BIG DATA ANALYTICS ON CLOUD ENVIRONMENT

M.Nandhakumar,
Asst Prof, Dept of BCA,VLB Janakiammal College of Arts and Science, kovaipudur, Coimbatore.

ABSTRACT

The beginning of the digital age has led to a rise in different types of data with every passing day. In fact, it is expected that half of the total data will be on the cloud by 2017. This data is complex and needs to be stored, processed and analyzed for information that can be used by organizations. Cloud computing provides an apt platform for big data analytics in view of the storage and computing requirements of the latter. This makes cloud-based analytics a viable research field. However, several issues need to be addressed and risks need to be mitigated before practical applications of this synergistic model can be popularly used. This paper explores the existing research, challenges, open issues and future research direction for this field of study.

Keywords: Cloud-based Big Data Analytics, Big Data, Big Data Analytics, Big Data Cloud Computing.

INTRODUCTION

With the advent of the digital age, the amount of data being generated, stored and shared has been on the rise. From data warehouses, webpages and blogs to audio/video streams, all of these are sources of massive amounts of data. The result of this proliferation is the generation of massive amounts of pervasive and complex data, which needs to be efficiently created, stored, shared and analyzed to extract useful information.

This data has huge potential, ever-increasing complexity, insecurity and risks, and irrelevance. The benefits and limitations of accessing this data are arguable in view of the fact that this analysis may involve access and analysis of medical records, social media interactions, financial data, government records and genetic sequences. The requirement of an efficient and effective analytics service, applications, programming tools and frameworks has given birth to the concept of Big Data Processing and Analytics.

Big data analytics has found application in several domains and fields. Some of these applications include medical research, solutions for the transportation and logistics sector, global security and prediction and management of issues concerning the socio-economic and environmental sector, to name a few.

Apart from standard applications in business and commerce and society administration, scientific research is one of the most critical applications of big data in the real world [30]. According to O'Driscoll, Daugelaite and Sleator [32], one of the main future applications of big data analytics and cloud computing lies in life sciences. Some of the identified high-impact areas include systems biology, structure and protein function prediction, personalized medicine and metagenomics. Besides this, one of the most relevant applications of big data analytics is to improve the existing business models for efficiency and customer satisfaction.

Big data, by definition, is a term used to describe a variety of data - structured, semi-structured and unstructured, which makes it a complex data infrastructure [18]. The complexity of this infrastructure requires powerful management and technological solutions. One of the commonly used models for explaining big data is the multi-V model. Figure 1 illustrates the multi-V model.

Some of the Vs used to characterize big data include variety, volume, velocity, veracity and value [4]. The different types of data available on a dataset determine variety while the rate at which data is produced determines velocity. Predictably, the size of data is called volume. The two additional characteristics, veracity and value, indicate data reliability and worth with respect to big data exploitation, respectively.
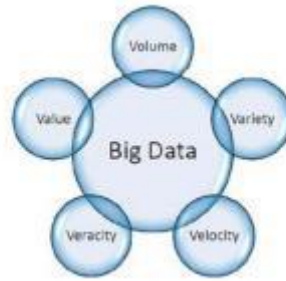
Fig. 1 – Big Data Characteristics [34]

In calculation, Wu, Zhu, Wu and Ding [26] gave another characterization called the HACE theorem. According to this theorem, big data has two main characteristics. Firstly, it has a large volume of data that comes from different and heterogeneous sources, which is complex in nature. Secondly, the data is decentralized and distributed in nature.

Data is the central element of communication and collaboration in Internet and all the applications that are built on this platform. The immense popularity of data intensive applications like Facebook, LinkedIn, Twitter, Amazon, eBay and Google+ contributes to increasing requirement of storage and processing of data in the cloud environment. Schouten [21] uses Gartner's estimation to predict that by the year 2016, half of the data will be on the cloud.

Furthermore, the data mining algorithms used for Big Data analytics possess high computing requirements. Therefore, they require high performance processors to do the job. The cloud provides a good platform for big data storage, processing and analysis, addressing two of the main requirements of big data analytics, high storage and high performance computing.

The cloud computing environment offers development, installation and implementation of software and data applications 'as a service'. Three multi-layered infrastructures namely, platform as a service (PaaS), software as a service (SaaS), and infrastructure as a service (IaaS), exist. Infrastructure-as-a-service is a model that provides computing and storage resources as a service. On the other hand, in case of PaaS and SaaS, the cloud services provide software platform or software itself as a service to its clients.

