Proc. of the Intl. Conf. on Advances in Computer Science and Electronics Engineering -- CSEE 2014

ISBN: 978-1-63248-000-2 doi: 10.15224/ 978-1-63248-000-2-32

Predictive Data Analysis in Cloud using Big Data Analytic Techniques

Mr.S.Ramamoorthy[1] Dr.S.Rajalakshmi[2] Ms.R.Poorvadevi[3]

***Abstract*:-All the companies are nowadays migrating their applications towards cloud environment, because of the huge reduce in the overall investment and greatest flexibility provided by the cloud. The Cloud provides the larger volume of space for the storage and different set of services for all kind of applications to the cloud customers. There is not much delay and major changes required at the client level. The large amount of user data and application results stored on the cloud environment, will automatically make the data analysis and prediction process very difficult on the different clusters of cloud. It is always difficult to process, whenever a user required to analyze the data stored on the cloud as well as frequently used service by other cloud customers for the same set of query on the cloud environment.The existing data mining techniques are insufficient to analyze those huge data volumes and identify the frequent services accessed by the cloud users. In this proposed scheme we are trying to provide an optimized data and service analysis model based on Map-Reduce algorithm along with BigData analytics techniques. Cloud services provider can Maintain the log for the frequent services from the past. The service history analysis on multiple clusters to predict the frequent service. Through this analysis cloud service provider can recommend the frequent services used by the other cloud customers for the same query. This scheme automatically increases the number of customers on the particular cloud environment and effectively analyze the data which is stored on the cloud storage**.

***Keywords*— Cloud, Cloud-Storage, BigData, Map-Reduce, Clusters.**

I. Introduction to Cloud Computing

Cloud computing, a new internet-based technology, has been widely envisioned as the most promising technology of IT enterprise. It manages and schedules the computing resources through network, and constitutes a large computing resources pool which can provide service to users on their demand. The network is called "cloud".

S.Ramamoorthy[1] , R.Poorvadevi[3]Assistant professor,

Department of Computer science and Engineering,

SCSVMV University, India .

Dr.S.Rajalakshmi[2],Professor,

SCSVMV University, India

Resources in cloud can be extended unlimitedly, got anytime, used on-demand and paid according to apply. This feature is often called using IT service as water or electricity.s It is a distributed processing, parallel processing and Grid Computing development. Together with this new technology, lots of business models which can be of "X as a service (Xaas)" where X could be infrastructure, platform, software etc [1]. The most representative commercial cloud platforms are Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage System [2], Google App Engine [3], and Microsoft Azure[4]. All of these service providers have achieved great success in business. The increasing network bandwidth and reliable yet flexible network connections make it even possible that users can now subscribe high quality services from data and software that reside solely on remote data centers.

cIn Cloud Computing Environment storage as a service is the one of the major services provided by the cloud. The processed data and applications of users, can be stored and maintained by the third party cloud service provider in the virtualized pools on the data centers. User needs to pay for the storage capacity which is consumed from the service provider. Users are provided an option to store the data on-premise or off-premise based on their data security concern. However cloud service provider can virtualized the resources depending on the user requirement of the storage on cloud.

A. *Cloud Data Maintenance*

User data stored on the cloud environment are divided in to number of chunks each chunks will be stored in the different cluster of different server on the cloud environment. Unique id will be provided for the data for the identity of the user. Based on the unique id the data can be distributed and maintained geographically in the different servers in different locations.

Cloud storage is built on the network computing environment. There are many benefits to move data into the cloud. For example, users do not have to care about the complexities of direct hardware management. But as users store their data in the cloud, data security is of concern.
Data analysis performed on the individual consumer and Identify the interest of the users by the Cloud Service provider on the data which stored on the Cloud storage.

II. *Problem Statement*

It consists of three different network entities which are users, cloud service provider and third party auditor [2]. Users are active participants. They have data to be stored in

Proc. of the Intl. Conf. on Advances in Computer Science and Electronics Engineering -- CSEE 2014

ISBN: 978-1-63248-000-2 doi: 10.15224/ 978-1-63248-000-2-32

the cloud and rely on the cloud for data maintenance and computation. Both individual consumers and organizations can be the users. Cloud service provider has significant storage space and computation resources to maintain the users' data. It also has expertise in building and managing distributed cloud storage servers and the ability to own and operate live cloud computing systems. The cloud service provider unable to predict the customer expectation and interest on the specific resources. Users who store their large data files in the cloud storage servers can be relieved the burden of storage and computation [3].

A. *Existing work*

Data mining Techniques is less accurate of data analysis because of huge number of clusters used to store data on the cloud environment. Insufficient techniques are used for the frequent services analysis. Classical database management systems are insufficient to manage the huge volume of data from the cloud customers.

