Introduction to Big Data Analytics using Cloud Computing

Chhaya Zala*, Pruthvi Patel

Department of Computer Engineering, GCET, India

Department of Information Technology, Shantilal Shah Engineering College, India

Abstract: The Massive increment in the volume and detail of the information collected by organizations has created enormous stream of data in either structured or unstructured form which has been defined as Big Data. For the storage of this kind of data, its processing and analysis, Cloud computing provides a good platform. It provides complex computing platform and helps to distribute the need of costly processing requirement, programming and storage as a latest innovation. There are some issues while using this cloud based data analysis platform that need to be solved before utilizing it. This paper provides the information about the current research, some open issues and future scope in this field.

Keywords: *Big Data, Cloud Computing, Big Data Analytics, Big Data Analytics in Cloud Environment, Cloud-based Big Data Analytics*

1.1. Introduction

With the beginning of the digital era, data are being continuously generated, shared, stored and manipulated across the world. All of these data are coming from the variety of sources like audio or video data, the WebPages and various data warehouses. The result of this continuous data generation increases the complexity. All these data should be efficiently and effectively managed, shared and analyzed to extract meaningful information. These requirements of providing analytics service, programming environments and tools can be fulfilled with big data analytics.

Big Data Analytics can be used in many applications like medical research, transportation, social media (for security and management issues) etc. It is also seful in economics, scientific and environmental sector [1]. Applications like Face book, LinkedIn, Twitter, Amazon, eBay, Google+ etc. requires high storage and processing capacity of their data. Additionally, the data mining algorithms that are used for analytics of the data should have high performing processors. To fulfill all these requirements one of the popular and important solution is cloud. Big data using cloud is therefore widely use now a days.

1.2. Characteristics of big data

Big data means not only just the size of data. It covers other characteristics as follows:

Volume: It is the quantity of data produced from diverse sources.

Variety: It states that Big Data can be either structured or unstructured like text, audio, video, log files etc.

Velocity: It is stated as the speed of generation or transmission of data.

Veracity: Forming trust and reliability in Big Data can be referred as veracity.

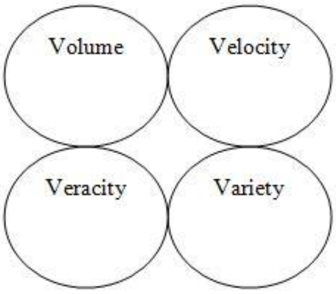

Fig.1.1. Characteristics of Big Data

1.3. Cloud Computing

The cloud computing provides platform for developing, installing and implementing the software and data applications 'as a service'. Mainly cloud providers offer three different services IaaS, PaaS and SaaS that is Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) [2]. Infrastructure which inlcudes storage, processing power, and virtual machines are provided by IaaS. The cloud provider fulfills the needs of the client by virtual resources according to the need; PaaS is constructed on top of IaaS, here users can install cloud applications and run time environments will be supported by cloud service provides. It is at this level where big data DBMS are implemented and SaaS comprises of applications running straight in the cloud. It is the most popular model of cloud services.

These basic services are thoroughly related as SaaS is established over PaaS and eventually PaaS is built on a top of IaaS. One more service is DaaS. DaaS is Data as a Service. It provides the public data sets to perform analytics on data. The storage cost has extraordinarily reduced by using the cloud environment. Furthermore, the ’pay-as-you-go’ model permit captivating and appropriate handling of widespread data, proposing rise to the idea of big data as a service. An example of one such phase is Google BigQuery. It is used to provide the Big Data in the Cloud environment [3].

Cloud computing and Big Data are well connected to each other. Big Data proposes users the ability to use commodity computing to process distributed queries across multiple datasets and after that it returns the resultant sets in a well-timed manner whereas cloud computing gives the underlying engine through the use of Hadoop. The Hadoop is a class of distributed data-processing platforms [4]. It is an open source programming structure. It is used for processing and storing big data in an adopted style on huge groups of item equipment.

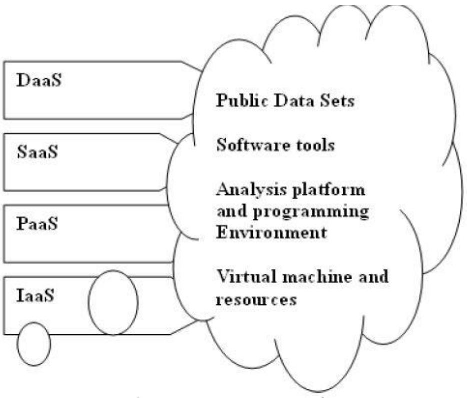
DaaS
SaaS
PaaS
IaaS
Public Data Sets
Software tools
Analysis platform and programming Environment
Virtual machine and resources

Fig.1.2. Cloud computing services [5]

The Hadoop environment contains a varied range of tools. Hadoop Distributed File System and Hadoop MapReduce are the center parts of Hadoop. Hadoop Distributed File System (HDFS) is a virtual document framework that resembles some other record framework [6]. With the exception of that when you move a record on HDFS; document is divided into numerous little documents. For the adaptation to internal failure requirements, each of those records is reproduced and put away on (ordinarily, might be altered) three servers [6].

