An Overview of Data Management in Cloud Computing

K.Yogitha Lakshmi, S.Dhanalakshmi, B.G.Obula Reddy

Abstract: ***As we all familiar with cloud computing, it's not a latest technology, rather we can mention it as an emerging technology where most of the industry is trying to store not only its crucial data for redundancy but also looking for the service management. In that scenario first thing comes in mind is management of data in most efficient way possible. So here we tried to showcase two technologies of cloud data management namely Cloud BigTable and Cloud DataStore which they have their own way of working environments. It makes so much importance to choose the right technology for the right nature of work.***

Keywords: Cloud Storage, Data Management, Virtualization, Google File Systems, Data Store

1. INTRODUCTION

Cloud computing referred as delivery of computing services such as networking, servers, databases, storage, virtualization, storage, software, business analytics and so on over the Internet as a utility just like using telephone/mobile services. It offers product innovations and flexible resources for the business like Pay per Use services from The Cloud. The advantages of cloud computing are: Flexible resources - On-demand services gives user a quick scale up or down of the resources. Metered service gives the liability to pay for what you use. Self service you can access all the IT resources without any assistance.

A. Deployment Models of Cloud Computing

Deployment models of cloud computing represents that public cloud, private cloud, hybrid cloud, community cloud and different services. The Fig.1. Shows the representation of cloud architecture in various models and its services.

Public Cloud: This infrastructure will be used by public cloud user in which some of the services will be unavailable. These resources will be provided and organized by a cloud service provider, academics or other organizations. The cloud server exists on the premises of the cloud provider. Eg: Google App Engine, Windows Azure etc.,

Private Cloud: This infrastructure is for exclusive use of a single organization with various services, it can be managed by the organization, a third party or sometimes both. It exists on or off the premises. Eg: VMWare, RedHat etc.,

Community Cloud: It can be managed, operated and owned by one or more organizations in the community, or a third party or combination of both. It may exists on/off premises. Eg: Salesforce community cloud etc.,

Hybrid Cloud: It is the combination of two or more types of infrastructure services that works under proprietary rules and standards. Eg: VMWare vCloud etc.,

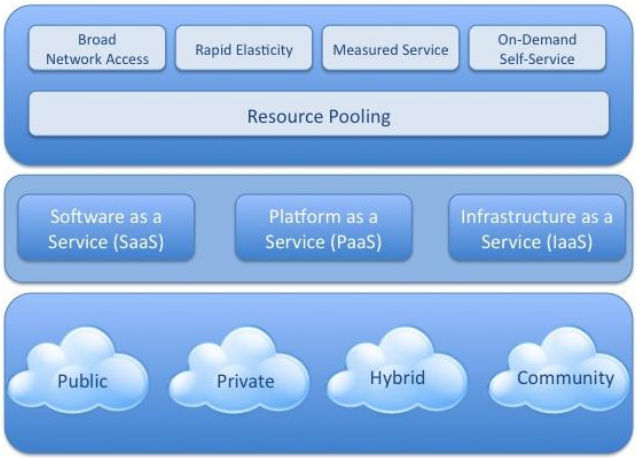

Fig.1. Cloud Architecture

2. THE ROLE OF VIRTUALIZATION IN CLOUD ENVIRONMENT

Virtualization is a multi-tenancy user infrastructure which is located at remote site and can perform function of multiple systems in one physical system by means of high speed internet. In cloud environment it comes under IaaS (Infrastructure as a Service), where the cloud consumer gets the service to use cloud based ready to access virtual storage and also some built in services. The pricing of these services depends on data storage no. of GB used per hour, network infrastructure used per hour etc., Fig.2. Represents the components stack of virtualization in hypervisor and hardware parts.

Virtualization Component Stack consists of Hardware, Operating Systems, Middleware and application layer. Operating Systems layer split into two parts:

a) Hypervisor is also called as virtual machine manager which allow user to have multiple OS in single hardware

b) Guest OS is a running Operating system within the Virtual machine.

Revised Manuscript Received on February 22, 2019.

