An Analytic Real Life Case Study of Google Cloud computing architecture Platform's for Big Data Analytics Products from AWS, Azure and GCP

Devendra Kumar Nagayach[1], Dr. Pushpneel Verma[2]
[1]*Research Scholar, Deptt. of Computer Science & Application, Bhagwant University, Rajasthan, India*
[2]*Associate Professor, Bhagwant University, Rajasthan, India*

ABSTRACT

A large amount of data, structured or unstructured, is called "Big Data", which mainly consists of five properties such as volume, velocity and variation, validation and value of which value is the most important whose main purpose is to extract relevant information. . The other four V's to solve this problem of polite data and analysis; various technologies have emerged in the last decades. "Cloud computing" is a platform where thousands of servers work together to meet various computing needs and billing is done as per 'pay-as-you-go' increases. This study will present a novel dynamic scaling methodology to improve the performance of big data systems. A dynamic scaling methodology will be developed to scale the system from a big data perspective. Furthermore, these aspects will be used by the algorithm of the supporting project, which can be broken down into smaller tasks to be processed by the system. These small bangles will be run on multiple virtual machines to perform parallel work to increase the runtime performance of the system.

Keyword: - *SCC, GFRSCC, Properties, and EFNARC etc.*

1. INTRODUCTION

Currently, we live in an era where big data has emerged and is attracting attention in many fields such as science, healthcare, business, finance and society. The continuous increase in the size of data in various regions has led to a huge flow of data in the last decade. Thus, many systems have encountered problems in analyzing, storing and processing large amounts of data, leading to performance failures or slow performance and processing. When systems that process massive amounts of data experience poor performance, the increased cost, reduced revenue, or both create a negative impact. Additionally, delays due to poor performance increase unprocessed data and response time. The question is: Does handling large amounts of data play a significant role or greatly affect performance? Also, how do we improve system performance when dealing with massive amounts of data? Under the exponential growth of data, the term Big Data means an increase in the amount of data and is difficult to store, process, and analyze through traditional databases. In addition, it is characterized in 4Vs: variation, price, velocity and volume.

Big data poses many challenges in performance, scalability, and capacity. With big data systems becoming more prevalent, a need exists to overcome the challenges and implications of huge data. Therefore, the purpose of this thesis is to improve performance in big data systems, and it focuses on the performance challenge from the perspective of big data systems.

What happens if 10 gigabytes (GB) of data needs to be sorted? Modern-day computers have enough memory to hold this amount of data and can easily process it through memory sorting algorithms such as QuickSort. What if 100 GB or One Terabyte (TB) of data needs to be sorted? High-end configuration servers are available to hold this large amount of data in memory, but as they are quite expensive, it is better to choose a disk-based system, but in this case, to sort the data like mergesert the algorithm can be used. However, what if 50TB, 100TB or more data needs to be sorted? This is only possible with many parallel disk systems but in this case, a different algorithm such as bitonic sort must be used. These scenarios clearly conclude that the same problem with different size of data needs a different solution [1]. The amount of new data in the world is increasing rapidly. This can be imagined with the fact that the data generated between the time and the beginning of the year 2000 is now generated every minute. With such polite data there is a problem with handling and processing data. "Big data" is the term used for such a large-scale data set, be it structured (eg RDBMS) or unstructured (eg social media, organization data, etc.). The analysis process of such large scale data or Big Data is known as Big Data Analytics.

2. CLOUD COMPUTING ARCHITECTURE

The cloud architecture consists of five different components that work together to provide on-demand services. Figure 1.1 is taken from the National Institute of Standards and Technology (NIST) cloud computing reference architecture [4]. It represents the cloud architecture and its five components i.e. cloud provider, cloud consumer, cloud carrier, cloud auditor and cloud broker.

