International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Serverless Computing Platform for Big Data Storage

Annie Christina. A, A.R. Kavitha

Abstract: ***This paper describes various challenges faced by the Big Data cloud providers and the challenges encountered by its users. This foreshadows that the Serverless computing as the feasible platform for Big Data application's data storages. The literature research undertaken focuses on various Serverless computing architectural designs, computational methodologies, performance, data movement and functions. The framework for Serverless cloud computing is discussed and its performance is tested for the metric of scaling in the Serverless cloud storage for Big Data applications. The results of the analyses and its outcome are also discussed. Thus suggesting that the scaling of Serverless cloud storage for data storage during random load increase as the optimal solution for cloud provider and Big Data application user.***

Keywords: Serverless computing, Big Data, Data Storage, Big Data Cloud, Serverless cloud

I. INTRODUCTOIN

In recent years it has been noted that there has been a drastic increase in the volume of data captured by organizations, Industries and the high devices used by a common man. In some examples such as the rise of Internet of Things (IoT), social media multimedia [12], smart homes, smart cities etc., this is resulting in an overwhelming flow of data both in structured or unstructured format. Data creation is taking place at a rapid rate [16], which is referred to as Big Data. Thus, Big Data has emerged as the widely recognized trend. Big data had drawn its attention from various domains such as the academia, government, and industry. Thus, Big data can be characterized into three simple aspects namely (a) data which are captured, and processed rapidly, (b) data which cannot be sort out into regular relational databases, and (c) data that are copious. Moreover, big data is transforming healthcare, science, engineering, finance, business, and eventually, the society. The advancements in data storage and mining technologies allow for the preservation of data held by organizations [12]. The un-malicious secure storage of data is staggering [14].

A. Big Data Cloud Providers

Cloud services are globally distributed, which ensures a high availability of resources, this is the assurance given by the most well-known organizations.

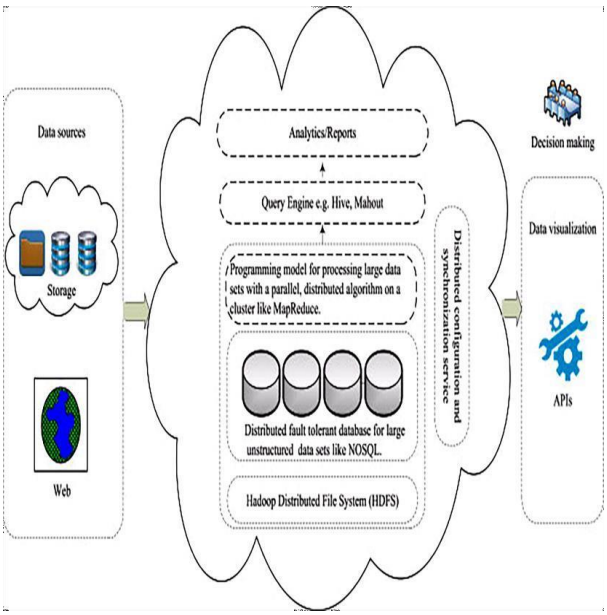

Fig.1.Big Data in cloud

Emerging cloud computing space like the Google Cloud Platform, Microsoft Azure, Rackspace, or Qubole etc [3] are discussed in Table.1. Recently, dominated by Amazon Web Services dominates the cloud computing space for Big Data, shown in Fig. 1. Foremost challenge for researchers and practitioners is that the growth rate of data exceeds their ability to design appropriate cloud computing platforms for data analytics and scale storage [13].

B. Challenges in Big Data Cloud Environment

Professional cloud storages require highly available, highly durable, and scalable system from few bytes to increasing sizes of data.

There are some prominent challenges Challenge 1:

Vertical scaling of the storage in cloud, demanding high storage elasticity to the users requirement.

