Storage Solution: A Virtual Distributed Storage And Migration Architecture For Big Data

Randle Oluwarotimi [1] ***, Matsebula Fezile*** [1] ***, Zuva Tranos*** [2]

Sol Plaatje University, South Africa [1] ,Vaal University of Technology, South Africa [2]

Keywords: Virtualization; data Migration; big data; distributed computing; cloud computing.

Abstract

The big data revolution has provided organizations with a large variety of data which assist in making decisions as well as providing data analysts with high volumes of data for prediction and pattern recognition. This data is stored in the cloud as it has proven to be a good storage environment due to its accessibility and security benefits .The cloud provides data for various applications such as gaming activities and data for prediction analysis in various sectors of the economy .The provision of these data to end-users such as data analysts is best provided through the use virtual desktops which require data regularly (real-time) and at a high speed, efficiency and performance levels. To achieve this users expect services to be hosted on virtual machines in interrelated data centres and that these virtual machines will migrate dynamically to locations best suited for the user as well as connect the new users .This leads to our current problem which is how can we provide high performance to manage large volumes of data from the cloud as well as how can data can be stored in such a manner that they can be easily retrieved and migrated between servers. We propose an Architecture in this paper by using Alluxio and our novel Dynamic Virtual Machine Server (DVMS) to speed up the process as well as ensure there is no delay. We further apply two plugins to Hadoop which are Sqoop and Network levitated Merge (NLM) which will assist to improve the transfer speed of data from the cloud to Hadoop to increase efficiency. The dynamic virtual machine manages the large and growing data load by categorising the data into 3 categories of pools called (1) Raw aggregated data pool, (2) Aggregated data to send and (3) Processed aggregated data pool which works in a loop to increase data migration speed as well as provide a medium to store data in preparation for new users.

1 Introduction

Most IT companies use the cloud as a base to store their resources and download their resources from the cloud. The cloud can be divided into public cloud where accessibility is provided freely to everyone through a remote interface, while private cloud is an environment designed to provide users with dynamic and agile cloud infrastructure which will run service workloads within administrative domains [1] and hybrid cloud are defined as an advancement of private cloud as they are a form of supplement for local infrastructure with computing capacity from external public cloud [1,2]. Virtualization is the process of imitation resources and it is an efficient way of providing resources to multiple users. Virtualization is an important component of the cloud as it reduces latency, which is a delay in the transfer of communication. The life cycle of a Virtual Machine (VM) consists of setting up networks dynamically for groups of VMs, storage requirements, deployment of VM disk images, configurable resource allocation [2].

Data migration is a critical component in the global technological world, as data is needed in different places and at massive or fast speed. There is a global agreement that there is a need to look for new ways to improve data migration as well as the need for a unique medium to retrieve massive amounts of data and use easily. Virtualization is a key component for distributed computing, as well as virtualization cannot be separated from distributed computing. services are now been instantiated by deploying service delivery systems on demand by dynamically allocating the necessary network and data centre resources which include on-demand gaming as well as network supported virtual desktop applications [3,4,5]. With these developments it becomes crucial that users do not experience delay when requiring more data as well as service delivery must be dynamic to adapting readily to the users' needs[5].Other issues include the capability of current cloud technologies to provide necessary capacity and high performance to address massive amounts of data , optimization of existing file systems for the volumes demanded by data mining applications, and how data can be stored in such a manner that they can be easily retrieved and migrated between servers[2,6].

2 Related Work

There have been proposals on how best to migrate data and some have suggested using dynamic DNS and tunnelling techniques [7], the unique technique was proposed by assigning the VM a new IP address at the final or target location. Through this process the VM retains both new and old IP addresses, whereby packets going to the old IP address are transferred through the old hosting machine to the new hosting machine, the derived limitation was that there was difficulty in bounding the transmission time as well as there was always a need to remodify the VM software since most VM migration procedures require the same IP address[7,8].Lakshaman et al., [9] proposed a technique which separates the forwarding elements from control elements over an open interface with a centralized route control. Fang et al have proposed a technique based on network-virtualization which relies on distributed forwarding elements and centralised control to enable seamless migration of virtual machines in enhancing delivery of cloud –based services

2.1 Definitions and Characteristics of Big Data

Data forms the foundational and fundamental part of analytics. According to Villars et al [14], a wide range of organizations, from small, medium, and large enterprises and government agencies are dealing with a flood of data. The data is collected through;

- Digitize business records and personal content, including the generation of ever-larger numbers of photos, movies, and medical images driven by continued advancements in device features, resolution, and processor power [14].
- Instrument devices such as set-top boxes, game systems, smartphones, smart meters. Also buildings, cities, and even entire regions to monitor changes in load, temperatures, location, traffic patterns and behaviours [14].
- Address governance, privacy, and regulatory compliance requirements that complicate the retention of business information [14].

