IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.105175

Cloud Computing for Big Data Processing: A Review

Aditya B S[1], Nagaraj G Cholli[2]

Student, Information Science and Engineering, R V College of Engineering, Bengaluru, India [1]

Associate Professor, Information Science and Engineering, R V College of Engineering, Bengaluru, India[2]

Abstract: Cloud computing can be described as an on-demand availability of computer resources to perform large-scale and complex computing which helps maintain dedicated space, computer hardware and software in an efficient manner. Big data cannot be easily handled and processed using traditional methods but can be utilized and controlled better with the help of the cloud. In this paper, we will see why big data handling and processing is better in a cloud environment in a sequence starting with simple illustration of how big data can be uploaded and linked to the cloud and on-premises applications using IPaaS, hybrid cloud solutions for encapsulation and data security and finally multi-cloud solutions for improving the overall performance of big data processing in the cloud environment with their advantages and disadvantages, respectively.

Keywords: Cloud computing, Big Data, Big data analytics, Hybrid cloud, multi cloud.

I. INTRODUCTION

Many companies, agencies and businesses make use of cloud services to store and access large amounts of data faster via the internet and process them. There are companies that provide cloud services to the clients or the product the client requests may use cloud architecture that makes it easy to access computer resources and multimedia contents. The service provider may in-turn ask for service charges called a subscription fee for the client to use the product or the service. There are various types of cloud services such as SaaS, PaaS, IaaS which help in satisfying a variety of requirements and these are known as service models. There is another type of model called the deployment model which contains four major types: public, private, community and hybrid. These models help with the deployment of the cloud service which can include infrastructure support needed for the usage, accessibility, and reliability of the services.

Big data can be described as large amounts of complex data which can be unstructured or structured generated by various sources and this data increases at rapid rates with respect to time. Big data processing cannot be done easily with traditional programs due to resource restrictions (time, computing resources) and need specialized algorithms and databases. Big data analytics can be used to process big data using advanced analytical techniques like data mining, machine learning, statistics, data fusion and integration etc. In this paper, we will look into how cloud computing can support processing big data and what cloud architectures can be used to do so.

II. UPLOADING BIG DATA INTO CLOUD

Big data and cloud computing can be said to be hand-in-hand. With the help of cloud, Big Data can be broken down into several parts and can be independently processed in distinct servers that help distribute data and reduce the workload. Related data can be brought together and processed in clusters of clouds. Cloud computing therefore can be said to be one of the best solutions for handling, processing, and analysing Big Data in time-bound, distributed computing requests. It provides necessary infrastructure for processing and secluding large amounts of possibly sensitive data from external threats and helps secure and sustain rapid growth for businesses and organizations.

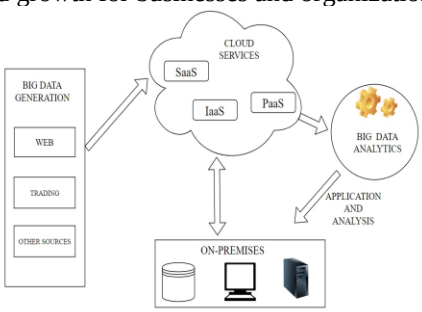

Figure. 1 Cloud architecture for Big Data

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.105175

Figure 1 shows an overview of cloud architecture for handling and processing big data. Large volumes of data are generated every second from web, trading, organizations, and various other sources as shown in 'Figure 1' which can be safely transferred to the cloud environment and can be manipulated and accessed when needed with the help of computers while being connected to the internet. It is a tedious task for many enterprises to maintain their own hardware infrastructure for computing and storage needed to handle and process data they generate on-premises (as shown in Figure 1) or remotely. Hence, they rely on third party organizations or companies that provide cloud services, in turn paying the cloud service providers for cloud storage or services they use dynamically.

Having a cloud environment for processing Big Data comes with a lot of benefits. Large databases and servers provided by cloud like Amazon Web Services, Microsoft Azure, Google Cloud Platform, MongoDB Atlas etc. helps introduce the scalability needed while computing and handling data. Real time data analysis and processing is possible with quick access to data via cloud. The necessary infrastructure provided by cloud is cost effective since it can be provisioned dynamically and can scale based on requirements. Finally, huge amounts of data stored in the cloud servers are encrypted and hence are secure.

