Overview of the Cloud Computing for Big Data

KPN Jayasena[1], Huazhu Song[2]

[1] *School of Computer science and Technology, Wuhan University of Technology*
Wuhan, Hubei, P.R.China
[2]*Department of Computing and information System, Sabaragamuwa University of Sri Lanka*
[a]pubudu.nuwanthika@gmail.com, [b]juliprn@sina.com

Abstract - *In the last few years, there has been a growing interest in Big Data technologies with cloud Infrastructure. Hadoop software tool is an open source platform that allows for the distributed processing of huge data sets across different clusters, which handles high volumes of the structured data, semi-structured data and unstructured data from various sources. Hadoop environment is scalable, and Hadoop adds new nodes without changing data formats or the application. The main objective of this research lies in summarizing the cloud computing and big data technologies, providing details of the most common infrastructures that have been developed, discussing several big data processing technologies and the presenting key problems of big data processing and the cloud computing platform. Finally the open issues and challenges are introduced and research directions in the future on big data processing are explored in cloud computing environments. Therefore cloud computing can be considered as an attractive technology platform for developing and deploying big data, and it has a good future.*

Keywords: Big data, Cloud computing, Hadoop, MapReduce, cloud services

I. INTRODUCTION

Cloud computing technology is becoming a powerful architecture to accomplish large scale and complex computing, and has modernized the way that computing infrastructure is intergraded and used (Judith Hurwitz, Alan Nugent,et.al., 2013.)

[2]Big data and cloud computing are both the fastest moving technologies identified in Gartner Inc.'s 2013 Hype Cycle for an emerging Technologies. Cloud computing is associated with new paradigm for the provision of computing infrastructure and big data processing method for all kinds of resources. Moreover, some new cloud-based technologies have to be adopted because dealing with big data for concurrent processing is difficult.

There are huge amount of unutilized resources present in the computing environment, which are required to be used in an effective and systematic manner (Big Data Explained 2014). These resources could be used to solve complex problems, intense scientific calculations, mathematical solutions and storage of data in the cloud. The goal of this paper is to provide the status of big data studies and related works, which aims at providing a general overview of cloud computing and big data.

This paper introduces several big data processing technologies from system and application aspects. First, it adds problems of big data processing in cloud computing infrastructure, explore the current issues and challenges, and deeply discuss the research directions in the future of big data processing in cloud computing environments.

The rest of the paper is organized as follows. In the section II, cloud computing concept is described. In the section III, the big data technology is illustrated. Cloud computing and big data concept explore in the section IV. Some applications and future direction in big data and cloud technology are discussed in the section V and the section VI. Finally the section VII gives the discussion of the paper.

II. CLOUD COMPUTING

Cloud computing offers a scalable and cost efficient solution to the big data challenge; largely defined and widely abused to represent anything that is 'online'. The National Institute for Standards and Technology (NIST)[3] defines Cloud computing as "a pay-per-use model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks,

[2] http://www.gartner.com/newsroom/id/2575515

[3] http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction".

Figure 1: Gartner Hype Cycle for Cloud computing identifies which fields of cloud computing are in the hype stage, applications technologies. Technologies at the overhyped stage comprise "big data", consumer 3D printing, gamification, and wearable user interfaces. Technologies that are climbing the slope or already becoming productive include gesture controls, biometric authentication systems, speech recognition and predictive analytics.[4]

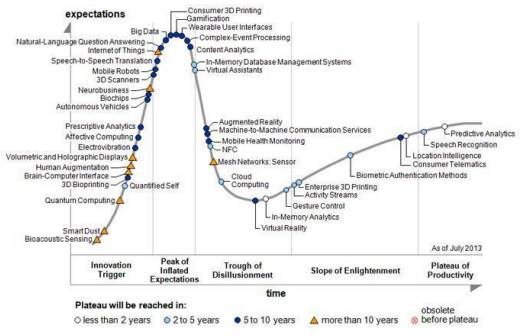


Figure 1: Hyper Cycle for Emerging technologies

Figure 2 illustrates the future of cloud computing, which shows future cloud computing mainly focus on big data and distributed sharing.