K.Yogitha Lakshmi, Assistant Professor, Department of IT, Malla Reddy Engineering College(A),Telangana,India

S.Dhanalakshmi, Professor, Department of CSE, Malla Reddy Engineering College(A),Telangana,India

B.G.Obula Reddy, Professor, Department of CSE, Malla Reddy Engineering College(A),Telangana,India

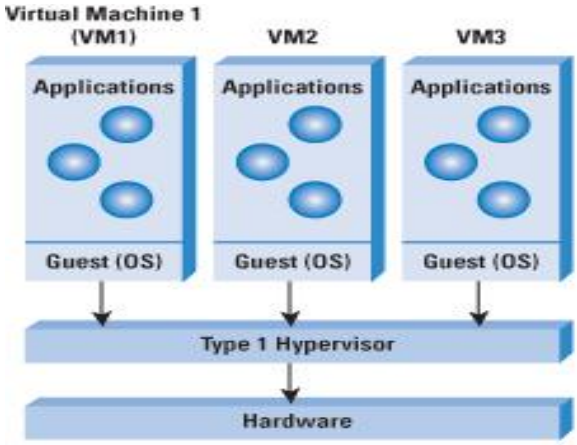

Fig.2. Virtualization Component Stack

The three goals of virtual machine are;

Equivalence: To poses unbiased hardware performance among all the VM's(Virtual Machine)

Resource Control: The VM's should be in complete control of any virtualized resources

Efficiency: The VM's instructions should be executed from its CPU rather than involving hypervisor.

3. TYPES OF VIRTUALIZATION

There are mainly three types of virtualizations namely Server virtualization, Client Virtualization and Storage Virtualization. The Fig.3. Shows the different categories of virtualizations.

Server Virtualization: It is the most common type of virtualization in cloud computing, where it gives optimum usage of server by running multiple applications on multiple operation systems at the same time on single server with the use of hypervisor by controlling CPU, Memory and other components without need of source code.

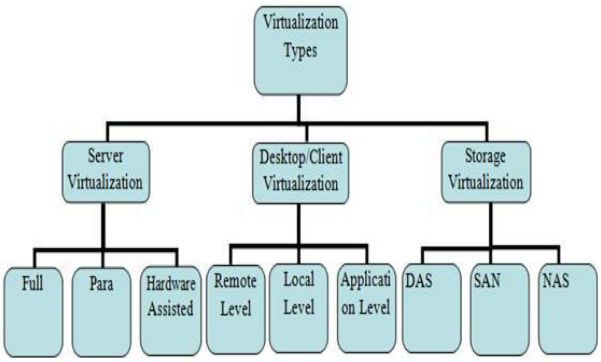

Fig.3. Types of Virtualization

Client Virtualization: In client virtualization, the administrator can manage and control the operations of client machine like personal devices. Here we need to have glance at three types of client virtualizations. First is remote level, where consumers can able to access cloud server which is located remotely anywhere and anytime across a network. Second, local level which runs on local server for the purpose of security. Third, application level virtualization which allows applications to run on isolated or private environment which is accessible by providing authentication.

Storage Virtualization: In Storage virtualization, a single storage device manages multiple network storage resources. It provides efficient storage management in large IT sector and reduces downtime. Three types of storage virtualizations are DAS(Direct Attached Storage), SAN(Storage Area Network), NAS(Network Attached Storage). DAS is a primary way of data storage, where the storage drivers are directly connected to the server. NAS follows a method of storage called sharded method which connects through the networks and it is used for file sharing, device sharing and scheduled/ Ad-hoc backup of the server. SAN is a technique of storing data in a device that is shared among different servers over a high speed network.

4. DATA MANAGEMENT IN CLOUD COMPUTING

In cloud computing data management itself is a big challenge in processing large quantity of data for the purpose of data storage, parallel processing of data execution, analytical processing and online query execution all by ensuring consistency and durability under peak loads.

Some of the cloud based analytical data management systems are: BigTable, HBase, HyperTable, Hive and HadoopDB. PNUTS and Cassandra are the web based data management systems. Here, in this paper the working nature of BigTable and Dynamo will be discussed

A.GFS (Google File System)

It is designed to manage large files in distributed networks of servers which is connected by a high speed internet. It provides atomicity during read/ writes operations of individual files. Supports read/ write and update operations simultaneously by multiple client programs.