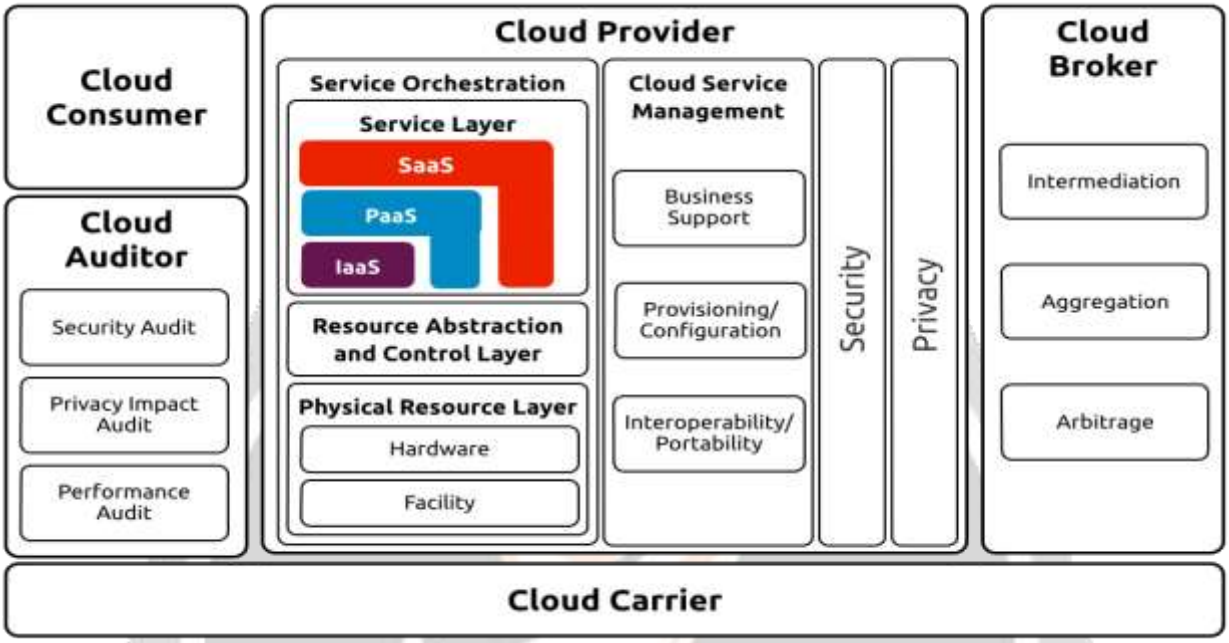


Figure 1.1: Cloud Architecture [4]

3. LITERATURE REVIEW

The most difficult problem that needs to be solved to handle big data effectively is storage; It is not easy to deal with large volumes of data and varieties (Elgendi and Elragal, 2014; Zhong, et al., 2016; Lu, Z. et al., 2017).
There are several large data storage and analysis models. Where large amounts of data are caused by a wide variety of users and devices, a data center may be necessary to store and process data. Establishing a network infrastructure is essential to help gather data that is generated rapidly, which is then sent to the data center before being accessed by users (Love et al., 2017).
Research by Yi et al. (2014) identifies the components of the network that must be established, such as a basic data network, the bridge used to connect and transmit data centers, and at least one data center.
Enterprise Data Warehouse (EDW) A big data environment requires analysis skills as opposed to traditional environments (Hartmann, T. et al., 2019).
The big data environment accepts and demands all possible data sources. On the other hand, EDW approaches data sources with caution, as it is more streamlined towards supporting structured data (Elgendy and Elragal, 2014); Hartmann et al., 2019).
Due to the increasing number of data sources and data analysis possible, agile databases are needed for large data storage, allowing analysts the opportunity to create and adapt data easily and quickly (Elgendy and Elragal, 2014; Hartmann et al., 2019).