Table.1. Outlook on recent Big Data Cloud platforms

Requirements given below	Google	Microsoft	Amazon	Cloudera
Big Data Storage	Google Cloud Services	Azure	S3	---
MapReduce	AppEngine	Hadoop on Azure	Elastic MapReduce (Hadoop)	MapReduce YARN
Big Data Analytics	BigQuery	Hadoop on Azure	Elastic MapReduce (Hadoop)	Elastic MapReduce (Hadoop)
Relational Database	Cloud SQL	SQL Azure	MySQL or Oracle	MySQL, Oracle, PostgreSQL
NoSQL Database	AppEngine Datastore	Table Storage	DynmoDB	Apache Accumlo
Streaming Processing	Search API	Stream-insight	Nothing Prepackaged	Apache Spark

Manuscript published on 30 September 2019
* Correspondence Author
Annie Christina .A, Computer Science and Engineering, Vel Tech Multitech Dr.Rangarajan Dr.Sakunthala Engg. College, Chennai, India anniechristinaa@gmail.com
Kavitha A. R., Computer Science and Engineering, Saveetha Engineering College, Chennai, India arkavithabalaji@gmail.com

Challenge 2:
The interchangeability of the resources with distributed software design and scaling of virtual computing instances.
Challenge 3:
Data mining operations should be mindful of denial of service attack during the process.
Challenge 4:
Data replication with zero room for error, if not it can affect the data analytics stage. Cost is high in dedicated cloud server [9].
In recent years, major cloud vendors have adopted Serverless [17] platforms such as Amazon Lambda, Google Cloud Functions, Microsoft Azure Functions, IBM Open Whisk [15] etc.
Serverless computing is a cloud execution model, suggested and used by the cloud providers to simplify allocation and management of resources. Serverless cloud computing is an emerging model where the user-defined functions are seamless and transparent. It's hosted and managed by a distributed platform [8].

II. LITERATURE RESEARCH

The literature research is about Serverless cloud platform, design and architecture, which are discussed below.
Eun-Sung Jung et al [11] proposed Serverless data movement architecture, in which data transfer nodes are bypassed, the data is moved to the file system stack, and the host system stack. Thus the data is directly moved from one disk array controller to another in order to obtain the highest end-to-end performance. Under the current data movement architecture followed in cloud, separate data transfer nodes arbitrate data transfer by input/output to parallel file systems over local area networks. Parallel file systems read/write actual data from/to disks done through disk controllers. The proposed Serverless cloud architecture embeds parallel file systems and data transfer processes into a disk controller to eliminate the data path between a data transfer process and a parallel file system. This prevents the network between those two entities from being a in end-to-end data transfer bottleneck between data transfer processes and parallel file system servers.
Garrett McGrath et al [5] developed a Serverless platform to learn about its considerations during implementation and endow a base-line for the existing platform. .NET was implemented in the platform, and then deployed to Microsoft Azure [4] with simple design and small feature-set.
The Azure Storage is used in messaging layer. The implementation consists of two components, the web services and worker services. The web service exposes the public REST API of the platform and the worker service manages and also executes the function containers. The web service discovers all the available workers through the messaging layer which consist various Azure Storage queues.
Azure Storage tables stores the Function metadata, and the Azure Storage blobs stores the Function code. An overview of the platform's components such as the web services messaging layer, worker services, code & meta base storage and messaging layer are shown in Fig. 2. High scalability and low-latency of storage was achieved by using Azure Storage. Azure Storage was used to provide primitives through a simple API that aligns well with the goals of the implementation.