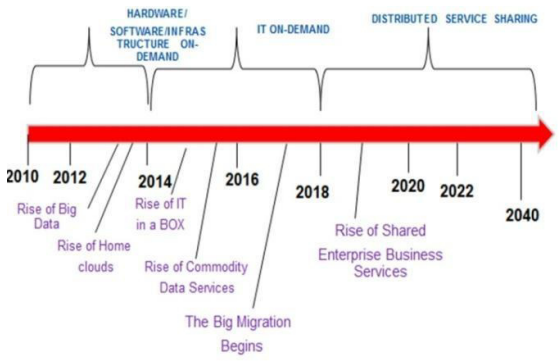


Figure 2: Future cloud computing trends

A. Overview on Cloud services

There is growing number of Cloud technologies, in particular software solutions for the management of cloud systems in different variations. This includes No-SQL databases (e.g. Couch DB), virtualization software (eg.. VM Ware) , distributed caching (e.g. Oracle Coherence), infrastructure management (e.g. Open Stack) and integration solutions (e.g. Cloudswitch). It is the trend to use Open Source solutions like Open Stack or Open Nebula, which are supported by main suppliers like IBM or Google. There is also a growing market for essential hardware, which is most likely a sub segment of the data center hardware market lead by companies like IBM, Dell, HP, Huawei, Cisco and others. As some of them are also suppliers of the Cloud Computing technology and services, they are able to offer fully integrated services to their customers. Another trend supporting this development is the growing number of solutions for modular data center server platform combining server hardware, switches, management, and virtualization software in a bundle.

Infrastructure as a Service (IaaS) is computing infrastructure (e.g. Amazon EC2 (Amazon EC2 2014)), storage infrastructure (e.g. Rackspace Cloud Files) backup infrastructure (e.g IBM Smart Cloud Managed Backup (Cloud-managed backup 2013)) or brokerage infrastructure (e.g Gravitant). Additionally also load balancing, content delivery infrastructure (e.g Amazon CloudFront) or management infrastructure (e.g. Amazon Cloud Watch) are offered as IaaS services.

Platform as a Service (PaaS) offers can be differentiated into four types: general purpose platforms (e.g. Microsoft Azure Platform); development platforms (e.g. IBM Rational Software Services); database platforms (e.g. Amazon Dynamo DB); and integration platforms (e.g. Informatica Cloud). One recent trend in this segment are Business Intelligence Platforms that provide collections of tools for analyzing different types of data from normal business data to big data collections.

Software as a Service (SaaS) offers a broad variety of services similar to the normal application landscape. Typical examples are customer relation management (CRM) (e.g. Salesforce CRM), enterprise resource planning (ERP) (e.g. SAP by Design), business intelligence (e.g. Datameer), collaboration tools (e.g. Jive Social Business Software, supply chain management (SCM) (e.g. Aravo) or human resources management (e.g. Workday)

Cloud technology provides organizations the

[4] Gartner, Hype Cycle for Cloud Computing,. Disclaimer: The Hype Cycle is copyrighted 2013 by Gartner, Inc

capability to store, share and analyze data more efficiently and also help companies grow. Cloud computing developments indicate that this technology is still changing and adapting to be even more beneficial for those who use it for data visualization and advanced analytics.

Cloud technology help different companies to collect and store big data in a secure location, but many of them are still risks associated with moving to the cloud. Computerworld reported that businesses need to take safety protections when exchanging data to a cloud.

Different cloud providers offer various security measures. For instance, the hybrid cloud, which combines both public and private clouds, provides new opportunities for scalable security, according to CIOL (Chartered Institute of Linguists). This allows for tiered security with features like access control policies that will keep information secure.

CIOL illustrates private cloud helps companies determine whether to run processes internally or externally and provide more control over the cloud. This technology provides real-time analytics and scalability that will adjust with technological advances and company growth.