4. RESEARCH METHODOLOGY

This paper illustrates a novel dynamic scaling method to improve the performance of big data systems. The methodology consists of three steps: segmenting and modifying systems that require performance improvement, selection of deployment environments, and changing the database, and a guideline of how to implement these steps to increase system performance. Development tools for the cloud tools required for cloud computing refer to the aggregation of components such as middleware and database services, which is helpful in developing, deploying,

and managing cloud-based applications, as a result it is dynamically allocated Creates large scale influential paradigms of resources and their complex computing. Big Data Analytics (BDA) provides data management solutions in cloud architecture for storage, analysis and processing of large amounts of data. This study presents a survey for performance-based an Analytic Real Life Case Study of Google Cloud computing architecture Platform's for Big Data Analytics Products from Amazon Web Services (AWS), Google Cloud Platform (GCP) and Microsoft Azure.

The three steps are:

- Scaling the System
    - Helper Algorithm
        - Consider categories of data in splitting
        - Consider analyzing data in splitting
        - Consider volume of data in splitting
- Modify the Current System
- Deployment Environment
- Database Transformation

Research Objective

1. To study the solutions provided by the top three cloud leaders i.e. Amazon Web Services, Microsoft Azure, Google Cloud Platform. A demo is performed to understand how easy it is to run a Big Data Analytics project in a cloud computing environment.
2. The study includes a demo of a Big Data Analysis on a major cloud computing platform to publicly validate the power of cloud computing using data sets.

5. RESULT AND DISCUSSION

5.1 Compare AWS and Azure services to Google Cloud

Compute Services

Services	AWS	Azure	GCP
IaaS	Amazon Elastic Compute Cloud	Virtual Machines	Google Compute Engine
PaaS	AWS Elastic Beanstalk	App Service and Cloud Services	Google App Engine
Containers	Amazon Elastic Compute Cloud Container Service	Azure Kubernetes Service (AKS)	Google Kubernetes Engine
Serverless Functions	AWS Lambda	Azure Functions	Google Cloud Functions

Database Services

Services	AWS	Azure	GCP
RDBMS	Amazon Relational Database Service	SQL Database	Google Cloud SQL
NoSQL: Key–Value	Amazon DynamoDB	Table Storage	Google Cloud Datastore Google Cloud Bigtable
NoSQL: Indexed	Amazon SimpleDB	Azure Cosmos DB	Google Cloud Datastore

Storage Services

Services	**AWS**	**Azure**	**GCP**
Object Storage	Amazon Simple Storage Service	Blob Storage	Google Cloud Storage
Virtual Server Disks	Amazon Elastic Block Store	Managed Disks	Google Compute Engine Persistent Disks
Cold Storage	Amazon Glacier	Azure Archive Blob Storage	Google Cloud Storage Nearline
File Storage	Amazon Elastic File System	Azure File Storage	ZFS/Avere
Services	AWS	Azure	GCP
Virtual Network	Amazon Virtual Private Cloud (VPC)	Virtual Networks (VNets)	Virtual Private Cloud
Elastic Load Balancer	Elastic Load Balancer	Load Balancer	Google Cloud Load Balancing
Peering	Direct Connect	ExpressRoute	Google Cloud Interconnect
DNS	Amazon Route 53	Azure DNS	Google Cloud DNS

An analysis of the above comparisons shows the following results which are categorized below:

General:

o Microsoft Azure is the most comprehensive cloud platform out of the three.

o Google Cloud Platform provides direct IDE support named Cloud9.

o Amazon EC2 is the oldest of the three with a firm grasp on the IaaS service model.

Database and Virtualization:

o Azure offers the least number of database options while Google Cloud Platform offers the most.

o Amazon EC2 offers the most virtualization options.

Pricing:

o All three have customer specific pricing plans that depend on usage.

