ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (1005-1009), Month: April - June 2015, Available at: www.researchpublish.com

Handling of Big Data by using Big Table

Nilesh Jaiswal

Department of MCA 3rd Year, Mumbai University, IMCOST, Thane, India

Abstract: **Cloud computing provides services to internet by utilizing resources of the computing infrastructure to provide different services of the internet. It allows consumers and businesses to use applications without installation and access their personal files at any computer with internet access. A distributed storage system for managing structured data at Google called Big table. Big table is designed to reliably scale to peta bytes of data and thousands of machines. Big table has achieved several goals: wide applicability, scalability, high performance, and high availability. Big table is used by more than sixty Google products and projects, including Google Analytics, Google Finance, Personalized Search, Google Earth and many more. In this paper a review is done to analyze the cloud performance on data stored at data centers.**

Keywords: **Cloud data, data center, map reduce, distributed file system.**

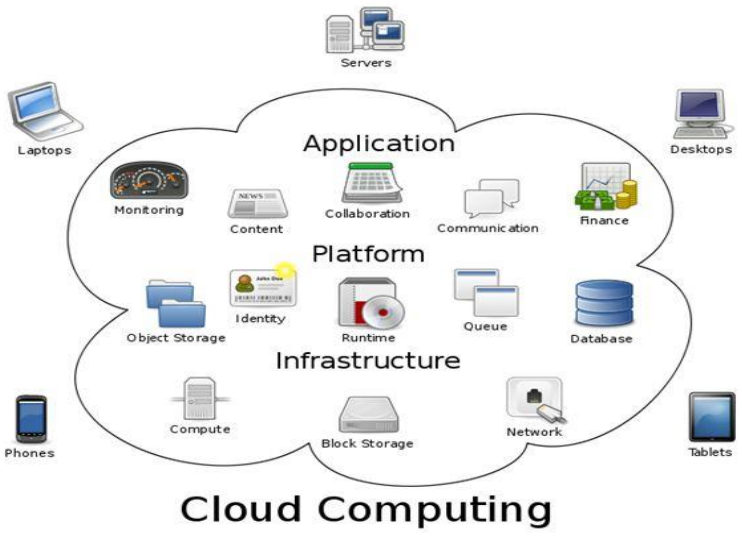
I. INTRODUCTION

The idea behind the Cloud is that users can use the service anytime, anywhere through the Internet, directly through the use of browser. In cloud computing data is stored in virtual space as it uses the browsers to use network services. Since, networks are associated so main concern is of security of data.

What is Cloud?

Cloud computing is a computing term or metaphor that evolved in the late 2000s, based on utility and consumption of computer resources. Cloud computing involves deploying groups of remote servers and software networks that allow centralized data storage and online access to computer services or resources. Clouds can be classified as public, private or hybrid.

In 2012, Gartner updated its definition as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."



A **Cloud Database** is a database that typically runs on a cloud computing platform, such as Amazon EC2, Go Grid, Sales force, Rack space, and Microsoft Azure. There are two common deployment models: users can run databases on the cloud independently, using a virtual machine image, or they can purchase access to a database service, maintained by a cloud database provider. Of the databases available on the cloud, some are SQL-based and some use a No SQL data model.

A **Data Center is** a facility used to house computer systems and associated components, such as telecommunications and storage systems. It generally includes redundant or backup power supplies, redundant data communications connections, environmental controls (e.g., air conditioning, fire suppression) and various security devices. Large data centers are industrial scale operations using as much electricity as a small town.

a) Big Data:

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many Peta bytes of data. Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale.

In a 2001 research report and related lectures, META Group (now Gartner) analyst Doug Laney defined data growth challenges and opportunities as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). Gartner, and now much of the industry, continue to use this "3Vs" model for describing big data.

Additionally, a new V "Veracity" is added by some organizations to describe it.

If Gartner's definition (the 3Vs) is still widely used, the growing maturity of the concept fosters a more sound difference between big data and Business Intelligence, regarding data and their use:

Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends etc.;

Big data uses inductive statistics and concepts from nonlinear system identification to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density to reveal relationships, dependencies and perform predictions of outcomes and behaviors.

A more recent, consensual definition states that "Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value".

Characteristics:

Big data can be described by the following characteristics:

Volume – The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered Big Data or not. The name 'Big Data' itself contains a term which is related to size and hence the characteristic.