Hadoop MapReduce is an approach that divides every task into smaller tasks which are sent to numerous smaller servers, permitting a really adaptable utilization of CPU power [6]. There are not many hands-on applications of big data analytics that utilizes the cloud. Information security and Data privacy are two main issues in research emphasis towards big data analysis using cloud.

1.4. Big data analytics in cloud environment

Big Data Analytics could not be performed on the traditional data management tools or data mining techniques because of the large volume of data and capacity of the datasets.

In 1980s many artificial intelligence-based algorithms were developed for data minin. Kumar, Wu, Ghosh, Quinlan, Motoda, Yang, et al mention the ten most influential data mining algorithms k-means, C4.5, Apriori, PageRank, SVM, AdaBoost, CART, naive bayes and kNN (k-Nearest Neighbours) [3]. Most of them have been use commercially as well. Alam and Shakil [7] propose architecture for management of data through cloud techniques. Apache’s Hadoop Distributed File System (HDFS) is emerging as a widespread programming fragment for cloud computing, joined alongside integrating parts, for example, Map Reduce. Hadoop and cloud computing has a substantial cost benefits. It is also facilitating quicker and improved decision making as the future depends mostly on data driven decisions and also a lot of space in enhancement either with new applications or services [6]

MapReduce is most popular model to process the data on the clusters of computers. Quadir, Jackson, Bharathi and Vijayakumar provide a survey on the programming models that support big data analytics [4]. They identify MapReduce as the most productive model for Big Data Analytics. These frameworks are used for storing and processing of data. To store this data, which may be is any of the structure databases like BigTable, HBase, and HadoopDB can be used. The Pig and Hive technologies can be used when it comes to data processing.

The frameworks are used for storing and processing of data. To store this data, which may be is any of the structure databases like BigTable, HBase, and HadoopDB can be used. The Pig and Hive technologies can be used when it comes to data processing.

Cloud providers preferred the MapReduce model because it accelerates the processing of large amounts of data in a cloud. In the server’s cluster, an interface that allows distributed computing and parallelization is provided by MapReduce [9]. It is one of the popular cloud computing framework that helps in scalable distributed applications robotically

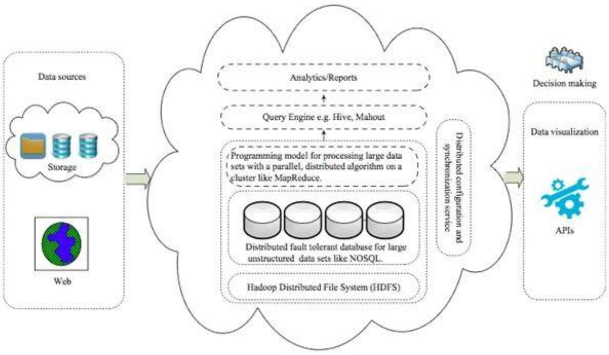
Data sources
Storage
Web
Analytics/Reports
Query Engine e.g. Hive, Mahout
Programming model for processing large data sets with a parallel, distributed algorithm on a cluster like MapReduce.
Distributed fault tolerant database for large unstructured data sets like NOSQL.
Hadoop Distributed File System (HDFS)
Distributed configuration and synchronization service
Decision making
Data visualization
APIs

Fig.1.3. Data Analysis in Cloud platform [5]

1.5. Related work

Industries and organizations can take help of data stored in a cloud-based database for their decision-making processes. Data analysts have large amount of data to work with along with high processing power to handle it while working with cloud base data. They can handle large numbers of records with many attributes which increases predictability. This cloud based big data analysis discovers a new behavioral data such as websites visited or location on a daily basis to create a big data management framework for the cloud. Some Research efforts have been made in this domain too.

Wang, Zhao, Sun, Tao, Kolodziej, Chen, Ranjan, Streit and Geor-gakopoulos suggested a G-Hadoop based framework for security [10]. As security is one of the main concerns they used SSL and public key cryptography the security of big data available on distributed cloud data centers. A model which provides a schema for data in cloud is proposed by Naqvi, Alam ,Rizvi and khan [11]. They also mentioned the easy process of querying data for the user. Balachandran, Bala M and Shivika Prasad [2] presents Cloud-based big data analytics service model. In this model, a public or private cloud is used to provide elements of the big data analytics process.

Jain, Vinay Kumar, and Shishir Kumar [12] says that Cloud computing provides the support for Big Data deployment. Bandwidth and integration issues can become major obstacles for the use of Big Data on the cloud. They also have listed the various problems with Big Data like security, privacy, widespread deployment. They have discussed the various techniques of computation of Big Data in cloud environment along with their benefits. Big Data Analysis in real time become interesting domain and it has gathered the attention of the research community all over the world. Real-time analysis is also provided by many commercial cloud service providers. AWS based-solutions for real-time stream processing are AWS Kinesis, Apache S4, IBM InfoSphere Streams and Storm is some of the frameworks.

