International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A Review on Big Data Cloud Computing

Neenu Daniel
CSE Department, VJCET,Ernakulam

Abstract-Big Data Cloud Computing has become one of the industry buzz words and a major discussion thread in the IT world. Big data is the term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications. In the IT industry as a whole, the rapid rise of Big Data has generated new issues and challenges with respect to data management and analysis. As for cloud Computing, it has become a significant research topic of the scientific and industrial communities . The foundation of cloud computing is the delivery of services, software and processing capacity over the Internet, reducing cost, increasing storage, automating systems, decoupling of service delivery from underlying technology, and providing flexibility and mobility of information. Cloud Computing provides strong storage, computation and distributed capability in support of Big Data processing. This paper presents a review on the background, challenges and benefits about managing big data on the cloud.

Keywords— Cloud Computing , BigData,Security, Privacy

I. INTRODUCTION

Big Data is a data analysis methodology enabled by a new generation of technologies and architecture which support high-velocity data capture, storage, and analysis (Villars, Olofson, & Eastwood, 2011). Big Data requires huge amounts of storage space. Data storage using cloud computing is a viable option for small to medium sized businesses considering the use of Big Data analytic techniques. During the last couple of years, improvements in the area of network based computing and also applications on requirement have generated an explosive development of application models for instance cloud computing, software as a service, community network, web store, etc. While a significant application model in the period of the Internet, Cloud Computing is now an important research field of the scientific together with industrial groups since 2007. Generally, cloud computing is called numerous services that are offered by an Internet-based cluster system. This kind of cluster systems include an assortment of economical servers or even Personal Computers, coordinating the several sources of the computers based on a particular administration method, and rewarding secure, dependable, fast, convenient and unambiguous services just like information storage, accessing and computing to clients . Cloud computing provides an apt platform for big data analytics in view of the storage and computing requirements of the latter. Big data needs clusters of servers for processing, which clouds can readily provide.

II. BIG DATA

A. Big Data Definition And Characteristics

Big data is a term that describes the large volume of data – both structured and unstructured – that overwhelm a business on a day-to-day basis[1]. But it's not the amount of data that's important. It's what organizations do with the data that matters. This concept gained momentum in the early 2000s when industry analyst Doug Laney articulated the now-mainstream definition of big data as the three Vs:**Volume.** Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. In the past, storing it would've been a problem – but new technologies (such as Hadoop) have eased the burden. **Velocity.** Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. **Variety.** Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.

B. Big Data Technologies

Now an increasing number of technologies and tools are available for processing Big Data on the cloud.

1) *Column-Oriented Databases*: With nowadays continuously growing data volumes, conventional DBMS are struggling with SQL queries over such large data sets. For a table with several tens of thousands of records you have to create indexes to get acceptable performance. Column adding or deletion is almost impossible and requires using special techniques for large tables. Column-oriented databases are here to address two issues — performance of complex queries over large data sets and changing table structure in production. Column-oriented databases store data with a focus on columns, instead of rows, allowing for huge

data compression and very fast query times.

2) *No SQL Databases:* There are several database types that fit into this category, such as key-value stores and document stores, which focus on the storage and retrieval of large volumes of unstructured, semi-structured, or even structured data. They achieve performance gains by doing away with some (or all) of the restrictions traditionally associated with conventional databases, such as read-write consistency, in exchange for scalability and distributed processing.

3) *Map Reduce:* This is a programming paradigm that allows for massive job execution scalability against thousands of servers or clusters of servers. Any MapReduce implementation consists of two tasks:The "Map" task, where an input dataset is converted into a different set of key/value pairs, or tuples; The "Reduce" task, where several of the outputs of the "Map" task are combined to form a reduced set of tuples (hence the name).

4) *Hadoop:* Hadoop is by far the most popular implementation of MapReduce, being an entirely open source platform for handling Big Data. It is flexible enough to be able to work with multiple data sources, either aggregating multiple sources of data in order to do large scale processing, or even reading data from a database in order to run processor-intensive machine learning jobs. It has several different applications, but one of the top use cases is for large volumes of constantly changing data, such as location based data from weather or traffic sensors, web-based or social media data, or machine-to-machine transactional data.

5) *HDFS(Hadoopdistributedfilesystem):* HDFS is a java based file system that is used to store structured or unstructured data over large clusters of distributed servers. The data stored in hdfs has no restriction or rule to be applied, the data can be either fully unstructured of purely structured. In hdfs the work to make data senseful is done by developer's code only.

III.CLOUD COMPUTING

A. *Importance Of Cloud Computing*

The cloud refers to the datacenter hardware and software that supports a clients needs, often in the form of data stores and remotely hosted applications [3]. These infrastructures enable companies to cut costs by eliminating the need for physical hardware, allowing companies to outsource data and computations on demand. Developers with innovative ideas for Internet services no longer need large capital outlays in hardware to deploy their services; this paradigm shift is transforming the IT industry. The operation of large scale, commodity computer datacenters was the key enabler of cloud computing, as these datacenters take advantage of economies of scale, allowing for decreases in the cost of electricity, bandwidth, operations, and hardware. Cloud Computing uses networks of a large group of servers with specialized connections to distribute data processing among the servers.

