ISSN:0975-9646

AN EFFICIENT DATA CU-RATION FOR BIG DATA CRYPTOGRAPHY

Jayanthi Vagini K[1], Manimekalai M[2] ,Ramesh Kumar B[3]

[1]*Assistant Professor, Department of Master of Computer Science, AJK College of Arts & Science, Coimbatore*
[2]*Assistant Professor, Department of Master of Computer Science, AJK College of Arts & Science, Coimbatore*
[3]*Head of the Department of Master of Computer Science, AJK College of Arts & Science,Coimbatore*

Abstract*: Mathematical cryptography hasn't gone out of style; in fact, it's gotten far more advanced. By constructing a system to search and filter encrypted data, such as the searchable symmetric encryption (SSE) protocol, enterprises can actually run Boolean queries on encrypted data. After that's installed, the CSA recommends a variety of cryptographic techniques. Relational encryption allows you to compare encrypted data without sharing encryption keys by matching identifiers and attribute values. Identity -based encryption (IBE) makes key management easier in public key systems by allowing plaintext to be encrypted for a given identity. Attribute-based encryption (ABE) can integrate access controls into an encryption scheme. Finally, there's converged encryption, which uses encryption keys to help cloud providers identify duplicate data.*

Keywords: *Big Data, Identity, Attribute, SAS technology, Encryption, and Cryptography.*

I. INTRODUCTION

Recent development of various areas of information and communication technologies has contributed to an explosive growth in the volume of data. These data sets used in the above example are so-called big data. This term refers to data sets that so large or complex that traditional data processing applications are inadequate. Big data usually has three V characteristics: volume (the quantity of generated and stored data may not be easily handled by con-ventional databases), velocity (the speed at which the data is generated and processed to meet the de-mands and challenges that lie in the path of growth and development), and variety (data generated are from multiple sources and in multiple formats) these techniques include file systems for big data [e.g., Hadoop distributed file system (HDFS)], noSQL databases (e.g., HBase), data processing models for big data (e.g., MapReduce), streaming techniques for big data (e.g., Storm), query engines (e.g., Impala), big data architecture (e.g., lambda architecture), and so forth..

The technology advances in big data are a) rapidly decreasing cost of storage in CPU b) the flexibility and cost effectiveness in datacenters c) the development of new upcoming framework.

II. REALTED WORK

A. *Drivers of Big Data*

Technical factors driving Big data adoption is done by three level. 1. Decreasing cost of storage 2. Flexibility and cost effectiveness of data centers and cloud computing 3. Development of new frameworks such as (Hadoop framework). Storage cost has dramatically decreased in the last few years. Therefore, while traditional data warehouse operations retained data for a specific time interval, Big Data applications retain data indefinitely to understand long historical trends. Big Data tools such as the Hadoop ecosystem and NoSQL databases provide the technology to increase the processing speed of complex queries and analytics. Extract, Transform, and Load (ETL) in traditional data warehouses is rigid because users have to define schemas ahead of time. As a result, after a data warehouse has been deployed, incorporating a new schema might be difficult. With Big Data tools, users do not have to use predefined formats. They can load structured and unstructured data in a variety of formats and can choose how best to use the data.

Big Data technologies can be divided into two groups: *batch processing*, which are analytics on data at rest, and *stream processing*, which are analytics on data in motion. Real-time processing does not always need to reside in memory, and new interactive analyses of large-scale data sets through new technologies like Drill and Dremel provide new paradigms for data analysis. Hadoop is one of the most popular technologies for batch processing. The Hadoop framework provides developers with the Hadoop Distributed File System for storing large files and the Map Reduce programming model, which is tailored for frequently occurring large-scale data processing problems that can be distributed and parallelized.

B. *Governance and Privacy*

It is also committed to also identifying the best practices in Big Data privacy and increasing awareness of the threat to private information. The preservation of privacy largely relies on technological limitations on the ability to extract, analyze, and correlate potentially sensitive data sets. However, advances in Big Data analytics provide tools to extract and utilize this data, making violations of privacy easier. As a result, along with developing Big Data tools, it is necessary to create safeguards to prevent abuse. In addition to privacy, data used

for analytics may include regulated information or intellectual property. System architects must ensure that the data is protected and used only according to regulations. The widely storage approaches could be categorized into three classes: local storage, external storage, and data-centric storage. Three different storage approaches where the cloud-like shape denotes the event, the arrows are the way sensor nodes relaying data, and the circles are the place storing data. Local storage refers to the sensor node which measures the physical phenomenon that stores the data. Then, some protocol should be defined in order to allow potential consumers of those data to find and access the nodes where they are stored.

