Volume 4 Issue IV, April 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Big Data and Cloud Computing: Challenges and Issues in Present Era

K.Abhinayalalitha[1], P.Eswari[2], B.Kavipriya[3], A.S.Sai Venkatesh4
1-ME CSE, Parisutham Institute of Technology & Science

Abstract --Big data is a data analysis methodology enabled by recent advances in technologies and architecture. However, big data entails a huge commitment of hardware and processing resources, making adoption costs of big data technology prohibitive to small and medium sized businesses. Cloud computing is a set of it services that are provided to a customer over a network on a leased basis and with the ability to scale up or down their service requirements. It advantages includes scalability, resilience, flexibility, efficiency and outsourcing non-core activities. Addressing big data is a challenging and time-demanding task that requires a large computational infrastructure to ensure successful data processing and analysis. The rise of big data in cloud computing is reviewed in this study. The definition, characteristics, and classification of big data along with some discussions on cloud computing are introduced. The relationship between big data and cloud computing, big data storage systems, and Hadoop technology are also discussed. Furthermore, research challenges are investigated, with focus on scalability, availability, data integrity, data transformation, data quality, data heterogeneity, privacy, legal and regulatory issues, and governance. Lastly, open research issues that require substantial research efforts are summarized.
Keywords: cloud computing; big data; storage and sharing; security, Information system application.

I. INTRODUCTION

Cloud Computing is a technology that uses the internet and central remote servers to maintain data and applications. Cloud computing allows consumers and businesses to use applications without installation and access their personal files at any computer with internet access. This technology allows for much more efficient computing by centralizing data storage, processing and bandwidth. The continuous increase in the volume and detail of data captured by organizations, such as the rise of social media, Internet of Things (IoT), and multimedia, has produced an overwhelming flow of data in either structured or unstructured format. Data creation is occurring at a record rate [1], referred to herein as big data, and has emerged as a widely recognized trend. Big data is eliciting attention from the academia, government, and industry. Bigdata are characterized by three aspects:(a) data are numerous

(b) data cannot be categorized into regular relational databases, and (c) data are generated, captured, and processed rapidly. Moreover, big data is transforming healthcare, science, engineering, finance, business, and eventually, the society. The advancements in data storage and mining technologies allow for the preservation of increasing amounts of data described by a change in the nature of data held by organizations [2]. The rate at which new data are being generated is staggering [3]. A major challenge for researchers and practitioners is that this growth rate exceeds their ability to design appropriate cloud computing platforms for data analysis and update intensive workloads. Cloud computing is one of the most significant shifts in modern ICT and service for enterprise applications and has become a powerful architecture to perform large-scale and complex computing. The advantages of cloud computing include virtualized resources, parallel processing, security, and data service integration with scalable data storage. Cloud computing can not only minimize the cost and restriction for automation and computerization by individuals and enterprises but can also provide reduced infrastructure maintenance cost, efficient management, and user access [4]. As a result of the said advantages, a number of applications that leverage various cloud platforms have been developed and resulted in a tremendous increase in the scale of data generated and consumed by such applications. Some of the first adopters of big data in cloud computing are users that deployed Hadoop clusters in highly scalable and elastic. computing environments provided by vendors, such as IBM, Microsoft Azure, and Amazon AWS [5]. Virtualization is one of the base technologies applicable to the implemen- tation of cloud computing. The basis for many platform attributes required to access, store, analyze, and manage distributed computing components in a big data environment is achieved through virtualization. Virtualization is a process of resource sharing and isolation of underlying hardware to increase computer resource utilization, efficiency, and scalability. The goal of this study is to implement a comprehensive investigation of the status of big data in cloud computing environments and provide the definition, characteristics, and classification of big data along with some discussions on cloud computing. The relationship