The cost of storage has considerably reduced with the advent of cloud-based solutions. In addition, the 'pay-as-you-go' model and the concept of commodity hardware allow effective and timely processing of large data, giving rise to the concept of 'big data as a service'. An example of one such platform is Google BigQuery, which provides real-time insights from big data in the cloud environment [12]. Shakil, Sethi, and Alam[37] demonstrates the application of cloud for management of Big Data in educational institutions which special focus on University-level data.

Though, there have not been many practical applications of big data analytics that make use of the cloud. This has led to an increasing shift of research focus towards cloud-based big data analytics. An issue that is evident in this arrangement is information security and data privacy. As part of the cloud services, trust in data is also defined as a service. There shall be a considerable decrease in trust in view of the fact that the chances of security breaches and privacy violation will significantly rise upon implementation of big data strategies in the cloud. In addition, another important issue of ownership and control will also exist.

Conversely, the potential of cloud-based big data analytics has compelled researchers to look into the existing issues to explore solutions. This paper discusses the different facets and aspects of data mining techniques/strategies adoption in the cloud environment for big data analytics. Moreover, it also looks into the existing research, identified challenges and future research directions in cloud-based big data analytics.

BACKGROUND

Established data management tools and data processing or data mining techniques cannot be used for Big Data Analytics for the large volume and complexity of the datasets that it includes. Conventional business

intelligence applications make use of methods, which are based on traditional analytics methods and techniques and make use of OLAP, BPM, Mining and database systems like RDBMS.

It was in the 1980s that artificial intelligence-based algorithms were developed for data mining. Wu, Kumar, Quinlan, Ghosh, Yang, Motoda, McLachlan, Ng, Liu, Yu, Zhou, Steinbach, Hand and Steinberg [25] mention the ten most influential data mining algorithms k-means, C4.5, Apriori, Expectation Maximization (EM), PageRank, SVM (support vector machine), AdaBoost, CART, a ve Bayes and kNN (k-nearest neighbors). Most of these algorithms have been used commercially as well. Alam and Shakil [38] propose architecture for management of data through cloud techniques.

One of the most popular models used for data processing on cluster of computers is MapReduce. Jackson, Vijayakumar, Quadir and Bharathi [33] provide a survey on the programming models that support big data analytics. It identifies MapReduce/Hadoop as the most productive model for Big Data Analytics yet mentions that languages and extensions like HiveQL, Latin and Pig have overpowering benefits for this use.

Hadoop is simply an open-source implementation of the MapReduce framework, which was originally created as a distributed file system. According to Neaga and Hao [19], the evolution of Hadoop as a complete ecosystem or infrastructure that works alongside MapReduce components and includes a range of software systems like Hive and Pig languages, a coordination service called Zookeeper and a distributed table store called HBase.

For cloud-based big data analytics, several frameworks like Google MapReduce, Spark, Haloop, Twister, Hadoop Reduce and Hadoop++ are available. Figure 2 gives a pictorial representation of the use of cloud computing in big data analytics. These frameworks are used for storing and processing of data. In order to store this data, which may be of any structure,databases like HBase, BigTable and HadoopDB may be used. When it comes to data processing, the Pig and Hive technologies come into the picture.