III. BIG DATA

The amount of data generated annually over the internet has exceeded the Zetabytes level. For an example ,You Tube users upload 48 hours of new video every minute of the day, 100 terabytes of data uploaded daily to facebook and Walmart handles more than 1 million customer transactions every hour and databases more than 2.5 petabytes of data. IDC Digital Universe study, sponsored by EMC predicts that between 2009 and 2020, digital data will grow 44 folds to 35 Zetabytes per year. Processing data with such high volume far exceeds the computational capability of today's datacenters and computers given the rise to the term 'Big Data'. Big Data is a technologies and techniques for working productively with data at any scale.

C., Paul, D. deRoos, (2013) indicates that Big data is the recognition of greater business intelligence by storing, processing, and analyzing data that was previously ignored due to the limitations of traditional data management technologies

The convergence of dimensions of big data helps both to define and distinguish big data. Volume is the amount of data and refers to the huge quantities of data that several organizations are trying to attach to improve decision-making through the enterprises. Velocity means quickly moving data, milliseconds to seconds to respond. The faster can collect and process data, the more opportunity have to control the information for competitive advantage. Variety means structured, semi-structured, unstructured, images, etc. Other data types include geo-spatial and location, log data, machine data, metrics, mobile, RFIDs, search, streaming data, social, text and so on. Veracity means Big data is often not verified, verifiable or validated, whose analysis, can't always be duplicated simply as data keeps changing and duplication, omission, and general incompleteness expected. Big data is a combination of these characteristics and creates an opportunity for industry to gain viable advantage in the digitized marketplace.

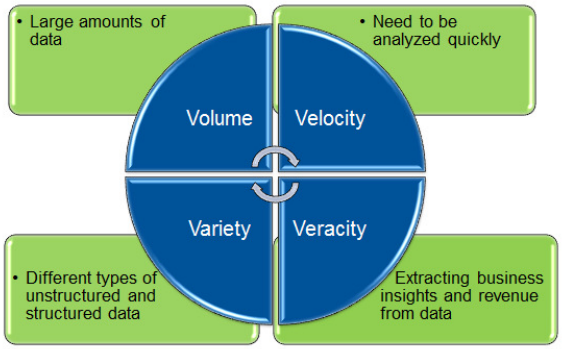


Figure 3: Four dimensions of big data

Table 1: The key drivers in the growth of big data

Data Availability	Ability to process	Cloud consumption model
Applications at the heart of business interactions	New programming models	Easy on-ramp cost effective experimentation
Devices and sensors	New scale and capabilities for SQL	Unlimited scale, low TCO
Lower cost of storage and ingestion	Easily available software(Open source)	Combine Open Source Software and platform services

A. Challenges of Big data

According to Lavastorm, (2013) and sas.com, number of challenges associated with big data exist.

- Skills shortage-There is currently a lack of proficiency within the industry to deliver big data projects.
- **Many organisations do not have the capability to derive quality information and insight from big data sources. Companies need to have data governance or information management process in place to ensure the data is clean.**
- Big data does not live in isolation of existing BI architecture and capability.
- **The current cost of building big data solutions is high (people and training costs are the primary costs) and the technology is still evolving.**
- Desktop and server configurations provide a cost-effective solution for low-volume and high-volume applications.[5]

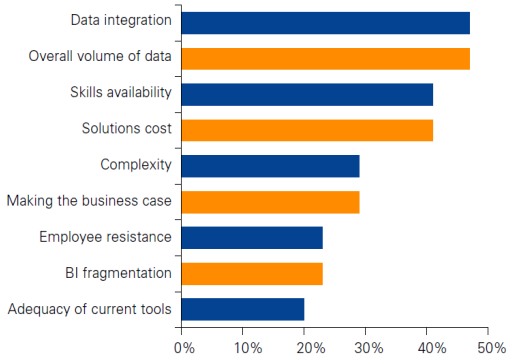


Figure 4: Biggest challenges for success in Big data and analytics source TM forum 2014

B. Big data and Hadoop

Big Data and the Apache Hadoop open source project are rapidly emerging as the ideal solution to address business and technology trends that are disrupting traditional data management and processing. Hadoop open source platform is a scalable fault tolerant distributed system for big data storage and processing. Hadoop has two main components those are Hadoop Distributed Fie Systems (HDFS) (Reliable redundant distributed file system optimized for large files) and MapReduce fault tolerant distributed processing (Mapping inputs to outputs and reducing the output of multiple mappers). Hadoop can operate on unstructured and structured data, which is an open source under the friendly Apache. The main advantage of the framework of Hadoop and MapReduce is that non-expert users can easily execute analytical tasks over big data. Users can control on how input datasets are managed. Users code their queries using Java rather than SQL. So, it is easy to handle for a larger number of developers.

