International Journal of Research in Engineering and Innovation (IJREI)
journal home page: http://www.ijrei.com
ISSN (Online): 2456-6934

Research on amazon AWS cloud computing seller data security analysis under big data

Muhammad Talha[1], Mishal Sohail[2], Hajar Hajji[3]

[1, 2]*Department of Computer Science, Superior University Lahore, Pakistan*
[3]*School of Natural and Applied Science, Bahcesehir University, Turkey*

Abstract

US financial media CNBC reported that Amazon sent users an email on Wednesday saying that due to technical problems, some users' names and email addresses were leaked, but the problem has been resolved. An Amazon spokesperson said in a statement: "We have addressed this issue and notified potentially affected users. "Amazon not only declined to disclose how many users are currently affected or how long user information has been exposed, nor did it say where the data was leaked, so the outside world can't estimate the severity of the matter. In the field of network and information security, various challenges are being faced. On the one hand, the security architecture of enterprises and organizations is becoming more and more complicated, and multiple types of security data are increasing. The traditional analysis capabilities are weak. On the other hand, the emergence of new threats, the deepening of internal control and compliance, and conventional analysis methods have many shortcomings; more and more security information needs to be analyzed, and decisions and responses made more quickly. Information security also faces challenges from big data.

Keywords: Big data, AWS cloud computing, Amazon, seller data, data security.

1. Introduction

There is a lot of intersection between cloud computing and big data. The companies that mainly do cloud in the industry, such as Google and Amazon, have a lot of big data. Data experts emphasize that big data applications must run on cloud facilities. This is the relationship between the two-big data cannot be separated from the cloud. At the same time, the underlying principles supporting big data and cloud computing are the same, namely scale, automation, resource allocation, and self-healing. These are the underlying technical principles [1].

With the surge in the amount of information on the Internet, a user's single data set has reached terabytes, and some customers have even reached the Pera level (1000 Tera). The existing storage system structure is used to process data with a smaller amount of data, and it can only process data from a single data source, facing the pressure of big data. The ability to handle significant levels of data and multiple data sources is fragile. The challenge of big data is not only storage and protection. The strength of data analysis capabilities will become the critical point of this era: we have solved the problem of data storage and security. All it takes is time, but the issue of mass data analysis. We are not ready for big data [1].

2. Amazon Data Security Risks

In September 2018, Amazon announced that it had received a related report and said that it had launched an internal investigation. The focus of the inquiry was on internal employees receiving bribes to disclose company data and other confidential information, establishing an advantage for merchants who purchased the data. According to media reports, this practice is particularly evident in China, where the number of sellers is soaring. At the same time, the relatively low salaries of Amazon employees in China may encourage them to take risks. There are even Shenzhen Amazon employees who can provide internal sales data and email addresses of reviewers, as well as services to delete negative reviews and restore banned Amazon accounts, with prices ranging from $ 80 to $ 2,000.

Also, according to several foreign media reports, the latest

Corresponding author: Muhammad Talha
Email Address: talhashoaibt@yahoo.com
https://doi.org/10.36037/IJREI.2020.4207

disclosed court documents showed that in November 2018 Amazon requested a UK judge to approve the search of account information of Barclays Bank and MasterCard-owned Prepay Technologies. Amazon believes it is being attacked on a large scale by unidentified hackers who stole funds from Amazon seller accounts in the UK within six months last year [2].

The documents stated that the relevant criminal activities occurred between May and October 2018. Hackers broke into about 100 Amazon seller accounts and changed the seller's bank account to the hacker's account information on Amazon's Seller Center platform. The amount is unknown. Of third-party seller loans or sales, proceeds were remitted to Barclays and Prepay Technologies accounts.

An Amazon spokesperson told the media that fraudsters might have targeted sellers through phishing emails in an attempt to steal passwords and other data. However, Amazon did not explain how hackers could modify seller account information in the background. A Barclay's spokesman declined to comment specifically on the case but said the bank sought to close criminal accounts to help protect customers quickly.

3. Amazon's AWS Cloud Computing

Amazon uses AWS cloud computing to implement data analysis and mining, with a high level of security protection. The data is on the cloud, which avoids security risks and reduces the probability of leaking users' privacy. The data stored in the AWS cloud is much less likely to be lost and tampered with. AWS cloud technology is a leap forward for Amazon enterprise information security [3].

