Computing Big Data using Cloud

Deepa T Kadkol[1] and Shruthi P[2]
1.Student,T John College Of Engineering,Bangalore
2.Student,T John College Of Engineering,Bangalore

***Abstract*— Effective management and analysis of large-scale data poses an interesting but critical challenge. Recently, big data has attracted a lot of attention from academia, industry as well as government. This paper introduces several big data processing technics from system and application aspects. First, from the view of cloud data management and big data processing mechanisms, we present the key issues of big data processing, including cloud computing platform, cloud architecture, cloud database and data storage scheme. Following the MapReduce parallel processing framework, we then introduce MapReduce optimization strategies and applications reported in the literature. Finally, we discuss the open issues and challenges, and deeply explore the research directions in the future on big data processing in cloud computing environments.**

***Keywords*-Big Data; Cloud Computing; Data Management; Distributed Computing.**

I. INTRODUCTION

In the last two decades, the continuous increase of computational power has produced an overwhelming flow of data. Big data is not only becoming more available but also more understandable to computers. For example, The famous social network Website, Facebook, serves 570 billion page views per month, stores 3 billion new photos every month, and manages 25 billion pieces of content2.Google's search and ad business, Facebook, Flickr, YouTube, and Linkedin use a bundle of artificial-intelligence tricks, require parsing vast quantities of data and making decisions instantaneously. Multimedia data mining platforms make it easy for everybody to achieve these goals with the minimum amount of effort in terms of software, CPU and network.

On March 29, 2012, American government announced the "Big Data Research and Development Initiative", and bigdata becomes the national policy for the first time3. All these examples showed that daunting big data challenges and significant resources were allocated to support these data intensive operations which lead to high storage and data processing costs. The current technologies such as grid and cloud computing have all intended to access large amounts of computing power by aggregating resources and offering a single system view. Among these technologies, cloud computing is becoming a powerful architecture to perform large-scale and complex computing, and has revolutionized the way that computing infrastructure is abstracted and used. In addition, an important aim of these technologies is to deliver computing as a solution for tackling big data, such as largescale, multi-media and high dimensional data sets.

The first documented use of the term "big data" appeared in a

1997 paper by scientists at NASA, describing the problem they had with visualization (i.e. computer graphics) which "provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data. When data sets do not fit in main memory (in core), or when they do not fit even on local disk, the most common solution is to acquire more resources."

Hence, Bigdata can be defined in various perspectives such as:

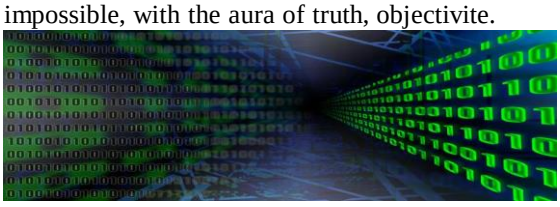
1.Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Although big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data.

2."Big data refers to a process that is used when traditional data mining and handling techniques cannot uncover the insights and meaning of the underlying data. Data that is unstructured or time sensitive or simply very large cannot be processed by relational database engines. This type of data requires a different processing approach called big data, which uses massive parallelism on readily-available hardware."

3.We define Big Data as a cultural, technological, and scholarly phenomenon that rests on the interplay of:

(A)Technology: Maximizing computation power and algorithm accuracy to gather, analyze, link, and compare large data sets. (B) Analysis: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.

(C) Mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivite.



The goal of this paper is to provide the status of big data studies and related works, which aims at providing a general view of big data management technologies andapplications. We give an overview of major approaches and classify them with respect to their strategies including big data management platform, distributed file system, big data storage, MapReduce application and optimization. However, maintaining and processing these large-scale data sets is typically beyond the reach of small businesses and it is increasingly posing challenges even for large companies and institutes. Finally, we discuss the open issues and challenges in processing big data in three important aspects: big data storage, analysis and security.

