ISSN No. 0976-5697

Volume 3, No. 6, Nov. 2012 (Special Issue)

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Impact of Cloud Computing on Data Mining System

N Raveendran*
Scientist, NIC GoI,
New Delhi, India,
raveeng@gmail.com

Dr Antony Selvadoss Dhanamani
Dept. Of CS, NGM College, Bharathiar University, Pollachi, India
selvdoss@gmail.com

Abstract—Due to the rapid growth in complex IT system, there is a huge demand to manage and maintain the big data. The larger IT organisations invest a lot of resources (hardware and software) in the usage of data mining processes to get the hidden patterns from the data which is very costly affair for these organisations. Cloud computing has emerged as a popular computing model for processing the larger volumetric data using clusters of heterogeneous computer systems. This paper analyses the usage of various models of cloud computing that are beneficial for data mining.

Keywords: Business Intelligence, Cloud Service Provider, Quality of Service

I. INTRODUCTION

The basic understanding of data mining system is that it would carry out the automatic or semi-automatic analysis of large quantities of data to extract the unknown interesting patterns which can be very much useful in further decision making. Historically, high performance data mining systems have been designed to take advantage of powerful, but shared pools of hardware components. Generally, data is scattered to the various computer systems, the computation is performed using a message passing or grid services library. This process requires a highly efficient computing model which can suffice the necessity of big data churning.

The Cloud computing is a resilient computing model that the users can lease the resources from the rentable infrastructure which can be very well used for data mining activities. The cloud computing is a broader concept of converged infrastructure and shared services[1] that can be used for the benefits of the organisations. The Cloud computing providers offer cloud services according to three fundamental models[2] namely; Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS).

Infrastructure as a Service (IaaS) - Provides the capability to process, store, network, and other fundamental computing resources where the end user is able to deploy and run arbitrary software. The end users need not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of selected networking components.

Platform as a Service (PaaS) - Provides the capability to deploy the applications onto the cloud infrastructure. The end users need not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.

Software as a Service (SaaS) - Provides the capability to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, or a program interface. The end users need not manage or control the underlying cloud infrastructure. The cloud can be implemented in various available cloud models such as private, public and hybrid clouds.

II. CLOUD COMPUTING FOR DATA MINING

Storage Cloud - The cloud is a new business model wrapped around new technologies such as server virtualization that take advantage of economies of scale and multi-tenancy to reduce the cost of using information technology resources. The standard interfaces, coordinated between different organizations can meet the emerging needs for interoperability and portability of data between clouds. Cloud Storage has been increasing in popularity recently due to many of the same reasons as Cloud Computing.

Cloud Storage delivers virtualized storage on demand, over a network based on a request for a given quality of service (QoS). There is no need to purchase storage or in some cases even provision it before storing data. The end user only pays for the amount of storage data what he is actually consuming. Cloud storage is used in many different ways. For example: local data (such as on a laptop) can be backed up to cloud storage; a virtual disk can be synched to the cloud and distributed to other computers; and the cloud can be used as an archive to retain data for other purposes.

Data Cloud - Applications and experiments in all areas of science are becoming increasingly complex and more demanding in terms of their computational and data requirements. Some applications generate data volumes reaching hundreds of terabytes and even petabytes. As scientific applications become more data intensive, the management of data resources and dataflow between the storage and compute resources is becoming the main bottleneck. Analyzing, visualizing, and disseminating these large data sets has become a major challenge and data intensive computing is now considered as the "fourth paradigm" in scientific discovery after theoretical, experimental, and computational science.

Data storage in a cloud is a process where a user stores his data through a Cloud Service Provider (CSP) into a set of cloud servers, which are running concurrently, cooperated and in distributed manner. Data redundancy can be employed with technique of erasure-correcting code to

CONFERENCE PAPER
II International Conference on Issues & Challenges in Networking, Intelligence & Computing Technologies
Organized by Krishna Institute of Engineering and Technology (KIET) Ghaziabad, India

further tolerate faults or server crash as user's data grows in size and importance. Thereafter, for application purposes, the user interacts with the cloud servers via CSP to access or retrieve his data. In some cases, the user may need to perform block level operations on his data. The most general forms of these operations that are considered are block revise, erase, insert and affix[6].

