Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 87 (2016) 128 – 133

Procedia Computer Science

4th International Conference on Recent Trends in Computer Science & Engineering

MetaCloudDataStorage Architecture for Big Data Security in Cloud Computing

Gunasekaran Manogaran[a], Chandu Thota[b]*, M. Vijay Kumar[c]

[a] VIT University, School of Information Technology and Engineering, Vellore, Tamil Nadu, India.
[*b]Infosys Hyderabad, Telangana, India
[c]Rayalaseema University, Kurnool, Andhra Pradesh, India.

Abstract

The cloud is increasingly being used to store and process the big data. Many researchers have been trying to protect big data in cloud computing environment. Traditional security mechanisms using encryption are neither efficient nor suited to the task of protecting big data in the Cloud. In this paper, we first discuss about challenges and potential solutions for protecting big data in cloud computing. Second, we propose MetaCloudDataStorage Architecture for protecting Big Data in Cloud Computing Environment. This framework ensures efficient processing of big data in cloud computing environment and gains more business insights.

Keywords: *Big Data Architecture, Big Data Security Architecture, Cloud Computing*

1. Introduction

1.1. Cloud computing

Cloud computing can defined as five attributes such as Massive Scalability, Multi-tenancy (Shared Resources), Elasticity, Pay as You go and Self-Provisioning of resources. Cloud computing enables user to access the remote servers hosted on the internet to store and process the data. Service models of cloud is classified into three types such as SaaS, PaaS, Iaas and different deployment models are classified into Private, Public, and Hybrid. Due to the high availability of cloud to all end users, cloud computing faces more security challenges. These challenges are classified into two broad categories as security issues faced by cloud providers and security issues faced by Customers.

* Corresponding author. Tel.: +91 8121813283;
E-mail address: chandutmca@gmail.com.

Peer-review under responsibility of the Organizing Committee of ICRTCSE 2016
doi:10.1016/j.procs.2016.05.138

1.2. Big Data and its applications

In general, Big Data is defined as a collection of huge size of data sets with different types so that it becomes difficult to process by using traditional data processing algorithms and platforms. Recently the number of data provisions has increased, such as social networks, sensor networks, high throughput instruments, satellite and streaming machines and these environments produce huge size of data. Big data used in many application health care [2][3], education, natural resources, social networking and so on.

1.3. Security challenges associated with big data in cloud computing

In order to secure big data, techniques such as logging, encryption, and honeypot detection must be necessary. In many organizations, the deployment of big data security framework is very attractive and useful. Big data analytics can be used to detect and prevent the malicious intruders and advanced threats [1]. Big data security in the cloud computing is essential due to the following issues such as: 1) To protect and prevent huge size of confidential business, government, or regulatory data from malicious intruders and advanced threats, 2) Lack of awareness and standards about how cloud service providers securely maintaining the huge disk space and erase existing big data, 3) Lack of standards about auditing and reporting of big data in public cloud, 4) Users who does not even work for the organization (malicious intruders), but may have full control and visibility into history of organization data (big data) [4]. Many researchers are developing big data security architecture and frameworks to protect huge data in cloud. For example preventing and detecting intrusion [5], Securing web application, Protecting confidential data in the cloud [6], Protecting GPS data from data mining based attacks and Securing the bank details in the cloud [7]. We proposed MetaCloudDataStorage Architecture for protecting Big Data in Cloud Computing Environment.

2. Proposed Framework

MetaCloudDataStorage security architecture is proposed in this section to protect big data against intruder. In this architecture organization data is stored in multiple cloud data centers based on the importance and scope. Data categorization is classified into three levels such as Sensitive, Critical and Normal. Each categorized data are supposed to be stored in different data center. Proposed MetaCloudDataStorage interface can efficiently redirect the user request to the appropriate datacenter in cloud provided by different vendors. AWS Cloud Trail is used in this proposed framework to process the log files. AWS Key Management Service (KMS) is integrated with AWS CloudTrail that delivers log files to an Amazon S3 bucket. CloudTrail can easily integrate with any application using proper API. AWS CloudTrail is capable of maintaining the time of the API call, IP address of the API caller, the request and response parameters the AWS service.