Variety - The next aspect of Big Data is its variety. This means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data analysts. This helps the people, who are closely analyzing the data and are associated with it, to effectively use the data to their advantage and thus upholding the importance of the Big Data.

Velocity - The term 'velocity' in the context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development.

Variability - This is a factor which can be a problem for those who analyze the data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

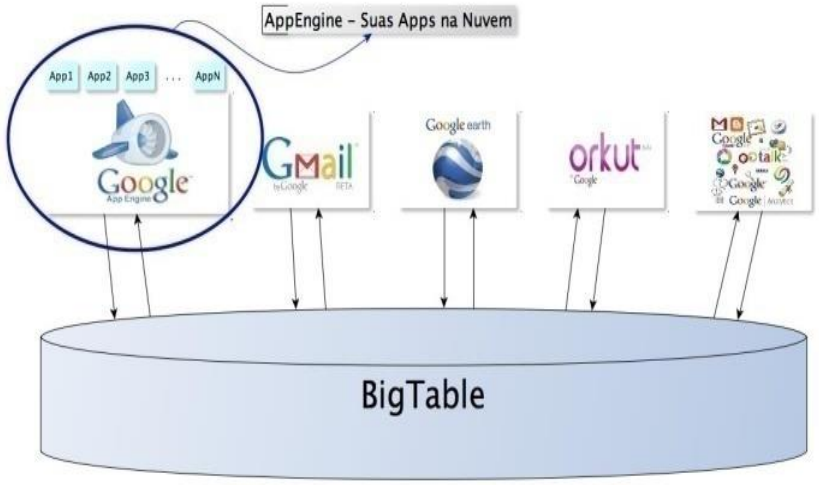
Veracity - The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (1005-1009), Month: April - June 2015, Available at: www.researchpublish.com

Complexity - Data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data. This situation, is therefore, termed as the 'complexity' of Big Data.

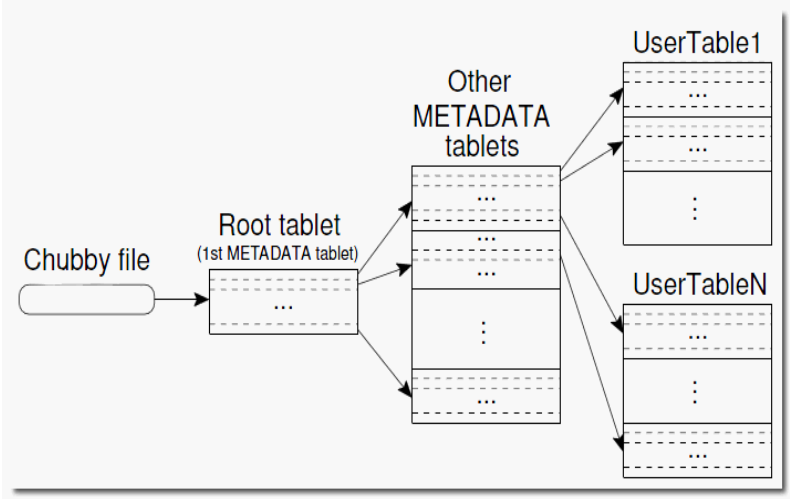
Big data analytics consists of 6 Cs in the integrated industry 4.0 and Cyber Physical Systems environment. 6C system, that is, consist of connection (sensor and networks), Cloud (computing and data on demand), Cyber (model and memory), content/context (meaning and correlation), community (sharing and collaboration), and customization (personalization and value). In this scenario and in order to provide useful insight to the factory management and gain correct content, data has to be processed with advanced tools (analytics and algorithms) to generate meaningful information. Considering the presence of visible and invisible issues in an industrial factory, the information generation algorithm has to capable of detecting and addressing invisible issues such as machine degradation, component wear, etc. in the factory floor.

b) Big table:



Big Table is a compressed, high performance, and proprietary data storage system built on Google File System, Chubby Lock Service, SS Table (log-structured storage like Level DB) and a few other Google technologies. It is not distributed outside Google, although it underlies Google Data store, which is available as a part of the Google Cloud Platform

Design:



Big Table maps two arbitrary string values (row key and column key) and timestamp (hence three-dimensional mapping) into an associated arbitrary byte array. It is not a relational database and can be better defined as a sparse, distributed multi-dimensional sorted map. Big Table is designed to scale into the Peta byte range across "hundreds or thousands of machines, and to make it easy to add more machines to the system and automatically start taking advantage of those resources without any reconfiguration".

