International Journal of Management, IT & Engineering
Vol. 8 Issue 6, June 2018,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

THE SYSTEMATIC COMPUTATIONAL ANALYSIS OF CLOUD-BASED BIG DATA ANALYTICS – A STUDY OF PRESENT AND FUTURE ORIENTATION

Dr. Pushpendra KumarVerma,

Abstract

The use of data has been growing rapidly in the era filled with this technology. Starting from the many conveniences that can be enjoyed every day to utilize it for business analysis. The development of these technologies greatly affects the amount of usage of our data today, for example; formerly we only use the diskette that is only able to load about 5MB, because in earlier times data for it can already be used for its needs at that time. But look now, the data of that size is probably only enough for 1 music file, therefore the development of technology will be directly proportional to the amount of data usage. The number of transactions that occur to date is many times over the past, this is due to the easy spread of data, namely through the internet. With the internet everyone can connect and give each other data, without apart distance and time. This development culminate in the collected data is too large to be what we call the big data. Initially this big data can not be processed because it is too big, but now it has been overcome by the emergence of new ways to process big data. Cloud computing provides a platform for big data analysis considering the storage and computing needs of the latter. This makes cloud-based analysis a searchable field of research. However, many problems need to be addressed and the risks mitigated before the practical applications of this synergistic model can be commonly used. This document explores existing research, challenges, open issues and the direction of future research for this field of study.

Keywords:

Cloud-based Big Data Analytics,
Big Data,
Big Data Analytics,
Big Data CloudComputing

Assistant Professor, Department of CS, Swami VivekanandSubharti University, Meerut, U.P., INDIA
Correspondence: dr.pkverma81@gmail.com

I. 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:

Currently, *Big Data is* often a hot conversation because most companies have used this technology as a "solution" in conducting analysis to improve the company's performance in the eyes of *stakeholders* . *Head of Research* from Non-Fiction company, Andres Cristian said that *Big Data*is a huge collection of data that includes various things, ranging from profiles, characteristics, to behavior. Therefore, large data can be processed and studied, to understand patterns of behavior and trends.

One of the effects of rapid technological developments is speed in decision making. But how to decide a decision in a short time but the result is right? Most companies think *Big Data* is the solution. With high data accuracy but in a fast time duration, the company will be able to generate greater investment value.

Furthermore, Andres exemplifies the case of *online* shopping sites that successfully use *big data* to know the characteristics of consumers. He said, with the *big data* that has been processed, *online* shopping sites are able to find out what items are most sought by visitors of various characteristics.

For example, during Lebaran, with the *big data* being processed, an *online*shopping site can already predict which products will be sought after-such as mukena, prayer mat, and date palm. When the long holiday moments, consumers will find more tickets for the streets, until the holiday equipment-all this information obtained thanks to data that has been processed in the previous period so the company can predict appropriately.

In the future, Andres predicts that there will be even more demanding data usage needs, even in a much narrower timeframe. Thanks to technology as it is now, the data needs for decision making will be higher and faster. No wonder, if at this time, start many *big data* service providers that will accommodate a lot *insight* about the characteristics of a population.

