Data Security in Cloud-Based Analytics

Charru Hasti and Ashema Hasti

Abstract Cloud computing platforms have grown in prominence in last few years, as they have made business applications and information accessible on the move without the need to purchase, set up, and maintain necessary hardware and software. The organizations are churning enormous gains due to scalability, agility, and efficiency achieved through the use of clouds. Data analytics involves voluminous data crunching to determine trends and patterns for business intelligence, scientific studies, and data mining. The incessant outburst of data from multiple sources such as web applications, social media, and other Internet-based sources motivate leveraging cloud technology for data analytics. Different strategies are being studied and incorporated to use the subscription-based cloud for serving analytics systems. The paper focusses on understanding the security threats associated with cloud-based analytics and approaches to cloud security assurance in data analytics systems.

Keywords Analytics as a Service (AaaS) · Multi-tenancy · Cryptography · VMWare · Trusted third-party auditor

1 Introduction

Cloud computing service technology inherently involves pooling and sharing of resources thus helping cut the costs of enterprises in the investment and maintenance of the technology and infrastructure. The delivery models encompass infrastructure as a service, software as a service, application cloud as a service, business process as a service, and now analytics as a service as well. Analytics is one of the areas getting enthused by cloud benefits today.

C. Hasti (✉)
Delhi Institute of Advanced Studies, Delhi, India
e-mail: charru.h1@gmail.com

A. Hasti
Mewar University, Rajasthan, India
e-mail: ashema.hasti@gmail.com

V.B. Aggarwal et al. (eds.), *Big Data Analytics*, Advances in Intelligent Systems and Computing 654, https://doi.org/10.1007/978-981-10-6620-7_11

Multiples of cloud vendors such as AWS and Cloud1010 offer services for a distributed data infrastructure and in-memory accelerating technologies for data access for the data analytics projects. Cloud vendors focus on one or more of the aspects viz. data models, processing applications, data storage, and computing power, to enable demand-based analytics. In-house business intelligence (BI) applications are capable of providing fast solutions to the users yet they require a lot of attention in terms of installation, execution, and maintenance. A cloud-based BI, on the other hand, is useful to the firms which otherwise cannot afford fully owned applications to analyze their data. Moreover, firms can extend the scale of data usage at much less cost. Remote provisioning of the warehousing facilities and sharing of resources through virtualization in areas such as social media analytics further enables focus on monetizing data via data exploration and gaining useful insights without concerning about maintaining and organizing exabytes of data.

2 Cloud Implementation Approaches in Data Analytics

Cloud framework involves three tiers: infrastructure, software, and platform. Infrastructure as a Service (IaaS) lets the companies such as Amazon and Google let the OS, storage, apps, and selected network components managed by the users. This has led to a cut down on server costs that the firms initially invested in a lot. Moreover, the exponential growth in the data has swamped the firms and made them think about diverting focus away from infrastructure through subscription-based IaaS. On the other hand, when an organization simply needs to get instantaneous and synchronized data and figures such as hits on an application, they simply employ a dashboard provided by a cloud vendor through Software as a Service (SaaS). The data dealt with can range from enterprise resource planning systems to inventory systems to customer relationship management systems. For instance, customer actions are captured as potential data to understand and analyze the behaviors and patterns in customers. This looks a bit scary in terms of customer privacy though is quite useful as well in order to make it customer friendly. Another opportunity that comes with cloud SaaS is that multiple users accumulating the data can benefit from each other by aggregating results from multiple sources. Platform as a Service (PaaS) cloud involves plugging in the BI or analytic capabilities into the applications for the customers. Independent software vendors such as Salesforce.com develop improved analytical applications as a software service or analytical engines as a platform service. These are then deployed by the client organizations to get an edge over others in the analytics market. These sources may provide structured forms of data or unstructured forms of data. The functional requirements of drawing useful trends and patterns through these data drive firms toward taking services from cloud vendors who focus on making data farms and warehouses for different needs. Therefore, the firm can build its own data center in the private cloud and follow the best practices of a public cloud environment or

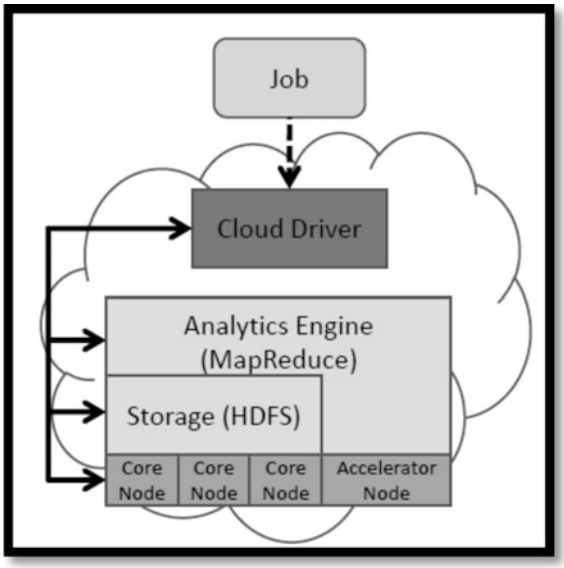

Fig. 1 Data analytic cloud architecture [12]

acquire a private cloud service from a cloud vendor. A cloud analytics service can also encompass organizations facilitating the consumption of analytic applications that have already been created [1] (Fig. 1).