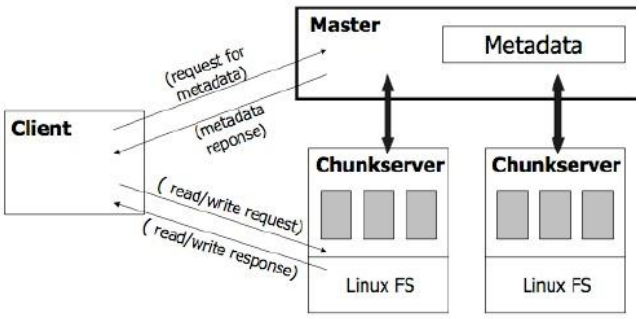

Fig.4. GFS Architecture

A single Master controls the namespace. A large file will be cut down to chunks or blocks with the size of 64MB. These Chunks (GFS) or Name nodes(HDFS) are stored on a servers called Chunk Server. The main functionality of this server is to replicate these chunks three times on different physical racks and network segments.

Read Operations in GFS: a) Client program request for metadata by sending its full path and offset of a file to MASTER or NAMENODE. b) MASTER replies back with metadata from one of its replica chunk where the data is found.

Write/Append Operations in GFS: For initiating write/append operation the process is same as read operation along with some extra steps. a) Client sends its data to be

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

appended to all its chunk servers b) Chunk server acknowledge the receipt of this data. c) Among all its chunk server replicas, MASTER chooses a PRIMARY Chunk server which is responsible to append the client data to its secondary chunk servers.

Fault Tolerance in GFS: MASTER bind its synchronization with its replicas by sending regular Heart-Beat messages. In case of failure, chunk server meta-data will be sent to MASTER and it will choose a new Primary Chunk server.

5. MANAGING DATA IN CLOUD ENVIRONMENT

Managing data in cloud environments can be provided different storage techniques in google platforms. Its specify features of Google Cloud BigTable and Google Cloud Datastore.

A.Google Cloud BigTable

BigTable is a distributed storage system that store large amount of data such as petabytes in NoSQL cloumn-oriented way of data store developed by Google Inc. to manage its internet search and web service functions. It works on powerful database servers which gives the benefits of scalability, easy administration and maintain elasticity of cluster without any down time.

BigTable is used to store and query the following types of data:

Time Series Data
Marketing Data
Financial Data
Internet of things Data
Graph Data

The Fig.5. Represents the BigTable Storage Model of rows and columns specifications. Each column store arbitrary value as name-value pair in form of column family. At the time of table creation initial value of no.of column families will be fixed. Labels of column families can be created at any point of time. Each BigTable cell can contain multiple versions of data in decreasing order of timestamp.

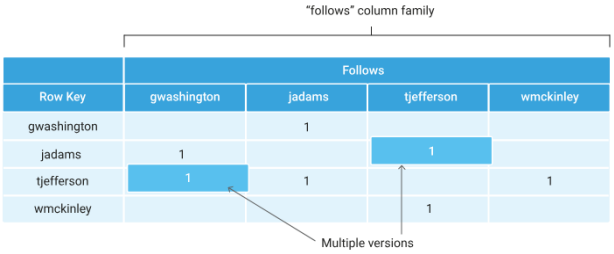

Fig.5. BigTable storage Model

Each table in BigTable will be divided into different row ranges called tablets. These Tablets will be maintained by a server called tablet server. It stores each column family in an allocated row range inside a distributed file called SSTable. BigTable maintains its meta-data table in a single meta-data server which is used to locate the user tablets in response to their read/write operations. The meta-data table itself will be divided into no.of tablets to support its large amount of data in most effective way . Root Table will help point out other meta-data tablets. It supports large parallel reads and inserts operations simultaneously on the same table.

B.Google Cloud DataStore

Google cloud Datastore is a NoSQL document database developed for incredible scalability, high performance and to support application development. The most appreciated feature in cloud datastore is to provide high performance to its subscriber even in the high incoming data traffic situation. It Maintain ACID properties and also it give high availability.