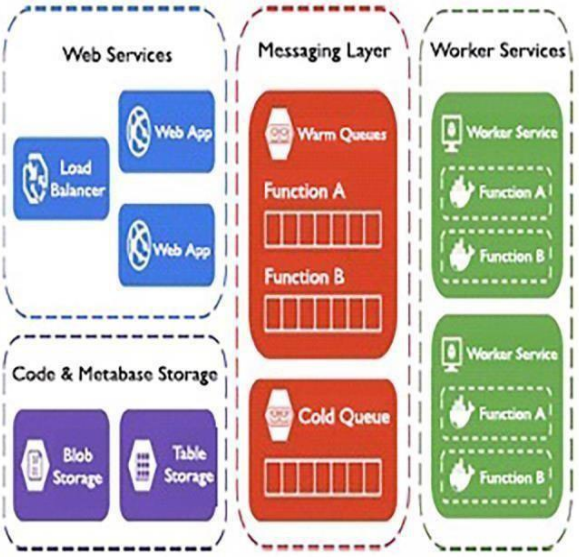

Fig.2.Shows an overview of the platform's components

Afifa Ghenai et al [1] propose that the Multi Cloud is an efficient solution; as it combines the diverse benefits from the different platforms complement each other on behalf the client applications rather than the traditional cloud solutions which consists of hiring resources from only a single Cloud provider. Serverless Function technology is a Cloud tool that hides the needless infrastructure management details, and thus allowing the developers to solely focus on the functional code. The paper discusses that the Serverless Functions limited to the provider and it does not adopt in a Multi Cloud context. In this paper, limit is tackled by suggesting a distributed architecture which extends the Serverless technology advantages to a wider scope, permitting the client to get at time the Multi Cloud and Serverless strengths.
Wes Lloyd1 et al [2] discuss the performance implications Serverless computing Infrastructure used for micro-service hosting. In particular, they profile the micro-service hosting and its implications during infrastructure elasticity, load balancing, provisioning variation, infrastructure retention, and memory reservations. Extra infrastructure was provisioned for providing elasticity and compensate the initialization overhead of COLD service requests. Docker container initialization causes significant overhead which burdens Serverless computing platforms, especially for VM cold initialization. During service requests against WARM infrastructure, reuse of extraneous infrastructure created in response to COLD initialization stress should be avoided. Higher reuse rates corresponding with higher stress levels of micro service requests, with respect to load balancing of requests against Serverless infrastructure needs to be well balanced. Distribution across containers is done in host VMs for COLD service invocations and in WARM service invocations at higher calculation stress levels. For low stress WARM service invocations, the load distribution across the hosts is uneven. This uneven use of infrastructure should not be used, for it would lead to early deprecation when the client workloads do not utilize all the nodes.

Serverless data movement architecture using parallel file system servers to prevent bottle neck, performance-oriented Serverless computing platform to comprehend Serverless implementation using Azure, Serverless Function Technology which permits Mutli-cloud with Serverless strengths and finally, the micro-service hosting and implications of infrastructure elasticity are discussed above in this section. Thus, provides understanding in various architecture, design and operations of Serverless cloud.

III. SERVERLESS FRAMEWORK

A Serverless framework is proposed for Big Data applications data storage at random intervals. The performance of the framework was considered for applications on AWS Lambda, Microsoft Azure and Google Cloud to provide FaaS (Function as a Service). FaaS cloud platform allows the customers to develop, run, and manage application functionalities involved in building and launching an App. FaaS platform is one way of achieving a Serverless architecture.

The databases used in Serverless Framework are RDBMS (Relational Database Management System). RDBMS liminates the need for provisioning or scaling the virtualized/physical database hardware. Examples of Serverless databases are: Azure Data Lake is a highly scalable data storage and used for analytics service, Google Cloud Data store which is a component of Google App Engine, Firebase is a hierarchical Google database and Amazon Aurora Serverless is an on-demand auto-scaling configuration etc.

A. The benefits of Serverless cloud

a. No server management
b. Flexible scaling
c. High availability
d. Zero idle capacity and
e. Reduced operational cost

These are some of the benefits of the Serverless cloud [7].

B. Applications Using Serverless Framework

Application1: Multiple devices accessing various file types [4] Serverless Framework can be applied to applications that involve the use of multiple devices accessing various file types, such as mobile phones an PCs uploading images, videos, and text files etc.

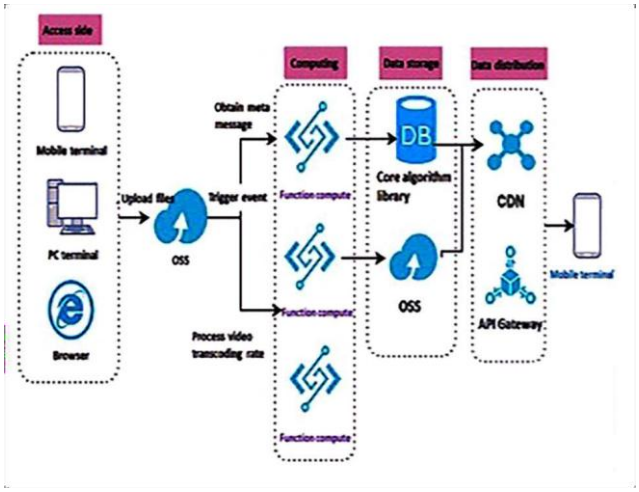

Fig.3. Application on multimedia device using Serverless Framework

Application 2: IoT enabled systems Serverless architecture [10] is feasible and can be applied to Internet of Things (IoT). It is best optimal for data processing applications like the Smart home, Wearable's, Smart City, and Smart grids etc.