Some of the recent research breakthroughs and milestones in cloud-based big data analytics are discussed here. Lee [16] elaborates on the advantages and limitations of MapReduce in parallel data analytics. A Hadoop-based data analytics system, created by Starfish [13], improves the performance of the clusters throughout the cycle of data analytics. additionally, the users are not required to understand the configuration details.
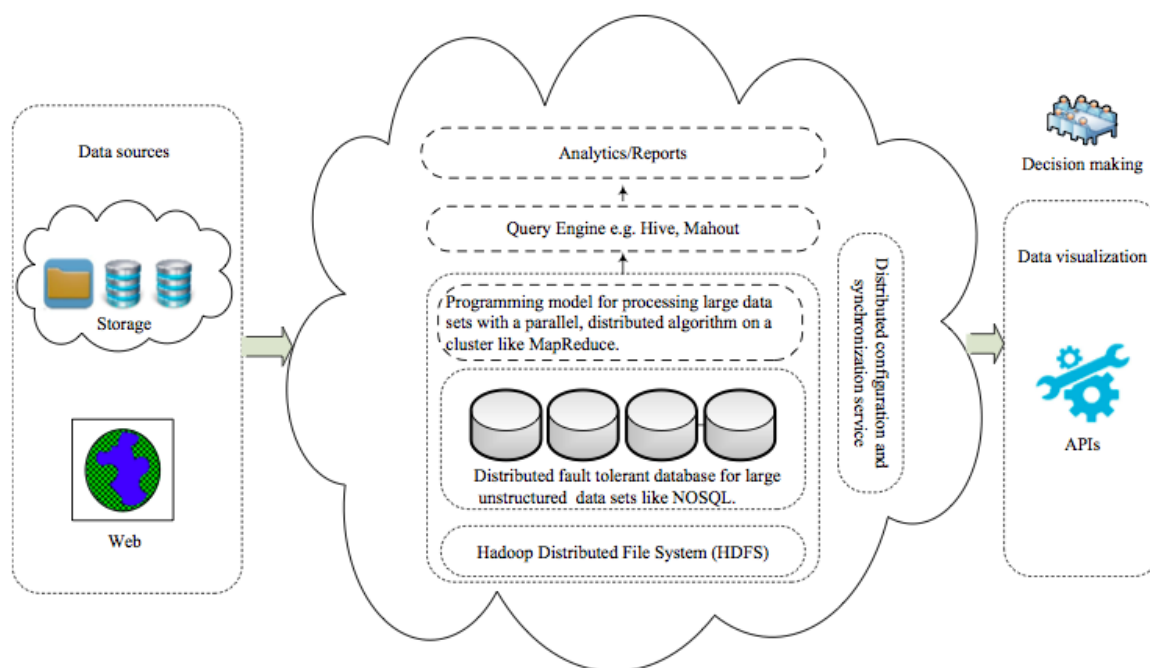

Fig. 2 – Use of Cloud Computing in Big Data [27])

In recent times, lack of interactivity has been identified as a major issue and several efforts have been made in this area. Borthakur, Gray, Sarma, Muthukkaruppan, Spiegelberg, Kuang, Ranganathan, Molkov,

Menon, Rash, Schmidt and Aiyer [5] optimize the HBase and HDFS implementation for better responsiveness. Strambei [23] evaluates the viability of OLAP Web Services for cloud-based architectures, with the specific objective to allow open and wide access to web analytical technologies.

Research efforts have been made to create a big data management framework for the cloud. Khan, Naqvi, Alam and Rizvi [35] propose a data model and provides a schema for big data in the cloud and attempts to ease the process of querying data for the user. Moreover, an important subject of research has been performance and speed of operation. Ortiz, Oneto and Anguita [28] explore the use of a proposed integrated Hadoop and MPI/OpenMP system and how the same can improve speed and performance.

In view of the fact that data needs to be transferred between data centers that are usually located distances apart, power consumption becomes a crucial parameter when it comes to analyzing efficiency of the system. A network-based routing algorithm called GreeDi can be used for finding the most energy efficient path to the cloud data center during big data processing and storage [29].

There are several practical simulation-enabled analytics systems. One such system is given by Li, Calheiros, Lu, Wang, Palit, Zheng and Buyya [17], which is a Direct Acrylic Graph (DAG) form analytical application used for modeling and predicting the outbreak of Dengue in Singapore.

Online risk analytics and the need for an infrastructure that can provide users the programming resources and infrastructure for carrying out the same have also appeared in the form of Aneka [6] and CloudComet [15]. Chen [7] investigates the concept of CAAAS or Continuous Analytics As A Service, which is used for predicting the behavior of a service or a user.

The last topic under Big Data Analysis that has caught the attention of the research community is Real-time Big Data Analysis. Many commercial cloud service providers are providing solutions for real-time analysis. AWS based-solutions for real-time stream processing is AWS Kinesis [2]. Many frameworks and software systems have also been introduced for this purpose, some of which are Apache S4 [3], IBM InfoSphere Streams [14] and Storm [22].

CHALLENGES AND ISSUES

In turn to move beyond the existing techniques and strategies used for machine learning and data analytics, some challenges need to be overcome. NESSI [20] identifies the following requirements as critical.

1. In order to select an adequate method or design, a solid scientific foundation needs to be developed.

2. New efficient and scalable algorithms need to be developed.

3. For proper implementation of devised solutions, appropriate development skills and technological platforms must be identified and developed.