3.1 Amazon Data Integrity

In the process of data transmission and storage, cloud computing ensures that data is not tampered with by unauthorized users or can be quickly discovered by the system after tampering. This refers to ensuring the integrity of the data. AWS cloud computing provides security guarantees for data transmission through security technology so that no changes or changes are made to the data during the transmission process. At the same time, the identity of the sender and receiver of the data transmission can be confirmed [2].

3.2 Data Availability

The condition that data is not unusable due to hacking or physical equipment failure is called data availability. Generally, if you worry that essential data will be lost due to computer viruses and power failures, cloud computing will take data backup measures to prevent it. In the Amazon AWS cloud solution, a redundant backup method is used to ensure availability, which uses the parallel model of the system to improve system reliability [4]. AWS supports the following RDS

- MySQL 5.1, 5.5, 5.6, community Edition with InnoDB engine, Multi-AZ
- Postgre SQL 9.3, 9.4 9.5, Multi-AZ
- Maria DB 10.0.17, MySQL completed, Multi-AZ
- Avamar Aurora, Multi-AZ, compatible with MySQL and Postgre SQL, five times the performance of MySQL, is the database recommended by AWS
- Oracle 11g, 12c, there are three versions, all support Multi-AZ
- Standard One, Multi-AZ, Included License, KMS
- Standard, Multi-AZ, BYOL, KMS
- Enterprise, Multi-AZ, BYOL, KMS and TDE
- SQL Server 2008R2, 2012, 2014, there are four versions Express, does not support Multi-AZ, Included License, KMS Web does not support Multi-AZ, Included License, KMS
- Standard, Multi-AZ, KMS
- Enterprise, Multi-AZ, KMS and TDE
- Storage Options
- Magnetic, the worst performance, the cheapest, used in scenarios that require very little IOPS
- General Purpose SSD, general-purpose, can be used in most scenarios
- Provided IOPS SSD, the highest configuration, the most expensive, used in scenarios that require high IOPS

3.3 Two types of RDS databases

Online Transaction Processing (OLTP), a transactional database with high transaction requirements and high data consistency requirements

Online Analytical Processing (OLAP), an analytical database, with high requirements for computing and processing data and high read performance

RPO / RTO

The maximum amount of data that can be lost in the event of an RPO (Recovery Point Objective) accident

RTO (Recovery Time Objective) The maximum allowable downtime in the event of an accident

Backup/Recovery

Automatic backup, Amazon RDS will automatically create snapshots for DB's Storage to back up data, but each backup is only retained for one day by default and can be set to retain up to 35 days

Manual backup, the manual backup will not be deleted automatically

Recovery, the original DB will not be affected when recovering the DB, only a new DB instance will be created

High Availability with Multi-AZ

Except for SQL Server Express / Web does not support Multi-AZ, other DBs support Multi-AZ

Multi-AZ refers to creating a DB Instance on a different AZ. If the primary Instance dies, AWS will automatically transfer the connection to the secondary Instance without any user action.

Multi-AZ does not improve the performance of DB, just to increase HA

The master DB can be simulated by rebooting the master instance.

3.4 Data Privacy

All links of mass data transmission, storage and processing must protect personal user data and their information. Data privacy is

an essential dimension of information security. Amazon AWS cloud computing protects data privacy through authentication methods such as key technology, new algorithms, and encryption algorithms while enhancing the protection of the data itself. Data is encrypted at various stages of data transmission, storage, and processing. AWS uses cloud technology to process information to achieve information, hiding and protect user data security [5].
With the AWS Data Pipeline, customers can define data-driven workflows so tasks can rely on the successful execution of previous jobs. Customers can define parameters for data transformation, and AWS Data Pipeline will implement the logic set by the customer. The pipeline schedules and runs tasks by creating Amazon EC2 instances to perform defined work activities. The customer uploads the pipeline definition to the pipeline and then activates the pipeline. Customers can edit the pipeline definition of a running pipeline and reactivate the pipeline for it to take effect. Customers can deactivate the pipeline, modify the data source, and then reactivate the pipeline, which can be deleted after using the pipeline [6].
Task Runner polls the tasks and executes them. For example, Task Runner can copy log files to Amazon S3 and then launch an Amazon EMR cluster. Task Runner is installed and will run automatically on the resources created by the pipeline definition. Customers can write custom task runner application applications or use the Task Runner application provided by AWS Data Pipeline.