This framework is used to simplify the processes of submitting job and authenticating users. Areas like programming abstracts or scalable high-level models and tool; and its development is suggested by Talia [5]. Computing inter-operability issues of data, integration of big data analytics frameworks is mentioned by him. Big data mining provenance and its techniques are available in [5]. To offer cloud-based data analytics services and its application a subscription-based pay-per-use pricing model is widely used. Cloud Analytics as a Service (CLAaaS) model is becoming popular as SaaS[13].In this model, analytics is voluntarily accessible through a cloud computing platform. Anytime anywhere basis automation processes can be achieved in businesses by using such a cloud-based data analytics service.

1.6. Future Scope

Security is one of the greatest concerns while using big data analytics based on cloud computing in an integrated model. Actually security is the reason why this aspect of big data analytics using cloud, implementation and its practical usage has pulled in large consideration [14].

Additional effort must be employed in standardizing data types and developing security mechanisms that ensure that data is accessed quickly and that encryption does not affect processing times so badly. New and secure QoS (quality of service) based data uploading mechanisms can be used to ease data uploading onto the cloud. Fully automatic reactive and proactive systems development should be the major concern to deal with load requirements automatically.

Cost-effectiveness, easy setting, easy testing, easy accessibility are the main reasons why cloud based big data analytics are important today. Cost effective and efficient services using cloud-based big data analytics can be achieve if service providers provides a provision for the availability of these data analytics on the cloud. Some of the main research directions include evolution of analytics and information management with respect to cloud-based analytics, alteration and advancement of techniques and strategies to improve efficiency, analysis and adaptation of legal and ethical practices with respect to the changing viewpoint, impact and effects of technological advances.

1.7. Conclusion

With data increasing on a daily base, big data systems and analytic tools have become a major strength of innovation that provides a way to store, process and get information over peta byte datasets. Cloud environments strongly control big data solutions by providing fault-tolerant, scalable and available environments to big data systems.

Big data analysis and analytics using cloud have become more and more relevant if we consider the rate of data creation in this digital world. Shifting big data analytics to cloud framework is a viable option causes most of the data is stored on clod itself. While big data systems are powerful systems that enable both enterprises and science to get insights over data, there are some concerns that need further exploration. Selection and implementation of effective big data solution with cloud architecture is somehow risky because it needs to be secure.

References:

[1] Chen, CL Philip, and Chun-Yang Zhang: Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences 275 (2014): 314-347.

[2] Balachandran, Bala M., and Shivika Prasad : Challenges and Benefits of Deploying Big Data Analytics in the Cloud for Business Intelligence. Procedia Computer Science 112 (2017): 1112-1122.

[3] Wu, Xindong, et al. "Top 10 algorithms in data mining." Knowledge and information systems 14.1 (2008): 1-37.

[4] Jackson, J. Christy, et al.: Survey on programming models and environments for cluster, cloud, and grid computing that defends big data. Procedia Computer Science 50 (2015): 517-523).

[5] Talia, Domenico. : Clouds for scalable big data analytics. Computer 46.5 (2013): 98-101.

[6] Yetis, Yunus, et al.: Application of big data analytics via cloud computing. World Automation Congress (WAC), 2016. IEEE, 2016.

[7] Alam, Mansaf and Kashish Ara Shakil. : Cloud database management system architecture. UACEE International Journal of Computer Science and its Applications 3.1 (2013): 27-31.

[8] Khan, Samiya, Kashish Ara Shakil, and Mansaf Alam. : Cloud-Based Big Data Analytics - A Survey of Current Research and Future Directions. Big Data Analytics. Springer, Singapore 595-604.

[9] Hashem, Ibrahim Abaker Targio, et al.: The rise of "big data" on cloud computing: Review and open research issues. Information Systems 47 (2015): 98-115.

[10] Zhao, Jiaqi, et al.: A security framework in G-Hadoop for big data computing across distributed Cloud data centres. Journal of Computer and System Sciences 80.5 (2014): 994-1007.

[11] Khan, Imran, et al.: Data model for Big Data in cloud environment. Computing for Sustainable Global Development (INDIA Com), 2015 2nd International Conference on IEEE, 2015.

[12] Jain, Vinay Kumar, and Shishir Kumar. : Big Data Analytic Using Cloud Computing. Advances in Computing and Communication Engineering (ICACCE), 2015 Second International Conference on. IEEE, 2015.

[13] Zulkernine, Farhana, et al.: Towards cloud-based analytics-as-a-service (claaas) for big data analytics in the cloud. Big Data (Big Data Congress), 2013 IEEE International Congress on. IEEE, 2013.

[14] Liu, Chang, et al.: External integrity verification for outsourced big data in cloud and IoT: A big picture. Future Generation Computer Systems 49 (2015): 58-67.

[15] Neves, Pedro Caldeira, et al.: Big Data in Cloud Computing: features and issues. Conference: International Conference on Internet of Things and Big Data. 2016.