International Journal of Academic Multidisciplinary Research (IJAMR)
ISSN: 2000-006X
Vol. 2 Issue 4, April – 2018, Pages: 12-18

A Cloud Based Approach For Big Data Analysis: A Comprehensive Review

[1]Lusekelo Kibona, [2]Hassana Ganame, [3]Kazi Md Shahiduzzaman

[1]Department of Computer Science
Ruaha Catholic University (RUCU), Tanzania
lusekelo2012@gmail.com

[2]Department of Information and Telecommunication
School of Engineering of Bamako, Mali
ganame_hassana@yahoo.fr

[3]Department of Electrical and Electronic Engineering
Jatiya Kabikazi Nazrul Islam University, Trishal, Mymensingh, Bangladesh
kazi_eee05@yahoo.com

Abstract: *The quick escalation of the Internet and the digital economy has driven an exponential rise in response for data storage and analytics, and most organizations are facing difficult challenge in protecting and analysing these increased volumes of information. Big data and cloud computing are both the fastest-moving technologies from the last decade, so there is a need to integrate the two technologies together to come up with a new defined technology which will solve the present problems in either big data alone or cloud computing alone. The main aim of this study was to visit different literature in cloud-based approach on big data analysis and find out some challenges countered on the cloud computing approach in big data analysis. It was found that there are still difficulties in using cloud based approach in big data analysis especially in terms of security and noisy data retrieved from cloud server. It was recommended that further research must be done on ensuring security on the large amount of data stored in the cloud servers.*

Keywords— Cloud computing, big data analysis, Cloud based server, digital economy, Big Data-as-a-Service

1. INTRODUCTION

The quick escalation of the Internet and the digital economy has driven an exponential rise in response for data storage and analytics, and most organizations are facing difficult challenge in protecting and analyzing these increased volumes of information.

Generating, collecting, distributing, processing and analyzing extraordinary amounts of diverse data has become a core topic in different industries and research disciplines as well as for society as a whole [1]. The term ‘‘big-data’’ was created to find the thoughtful meaning of this data-explosion trend [2].

It is useful in data collections whose proportions or type is outside the ability of traditional relational databases to capture, manage, and process the data with low-latency and it has one or more of the following characteristics – high volume, high velocity, or high variety [3], also is a term that can only be defined relative to something and even when care is taken in its definition, gaps readily form that lead to discrepancies [4].

Big data analytics is the process of scrutinizing outsized and diverse data sets i.e., big data to expose unseen patterns, unidentified correlations, market trends, customer likings and other useful information that can help organizations make more-informed business decisions [5].

According to IBM [3], big data analytics is the use of forward-looking analytic techniques against very large, varied data sets that include different types such as organized/free and streaming/batch, and different sizes from terabytes to zettabytes, it mirrors the challenges of data that are too massive, too unstructured, and too fast moving to be accomplished by traditional ways and means.

Big data analysis can be described as the sub-area of big data concerned with adding structure to data to back choice making as well as supporting domain-specific usage scenarios [6].

Gantz *et* al [7], defined ‘*Big data technologies describes a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis’.*

A part from four characteristics of big data (volume, variety, value and velocity) [8], added another characteristic which is veracity, so now there are 5 V’s of big data characteristics which are summarized as follows:

- ✓ Volume- *Measures the amount of data available to an organization, which does not necessarily have to own all of it as long as it can access it.*
- ✓ Variety - *Measures the richness of the data representation – text, images video, audio, etc. From an analytic perspective, it is probably the*

biggest obstacle to effectively using large volumes of data.

- ✓ Velocity - *Refers to the speed of data transfer. It measures the speed of data creation, streaming, and aggregation.*
- ✓ Veracity - *A lot of data generated are noisy, e.g., data from sensors. Data are often incorrect. For example, many websites accessed may not have the correct information. It is difficult to be certain about the veracity of big data.*
- ✓ Value - *Refers to the process of discovering huge hidden values from large datasets with various types and rapid generation. Data by itself is of no significance unless it is processed to acquire information.*

Big data exploits dispersed storage technology based on cloud computing rather than local storage devoted to a computer or electronic device and cloud computing not only does it provides facilities for the computation and processing of big data but also serves as a service model [9].