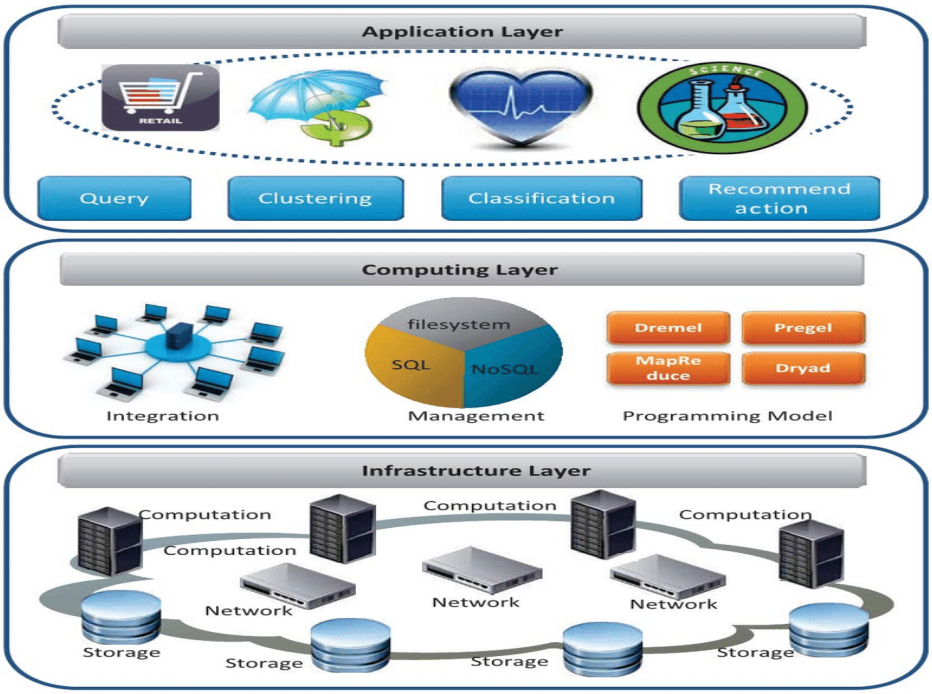

Fig. 1: Layered Architecture of Big Data System

Challenges include acquisition, duration, storage, search, sharing, transfer, analysis , and data visualization . The growing trend of data sets is due to the increase of information from large interrelated sets, compared to other small sets of the same total data. New correlations can be found in the analysis of the data set to "look at business trends, determine the quality of research , prevent disease, fight crime , and know the condition of road traffic in real time. 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. In addition, 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.

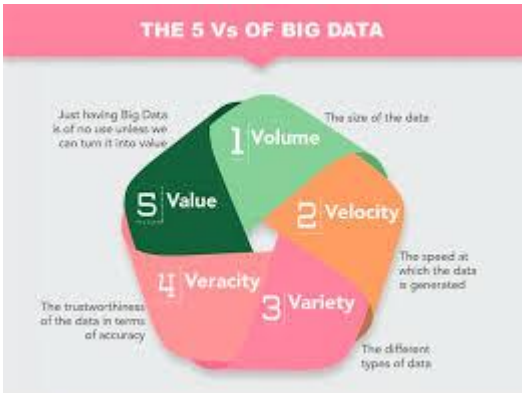

Fig. 2– Big Data Characteristics

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. Schoutenuses Gartner's estimation to predict that by the year 2016, half of the data will be on the cloud.

Moreover, 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.

However, 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.

However, 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.

II. BACKGROUND

Traditional data management tools and data processing or data mining techniques can not be used for Big Data analysis due to the high volume and complexity of the data sets it includes. Conventional business intelligence applications make use of methods, which are based on traditional analysis techniques and methods and make use of OLAP, BPM, extraction systems and databases, such as RDBMS.

It was in the 80s when algorithms based on artificial intelligence for data mining were developed. Wu, Kumar, Quinlan, Ghosh, Yang, Motoda, McLachlan, Ng, Liu, Yu Zhou, Steinbach, Mano and Steinberg [25] mention the ten most influential k-means mining algorithms, C4.5, Apriori, expectation maximization (EM), PageRank, SVM (vector machine support), AdaBoost, CART, Bayes and kNN (k-nearest neighbors). Most of these algorithms have also been used commercially. Alam and Shakil [38] propose architecture for data management through cloud techniques.

MapReduce is one of the most popular models used for data processing in computer clusters. Jackson, Vijayakumar, Quadir and Bharathi [33] provide an investigation into the programming models that support big data analysis. Identifies MapReduce / Hadoop as the most productive model for Big Data Analytics, but mentions that languages and extensions like HiveQL, Pig Latin and have benefits overflowing 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 in conjunction with the components of MapReduce and includes a range of software systems such as Hive and pork tongues, coordination service called Zookeeper and a store tables distributed calls HBase.

For analysis of large, cloud-based data, multiple frames such as Google MapReduce, Spark, Haloop, Twister, Reduce and Hadoop Hadoop ++ are available. Figure 2 provides a graphical representation of the use of cloud computing in big data analysis. These frames are used to store and process data. To store this data, which can be of any structure, databases such as HBase, BigTable and HadoopDB can be used. When it comes to data processing, Pig and Hive technologies come into play.