In this proposed framework datacenters will be separated into a sequence of n parts, where each part can be denoted by part i (i (1, n)), and they will be stored at m different storage providers, where each provider is identified as provider j (j (1, m)). In general, (parts of the datacenter) n is always far greater than (number of provide) m, these m storage providers belong to different organizations, such as Amazon, Google and Salesforce. Data parts are stored on certain cloud storage providers will be allocated to some physical storage media that belongs to the storage provider. When big data is stored in the datacenter it will form a unique storage path given as: Mapping Storage_Path = {Data((P_1(M_1,M_2 ... M_r))(P_2(M_1,M_2 ... M_s)) ... (P_n(M_1,M_2 ... M_t))};where, P denotes the storage provider and M denotes the physical storage media. Big data are always enormous and impossible to encrypt them as a whole, so propose framework encrypt the storage path of the big data, and get a cryptographic value which can be called cryptographic virtual mapping of big data. So instead of protecting the big data itself, proposed framework protects mapping of the various data elements to each provider using MetaCloudDataStorage interface.

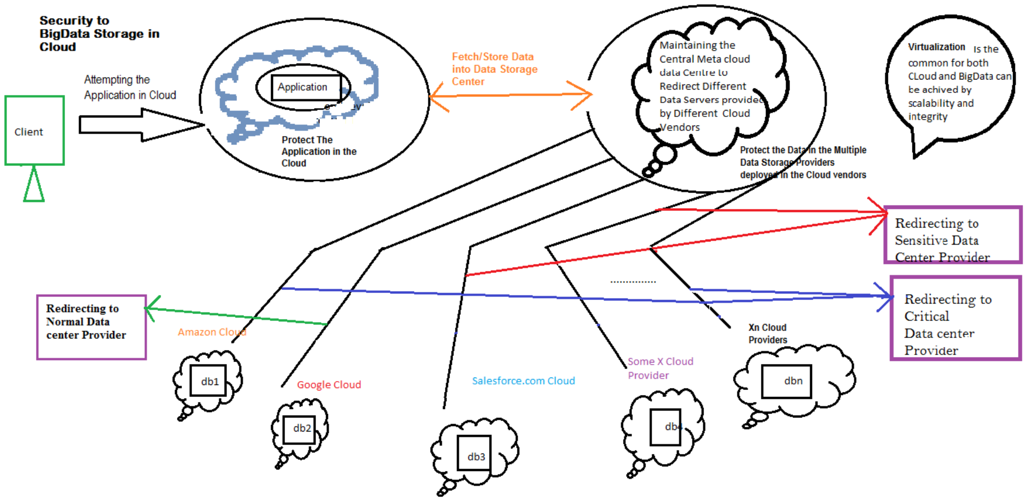
Security to BigData Storage in Cloud
Attempting the Application in Cloud
Client
Application
Protect The Application in the Cloud
Fetch/Store Data into Data Storage Center
Maintaining the Central Meta cloud data Centre to Redirect Different Data Servers provided by Different Cloud Vendors
Protect the Data in the Multiple Data Storage Providers deployed in the Cloud vendors
Virtualization Is the common for both CLoud and BigData can be achived by scalability and integrity
Redirecting to Sensitive Data Center Provider
Redirecting to Critical Data center Provider
Redirecting to Normal Data center Provider
Amazon Cloud
db1
Google Cloud
db2
Salesforce.com Cloud
db3
Some X Cloud Provider
db4
Xn Cloud Providers
dbn
....

Fig. 1. End User Accessing Applications and Data in Distributed Cloud

Although, the proposed framework will distribute all data parts in different storage service providers, and each provider holds some of the data parts. In order to provide high availability and robustness, the proposed framework will store multiple copies of same data on different cloud storage providers. Though big data is split and stored in different data center, the administrator of the entire system will keep the storage index information for each data parts. When there is a problem in some data parts on the cloud storage, propose framework can find another copy of the data parts according to their storage index information. Figure 1 shows how the end user will access the applications and data in distributed cloud. The proposed security algorithm which is shown below that protects the unauthorized access when trying to login into application which has been deployed in cloud. Although this algorithm updates the following tables such as 1) Threat_updated Table, 2) Meta Data Storage Cloud Table, 3) Amazon Cloud Data Storage Table, 4) Google Cloud Data Storage Table, 5) Xcloud Data Storage Table, 6) Xncloud Data Storage Table. Threat_updated table will store the entry related to malicious attempt, whereas Meta Data Storage Cloud table stores information regarding the data storage entry of different vendors. Critical, Sensitive and non-sensitive data are stored in other tables.

```
/*Process the Audit Log File to check the logged in user’s details like his/her/AI Machine track status, last
Logged in time etc.*/
Select customerid, password, last_loggedin_time, current status, geolocation, browser type, ipconfig, sysdate
from updated_audit_log;
//If any malicious user tried to login application
{
Insert into Threat_updated values (customerid, password, geolocation, browsertype, geolocation, sysdate);
return -1;
}
else {
List l = Select DataStorageproviderName, DataScope from MetaDataStorageCloud where
Customer_loggedin_ApplicationId=? ;
If (l.DataStorageproviderName ==’’amzoncloud” && l.datascope==”critical”)
{
Goto →Amazon Cloud Data Storage
Select * from AmazonCloudDataStorage where Customer_loggedin_ApplicationId=? ;
}
elseif(l.DataStorageproviderName ==”googlecloud” && l.datascope==”sensitive”)
{
Goto → google Cloud Data Storage
Select * from GoogleCloudDataStorage where Customer_loggedin_ApplicationId=? ;
}
elseif (l.DataStorageproviderName ==”xcloud”&& l.datascope==”normal”)
{
Goto →x clod data storage;
Select * from XcloudCloudDataStorage where Customer_loggedin_ApplicationId=? ;
}
}
}//end else
```

