ISSN No. 0976-5697

Volume 2, No. 4, July-August 2011

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

An Information Retrieval Framework for Cloud Computing Environment

Sanjib Kumar Sahu
Assistant Registrar
Guru Gobind Singh Indraprastha University,
New Delhi, India
sahu_ sanjib@rediffmail.com

Arti Gupta*
Student, M.Tech (IT)
University School of Information Technology (USIT),
New Delhi, India
artigupta_84@yahoo.com

Abstract: Cloud computing has gained significant attraction in recent years as it has provide many services as Paas, Iaas, Saas, Caas (Communication) and Daas (Storage) which has increased the use of IT services in the industry and also helpful in monetarily issues. Companies such as Google, Amazon, IBM and Microsoft have been building massive data centers over the past few years. These data centers tend to be built out of large number of computers managed by these companies, with a large amount of servers which tend to be extremely voluminous. The challenge is to provide integrity and security in a framework. In the light of above discussion this paper tries to develop delve deep into area such as framework for data centers to retrieve information without redundant data, its role and relevance to achieve security and integrity. To an extend we have also analyzed the recent research regarding data centers of clouds . All in all this paper present top Paas providers, recent work, a proposed data framework and its analysis for future computing environment.

Keywords: Cloud computing, Information Retrieval System, Data model, Framework.

I. INTRODUCTION

Cloud Computing promises an easy way to use and access to a large pool of virtualized resources (such as hardware, development platforms and/or services) that can be dynamically provisioned to adjust to a variable workload, allowing also for optimum resource utilization. Cloud computing can be defined as a new style of computing in which dynamically scalable and often virtualized resources are provided as a services over the Internet. Cloud computing has become a significant technology trend, and many experts expect that cloud computing will reshape information technology (IT) processes and the IT marketplace [1]. At the PaaS level, what the service provider's offer is packaged IT capability, or some logical resources, such as databases, file systems, and application operating environment. Currently, actual cases in the industry include Rational Developer Cloud of IBM, Azure of Microsoft and AppEngine of Google. Core technology is large-scale distributed application operating environment. It refers to scalable application middleware, database and file system built with a large amount of servers. The datasets managed by these systems tend to be extremely voluminous. It is not unusual for these datasets to be several terabytes.

In this paper we have discussed about the top data base providers, some related work and on the basis of that we have proposed an information retrieval model which can be implemented in any industry to retrieve data from cloud large database. As this paper include 3 sections, first section contain introduction and top cloud data center providers, second has literature review, third section has proposed architectural description and fourth section include data analysis in contrast with previous method.
Some of the Top database providers:

A. Force.Com

Salesforce.com has declared Database.com as its relational database service which offers stand-alone, cloud-hosted databases.

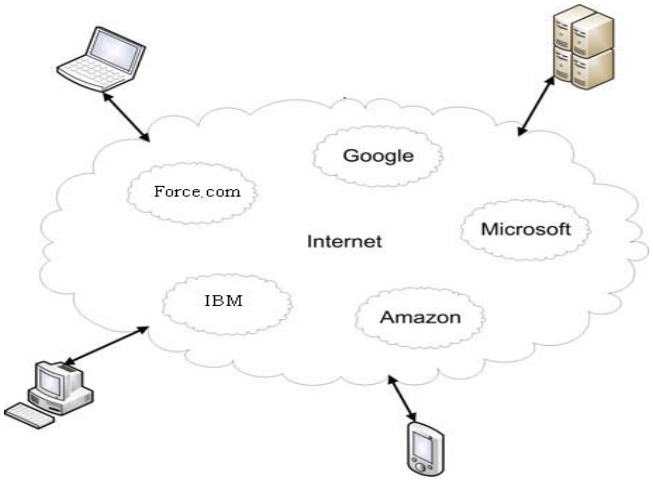

Figure 1: Some Common Clouds Providers

B. Microsoft SQL Azure Database

Azure provides an application platform which host and run Microsoft Datacenter, it also contributes in other services as development, management and hosting managed applications [2].