3.5 Subsequent Pages

For pages other than the first page, start at the top of the page, and continue in double-column format. The two columns on the last page should be as close to equal length as possible.

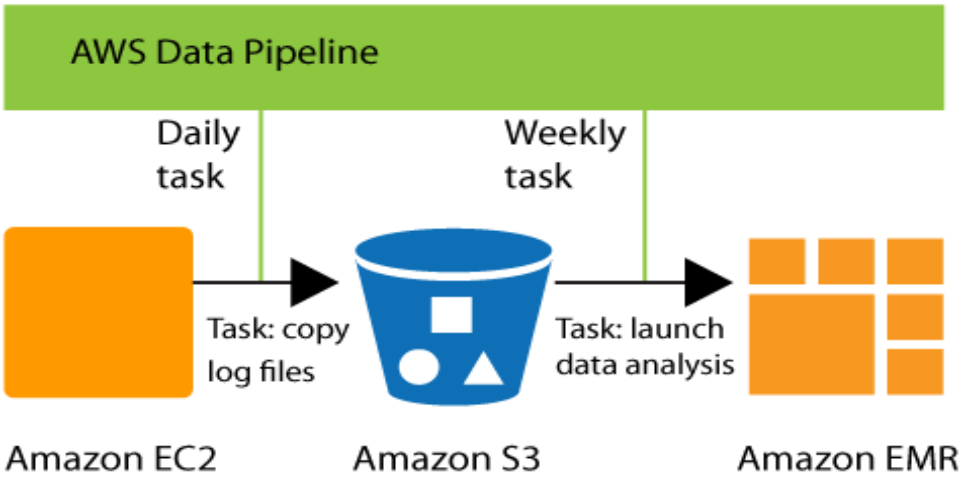

Figure 1: AWS Data Pipeline

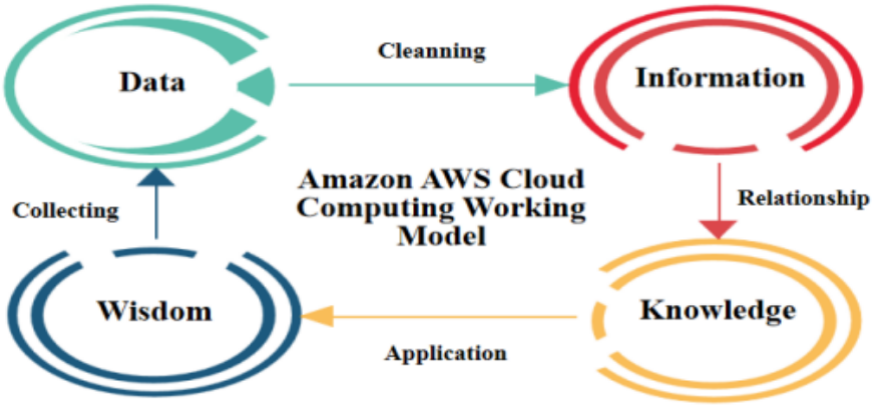

Figure 2: AWS Working Frame

4. Analysis of Hidden Dangers of Data Security in the Context of Big Data Cloud Compute

For big data, enterprises also need to consider how to deal with the risk of data leakage, and establish detailed plans, because big data has increased the requirements for analysis and calculation performance, but also brought more security risks. In the context of cloud computing, big data will use random forest algorithms in data security cracking [7]. Leo Breiman proposes the random forest. It uses bootstrap resampling technology to repeatedly extract k samples randomly from the original training sample set N to generate a new training sample set. Then, k classification

trees are produced according to the self-help sample set to form a random forest. The classification result of the new data is determined by the number of scores composed by the classification tree. Its essence is an improvement on the decision tree algorithm. Multiple decision trees are merged. The establishment of each tree depends on an independent sample. Each tree in the forest has the same distribution. The classification error depends on each the classification ability of a tree and the correlation between them. Feature selection uses a random method to split each node and then compares the errors generated in different situations. The number of features selected can be determined by the inherent estimation errors, classification capabilities, and correlations that can be detected [7].

