IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-5, Issue-9 E-ISSN: 2347-2693

A Survey on Various Issues Big Data in Cloud Computing

Gaurav Jain[1*], Kunal Gupta[2], Arpit Kushwah[3], Abhishek Agrawal[4]

[1*]Department of CSE and IT, M.I.T.S. Gwalior, India
[2]Department of CSE and IT, M.I.T.S. Gwalior, India
[3]Department of CSE and IT, M.I.T.S. Gwalior, India
[4]Department of CSE and IT, M.I.T.S. Gwalior, India

[*]*Corresponding Author: gauravjainmitsgwalior6@gmail.com, Tel.: +91-8989217678*

Available online at: www.ijcseonline.org

Received: 19/Aug/2017, Revised: 28/Aug/2017, Accepted: 12/Sep/2017, Published: 30/Sep/2017

***Abstract*—** Big data is a key concept that cannot be over looked in the IT world considering the prominent increase in data, and data related services, it is important to explore this field and look at ways to improve data service delivery especially in the cloud. (CC) on the other hand helps in tackling the issue of storage and data service. This research focus on the two key concept big data and (CC) and some of the issues and challenges that are inherent with the deployment of cloud services and big data. The shows study that security challenges are among the most prominent issue in the cloud and big data services. The plumbing issue and some other issues such as the issue of the cost to run cloud services in handling big data were observed. Also the issues of service level agreement which gives an organization the assurance of enjoying all services rendered by the organization running the cloud services. After considering some of the issues associated with big data and (CC), some solution was suggested towards improving the two key concepts which will go a long way in increasing the adoption rate of (CC) by organizations. It is important for organizations to consider the nature of how their data will grow in the future before deploying any cloud service in their business.

***Keywords*—** Big Data, Cloud Computing, Data Mining

I. INTRODUCTION

The term 'Big Data' appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the Next Wave of Infra Stress" [1]. It is the term for data sets so large and complicated that it becomes difficult to process using traditional data management tools or processing applications. The origin of the term 'Big Data' is due to the fact that we are creating a huge amount of data every day. At the KDD Big Mine 12 Workshop Usama Fayyad in his invited talk presented amazing data numbers about internet usage, among them are the following: each day Google has more than 1 billion queries, Twitter has more than 250 million tweets per day, per day Facebook has more than 800 million updates, and YouTube has more than 4 billion views per day. Big Data is a heterogeneous mix of data both structured (traditional datasets –in rows and columns like DBMS tables, CSV's and XLS's) and unstructured data like PDF documents, e-mail attachments, images, manuals, medical records such as x-rays, ECG and MRI images, forms, rich media like graphics, video and audio, contacts, forms and documents. Businesses are primarily concerned with managing unstructured data because about 80 percent of enterprise data is unstructured [1]. Google has introduced Map Reduce [2] framework for processing large amounts of data on commodity hardware. Apache's Hadoop distributed file system (HDFS) is evolving as a superior software component for (CC) combined along with integrated parts such as Map Reduce.

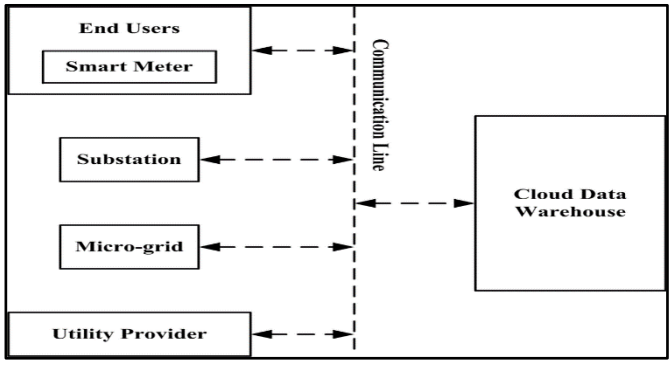

Fig.1 CC Big data

II. IMPORTANCE OF BIG DATA

The government's emphasis is on how big data creates "value" –both within and across disciplines and domains. Value arises from the ability to analyses the data to develop actionable information. The survey of the technical literature

[6] suggests five generic ways that big data can support value creation for organizations.