Some of the recent advances in research and milestones in cloud-based big data analytics are discussed here. Lee [16] explores the advantages and limitations of MapReduce in parallel data analysis. A Hadoop-based data analysis system, created by Starfish [13], improves cluster performance during the data analysis cycle. Furthermore, users are not required to understand configuration details.

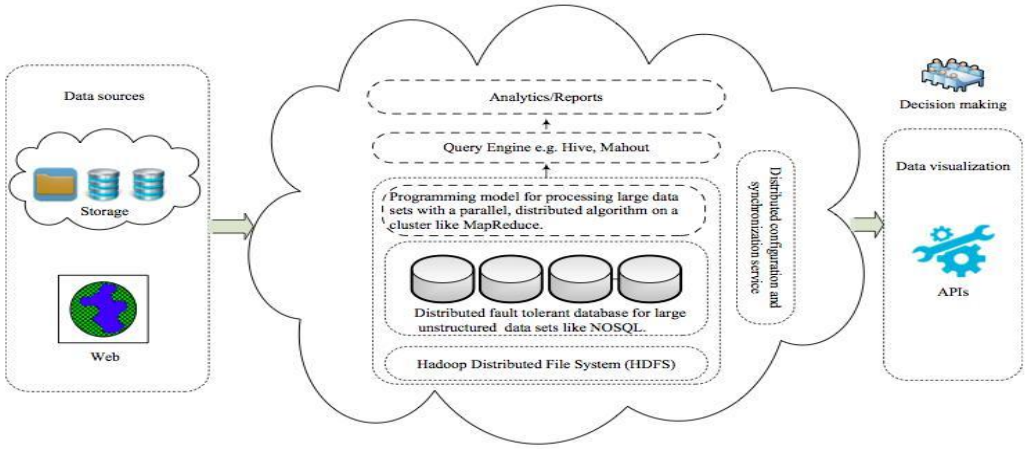

Fig. 3 – Use of Cloud Computing in Big Data

In recent times, the lack of interactivity has been identified as an important problem 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 performance of HBase and HDFS for better response. Strambei [23] assesses the feasibility of OLAP web services for cloud-based architectures, with the specific goal of allowing open and broad access to web analytics technologies.

Research efforts have been made to create a comprehensive data management framework for the cloud. Khan, Naqvi, Alam Rizvi [35] propose a data model and provide a framework for large data in the cloud and try to facilitate the process of consultation data for the user. Furthermore, an important research topic was represented by performance and operational speed. Ortiz, Oneto and Anguita [28] explore the use of an integrated Hadoop and MPI / Open MP system and how it can improve speed and performance.

In view of the fact that data must be transferred between data centers that are generally located at separate distances, energy consumption becomes a crucial parameter when it comes to analyzing the efficiency of the system. A network-based routing algorithm called GreeDi can be used to find the most efficient path force at the cloud data center when processing and storing large data [29].

There are several practical analysis systems enabled for simulation. A system of this type is given by Li, Calheiros, Lu, Wang, Palit, Zheng and Buyya [17], which is a direct acrylic graphical analytical application (DAG) used to model and predict the dengue outbreak in Singapore.

Online risk analysis and the need for an infrastructure to provide users with programming resources and infrastructure to implement it 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, used to predict the behavior of a service or a user.

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

Problem or Challenges associated with Big Data Processing

The challenges in Big Data are usually the real hurdles which require immediate attention. Any implementation without handling these challenges may lead to the failure of the technology implementation and some unpleasant results

A. Size: The first thing anyone thinks of with Big Data is its size. The word "big" is there in the very name. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore's law, to provide us with the resources needed to cope with increasing volumes of data. But, there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are static.

B. Privacy and Security: It is the most important challenges with Big data which is sensitive.

• The personal information (e.g. in database of social networking website) of a person when combined with external large data sets, leads to the inference of new facts about that person and it's possible that these kinds of facts may be secretive and the person might not want the data owner to know or any person to know about them.