4. Lastly, the business value of the solutions must be explored just as much as the data structure and its usability.

In view of cloud-based big data analytics, additional challenges like adoption and implementation of effective big data solutions using cloud architecture and mitigating the security and privacy risks also exist. One of the biggest concerns while using big data analytics and cloud computing in an integrated model is security. This is perhaps the reason why this aspect of cloud-based big data analytics and its practical usage and implementation has attracted immense attention.

Liu, C., Yang, C., Zhang, X. and Chen, J. [31] provides a summary, analysis and comparison of authenticator-based data integrity verification techniques on cloud and Internet-of-things data. This paper suggests that any future developments in this area needs to look at three main aspects namely, efficiency, security and scalability/elasticity.

A G-Hadoop based security framework is proposed by Zhao, Wang, Tao, Chen, Sun, Ranjan, Kołodziej, Streit and Georgakopoulos [36], which makes use of solutions like SSL and public key cryptography for ensuring security of big data resident on distributed cloud data centers. In addition to several security mechanisms, this framework also aims to simplify the processes of submitting job and authenticating users. Talia [24] suggests further research and development in the following areas:

1. Programming abstracts or scalable high-level models and tools.

2. Solutions for data and computing interoperability issues.

3. Integration of different big data analytics frameworks

4. Techniques for mining provenance data

FUTURE RESEARCH DIRECTIONS

Several open source data mining techniques, resources and tools exist. Some of these include R, Gate, Rapid-Miner and Weka, in addition to many others. Cloud-based big data analytics solutions must provide a provision for the availability of these affordable data analytics on the cloud so that cost-effective and efficient services can be provided. The fundamental reason why cloud-based analytics are such a big thing is their easy accessibility, cost-effectiveness and ease of setting up and testing. In view of this, some of the main research directions identified by Neaga and Hao [19] include:

1. Evolution of analytics and information management with respect to cloud-based analytics.

2. Adaptation and evolution of techniques and strategies to improve efficiency and mitigate risks.

3. Formulate strategies and techniques to deal with the privacy and security concerns.

4. Analysis and adaptation of legal and ethical practices to suit the changing viewpoint, impact and effects of technological advances in this regard.

With this said, the research directions are not limited to the above-mentioned points. The main goal is to transform the cloud from being a data management and infrastructure platform to a scalable data analytics platform.

CONCLUSION

This is an age of big data and the emergence of this field of study has attracted the attention of many practitioners and researchers. Considering the rate at which data is being created in the digital world, big data analytics and analysis have become all the more relevant. Moreover, most of this data is already on the cloud. Therefore, shifting big data analytics to the cloud framework is a viable option.

Moreover, the cloud infrastructure suffices the storage and computing requirements of data analytics algorithms. On the other hand, open issues like security, privacy and the lack of ownership and control exist. Research studies in the area of cloud-based big data analytics aim to create an effective and efficient system that addresses the identified risks and concerns.

REFERENCES

[1] Agarwal, D., Das, S. and Abbadi, A. (2011). Big Data and Cloud Computing: Current State and Future Opportunities.ACM978-1-4503-0528-0/11/0003.Retrievedfrom: http://www.edbt.org/Proceedings/2011-Uppsala/papers/edbt/a50-agrawal.pdf

[2]AmazonKinesis.(n.d.).DeveloperResources.Retrievedfrom: http://aws.amazon.com/kinesis/developer-resources

[3]ApacheS4.(n.d.).Distributed Stream Computing Platform. Retrievedfrom: http://incubator.apache.org/s4

[4] Assuncao, M. D., Calheiros, R. N., Bianchi, S. and Netto, M. A. S. (2015). Big Data Computing and Clouds: Trends and Future Directions. J. Parallel Distrib. Computing, 79-80 (2015) 3-15. Retrieved from: http://www.buyya.com/papers/BDC-Trends-JPDC.pdf

[5] Borthakur, D., Gray, J., Sarma, J. S., Muthukkaruppan, K., Spiegelberg, N., Kuang, H., Ranganathan, K., Molkov, D., Menon, A., Rash, S., Schmidt, R. and Aiyer, A. (2011). Apache Hadoop Goes Real-time at Facebook, in: Proceedings of the ACM SIGMODInternational Conference on Management of Data (SIGMOD 2011), ACM, New York, USA, 2011, pp. 1071–1080. Retrieved from: http://cloud.pubs.dbs.uni-leipzig.de/sites/cloud.pubs.dbs.uni-leipzig.de/files/RealtimeHadoopSigmod2011.pdf