Social media, medical, computer, and other data set have been growing tremendously. Data flow cannot be handled properly and the prediction with the existing relational database and other mining techniques are insufficient.

B. *Prior Related work*

Traditional Data mining Techniques are insufficient to analysis the continuous grow of data from the flow of stream from the multiple data collection. Hadoop Acceleration in an Open Flow-Based Cluster model can analyze the open flow of data from the different sources. It provide the advantage for the single node cluster model but inefficient for the multiple node clusters.[8]. As per Kanthaka, Big Data Caller Detail Record (CDR) Analyzer for Near Real Time Telecom Promotions is used as a key technique to provide the effective data analysis for the cloud frequent service to improve the competitive market among the number of cloud service providers.

The data collection and analysis can provide the better performance comparatively from the previous technique [9].

The data Storage and access control on the cloud database is effectively done by the Ensuring Data Storage Security in Cloud Computing by Cong Wang, Qian Wang and achieved data security among the number of clusters used to store the cloud user data.[3]

C. *Proposed Scheme*

Map-Reduce algorithm analyzes the different cloud clusters and recommends the client for the frequent set of services used by the other users for the similar type of task.

This will reduce the complexity and ambiguity of user to analyze the services provided by the cloud.

Instead of analyzing the different set of services on the cloud, user can directly select the suitable service from the provided option based on the recommended services by the cloud using frequent services used by other users.

It reduces the overall time required by the user to analyze the data and services on the cloud environment, and effectively user can retrieve and identify the data from the cloud storage environment. It provides sufficient knowledge to the user on the required services. This proposed scheme reduces the overall investment of the customer by selecting the optimized service from the cloud. It will also increase the profit of the Cloud Service provider and they can effectively compete with other Service providers.

III. Proposed Algorithm

The *Map* and *Reduce* functions of *MapReduce* are both defined with respect to data structured in (key, value) pairs. *Map* Reduce function have a pair of data with a type in one data in one particular area. It returns a list of pairs in a different domain:

$$Map(k1,v1) \rightarrow list(k2,v2)$$

The *Map* function is applied parallel to every pair in the input dataset. This produces a list of pairs for each iteration. After that, the Map Reduce framework collects all pairs with the same set of keys from all lists and groups them together, creating one group for each key. The *Reduce* function is applied to the set of keys returned by the map function earlier. Each group, which in turn produces a collection of values in the same domain:

$$Reduce(k2, list\ (v2)) \rightarrow list(v3)$$

Each Reduce call typically produces either one value v3 or an empty return, The returned values of all iterations are collected as the desired result list.

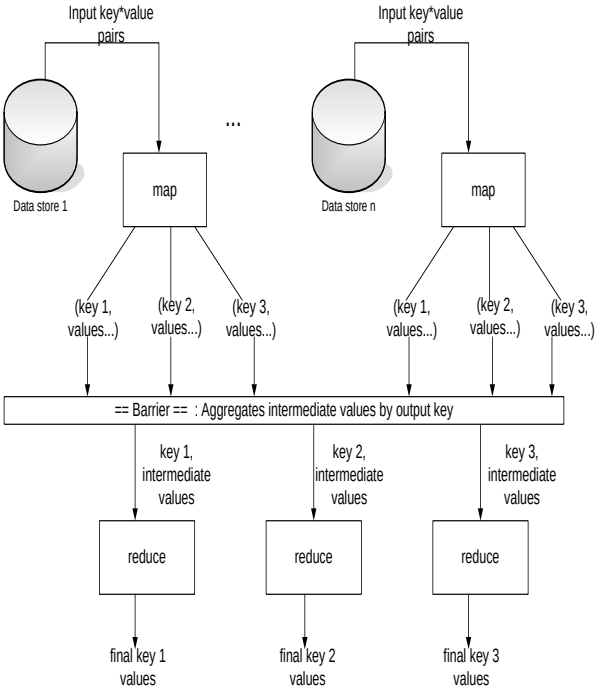
Thus the Map Reduce framework transforms a list of (key, value) pairs into a list of values. This behavior is different from the typical functional programming map and reduce combination, which accepts a list of arbitrary values and returns one single value that combines *all* the values returned by map.