between big data and cloud computing, big data storage systems, and Hadoop technology are discussed. Furthermore, research challenges are discussed, with focus on scalability, avail- ability, data integrity, data transformation, data quality, data heterogeneity, privacy, legal and regulatory issues, and governance. Several open research issues that require substantial research efforts are likewise summarized. Cloud computing is an extremely successful paradigm of service oriented computing, and has revolutionized the way computing infrastructure is abstracted and used. Three most popular cloud paradigms include: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). The concept a however can also be extended to Database as a Service or Storage as a Service. Elasticity, pay-per-use, low upfront investment, low time to market, and transfer of risks are some of the major enabling features that make cloud computing a ubiquitous paradigm for deploying novel applications which were not economically feasible in a traditional enterprise infrastructure settings. This has seen a proliferation in the number of applications which leverage various cloud platforms, resulting in a tremendous increase in the scale of the data generated as well as consumed by such applications. Scalable database management systems (DBMS)—both for update intensive application workloads, as well as decision support systems—are thus a critical part of the cloud infrastructure. Scalable and distributed data management has been the vision of the database research community for more than three decades. Much research has focused on designing scalable systems for both update intensive workloads as well as ad hoc analysis workloads. Initial designs include distributed databases [7] for update intensive workloads, and parallel database systems 19] for analytical workloads. Parallel databases grew beyond prototype systems to large commercial systems, but distributed database systems were not very successful and were never commercialized – rather various ad-hoc approaches to scaling were used. Changes in the data access patterns of applications and the need to scale out to thousands of commodity machines led to the birth of a new class of systems referred to as *Key-Value* stores [8, 10, 12] which are now being widely adopted by various enterprises. In the domain of data analysis, the MapReduce paradigm [10] and its open-source implementation Hadoop [8] has also seen widespread adoption in industry and academia alike. Solutions have also been proposed to improve Hadoop based systems in terms of usability [5,8] and performance. In summary, the quest for conquering the challenges posed by management of big data has led to a plethora of systems. Furthermore, applications being deployed in the cloud have their own set of desideratum which opens up various possibilities in the design space.

II. EXISTING SOLUTIONS AND RELATED WORKS

Cloud computing and big data are conjoined. 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. Cloud computing provides the underlying engine through the use of Hadoop, a class of distributed data-processing platforms. Large data sources from the cloud and Web are stored in a distributed fault-tolerant database and processed through a programming model for large data sets with a parallel distributed algorithm in a cluster. 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. The use of cloud computing in big data is shown in the following figure. The main purpose of data visualization is to view analytical results presented visually through different graphs for decision making. Therefore, cloud computing not only provides facilities for the computation and processing of big data but also serves as a service model. Cloud computing is correlated with a new pattern for the provision of computing infrastructure and big data processing method for all types of resources available in the cloud through data analysis.

Comparison of several big data cloud platforms.

	Google	Microsoft	Amazon	Cloudera
Big data storage	Google cloud services	Azure	S3	
MapReduce	AppEngine	Hadoop on Azure	Elastic MapReduce (Hadoop)	MapReduce YARN
Big data analytics	BigQuery	Hadoop on Azure	Elastic MapReduce (Hadoop)	Elastic MapReduce (Hadoop)
Relational database	Cloud SQL	SQL Azure	MySQL or Oracle	MySQL, Oracle, PostgreSQL
NoSQL database	AppEngine Datastore	Table storage	DynamoDB	Apache Accumulo
Streaming processing	Search API	Streaminsight	Nothing prepackaged	Apache Spark
Machine learning	Prediction API	Hadoop + Mahout	Hadoop + Mahout	Hadoop + Oryx
Data import	Network	Network	Network	Network
Data sources	A few sample datasets	Windows Azure marketplace	Public Datasets	Public Datasets
Availability	Some services in private beta	Some services in private beta	Public production	Industries