Fig. 3 shows the work flow of the multimedia device application using the Server- less framework. The work flow is explained through 4 simple stages: access side, computing, device storage and device distribution.

IV. PERFORMANCE OF COMPUTING SERVERLSS PLATFORM FOR BIG DATA STORAGE

Serverless computing platform was analyzed to check its performance in the storage of Big Data set. This was monitored under various workloads. The workload can be of various sized Big Data set clusters. Performance of various clustered was monitored through cloud simulation tools. Simulation tools are easier for understanding complex scenarios, large environments simulation that represents both non-existent and existing cloud architectures. There are various well known cloud simulation platforms like CloudSim, CloudAnalyst, GreenCloud etc. The tool which we used for analyzing various sized Big Data sets is iCan Cloud simulator.

iCanCloud simulator is used to simulate cloud computing systems. This simulation platform is built on OMNeT++ and INET frameworks, both are required to operate the simulator. Existing and non-existing cloud computing architectures can be modeled with an easy method of integration and can be simulated for results in iCanCloud simulator. The GUI of iCanCloud is a user-friendly platform, this GUI gives the ease in generation and customizationof large distributed models.

A. Evaluation Scenarios

One of the major concerns for using cloud for Big Data is scaling of storage, which is considered feasible in Serverless computational platform. This was analyzed by creating cloud clusters under different data sizes. There are 3 cloud clusters which were used: large cluster, medium cluster and small Clusters with nodes being allocated for storage and computation.

B. Observation: While simulating the small clusters, it was noted that the small clusters performed well for small size data sets of multiple formats. The small cluster was could not handle as the workload began to rise and there was a time delay. This indicated that small Serverless cloud cluster required scaling. On this observation, the small cluster was scaled by adding racks and nodes for storage, as seen in Fig.4.

This step taken prevented time delay in processing the request. Thus, enabling the Serverless cloud users to scale the storage capacity as required on demand. Another benefit of this process is that the Serverless cloud users don't have to pay for unused racks, they can pay as per their use and scale as they need.

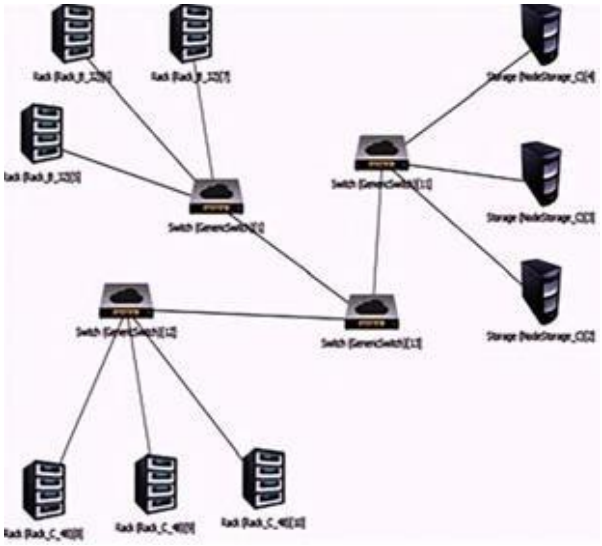

Fig.4. iCan Cloud simulation of scaling small cluster

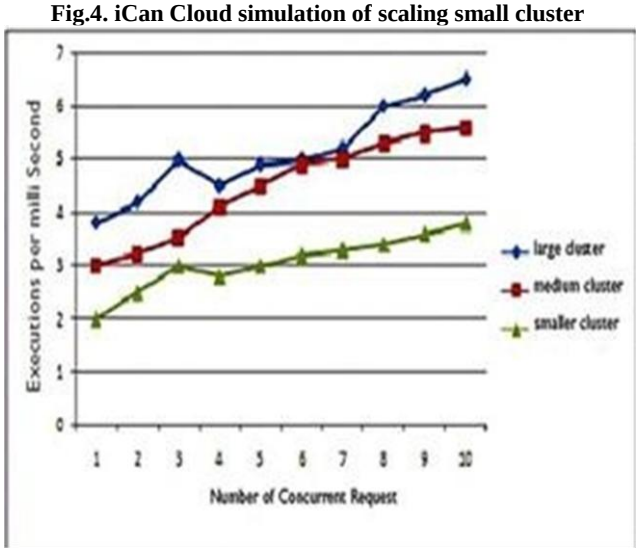

Fig.5. Performance of Serverless Cloud on scaling the cloud storage due to increase in data load

Results obtained while scaling small cluster of Big Data datasets are shown in Fig.5. The graph shows the performance of Serverless cloud under rapid increase in data load to the cloud via concurrent requests. The small cluster was scaled and increased to medium and large cluster of cloud storage as the workload increases. The performance was observed to be good and with no notable data loss. This result shows that the elasticity of Serverless cloud computing platform is ideal for storing Big Data.