The main characteristics that cloud computing offers today are:

1) *On-Demand Self Service*: Provision computing capabilities, computer service such as network, email, application. It also provision server service without requiring human interaction with each service provider.

2) *Broad Network Access:* Cloud capabilities are available over the network. It can access business management solution using their tablets, smart phone, laptops, and office computers. Network access includes private cloud that operates within a company's firewall, public clouds or a hybrid cloud.

3) *Resource Pooling*: The provider's computing resource are pooled to serve multiple consumers using a multi tenant model, with different physical and virtual resources dynamically assigned and reassigned according to the consumer demand. The cloud enables to enter and use data within the business management software hosted in the cloud at a same time, at any time and from any location.

4) *Rapid Elasticity*: The cloud is flexible and scalable to submit needs. The capacity and can be shrinked very quickly. The self service and resource pooling make rapid elasticity possible. The service provider can automatically allocate more or less resources from the available pool.

5) *Measured Service*: cloud computing resource usage can be measured, controlled, and reported providing transparency. Cloud systems automatically control and optimize resource use.

B. *Cloud Computing Service Models*

Cloud computing service models are classified as:

1) *Software As A Service (SAAS):* It is the delivery of application. In SaaS a complete application is provided to user which is running on cloud infrastructure. As software is hosted by provider, users do not need to buy, install or manage hardware for it. In

SaaS instances of a software application are shared as a service. Examples of SaaS are Google Docs, Cloud Drive, and Salesforce.com CRM application.

2) Platform As A Service (PAAS): PaaS enables developers to deploy their application on the cloud. The consumer can control their application but do not have any control over underlying infrastructure. It provides user an integrated set of software through the internet. PaaS is a delivery of computing platform as a service. Examples of PaaS are Google App Engine, Amazon Web Services, and Microsoft Azure.

3) Infrastructure As A Service (IAAS): Using IaaS user get access to resources like storage, server, networks, data center space. It shares pool of computing resources. User can deploy and run both application and operating system on IaaS. It frees user from buying or managing underlying software and hardware. Example of IaaS is Amazon EC2.

IV. BIG DATA AND THE CLOUD

Big Data and Cloud, two of the trends that are defining the emerging Enterprise Computing, show a lot of potential for a new era of combined applications. The provision of Big Data analytical capabilities using cloud delivery models could ease adoption for many companies, and in addition to important cost savings, it could simplify useful insights that could provide them with different kinds of competitive advantage. . 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[5].Big data provides users the ability to use commodity computing to process distributed queries across multiple datasets and return resultant sets in a timely manner. Big data utilizes distributed storage technology based on cloud computing rather than local storage attached to a computer or electronic device. Big data evaluation is driven by fast-growing cloud-based applications developed using virtualized technologies. Therefore, cloud computing not only provides facilities for the computation and processing of big data but also serves as a service model. Cloud computing provides an environment for small to medium sized businesses to implement big data technology. Benefits that businesses can realize from big data include performance improvement, decision making support, and innovation in business models, products, and service[8]. Three major reasons for small to medium sized businesses to use cloud computing for big data technology implementation are the ability to reduce hardware costs, reduce processing costs, and to test the value of big data before committing significant company resources

V. SECURITY AND PRIVACY CHALLENGES FOR BIG DATA CLOUD COMPUTING

Although our data capacity is growing exponentially, we have imperfect solutions for the many security issues. The challenges associated with big data privacy issues which can be divided into four groups[4]:

A. Infrastructure Security

Secure computations in distributed programming frameworks as well as in non relational data stores. Distributed programming frameworks process big data with parallel computation and storage techniques. In such frameworks, unauthenticated or modified mappers which divide huge tasks into smaller sub-tasks so that the tasks can be aggregated to create a final output — can compromise data. Faulty or modified worker nodes which take inputs from the mapper to execute the tasks — can compromise data by tapping data communication between the mapper and other worker nodes. Most cloud-based data frameworks use the NoSQL database. The NoSQL database is beneficial for handling huge, unstructured data sets but from a security perspective, it is poorly designed. NoSQL was originally designed with almost no security considerations in mind. One of the biggest weaknesses of NoSQL is transactional integrity. It has poor authentication mechanisms, which makes it vulnerable to man-in-the-middle or replay attacks. To make things worse, NoSQL does not support third-party module integration to strengthen authentication mechanisms. Since authentication mechanisms are rather no strict, data is also exposed to insider attacks. Attacks could go unnoticed and untracked because of poor logging and log analysis mechanisms.