Several cloud-based technologies have to cope with this new environment because dealing with big data for concurrent processing has become increasingly complicated [8]. Map Reduce [7] is a good example of big data processing in a cloud environment; it

allows for the processing of large amounts of datasets stored in parallel in the cluster. The rapid growth of data has restricted the capability of existing storage technologies to store and manage data. Over the past few years, traditional storage systems have been utilized to store data through structured RDBMS [13]. However, almost storage systems have limitations and are inapplicable to the storage and management of big data. A storage architecture that can be accessed in a highly efficient manner while achieving availability and reliability is required to store and manage large datasets. Several storage technologies have been developed to meet the demands of massive data. Existing technologies can be classified as direct attached storage (DAS), network attached storage (NAS), and storage area network (SAN). In DAS, various hard disk drives (HDDs) are directly connected to the servers. Each HDD receives a certain amount of input/output (I/O) resource, which is managed by individual applications. Therefore, DAS is suitable only for servers that are interconnected on a small scale. Given the afore said low scalability, storage capacity is increased but expandability and upgradeability are limited significantly. NAS is a storage device that supports a network. NAS is connected directly to a network through a switch or hub via TCP/IP protocols. In NAS, data are transferred as files. Given that the NAS server can indirectly access a storage device through networks, the I/O burden on a NAS server is significantly lighter than that on a DAS server. NAS can orient networks, particularly scalable and bandwidth-intensive networks. Such networks include high-speed networks of optical-fiber connections. The SAN system of data storage is independent with respect to storage on the local area network (LAN).Multipath data switching is conducted among internal nodes to maximize data management and sharing. The organizational systems of data storages (DAS, NAS, and SAN) can be divided into three parts: (i)disc array ,where the foundation of a storage system provides the fundamental guarantee,(ii) connection and network sub systems, which connect one or more disc arrays and servers, and(iii) storage management software, which oversees data sharing, storage management, and disaster recovery tasks for multiple servers. Classic architecture's potential bottleneck is the database server while faced with peak workloads. One database server has restriction of scalability and cost, which are two important goals of big data processing. In order to adapt various large data processing models, D. Kossmann et al. presented four different architectures based on classic multi-tier database application architecture which are partitioning, replication, distributed control and caching architecture[3]. It is clear that the alternative providers have different business models and target different kinds of applications: Google seems to be more interested in small applications with light workloads whereas Azure is currently the most affordable service for medium to large services. Most of recent cloud service providers are utilizing hybrid architecture that is capable of satisfying their actual service requirements. In this section, we mainly discuss big data architecture from three key aspects: distributed file system, non-structural and semi-structured data storage and open source cloud platform.

III. PROPOSED MODEL

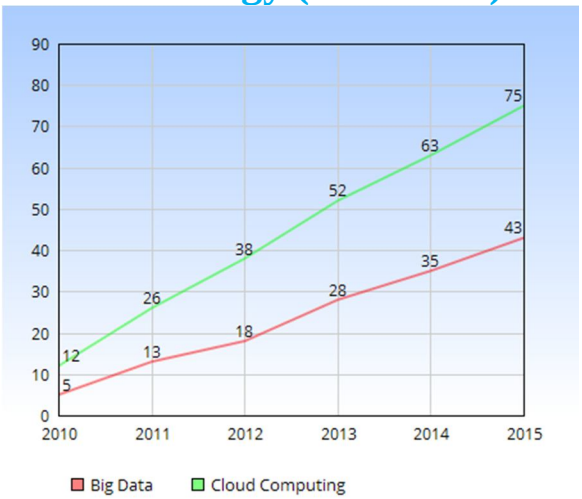
With the success of the Web 2.0, more and more IT companies have increasing needs to store and analyze the ever growing data, such as search logs, crawled web content, and click streams, usually in the range of petabytes, collected from a variety of web services. However, web data sets are usually non-relational or less structured and processing such semi-structured data sets at scale poses another challenge. Moreover, simple distributed file systems mentioned above cannot satisfy service providers like Google, Yahoo!, Microsoft and Amazon. All providers have their purpose to serve potential users and own their relevant state of-the-art of big data management systems in the cloud environments. Bigtable [10] is a distributed storage system of Google for managing structured data that is designed to scale to a very large size (petabytes of data) across thousands of commodity servers. Bigtable does not support a full relational data model. However, it provides clients with a simple data model that supports dynamic control over data layout and format. PNUTS[11] is a massive scale hosted database system designed to support Yahoo!'s web applications. The main focus of the system is on data serving for web applications, rather than complex queries. Upon PNUTS, new applications can be built very easily and the overhead of creating and maintaining these applications is nothing much. The Dynamo[12] is a highly available and scalable distributed key/value based data store built for supporting internal Amazon's applications. It provides a simple primary-key only interface to meet the requirements of these applications. However, it differs from key-value storage system. Facebook proposed the design of a new cluster-based data warehouse system, Llama[13], a hybrid data management system which combines the features of row-wise and column-wise database systems. They also describe a new column-wise file format for Hadoop called CFile, which provides better performance than other file formats in data analysis. The main idea behind data center is to leverage the virtualization technology to maximize the utilization of computing resources. Therefore, it provides the basic ingredients such as storage, CPUs, and network bandwidth as a commodity by specialized service providers at low unit cost. For reaching the goals of big data management, most of the research institutions and enterprises bring virtualization into cloud architectures. Amazon Web Services (AWS), Eucalptus, Opennebula, Cloudstack and Openstack are the most popular cloud