With that being said, big data is about the growing challenge that organizations face as they deal with large and fast growing sources of data and information that also present a complex range of analysis and use problems [14]. These include:

- having a computing infrastructure that can ingest, validate, and analyse high volumes of data
- assessing structured and unstructured data from multiple data sources
- dealing with unpredictable content with no apparent schema or structure
- Enabling real-time or near-time collection, analysis, and answers.

Big data is a term used to refer to the increase in the volume of data that are difficult to store, process and analyse through traditional databases. Currently several definitions of big data exists in literature. [15] Referred to big data as a large volume of scientific data for visualization. Whilst [16] defined big data as the amount of data just beyond technologies capability to store, manage and process efficiently. The term big data was introduced by Gartner as characterized by 3Vs; Volume, Velocity and Variety. Although Berman [17] pointed out that big data cannot be characterized by 3Vs, Value also characterizes big data. Hashem et al [2] proposed the definition of big data as a set of techniques and technologies that require new forms of integration to uncover large hidden values from larger datasets that are diverse, complex and of a massive scale.

a. Volume – refers to the enormous amount of all data types collected from multiple data sources and continue to grow.

b. Variety – refers to different types of data collected from different sources, such as, sensors, smartphones, or social networks. These data can be structured or unstructured data, such as video, images, text, audio and data logs.

c. Velocity – refers to the speed of data transfer. [17] Asserts that the contents of data constantly change because of the absorption of complementary data collections, introduction of previously archived data and streamed data arriving from multiple sources.

d. Value refers to the process of discovering hidden values from larger data sets with various types and rapid generation [18].

2.2 Cloud Computing

The key to understanding and using cloud computing effectively is to realize that it is ultimately a bi-directional service; valuable data, information and knowledge must flow easily and securely to and from the user and these valuable assets must be saved securely and be adequately backed up and protected from disasters [19]. Cloud services has become a powerful architecture to perform complex large-scale computing tasks and span a range of IT functions from storage and computation to database and application services [2]. [20] Points out that combining the cloud computing utility model and a rich set of computations, infrastructures, and storage cloud services offers a highly attractive environment where scientists can perform their experiments.

PaaS, SaaS, and IaaS are typical cloud service models [2];

- PaaS refers to different sources operating on a cloud to provide platform computing for end users.
- SaaS refers to applications operating on a remote cloud infrastructure offered by the cloud provider as services that can be accessed through the internet.
- IaaS refers to hardware equipment operating on a cloud provided by service providers and used by end users upon demand.

In addition as mentioned by [21], cloud service providers have begun to integrate frameworks for parallel data processing in their services to help users' access cloud resources and deploy their programs.

3 Proposed VDSM Architecture

The VDSM architecture is an architecture designed to increase the download speed and manage data migration between virtual servers. The architecture was designed after studies such as [2, 6] indicated that there is a need for an easier way to migrate data between servers and can easily be retrieved. To understand how the architecture works, the architecture consists of four major sections which are cloud environment, data acceleration process, virtual server and dynamic virtual machine servers. Figure 1 below provides a pictorial explanation of the architecture.