C. Google App Engine Data Store

Google is a huge service provider which develops and host web application managed by Google data centers. It basically use python language to write web applications and uses an API for storage, URL fetch, and email services. As Google stores all the data in a Table named as Bigtable which includes web indexing for retrieving of data. [3]

D. Amazon Elastic Block Store

The Simple Storage Service (S3) from Amazon is used by home users, small businesses, academic institutions, and large enterprises. It is an online storage web service which offers low latency, infinite data durability, high availability

and reliable storage volumes (from 1 GB to 1 TB) that can be attached to a running Amazon EC2 instance and exposed as a device within the instance to provide cloud infrastructure.

Amazon has another storage services as EBS (Amazon Elastic Block Store) is particularly suited for applications that require a database, file system, or access to raw block level storage. It help in prevent data loss while hardware failure by recovery of data. It has many users such as SmugMug, Slideshare and Twitter use it for hosting images; Apache Hadoop uses S3 to store computation data, and online synchronization utilities such as Dropbox [2].

E. Rackspace Cloud

Rackspace Cloud provides services including a cloud server, cloud files, and cloud site. The cloud files service is a cloud storage service providing unlimited online storage and a Content Delivery Network (CDN) for media on a Utility computing basis. It provide replica of 3 copies of data for multiple computer in multi zones top deal with security.

F. Cisco VFrame Data Center

It is a cross-functional, rapid services orchestration and provisioning platform, which offers a modular approach for deploying and configuring services over the data centers. Data centers provides orchestrated virtualized set of hosting services which facilitate with security, content load balancing and high availability data.

According to currently discussed cloud providers, some of the top Paas provider which works over different languages, infrastructure and application are mention in table 1.

Table 1: Top 4 Paas providers:

Basis	Paas provider			
	Azure	**Google**	**Force. Com**	**Heroku**
Programming	.Net	Python, Java	Apex programming and javal	Ruby
Infrastructure	Microsoft data center	Google data center	Saleforce.com	AmazonEC2 and S3
Application Host/ Application	Pin Point	Gigapan	The wall street	Ubermind
Execution Engine	Dryad	Google Map Reduce	Hadoop, Map Reduce	Elastic Compute Cloud (ECC), Elastic Map Reduce, Hadoop
Distributed data storage (Unstructured)	Azure,Cosmos	Google File System (GFS)	Hadoop Distributed File system (HDFS)	Elastic Block Storage (EBS), Simple Storage Service (S3)
Distributed data storage (Structured)	SQL Azure	BigTable	HBase, Hyper table	Dynamic Simple DB, Relational DB Service (RDS)

II. LITERATURE REVIEW

Some of the related work which had already implemented over cloud data base is discussed as follows:

Dashi et all has implemented jena semantic web framework for data preprocessing and pellet OWL Reasoner for query execution. It contain two components as data generator preprocessor and query processing which further has 3 more sub components as LUBM which creates data in RDF/XML serialization format. Then this data is converted into N Triple serialization using N-Triple converter component and Jena is used to convert the data. After this Predicate Based file splitter takes the data and splits it into predicate file ,then these files are fed into objects type based file splitter system which split the predicate file based on type of objects then the output is fed into the HDFS. Now, Map Reduce [4] framework has 3 components SPARQL which take the query from user and then passes it to job decider and input selector after that number of jobs have been selected and passes it to the job handler which submits jobs to Hadoop. The query requires inferencing which is possible by using pellet OWL Reasoner [5].

Another concept by wang, shang and zhou regarding the spatial data retrieving system which retrieve the data by storing well known bits (WKB) objects in BLOB entity and well known text (WKT) objects in CLOB entity. The primary tasks is to search the spatial data objects in a CLOB/BLOB virtual objects in cloud, and implements some objects to manipulate the data (such as operators are- Add, delete, update). Here model work with spatial data object in compliance with OGC (Open GIS Consortium) simple feature specification spatial objects is encapsulated in the persistent layer through which spatial objects is encoded into WKB (Well known bits) and WKT (Well known Text) format. Then access BLOB/CLOB virtual entity through APTs cloud from cloud storage and it consider ordinary data type as string numbers or others [6].