management platforms for infrastructure as a service (IaaS). AWS is not free but it has huge usage in elastic platform. It is very easy to use and only pay-as-you-go. The Eucalyptus[14] works in IaaS as an open source. It uses virtual machine in controlling and managing resources. Since Eucalyptus is the earliest cloud management platform for IaaS, it signs API compatible agreement with AWS. It has a leading position in the private cloud market for the AWS ecological environment. OpenNebula[15] has integration with various environments. It can offer the richest features, flexible ways and better interoperability to build private, public or hybrid clouds. OpenNebula is not a Service Oriented Architecture (SOA) design and has weak decoupling for computing, storage and network independent components. CloudStack10 is an open source cloud operating system which delivers public cloud computing similar to Amazon EC2 but using users' own hardware. CloudStack users can take full advantage of cloud computing to deliver higher efficiency, limitless scale and faster deployment of new services and systems to the enduser. At present, CloudStack is one of the Apache open source projects. It already has mature functions. However, it needs to further strengthen the loosely coupling and component design. OpenStack11 is a collection of open source software projects aiming to build an open-source community with researchers, developers and enterprises. People in this community share a common goal to create a cloud that is simple to deploy, massively scalable and full of rich features. The architecture and components of OpenStack are straightforward and stable, so it is a good choice to provide specific applications for enterprises. In current situation, OpenStack has good community and ecological environment. However, it still have some shortcomings like incomplete functions and lack of commercial supports. In this age of data explosion, parallel processing is essential to perform a massive volume of data in a timely manner. The use of parallelization techniques and algorithms is the key to achieve better scalability and performance for processing big data. At present, there are a lot of popular parallel processing models, including MPI, General Purpose GPU (GPGPU), MapReduce and MapReduce-like. MapReduce proposed by Google, is a very popular big data processing model that has rapidly been studied and applied by both industry and academia. MapReduce has two major advantages: the MapReduce model hide details related to the data storage, distribution, replication, load balancing and so on. Furthermore, it is so simple that programmers only specify two functions, which are map function and reduce function, for performing the processing of the big data. We divided existing MapReduce applications into three categories: partitioning sub-space, decomposing sub-processes and approximate overlapping calculations. While MapReduce is referred to as a new approach of processing big data in cloud computing environments, it is also criticized as a "major step backwards" compared with DBMS[16]. We all know that MapReduce is schema-free and index-free. Thus, the MapReduce framework requires parsing each record at reading input. As the debate continues, the final result shows that neither is good at the other does well, and the two technologies are complementary[12]. Recently, some DBMS vendors also have integrated MapReduce front-ends into their systems including Aster, HadoopDB[13], Greenplum[15] and Vertuca. Mostly of those are still databases, which simply provide a MapReduce front-end to a DBMS. HadoopDB is a hybrid system which efficiently takes the best features from the scalability of MapReduce and the performance of DBMS. The result shows that HadoopDB improves task processing times of Hadoop by a large factor to match the sharednothing DBMS. Lately, J. Dittrich et al. propose a new type of system named Hadoop++[16] which indicates that HadoopDB has also severe drawbacks, including forcing user to use DBMS, changing the interface to SQL and so on. There are also certain papers adapting different inverted index, which is a simple but practical index structure and appropriate for MapReduce to process big data, such as [15] etc. We also do intensive study on large-scale spatial data environment and design a distributed inverted grid index by combining inverted index and spatial grid partition with MapReduce model, which is simple, dynamic, scale and fit for processing high dimensional spatial data.