Cloud computing is an information technology (IT) archetype that permits worldwide access to shared pools of configurable system possessions and higher-level services which can be rapidly provisioned with minimal management effort, often over the Internet or it can be defined as the delivery of computing services such as servers, storage, databases, networking, software, analytics and more over the Internet ("the cloud"), cloud computing can be deployed as private, public or hybrid as explained below [10-13].

- ❖ ***Public cloud:*** Are owned and operated by companies that offer rapid access over a public network to affordable computing resources.
- ❖ ***Private cloud:*** Are operated solely for a single organization, whether managed internally or by a third party and hosted internally or externally.
- ❖ ***Hybrid cloud:*** Combines public and private clouds, bound together by technology that allows data and applications to be shared between them

The main services provided by cloud computing includes Infrastructure as a service (IaaS), Platform as a service (PaaS) and Software as a service (SaaS) that run on servers reachable through the Internet rather than be present on the desktop and are briefly explained below [10-13].

- ✓ IaaS - *Provides companies with computing resources including servers, networking, storage, and data center space on a pay-per-use basis.*
  Users have an allocated storage capacity and can start, stop, access and configure the VM and storage as desired.
- ✓ PaaS - *Refers to cloud computing services that supply an on-demand environment for developing, testing, delivering and managing software applications.*
  Users access developing tools over the internet using APIs, web portals or gateway software.
- ✓ SaaS - *Refers to a method for delivering software applications over the Internet, on demand and typically on a subscription basis.*
  Users can access SaaS applications and services from any location using a computer or mobile device that has internet access.

Cloud computing plays a main role for Big Data; not because it provides infrastructure and tools, but also because it is a business prototypical that Big Data analytics can follow (e.g. Analytics as a Service (AaaS) or Big Data as a Service (BDaaS)) [14].

Big data in cloud atmospheres can help organizations define, evaluate, and act on data aggregates in a simple and convenient way without requiring much physical server space.

Big data bases from the cloud and Web are warehoused in a distributed fault-tolerant database and managed through a programing model for large datasets with a parallel-distributed algorithm in a cluster, the following diagram illustrates the use of cloud computing in big data analysis.