III. ARCHITECTURE DESCRIPTION

Data center implemented by Google- Bigtable is less secure and data integrity is difficult to maintain [2] [7]. These problems are overcome in this model by implementing security at the login session and integrity is maintained by master and slave operations [8]. In order to facilitate a seamless integration of various organization databases, the schemas of the distributed database residing on cloud(s) will imitate the existing schema of specific organization. And along with this some of the basic standards are used in each

organizations to build a successful system like in medical, Database standards [1] are : Health Language 7: enabling communication between applications provided by different vendors, using different platforms, operating systems, application environments (e.g. programming languages, tools). Electronic Data Interchange: is a data format based on ASC (Accredited Standards Committee) X12 standards. It is used to exchange specific data between two or more trading partners. like invoice, purchase order, healthcare claim, etc. Health Insurance Portability and Accountability*:* HIPAA regulations were established to protect the integrity and security of health information, including protecting against unauthorized use or disclosure of the information. Disaster Recovery: there should be a provision for backing up of entire system data [6].

This paper has proposed system architecture for information retrieval from distributed database for a cloud system. As shown in fig2 this model includes many modules such as the first module which have the information about organization standards, EDI, security policies, and other backup procedures while retrieving the data from cloud data center, other modules are web service Integration client, web service, distributed database and cache.

Now, Each of these module are discussed separately, As during initial setup of the web-service clients, the schema of the existing Company Information system will mapped to the proposed cloud based distributed database schema.

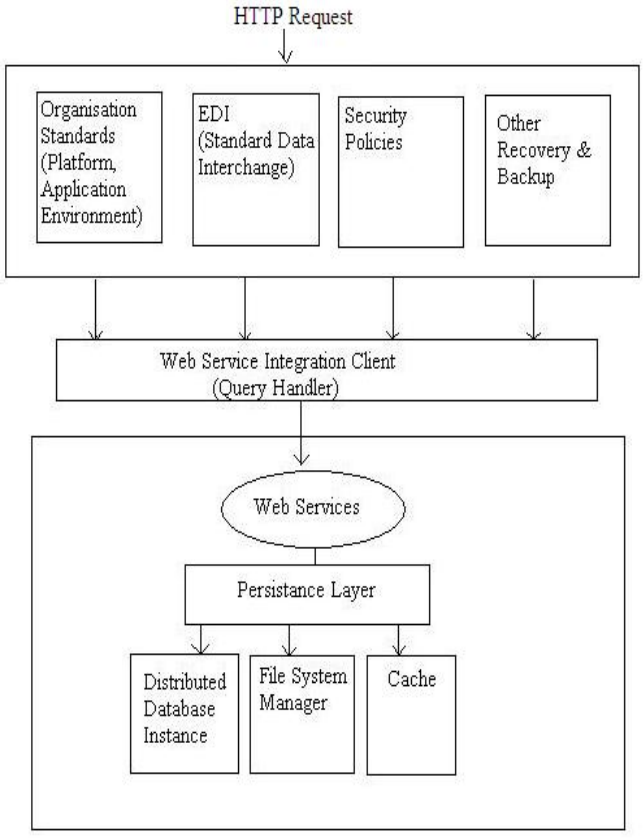

Figure 2: Information Retrieval Model/ Framework

This would allow the agent to periodically query the client databases through established connections which will facilitate the transfer of data to the cloud(s) over secure HTTP connection [1]. Which also store the information regarding type of platform, data base provider, security policies, EDI format and backup recovery schemes though connecting with cloud database.The next is the web service integration layer which is a query handler, gets input from user as a text or word then match it with the master table to find correct options from it.

Then web services (WS) layer which is a major service oriented and connectionless technology which is mostly open and specification based. As its main advantage over existing middleware is language, platform independent and location transparency [9], also it is easy for an organization to work over interest as it allow tunneling through HTTP that bypass corporate firewall. It will also provide technology like XML Signature, XML Encryption and WS Security to quality the protection for message integrity and confidentiality.