Cloud DataStore is used for applications like:

Product Catalog where it provides real-time inventory

User Profiles where the retailer can view the preferences of the user based on past interests.

Bank Transactions where ACID property will guarantee the transaction of transferred funds.

All the data in Datastore stores in one bigtable called as Entity Table. It stores data horizontally across its disks in which it is called as shared and key values are sorted lexicographically.

It can handle multiple queries at a time by various users with the help of multiple index tables. For every data set they have entity sets from where user gets the results back. for example a query will have a defined set of results say 100 entities, because of this scenarios some queries would not get support in cloud datastore. Like in traditional RDBMS cloud datastore doesn't support schema and it is a schemaless database. Cloud Datastore do not support join operations, it won't filter data from a table with multiple keyed properties or by the result of a subquery. Cloud Datastore doesn't do justice for analysis of data but it can provide assurance for a transactional data.

6. CONCLUSION

In cloud computing environment, without the virtualization technique it would not be possible to use single hardware device among the users. It is the basic service of any development in cloud computing. Data management in cloud computing shows the rapid growth of deployment in remote servers for the purpose of storage and cloud services. Cloud BigTable is mainly used for the non-transactional data where it does not give any redundancy for the data. It can be used for data analytics where you can get the results by querying historical data. Cloud DataStore is built on BigTable but they are completely different from each other, where it supports ACID properties of the transaction and it is used on transactional data. It features are similar to SQL but it cannot perform some operations.

7. REFERENCES

1. Hamlen, K. Kantarcioglu, M. Khan, L. Thuraisingham, B. (2010). Security Issues for Cloud Computing.International Journal of Information Security and Privacy, 4(2), 36-48.
2. Bernardo Ferreira, Henrique Domingos (2012). Management and Search of Private Data on Storage Clouds.Center for Informatics and Information Technologies.SDMCMM'12, December 3-4, 2012, Montreal, Quebec, Canada.

3. RizwanMian, Patrick Martin (2012). Executing data-intensive workloads in a Cloud.ACM International Symposium on Cluster 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing.
4. Xiao-Bai Li, SumitSarkar (2006). Privacy Protection in Data Mining: A Perturbation Approach for Categorical Data Information Systems Research. (17) 3, 254–270
5. Iyengar, V. S. (2002). Transforming data to satisfy privacy constraints. Knowledge Discovery DataMining.ACM Press, New York, 279–288.
6. Daniel J. Abadi (nd) Data Management in the Cloud: Limitations and Opportunities. IEEE Computer Society Technical Committee on Data Engineering
7. B. Siddhisena, Lakmal Wruasawithana, Mithila Mendis, ―Next generation muti tenant virtualization cloud computing platform‖, In: Proceedings of 13th International conference on advanced communication technology(ICACT), vol. 12, no.3; 2011. p.405–10.
8. Z. Xiao and Y. Xiao, ―Security and Privacy in Cloud Computing‖, IEEE Communications Surveys & Tutorials, vol. 15, no. 2, pp. 843–859, 2013.
9. Sunilkumar S.Manvi, Gopal Krishna Shyam, "Resource anagement for Infrastructure as a Service(IaaS) in cloud computing: A survey", Journal of Network and Computer Applications 41, (2014) 424–440.
10. Chase JS, Darrell C Anderson, Prachi N Thakar, Amin M Vahdat, ―Managing energy and server resources in hosting centers‖, In: Proceedings of 11th IEEE/ACM international conference on grid computing (GRID), vol.12, no.4; 2010. p.50–2.
11. B. Urgaonkar, P. Shenoy, A. Chandra, P. Goyal, T. Wood, ―Agile dynamic provisioning of multi-tier Internet applications‖, ACM Trans Auton Adaptive Syst 2010; 5 (5):139–48.
12. Vaquero LM, Luis Rodero-Merino, Rajkumar Buyya, ―Dynamically scaling applications in the cloud‖, In: Proceedings of the ACM SIGCOMM computer communication review, vol.41, no.1; 2011. p.45–52.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication