ISSN 2320 – 2602

Volume 4, No.3, March 2015

International Journal of Advances in Computer Science and Technology

Available Online at http://warse.org/pdfs/2015/ijacst04432015.pdf

STORING AND QUERYING OF BIG DATA WITH DATA SECURITY

S.Md.Mujeeb, Dr.V.S.Giridhar Akula
Assistant Professor, Malla Reddy Institute of Technology & Science, Hyderabad
mujeeb.smd@gmail.com
Professor & Principal, Methodist College of Engineering & Technology, Hyderabad
seshagiridhar.a@gmail.com

ABSTRACT

The present day organizations are managing large amounts of data, 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. Big data 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]. We have presented all issues related big data and cloud commuting and the proposed work specifies the network servers used to handle big data applications of meta-data.

Key words: *Meta data, Big Data, Cloud computing, next generation networks, wireless networks, SQL.*

1. INTRODUCTION

Cloud computing is a fast-growing technology that has established itself in the next generation of IT industry and business. Cloud computing promises reliable software, hardware and IaaS delivered over the Internet and remote data centers [5]. Cloud services have become a powerful architecture to perform complex large-scale computing tasks and span a range of IT functions from storage and computation to database and application services. The need to store, process, and analyze large amounts of datasets has driven many organizations and individuals to adopt cloud computing [6]. A large number of scientific applications for extensive experiments are currently deployed in the cloud and may continue to increase because of the lack of available computing facilities in local servers, reduced capital costs and increasing volume of data produced and consumed by the experiments [7]. In addition, cloud service providers have begun to integrate frame works for parallel data processing in their services to help user's access cloud resources and deploy their programs [8]. Cloud computing "is a model for allowing ubiquitous, convenient and on-demand network access to a number of configured computing resources (e.g., networks, server, storage, application, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [9].

2. SOFTWARE RESOURCES USED TO HANDLE DATA

SwiftKey is a language technology founded in London in 2008. This language technology aids touch screen typing by providing personalized predictions and corrections. The company collects and analyzes terabytes of data to create language models for many active users. Thus, the company needs a highly scalable, multilayered model system that can keep pace with steadily increasing demand and that has a powerful processing engine for the artificial intelligence technology used in prediction generation.

Mining Twitter in the cloud, Noordhu et al. [11] used cloud computing to analyze large amounts of data on Twitter. The author applied the Page Rank algorithm on the Twitter user base to obtain user rankings. The Amazon cloud infrastructure was used to host all related computations. Computations were conducted in a two-phase process: in the crawling phase, all data were retrieved from Twitter. In the processing phase, the Page Rank algorithm was applied to compute the acquired data.

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 [4]. 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. The storage media currently employed in enterprises are discussed. 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.

Hadoop [12] is an open-source Apache Software Foundation project written in Java that enables the distributed processing of large datasets across clusters of commodity. Hadoop has two primary components, namely, HDFS and Map Reduce programming framework. The most significant feature of Hadoop is that HDFS and Map Reduce are closely related to each other; each are co-deployed such that a single cluster is produced [12]. Therefore, the storage system is not physically separated from the processing system.

Map Reduce [10] is a simplified programming model for processing large numbers of datasets pioneered by Google for data-intensive applications. The Map Reduce model was developed based on GFS [13] and is adopted through open-source Hadoop implementation, which was popularized by Yahoo. Apart from the Map Reduce framework, several other current open-source Apache projects are related to the Hadoop ecosystem, including Hive, Hbase Mahout, Pig, Zookeeper, Spark, and Avro. Twister [14] provides support for efficient and iterative Map Reduce computations. Currently, many alternative solutions are available to deploy MapReduce in cloud environments; these solutions include using cloud MapReduce in runtime, that maximize cloud infrastructure services, using MapReduceBase or setting up one's own MapReduce cluster in cloud instances [15]. Several strategies have been proposed to improve the performance of big data processing. Moreover, effort has been exerted to develop SQL interfaces in the MapReduce framework to assist programmers who prefer to use SQL as a high-level language to express their task while leaving all of the execution optimization details to the backend engine [16].

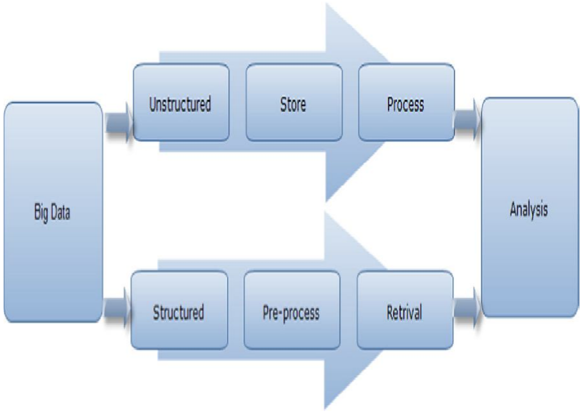

Figure 1 Big Data Analysis

Cloud vendors must ensure that all service level agreements are met. Recently, some controversies have revealed how some security agencies use data generated by individuals for their own benefit without permission. Therefore, policies that cover all user privacy concerns should be developed. Traditionally, the most common technique for privacy and data control is to protect the systems utilized to manage data rather than the data itself; however, such systems have proven to be vulnerable. Utilizing strong cryptography to encapsulate sensitive data in a cloud computing environment and developing a novel algorithm that efficiently allows for key management and secure key exchange are important to manage access to big data, particularly as they exist in the cloud independent of any platform.

3. CONCLUSION

The velocity of data generation and growth is increasing because of the proliferation of mobile devices and other device sensors connected to the Internet. These data provide opportunities that allow businesses across all industries to gain real-time business insights. The use of cloud services to store, process, and analyze data has been available for some time; it has changed the context of information technology and has turned the promises of the on-demand service model into reality. In this study, we presented a review on the rise of big data in cloud computing. We proposed a classification for big data, a conceptual view of big data, and a cloud services model. This model was compared with several representative big data cloud platforms. We discussed the background of Hadoop technology and its core components, namely, MapReduce and HDFS. We presented current MapReduce projects and related software. We also reviewed some of the challenges in big data processing. The review covered volume, scalability, availability, data integrity, data protection, data transformation, data quality/ heterogeneity, privacy and legal/ regulatory issues, data access and governance. Furthermore, the key issues in big data in clouds were highlighted.

REFERENCES

[1]L.Villars, C.W.Olofson, M.Eastwood, Bigdat a: what it is and why you shouldcare,WhitePaper,IDC,2011,MA,USA.

[2]RCumbley, P.Church, Is BigData Creepy/ Comput. LawSecur.Rev.29 (2013)601–609.

[3]Kaisler, F.Armour, J.A.Espinosa, W.Money, BigData: Issues and Challenges Moving Forward, System Sciences(HICSS), 2013,in: Proceedings ofthe46th Hawaii International Conference on, IEEE, 2013,pp.995–1004.

[4]MChen, S.Mao, Y.Liu, Bigdata:asurvey, Mob.Netw.Appl.19(2) (2014)1–39.

[5]M.Armbrust,A.Fox,R.Griffith,A.D.Joseph,R.Katz,A.Konwinski,G.Lee,D.Patterson, A.Rabkin,I.Stoica,M.Zaharia,Aviewofcloud computing, Commun.ACM53(2010)50–58.

[6]L.Huan, Bigdata drives cloud adopt ion in enterprise, IEEE Internet Computing. 17(2013)68–71.

[7]S.Pandey,S.Nepal, Cloud computing and scientific applications — big data, Scalable And Beyond, Futur.Gener.Comput.Syst.29 (2013)1774–1776.

[8]D.Warneke, O.Kao, Nephele: Efficient parallel data processing in the cloud, in: Proceedings of the 2nd work shop on many-task computing on grids and supercomputers, ACM,2009,p.8.

[9]P.Mell,T.Grance, The NIST definition of cloud computing(draft), NIST Spec. Publ.800(2011)7.

[10]J.Dean, S.Ghemawat, MapReduce: simplified data processing on large clusters, Commun. ACM51(2008)107–113.

[11]P.Noordhuis, M.Heijkoop, A.Lazovik, Mining twitter in the cloud: A case study, Cloud Computing (CLOUD), 2010, in: Proceedings of IEEE 3rd International Conferenceon,IEEE, Miami,FL,2010, pp. 107–114.

[12]T.White, Hadoop: The Definitive Guide, O'ReillyMedia, Sebastapol, CA,2009.

[13]S.Ghemawat, H.Gobioff, S.-T.Leung, The Google file system, ACM SIGOPS Oper.Syst.Rev.ACM37(5)(2003)29–43.

[14]J.Ekanayake, H.Li, B.Zhang, T.Gunarathne, S.H.Bae, J.Qiu, G.Fox, Twister: A run time for iterative MapReduce, in: Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, ACM, 2010, pp.810–818.

[15]T.Gunarathne, T.-L.Wu, J.Qiu, G.Fox, MapReduce in the Clouds for Science, IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom), 2010,pp.565–572.

[16]S.Sakr,A.Liu,A.G.Fayoumi,The family of MapReduce and large scale data processing systems, ACM Computing Surveys (CSUR)46 (2013)11.