1. Creating transparency by making big data openly available for business and functional analysis (quality, lower costs, reduce time to market, etc.)
2. Supporting experimental analysis in individual locations that can test decisions or approaches, such as specific market programs.
3. Assisting, based on customer information, in defining market segmentation at more narrow levels.
4. Supporting Real-time analysis and decisions based on sophisticated analytics applied to data sets from customers and embedded sensors.
5. Facilitating computer-assisted innovation in products based on embedded product sensors indicating customer responses. C.

III. TYPES OF BIG DATA AND SOURCES

There are two types of big data: structured and unstructured.

1. Structured Data

Structured Data are numbers and words that can be easily Categorized and analysed. Things like network sensors embedded in electronic devices, smart phones, and global positioning system (GPS) devices generate these data. Structured data also include things like sales figures, account balances, and transaction data.

2. Unstructured Data

Unstructured Data include more complex information, such as customer reviews from commercial websites, photos and other multimedia, and comments on social networking sites. These data cannot easily be separated into categories or analysed numerically. The explosive growth of the Internet in recent years means that the variety and amount of big data continue to grow. Much of that growth comes from unstructured data.

Fig.2 - Big Data Store Types

IV. MAJOR OPEN PROBLEMS

In this concluding part of the tutorial, we identify some of the major open problems that must be addressed to ensure the success of data management systems in the cloud. In summary, a single perfect data management solution for the cloud is yet to be designed. Different systems target different aspects in the design space, and multiple open problems still remain. With respect to *Key-Value* stores, though these systems are popular, they only support. Providing support for ad-hoc querying on top of a *Key-Value* store [4] or providing consistency guarantees at different access granularities [14] are some research efforts targeted towards enriching the functionality supported by *Key-Value* stores. Further research, however is needed to generalize these proposals to different classes of applications and different *Key-Value* stores. Similarly, extending the *Key-Value* stores for supporting rich set of applications is also an important research challenge. On the other hand, in the domain of relational database management, an important open problem is how to make the systems *elastic* for effectively utilizing the available resources and minimizing the cost of operation. Furthermore, characterizing the different consistency semantics that can be provided at different scales, and effective techniques for load balancing are also critical aspects of the system. Designing scalable, elastic, and autonomic multitenant database systems is another important challenge that must also be addressed. In addition, ensuring the security and privacy of the data outsources to the cloud is also an important problem for ensuring the success of data management systems in the cloud.

Following are the learning outcomes:

• State-of-the-art in scalable data management for traditional and (CC) infrastructures for both update heavy as well as analytical workloads. Summary of current research projects and future research directions.
• Design choices that have led to the success of the scalable systems, and the errors that limited the success of some other systems.
• Design principles that should be carried over in designing the next generation of data management systems for the cloud.
• Understanding the design space for DBMS targeted to Supporting update-intensive workloads for supporting large single tenant systems and large multitenant systems.
• Understanding the different forms of multi-tenancy in the Database layer.
• A list of open research challenges in cloud data management That must be addressed to ensure the continued success of DBMSs.

1. Analytics Architecture

It is not clear yet how an optimal architecture of an analytics system should be to deal with historic data and with real-time data at the same time. An interesting proposal is the Lambda architecture of Nathan Marz. The Lambda Architecture solves the problem of computing arbitrary functions on arbitrary data in real time by decomposing the problem into three layers: the batch layer, the serving layer, and the speed

layer. It combines in the same system Hadoop for the batch layer, and Storm for the speed layer.

2. Statistical Significance
It is important to achieve significant statistical results, and not be fooled by randomness. As efron explains in his book about Large Scale Inference it is easy to go wrong with huge datasets and thousands of questions to answer at once.
3. Distributed Mining
Many data mining techniques are not trivial to paralyze. To have distributed versions of some methods, a lot of research is needed with practical and theoretical analysis to provide new methods.