• Information regarding the people is collected and used in order to add value to the business of the organization. Another important consequence arising would be Social sites where a person would be taking advantages ofthe Big data predictive analysis and on the other hand underprivileged will be easily identified and treated worse.

• Big Data increase the chances of certain tagged people tosuffer from adverse consequences without the ability to fight back or even having knowledge that they are being discriminated.

C. Data Access and Sharing of Information: Due to huge amount of data, data management and governance process is bit complex adding the necessity to make data open and make it available to government agencies in standardized manner with standardized APIs, metadata and formats. Expecting sharing of data between companies is awkward because of the need to get an edge in business. Sharing data about their clients and operations threatens the culture of secrecy and competitiveness

D. Analytical Challenges: The main analytical challenging questions are as:

• What if data volume gets so large and varied and it is not known how to deal with it?
• Does all data need to be stored?
• Does all data need to be analyzed?
• How to find out which data points are really important?
• How can the data be used to best advantage?

Big data brings along with it some huge analytical challenges. The type of analysis to be done on this huge amount of data which can be unstructured, semi structured or structured.

E. Human Resources and Manpower: Since Big data is an emerging technology so it needs to attract organizations and youth with diverse new skill sets. These skills should not be limited to technical ones but also should extend to research, analytical, interpretive andcreative ones. These skills need to be developed in individuals hence requires training programs to be held by the organizations. Moreover the Universities need to introduce curriculum on Big data to produce skilled employees in this expertise.

F. Technical Challenges

1. Fault Tolerance

With the incoming of new technologies like Cloud computing andBig data it is always intended that whenever the failure occurs thedamage done should be acceptable.Fault-tolerant computing is extremely hard, involving intricatealgorithms. Thus the main task is to reduce the probability offailure to an "acceptable" level.Two methods which seem to increase the fault tolerance in Bigdata are as:

a) First is to divide the whole computation being done into tasks and assign these tasks to different nodes for computation.
b) Second is, one node is assigned the work of observing that these nodes are working properly. If something happens that particular task is restarted. But sometimes it's quite possible that that the whole computation can't be divided into such independent tasks. There could be some tasks which might be recursive in nature and the output of the previous computation of task is the input to the next computation. Thus restarting the whole computation becomes cumbersome process. This can be avoided

by applying Checkpoints which keeps the state of the system at certain intervals of the time. In case of any failure, the computation can restart from last checkpoint maintained.

2. Scalability: The scalability issue of Big data has lead towards cloud computing, which now aggregates multiple disparate workloads with varying performance goals into very large clusters. This requires a high level of sharing of resources which is expensive and also brings with it various challenges like how to run and execute various jobs so that we can meet the goal of each workload cost effectively. It also requires dealing with the system failures in an efficient manner which occurs more frequently if operating on large clusters. These factors combined put the concern on how to express the programs, even complex machine learning tasks. There has been a huge shift in the technologies being used. Hard Disk Drives (HDD) are being replaced by the solid state Drives and Phase Change technology which are not having the same performance between sequential and random data transfer. Thus, what kinds of storage devices are to be used; is again a big question for data storage.

3. Quality of Data: Big data basically focuses on quality data storage rather than having very large irrelevant data so that better results and conclusions can be drawn. This further leads to various questions like how it can be ensured that which data is relevant, how much data would be enough for decision making and whether the stored data is accurate or not to draw conclusions from it etc.

4. Heterogeneous: Data Unstructured data represents almost every kind of data being produced like social media interactions, to recorded meetings, to handling of PDF documents, fax transfers, to emails and more. Working with unstructured data is a cumbersome problem and of course costly too. Converting all this unstructured data into structured one is also not feasible. Structured data is always organized into highly mechanized and manageable way.

It shows well integration with database but unstructured data is completely raw and unorganized.

III. CHALLENGES AND ISSUES

In the *Big Data* utilization effort there can be many obstacles and challenges, some of which relate to data which involves acquisition, sharing and data privacy, as well as in data analysis and processing

- *Privacy*

Privacy is the most sensitive issue, with conceptual, legal, and technological, Privacy can be understood in a broad sense as a business enterprise to protecting their competitiveness and their customers .The data used / stored as big data

- *Access* and *sharing*

Access to data, both old and new data can be an obstacle in getting data for big data, especially on old data where data are stored in different forms and different or physical form, access to new data also requires more business as it requires licenses and licenses to access non-public data legally.