B. Data Privacy

Secure the data itself using a privacy-preserving approach for data mining and analytics. Also, protect sensitive data through the use of cryptographically enforced data centric security and granular access control. The amount of information collected on each individual can be processed to provide a surprisingly complete picture. As a result, organizations that own data are legally responsible for the security and the usage policies they apply to their data. Attempts to anonymous specific data are not successful in protecting privacy because there is so much available that some data can be used as a correlation for identification purposes. Users' data are also constantly in transit, being accessed by inside users and outside contractors, government agencies, and business

partners sharing data for research.

C. Data Management

Manage the enormous volume of data using scalable, distributed solutions to secure data stores and enable efficient audits and data provenance. There are specific vulnerabilities associated with big data storage: Confidentiality and integrity, data provenance, and consistency.

D. Integrity/ Reactive Security

Use endpoint validation and filtering to check the integrity of streaming data, and real-time security monitoring and analytics to help prevent and address security problems. Because of the breadth of data sources, including endpoint collection devices, a major challenge facing big data schemes is whether the data is valid from the point of input. Given the size of the data pool, how can we validate the sources? How can we be sure that a source of input data is not malicious, or simply incorrect? In addition, how can we filter out malicious or unreliable data? Both data collection devices and programs are susceptible to attack. An infiltrator may spoof multiple IDs and feed fake data to the collection system.

VI. BENEFITS OF BIGDATA ON THE CLOUD

A. Cost Reduction

Cloud computing offers a cost-effective way to support big data technologies and the advanced analytics applications that can drive business value. Enterprises are looking to unlock data's hidden potential and deliver competitive advantage. Big data environments require clusters of servers to support the tools that process the large volumes, high velocity, and varied formats of big data. IT organizations should look to cloud computing as the structure to save costs with the cloud's pay-per-use model.

B. Reduce Overhead

Various components and integration are required for any big data solution implementation. With cloud computing, these components can be automated, reducing complexity and improving the IT team's productivity.

C. Rapid Provisioning/Time To Market

Provisioning servers in the cloud is as easy as buying something on the Internet. Big data environments can be scaled up or down easily based on the processing requirements. Faster provisioning is important for big data applications because the value of data reduces quickly as time goes by.

D. Flexibility/Scalability

Big data analysis, especially in the life sciences industry, requires huge compute power for a brief amount of time. For this type of analysis, servers need to be provisioned in minutes. This kind of scalability and flexibility can be achieved in the cloud, replacing huge investments on super computers with simply paying for the computing on an hourly basis.

VII. CONCLUSIONS

Big Data provided through Cloud computing is an absolutely necessary characteristic for today's businesses to make proactive, knowledge driven decisions, as it helps them have future trends and behaviours predicted. This paper provided an overview of the necessity and utility of Big Data Cloud computing. With emergence of big data systems, the ability of integrating them in cloud computing becomes more and more necessary. In this paper basics of Big Data on cloud Computing, issues and its challenges are discussed.

REFERENCES

[1] Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao Athanasios V. Vasilakos, Big data analytics: a survey,Journal of BigData,oct2015 .
[2] M. D. Assuncao, R. N. Calheiros, S. Bianchi, M.A. Netto, R. Buyya, "Big Data Computing and Clouds: Challenges,Solutions, and Future Directions," arXiv preprint arXiv:1312.4722, .2013.
[3] Amit goyal and sara dadizadeh " Asurvey on Cloud Computing University of BritishColumbia, Technical Report for CS 508, December 2009
[4] http://www.business.com/technology/privacy-and-security-issues-in-the-age-of-big-data.
[5] L. Zhang, C. Wu, Z. Li, C. Guo, M. Chen, and F. C. M. Lau, "Moving Big Data to The Cloud: An Online Cost-Minimizing Approach," IJCA, Vol.31- No. 12,

Dec 2013.

[6] Venkata Narasimha Inukollu, Sailaja Arsi and Srinivasa Rao Ravuri ―Security Issues Associated With Big Data In Cloud Computing‖ International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014

[7] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li ―Big Data Processing in Cloud Computing Environments‖ 2012 International Symposium on Pervasive Systems, Algorithms and Networks.

[8] Manyika, J., Chui,. M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011, June). Big data: The next frontier for innovation, competition,andproductivity.McKinseyGlobalInstitute.Retrievedromhttp://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_ innovation.

[9] Dr.M. Moorthy, R. Baby, S. Senthamaraiselvi,"An Analysis for Big Data and its Technologies"| International Journal of Computer Science Engineering and Technology(IJCSET) " Dec 2014 , Vol 4, Issue 12,412-418

[10] H.E.miller, H. Miller, "Big-data in cloud computing: a taxonomy of risks",Inf. Res., 18 (2013), p. 571

[11] Villars, R. L., Olofson, C. W., & Eastwood, M. (2011, June). Big data: What it is and why you should care. IDC White Paper. Framingham, MA: IDC