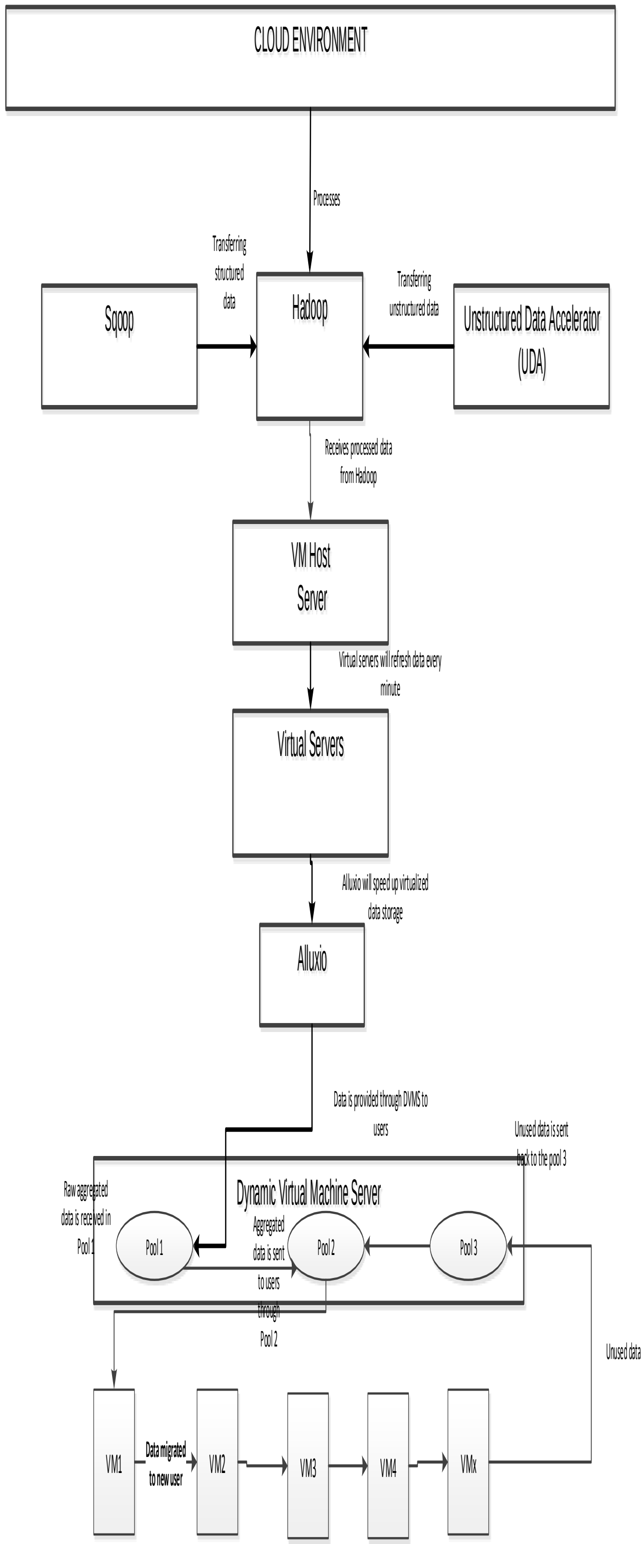
CLOUD ENVIRONMENT
Processes
Transferring structured data
Sqoop
Hadoop
Transferring unstructured data
Unstructured Data Accelerator (UDA)
Receives processed data from Hadoop
VM Host Server
Virtual servers will refresh data every minute
Virtual Servers
Alluxio will speed up virtualized data storage
Alluxio
Data is provided through DVMS to users
Unused data is sent back to the pool 3
Dynamic Virtual Machine Server
Raw aggregated data is received in Pool 1
Pool 1
Aggregated data is sent to users through Pool 2
Pool 2
Pool 3
Unused data
VM1
Data migrated to new user
VM2
VM3
VM4
VMx

Figure 1: A Virtual Distributed Storage and Migration Architecture for Big Data

4 Components of the Proposed Architecture

The components of the proposed architecture are explained below;

4.1 Cloud environment

This is the unique process of providing shared resources such as development, deployment, business protocols, and storage through computing [1]. The cloud environment is now needed by businesses and with big data it has become an integral component of the big data environment. The cloud has become the new data warehouse for billions of data which help companies with their analysis, prediction and data storage.

4.2 Alluxio

Speeds virtual distribution storage platform, sits between the storage and processing architecture and improves performance. It is a distributed system that is memory centric and distributed .It has the ability to sit with or besides existing file systems. It is also defined as the glue that holds together increasing dissimilar and disjointed big data architectures [10]. Alluxio main function in this proposed architecture is to manage the virtualized data as well as provide the link dissimilar data as well as work with the Hadoop distributed file system (HDFS).