Map Reduce function for Identifying the repeated number of Cloud services among the multiple cloud clusters:

```
function map(String name, String document):
  // name: cluster name
  // document: cloud document contents
  for each service s in document:
    emit (s, 1)
function reduce(String word, Iterator partialCounts):
  // service: a frequent service
  // partialCounts: a list of aggregated partial counts
  sum = 0
  for each pc in partialCounts:
    sum += ParseInt(pc)
  emit (service, sum)
```

Proc. of the Intl. Conf. on Advances in Computer Science and Electronics Engineering -- CSEE 2014

ISBN: 978-1-63248-000-2 doi: 10.15224/ 978-1-63248-000-2-32



3.1.Map Reduce Function

Estimated relative frequencies from the number of counts of the frequent services can be identified from the map-reduce function. Using this knowledge the cloud service provider can offer multiple numbers of services from the cloud environment based on the interest of different customers for the same set of services.

$$f(B \mid A) = \frac{count(A,B)}{count(A)} = \frac{count(A,B)}{\sum_{B'} count(A,B')} \quad (1)$$

Using the above equation the relative frequency of the frequent service can be identified from the different data set items.

A. *Big Data Analytics*

Map Reduce framework is partitioned in to six phases to process the given data set to find the frequent item set from the multiple clusters [5].

1) Input reader

The *input reader* divides the cloud user document input into appropriate size into either 16 MB to 128 MB and the framework assigns one split to each *Map* function. The *input reader* reads data from cloud storage system and generates key/value pairs.

2) Partition function

Each *Map* function output is allocated to a particular *reducer* by the application's *partition* function for sharing purposes. The *partition* function is given the key and the number of reducers and returns the index of the desired *reduce*.

A typical default is to hash the key and use the hash value modulo the number of *reducers*. It is important to pick a partition function that gives an approximately uniform distribution of data per shared for load-balancing purposes, otherwise the Map Reduce operation can be held up waiting for slow reducers (reducers assigned more than their share of data) to finish.

Between the map and reduce stages, the data is *shuffled* in order to move the data from the map node that produced it, to the shard in which it will be reduced.

Comparison function

The input for each *Reduce* is pulled from the machine where the *Map* ran and sorted using the application's *comparison* function.

3) Reduce function

The framework calls the application's *Reduce* function once for each unique key in the sorted order. The *Reduce* can iterate through the values that are associated with that key and produce zero or more outputs.

4) Output writer

The *Output Writer* writes the output of the *Reduce* to the cloud storage in the cloud cluster .

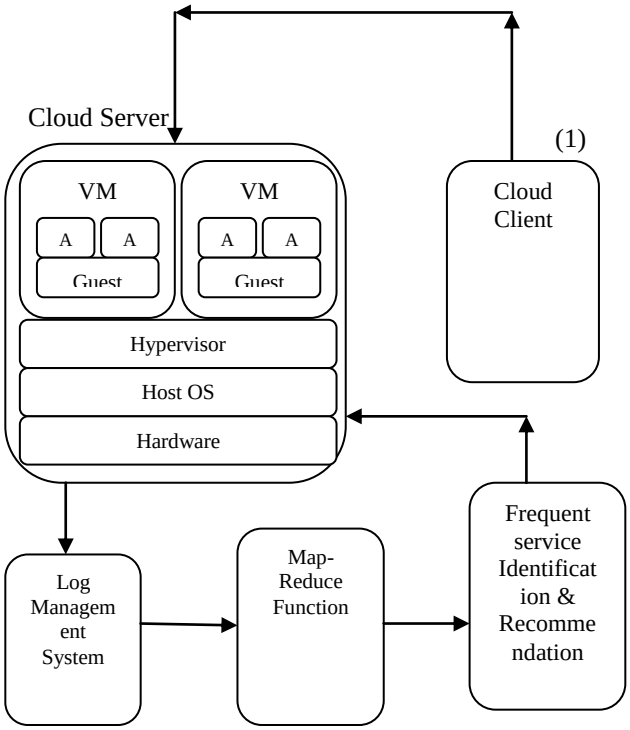


Figure 3.2 Proposed Architecture

B. *Definition and notation*

Each user is allocated a session key and the signature which are necessary for the cloud access control. Bilinear

Proc. of the Intl. Conf. on Advances in Computer Science and Electronics Engineering -- CSEE 2014

ISBN: 978-1-63248-000-2 doi: 10.15224/ 978-1-63248-000-2-32

Diffie-Hellman protocol is used to exchange these keys and signatures between users and cloud service provider.

TABLE A. DESCRIPTION OF NOTATION

S.NO	Notations Used and Definitions	
	Notation	***Explanation***
1.	CUID	Customer Identity
2.	SID	Server Identity
3.	SK	Session Key Identity
4.	SGID	Signature Identity
5.	QY	Request Query
6.	PWD	Password

For each data file, users add a message header before sending it to cloud. RSA is used to encrypt the data packet with the allocated keys[2].