Facebook uses Hadoop to analyse user behaviour and the effectiveness of ads on the site. The tech team at 'The New York Times' rented computing power on Amazon's cloud and used Hadoop to convert 11 million archived articles, dating back to 1851, to digital and searchable documents.

1) Hadoop Distributed File System(HDFS):

HDFS consists of a single master node multiple nodes. Files are split up blocks (typically 64MB), blocks are spread across data nodes in the cluster, each block is replicated multiple times to different data nodes in the cluster and master node keeps track of which blocks belong to a file.

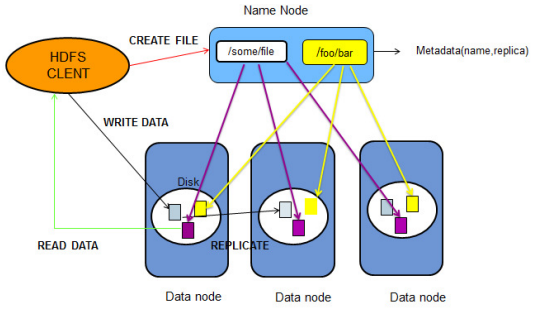


Figure 5: HDFS Architecture

Main Properties of the HDFS

- Large: HDFS instance consist of thousands of server machines, each storing part of the file system's data
- Replication: Each data block is replicated many times (default is 3)
- Failure: Failure is the norm rather than exception
- Fault Tolerance: Detection of faults and quick, automatic recovery from them is a core architectural goal of HDFS
- Namenode is consistently checking Datanodes

2) MapReduce :

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a map

[5] http://www.tmforum.org/Big-Data-InFocus-2014/15491/home.html

function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key.

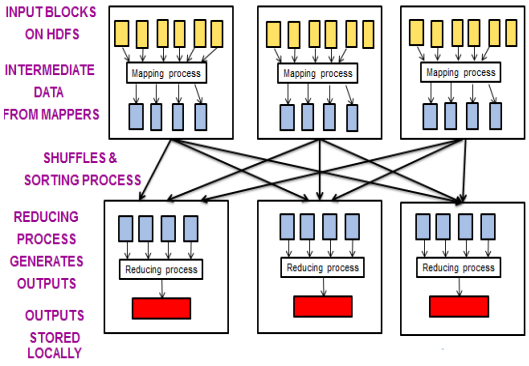


Figure 6: MapReduce Process

IV. BIG DATA TECHNOLOGY AND CLOUD COMPUTING

The recent survey conducted by GigaSpaces summarized that 80% of IT executives think big data processing is important and moving big data analytics to cloud delivery models.

Data is becoming more valuable than past decade. Today the conversation is shifting from "What data should we store?" to "What can we do with the data?. Gartner predicts that enterprise data will grow by 800 percent from 2011 to 2015, with 80 percent unstructured (for example, e-mails, documents, video, images, and social media content) and 20 percent structured (for example, credit card transactions and contact information). Companies must have to find new directions to processing, managing, and analyzing their data whether it is structured data originates in traditional relational database management systems (RDBMSs) or more diverse unstructured formats.

Data analytics is moving from batch processing to real time. Intel's 2012 survey of 200 IT managers in large enterprises found that while the amount of batch versus real-time processing is split evenly today, the trend is toward increasing real time to two-thirds of total data management by 2015. At the same time, the technology for processing real time or near-real time information is moving past hype to early stages of maturity.

IT organizations are looking to cloud computing as the best structure to support their big data projects because cloud infrastructure to analyse big data is efficient and cost effective. Big data may mix internal and external sources. While enterprises often keep their most sensitive data in-house, huge volumes of big data (owned by the organization or generated by third-party and public providers) may be located externally and moving relevant data sources behind firewall can be a significant commitment of resources. Data services are needed to extract value from big data. Depending on requirements and the usage scenario, the best use of your IT budget may be to focus on analytics as a service (AaaS)—supported by internal private cloud, a public cloud, or a hybrid model.