II. BIG DATA MANAGEMENT SYSTEM

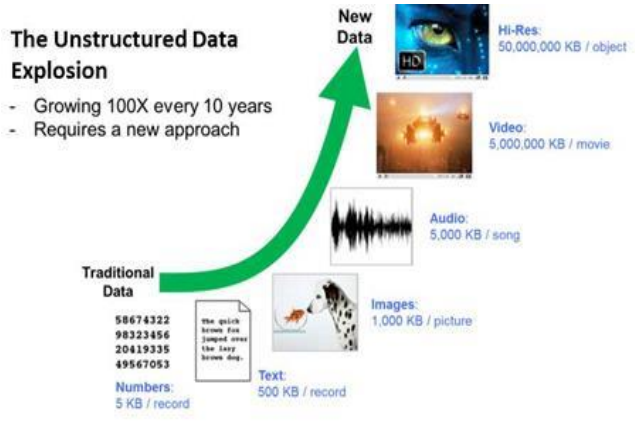
Many researchers have suggested that commercial DBMSs are not suitable for processing extremely large scale data.Big Data Management System" is a totally generic term: it's what many organizations need to run their business in this new era of big data; and it's what vendors need to deliver or help their customers to acquire and build.Most big data environments go beyond relational databases and traditional data warehouse platforms to incorporate technologies that are suited to processing and storing nontransactional forms of data. The increasing focus on collecting and analyzing big data is shaping new platforms that combine the traditional data warehouse with big data systems in a logical data warehousing architecture. As part of the process, the must decide what data must be kept for compliance reasons, what data can be disposed of and what data should be kept and analyzed in order to improve current business processes or provide a business with a competitive advantage. This process requires careful data classification so that ultimately, smaller sets of data can be analyzed quickly and productively. 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.

A. Distributed File System

A method of storing and accessing files based in a client/server architecture. In a distributed file system, one or more central servers store files that can be accessed, with proper authorization rights, by any number of remote clients in the network .The distributed system uses a uniform naming convention and a mapping scheme to keep track of where files are located. When the client device retrieves a file from the server, the file appears as a normal file on the client machine, and the user is able to work with the file in the same ways as if it were stored locally on the workstation. When the user finishes working with the file, it is returned over the network to the server, which stores the now-altered file for retrieval at a later time. Distributed file systems can be advantageous because they make it easier to distribute documents to multiple clients and they provide a centralized storage system so that client machines are not using their resources to store files.

B. Unstructured and Semi-structured Data Storage

"Unstructured data" usually refers to information that doesn't reside in a traditional row-column database. As you might expect, it's the opposite of structured data -- the data stored in fields in a database.Unstructured data files often include text and multimedia content. Examples include e-mail messages, word processing documents, videos, photos, audio files, presentations, webpages and many other kinds of business documents. Note that while these sorts of files may have an internal structure, they are still considered "unstructured" because the data they contain doesn't fit neatly in a database. Semi-structured data is data that has not been organized into a specialized repository, such as a database, but that nevertheless has associated information, such as metadata, that makes it more amenable to processing than raw data.n semi-structured data, the entities belonging to the same class may have different attributes even though they are grouped together, and the attributes' order is not important.Semi-structured data is increasingly occurring since the advent of the Internet where full-text documents and databases are not the only forms of data any more and different applications need a medium for exchanging information. In object-oriented databases, one often finds semi-structured data.

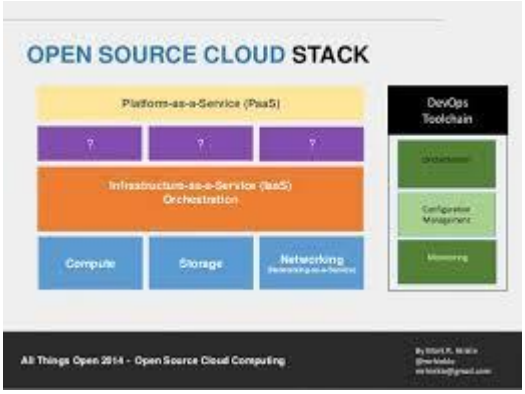


C. Open Source Cloud Platform

The main idea behind data center is 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 serviceIt is very easy to use and only pay-as-you-go. The Eucalyptus 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 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.



III. APPLICATIONS AND OPTIMIZATION

A. Application

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 algorithmic 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), Map Reduce 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 Map Reduce applications into three categories: partitioning sub-space, decomposing sub- processes and approximate overlapping calculations. MapReduce has received a lot of attentions in many fields, including data mining, information retrieval, image retrieval, machine learning, and pattern recognition. For example, Mahout12 is an Apache project that aims at building scalable machine learning libraries which are all implemented on the Hadoop. However, as the amount of data that need to be processed grows, many data processing methods have become not suitable or limited.