3 Security Issues in Cloud Analytics

Several security risks and vulnerabilities have to be addressed by both the parties: providers as well as subscribers. In general, all information technology assets need to be safeguarded against the privacy threats irrespective of the service delivery model employed. Some of the common security risks that have been identified by researchers are listed below that encompass any cloud application [2, 3].

1. User access control and authentication
2. Data confidentiality conservation during data movement between clouds
3. Business continuity, recovery, and disaster management
4. Data location control
5. Data segregation by encryption
6. Controlled resource allocation
7. Investigative support for any illegal activity
8. Long-term viability of data
9. Virtual machine monitoring and security
10. Regulatory compliance.

These issues encompass but may not only include concerns of cloud service users such as ambiguity in responsibility, loss of governance and trust, switching service provider, data leakage, and concerns of CSPs such as protection inconsistency, un-administered APIs, license risks and so on [4].

4 Current Strategies for Cloud Security Assurance

Several solutions have been under research in lieu of standardizing cloud security assurance protecting against one or a class of security threats. Many CSPs, as well as research organizations, have been working toward achieving the trust of cloud users. Major work is focused toward identifying techniques to ensure data is protected from unwanted access by using appropriately designed encryption and decryption techniques. The areas such as secured browsers, proofs of the location of data, authentication of sensitive information are rigorously studied by researchers worldwide. The role of a third-party auditor has also been found over various studies as essential to maintain dynamic auditing control as well as compliance.

Boyd [5] has explained the role of cryptography in securing data. The techniques reviewed are searchable encryption, homomorphic encryption, proofs of storage, and proofs of location. These methods enable data encryption, integrity, location check. The traditional cryptography algorithms may not be useful to cloud users since the data is entrusted upon, stored, and processed at a distant location. Various cryptographic measures have been reviewed by Boyd to provide data assurance to users without entrusting the cloud provider. Certain improved storage schemes have been researched, for instance in the works of Shacham and Waters [6]. Proofs of location are discussed by Boyd further for achieving assurance of location to the users that compute the total time taken by data to reach the user from the requested site. Cloud Geolocation protocols were also exemplified that determine file locations on clouds using distance bounding protocols, based on the fact that an economically rational provider does not want to incur an additional cost of saving additional copies of the same file in multiple places. Distance Bounding Protocols estimate data transmission delays on the Internet. Proofs of Redundancy have also been shown as significant. These techniques are also based on timing methods and estimate delay in retrieving data. The research has suggested a geographic separation of multiple copies in order to provide assurance of more robust backup. Methods to process encrypted data that have been considered include searchable encryption (enabling server to identify requested file among all stored data without knowing contents of the file), homomorphic encryption (using such an encryption scheme that allows computations to be performed on data without the need of decrypting it), and homomorphic authentication (authenticating the output of computations on encrypted data).

Sharma and Gupta [2] have discussed the need of designing an architecture-independent policy that works well with all delivery models and cater to the underlying requirements of an organization or business purpose. They

explained certain security framework policies that ensure the cross-domain accesses. These included Security Assertion Markup Language (SAML) developed by OASIS, designed to solve Single Sign-on problem; eXtensible Access Control Markup Language (XACML), part of XML, used to set up rules according to the policies in order to check the authorization requests; OpenID decentralized authentication protocol used for authenticating a user to access many web applications through a single username and password; and WS-Security, WS-Trust, WS-Secure Conversation, WS-Federation, WS-Security Policy, Privacy and Identity Management for Europe (PRIME) that maintains and protects user information through a single console. Besides these policies in place, it was shown that a trusted third-party security analyst (T2PA) is required to confirm data consistency and integrity during inter-cloud data movement. This mechanism was concluded to be easy to implement by the data owner and reduced computational cost for the customer who cannot afford high-end security mechanism in place. Additionally, the traditional encryption and decryption algorithms were proposed to be deployable such as XML file signature in each file and AES (Advanced Encryption Standard) keys. AES algorithm has been found to be the most efficient in terms of speed, time, and throughput for cloud services. The trusted third-party analyst is used to resolve any kind of inconsistency between cloud service provider and client.