In cloud service provider, there are very small number of servers which are responsible for keeping the whole access keys called trustful organization's servers. They are maintained by a trustful organization and cloud service provider and users cannot get any authentication information from them without a Specific Authentication module. All of the cloud storage servers have the specific Authentication module. It is used to authenticate the users. Also it has the ability to assign and update keys for the users. Each authentication module can communicate with each other and the trustful organization's servers[3]. Figure 3.2 gives the description of notation to be used in the scheme .

IV. Scheme Description

The interoperation between the users and the cloud storage servers in the proposed scheme are designed as follows.

System Setup In this operation, user sends a request to any one cloud storage server for the purpose to get the service from cloud. A Cloud Server checks the requested Query(QY) and Verification protocol gets the request and creates a pair of SK and PWD and a SID for the user uniquely. Then it sends QY,SK, PWD and SID of itself back to the user. Also it will store the requested Query in the Log management system to tracing the frequent services in the future. Now the request will be given to the Map-Reduce function to analyze the frequency of user using the services for the same set of Query. Through Bigdata analytics techniques the cloud server can identify the frequent service and recommend to the user for the feasible options [1] .

New Signature Creation before Login into Server:

The signature Creation of user deals with the following two steps:1)Create new self signature with the given session key and make the cryptographic digital signature with hash function .

2)Frequent set of service identification from the map reduce function

3)Recommend the user for the same set of services Encrypt the Signature and Session key (include H-x) with CUID. Then send the generated new Signature. The cloud storage server firstly checks up the H-x of the Signature and picks up the SID information. Then it searches the SID both in its SL and the trustful organization's servers. If the SID is not found, the server will discard the request. To the contrary, it will communicate with the SID server with the CUID in order to get the available information to the user.[2]

V. Performance Evaluation and Results

It is possible to run Hadoop on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service. Amazon Elastic MapReduce using Hadoop cluster run and terminate jobs. Data transfer between EC2 and S3 are automated by Elastic MapReduce. Apache Hive, which is built on top of Hadoop for providing data warehouse services, is also offered in Elastic MapReduce.

Let us consider the 'N' number of clusters used in the cloud environment, from this multiple clusters all the service request and data can be collected and it will be given as a input for the map reduce functions for the frequent service set identification. Apache Hadoop distributed File System(HDFS) provides the scalability and reliability among the clusters in the cloud environment.

Using apache Hadoop various clusters can be analyzed and frequent service set can be identified .

$$\text{Frequent service set} = \frac{\text{Number of repetition of the same service}}{\text{Single Cluster}}$$

$$\text{Average Frequent Cloud Service} = \frac{\text{Frequent Service set}}{\text{Total Number of Clusters in the Cloud}}$$

avg = 9/10 =0.9 for 9 out of 10 user using the same service

$$\text{Frequent service} = \frac{\text{Average number of users}}{\text{Total number of cloud Database}}$$

Suppose, most of the cloud users are interested for server machine software windows server 2008; then the cloud service provider can offer the winserver2008 for the same set of user request from the client. Using frequent data item set identified through bigdata map reduce function, cloud server analyzes the entire cloud database for the same set of services used by other clients. It will recommend to the user for most user frequent service and can rapidly increase the number of users in the cloud environment.

C. Result Comparison

Using apache hadoop karma sphere studio tool in netbeans IDE the Map Reduce function frequent item set values are tested.

This will improve the performance for the multiple clusters on cloud which can be tested and simulated.

Proc. of the Intl. Conf. on Advances in Computer Science and Electronics Engineering -- CSEE 2014