V. CONCLUSION

This paper highlights the use of Serverless cloud platform for the storage of Big Data addressing the challenge of scalability of data storage. The performance of Serverless cloud platform for the storage of Big Data was analyzing using iCanCloud simulator. The simulation result was noticed that the challenges of scaling the storage cloud for Big Data are optimal through Serverless framework, for storage of large Data sets. Big Data generating organizations/industries can depend on Serverless platform for Big Data storage. Further research has to be done for data replication and data security for Big Data application. Data replication with zero error as it can affect the data analytics stage. It is fundamental to make the searching, sharing, transfer, storage, visualization and analytics of data in a best feasible manner as possible. Self-healing processes is needed to detect and repair errors and also device failures.

REFERENCES

1. Afifa Ghenai et al, "Towards Distributed Containerized Serverless Architecture in Multi cloud environment", Elsevier, Procedia Science Direct, Volume 134, 2018, pp. 121-128.
2. Wes Llyod et al, "Serverless computing: An Investigation of factors influencing microservice performance", *IEEE International Conference on Cloud Engineering (IC2E)*, April 2018.
3. Big Data cloud database and computing, 2018 [online] Available: https: //www. qubole.com/ resources/big-data-cloud-database-and-computin/
4. Leona Zhang, "4 use cases of Serverless architecture", Aug 2018, [online] Available: https://dzone.com/articles/ 4-use-cases-of-serverless-architecture.
5. Garrett et al, "Serverless Computing: Design, Implementation, and Performance", *IEEE 37th International Conference on Distributed computing Systems Workshops*, 2017.
6. Serverless Framework, February 2017, [online] Available: https://serverless.com.
7. Adam Eivy, "Be Wary of the Economics of Serverless Cloud Computing", *IEEE journal of Cloud Computing*, vol. 4, issue no. 2, pp. 6-12, March.– April. 2017.
8. Joe Weinman, "The Evolving Cloud", *Cloud Computing IEEE*, vol. 4, pp. 4-6, 2017.
9. G. McGrath, J. Short, S. Ennis, B. Judson, and P. Brenner, "Cloud event programming paradigms: Applications and analysis," 2016 *IEEE 9th International Conference on Cloud Computing (CLOUD)*, June 2016, pp. 400–406.
10. A. Avram, "FaaS, PaaS, and the Benefits of the Serverless Architecture,"[Online] Available: https://www.infoq.com/news /2016/06/faas-serverless-architecture
11. Eun-Sung Jung et al, "High-Performance Serverless Data Transfer over Wide Area Networks", 2015, *IEEE International Parallel and Distributed Processing Symposium Workshops*.
12. Ibrahim Abaker Targio et al, "The rise of big data on cloud computing: Review and open research issues" *Information Systems 47 (2015) Elsevier*, pp.98–115
13. R.Cumbley et al, " Is BigData creepy? ", Computer Law & Security Review, Volume 29, Issue 5, October 2013, pp.601-609
14. S.Kaisler et al, "BigData: Issues and challenges Moving Forward, 46th Hawaii International Conference on System Sciences, *IEEE Computer Society Digital Library*, 2013, pp.995–1004.
15. *Openwhisk*, [online] Available: https: //github.com /openwhisk / openwhisk.
16. .R.L.Villars et al, "Big Data: what it is and why you should care", White Paper, IDC, 2011, MA, USA.
17. Iuon-Chang Lin et al, "Lightweight and Serverless RFID Authentication and Search Protocol", 2009, *Second International Conference on Computer and Electrical Engineering*.

AUTHORS PROFILE

Annie Christina.A, M.Tech computer science, published in Asian Journal of Computer Science and Technology

Kavitha A. R., Ph.D (I&C), published papers in ACM, SCI, Web of Science, and Scopus indexed Transaction/Journals