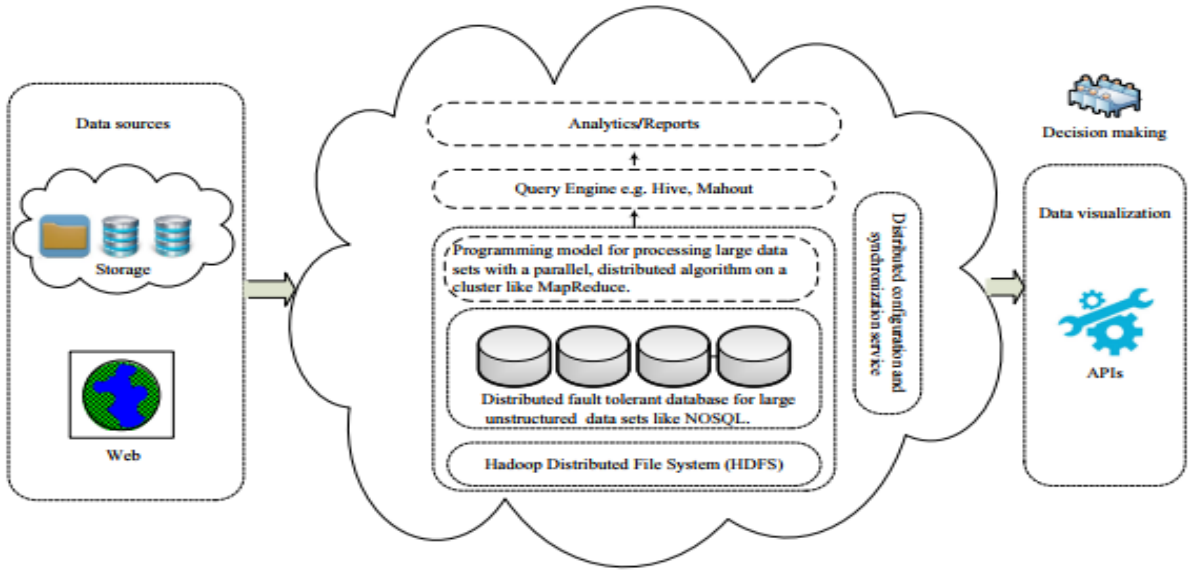

Figure 4: Diagram showing cloud computing usage in big data [9]

Big data and cloud computing are both the fastest-moving technologies from the last decade, so there is a need to integrate the two technologies together to come up with a new defined technology which will solve the present problems in either big data alone or cloud computing alone.

The main aim of this article is to comprehensively review the integration between cloud computing based approach on big data analytics by reviewing cloud computing services, types and its usage in big data analytics and also to find out some challenges countered on the cloud computing approach in big data analysis.

2. LITERATURE SURVEY

Storage and data transport are technology problems, which seem to be answerable in the near-term, but represent longterm challenges that require research and new paradigms [15].

According to [8], in his research titled 'Big data analytics', discussed the possibility of analyzing huge data collections (mostly of unstructured data such as emails, blogs, Twitter, Facebook posts, images, and videos estimated that in 2015, 8 Zettabytes (Zetta=1021)) with clusters of thousands of low-cost computers to discover arrays in the data that have many applications but he warned that analyzing gigantic volumes of data obtainable in the Internet has the potential of imposing on our confidentiality.

[16], provided a synopsis on the topic of Big Data on how the current problem can be addressed from the perspective of Cloud Computing and its programming frameworks, they focused on those systems for large-scale analytics grounded on the MapReduce scheme and Hadoop, its open-source implementation.

To address the challenge of storing, managing, and creating values from the service-oriented big data have become an important research challenge, to address this challenge Zheng et al [17], provided an overview on service-generated big data and Big Data-as-a-Service in which Big Data-as-a-Service, including Big Data Infrastructure-as-a-Service, Big Data Platform-as-a-Service, and Big Data Analytics Software-as-a-Service, is employed to provide common big data related services (e.g., accessing service-generated big data and data analytics results) to users to enhance efficiency and reduce cost.

High volumes of event data produced by the execution of processes during the business lifetime thwart business users from proficiently accessing timely analytics data, in their article Vera et al [18], presented a technological solution using a big data approach to provide business analysts with visibility on distributed process and business performance by proposing an architecture which allow users to evaluate business performance in highly distributed environments with a short time response.

Demchenko et al [19] discussed a nature of Big Data that may originate from different scientific, industry and social activity domains and proposes improved Big Data definition that includes the following parts: Big Data properties (also called Big Data 5V: Volume, Velocity, Variety, Value and Veracity), data models and structures, data analytics, infrastructure and security they also discussed paradigm change from traditional host or service based to data centric architecture and operational models in Big Data.

Schmidt et al [20], developed a framework that reckons the alternatives for executing Big Data applications using cloud-services and identified the strategic areas supported by these Alternatives, they created framework that clarified the options for Big Data initiatives using cloud-computing and thus improved the strategic alignment of Big Data applications.

In their research titled 'Towards Service-Oriented Enterprise Architectures for Big Data Applications in the Cloud', Zimmerman et al [21], proposed a new integration model for service-oriented Enterprise Architectures on basis of ESARC - Enterprise Services Architecture Reference Cube, which was their earlier developed service-oriented enterprise architecture classification framework, with MFESA - Method Framework for Engineering System Architectures - for the design of service-oriented enterprise architectures, and the systematic development, diagnostics and optimization of architecture artifacts of service-oriented cloud-based enterprise systems for Big Data applications.