When using cloud computing, it is essential to understand that security will not target all workloads. AWS emphasizes this model as "secure sharing."Secure Sharing" only provides security guarantees for AWS's physical data centers (virtual machines, storage, and even security functions), and whether to implement security measures on AWS's infrastructure depends on the users themselves.

4.1 Enable Two-Factor Authentication or Multi-Factor Authentication (MFA)

Enabling two-factor authentication (2FA) is a general way to prevent hackers from intruding into your account. Two-factor authentication means that the user provides two forms of authentication when logging in to the system. For example, the user needs to enter the set password and the random verification code. A free multi-factor authentication service is provided for free in AWS. It adds an extra layer of protection in addition to the username and password. After AWS MFA is enabled, when users log in to the AWS website, they will be required to enter a username and password (first security element-known to the user), and an authentication code from their AWS MFA device (second security element-the user already has). These multiple elements combine to provide greater security for your AWS account settings and resources [8].

JavaScript can implement 2FA's real code.
First, install this module.

```
$ npm install --save 2fa
```

Then, a 32-bit character key is generated.

```
var tfa = require ('2fa');
tfa.generateKey(32, function(err, key) {
console.log(key);
});
// b5jjo0cz87d66mhwa9azplhxiao18zlx
```

You can now generate the hash.

```
var tc = Math.floor(Date.now() / 1000 / 30);
var totp = tfa.generateCode(key, tc);
console.log (totp); // 683464
```

The advantage of two-factor authentication is that it is much more secure than a simple password login. Various password cracking methods are invalid for two-factor authentication. Two-factor authentication is just a way to protect security. To ensure security, it is more important to protect the confidentiality of key information of the enterprise. AWS has many forms of guaranteeing critical information, including HSM (Hardware Security Mode), which can be installed on the user's premises firewall, and its purpose is to help manage enterprise-critical information.

4.2 Monitoring suspicious information

We must not only increase the barriers for hackers and unauthorized users to enter the system, but also ensure the intrusion of unauthorized users. AWS Marketplace provides some free tools that can help users prevent hackers and unauthorized users from invading [9]. At the 2013 AWS Summit, Cloud Trail (which is in beta) was released to help users monitor suspicious information and analyze availability. Cloud Trail can help users create API-logs, which mainly report some usage status of user accounts.
There are many tools to find suspicious behavior in the market. Skyfence is one of the information agent systems that mainly monitors the operation of AWS. Skyfence warns users when they notice unusual behaviors, such as when users log in at suspicious times and unique IP addresses. The seller's login IP is changed too frequently, which will trigger a KYC audit, which is also one of the AWS cloud computing security measures.

4.3 Preventing intrusion by unauthorized users

If you have a tool that detects suspicious behavior, the next step is to detect intrusions by unauthorized users. Skyfence's delegation system function can close the AWS account and verify its identity before unauthorized users can access the management console. When changing data in the AWS Cloud, it must be authenticated by an authorized user. In the Code Spaces' case, this feature could prevent hackers from deleting data in the AWS cloud [9].

4.4 Encryption

There are other ways to prevent hackers from damaging the system after hacking into an AWS account. For example, encrypting data information in the AWS cloud. There are many different encryption service providers in the AWS marketplaces, such as SafeNet and Vormetric, can provide a variety of encryption services. AWS provides encryption and some other services for the Simple Storage Service (S3), but these services can only block most intruders and cannot guarantee the protection of the entire system. At the same time, after a hacker has successfully invaded, encryption cannot prevent the hacker from modifying the data.

4.5 Application of firewall

The invasion of DDoS puts Code Space in a dangerous situation and devours Code Space's cloud step by step. Using a firewall is an advantageous way to prevent DDoS intrusions. For example, Barracuda and Alert Logic in the Marketplace can provide monitoring to avoid hackers and identify and block suspicious behavior..