IV. RESULT AND DISCUSSION

We are now in the days of big data. We can gather more information from daily life of every human being. The top seven big data drivers are science data, Internet data, finance data, mobile device data, sensor data, RFID data and streaming data. Coupled with recent advances in machine learning and reasoning, as well as rapid rises in computing power and storage, we are transforming our ability to make sense of these increasingly large, heterogeneous, noisy and incomplete datasets collected from a variety of sources. So far, researchers are not able to unify around the essential features of big data. Some think that big data is the data that we are not able to process using pre-exist technology, method and theory. However, no matter how we consider the definition of big data, the world is turning into a "helplessness" age while varies of incalculable data is being generated by science, business and society.

Volume 4 Issue IV, April 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)



The above graph shows the difference in the growth of big data and cloud computing domains. Big data put forward new challenges for data management and analysis, and even for the whole IT industry. We consider there are three important aspects while we encounter with problems in processing big data, and we present our points of view in details as follows. Big Data Storage and Management: Current technologies of data management systems are not able to satisfy the needs of big data, and the increasing speed of storage capacity is much less than that of data, thus a revolution re-construction of information framework is desperately needed. We need to design a hierarchical storage architecture. Besides, previous computer algorithms are not able to effectively storage data that is directly acquired from the actual world, due to the heterogeneity of the big data. However, they perform excellent in processing homogeneous data. Therefore, how to re-organize data is one big problem in big data management. Virtual server technology can exacerbate the problem, raising the prospect of overcommitted resources, especially if communication is poor between the application, server and storage administrators. We also need to solve the bottleneck problems of the high concurrent I/O and single-named node in the present Master-Slave system model. Big Data Computation and Analysis: While processing a query in big data, speed is a significant demand[41]. However, the process may take time because mostly it cannot traverse all the related data in the whole database in a short time. In this case, index will be an optimal choice. At present, indices in big data are only aiming at simple type of data, while big data is becoming more complicated. The combination of appropriate index for big data and up-to-date preprocessing technology will be a desirable solution when we encountered this kind of problems. Application parallelization and divide-and-conquer is natural computational paradigms for approaching big data problems. But getting additional computational resources is not as simple as just upgrading to a bigger and more powerful machine on the fly. The traditional serial algorithm is inefficient for the big data. If there is enough data parallelism in the application, users can take advantage of the cloud's reduced cost model to use hundreds of computers for a short time costs. Big Data Security: By using online big data application, a lot of companies can greatly reduce their IT cost. However, security and privacy affect the entire big data storage and processing, since there is a massive use of third-party services and infrastructures that are used to host important data or to perform critical operations. The scale of data and applications grow exponentially, and bring huge challenges of dynamic data monitoring and security protection. Unlike traditional security method, security in big data is mainly in the form of how to process data mining without exposing sensitive information of users. Besides, current technologies of privacy protection are mainly based on static data set, while data is always dynamically changed, including data pattern, variation of attribute and addition of new data. Thus, it is a challenge to implement effective privacy protection in this complex circumstance. In addition, legal and regulatory issues also need attention.