Cloud solutions might not always work out. In many instances, it might not be a suitable solution. Some organizations cannot tolerate cloud failures as it can have a long-lasting effect on services provided by that organization and can result in the mix-up or loss of sensitive data. Many times, migrating data on-premises to the cloud can be difficult as it needs large bandwidth that can cost the organization time and money. Thus, overall expense of cloud services could exceed the originally estimated cost for the organization. The necessary infrastructure provided by the cloud should be monitored, maintained and swift actions and backup plan must be in place to handle cloud failures that can be caused due to power failures, hardware failure and server crashes which may need a lot of manpower and pile up on expenses.

A. IPaaS for Big Data Integration

Heterogeneity and volatility can be considered key properties of Big Data. We need an efficient and secure way to push data from multiple varying sources to the cloud for processing and managing it in a cost-effective manner. IPaaS (Integration Platform as a service) is one such way we can do this. IPaaS is a type of cloud service which can help integrate on-premises systems with the cloud platforms. With this, no specialized tools, services, or middleware are required connecting, uploading data, running computations, and performing analysis.

IPaaS integration involves two distinct connections. One type of connection is to link on-premises data sources with the cloud services. The other type of connection is between two or more clouds. This way, multiple cloud architectures can be put to work together, each handling the kind of data processing that it is specialized for. This gives rise to multiple possible solutions and an appropriate solution can be selected based on cost effectiveness, time and computing resource limitations or complexity of data.

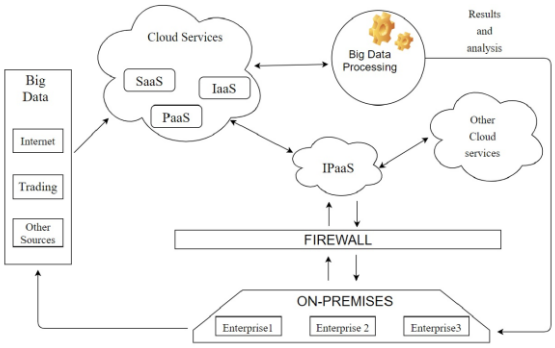

Figure. 2 IPaaS for big data processing

Since it helps integrate all sources, it is easy to monitor flow of data and ensure data security. The only major issue that can take place is accidental leakage of sensitive data. The probability of this happening is extremely low and can be caused mainly due to human error while interfering with the working of the system. This has a simple solution. Proper training can be given to the IT specialists handling data flow to reduce human error. Let us now see some ways big data can be handled and processed.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.105175

III. PROCESSING BIG DATA IN CLOUD ENVIRONMENT

A. Hybrid Cloud Solutions

Hybrid cloud is a cloud computing technique in which one private cloud and one public cloud are combined as shown in Figure 3 with ground (on-premises) applications for combined computation and storage. There is interaction in the hybrid cloud between its deployment models. The main goal here is to increase the agility of data processing. Here, the hybrid cloud uses benefits from both public and private clouds for processing data and handling data security.

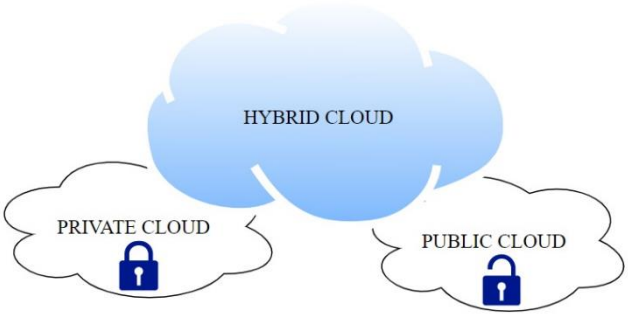

Figure. 3 Hybrid cloud

The hybrid design makes sure that there exist boundaries between private and public clouds and certain information that need not go public can be stored in the private (dedicated cloud) cloud space. Depending on the requirements, data processing can be done dynamically either in a public cloud or in the private cloud space. This ensures that sensitive data can be securely stored in the private cloud with a firewall creating a barrier and cannot be accessed by the public cloud.