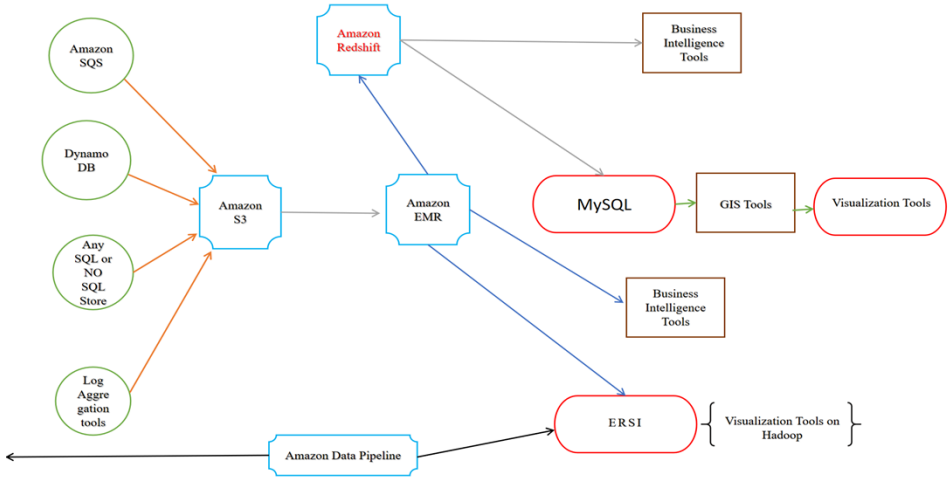

Figure 3: Amazon AWS Data Visualization Tools

4.6 Backup

Rob Ayoub of NSS Labs (the world's best-known independent security research and evaluation agency based in the United States) wrote in an AWS report that data backup is the best way to ensure security. Although backup cannot prevent hackers, data backup can make the database recover quickly.

If data is stored in the cloud, it is automatically backed up, which is a misunderstanding of the cloud by many people. Although this can be achieved in some services, backups will not be implemented in all services. For example, AWS's EBS and S3 are highly reliable. Because the AWS system will back up the data, which can ensure that the data will not be lost (after the user enters the management console, the data can be changed to make the built-in backup useless).For example, EC2 virtual machine instances are not automatically backed up. Therefore, when using the application, it is necessary to understand clearly what kind of guarantee each service will have [10].

If a hacker breaks into an account and causes damage, users can restore data from a backup. Users need to know what type of data they need to back up. Some companies will back up all their data, while others will only back up critical data. Some backups are data that is updated in real-time, while others can be backed up daily, weekly, and monthly or at any time according to user preferences.

AWS has many options for backup capabilities, including different storage methods and diverse database types, such as S3, EBS, and DynamoDB. A glacier is a service called "cold storage" at a little cost. However, compared to back up in the cloud, some users prefer to do backups in an internal environment.

4.7 Applying updates

AWS users have another misconception that applications in the cloud are automatically updated. Applications in SaaS can be automatically updated, but appeals in IaaS are not automatically updated. AWS provides essential application hosting services. It depends on the user's control of the virtual device. Many users fix bugs and update security through frequent software updates, and these features are only available on the latest version.

1. Each node is split into slices
-One slice per core
-DW1-2 SLICES ON xl, 16 on 8 XL
-DW2-2 SLICES ON xl, 16 on 8 XL

2. Each slice is allocated memory, CPU, and disk space

3.Each slice process a piece of the workload in parallel

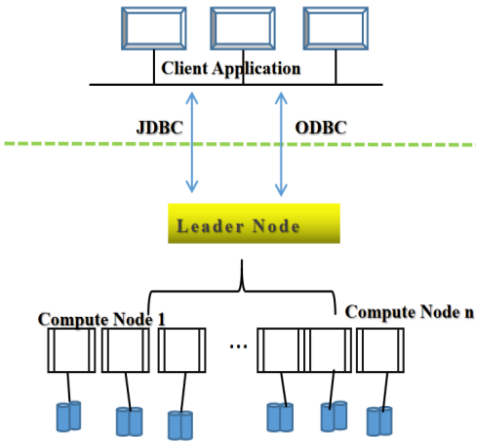

Figure 4: AWS Leader Node

The problem now is that many companies are unable to use suitable methods to secure their accounts. Although the cloud has many economic and practical advantages, for example, low cost, easy management and easy access. However, until a security issue is identified, no business will use the cloud to manage its data.