[6] Calheiros, R. N., Vecchiola, C., Karunamoorthy, D. and Buyya, R. (2012) The Aneka platform and QoS-driven resource provisioning for elastic applications on hybrid Clouds, Future Gener. Comput. Syst. 28 (6) (2012) 861–870. Retrieved from: http://www.buyya.com/papers/Aneka-QoS-ResourceProvisioning-FGCS.pdf

[7] Chen, Q., Hsu, M. and Zeller, H. (2011). Experience in Continuous analytics as a Service (CaaaS), in: Proceedings of the 14th International Conference on Extending Database Technology, ACM, New York, USA, 2011, pp. 509–514. Retrieved from: http://www.edbt.org/Proceedings/2011-Uppsala/papers/edbt/a46-chen.pdf

[8] Chen, H., Chiang, R. H. L. and Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly. Special Issue: Business Intelligence Research. Retrieved from:http://ai.arizona.edu/mis510/other/MISQ%20BI%20Special%20Issue%20Introduction%20Chen-Chiang-Storey%20December%202012.pdf

[9]Dean, J. and Ghemawat, S. (2004). OSDI 2004. Retrieved from: http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf

[10] Demirkan, H. and Delen, D. (2013). Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. Decision Support Systems 55 (2013) 412421.Retrievedfrom:http://www.crisismanagement.com.cn/templates/blue/down_list/llzt_dsj/Leveraging%20the%20capabilities%20of%20serviceoriented%20decision%20support%20systems%20Putting%20analytics%20and%20big%20data%20in%20cloud.pdf.

[11] GigaSpaces. (2012). Big Data Survey. Retrieved from: http://www.gigaspaces.com/sites/default/files/product/BigDataSurvey_Report.pdf

[12] Google Cloud Platform. (n.d.). Big Query. Retrieved from: https://cloud.google.com/bigquery

[13] Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F.B. and Babu, S. (2011). Starfish: A Self-tuning System for Big Data Analytics, in: Proceedings of the 5th BiennialConference on Innovative Data Systems Research (CIDR 2011), 2011, pp. 261–272. Retrieved from: https://www.cs.duke.edu/starfish/files/cidr11_starfish.pdf

[14] IBM InfoSphere Streams. (n.d.). InfoSphere Streams. Retrieved from: http://www.ibm.com/software/products/en/infosphere-streams.

[15] Kim, H., Abdelbaky, M. and Parashar, M. (2009). CometPortal: A Portal for Online Risk Analytics Using CometCloud. 17th International Conference on Computer Theory and Applications (ICCTA2009). Retrieved from: http://nsfcac.rutgers.edu/CometCloud/sites/nsfcac.rutgers.edu.CometCloud/files/pub/ICCTA2009.pdf

[16] Lee, K. H., Lee, Y. J., Choi, H., Chung, Y.D. and Moon, B. (2011). Parallel Data Processing with MapReduce: A Survey, SIGMOD Record 40 (4) (2011) 11–20. Retrieved from: http://www.cs.arizona.edu/~bkmoon/papers/sigmodrec11.pdf

[17] Li, X., Calheiros, R. N., Lu, S., Wang, L., Palit, H., Zheng, Q. and Buyya, R. (2012). Design and Development of an Adaptive Workflow-Enabled Spatial-Temporal Analytics Framework, in: Proceedings of the IEEE 18th International Conference on Parallel and Distributed Systems (ICPADS 2012), IEEE Computer Society, Singapore, 2012, pp. 862–867. Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.465.5453&rep=rep1&type=pdf.

[18] Manekar, A. and Pradeepini, G. (2015). A Review on Cloud-based Big Data Analytics. ICSES Journal on Computer Networks and Communication (IJCNC), May 2015, Vol. 1, No, 1. Retrieved from: http://www.i-cses.com/ijcnc/Archive/V1N1/IJCNC-V1N1-P0001.pdf

[19] Neaga, I. and Hao, Y. (2014). A Holistic Analysis of Cloud Based Big Data Mining. International Journal of Knowledge, Innovation and Entrepreneurship. Volume 2 No. 2, 2014, pp. 56—64. Retrieved from: http://ijkie.org/IJKIE_December2014_IRINA&HAO.pdf

[20] NESSI. (2012). Big Data: A New World of Opportunities. Retrieved from: http://www.nessi-europe.com/Files/Private/NESSI_WhitePaper_BigData.pdf

[21] Schouten, E. (2012). Big Data As A Service. Retrieved from: http://edwinschouten.nl/2012/09/19/bigdata-as-a-service

[22] Storm. (n.d.). Apache Storm: Distributed and fault-tolerant real-time computation. Retrieved from: http://storm.incubator.apache.org.

