A Survey on Big Data Storage in Cloud

Praywin Ebenezer.S[1] Thirunavukarasu.T[2]
[1]P.G. Scholar [2]Assistant Professor
[1,2]Department of Information Technology
[1,2]SNS College of Technology, Coimbatore, Tamilnadu, India

Abstract— In today's computing and digital world, generation and accessing of data information has been rapidly increased in several organization. Millions of people are generating huge amount of data for their need of organizations which is often referred as big data. Big Data is considered as a promising for increasing segment of IT industry and business applications includes namely health care, banking etc. Since the extent and volume of big data is large, the fixed amount of storage is not suitable for storing such dynamic generation of data. Earlier versions of technology used major storage devices such as solid-state devices, optical devices, magnetic tapes and magnetic disks. However these storage devices don't supports for dynamic data generation from various organizations. Therefore, recent technology uses cloud based storage and storage networking therewith improved data storage and mining techniques are needed to be incorporated and advanced for preserving this increased volume of data. This paper surveys various storage techniques used for storing large volume of generated data. This survey comprehensively analyses and classifies the several attributes of big data which generally includes characteristics, nature, rapid growth and volume. At last possible research directions for big data systems in near future are outlined.

Key words: Big Data, Cloud, Megastore, Dynamo

I. INTRODUCTION

Recently, Cloud computing is being exhaustively considered as one of the most leading innovations in information technology [1], [2]. By means of resource virtualization, cloud can offer computing services and resources in a pay-as-you-go mode, which is imagined to develop as suitable to use comparable to daily-life utilities such as gas, telephone, electricity, and water in the near future. These computing services can be categorized into Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) [3] and Infrastructure-as-a-Service (IaaS). Several international IT corporations now suggest authorized public cloud services to users on an extent from entity to enterprise in all over the world. Some of the authorized public cloud services are IBM Smart Cloud, Amazon AWS, and Microsoft Azure.

In today technology, big data is considered as a key area in research and utilization. Big Data analytics is defined as the procedure of analyzing and mining Big Data that could make business and operational facts at an extraordinary range and specificity. One of the main drivers for Big Data analysis tools is to analyze and influence data collected by businesses. Market research firm governed that increasing growth of data has averaged about 35 percent yearly in near future. This means that the amount of storage is essential to utilize all this information which eventually doubles on every two years, exclusive of the advances of new technologies. Recently, big data has concerned with it some difficulties. The problem is the storage and securing of huge information data in particular organization. Fore coming research issue is the unavailability of new storage technology to aid for such explosion of data. This means that existing technologies are stepped up by the usage of conventional hard drive. Actually, there is new progress that could considerably amplify storage capacity of hard drives. Cloud storage would offers directions for many organizations to hold increasing quantities of information. Usage of Object storage [4] would be efficient for storing of information to a large scale as needed. In addition solid-state drives (SSDs) provide better performance, which is significant for the various organizations that required to rapidly uses and make sense of their data. Still, several efforts and challenges are needed to be handled with the increasing volume of data.

Since various organizations are placing their huge volume of data in the cloud, it frequently places data through large data centres as a way to ease of managing, indexing, and securing objects stored in conventional, hierarchical file systems. The issue and challenge found in this conventional file system is preserving central data indices and hierarchical organization. Thus organizations are Changing to object storage, which stores variable-size objects as data as relatively as fixed-size blocks. Unlike traditional storage, the used of object storage discovers information in object storage systems related to its physical position on a disk drive. Instead, to locate and discover data objects, unique identifier has been used by object storage. This in turn presents near-infinite and non hierarchical address spaces. As a result, it has been observed that object-storage systems easily scale by without making it complex to discover information.

Particularly, the barrier occurs in storage is construction of more storage centers or improved storage technology which is nearly expensive, as people often go for cheap storage. The accepted fact is that the cost of storage isn't deteriorating quite fast. As a result, industries are dealing with this because of its dynamic behaviour. Since offering storage at very low prices would probably signify that convincing on performance and no organizations needs that. Also, rapid scaling of storage with no consequence on performance has been a challenging issue. Although, storing large volume of information in the cloud increases privacy concerns and security, then shifting to a new and improved storage environment is mandatory which is too expensive and difficult for some companies dealing with large datasets respectively.

The following literature describes various storage techniques used for storing large volume of data generated by various organization.