V. CONCLUSION

This paper described a systematic flow of survey on the big data processing in the context of cloud computing. We respectively discussed the key issues, including cloud storage and computing architecture, popular parallel processing framework, major applications and optimization of MapReduce. Big Data is not a new concept but very challenging. It calls for scalable storage index and a distributed approach to retrieve required results near real-time. It is a fundamental fact that data is too big to process conventionally. Nevertheless, big data will be complex and exist continuously during all big challenges, which are the big

opportunities for us. In the future, significant challenges need to be tackled by industry and academia. It is an urgent need that computer scholars and social sciences scholars make close cooperation, in order to guarantee the long-term success of cloud computing and collectively explore new 21 territory.

REFERENCES

[1] Abouzeid, K. B. Pawlikowski, D. J. Abadi, A. Rasin, and A. Silberschatz. HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads. *PVLDB*, 2(1):922–933, 2009.
[2] D. Agrawal, S. Das, and A. E. Abbadi. Big data and cloud computing: New wine or just new bottles? *PVLDB*, 3(2):1647–1648, 2010.
[3] D. Agrawal, A. El Abbadi, S. Antony, and S. Das. Data Management Challenges in Cloud Computing Infrastructures. In *DNIS*, pages 1–10, 2010.
[4] P. Agrawal, A. Silberstein, B. F. Cooper, U. Srivastava, and R. Ramakrishnan. Asynchronous view maintenance for vlsd databases. In *SIGMOD Conference*, pages 179–192, 2009.
[5] 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.
[6] P. Bernstein, C. Rein, and S. Das. Hyder – A Transactional Record Manager for Shared Flash. In *CIDR*, 2011.
[7] M. Brantner, D. Florescu, D. Graf, D. Kossmann, and T. Kraska. Building a database on S3. In *SIGMOD*, pages 251–264, 2008.
[8] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. Bigtable: A Distributed Storage System for Structured Data. In *OSDI*, pages 205–218, 2006.
[9] J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein, and C. Welton. Mad skills: New analysis practices for big data. *PVLDB*, 2(2):1481–1492, 2009.
[10] B. F. Cooper, R. Ramakrishnan, U. Srivastava, A. Silberstein, P. Bohannon, H.-A. Jacobsen, N. Puz, D. Weaver, and R. Yerneni. PNUTS: Yahoo!'s hosted data serving platform. *Proc. VLDB Endow.*, 1(2):1277–1288, 2008.
[11] C. Curino, E. Jones, Y. Zhang, E. Wu, and S. Madden. Relational Cloud: The Case for a Database Service. Technical Report 2010-14, CSAIL, MIT, 2010. http://hdl.handle.net/1721.1/52606.
[12] S. Das, S. Agarwal, D. Agrawal, and A. El Abbadi. ElasTraS: An Elastic, Scalable, and Self Managing Transactional Database for the Cloud. Technical Report 2010-04, CS, UCSB, 2010.
[13] S. Das, D. Agrawal, and A. El Abbadi. ElasTraS: An Elastic Transactional Data Store in the Cloud. In *USENIX HotCloud*, 2009.
[14] S. Das, D. Agrawal, and A. El Abbadi. G-Store: A Scalable Data Store for Transactional Multi key Access in the Cloud. In *ACM SOCC*, 2010.
[15] S. Das, S. Nishimura, D. Agrawal, and A. El Abbadi. Live Database Migration for Elasticity in a Multitenant Database for Cloud Platforms. Technical Report 2010-09, CS, UCSB, 2010.
[16] S. Das, Y. Sismanis, K. Beyer, R. Gemulla, P. Haas, and J. McPherson. Ricardo: Integrating R and Hadoop. In *SIGMOD*, 2010.
[17] J. Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters. In *OSDI*, pages 137–150, 2004.
[18] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels. Dynamo: Amazon's highly available key-value store. In *SOSP*, pages 205–220, 2007.
[19] D. J. Dewitt, S. Ghandeharizadeh, D. A. Schneider, A. Bricker, H. I. Hsiao, and R. Rasmussen. The Gamma Database Machine Project. *IEEE Trans. on Knowl. and Data Eng.*, 2(1):44–62, 1990.
[20] The Apache Hadoop Project. http://hadoop.apache.org/core/, 2009.