Syam Kumar and Subramanian [7] had proposed a protocol for dynamic data verification and remote data integrity checking in which a consumer generates some metadata which can be used later by the user to challenge the server for integrity of certain file blocks through challenge–response protocol.

Sasireka and Raja [8] explained an efficient approach to prevent data mining attacks and hence ensure data security. The suggested approach involves three steps: classification of data as sensitive, fragmentation of data into chunks, and distribution of data to different CSPs depending upon the reliability of cloud provider and data sensitivity.

Numerous authentications, as well as encryption policies, are undergoing study that needs an improvement to balance different risks associated with multiple delivery models. The CloudTrust Protocol (CTP) specified by Cloud Security Alliance [9] is a step in the direction of formalizing the assurance policies. It entrusts the responsibility of data security onto the CSPs and it is meant to create transparency between the providers and users to gain digital trust on the cloud.

5 Mapping Security Techniques to Cloud Analytics

There is a trade-off between on-premise analytics software or self-maintained warehouses versus cloud empowered analytics. It is concerned about privacy and latency versus total cost of ownership, ease of use versus less functionality, and

private cloud within versus user's firewall versus highly scalable pool of cloud data centers. The efficacy of cloud-based analytics increases if the security concerns are addressed well. Many techniques and strategies have been recently proposed by researchers in enhancing security in cloud analytics.

In the initial efforts, data have been stored in a layered set of servers which classified data depending upon their security restrictions and functions or categorization such as structured or unstructured. Therefore, mechanisms were put in place to secure the data stored in these silos. However, the inherent characteristic of data analytics on cloud is an efficient aggregation of data and results from across multiple sources for carrying out analysis. So the traditional approaches are not able to work well with cloud analytics since these segregated locations with varying access controls make it infeasible to access and collect data together at once. A workaround is to allow access between these layers for data sharing but this too becomes difficult as well as expensive to manage when scale of databases increases. Access management also becomes a challenge since source of data reads and updates is unrecorded most of the times.

The privacy-preserving analysis technique, proposed by Naganuma et al. [10], can be used to analyze data in encrypted form to provide data security when performing big data analysis on third-party cloud servers. The core of the proposed method is a searchable encryption technique that permits searching of data in encrypted form and can be used for statistical or correlation rule analysis of encrypted data. Because this privacy-preserving analysis technique only requires encrypted data and encrypted queries, it reduces the risk in the event of unauthorized access or a data leak (Fig. 2).

Conceived by Booz Allen Hamilton and the US government, the Cloud Analytics Reference Architecture [11] tags the data with security metadata once it enters the data lake repository. This framework aims at securely storing, analyzing, and maintaining data on the cloud. Firms can perform tagging using off-the-shelf tools in order to attach metadata to data on the cloud. Still, there can be challenges related to legal and political policies of sharing and aggregating data. Such issues need to be handled by the parties involved along with the decision makers corroborating with them. Hence, all firms subscribing to cloud analytics can establish their practices and then use the security metadata tagged to each of the data as per their mutually agreed rules for managing security in terms of compliance, authentication, and configuration (Fig. 3).

The cloud service vendors are constantly finding ways to keep their customer's trust and ensuring no information leaks by adopting standard cloud security techniques. Still, a lot of effort is required to make analytics as a foolproof service addressing all the security risks.

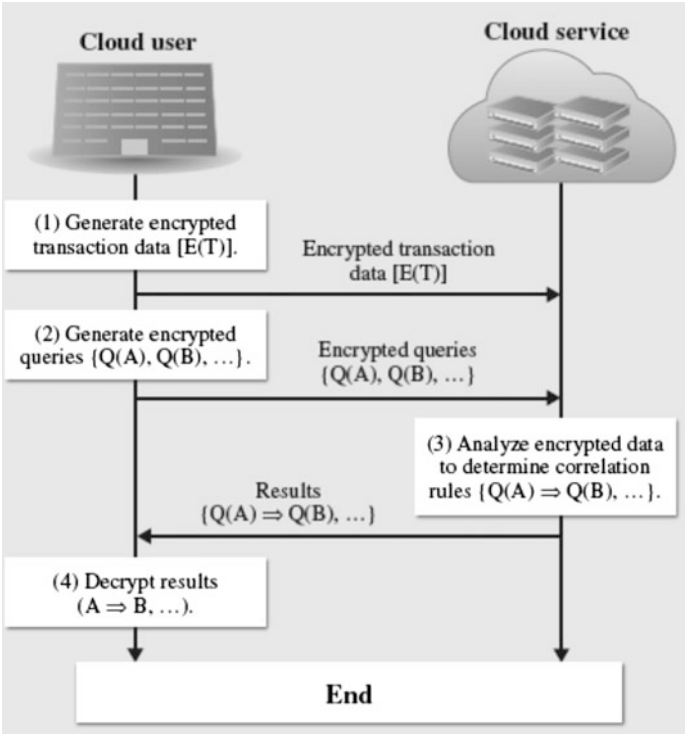

Fig. 2 Flowchart for correlation rule analysis of encrypted transaction data [10]