2.1. Security for Cloud Data and Applications

The security architecture for the MetaCloudDataStorage in Cloud is shown in the Figure 2.

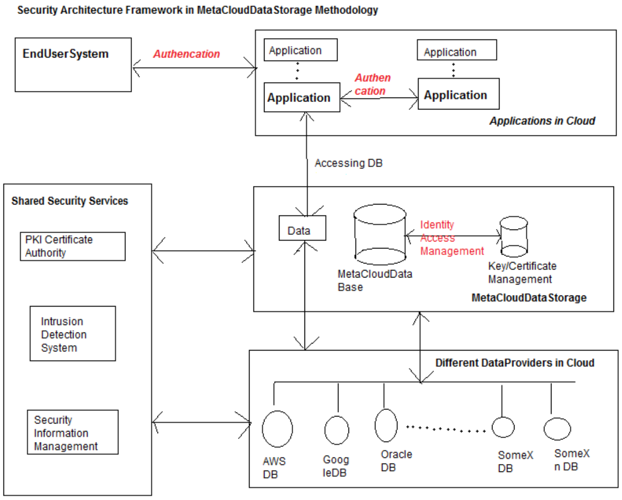
Security Architecture Framework in MetaCloudDataStorage Methodology
EndUserSystem
Authencation
Application
Application
Authen cation
Application
Application
Applications in Cloud
Accessing DB
Shared Security Services
PKI Certificate Authority
Intrusion Detection System
Security Information Management
Data
MetaCloudData Base
Identity Access Management
Key/Certificate Management
MetaCloudDataStorage
Different DataProviders in Cloud
AWS DB
Goog leDB
Oracle DB
SomeX DB
SomeX n DB

Fig. 2. Security Architecture for MetaCloudDataStorage in Cloud

2.2. Data Analytics: Map Reduce Algorithm for processing log files in distributed cloud data center

Map Reduce is a programming model or framework that process tasks in parallel across a huge size of systems. It contains two functions such as Map and Reduce. Map function splits the huge size of input data into <key, value> pairs. Intermediate <key, value> pairs will be created bases on aggregating several input key value pairs from the Map phase. Finally, Reduce takes the intermediate key value pairs and produces the output <key, value> pairs that can be easily understood by the end user. In this proposed architecture, Map Reduce framework is used to find the number of users who were logged in to the cloud data center. Proposed MapReduce pseudo code can efficiently process the huge size of log file in which it contains users who were logged in with date and the log in time duration. As shown in the Figure 3, the first process is map phase in which each date that represents the key is assigned a value of one initially. While reduce phase, the key values are summed up to find out the number of users logged in. For example, three users were logged in 01-02-2016, whereas two users were logged in 02-02-2016 [9].

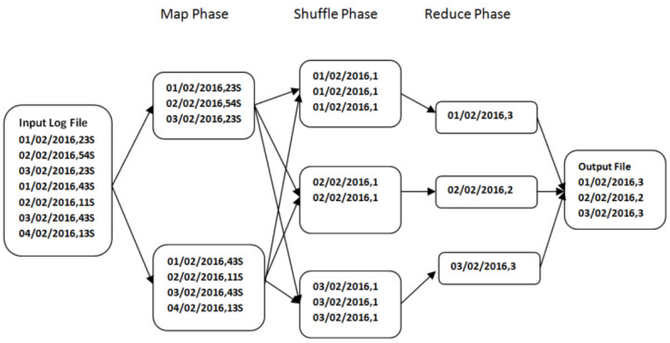

Fig. 3. MapReduce framework for processing log files

Mapper Function

```
public void Map(LongWritable key, Text value, OutputCollector output, Reporter reporter)
for each key ϵ value do
Emit(term key; count 1)
```