4.3 Sqoop

Sqoop has an efficiently transferring bulk data between apache Hadoop and structured databases [11]. It operates in a unique manner which includes copying the data quickly from external systems to Hadoop and enables data imports from external data sources. Sqoop provides different data formats for data import. In this proposed architecture we intend to use sqoop for only the transfer of structured data because of its ability to transfer structured data at a fast speed.

4.4 Unstructured Data Accelerator (UDA)

The proposed introduction of UDA is to assist to solve the current problems of the Hadoop Architecture which will in the long run affect negatively the data migration and distribution process which this study proposes which includes (a) the serialization between Hadoop shuffle/merge and reduce phases, (b) repetitive merges and disk access, (c) the lack of support for Remote Direct Memory Access (RDMA) interconnects [12]. Unstructured data Accelerator is a novel process or protocol which aims to increase the data analytics process of Hadoop, it is a scalable system that is deployed on clusters with three data nodes [12]. UDA is based on the network –levitated merge algorithm which works by overcoming the serialization process between the shuffle, merge and reduce phases by ensuring the RDMA speeds the data transfer process and reduces the CPU overhead thereby leading to increased CPU processing availability for data analytics [12].Our proposed system will use this technique to as a plug-in to assist Hadoop with unstructured data.

4.5 Hadoop

The Hadoop Distributed File System (HDFS) is an open source solution (inspired by Google File System and Map Reduce architecture) that is designed to store very large data sets and to stream these datasets to user applications ensuring a good performance and bandwidth. The HDFS architecture includes the 3 following components: a Name node that maintain the hierarchy and the file system metadata such, the Data nodes used to store the HDFS file data in local file system, and the HDFS Client which is a code library to export HDFS file system interface so application can access and use it. The MapReduce which is implemented by Hadoop, is an architecture designed to distribute the stored data and computation tasks across multiple servers to enable resources to scale according to the demand but still economize in the size. By using MapReduce developers can easily work with large scale parallel computation the functions can be used to aggregate large amounts of data [13].

4.6 Improved Virtual Machine (VM) host server
The VM host server is a component that handles the hardware that enables and provides the computing power resources which include memory, power for processing, network and disk I/O [8]. This component is vital because it assumes the role of the brain of all virtualized resources.

4.7 Proposed Dynamic virtual machine server
Our proposed dynamic virtual server has been designed to optimise migration speed through the use of 3 stages which includes stage 1- The initial set of data which is the raw data is received from the virtual server which will receive updated and refreshed data every 2 minutes. It then transfers it to stage 2 which is the aggregated data to send to users, which serves as the portal or point where data will be distributed to all the various users through their Virtual Machines (VM). Once the users receive their data they can decide to retain data or migrate the data which is not used to other VM, once the last user has received data it is then passed onto stage 3- Processed Aggregated data pool which then passes it to Stage 2 to redistribute the data through all the virtual machines user including new user. The proposed process is further explained in the pseudocode below.

4.8 PSEUDOCODE FOR DYNAMIC VIRTUAL MACHINE SERVER

```
initialize vm requirement
input raw data
initialize data data_pool_1, data_pool_2, data_pool_3
initialize user1,user2,user3
initialize cloud connection
while cloud connection is active
add raw data to data_pool_1
if vm does not have enough recourses
increase vm requirement
else
decrease vm requirement
move data from data_pool_1 to data_pool_2
if data_pool_3 is not empty
move data from data_pool_3 to data_pool_2
move data from data_pool_2 to user1
move data from user1 to user2
if user2 request data from data_pool_3
move data from data_pool_3 to user2
move data from user2 to user3
if user3 request data from data_pool_3
move data from data_pool_3 to user3
move data from user3 to data_pool_3
```

5 Conclusion and Discussion

This paper proposes an architecture for virtualized cloud storage and migration to speed up data process. Although the cloud is used globally for various purposes the critical issue of how to improve performance and how to migrate data easily to various users still remains a problem. The paper also proposes a layered architecture for cloud storage, migration and improved processing speed through virtualization. In the operation mechanism, a Dynamic Virtualized Machine Server (DVMS) is presented which deserves future investigation. As part of future work, the proposed virtual distributed storage and migration architecture for big data will be validated in a health observatory since the data generated by the observatory is being used by various stakeholders. The validation will involve the implementation of the framework using big data tools . It will also involve interviews with health practitioners who are involve with data and information management. Generalizability of the suggested framework could be confirmed by conducting follow-up research in other organizations such as higher education, banking sector and small medium enterprises.