Specifications:

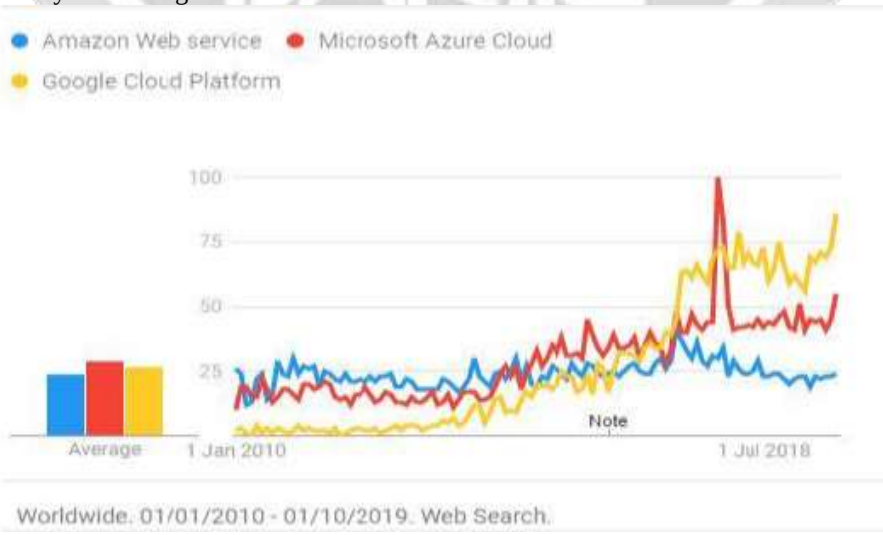
o While Microsoft Azure provides the largest number of ML framework support, Amazon EC2 has the largest number of pre-configured OSes.

o Google Cloud Platform boasts the highest runtime.

Support:

o All three platforms provide a lot of support in the form of forums and documents.

The three cloud platforms as compared to the above have their own merits that make them good in their own way. While Amazon EC2 is the oldest and has support for the maximum number of pre-configured operating systems, it is lacking in accessibility and support availability. Similarly, Google Cloud Platform supports maximum number of databases and has a large repository of in-built libraries, lacks SDK support and has a pay-to-help model that optimizes support delays according to the level of service received.

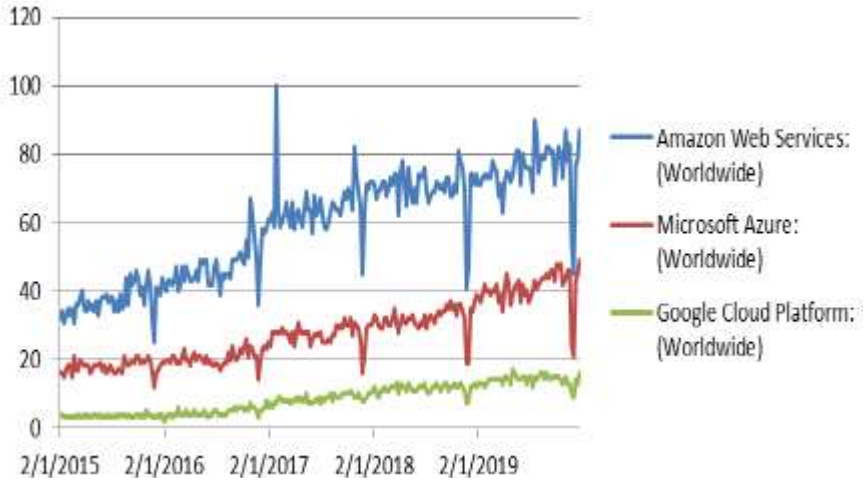


5.2 Google Cloud computing architecture Platform's for Big Data Analytics Products from AWS, Azure and GCP

AWS is one of the oldest cloud platform in the market and one of the well known cloud platform available. So AWS is widely available. Amazon Web Services (AWS) has 63 Availability Zones around the world.

Google will provide cloud computing services to its customers. Services provided by GCP include Storage, Big Data, Database, Analytics, Cloud AI, Network, Mobile Computing, Development Tools, Management Tools, Internet of Things, Cloud Security and Data Transfer. Google Cloud Platform (GCP) has 21 availability zones around the world [9] [10].