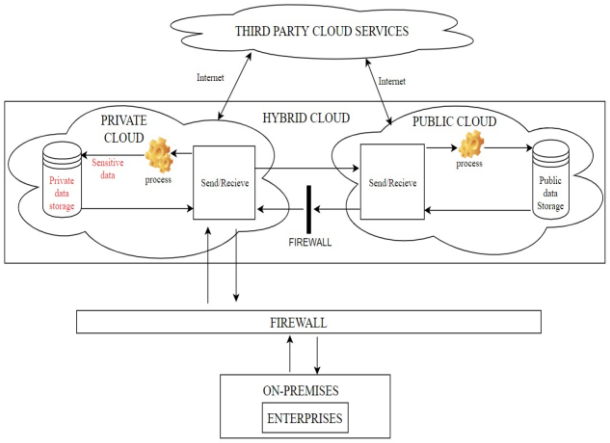

Figure. 4 Hybrid cloud architecture for big data processing

Figure 4 shows a possible hybrid architecture for big data processing. The on-premises data stored securely passes through a network security barrier (Firewall) as shown in 'Figure 4' and data is received on the private cloud space (also called dedicated cloud space). The data if not sensitive can be processed in the public cloud space, is sent to the public cloud, and is processed and stored there. If there is sensitive or high priority data that needs to be processed in the private cloud space it can go through a firewall and only the required data to be processed in the private cloud space is accepted and the other unnecessary data or the ones that can be done on the public side itself is denied entry as illustrated in 'Figure 4'. This way, sensitive data can be securely processed and stored in the dedicated cloud space, workload on the private cloud space is reduced and is made sure that no harm is done to the sensitive data.

If the data needs to be accessed by on-premises devices, it can be safely sent to the device requesting the data. There is also connection to third party cloud services via the internet, if needed for any computation which is usually connected from the public cloud.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.105175

B. Multi Cloud Solutions

There exists a special case here. In this Hybrid design, if there exists more than one public cloud and one or more private cloud(s), then it is called a multi-cloud solution. One of the main differences between the hybrid cloud and multi-cloud is that in hybrid cloud there exists only a single cloud service provider and a public cloud that can be connected to on-premises applications.

In multi-cloud, there can exist more than one cloud service provider and based on the demand, need, flexibility, scalability, agility, and the dynamic processing required, the multi-cloud can make use of other cloud services for certain data processing and particular problems can be fixed by the cloud service providers that specialize in solving that problem. This type of complex computing can enhance processing speed while maintaining integrity and security in a balanced and efficient manner and helps finding best solutions for problems. This can increase response time and optimize the workload and improve the overall performance of the system.

The main purpose of multi-cloud systems is to distribute, spread across data and its resources used for computing and to cut back on the downtime and curb data losses. Organizations exploit this solution by customizing it to meet their own business requirements and strategies and hence benefit from this.

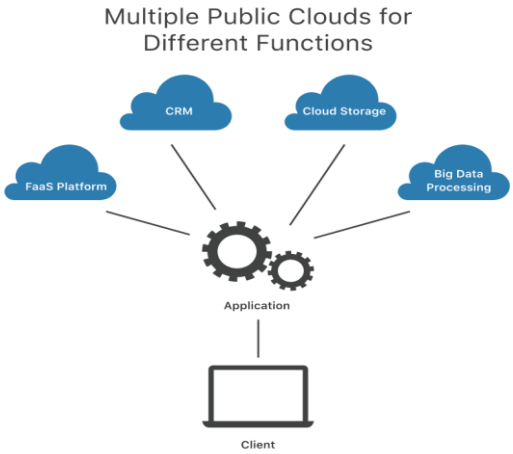

Figure. 5 Multi cloud solution for big data processing (Source: Cloudflare.com)

However, it comes with its own problems and limitations. Since there exists multiple public clouds in the multi-cloud architecture, many times not all the public clouds will be utilized to their fullest extent for processing data. This can result in extra expenses for public clouds that are not used for processing data and can be made use of elsewhere.