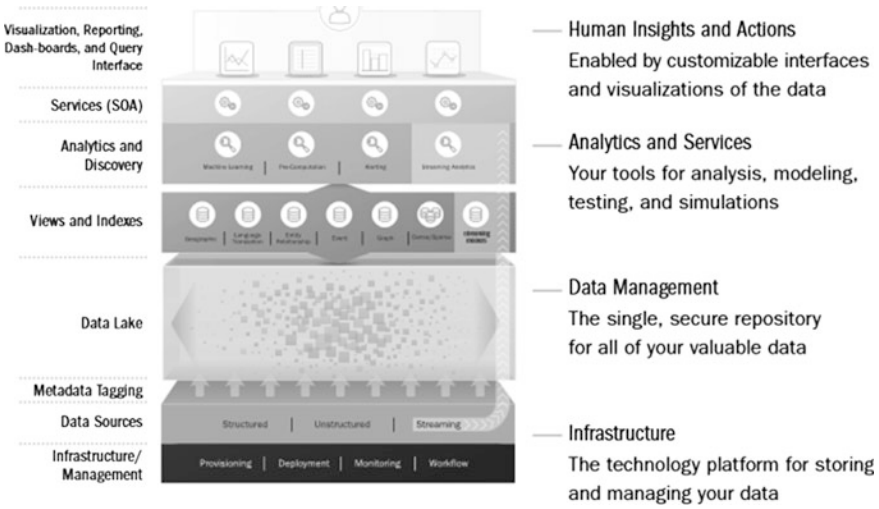

Fig. 3 Primary elements of the cloud analytics reference architecture [11]

6 Conclusion and Future Scope

Unification of analytics services and cloud computing has become a significant game changer. It is not only reshaping how firms deal with data number crunching, but also has started influencing how users make use of information technology. With numerous technical as well as monetary advantages associated with cloud analytics, there is an unquestionable need for meeting all security concerns due to relinquishing of data control as well as cross sharing of data lakes.

Efficient risk management systems need to be in place to ensure constant monitoring of data accesses and controls. The privacy-preserving analysis technique [10] and the Cloud Analytics Reference Architecture [11] are a few of the efforts in this direction. Yet there are multiple constraints to be addressed such as transparency, shifting away from more secure and expensive owned legacy systems. Any firm wanting to employ or provide cloud analytics must first analyze the benefits as well as security level that can be provided through the adoption of cloud-based security assurance techniques.

References

1. Schlegel, K., Sallam, R.L., Yuen, D., Tapadinhas, J.: Magic quadrant for business intelligence and analytics platforms. Gartner, 5 Feb 2013, G00239854
2. Sharma, P.D., Gupta, H.: An implementation for conserving privacy based on encryption process to secured cloud computing environment. Int. J. Eng. Sci. Res. Technol. (2014)
3. Hashizume, et al.: An analysis of security issues for cloud computing. J. Internet Serv. Appl. **4**, 5 (2013)
4. Jain, R., Singh, R.: A survey on current cloud computing trends and related security issues. Int. J. Res. Appl. Sci. Eng. Technol. **2**(I) 2014
5. Boyd, C.: Cryptography in the cloud: advances and challenges. J. Inf. Commun. Converg. Eng. (2013)
6. Shacham, H., Waters, B.: Compact proofs of retrievability. Advances in cryptology—ASIACRYPT 2008. Lecture notes in computer science, vol. 5350, (2008)
7. Syam Kumar, P., Subramanian, R.: An efficient and secure protocol for ensuring data storage security in cloud computing. Int. J. Comput. Sci. Issues **8**(6) (2011)
8. Sasireka, K., Raja, K.: An approach to improve cloud data privacy by preventing from data mining attacks. Int. J. Sci. Res. Publ. **4**(2) 2014
9. Cloud Security Alliance: https://cloudsecurityalliance.org/research/ctp/
10. Naganuma, K., Yoshino, M., Sato, H., Sato, Y.: Privacy-preserving analysis technique for secure, cloud-based big data analytics
11. http://www.boozallen.com/media/file/Enabling_Cloud_Analytics_with_Data-Level_Security.pdf
12. Leey, G., Chunz, B.-G., Katzy, R.H.: Heterogeneity-aware resource allocation and scheduling in the cloud. In: Proceedings of HotCloud, pp. 1–5 (2011)