Netflix, Facebook, BBC, Adobe, Twitter, BMW, Disney, Expedia and many more. Microsoft Azure also has list of clients like Delaware Resource Group, Erickson Advisors, Hudson River Fruit Distributors etc. Google Cloud Platform (GCP) clients are Spotify, HSBC, Snapshot, HTC, Philips, Coca Cola, Domino's, Sony Music, etc[13].



6. CONCLUSION

The cost-benefit between traditional data-center computing and cloud computing is a strong reason for its huge growth. One can immediately start a project with a small investment and scale according to the growth of the business. It reduces costs associated with setting up or renting a data center and purchasing servers and other IT infrastructure, which takes several weeks to begin. With cloud computing, a business can quickly spin hundreds of servers in minutes. That is why this data is needed. In cloud computing, resources such as servers, storage, networks, and many other business applications are easily provisioned without any human interaction. Billing of these services is calculated per minute or per hour based on different services offered by different cloud computing platforms. Companies offering these computing services are known as cloud service providers and typically, they charge for cloud computing services based on the same usage of electricity or water billing in homes. In 2015, the worldwide cloud computing market grew by 28% to 110B in revenue. Synergy Research Group found that public IaaS / Paa services achieved the highest growth rate of 51%, followed by private and hybrid cloud infrastructure services, which grew by 45%. The features offered by CSP companies are discussed as well as a comparison of AWS, GCP and Microsoft Azure. The purpose of CSPs is to compare and highlight the service features of Microsoft Azure, Google Cloud Platform and Amazon Web Services that are important to organizations when choosing a CSP.

REFERENCES

[1]. P. Srivastava and R. Khan, "A Review Paper on Cloud Computing," Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 8, p. 17, Jun. 2018.

[2]. [2] A. Mazrekaj and I. Shabani, "Pricing Schemes in Cloud Computing : An Overview," vol. 7, no. 2, pp. 80–86, 2016.

[3]. A. Behl and K. Behl, "An analysis of cloud computing security issues," 2012, pp. 109–114.

[4]. https://phoenixnap.com/blog/orchestration-vs- automation.

[5]. https://mindmajix.com/aws-architecture

[6]. B. Rochwerger et al., "The Reservoir model and architecture for open federated cloud computing,"IBM J. Res. Dev., vol. 53, no. 4, 2009.

[7]. X. Liu, "Cloud architecture learning based on social architecture," 2011, pp. 418–421.

[8]. M. Klems, J. Nimis, and S. Tai, "Do clouds compute? A framework for estimating the value ofcloud computing," vol. 22 LNBIP. Springer Verlag, Forschungszentrum Informatik (FZI), Haid-und-Neu-Str. 10-14, Karlsruhe 76131, Germany, pp. 110–123, 2009.

[9]. C. Zhang, A. Yin, Y. Wu, Y. Chen, and X. Wang, "Fast Time Series Discords Detection with Privacy Preserving," 2018, pp. 1129–1139.

[10]. T. C. Y. Chui, D. Siuta, G. West, H. Modzelewski, R. Schigas, and R. Stull, "On producing reliableand Affordable numerical weather forecasts on public cloud- computing infrastructure," J. Atmos.Ocean. Technol., vol. 36, no. 3, pp. 491–509, 2019.

[11]. X. Yang, S. Zhu, and X. Pan, "Improved verifiability scheme for data storage in cloud computing,"Wuhan Univ. J. Nat. Sci., vol. 16, no. 5, pp. 399–404, 2011.

[12]. J. Ellman, N. Lee, and N. Jin, "Cloud computing deployment: a cost-modelling case-study,"Wireless Networks. Springer New York LLC, Department of Computer and Information Sciences,Faculty of Engineering and Environment,