- *Analysis*

Working with new data sources brings a number of analytical challenges. the relevance and severity of the challenge will vary depending on the type of analysis being conducted, and on the type of decision that will eventually be informed by the data.

Depending on the type of data there are 3 categories in the data analysis

- *Determination of the correct picture*

This is usually found in *unstructured user-generated text-based data*. The data obtained are not necessarily true because of incorrect data or sources.

- *Interpreting Data*

Mistakes such as sampling selection bias are commonly found in existing data that cannot be used to represent all of the existing populations, and apothecia, see patterns even though they do not really exist due to large amounts of data, and errors in enterprises relationship in the data.

- Defining and detecting anomalies

the challenge of sensitivity to the specificity of monitoring systems. Sensitivity refers to the ability of the monitoring system to detect all cases already set to detect while specificity refers to its ability to detect only relevant cases. Failure to achieve the last result "Type I error decisions", also known as "false positives"; failure to achieve former "Type II error", or "false negative." Both unintended errors when trying to detect malfunctions or anomalies however are defined, for various reasons. False positives undermine the credibility of the system while false negatives are cast doubt on its relevance. But whether false negatives are more or less problematic than false positives depends on what is being monitored, and why it is being monitored.

IV. 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:

- Evolution of analytics and information management with respect to cloud-based analytics.
- Adaptation and evolution of techniques and strategies to improve efficiency and mitigate risks.
- Formulate strategies and techniques to deal with the privacy and security concerns.
- 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.

V. Good Practices for Big Data

• Creating dimensions of all the data being store is a good practice for Big data analytics. It needs to be divided intodimensions and facts.
• All the dimensions should have durable surrogate keysearning that these keys can't be changed by any business ruleand are assigned in sequence or generated by some hashingalgorithm ensuring uniqueness.
• Expect to integrate structured and unstructured data as allkind of data is a part of Big data which needs to be analyzedtogether.
• Generality of the technology is needed to deal with differentformats of data. Building technology around key value pairswork.
• Analyzing data sets including identifying information aboutindividuals or organizations privacy is an issue whoseimportance particularly to consumers is growing as the valueof big data becomes more apparent.
• Data quality needs to be better. Different tasks like filtering, cleansing, pruning, and conforming, matching, joining, anddiagnosing should be applied at the earliest touch pointspossible.
• There should be certain limits on the scalability of the datastored.
• Business leaders and IT leaders should work together to yieldmore business value from the data. Collecting, storing andanalyzing data comes at a cost. Business leaders will go for itbut IT leaders have to look for many things like technologicallimitations, staff restrictions etc. The decisions taken shouldbe revised to ensure that the organization is considering theright data to produce insights at any given point of time.
• Investment in data quality and metadata is also important asit reduces the processing time.

VI. 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. ACM 978-1-4503-0528-0/11/0003. Retrieved from: http://www.edbt.org/Proceedings/2011-Uppsala/papers/edbt/a50-agrawal.pdf
[2] Amazon Kinesis. (n.d.). Developer Resources. Retrieved from: http://aws.amazon.com/kinesis/developer-resources/
[3] Apache S4. (n.d.). Distributed Stream Computing Platform. Retrieved from: 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 SIGMOD International 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%20Ch en-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) 412-421. Retrieve from: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. Retrievedfrom: 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 Biennial Conference on Innovative Data Systems Research(CIDR2011),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).Retrievedfrom: http://nsfcac.rutgers.edu/CometCloud/sites/nsfcac.rutgers.edu.CometCloud/files/pub/ICCTA2 009.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
[25] 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=0CCYQgAMoADAAahUKEwi3
k4Hymv7GAhUHUKYKHdToBCM
[26] 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. KnowlInfSyst (2008) 14:1–37. DOI: 10.1007/s10115-007-0114-2. Retrieved from: http://www.cs.umd.edu/~samir/498/10Algorithms-08.pdf