Another layer is persistence layer which is an architectural layer whose job is to provide an abstract interface to information storage mechanism(s). An API to such an interface should be abstract, and independent of storage technologies and vendors. It would typically have features such as the following:

A. store / retrieve of whole object networks by key
B. logical database cursor abstraction for accessing some / all instances of a given type
C. transaction support open, commit, abort
D. session management
E. some basic querying support, e.g. how many instances of a given type exist etc

Distributed database and file system which store the client(s) data a multiple locations and cache is used to store the frequently accessed database and customer id which took less time in retrieving process. So cache become the most important module of this model, help in fast retrieving of data. In Distributed systems we have one master and some salve data bases. Master data store the information of client login id, session id and its database location (slave address), where as slave will store all the data regarding the client that will be accessed only by a key stored in master storage in respect of each client and then by use of that key we may access the database [8]. Through this we can maintain security too. Retrieving of data is allowed till the session is maintained if once session breaks down, the data base will save the client data by rollback /commit according to transaction for this we have to maintain a replica at slave side. So commit or rollback operations are performed over only slave replica not over main data to maintain data integrity.

IV. DATA ANALYSIS

On the basis of two Algorithms discussed in section 2, one of the algorithm by dashi etc all [5] used for string/text processing using jena semantic web framework and it also use techniques like N-Triple and LUMB for processing and execution of data. This algo doesn't provide much data security, whereas another algorithm by wang etc all [6] worked over multimedia data and it uses BLOB/CLOB techniques to store multimedia information with the help of Open GIS consortium and provides less data integrity.

In order to keep in mind security and integrity of data a model is proposed in this paper with other functionalities too. The modules implemented in this model are web service Integration client, web service, distributed database and cache which provide a better data center.

V. CONCLUSION

On the basis of above analysis it can be concluded cloud computing is a new style of computing in which dynamically scalable and often virtualized resources are provided as a services over the Internet. The model discussed in this paper is architecture for information retrieval which provides scalability, secure and frequent access, backup and recovery policies and data integrity too.

In Cloud computing there are some loopholes regarding actual location of data at a given point of time, resources provided to multiple customer dynamically according to there need / demand and they have no knowledge or control over the exact location of resource provider. So, these are some suggested areas where studies are required to be done so as to manage the resources to be cost effective.

VI. REFERENCES

[1] Sun Microsystems (June 2009), Introduction to cloud computing architecture. White Paper, Sun Microsystems

[2] Kathleen Ericson and Shrideep Pallickara, "Survey of Storage and Fault Tolerance Strategies Used in Cloud Computing", Handling of cloud computing, 2010, ISBN 978-1-4419-6523-3

[3] http://www.readwriteweb.com/cloud/2011/01/7-cloud-based-database-service.php

[4] Boom09 (2009). BOOM: Data-centric programming in the datacenter (UC Berkeley EECS Tech. Rep. No. UCB/EECS-2009-113 August 11, 2009).

[5] Husain, Dashi, Khanand and Thuraisinghan "Storage and retrieval of large RDF graph using Hadoop and Mapreduce", Cloud Computing: First International Conference on CloudCom, 2009 Beijing.

[6] Yonggang, Wang, Shang and Daliang zhou,"Retriving and Indexing Spatail data in cloud computing Environment", Cloud Computing: First International Conference on CloudCom, Pg-:324-327, 2009 Beijing

[7] Chang F., Dean J., Ghemawat, S.Hsieh, and W.C: Bigtable: A Distributed Storage System for Structured Data. In OSDI 2006: 7th Symposium on operating system design and Implementation pp15-25(2006).

[8] Fabrizio Marozzo, Domenico Talia, Paolo Trunfio, Peer to peer framework for supporting MapReduce applications in dynamic cloud, Cloud Computing: Principles, Systems and Applications by Lee Gillam, 2010, Ch-7, Pg:120-122, ISBN 978-1-84996-240-7(online).

[9] Gerald Kowalski, Information Retrieval Architecture and Algorithms, Ch-8 search optimization, page no.235-237, 2010, ISBN 978-1-4419-6523-3(online).

.