content and analyze it for malicious content. To download the content for analyzing is time consuming and consumes bandwidth. Therefore content based analysis services will be integrated with a blacklist or a cache mechanism to optimize the performance and avoid re-analyzing the same URL. However, content-based analysis methods are not a practical solution for the large volume of URLs and the speed at which new URLs can be created. We propose a filtering mechanism to be used before the content based analysis to remove the bulk of benign URLs, reducing the volume of URLs on which content-based analysis needs to be performed.

challenges:

• Large scale: the service provider averagely receives several million URLs every hour

• Extremely imbalanced data set: the malicious URLs are only around 0.01% of the total received URLs

• Sparse data set: the lexical features give us a very sparse data set To meet the goal of reducing the great amount of URL queries, the filtration system should be able to work in real time and achieve a high detection rate with a tolerable false alarm rate. The system should also be able to update the model using the feedback of the content-based analysis system.

In this paper, we introduce a framework to filter the no suspicious URLs using only features extracted from the URL string. We set the filtration as a binary classification problem. The proposed framework combines lexical information and static characteristics of URL strings together for classification.

II. Feature Extraction:

In order to successfully separate benign URLs from malicious ones, we design two feature sets to describe URLs. we design two feature sets to describe URLs.

TABLE I
COMPONENTS OF URL

Component	Example
URL	aneisig.es/vx/hstart.php?id=664&logon=141
Domain Name	aneisig.es
Path	vx/hstart.php
Sub-Directory	vx
Filename	hstart
File Extension	php
Argument	id=664&logon=141

TABLE II
THE DELIMITERS OF URL COMPONENT

Component	Delimiters
Domain Name	dash and slash
Path	dash, dot, underscore and back slash
Argument	ampersand and equal sign

These two feature sets represent the lexical information and some other characteristics of URLs.

A. Lexical Features

The lexical features use the bag-of-words model to describe URLs. We use three components for extracting lexical features: domain name, path and argument. Each component is split by specific delimiters. The details of delimiters are shown in table II. Additionally, we use a three character length sliding window on the domain name for generating the fixed length tokens. This method can identify a malicious website which subtly modifies its domain name.

Each word/feature generated from URLs is stored in a dictionary with a specific index, only the same word can get the same index. This dictionary could be dynamically updated with data streaming, but in practice the dictionary would consume a large amount of memory for the words generated from millions of URLs. For this reason, we use following methods to remove less useful words and to reduce memory usage.

1) Remove zero-weight words: Removing zero-weight words in the learned model is an easy and intuitive method.
2) Remove argument value words:
3) Replace IP address with an AS number:
4) Replace the digits in words with regular expression:
5) Keep the words generated in the last 24 hours only:

B. Descriptive Features

In descriptive features, we split the path component into sub-directory, filename and file extension to obtain more detailed information. To accurately extract the information on domain name, we remove the "www" (and www with any number), country code top-level domain (ccTLD) and generic top-level domain (gTLD). Removing these tokens can help us to focus on the remaining part chosen by the domain name owner. The descriptive features represent the static characteristics of the URL. These static characteristics are different from lexical features, which can be easily changed by slightly modifying the characters of URLs. We studied the URLs of phishing and malware sites to design these descriptive features. These features can help us to distinguish malicious and benign URLs.

1) Length: From our observation, the malicious URLs commonly need to add some keywords in its components. This action causes URLs to become noticeably lengthier. For detecting this situation, we record the length of each component. All these length values will be stored by the logarithm base 10 to avoid the normalization process being unduly affected by the large length value.

2) Length ratio: In addition to length, we also want to see the length ratio of different components. It can help to find the abnormal component. The length ratio feature uses the length of URL, domain name, path and argument component. We check all of the combinations of these components and compute the length division as follows:

- Domain Name divided by URL
- Path divided by URL
- Argument divided by URL
- Path divided by Domain Name
- Argument divided by Domain Name
- Argument divided by Path

Each pair of combination on the list contributes a descriptive feature in our feature set.

3) Letter-digit-letter and Digit-letter-digit: For detecting a phishing website masquerading as another website, we record the number of occurrences of the following two patterns: a letter between two digits (ex. award2o12) and a digit between two letters (ex. examp1e). They can help detect whether the URL is trying to deceive the user or not.

4) Delimiter count and Longest word length: We previously used delimiters to split the components in lexical features, now we create a descriptive feature by counting the number of delimiters in each component. The length of the longest word after splitting is also recorded.

5) Letter, Digit and Symbol count: Here, we categorize all characters as letter, digit and symbol and tally the frequency

We focus on a very practical problem, efficiently and effectively detecting malicious URLs.

The most challenging part of this problem is how to find an extremely small portion of malicious URLs out of a huge volume of URLs being generated at high speed. We proposed a novel method that uses the URL strings only for the detection. In the proposed method, we combine the lexical information and static characteristics of URL string. Without the host-based information and content-based analysis, we are able to deal with two millions URL strings in five minutes.

VI. CONCLUSION

We represented "Big data Security". Big data have various challenges related to security like-computation in distributed programming, security of data storage and transaction log, input filtering from client, scalable data mining and analytics, access control and secure communication. For tackling with such security challenges we used different security methods like Type Based keyword search for security of big data, use of hybrid cloud to provide privacy in big data .also we represent application of big data in malicious url filtering.

REFERENCES

[1] *"Cloud Security Alliance Top Ten Big Data Security And Privacy Challenges "by* CSA Big Data Working Group

[2] Yang Yang,Xianghan Zheng*"Type Based Keyword Search For Securing Big Data" in* 2013 International Conference On Cloud Computing And Big Data.

[3] Xueli Huang and Xiaojiang Du*"Achieving Big Data Privacy Via Hybrid Cloud"* in 2014 IEEE INFOCOM workshops:2014 IEEE INFOCOM workshop on security and privacy in Big Data

[4] Min-Sheng Lin,Chien-Yi Chiu,Yuh-Jye Lee and Hsing-Kuo Pao*"Malicious URL Filtering-A Big Data Application*" in 2013 IEEE International Conference on Big Data.

[5] Roger Schell"Security –*" A Big Question for Big Data"*in 2013 IEEE International Conference on Big Data

[6] Katina Michael, Keith W. **Miller "***Big Data: New Opportunities and New Challenges,"* Published by the IEEE Computer Society 0018-9162/13/$31.00 © 2013 IEEE