[23] Strambei, C. (2012). OLAP Services on Cloud Architecture. IBIMA Publishing. Journal of Software and Systems Development. Vol. 2012 (2012). DOI: 10.5171/2012.840273. Retrieved from: http://www.ibimapublishing.com/journals/JSSD/2012/840273/840273.pdf

[24] Talia, D. (2013). Clouds for Scalable Big Data Analytics. Published by IEEE Computer Society. Retrieved from: http://scholar.google.co.in/scholar_url?url=http://xa.yimg.com/kq/groups/16253916/1476905727/name/06515548.pdf&hl=en&sa=X&scisig=AAGBfm12aY-Nbu37oZYRuEqeqsdslzKfBQ&nossl=1&oi=scholarr&ved=0CCYQgAMoADAAahUKEwi3k4Hymv7GAhUHUKYKHdToBCM

[25] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A, Liu, B., Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D. J. and Steinberg, D. (2008). Top 10 algorithms in Data Mining. Knowl Inf Syst (2008) 14:1–37. DOI: 10.1007/s10115-007-0114-2. Retrieved from: http://www.cs.umd.edu/~samir/498/10Algorithms-08.pdf

[26] Wu, X., Zhu, X., Wu, G. and Ding, W. (2013). Data Mining with Big Data. Retrieved from: http://www.cs.umb.edu/~ding/papers/TKDE2013.pdf

[27] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A. and Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. Information Systems 47 (2015) 98-115.

[28] Ortiz, J. L. R., Oneto, L. and Anguita, D. (2015). Big Data Analytics in the Cloud: Spark on Hadoop vs MPI/OpenMP on Beowulf. P2015 INNS Conference on Big Data. Published in Procedia Computer Science. Volume 53, 2015, pp. 121-130.

[29] Baker, T., Al-Dawsari, B., Tawfik, H., Reid, D. and Nyogo, Y. (2015). GreeDi: An energy efficient routing algorithm for big data on cloud. Ad Hoc Networks 000 (2015) 1-14.

[30] Chen, C. L. P. and Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences 275 (2014) 314-347.

[31] Liu, C., Yang, C., Zhang, X. and Chen, J. (2015). External Integrity Verification for Outsourced Big Data in cloud and IoT: A Big Picture. Future Generation Computer System 49 (2015), pp. 58-67.

[32] O'Driscoll, A., Daugelaite, J. and Sleator, R. D. (2013). 'Big Data', Hadoop and Cloud Computing in Genomics. Journal of Biomedical Informatics. Volume 46, Issue 5, October 2013, pp. 774-781.

[33] Jackson, J. C., Vijayakumar, V., Quadir, M. A. and Bharathi, C. (2015). Survey on Programming Models and Environments for Cluster, Cloud and Grid Computing that defends Big Data. 2nd International Symposium on Big Data and Cloud Computing (ISBCC '15). Procedia Computer Science 50 (2015) 517-523.

[34] Elragal, A. (2014). ERP and Big Data: The Inept Couple. Procedia Technology, 16, 242-249.

[35] Khan, I., Naqvi, S.K. Alam, M. Rizvi, S.N.A. (2015). Data model for Big Data in cloud environment. Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference. pp. 582 – 585.

[36] Zhao, J., Wang, L., Tao, J., Chen, J., Sun, W., Ranjan, R., Kołodziej, J., Streit, A. and Georgakopoulos, D. (2014). A security framework in G-Hadoop for big data computing across distributed Cloud data centers. Journal of Computer and System Sciences 80 (2014), 994-1007.

[37] Shakil, K.A.; Sethi, S.; Alam, M., (2015). An effective framework for managing university data using a cloud based environment, Computing for Sustainable Global Development (INDIACom), 2nd International Conference on , vol., no., pp.1262,1266, 11-13.

[38] Alam, M., & Shakil, K. A. (2013). Cloud Database Management System Architecture. UACEE International Journal of Computer Science and its Applications, 3(1), 27-31.