Another big task in multi cloud computing is handling those clouds. In cases where several of both public and private clouds are used, it is not an easy task to manage these many even though they are all linked as separate cloud services to a single private link distribution connected to the on-premises data sources. Therefore, extra tools and software needs to be used by the private linking cloud for handling the various clouds and data flow traffic which makes the entire system complicated and makes it a tedious task during regular maintenance to fix issues.

There comes another issue with several public clouds being deployed. There can be several accidental data leaks and due to the presence of multiple public clouds and care must be taken to track the data being processed by each public cloud resource to pinpoint which public cloud had the security violation. Sometimes, if this accidentally leaked data turns out to be sensitive data, then it is liable to a risk of security breach and data corruption and to prevent such an issue special security measures need to be in place and proper protocol needs to be present.

Moreover, with more multi-cloud solutions being deployed, it reduces on-site maintenance that is needed for maintaining and running the system. The personnel responsible for these multi-cloud systems need to be experts or well trained. If they lack the necessary skill needed to run the system, then it can lead to human errors that can have long lasting effects on the entire system and it can become difficult to trace the problem and fix it in reasonable time.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 5, May 2021

DOI 10.17148/IJARCCE.2021.105175

IV. CONCLUSION

This paper explains why cloud computing is an essential tool for handling Big Data and contains different cloud computing techniques for storing and handling big data. Cloud solutions provide lots of benefits to organizations by providing them infrastructure for storage, computing and other cloud services which help in rapid data processing, fast access to data when needed and overall an economical solution. It is this solution adopted by organizations across the globe. In this paper, we start with basic cloud architecture for transferring Big Data to the cloud and IPaaS solution to connect Big Data with cloud services and on-premises applications that help with processing and handling Big Data. Then we move onto processing Big Data using hybrid cloud and multi-cloud solutions that provide a faster, safer and an efficient solution helping to increase the overall performance of the system.

Cloud computing techniques not only help in providing solutions to upload and process big data but also help to securely store and access them when needed. With ever increasing demand for cloud solutions for Big Data, a lot of research and development is going on in this domain for better cloud computing solutions needed by organizations.

REFERENCES

[1]. Shrivatsa D Perur, Venkatesh H, Nivedita Jalihal, "A Study on Use of Big Data in Cloud Computing Environment",Venkatesh H et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3), 2076-2078, 2015.
[2]. Samar Wazir, Subia Saif," Performance Analysis of Big Data and Cloud Computing Techniques: A Survey",International Conference on Computational Intelligence and Data Science, 2018.
[3]. Kenan Matawie, Bahman Javadi, Rodrigo Calherios,Rekha Nachiappan,"Cloud Storage Reliability for Big Data Applications: A State of the Art Survey ", Journal of Network and Computer Applications, http://dx.doi.org/10.1016/j.jnca.2017.08.011.
[4]. Dr.Ilango Paramasivam, Shanmugasundaram Palanimalai, "An enterprise oriented view on the cloud integration approaches – Hybrid cloud and Big Data", 2nd International Symposium on Big Data and Cloud Computing, 2015.
[5]. Gokay Saldamli, Lo'ai A. Tawalbeh,"Reconsidering big data security and privacy in cloud and mobile cloud Systems", Journal of King Saud University – Computer and Information Sciences, 2019.
[6]. L. Zhao, Z. Zhou," Cloud computing model for big data processing and performance optimization of multimedia communication",Computer Communications (2020),doi: https://doi.org/10.1016/j.comcom.2020.06.015.
[7]. Anna Ciampolini, Daniela Loreti,"A Hybrid Cloud Infrastructure for Big Data Applications",IEEE 17th International Conference on High Performance Computing and Communications (HPCC), 2015.
[8]. Anna Rozeva, Svetoslav Zhelev, "Big data processing in the cloud – Challenges and platforms",AIP Conference Proceedings 1910, 060013, 2017.