B. Optimization

In this section, we present details of approaches to improve the performance of processing big data with MapReduce.

1) Data Transfer Bottlenecks: It is a big challenge that cloud users must consider how to minimize the cost of data transmission. Consequently, researchers have begun to propose variety of approaches. Map-Reduce-Merge is a new model that adds a Merge phase after Reduce phase that combines two reduced outputs from two different MapReduce jobs into one, which can efficiently merge data that is already partitioned and sorted (or hashed) by map and reduce modules. Map-Join- Reduce is a system that extends and improves MapReduce runtime framework by adding Join stage before Reduce stage to perform complex data analysis tasks on large clusters.

2) Iterative Optimization: MapReduce also is a popular platform in which the dataflow takes the form of a directed acyclic graph of operators. However, it requires lots of I/Os and unnecessary computations while solving the problem of iterations with MapReduce. Twister proposed by J. Ekanayake et al. is an enhanced MapReduce runtime that supports iterative MapReduce computations efficiently, which adds an extra Combine stage after Reduce stage. Thus, data output from combine stage flows to the next iteration's Map stage.

3) Online: There are some jobs which need to process online while original MapReduce can not do this very well. MapReduce Online is desgined to support online aggregation and continuous queries in MapReduce. It raises an issue that frequent checkpointing and shuffling of intermediate
results limit pipelined processing. They modify MapReduce framework by making Mappers push their data temporarily stored in local storage to Reducers preiodically in the same MR job. In addition, Map-side pre-aggregation
is used to reduce
communication.

4) Join Query Optimization: Join Query is a popular problem in big data area. However a join problem needsmore than two inputs while MapReduce is devised for processing a single input.To reduce shuffling and computational costs, they desgine an effective mapping mechanism that exploits pruning rules for distance filtering.

IV. DISCUSSION AND CHALLENGES

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. 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.

Big Data Computation and Analysis: While processing a query in big data, speed is a significant demand. 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.

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 territory.

VI.REFERENCES

1. D. Kossmann, T. Kraska, and S. Loesing, "An evaluation of alternative architectures for transaction processing in the cloud," in *Proceedings of the 2010 international conference on Management of data*. ACM, 2010, pp. 579–590.
2. S. Sakr, A. Liu, D. Batista, and M. Alomari, "A survey of large scale data management approaches in cloud environments," *Communications Surveys & Tutorials, IEEE*, vol. 13, no. 3, pp. 311–336, 2011.
3. F. Chang, J. Dean, S. Ghemawat, W. Hsieh, D. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. Gruber, "Bigtable: A distributed structured data storage system," in *7th OSDI*, 2006, pp. 305–314.
4. B. Cooper, R. Ramakrishnan, U. Srivastava, A. Silberstein, P. Bohannon, H. Jacobsen, N. Puz, D. Weaver, and R. Yerneni, "Pnuts: Yahoo!'s hosted data serving platform," *Proceedings of the VLDB Endowment*, vol. 1, no. 2, pp. 1277–1288, 2008.
5. G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels, "Dynamo: amazon's highly available keyvalue store," in *ACM SIGOPS Operating Systems Review*, vol. 41, no. 6. ACM, 2007, pp. 205–220.
6. Y. Lin, D. Agrawal, C. Chen, B. Ooi, and S. Wu, "Llama: leveraging columnar storage for scalable join processing in the mapreduce framework," in *Proceedings of the 2011 international conference on Management of data*. ACM, 2011, pp. 961–972.
7. D. Jiang, B. Ooi, L. Shi, and S. Wu, "The performance of mapreduce: An in-depth study," *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 472–483, 2010.
8. R. Vernica, M. Carey, and C. Li, "Efficient parallel setsimilarity joins using mapreduce," in *SIGMOD conference*. Citeseer, 2010, pp. 495–506.
9. C. Zhang, F. Li, and J. Jestes, "Efficient parallel knn joins for large data in mapreduce," in *Proceedings of the 15th International Conference on Extending Database Technology*. ACM, 2012, pp. 38–49.
10. X. Zhou, J. Lu, C. Li, and X. Du, "Big data challenge in the management perspective," *Communications of the CCF*, vol. 8, pp. 16–20, 2012.

.