Reducer Function

```
public void reduce(Text key, Iterator values, OutputCollector output, Reporter reporter)
sum←0
for each v ϵ value do
sum←sum + v
Emit(key, sum)
```

2.3. Processing Big Data in Amazon Web Service with Apache Hive

Apache Hive is powerful open source software that runs on top of Hadoop in Amazon EMR. There is no special program or steps needed to create an interactive Hive Session in Amazon EMR webpage. Creating an interactive Hive Session does not require any steps to be added or configured in Amazon EMR webpage. In general, Hive and Pig are installed by default on every new Cluster of Amazon EMR. Once the user successfully connected with master node in the cluster, then it is easy to invoke the hive command directly on the EMR cluster [8]. It provides SQL-like queries to access the data. In this proposed architecture Hive is used to process the log files stored in the Amazon S3. Example comma-separated file (Log File) is shown below:

```
02:02:2016,43,address
02:02:2016,58,index
03:02:2016,33,sponsered
```

2.3.1 Storing Data in to AWS

```
Hive> CREATE EXTERNAL TABLE bankdetails(Date DATE, Time INT, Typeofdata
STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE
LOCATION 's3n://your_bucket/some_where/';
```

The EXTERNAL keyword is used to manage the table outside of Hive. It means that create and drop commands will not affect the data. However, overwriting data is possible at the location. Hence, Hive is used to read and process the data which are stored in Hadoop's HDFS or Amazon S3.The keywords ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' and STORED AS TEXTFILE represents the data in comma separated fields in a text file format. LOCATION keyword represents location of the data on S3 or HDFS.

2.3.1 Fetching Data from AWS

Hive commands are converted into a set of map and reduce jobs that are executed on cluster of computers against a distributed data set. WHERE clause is used to filter the huge data set in Hive and return the matching records. Select statement is used to obtain the number of users who were logged in during the specified time.

```
hive> select count(*) from bankdetails where Time >= 35;
```

3. Conclusion

In this paper we proposed MetaCloudDataStorage Architecture for protecting Big Data in Cloud Computing Environment. Map Reduce framework is used to find the number of users who were logged into the cloud data center. Proposed framework protects the mapping of various data elements to each provider using MetaCloudDataStorage interface. Though this proposed approach requires high implementation effort, it provides valuable information for cloud computing environment that can have high impact on the next generation systems. Our future work is to extend the proposed MetaCloudDataStorage Architecture for real time processing of streaming data.

References

[1] Victor, N., Lopez, D., & Abawajy, J. H.. Privacy models for big data: a survey. International Journal of Big Data Intelligence, 2016 *3*(1), 61-75.

[2] Lopez, D., & Gunasekaran, M. Assessment of Vaccination Strategies Using Fuzzy Multi-criteria Decision Making. In Proceedings of the Fifth International Conference on Fuzzy and Neuro Computing (FANCCO-2015) 2015: 195-208. Springer

[3] Lopez, D., Gunasekaran, M., Murugan, B. S., Kaur, H., & Abbas, K. M. Spatial big data analytics of influenza epidemic in Vellore, India. In 2014 IEEE International Conference on Big Data (Big Data), 2014, October: 19-24. IEEE.

[4] Thilagavathi, M., Lopez, D., & Murugan, B. S. Middleware for Preserving Privacy in Big Data. Handbook of Research on Cloud Infrastructures for Big Data Analytics, IGI Global, 2014.

[5] Marchal S, Jiang X, State R, Engel T. A Big Data Architecture for Large Scale Security Monitoring. In 2014 IEEE International Congress on Big Data (BigData Congress), 2014 Jun 27: 56-63. IEEE.

[6] Hongbing C, Chunming R, Kai H, Weihong W, Yanyan L. Secure big data storage and sharing scheme for cloud tenants. Communications, China. 2015 Jun;12(6):106-15.

[7] Subashini S, Kavitha V. A metadata based storage model for Securing data in cloud environment. In CyberC 2011 Oct 10:429-434.

[8] Schmidt K, Phillips C. Programming Elastic MapReduce: Using AWS Services to Build an End-to-end Application. O'Reilly Media, Inc.; 2013 Dec 10.

[9] Siddesh GM, Hiriyannaiah S, Srinivasa KG. Driving Big Data with Hadoop Technologies. Handbook of Research on Cloud Infrastructures for Big Data Analytics. IGI Global, 2014 Mar 31:232.