The main drawback is Data-driven information security dates back to bank fraud detection and anomaly-based intrusion detection systems. Fraud detection is one of the most visible uses for Big Data analytics. Credit card companies have conducted fraud detection for decades. However, the custom-built infrastructure to mine Big Data for fraud detection was not economical to adapt for other fraud detection uses. Big Data analytics can be employed to analyze financial transactions, log files, and network traffic to identify anomalies and suspicious activities, and to correlate multiple sources of information into a coherent view.

III. PROPOSED METHODOLOGY

We want to protect against, we need to describe the security goals. The three most fundamental security goals are confidentiality, integrity, and availability, through attributes and identity. Confidentiality: Confidentiality is the goal of keeping all sensitive data secret from an adversary. More formally, traditional definitions of confidentiality guarantee that an adversary should learn no information about the sensitive data, other than its length. Confidentiality is critical in big data applications to guarantee that sensitive data is not revealed to the wrong parties. Integrity: Integrity is the goal that any unauthorized modification of data should be detectable. That is, a malicious adversary should not be able to modify such data without leaving a trace. This is very important to help guarantee the veracity of data collected in big data applications. Availability: Availability is the goal of always being able to access one's data and computing resources. In particular, an adversary should not be able to disable access to critical data or resources. This is a very important security goal in big data processing, as the sheer volume and velocity of the data make guaranteeing constant access a difficult task.

A. *Basic Cryptographic Tools*

Encryption: The main tool for guaranteeing confidentiality of data is data encryption. Encryption takes a piece of data, commonly called the plaintext, together with a crypto-graphic attributes and identity to produces a scrambled version of the data called the cipher text. Using the attribute and identity it is possible to decrypt the data to recover the plaintext, but without the attribute and identity the cipher text hides all information about the original data, other than its length. This security property, commonly known as semantic security which guarantees that, without the attribute and identity, an adversary cannot learn any (potentially sensitive) property of the underlying data even if he has a lot of insight as to what the data may be.

Enc(a; p) = c - an encryption algorithm that uses attribute a to scramble the plaintext p into cipher text c,

Dec(i; c) = p - a decryption algorithm that uses the identity i to recover the plaintext p from the cipher text c.

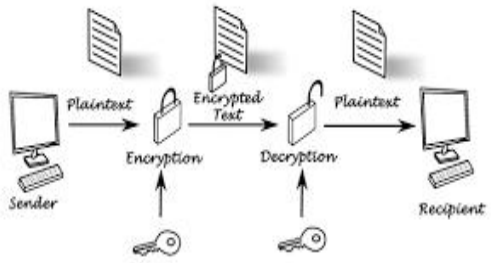

Fig. 1 Encryption and Decryption

B. *Batch Processing*

Map reduce in Hadoop: Map Reduce is a programming model for processing and generating large data sets. It contains two main processes: (1) map (k, v) -><k^0, v^0> and (2) reduce (k^0, <v^0>*) -><k^0, v^0>. The map takes input as key/ attribute value pair and produces another intermediate key/ identity value pair. On the other hand, Map Reduce is used to aggregate/summarize data—for example, to count the number of words appearing in a document. The map operation breaks content into words:

```
map( attribute key, attribute value):
  //  key: document identifier
  //  value: full text in the document for each word w
      in
      value:

      Emit intermediate(w,"1");
A reduce operation adds up counts for each word w:
reduce(attribute key, Iterator values):
   //   key: a word
   //   values: a list of counts
        int Result = 0;
        For each v in values: result
        += parseInt(v);
        Emit (AsString(result));
```

C. *Attribute based Access Control*

The Key management based solutions such as above have an inherent limitation. In order to share data with a set of users, it is necessary to know the identities (and keys) of all the authorized users. This is problematic in large systems or in systems with several organizational structures (as is very common in big data architectures where the data is collected, stored, and used in different environments) as the data owner is unlikely to know the identities of all the authorized users. An alternative approach to access control in such settings is a technique called attribute-based access control (ABAC). In ABAC, data is encrypted together with a policy describing the attributes of users authorized to access the data. The users receive keys for the attributes they possess and are able to access the data if and only if those attributes are authorized. This allows for enforcing access to data without knowing the full set of users with the authorized attributes. For example, to encrypt data for analysis by NIH scientists, a data provider could encrypt the data with the policy (\NIH" and \scientist") and only someone possessing both these attributes would be able to decrypt the data.