II. TECHNIQUES USED FOR BIG DATA STORAGE

A. Bigtable

In [5] Fay et.al presented a Bigtable which is one of the column oriented databases that stores and process data by column in place of by row. Bigtable is defined as a

distributed storage system for managing structured data which is built to the extent of a petabytes of data across thousands of servers. The fundamental structure of Bigtable is a distributed, sparse and determined multi-dimensional sorted map. The sorted map is indexed by a timestamp, row key and a column key. A timestamp is employed to distinguish deteriorations of a cell value. Then Rows are used in lexicographic order and are dynamically divided into tablets that correspond to the component of load balancing and distribution. Columns are aggregated by their prefix key into sets known as column families that characterize the fundamental unit of access control. The implementation of Bigtable's comprises of three main components namely client library, tablet server and master server. One master server is assigned for each runtime of Bigtable and is accountable for allocating tablets to servers of tablet by identifying added and disconnected tablet servers, and providing the workload across tablet servers. In addition, the processes of master server vary in the Bigtable schema, during the generation of column families and tables, and then collect garbage. Here garbage denotes expired and deleted files that are accumulated in GFS (Google file system) for the particular instances of Bigtable. At last, a set of tablets are managed by each tablet server which efficiently handles write and requests for tablets. In addition, a client library is offered for applications with Bigtable instances to interact with each other.

B. Dynamo

In [6] Giuseppe et.al presented Amazon's Dynamo which is one of the Key-value databases that have appeared in near future. Key-value stores contain a data model wherein data are stored as a key-value pair. Since each key contains unique characteristics, the clients put on values for each key. In particular, Dynamo is defined and referred as a scalable and excessively available data store which is utilized for storing state of an amount of central part services of Amazon.com e-commerce platform. Unlike traditional storage mechanism, Dynamo offers the required levels of performance and availability as well as effectively handles network partitions, server failures and data centre failures. Based on the request load of service owners, dynamo increases the scalability by scaling up and down methods. In addition, it makes the owners to modify their storage system to gather their required performance, consistency SLAs and robustness by permitting them to alter the parameters. The past use of dynamo in terms of production illustrated that the decentralized methodology can be merged which results a distinct highly-available system.

C. Cassandra

In [7] Avinash et.al presented Cassandra which is one of the distributed storage system for managing huge amount of structured data extended across numerous commodity servers, while gives that extremely available service by without failure on single point. Cassandra used together with the distribute system from dynamo and Bigtable data modal. Particularly, table in Cassandra is considered as a distributed multi-dimensional map structured with four categories namely rows, columns, column families and super columns. The goal of Cassandra is to be used on the peak of a hundreds of nodes infrastructure. During this stage, failure occurs continuously in small and large components. At this failure stage, Cassandra manages the constant state and drives the scalability and reliability of the software system which using this service. In several ways, Cassandra is similarly viewed as a database and shares several design and implementation approaches. It does not hold as complete relational data model, instead it gives clients an easy data model which in term holds dynamic strategy over data format and layout. In particular, the design of Cassandra made to execute on low-cost commodity hardware and holds throughput of write and not declining the read efficiency respectively.

D. Megastore

In [8] Jason et.al presented Megastore which is one of the storage system designed to entails the storage prerequisites of recent interactive online services. Megastore brings together the scalability of a NoSQL data in which it store with the conventional RDBMS. In order to achieve consistent view and high availability of data, Megastore uses synchronous replication. In short, it presents completely serializable ACID semantics above distant models with low sufficient latencies to sustain interactive applications. This can be attained by choosing a centre ground in the RDBMS versus NoSQL design gap. By this, the replica and data store are partitioned separately by giving complete ACID semantics inside partitions, however only limited consistency promises across them. Later, conventional database features are provided namely secondary indexes, however only those features that cans extent inside user-acceptable latency limits, and merely with the semantics that this partitioning system is able to hold up. By this way, Megastore results that the data for most Internet services can be suitably partitioned to construct this approach feasible, with a set of features that can considerably simplifies the difficulties of developing cloud applications.

E. Extent Based Dynamic Tiering

In [9] Jorge et.al presented Extent Based Dynamic Tiering for providing storage facility by minimizing operating cost and power for huge data storage management. Extent-based Dynamic Tiering (EDT) system incorporates a Configuration Adviser tool EDT-CA to compute optimized cost mixes of devices which deal a customer's workload service. And, the data can be placed by dynamically migrating scales to most appropriate tiers for a given workload by using Dynamic Tier Management EDTDTM component which can be executed in the configured storage system. EDT- Configuration Adviser tool operates by simulating the dynamic assignment of areas inside tiers that present the lowest cost to obtain an extent's Input/output requirements as they vary over time, as a result it suitable for each size of tier. EDT-DTM manages extent placement and migration and observes active workload by the way of meeting the optimized operating cost with its achieved performance which is considered as feasible by consolidating data into smaller devices in each tier and rest tiers are switched off.

F. ObliviStore

In [10] Emil et.al designed and constructed ObliviStore which is a high performance distributed ORAM-based cloud data store which protects the system in the malicious model. ObliviStore is the best and well-known ORAM

implementation runs faster by more than 10X in contrast with the efficient ORAM implementation. This storage attains high throughput by asynchronous building of Input or output operations. This type of asynchronous brings in security challenges that is one should avert information leakage not only in the course of access patterns, but also from the side of Input or output events timing. With this, several practical optimizations are presented which are considered as primary steps for attaining high performance with the data centre techniques which dynamically scales up a distributed ORAM.