V. ADVANCEMENTS & CONCLUSION

Streaming algorithms represent an alternative programming model for dealing with large volumes of data with limited computational and storage resources. Stream processing is very attractive for working with time-series data (news feeds, tweets, sensor readings, etc.), which is difficult in Map Reduce (once again, given its batch-oriented design). Another system worth mentioning is Pregel, which implements a programming model inspired by Valiant's Bulk Synchronous Parallel (BSP) model. Pig, which is inspired by Google, can be described as a data analytics platform that provides a lightweight scripting language for manipulating large datasets. Similarly, Hive, another open-source project, provides an abstraction on top of Hadoop that allows users to issue SQL queries against large relational datasets stored in HDFS. Therefore, the system provides a data analysis tool for users who are already comfortable with relational databases, while simultaneously taking advantage of Hadoop's data processing capabilities. Map Reduce is certainly no exception to this generalization, even within the Hadoop/HDFS/ Map Reduce ecosystem; it is already observed the development of alternative approaches for expressing distributed computations.

For example, there can be a third merge phase after map and reduce to better support relational operations. Join processing mentioned n the paper can also tackle the Map Reduce tasks effectively. Big data is the "new" business and social science frontier. The amount of information and knowledge that can be extracted from the digital universe is continuing to expand as users come up with new ways to massage and process data. Moreover, it has become clear that "more data is not just more data", but that "more data is different". "Big data" is just the beginning of the problem. Technology evolution and placement guarantee that in a few years more data will be available in a year than has been collected since the dawn of man. If Facebook and Twitter are producing, collectively, around 50 gigabytes of data per day, and tripling every year, within a few years (perhaps 2-4) we are indeed facing the challenge of "big data becoming really big data".

In this work, we have done in-depth reviews on recent efforts dedicated to big data and big data networking. We have reviewed the progress in fundamental big data technologies, important aspects of big data networking, and security in (CC) such as new challenges and opportunities, resource management and performance optimizations are also introduced and discussed with independent viewpoints.

VI. REFERENCES

[1] Muhammad Yasir Shabir, Asif Iqbal, Zahid Mahmood, AtaUllah Ghafoor, "Analysis of Classical Encryption Techniques in CC", ISSN:ll1007-0214ll09/10ll, pp102-113 Volume 21, Number 1, February 2016.

[2] Rajarshi Roy Chowdhury, "Security in CC", International Journal of Computer Applications (0975 – 8887) Volume 96– No.15, June 2014.

[3] Jawahar Thakur and Nagesh Kumar, 'DES, AES, Blowfish: Symmetric Key Cryptography Algorithm Simulation Based Performance Analysis', International Journal of Emerging Technologies and Advanced Engineering (IJETAE). December (2011), ISSN: 2250-2459 Vol. 1, Issue 2.

[4] Neha Jain and Gurpreet Kaur, 'Implementing DES Algorithm in Cloud for Data Security', VSRD International Journal of CS & IT. (2012), Vol.2 Issue 4, pp. 316-321

[5] Er. Ashima Pansotra1 and Er. Simar Preet Singh, "Cloud Security Algorithms",Vol.9, No.10 (2015), pp.353-360 /ISSN: 1738-9976 IJSIA Copyright ⓒ 2015 SERSC.

[6] Rashmi Nigoti, Manoj Jhuria and Dr. Shailendra Singh, 'A survey of Cryptographic Algorithm for (CC)', International Journal of Emerging Technologies in Computational and Applied Science.(2013) ISSN(Print): 2279-0047, (Online): 2279-0055.

[7] B.Persis Urbana Ivy, Purshotam Mandiwa and Mukesh Kumar, 'A Modified RSA Cryptosystem Based on 'n' Prime Number', International Journal of Engineering and Computer Science. Nov (2012) ISSN: 2319-7242 Volume 1 Issue 2.

[8] Shakeeba S. Khan, Prof.R.R. Tuteja, "Security in CC using Cryptographic Algorithms", ISSN(Online): 2320-9801 ISSN (Print): 2320-9798/Vol. 3, Issue 1, January 2015.