4.8 AWS Cloud Watch

AWS Cloud Watch can trigger operations including starting, terminating, restarting EC2, increasing or decreasing auto Scaling groups, and sending messages to SNS.AWS Cloud Watch supports monitoring and specifying metrics for most AWS services, including Auto Scaling, Amazon Cloud Front, Amazon Cloud Search, Amazon Dynamo DB, Amazon EC2, Amazon EC2 Container Service (Amazon ECS), Amazon Elasti Cache, Amazon Elastic Block Store (Amazon EBS), Elastic Load Balancing, Amazon Elastic Map reduce (Amazon EMR), Amazon Elastic search Service, Amazon Kinesis Streams, Amazon Kinesis Fire hose, AWS Lambda, Amazon Machine Learning, AWS Ops Works, Amazon Redshift, Amazon Relational Database Service (Amazon RDS), Amazon Route 53, Amazon SNS, Amazon Simple Queue Service (Amazon SQS), Amazon S3, AWS Simple Workflow Service (Amazon SWF), AWS Storage Gateway, AWS WAF, and Amazon Work Spaces.

4.9 AWS Cloud Watch monitoring frequency

Essential monitoring is collected every 5 minutes as a data point, and a limited number of indicators and supervision are provided for free. Detailed tracking is managed every minute as a data point, and signs can be customized. Supports finer-grained high-resolution indicators, collected every 1s. Cloud Watch supports cross-AZ aggregation and retrieval but does not support cross-region collection .CloudWatch can only monitor performance indicators and cannot track changes [8]

5. Conclusions

In addition to big data itself, in addition to data collection, collection, and aggregation of a certain amount of data, it is more critical in the process of data processing, mining, analysis, visualization, and application.

The topic around big data is basically around three issues: first, where the data comes from, second how the data is analyzed, and third how the information is commercialized. Any big data is application-oriented. In the future, the accurate mining of multi-dimensional and multi-complex big data will provide the most critical business solutions.

The three sources of data are government, corporate industry, and personal consumption. Government data was authorized, but government data was abused due to incomplete laws and other aspects. Consumer data comes from telecommunications, finance, or large BAT-like companies. The data at the traffic entrance will be automatically captured. The data provider can provide data in all dimensions, but each is local.

If data optimizers want to develop in the big data industry chain for a long time, they must be proficient in big data models, algorithms, and data characteristics, and at the same time have apparent sensitivity to the industry and ecology. And if algorithm providers rely on simple algorithms alone, they will become a weakness in the future. Application providers are closest to customers and are most familiar with customer needs. What they do at the same time is final data integration, which may have more room for development in the industry chain.

References

[1] Lin Gang. (2019). Research on Railway Intelligent Operation and Maintenance System Technology Based on Big Data Cloud Computing. Railway Communication Signals (5), 37-41.
[2] Xu Ziming. (2018). Data Security Analysis in Big Data Cloud Computing Environment. Electronic Technology and Software Engineering (20).
[3] Tang Zhuo, Chen Jianguo, Li Kenli, Lu Bin, Chen Junjie, & Xiao Jinbo: Random forest parallel machine learning method for big data in Spark cloud service environment.
[4] Zhang Xing. Energy saving analysis of thermal power plants based on Spark big data platform. (Doctoral dissertation, Taiyuan University of Technology).
[5] Hong Hanshu, Sun Zhixin,. Research on Security of Big Data Storage Based on Cloud Computing. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition) (4), 32.
[6] Xing Wei, Su Shengkui. Data Security Analysis in Big Data Cloud Computing Environment. Communication World (13), 25-25.
[7] Kong Lingtao, & Zhao Hui. Data Security Analysis in Big Data Cloud Computing Environment. Network Security Technology and Application (9).
[8] Liu Enjun. (2019). Data Security Analysis in Big Data Cloud Computing Environment. Network Security Technology and Application (5).
[9] Li Hao. (2019). Discussion on Data Security in Big Data Cloud Computing Environment. Shandong Industrial Technology (20), 129-129.
[10] Open Source Initiatives for Big Data Governance and Security: A Survey[J]. HU Baiqing,WANG Wenjie,Chi Harold Liu. ZTE Communications. 2018(02).

Cite this article as: Muhammad Talha, Mishal Sohail, Hajar Hajji, Research on amazon AWS cloud computing seller data security analysis under big data, International Journal of Research in Engineering and Innovation Vol-4, Issue-2 (2020), 124-129. https://doi.org/10.36037/IJREI.2020.4207.