G. The Non-Intrusive Load Monitor Database

In [11] James et.al presented The Non-Intrusive Load Monitor Database (NilmDB), a widespread framework designed to get involved for the problem of "big data" of non-intrusive load examining and analytics. The NilmDB data storage and management structure correspond to a change in the load monitoring systems design and implementation. It offers a whole structured, reliable, network- conscious design that facilitates the growth of actionable analytics across an extensive range of systems. This affords these services by reducing network communication demands of resources. NilmDB arranges and regulates the set and processing steps which allowing reusable and modular filter mechanism to make more efficient and simplify the operation of monitoring systems. By means of NILM Manager, NilmDB presents the solution to the analytic problems of "big data" for high-scale power monitoring system. It facilitates advanced NILM techniques that can disaggregate and account the operating list of individual loads exactly from estimations of collective existing consumption, whereas preserving low network bandwidth conditions and flexible options for computing.

H. Swift Array

In [12] Yifeng et.al presented SwiftArray which is a new storage design with indexing methods which fasten the queries with dimension and its value of sub setting conditions. SwiftArray split the multidimensional array into several blocks and then sorted values in each block are stored so as to employ the binary search to determine the necessary values for value subletting state rapidly. A Hilbert space-filling curve is selected as the block storage layout for attaining improved locality of data for dimension sub setting queries. After that a range index has been constructed by using the maximum and minimum values in each block for determining the necessary values in particular block. For this purpose, 2-D-Bin method is presented which is one of the named efficient indexing method which is powerful than sequential and 1-D-Bin methods. At last, in order to balance the performance for both dimension and value sub setting queries, an appropriate size of block has been chosen for SwiftArray.

III. CONCLUSION

The present survey presented the concepts of different storage techniques used for storing large amount of data known as big data generated by various organizations. Generally, big data cycle comprises of four stages namely Generation of data, acquisition of data, data storage and analysis of data respectively. The present survey mainly focuses on the storage of data in cloud based platform. Keeping this point, the above literature described various storage techniques namely, Bigtable, Dynamo, Cassandra, Megastore, EDT, ObliviStore, NilmDB and SwiftArray. Each storage techniques present unique characterization of storing large volume of data in cloud environment. Since data management is gaining significant role in digital data growth, the advancement of storage system technology is still need to be enhanced and improved in future IT system.

REFERENCE

[1] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, ''Cloud Computing and Emerging IT Platforms: Vision, Hype, Reality for Delivering Computing as the 5th Utility,'' Future Gen. Comput. Syst., vol. 25, no. 6, pp. 599-616, June 2009.

[2] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, ''A View of Cloud Computing,'' Commun. ACM, vol. 53, no. 4, pp. 50-58, Apr. 2010.

[3] Customer Presentations on Amazon Summit Australia, Sydney, 2012, accessed on: March 25, 2013. [Online]. Available: http:// aws.amazon.com/apac/awssummit-au/.

[4] Neal Leavitt," Storage Challenge: Where Will All That Big Data Go?," Published by the IEEE Computer Society,2013.Available: http://ComputingNow.computer.org.

[5] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber,``Bigtable: A distributed storage system for structured data," ACM Trans. Comput. Syst., vol. 26, no. 2, pp. 4:1-4:26, Jun. 2008.

[6] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall and Werner Vogels,``Dynamo: Amazon's highly available key-value store," SIGOPS Oper. Syst. Rev., vol. 41, no. 6, pp. 205-220, 2007.

[7] Lakshman and P. Malik, ``Cassandra: Structured storage system on a p2p network," in Proc. 28th ACMSymp. Principles Distrib. Comput. 2009, p. 5.

[8] Baker et al., ``Megastore: Providing scalable, highly available storage for interactive services," in Proc. Conf. Innov. Database Res. (CIDR), 2011, pp. 223-234.

[9] Guerra, H. Pucha, J. S. Glider, W. Belluomini, and R. Rangaswami, ``Cost effective storage using extent based dynamic tiering," in Proc. 9th USENIX Conf. File Storage Technol. (FAST), 2011, pp. 273-286.

[10] Emil Stefanov, Elaine Shi," ObliviStore: High Performance Oblivious Cloud Storage," 2013 , IEEE Symposium on Security and Privacy, pp. 1081-6011

[11] James Paris, John S. Donnal, and Steven B. Leeb," NilmDB: The Non-Intrusive Load Monitor

Database," IEEE transactions on smart grid, vol. 5, no. 5, September 2014

[12] Yifeng Geng, Xiaomeng Huang, and Guangwen Yang," SwiftArray: Accelerating Queries on Multidimensional Arrays," Tsinghua Science and Technology, ISSN 1007-0214 12/13 pp 521-530 Volume 19, Number 5, October 2014

