Integration of Big Data and Cloud Computing

Charlotte Castelino[1], Dhaval Gandhi[1], Harish G. Narula[2], Nirav H. Chokshi[1]

[1]*Undergraduate Student, Electronics And Telecommunication Engineering Department, Dwarkadas J. Sanghvi College of Engineering, Mumbai, India.*

[2]*Senior Lecturer, Computer Engineering Department, Dwarkadas J. Sanghvi College of Engineering, Mumbai, India*

***Abstract* - Big data and cloud computing are both emerging technologies whose rate of adoption by businesses has been increasing rapidly over the past decade. This paper stresses on the integration of big data with cloud computing, which can serve as a driving force for the business and IT industry, as well as, for data analytics in general. We discuss the methodologies, challenges faced, possible solutions, and the benefits of integrating the two.**

***Keywords* - Big data, cloud environments, data analytics, cloud services, cloud computing, cloud security**

I. INTRODUCTION

A. What is big data?

Big data is a new phenomenon which has been introduced due to the complex and vast data that we interact with, today. It may include structured data in traditional databases as well as semi-structured or unstructured data, e-mails, audio, video, etc all combined together. Apart from the vast size of data, it also includes heterogeneous data present in various formats.

This data arrives at high speeds from multiple sources such as social media, transactions, interactions with other web pages, etc in a random fashion. Managing, harvesting and analyzing the information obtained from such data is thus posing a challenge to organizations.

B. Why has big data processing become essential to implement?

The volume, variety and velocity associated with big data causes performance problems when being analyzed using the conventional data processing techniques. Hence, making use of conventional relational database management systems is not feasible. However, an efficient analysis of this data by using suitable processing techniques is essential as it can lead to a wealth of information and when mined, can have potential benefits such as cost cutting, increase in revenues, etc.

C. What is cloud computing?

Cloud computing refers to the use of remote servers for the storage and processing of data. These services are offered to a company/user by some third party as a resource. Generally, these resources are shared by multiple users and may also be dynamically allocated among users as per the demand. This helps to increase the usability of resources.

II. WHY INTEGRATE BIG DATA AND CLOUD COMPUTING

With cloud computing, multiple users can access a single server to retrieve and update their data without purchasing licenses for different applications. [4] Also, the user need not worry about technical issues as these are handled by the third party and can focus on other tasks.

There are three types of clouds: private, public and hybrid. After considering various factors such as cost, security, workload, etc, an organization can decide on the infrastructure to be deployed and accordingly, avail the necessary services using an appropriate cloud environment. It is also possible to implement hybrid models which combine certain selective features of private clouds, such as privacy with those of public clouds, such as scalability or inter-operatibility.

Big data, due to its vast size, variability and high velocity requires the use of multiple servers that work in a parallel manner. Since cloud environments already make use of multiple servers and allocate resources on demand, it would be highly beneficial for organizations to make use of cloud services for big data analysis. The facility of parallel computing offered by cloud services can augment the efficiency with which big data is processed. It also has the potential to replace most batch processing systems with real time processing systems.

The intersection between cloud and big data is still relatively untapped. Yet, utilizing a cloud system to store big data has long term benefits to both, the insights yielded, as well as, the performance of the IT sector. Data without analysis is worthless. Big data requires advanced analytic techniques to deal with the extensive amounts of data. Cloud systems are typically based on remote servers, which are able to handle extensive amounts of data with rapid response time for real time processes.

A. Infrastructure-as-a-Service (IaaS)

Utilizing IaaS, cloud systems further reduces data centre rental costs. Thus, Cloud computing infrastructure

enables more efficient use of hardware and software investments. Pooling these resources forces costs down and improves utilization. Problems of analysis and storage can be solved through a hosted cloud system which provides secure, scalable solutions for managing data.

Elasticity, pay-per-use, low upfront investment, and low security risks are some of the major features that make cloud computing an ideal platform for big data analysis, which would not have been economically viable on traditional infrastructure.

III. INTEGRATED CLOUD ENVIRONMENTS FOR BIG DATA

Big Data solutions will be a hybrid of traditional databases, data appliances and the enterprise system. The combination of high speed SQL access and the enterprise systems works very well. Data only needs to be shared between the data sources which are employed.

ConPaaS [1] is an integrated cloud environment for big data, which has been developed by Vrije Universiteit Amsterdam, the University of Rennes 1, Zuse-Institut Berlin and XLAB. ConPaaS makes it easy to write scalable cloud applications without worrying about the complexity of the cloud. It provides a runtime environment that facilitates deployment of end-user applications in the cloud. ConPaaS currently provides services for web hosting (PHP and Java), SQL and NoSQL databases (MySQL and Scalaris), data storage (XtreemFS). It contains two services that are specifically dedicated to Big Data: MapReduce and TaskFarming.

MapReduce provides users with parallel programming facilities, whereas TaskFarming allows the automatic execution of a large collection of independent tasks for the user.

For reaching the goals of big data management, most of the research institutions and enterprises bring virtualization into cloud architectures [5]. Therefore, the basic elements such as storage and network bandwidth are taken care of, by specialized service providers. Some of the popular cloud management platforms such as Amazon web services (Amazon EC2 and Amazon S3), Cloudstack, Openstack and others are looking to harness the benefits of big data analytics in business.