ISBN: 978-1-63248-000-2 doi: 10.15224/ 978-1-63248-000-2-32

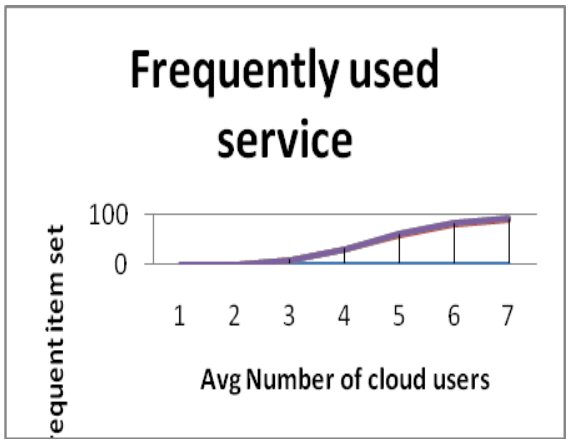


Figure.5.1 Frequent Service used in cloud Database

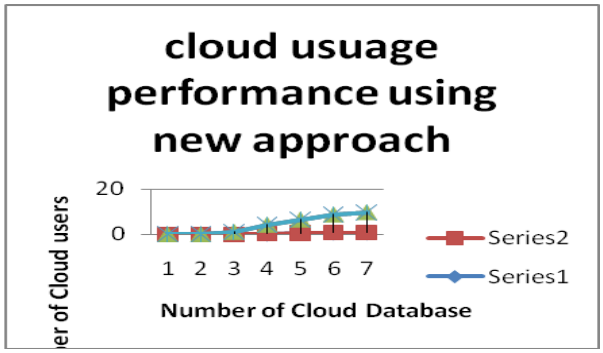


Figure.5.2 Cloud usage pcerformance using new technique

VI. Conclusions

This paper aims to provide an optimized data analysis technique on the cloud storage for multiple cloud clusters. Also proposed scheme is trying to analyze the trend resource by identifying frequent service used by the other users using Map-reduce algorithm. Utilizing the new scheme will increase the number of cloud users on cloud environment and also reduce the complexity to find frequent services in cloud computing, In the proposed scheme using Map-Reduce algorithm and BigData analysis techniques will improve the performance and profit of the cloud service provider compared to the existing techniques.

Acknowledgment

This work was supported by my research guide and different cloud computing international journals

References

[1] Shucheng Yu, Cong Wang, Kui Ren, Wenjing Lou, "Achieving Secure, Scalable, and Fine-grained Data Access Control in Cloud Computing", IEEE INFOCOM 2010 proceedings.
[2] Qian Wang, Cong Wang, Jin Li, Kui Ren, Wenjing Lou, "Enabling Public Verifiability and Data Dynamics for Storage Security in Cloud Computing", ESORICS 2009
[3] Cong Wang, Qian Wang, Kui Ren, Wenjing Lou,"Ensuring Data Storage Security in Cloud Computing", IEEE, IWQoS. 17th International conference 2009.
[4] Lifei Wei, Haojin Zhu, Zhenfu Cao, Weiwei Jia, thanasios V. Vasilakos, "SecCloud: Bridging Secure Storage and Computation in Cloud", IEEEICDCSW,2010
[5] Big Data for Development: From Information- to Knowledge Societies by Martin Hilbret,University of Southern California - Annenberg School for Communication; United Nations ECLAC,2013
[6] BigData Beyond the Hype, White paper,Datastax corporation 2012
[7] Synchronous Parallel Processing of Big-Data Analytics Services to Optimize Performance in Federated Clouds in Cloud Computing (CLOUD), 2012 IEEE
[8] Hadoop Acceleration in an Open Flow-Based Cluster By Narayan, Sandhya ,Bailey, Stuart, Daga, Anand in High Performance Computing, Networking, Storage and Analysis (SCC), IEEE 2012
[9] Kanthaka: Big Data Caller Detail Record (CDR) Analyzer for Near Real Time Telecom Promotions By Jayawardhana.,Kumara, Perera, Paranawithana, in Intelligent Systems Modelling & Simulation (ISMS), IEEE 2013
[10] A SLA-based method for big-data transfers with multi-criteria optimization constraints for IaaS by Chilipirea, Dobre, Pop,in. Roedunet International Conference (RoEduNet),IEEE- 2013
[11] A Solution for Privacy Protection in MapReduce by Quang Tran , Sato, H. in Computer Software and Applications Conference (COMPSAC), 2012 IEEE
[12] Stream processing with BigData: SSS-MapReduce by Nakada, Ogawa, Kudoh, T. In Cloud Computing Technology and Science (CloudCom), 2012 IEEE

About Author(s)

Professor Mr.S.Ramamoorthy. B.E(CSE), M.E(CSE),(phd). He is currently pursing P.hD in SCSVMV University, India. His areas of interest Cloud Computing, BigData, Network Security, Data Mining, Mobile Communication. He has got around 6 years of teaching experience in SCSVMV University. Published various papers and journals on Cloud Computing and Network security. Currently working in the project for Education Cloud.

Professor Dr.S.Rajalakshmi, Phd. Currently working as a Professor and Director for Advanced Computing Center in SCSVMV University, India. Having 20 years of teaching experience and headed the department for the past 10 years. Under her valuable guidance more than 15 members doing research in Computer science and Communication Technology.

Professor Ms.R.Poorvadevi B.E(CSE), M.E(CSE),(phd). He is currently pursing P.hD in SCSVMV University, India. Her areas of interest Cloud Computing, Data Mining and Artificial Intelligence. she has got around 5 years of teaching experience in SCSVMV University. Published various papers and journals on Cloud Computing and Network security.