One approach to ABAC that has gained traction in government and commercial uses is to have a trusted server evaluate the access policy over a user's attributes and grant or restrict access to data accordingly. However, this requires trusting the server to correctly administer these permissions and is problematic in scenarios where there is no such trusted entity, such as in outsourced storage. We instead focus on solutions for ABAC that do not require a trusted server to evaluate the access policies. Specifically, we present a powerful cryptographic technique known as attribute-based encryption (ABE) that can be used to solve this problem cryptographically. This approach therefore offers trustiness and reliability of sensed data. Cooperative data integration is used when independent sources are fused their data to produce a new piece of data. Cooperative data integration is suitable for the body sensor networks (i.e., data integration should be done in a centralized manner).

D. Role of SAS

SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market.

SAS Grid Manager allows organizations to create a managed, shared environment for processing large volumes of data and analytic programs. This makes it a perfect solution for managing multiple SAS users and jobs while enabling efficient use of IT resources and lower-cost commodity hardware.

SAS In-Database processing is a flexible, efficient way to get more value from increasing amounts of data by integrating select SAS technologies into your databases or data warehouses. SAS Scoring Accelerator takes models and publishes them as scoring functions inside a database. This exploits the parallel processing architecture offered by the database to achieve faster results.

SAS Analytics Accelerator for Teradata is designed for users who want to build predictive and descriptive models for executing directly within the database environment. In-database analytics shortens the time needed to build, execute and deploy models, improving productivity for both analytic professionals. They also help tighten data governance processes by giving analytic professionals access to consistent, fresh data for faster, more accurate results.

SAS In-Memory Analytics enables you to tackle previously unsolvable problems using big data and sophisticated analytics. It allows complex data exploration, model development and model deployment steps to be processed in-memory and distributed in parallel across a dedicated set of nodes. Because data can be quickly pulled into the memory, requests to run new scenarios or new analytical computations can be handled much faster and with better response times.

SAS High-Performance Analytics is the only in-memory offering on the market that processes sophisticated analytics and big data to produce time-sensitive insights very quickly. SAS High-Performance Analytics is truly about applying high-end analytical techniques to solve complex business problems – not just about using query, reporting and descriptive statistics within an in-memory environment. For optimal performance, data is pulled and placed within the memory of a dedicated database appliance for analytic processing. Because the data is stored locally in the database appliance, it can be pulled into memory again for future analyses in a rapid manner.

SAS Visual Analytics is a high performance, in-memory solution that empowers all types of users to visually explore big data, execute analytic correlations on billions of rows of data in minutes or seconds, gain insights into what the data means and deliver results quickly wherever needed. It analyzes massive amounts of data in parallel and enables retailers to identify and implement optimal pricing strategies. Retailers can quickly determine which products to mark down, how much to mark them down, and when and where to adjust pricing to maximize revenues.

IV. DATA CU-RATION

A. Performance Evaluation

Map Reduce programs are not guaranteed to be fast or a panacea for every problem. It is concluded that relational databases still have advantages for several scenarios. However, with the release of Apache Hadoop project, one of the most popular frameworks to support the Map Reduce paradigm, Map Reduce has been extensively adapted to deal with the big data challenge.

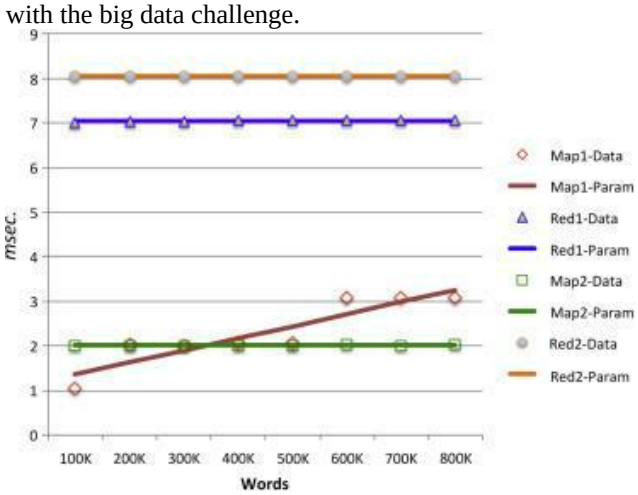

Fig 2. Map Reduce strategy work

New Big Data technologies, such as databases related to the Hadoop ecosystem and stream processing, are enabling the storage and analysis of large heterogeneous data sets at an unprecedented scale and speed. These technologies will transform security analytics by: (a) collecting data at a massive scale from many internal enterprise sources and external sources such as vulnerability databases; (b) performing deeper analytics on the data; (c) providing a consolidated view of security-related information; and (d) achieving real-time analysis of streaming data. It is important to note that Big Data tools still require system architects and analysts to have a deep knowledge of their system in order to properly configure the Big Data analysis tools.