Table 2: how big data and cloud work together

Big Data	Cloud Computing
Variety, volume and velocity require new tools.	Variety of compute/storage and network options.
Potentially massive data sets	Massive/virtually unlimited capacity.
Iterative/experimental style of data manipulation and analysis.	Iterative/experimental style of infrastructure deployment usage.
Frequently not a steady state workload peaks and valleys	At its most efficient with highly variable workloads
Absolute performance not as critical as "time to result"; shared resources are a bottleneck.	Parallel compute projects allow each workgroup to have more autonomy get faster results.

A. Issues of cloud computing and Big data

These are some of the issues that integrated with cloud services when it goes to deliver an economical solution to big data needs. (Judith Hurwitz 2013)

- Data integrity: Different providers have the right controls in place to ensure that the integrity of the data is maintained.
- Compliance: Make sure that provider can comply with any compliance issues particular to the company or industry.
- Data transport: Be sure that definitions of service-level agreements exist for availability, support, and performance.
- Data access: what forms of secure access control are in place and it includes identity management.

- Location: Where will be the data be located Regulatory issues prevent data from being stored or processed on machines in a different country.

V. APPLICATIONS IN CLOUD COMPUTING AND BIG DATA

IBM big data platform that enables comprehensive search and industrial strength industrial-strength platform that includes Hadoop that address the full spectrum of big data business challenges. IBM big data platform include Hadoop-based analytics (Enables distributed processing of large data sets across commodity server clusters), High performance stream computing which enables low latency analytics for stream data, IBM industry leading platform delivering scalable and high performance data warehousing and business analytics and Information integration and governance which deliver a higher level of information confidence in the era of big data. (IBM,2013)

Google made big waves in cloud computing launching its own IaaS service.Google App Engine is the Google's Platform-as-a-Service (PaaS), develop application easily using built-in services that make more productive. Google BigQuery is a web service and REST API tool for quickly analyzing very large datasets.

Oracle Big Data brings big data solutions to mainstream enterprises and it is the first vendor to offer a complete and integrated solution to big data requirements. New big data technologies, such as Hadoop and Oracle NoSQL database, run alongside Oracle data warehouse to deliver business value and address big data requirements.

Apache Hadoop is an open-source software framework first used by Yahoo and Facebook. Yahoo is the largest tester and contributor to Hadoop. Facebook recently installed Facebook Messages, its user-facing application Built on the Apache Hadoop platform. Cloudera introduced commercial support for enterprises in 2008, and MapR and Hortonworks piled on in 2009 and 2011, respectively, IBM and EMC-spinout Pivotal each has introduced its own Hadoop distribution. Microsoft and Teradata offer complementary software and first-line support for Hortonworks' platform.

VI. FUTURE TRENDS IN BIG DATA AND CLOUD COMPUTING

Cloud computing and big data is the next innovation wave. It will change every industry. Amazon Web Services, Hortonworks and Microsoft have been playing a big role in the enterprises to bring Big Data into the cloud platform. At BigDataSV 2014 report mentions, Kinesis is a streaming data framework for real-time applications released by Amazon web services, and RedShift fully managed, data warehouse service. Hortonworks versatile Hadoop data platform and Microsoft HDInsights delivers Hortonworks Data Platform (HDP) on Microsoft's Azure cloud.

The following are the top cloud computing trends. Cloud and big data will offer new field in security platform can be more effectively governed, managed and controlled at the network. Better supply chain optimization and visibility possible with cloud and big data platform which will be centralized and share common view.

VII CONCLUSION & DISCUSSION

Cloud technology is powerful IT resource pool (public cloud), or management tool (private cloud). Its elastic property agrees application resizing itself to any scale. The almost important advantage of using cloud is paying and using the IT resource as you go. How much you have used, how much you should pay. These features of cloud enable small company or organization like school running large scale application or experiment with a reasonable cost.