References

[1]S. Borja, M. S. Ruben, M. T. Ignacio, －An Open Source Solution for Virtual Infrastructure Management in Private and Hybrid Clouds,‖ IEEE Internet Computing, Vol. 1, pp. 14-22, 2009

[2] Hashem, I.B.H, Yaqoob, I.,Anuar, N.B, Mokhtar .S, Gani. A, Khan.S.U. 2015. The rise of “big data” on cloud computing: Review and open research issues. Elsevier, Information Systems, 47, 98-115

[3]mokafive. http://www.mokafive.com

[4]onlive.http://www.onlive.com

[5]Fang,H.Laksham,T.V,.Mukherjee,S and Song,H. Enhancing Dynamic Cloud-based Services using Network Virtualization. ACM SIGCOMN Computer Communications review, Vol 40,pp 67-74

[6]Leavitt, N 2013.Storage challenge: where will all that big data go? Computer 46,22–25.

[7]R. Bradford, E. Kotsovinos, A. Feldmann, and H. Schioberg. Live wide-area migration of virtual machines including local persistent state. In ACM/Usenix International Conference On Virtual Execution Environments, 2007.

[8]M. Caesar, D. Caldwell, N. Feamster, J. Rexford, A. Shaikh, and J. van der Merwe. Design and Implementation of a Routing Control Platform. In Networked Systems Design and Implementation, 2005

[9]T.V. Lakshman, T. Nandagopal, R. Ramjee, K. Sabnani, and T. Woo. The SoftRouter Architecture. In ACM HOTNETS, 2004

[10]Woodie, A. 2016. Meet Alluxio, the distributed file system formerly known as Tachyon.

https://www.datanami.com/2016/02/23/meet-alluxio-the-distributed-file-system-formerly-known-as-tachyon/

[11]Aravinth,S.S.,Begam,A.,Shanmugapriyya,S.,Sowmya,S., Arun ,E. 2015. An Efficient Hadoop frameworks sqoop and Ambari for big data processing . International Journal for Innovstive Research in Science and technology , Vol, 1, no 10, pp 252-255

[12]W. Zeng, Y. Zhao, K. Ou, and W. Song, “Research on cloud storage architecture and key technologies,” in Proc. ACM ICIS ’09, Seoul, Republic of Korea, 2009, pp. 1044–1048.

[13]Dede, E., Govindaraju, M., Gunter, D., Canon, R. S., Ramakrishnan, L., 2013. Performance evaluation of a mongodb and hadoop platform for scientific data analysis. 4 th Workshop on Scientific Cloud Computing, ACM, pp. 13–20.

[14] R.L. Villars, C.W. Olofson, M. Eastwood (2011). Big data: What it is and why should you care. White Paper

[15] M. Cox, D. Ellsworth (1997). Managing big data for scientific visualization. ACM Siggraph, MRJ/NASA Ames Research Centre

[16] J.M. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A.H. Byers. (2011). Big Data: The next frontier for innovation, competition and productivity.

[17] J.J. Berman. (2013). Introduction in: Principles of Big data, Morgan Kaufman, Boston.

[18] S. Kaisler, F. Armour, J.A. Espinosa, W. Money. (2013). Big data: Issues and challenges moving forward, system sciences (HICSS)

[19] M. Armburst, A. Fox, R. Graffith, A.D. Joseph, R.Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, M. Zaharia (2010). A view of cloud computing. Commun. ACM

[20] T. Gunarathne, B. Zhang, T-L. Wu, J. Qui (2013). Scalable parallel computing on clouds using Twister4Azure iterative MapReduce. Futur.Gener .Comput.Syst.

[21] D. Warneke, O.Kao. (2009). Nephele: Efficient parallel data processing in the cloud, in Proceedings of the 2nd workshop on many-task computing on grids and super-computers. ACM