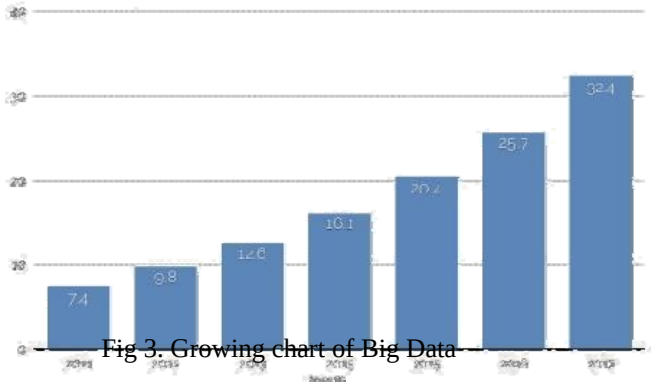

Fig 3. Growing chart of Big Data

Organizations are constantly seeking more effective ways to make decisions, relying increasingly on facts derived from a variety of data assets. But difficulties arise when data volumes grow ever larger and there are hundreds or thousands of decisions to make each day. Let SAS be the trusted adviser you turn to when you need to solve big data problems with big analytics. With high performance analytics options from SAS, you can use more granular data, perform more sophisticated analysis, better manage your IT resources and reduce the time to results, enabling faster, more confident decision making. The scalability of SAS to handle huge volumes of data is unsurpassed. And SAS Analytics is considered best in-class by both our customers and industry analysts. These advantages, combined with high-performance analytics, enable you to quickly exploit high-value opportunities from big data, while making the most of your existing investments or the latest advances in analytics infrastructure.

V. CONCLUSION

This section will discuss some popular technologies that were developed for the age of big data. And by this we start data storage, followed by batch data processing, streaming data processing and end with a discussion on popular architecture design.

.As the need to solve larger problems and tackle more complex scenarios evolves, high-performance analytics enables organizations to take advantage of the latest hardware advances and a variety of processing options to make the best use of all available resources with Hadoop framework.

The future works could aim at taking a good balance between centralization (get computation back to big data systems) and decentralization (put computation down to sensor nodes).

REFERENCES

[1] 2013 IEEE Symposium on Security and Privacy, SP 2013, Berkeley, CA, USA, May 19-22, 2013. IEEE Computer Society, 2013.

[2] Ariel Hamlin.Y, Nabil Schear.Y, Emily Shen.Y, Mayank Varia.Z, Sophia Yakoubov.Y, Arkady Yerukhimovich.Y.X, on Cryptography for Big data Security, Book chapter for Big data: Storage, Sharing, and Security(3S) December 17, 2015.

[3] Big Data Management on Wireless Sensor Networks, Chih-Chieh Hung, Chu Cheng Hsieh, Tamkang university, New taipei city, Taiwan Slice Technologies Inc., San Maeto, CA, United states.

[4] Manadhata, P.K., W. Horne, & P. Rao. (Forthcoming). Big Data for Security: Processing Very Large Enterprise Event Datasets. In B. Furht and A. Escalante (Eds.), Handbook of Big Data Analytics. s.l: Springer.

[5] Verizon Inc. (2010). 2010 Data Breach Investigation Report. Retrieved July 15th, 2013, from http://www.verizonenterprise.com/resources/reports/rp_20 10-data-breach-report_en_xg.pdf.

[6] An Efficient way to Gather Big Data in WSN using Mobile Sink Routing by vijayalaxmi, vol 4, Issue 8, august 2015.

Ben A. Fisch, Binh Vo, Fernando Krell, Abishek Kumarasubramanian, Vladimir Kolesnikov, Tal Malkin, and Steven M. Bellovin. Malicious-client security in Blind Seer: A scalable private DBMS. In IEEE Symposium on Security and Privacy, pages 395{410. IEEE Computer Society, 2015.

[7] Yarkin Dor•oz, Yin Hu, and Berk Sunar. Homomorphic AES evaluation using NTRU. IACR Cryptology ePrint Archive, 2014:39, 2014.

[8] Srinivas Devadas, Marten van Dijk, Christopher W. Fletcher, and Ling Ren. Onion ORAM: A constant bandwidth and constant client storage ORAM (without FHE or SWHE). IACR Cryptology ePrint Archive, 2015:5, 2015.

[9] SAS Analytics – General Survey Information.

[10] Toshniwal A, et al. Storm@ twitter. In: Proceedings of the 2014 ACM SIGMOD international conference on management of data. Utah, USA:ACM;2014.