Zhang et al [22] proposed two online algorithms: an online lazy migration (OLM) algorithm and a randomized fixed horizon control (RFHC) algorithm , for optimizing at any given time the choice of the data center for data aggregation and processing, as well as the routes for transmitting data there and careful comparisons among these online and offline algorithms in accurate settings were conducted through widespread experiments, which determine close-to-offline-optimum performance of the online algorithms.

According to [23], the main objective of cloud computing is to share resources, which consist of infrastructure, platform, software, and business process and when those resources are provisioned as services, the value of cloud computing is realized, this also comes in servicelization which is the way of offering social networking services, big data analytics, and mobile Internet services, in short, "everything as a service" is creating a Big Services age due to the foundational architecture (i.e., Service-Oriented Architecture) of services computing.

Bi et al [24],presented a work which clarified the necessities of predictive systems, and identified research challenges and opportunities on BDA to support cloud-based information systems.

As per Rajinder et al [25], a global architecture was proposed for QoS based scheduling for big data application to distributed cloud datacenter at two levels which are coarse grained and fine grained, at coarse grain level, appropriate local data center was chosen based on network distance between user and data center, network throughput and total available resources using adaptive K nearest neighbor algorithm while at fine grained level, probability triplet (C, I, M) is predicted using naïve Bayes algorithm which provides probability of new application to fall in compute intensive (C), input/output intensive (I) and memory intensive (M) categories, each data center was then changed into a pool of virtual clusters capable of executing specific category of jobs with specific (C, I, M) requirements using self-organized maps.

Shang et al [26], proposed a lightweight approach for detecting differences between pseudo and large-scale cloud deployments, their approach made use of the readily-available yet rarely used execution logs from these platforms, the approach abstracts the execution logs, recovers the execution sequences, and compares the sequences between the pseudo and cloud deployments.

Grolinger et al [27], identified challenges that were grouped into four main categories corresponding to Big Data tasks types: data storage (relational databases and NoSQL stores), Big Data analytics (machine learning and interactive analytics), online processing, and security and privacy. Additionally, current efforts aimed at improving and extending MapReduce to address identified issues and challenges MapReduce faces when handling Big Data.

According to Zhang et al [28], a scalable multidimensional anonymization method for big data privacy protection was proposed by using Map Reduce on cloud, in their approach, a highly scalable median-finding algorithm merging the idea of the median of medians and histogram technique is proposed and the recursion granularity was controlled to achieve cost-effectiveness.

Suciu et al [29], analyzed current mechanisms and methods of firmly incorporating big data processing with cloud Machine to Machine (M2M) systems based on Remote Telemetry Units (RTUs) and proposed a converged E-Health architecture built on Exalead Cloud View, a search based application.

As per Terzo et al [30], a DaaS approach for intelligent sharing and processing of large data gatherings with the aim of abstracting the data location (by making it suitable to the needs of sharing and retrieving) and to fully decouple the data and its processing was proposed, the aim of their methodology was to build a Cloud computing platform, offering DaaS to support large communities of users that need to share, access, and process the data for collectively building knowledge from data.

Patel et al [31], reported the experimental work on big data problem and its optimal solution using Hadoop cluster, Hadoop Distributed File System (HDFS) for storage and using parallel processing to process large data sets using Map Reduce programming framework, they prototyped implementation of Hadoop cluster, HDFS storage and Map Reduce framework for processing large data sets by considering prototype of big data application scenarios.

Assunção et al [14], critised AaaS/BDaaS, as it brings several challenges because the customer and provider's staff are much more involved in the loop than in traditional Cloud providers offering infrastructure/platform/software as a service.

By what means cloud computing can give a clarification for big data with cloud services and open-source cloud software tools for treatment of big data issues have been clearly explained by Bahram et al [32], they also explained the role of cloud architecture in big data, the role of major cloud service layers in big data, and the role of cloud computing systems in handling big data in business intelligence models.

Ahuja et al [33], scrutinized the present developments and features of Big Data, its examination and how these are offering challenges in data collection, storage and management in cloud computing.