In data cloud, users store their data in the cloud and no longer possess the data locally. Thus, the correctness and availability of the data files being stored on the data cloud must be guaranteed. One of the key issues is to effectively detect any unauthorized data modification and corruption, possibly due to server compromise and/or random Byzantine failures. Besides, in the distributed case when such inconsistencies are successfully detected, to find which server the data error lies in is also of great significance, since it can be the first step to fast recover the storage errors. To address these problems, main scheme for ensuring cloud data storage is presented. This section is devoted to a review of basic tools from coding theory that is needed for file distribution across cloud servers. Then, the homomorphic token is introduced. The token computation function we are considering belongs to a family of universal hash function, chosen to preserve the homomorphic properties, which can be perfectly integrated with the verification of erasure-coded data [6].

Subsequently, it is also necessary that how to derive a challenge response protocol for verifying the storage correctness as well as identifying misbehaving servers. Finally, the procedure for file retrieval and error recovery based on erasure-correcting code is outlined [6].

Compute Cloud - Cloud services such as Amazon's Elastic Compute Cloud and IBM's SmartCloud are quickly changing the way organizations are dealing with IT infrastructures and are providing online services. Today, if an organization needs computing power, it can simply buy it online by instantiating a virtual server image on the cloud. Servers can be quickly launched and shut down via application programming interfaces, offering the user a greater flexibility compared to traditional server rooms. A popular approach in cloud-based services is to allow users to create and share virtual images with other users. In addition to these user-shared images, the cloud providers also often provide virtual images that have been pre-configured with popular software such as open source databases and web servers.

A popular approach in cloud-based services is to allow users to create and share virtual images with other users. For example, a user who has created a legacy image may decide to make the image public so that other users can easily reuse it. In addition to user-shared images, the cloud service provider may also provide customized public images based on common needs of their customers. This allows the customers to simply instantiate and start new servers, without the hassle of installing new software themselves. Unfortunately, while the trust model between the cloud user and the cloud provider is well-defined the trust relationship between the provider of the virtual image and the cloud user is not as clear.

III. EXAMPLES

There are hundreds of cloud storage providers on the Web, and their numbers seem to increase every day. Not only are there a lot of companies competing to provide storage, but also the amount of storage each company offers to clients seems to grow regularly[5].

One may be familiar with several providers of cloud storage services, though they might not think of them in that way. Here are a few well-known companies that offer some form of cloud storage:

a. Google Docs allows users to upload documents, spreadsheets and presentations to Google's data servers. Users can edit files using a Google application. Users can also publish documents so that other people can read them or even make edits, which means Google Docs is also an example of cloud computing.
b. Web e-mail providers like Gmail, Hotmail and Yahoo! Mail store e-mail messages on their own servers. Users can access their e-mail from computers and other devices connected to the Internet.
c. Sites like Flickr and Picasa host millions of digital photographs. Their users create online photo albums by uploading pictures directly to the services' servers.
d. YouTube hosts millions of user-uploaded video files.
e. Web site hosting companies like StartLogic, Hostmonster and GoDaddy store the files and data for client Web sites.
f. Social networking sites like Facebook and MySpace allow members to post pictures and other content. All of that content is stored on the respective site's servers.
g. Services like Xdrive, MediaMax and Strongspace offer storage space for any kind of digital data.

IV. WHY CLOUD FOR DATA MINING

One of the primary concepts in cloud computing is low cost scalability systems that can grow to handle larger volumes of users and data by adding more low cost hardware. Business Intelligence for data mining is all about making better decisions from the data we have. However, the data we have is difficult to process by typical BI tools. High volume datasets can be mastered by simply throwing more hardware and software at the problem like larger servers, cluster licenses, faster networks, bigger memories, faster disks and this can be achieved by entering into the cloud.

Google's entire infrastructure is built on this approach of distributing work out to thousands of inexpensive servers, instead of relying on centralized "supercomputers" to provide the horsepower. The scalability strategy that Google uses is called MapReduce [8]. The MapReduce model provides a conceptual framework for dividing work up into small, manageable sets that can be distributed across 1 or 10 or 100 or 1000 or even 10000 servers, which can all work in parallel. This technology can be used with BI to meet the challenge of large scale, messy data, but one can't use Google's infrastructure to run his own MapReduce system. Luckily, there's Hadoop – an open source implementation of the Google MapReduce system[4].

V. ISSUES WITH CLOUD COMPUTING WHEN USED FOR DATA MINING

The two biggest concerns about cloud storage are *reliability* and *security*. Clients aren't likely to entrust their data to another company without a guarantee that they'll be

CONFERENCE PAPER
II International Conference on Issues & Challenges in Networking, Intelligence & Computing Technologies
Organized by Krishna Institute of Engineering and Technology (KIET) Ghaziabad, India

able to access their information whenever they want and no one else will be able to get at it[5].