Big data puts forward new challenges for data management and analysis, and even for the whole IT industries. Hadoop & MapReduce is one of the very powerful frameworks that enable easy development on data-intensive application. Furthermore, challenges and issues in cloud computing and big data will require solutions, and it will not be addressed naturally by the next generation of industrial products. Industry must support and encourage research towards addressing these technical challenges to achieve the promised benefits of Big Data.

REFERENCES

Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., …Widom, J., (2012). Challenges and Opportunities with Big Data Challenges and Opportunities with Big Data. Available at: http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf

Amazon web services, 2014. Amazon EC2 [online]. Available from: http://aws.amazon.com/ec2/ [Accessed 10 June 2014].

Buyya, R., Broberg, J. & Goscinski, A., (2011). Cloud Computing Principles and Paradigms, A JOHN WILEY & SONS,INC.

C., Paul, D. deRoos, (2013). Harness the power of Big Data, McGraw Hill

Dean, J. & Ghemawat, S., (2008). MapReduce : Simplified Data Processing on Large Clusters L. P. Daniel, ed. Communications of the ACM, 51(1), pp.1–13. Available at: http://portal.acm.org/citation.cfm?id=1327492.

Ekanayake, J. & Fox, G., (2010). High performance parallel computing with clouds and cloud technologies. Cloud Computing, pp.20–38. Available at: http://link.springer.com/chapter/10.1007/978-3-642-12636-9_2 [Accessed March 31, 2014].

Ganesan, H., & Olety, V. (2012). Introduction to Big Data Hadoop Ecosystem. Availabe at: http://cloudstory.in/2012/04/introduction-to-big-data-hadoop-ecosystem-part-1/

Guo, S., (2013). Hadoop Operations and Cluster Management Cookbook, Packt. [Accessed April 17, 2014]

Gupta, A. et al., (2013). The Who , What , Why and How of High Performance Computing Applications in the Cloud.

Ji, C. et al.,(2012). Big Data Processing in Cloud Computing Environments. 2012 12th International Symposium on Pervasive Systems, Algorithms and Networks. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6428800 [Accessed June 1, 2014].

Judith Hurwitz, Alan Nugent,et.al., (2013). Blg Data for Dummies, John Wiley & Sons.

IBM, (2013). The IBM big data platform, Available at: http://www-01.ibm.com/software/data/bigdata/.

IBM, 2013. Cloud-managed backup [online]. Available from: http://www-935.ibm.com/services/us/en/it-services/business-continuity/cloud-managed-backup/index.html [Accessed 10 June 2014].

Lavastorm, 2013. The Top Challenges in Big Data and Analytics,

Marcos D. Assuncao, Rodrigo N. Calheiros, Silvia Bianchi, Marco A. S. Netto, R.B., (2013). Big Data Computing and Clouds : Challenges , Solutions , and Future Directions, Cloud Computing and Distributed Systems Laboratory, The University of Melbourne.

Mahmood, Z. & Approaches, P.,(2013). Cloud Computing Methods and Practical Approches, Springer.

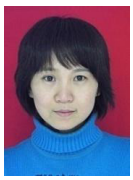
mongoDB ,2014. Big Data Explained [online]. Available from:http://www.mongodb.com/big-data-explained [Accessed 10 June 2014]

SAS, (2012). Five big data challenges Data visualization is becoming an increasingly important component of analytics in the age of big data ., Available at: http://www.sas.com/content/dam/SAS/en_us/doc/other1/five-big-data-challenges-106263.pdf.

Velte, T., Velte, A. & Elsenpeter, R.,(2010). Cloud computing, a practical approach, McGraw Hill. Available at: http://dl.acm.org/citation.cfm?id=1594816 [Accessed June 17, 2014].

BIOGRAPHY OF AUTHORS

[1]K.P.N Jayasena is an currently a probationary lecturer attached to the department of Computing and Information Systems, Faculty of Applied Sciences,Sabaragamuwa University of Sri Lanka. She is currently pursuing MSc in Computer Science and Technology, Wuhan University of Technology China.



[2]Huazhu Song is a PhD, Associate professor in the Computer Science and Technology, Wuhan University of Technology China.