Cloud Computing provides scalability with respect to the use of resources, low administration effort, flexibility in the pricing model and mobility for the software user, due to these circumstances, it is understandable that the Cloud Computing paradigm benefits large projects, such as the ones related with Big Data and BI [34].

3. METHODOLOGY

The methodology adopted by this study was 'Literature readings from previous researchers and then analysed what have been previously done in the field of Cloud computing merged with big data'. The study consulted different sources on the Internet to establish evidence and facts about big data analytics and cloud computing, the author went further on finding the sources about the integration of cloud computing approach on the big data analysis. Where possible the websites of the specific resource were visited, for example website of some journals that only put materials in html format rather than pdf or documents. The reviewed literatures are mostly available on the Internet and where possible in some areas algorithm were modified to facilitate the discussion. So generally secondary source of data were mainly used in a large part to come up to conclusion.

4. DISCUSSION AND CHALLENGES

The discussion, technology about integrating cloud based approach on big data analysis, concepts and models from different literatures involves different analysis and interpretation.

Integrating Cloud computing with big data seems to be a savior approach to the current crisis of data storage which requires a huge servers or storage devices with maximum care about data loss.

The approach itself has some drawbacks as it have been observed in some literatures, like the use for example of MapReduce which seems to be a very new approach in big data processing in cloud computing has the weakness in supporting updates and recapture from disasters during updating of the stored data, in transforming input data in files into a format needed by a certain DBMS and so many other weaknesses.

Another new approach for cloud based computation for big data analysis is HADOOP database which seems to cover some weaknesses of MapReduce but itself has got some drawbacks like forcing users to use certain type of DBMS like mostly users must be required to use SQL rather than any other DBMS, which for current technological advancement is not necessary so it is required to develop an API (Application Program Interface) which can be integrated to any DBMS.

Storing a large amount of big data in cloud may also pose another challenge in retrieving them because upon retrieving you may encounter another amount of data embedded with the data you stored which are not relevant to you, those kind of undesired data are commonly termed as noisy data (unrequired data which are retrieved with the significant or needed data) a real example of this is shown in figure below

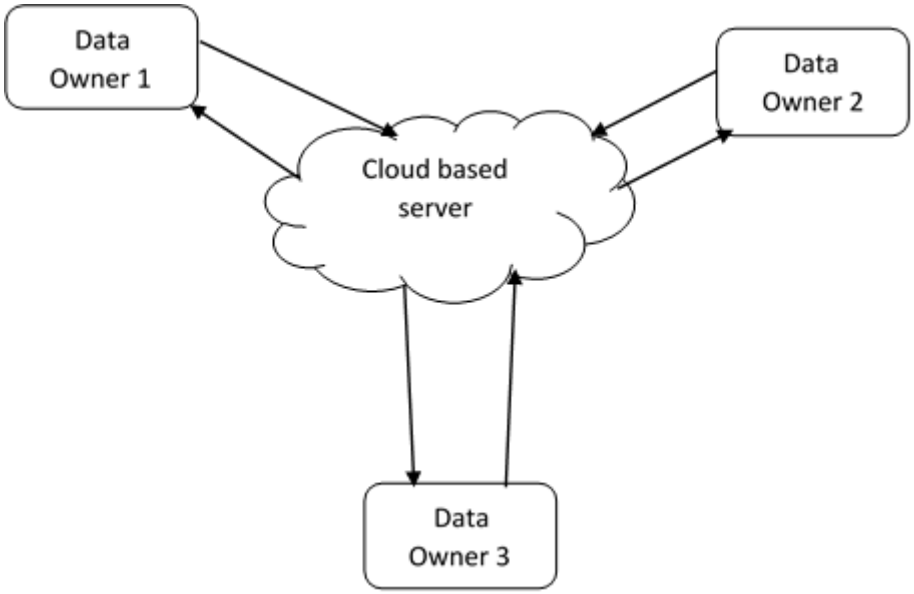

Figure 2: showing three data sources (Owners), sending and retrieving data from Cloud based server.