[9] Mr.V.Biksham, Dr. D.Vasumathi, "Query based computations on encrypted data through homomorphic encryption in CC security",978-1-4673-9939-5/16©2016 IEEE.

[10] Peidong Sha, Zhixiang Zhu, "the Modification Of Rsa Algorithm To Adapt Fully Homomorphic Encryption Algorithm In CC", 978-1-5090-1256-5/16©2016 ieee.

[11] Viney Pal Bansal, Sandeep Singh, "A Hybrid Data Encryption Technique using RSA and Blowfish for CC on FPGAs", 978-1-4673-8253-3/15©2015 IEEE

[12] Adil Bouti, J¨org Keller, "Towards Practical Homomorphic Encryption in CC",978-1-4673-7741-6/15©2015 IEEE

[13] Atewologun Olumide, Abeer Alsadoon, P.W.c. Prasad, Linh Pham, "A Hybrid Encryption model for Secure CC", 978-1-4673-9190-0/15©20 15 IEEE

[14] Lt. Col. Jatinder Paul Singh1, Dr. Mamta2 and Sunil Kumar, " Authentication and Encryption in CC", 978-1-4799-9855-5/15©2015 IEEE

[15] D. Agrawal, A. El Abbadi, S. Antony, and S. Das. Data Management Challenges in (CC) Infrastructures. In *DNIS*, pages1–10, 2010.

[16] P. Agrawal, A. Silberstein, B. F. Cooper, U. Srivastava, and R. Ramakrishnan. Asynchronous view maintenance for vls databases. In *SIGMOD Conference*, pages 179–192, 2009.

[17] S. Aulbach, D. Jacobs, A. Kemper, and M. Seibold. A comparison of flexible schemas for software as a service. In *SIGMOD*, pages 881–888, 2009.

Authors Profile

Mr. Gaurav Jain pursed Bachelor of Computer Science and Engineering from LNIT Gwalior, RJPV Bhopal in the year 2009- 2013 and Masters in Cyber Security from Madhav Institute of Technology and Science in the year 2017. He has published several research papers in reputed international journals including IEEE and UGC approved Journals. His main research work focuses on Cryptography Algorithms, Network Security, Cloud Security and Privacy, Big Data Analytics, Data Mining based education. He has 2 years of Research Experience.

Mr Kunal Gupta pursed Bachelor of Computer Science and Engineering from SRM University, Chennai in 2015 after that He Worked in Wipro Technologies in IT Security for the span of a year after which he did Masters in Cyber Security from Madhav Institute of Technology and Science in the year 2017. He is an ex-Chairman of IET of MITS, Student Chapter and has papers in IEEE computer society and other UGC approved Journals and it's also available online. His main research work focuses on Cryptography Algorithms, Network Security, Cloud Security, and Privacy, Big Data Analytics, Data Mining, He is Certified Ethical Hacker v9 by EC-Council and also has 2 years of Research Experience.

Mr. *Arpit Kushwah* pursed Bachelor of Computer Science and Engineering from Amity University, Gwalior in year 2011- 2015 and Masters in Computer Science from Madhav Institute of Technology and Science in the year 2017. He has published several research papers in reputed international journals including IEEE and other UGC approved Journals. His main research work focuses on Image Segmentation, Network Mining, Cloud Computing, Big Data Analytics and Data Mining based education. He has 2 years of Research Experience.

Mr. *Abhishek Agrawal* pursed Bachelor of Computer Science and Engineering from Vikrant College RGPV, Gwalior in year 2011- 2015 and Masters in Cyber Security from Madhav Institute of Technology and Science in the year 2017. He has published several research papers in reputed international journals including IEEE and other UGC approved Journals. His main research work focuses on Network Mining, Cloud Computing and Data Mining based education. He has 2 years of Research Experience. Cryptography Algorithms, Network Security, Cloud Security and Privacy and Data Mining based education. He has 2 years of Research Experience.