Each table has multiple dimensions (one of which is a field for time, allowing for versioning and garbage collection). Tables are optimized for Google File System (GFS) by being split into multiple tablets – segments of the table are split along a row chosen such that the tablet will be ~200 megabytes in size. When sizes threaten to grow beyond a specified limit, the tablets are compressed using the algorithm BM Diff and the Zippy compression algorithm publicly known and open-sourced as Snappy, which is a less space-optimal variation of LZ77 but more efficient in terms of computing time. The locations in the GFS of tablets are recorded as database entries in multiple special tablets, which are called "META1" tablets. META1 tablets are found by querying the single "META0" tablet, which typically resides on a server of its own since it is often queried by clients as to the location of the "META1" tablet which itself has the answer to the question of where the actual data is located. Like GFS's master server, the META0 server is not generally a bottleneck since the processor time and bandwidth necessary to discover and transmit META1 locations is minimal and clients aggressively cache locations to minimize queries.

Issues in Big Table for Achieving ACID

Since Big Table is referred as a No SQL Data Base because it does not contain any relations among the tables. No SQL Cloud data services provide scalability and high availability properties for web applications but at the same time they sacrifice data consistency. However, many applications cannot afford any data inconsistency.

Transactions characterized by the ACID properties of Atomicity, Consistency, Isolation, and Durability are the most tractable and powerful construct for managing concurrency, allowing multiple clients to simultaneously read and write data without fear of conflicts. Transactions serve as a building block for abstractions upon which additional functionality can be constructed.

So Big Table Fail to achieve the Acid Properties.

Solution for Achieving ACID

In Cloud TPS, applications issue transactions to a Transaction Processing System (TPS), which corresponds to the transaction manager of a conventional system. The TPS is composed of a number of Local Transaction Managers (LTMs), each of which is responsible for a subset of data and which thereby distribute the load of transaction processing in a scalable manner. TPS employs the Two-Phase Commit (2PC) protocol. Its designers "observe that Cloud TPS is mostly latency-bound." Relating to the use of 2PC, the "main factors influencing performance are the network round-trip times and the queuing delays inside LTMs. Cloud TPS is therefore best suited for deployments with a single data center."

II. BASIC ALGORITHM

Commit request phase or voting phase:

1. The coordinator sends a query to commit message to all cohorts and waits until it has received a reply from all cohorts.
2. The cohorts execute the transaction up to the point where they will be asked to commit. They each write an entry to their undo log and an entry to their redo log.
3. Each cohort replies with an agreement message (cohort votes Yes to commit), if the cohort's actions succeeded, or an abort message (cohort votes No, not to commit), if the cohort experiences a failure that will make it impossible to commit.

Commit phase or Completion phase:

Success

If the coordinator received an agreement message from all cohorts during the commit-request phase:

1. The coordinator sends a commit message to all the cohorts.
2. Each cohort completes the operation, and releases all the locks and resources held during the transaction.
3. Each cohort sends an acknowledgment to the coordinator.
4. The coordinator completes the transaction when all acknowledgments have been received.

Failure

If any cohort votes No during the commit-request phase (or the coordinator's timeout expires):

1. The coordinator sends a rollback message to all the cohorts.
2. Each cohort undoes the transaction using the undo log, and releases the resources and locks held during the transaction.
3. Each cohort sends an acknowledgement to the coordinator.
4. The coordinator undoes the transaction when all acknowledgements have been received.

III. CONCLUSION

In this paper, we have discussed big table, big data in cloud, Types of big data. Research findings of various authors have been studied and their results are discussed. Big data is a useful technology which makes the work easy to handle and also discussed about the big table and how it can be achieve ACID properties.

REFERENCES

[1] Fay Chang, "Big table: A Distributed Storage System for Structured Data,"

[2] Arjun Kumar1 "Efficient and Secure Cloud Storage for Handling Big Data,"

[3] Linquan Zhang, "Moving Big Data to The Cloud: An Online Cost-Minimizing Approach,"