In figure 2, it can be easily understood that the data owners 1, 2, 3 can send data to be stored in cloud server in what we can say they send clean data, but upon retrieving by using either keywords or index words, it is possible for the data from data owner 1 be retrieved by data owner 2 if and only if the keywords are similar among the two data owners, so the data retrieved by owner 2 but belongs to data owner 3 are termed as noisy data by data owner 2. So this is one of the challenge big data face in cloud computing.

Even though cloud computing provides minimal administrative costs of big data management, it also has some flaws in terms of security as it becomes very difficult to ensure maximum security in data stored in clouds.

5. CONCLUSION

According to visited literature reviews, which brings about the secondary data sources and some few primary data sources, it seems that there are still some challenges in totally integrating cloud computing and big data analysis. The main aim of this study was to visit different literature in cloud-based approach on big data analysis and find out some challenges countering the integration between the cloud computing and big data analysis. The results of the literature review visited shown that there are still difficulties in using cloud based approach in big data analysis especially in terms of security and noisy data retrieved from cloud server as one can retrieve undesired data which were not intended to be retrieved but they are retrieved because of sharing the same keyword or index word, this makes duplication of data which may sometimes occupy unnecessary space required to be occupied by the required data item sets.

Even if this paper has not fixed the complete theme about this important topic, confidently it has delivered some valuable argument and an outline for scientists.

6. RECOMMENDATION AND FUTURE WORK

In the future, research must be done on:

- ✓ Ensuring security on the data stored in the cloud servers, and
- ✓ Ensuring that the data stored in the cloud servers upon retrieval they must be obtained without the noisy data, which are termed so to mean data unrequired to be accessed at the time of retrieval.

REFERENCES

[1] E. Rahm, "Big Data Analytics," it-Information Technology, vol. 58, pp. 155-156, 2016.

[2] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," IEEE access, vol. 2, pp. 652-687, 2014.

[3] IBM. (2017). What is Big Data Analytics? Available: https://www.ibm.com/analytics/us/en/technology/hadoop/big-data-analytics/

[4] T. V. Kumar, D. JNU, P. S. Rana, M. S. Sinha, H. Tagra, M. B. Misra, et al., "Big Data Analytics," 2017.

[5] TechTarget. (2017). Big Data Analytics. Available: http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

[6] J. Domingue, N. Lasierra, A. Fensel, T. van Kasteren, M. Strohbach, and A. Thalhammer, "Big data analysis," in New Horizons for a Data-Driven Economy, ed: Springer, 2016, pp. 63-86.

[7] J. Gantz and D. Reinsel, "Extracting value from chaos," IDC iview, vol. 1142, pp. 1-12, 2011.

[8] V. Rajaraman, "Big data analytics," Resonance, vol. 21, pp. 695-716, 2016.

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

[10] Wikipedia. (2017). Cloud computing. Available: https://en.wikipedia.org/wiki/Cloud_computing

[11] TechTarget. (2017). Cloud computing. Available: http://searchcloudcomputing.techtarget.com/definition/cloud-computing

[12] IBM. (2017). What is cloud computing? Available: https://www.ibm.com/cloud/learn/what-is-cloud-computing

[13] Microsoft. (2017). What is cloud computing? Available: https://azure.microsoft.com/en-gb/overview/what-is-cloud-computing/

[14] M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. Netto, and R. Buyya, "Big Data computing and clouds: Trends and future directions," Journal of Parallel and Distributed Computing, vol. 79, pp. 3-15, 2015.

[15] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big data: Issues and challenges moving forward," in System Sciences (HICSS), 2013 46th Hawaii International Conference on, 2013, pp. 995-1004.

[16] A. Fernández, S. del Río, V. López, A. Bawakid, M. J. del Jesus, J. M. Benítez, et al., "Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 4, pp. 380-409, 2014.

[17] Z. Zheng, J. Zhu, and M. R. Lyu, "Service-generated big data and big data-as-a-service: an overview," in Big Data (BigData Congress), 2013 IEEE International Congress on, 2013, pp. 403-410.