To secure data, most systems use a combination of techniques such as *Encryption, Authentication* Processes and Authorization Practices. Even with these protective measures in place, many people worry that data saved on a remote storage system is vulnerable. There's always the possibility that a hacker will find an electronic back door and access data. Hackers could also attempt to steal the physical machines on which data are stored. A disgruntled employee could alter or destroy data using his or her authenticated user name and password. Cloud storage providers invest a lot of money in security measures in order to limit the possibility of data theft or corruption.

The other big concern, *reliability,* is just as important as security. An unstable cloud storage system is a liability. No one wants to save data to a failure-prone system, nor do they want to trust a company that isn't financially stable. While most cloud storage systems try to address this concern through redundancy techniques, there's still the possibility that an entire system could crash and leave clients with no way to access their saved data.

Cloud storage companies live and die by their reputations. It's in each company's best interests to provide the most secure and reliable service possible. If a company cannot meet these basic client expectations, it does not have much of a chance -- there are too many other options available on the market.

VI. BENEFITS WITH CLOUD COMPUTING WHEN USED FOR DATA MINING

Cloud computing offers substantial benefits [7] that we are discussing here.

A. Reduced Cost:

Cloud technology is paid incrementally, saving organizations money.

B. Increased Storage:

Organizations can store more data than on private computer systems.

C. Highly Automated:

No longer do IT personnel need to worry about keeping software up to date.

D. Flexibility:

Cloud computing offers much more flexibility than past computing methods.

E. More Mobility:

Employees can access information wherever they are, rather than having to remain at their desks.

F. Allows IT to Shift Focus:

No longer having to worry about constant server updates and other computing issues, government organizations will be free to concentrate on innovation.

VII. CONCLUSION

Cloud computing implementation for Data Mining is very beneficial to all the involved parties. But the Cloud providers need to safeguard the privacy and security of personal data that they hold on behalf of organizations and users. In particular, it is essential for the adoption of public cloud systems so that consumers and citizens are reassured that privacy and security will not be compromised. Responsible management of personal data is a central part of creating the trust that underpins adoption of cloud based services.

The advantages of cloud computing for data mining is, its ability to scale rapidly, store data remotely (in unknown places), and share services in a dynamic environment which can become disadvantages in maintaining a level of assurance, sufficient to sustain confidence in potential customers.

In this paper we have discussed the usage of cloud computing for Data Mining, considering its benefits and issues. We strongly believe that there will be very strong and positive impact of Cloud computing in Data Mining field. Cloud computing for Data Mining has the potential to become a frontrunner in promoting a secure, virtual and economically viable IT solution and future work and progress lies in standardizing Cloud computing accessibility and security protocols.

VIII. REFERENCES

[1] "Kerravala, Zeus, Yankee Group, "Migrating to the cloud is dependent infrastructure," Tech Target". Convergedinfrastructure.com. Retrieved 2011-12-02.

[2] "The NIST Definition of Cloud Computing". National Institute of Science and Technology. Retrieved 24 July 2011.

[3] Cloud Storage for Cloud Computing, ogf.org/Resources/documents/CloudStorageForCloudComputing.pdf, SNIA, September, 2009

[4] Data Mining in the Swamp: Taming Unruly Data With Cloud Computing http://www.infoq.com/articles/data-mine-cloud-hadoop Posted by John Brothers on Aug 13, 2010

[5] How Cloud Storage works - Jonathan Strickland, http://computer.howstuffworks.com/cloud-computing/cloud-storage.htm

[6] Secure Data Storage In Cloud Computing by B. Shwetha Bindu, B. Yadaiah. International Journal of Research in Computer Science eISSN 2249-8265 Volume 1 Issue 1 (2011) pp. 63-7, White Globe Publications www.ijorcs.org

[7] Privacy in Cloud -Risks and Benefits of Cloud Computing http://www.princeton.edu/~ddix/risks-benefits.html

[8] MapReduce: Simplified Data Processing on Large Clusters by Jeffrey Dean and Sanjay Ghemawat - OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004.

CONFERENCE PAPER
II International Conference on Issues & Challenges in Networking, Intelligence & Computing Technologies
Organized by Krishna Institute of Engineering and Technology (KIET) Ghaziabad, India