IV. CLOUD SELECTION FOR BIG DATA

Many organizations are apprehensive to place sensitive data on a public cloud. Such organizations prefer investing in their own infrastructure, even though it may be more expensive as compared to using resources on the cloud. Some organizations, however, that make use of big data, only for the purpose of decision support or other secondary purposes, may readily adopt cloud services offered by third parties.

Industry observers are optimistic about big data deployments on the private cloud, though.

Sometimes there may be a tradeoff involved with privacy when an organization wishes to take advantage of the flexibility and scalability that the cloud offers. Ethical considerations are equally important while placing big data in the cloud. With the permission of the end user, big data analytics that are more invasive are being increasingly employed.

V. POTENTIAL BENEFITS

Enterprises are beginning to realize the benefits of using the cloud to harness the business intelligence (BI) present in big data. Some of the potential benefits of using cloud services while employing big data analytics are as follows:

A. Parallel computing

One of the major benefits of using cloud services for big data analytics is the facility of parallel computing that they offer. Big data cannot generally be stored or analyzed on a single machine. Deploying cloud services with multiple machines, data can be processed faster and with much more efficiency.

B. Scalability

Another benefit of placing big data on the cloud is the scalability, which is also one of the most touted benefits of cloud computing. For businesses, it provides the ability to add or remove users or resources as required. This facility is especially useful when contracted or seasonal resources are being used. Also, the needs of any business are constantly fluctuating as per the market. As such, scalability is an excellent feature.

C. Elasticity

Big data analysis requires a large number of machines working together at a particular time. It is also entirely possible that most of these machines would be idle for some period of time. Since cloud platforms allocate resources on demand, there is no wastage of resources which helps cost saving and optimization.

D. Inexpensive

Deploying multiple machines for the processing of big data means an increased cost. Making use of scalable cloud resources for the same, however, will incur only a fixed cost which will not depend on the size of data being processed, as cloud platforms are capable of supporting parallel computing.

VI. SECURITY AND PRIVACY CHALLENGES

Mining of big data can provide significant information. However, this may be done at the cost of user privacy. In the case of cloud computing, since resources are distributed among various clients by a service provider, there may be serious lapses in preserving the privacy of users. It is possible that an unauthorized user may gain access to sensitive information or the data administrator of the cloud network could also manipulate or misuse sensitive data belonging to clients.

In order to protect the data being stored in the cloud from possible misuse, data must be encrypted using complex encryption techniques. These encryption techniques, however, require a higher bandwidth and increase processing time. This, in turn, increases the cost for the end user.

A. Potential solution

Some alternative techniques which have been deployed by cloud service providers include:

i) Distributed file systems: Individual files that are part of a larger database are distributed among various machines present at different locations. Hence, access to any individual part of the database does not pose a significant threat, if the relation with others is not known.

ii) Data obfuscation/ masking: In this technique, data is obfuscated by making use of a special key known only to the user. Thus, even the service provider would not be able to de-obfuscate the data.

In 2012, the Cloud Security Alliance (CSA) teamed up with Fujitsu Laboratories of America for the launch of the Big Data Working Group, specifically to address the privacy and security concerns associated with big data, especially data which is placed on the cloud. The Big Data Working Group cited six areas it will focus on: big data-scale cryptography, cloud infrastructure, data analytics for security, framework and taxonomy, policy and governance, and privacy [7].

VII. PERFORMANCE CHALLENGES

A. Data transfer limitations

Organizations generally retain most of their data on-premise. Transferring data onto the cloud servers, thus introduces latency within the system. Router hops and network inter connections may further increase the latency

B. Data retention

In the case of cloud computing, data belonging to a particular user may be stored on multiple servers. Therefore, data retention is governed by multiple laws. Also, determining what data is to be retained for which user is a challenge.

C. Isolation management

Cloud computing makes use of application sharing and multi-tenancy. Although advanced techniques are used for isolating data and resources belonging to each user, there is still a good chance that data belonging to one user will be commingled with that of another.

D. Disaster recovery

Since data is generally scattered among multiple servers, it may be difficult to pin-point the exact location of data at a particular time, thus making rapid data retrieval difficult in the face of a disaster.

VIII. FUTURE SCOPE

An increasing number of organizations are already making use of big data analytics as well as cloud platforms for storage. Companies analyzing big data have determined that they are able to reach the right audience more frequently, and recognize a greater return on their investment. At the same time, making use of cloud services has enabled cost cutting.

However, it may be a while before big data can be reliably placed on the cloud and simultaneously analyzed. Basic platform capabilities, such as security, access control, virtualization, availability, etc. will have to be standard before organizations would adopt the cloud completely.

REFERENCES

[1] http://www.conpaas.eu/

[2]http://www.intel.com/content/www/us/en/big-data/big-data-cloud-technologies-brief.html

[3]http://www.infosys.com/cloud-services/resources/Pages/big-data-spectrum.aspx

[4] http://en.wikipedia.org/wiki/Cloud_computing

[5] Big Data Processing in Cloud Computing Environments- Changqing Ji, Yu Li,Wenming Qiu, Uchechukwu Awada, Keqiu Li

[6]http://searchcloudcomputing.techtarget.com/feature/Big-data-analytics-CloudStack-top-hot-cloud-computing-trends-in-2013

[7]http://searchcloudsecurity.techtarget.com/news/2240162533/Cloud-Security-Alliance-tackles-big-data-security

[8] https://gigaom.com/2009/10/25/the-future-is-big-data-in-the-cloud/