[18] A. Vera-Baquero, R. Colomo-Palacios, and O. Molloy, "Business process analytics using a big data approach," IT Professional, vol. 15, pp. 29-35, 2013.

[19] Y. Demchenko, C. De Laat, and P. Membrey, "Defining architecture components of the Big Data Ecosystem," in Collaboration Technologies and Systems (CTS), 2014 International Conference on, 2014, pp. 104-112.

[20] R. Schmidt and M. Mohring, "Strategic alignment of cloud-based architectures for big data," in Enterprise Distributed Object Computing Conference Workshops (EDOCW), 2013 17th IEEE International, 2013, pp. 136-143.

[21] A. Zimmermann, M. Pretz, G. Zimmermann, D. G. Firesmith, I. Petrov, and E. El-Sheikh, "Towards service-oriented enterprise architectures for big data applications in the cloud," in Enterprise Distributed Object Computing Conference Workshops (EDOCW), 2013 17th IEEE International, 2013, pp. 130-135.

[22] L. Zhang, C. Wu, Z. Li, C. Guo, M. Chen, and F. C. Lau, "Moving big data to the cloud: An online cost-minimizing approach," IEEE Journal on Selected Areas in Communications, vol. 31, pp. 2710-2721, 2013.

[23] L.-J. Zhang, "Big services era: Global trends of cloud computing and big data," IEEE Transactions on Services Computing, vol. 5, pp. 467-468, 2012.
[24] Z. Bi and D. Cochran, "Big data analytics with applications," Journal of Management Analytics, vol. 1, pp. 249-265, 2014.
[25] R. Sandhu and S. K. Sood, "Scheduling of big data applications on distributed cloud based on QoS parameters," Cluster Computing, vol. 18, pp. 817-828, 2015.
[26] W. Shang, Z. M. Jiang, H. Hemmati, B. Adams, A. E. Hassan, and P. Martin, "Assisting developers of big data analytics applications when deploying on hadoop clouds," in Proceedings of the 2013 International Conference on Software Engineering, 2013, pp. 402-411.
[27] K. Grolinger, M. Hayes, W. A. Higashino, A. L'Heureux, D. S. Allison, and M. A. Capretz, "Challenges for mapreduce in big data," in Services (SERVICES), 2014 IEEE World Congress on, 2014, pp. 182-189.
[28] X. Zhang, C. Yang, S. Nepal, C. Liu, W. Dou, and J. Chen, "A MapReduce based approach of scalable multidimensional anonymization for big data privacy preservation on cloud," in Cloud and Green Computing (CGC), 2013 Third International Conference on, 2013, pp. 105-112.
[29] G. Suciu, V. Suciu, A. Martian, R. Craciunescu, A. Vulpe, I. Marcu, et al., "Big data, internet of things and cloud convergence–an architecture for secure e-health applications," Journal of medical systems, vol. 39, p. 141, 2015.
[30] O. Terzo, P. Ruiu, E. Bucci, and F. Xhafa, "Data as a service (DaaS) for sharing and processing of large data collections in the cloud," in Complex, Intelligent, and Software Intensive Systems (CISIS), 2013 Seventh International Conference on, 2013, pp. 475-480.
[31] A. B. Patel, M. Birla, and U. Nair, "Addressing big data problem using Hadoop and Map Reduce," in Engineering (NUiCONE), 2012 Nirma University International Conference on, 2012, pp. 1-5.
[32] M. Bahrami and M. Singhal, "The role of cloud computing architecture in big data," in Information granularity, big data, and computational intelligence, ed: Springer, 2015, pp. 275-295.
[33] S. P. Ahuja and B. Moore, "State of big data analysis in the cloud," Network and Communication Technologies, vol. 2, p. 62, 2013.
[34] K. Shim, S. K. Cha, L. Chen, W.-S. Han, D. Srivastava, K. Tanaka, et al., "Data management challenges and opportunities in cloud computing," in Proceedings of the 17th international conference on Database Systems for Advanced Applications-Volume Part II, 2012, pp. 323-323.