Big Data in Cloud Computing: A Review of Key Technologies and Open Issues

Elena Canaj[1(✉)] and Aleksandër Xhuvani[2]

[1] Faculty of Information Technology, Graduate School of ICT, Polytechnic University of Tirana, Tirana, Albania
elenacanaj@gmail.com
[2] Department of Computer Engineering, Faculty of Information Technology, Polytechnic University of Tirana, Tirana, Albania
axhuvani@fti.edu.al

Abstract. Academia, industry and government as well, are involved in big data projects. Many researches on big data applications and technologies are actively being conducted. This paper presents a literature review of recent researches on key technologies and open issues for big data management via cloud computing. Its goal is to identify and evaluate the main technology components and their impacts on cloud-based big data implementations. This is achieved by reviewing 40 publications published in the latest four years, 2014–2017. We classified the results based on the main technical aspects: frameworks, databases and data processing techniques, and programming languages. This paper also provides a reference source for researchers and developers, to determine the best emerging technologies for big data project implementation.

1 Introduction

In recent years, big data management has attracted a lot of attention. Big data, as a variety of data, structured, semi-structured and unstructured, requires high storage and high performance computing. The processing of big data is easier via cloud computing due to its distributed computational paradigm. The cloud computing architecture provides a good solution for large-scale data storage and processing, addressing two of the main requirements of big data. Although big data management as a service in cloud computing has solved many of the big data requirements, it also has raised many important issues, related to the data migration in cloud such as, data security and data privacy.

Many researches on the new methodologies and technologies for both cloud computing and big data are proposed and developed recently. In this paper we present a literature review of recent researches for big data management via cloud computing. This paper identifies and evaluates the key technologies and open research issues of big data management via cloud computing. Its contribution in that respect may be summarized as follows:

- A literature review of key technologies for big data deployment in cloud computing with respect to frameworks, databases and data processing techniques, and programming languages;

L. Barolli et al. (Eds.): EIDWT 2018, LNDECT 17, pp. 504–513, 2018.
https://doi.org/10.1007/978-3-319-75928-9_45

- An overview over the open research issues and challenges of big data management via cloud computing.

Within the context of this paper we provide a reference source for researchers and developers, to determine the best emerging technologies for big data project implementation via cloud computing. For a researcher who is exploring the big data deployment in cloud, is really critical to determine which tools to use during project implementation and the concerns that should be taken into considerations.

The paper consists in: Sect. 2, in which the research methodology is deployed; Sect. 3, in which results and analysis of the searches in a quantitative perspective are presented; Sect. 4, that gives a detailed description and evaluation of the reviewed papers; Sect. 5, that provides an overview of open issues and challenges for big data management in cloud computing; and the final section is the epilogue concluding our work.

2 Methodology

The objective of this paper is to identify and evaluate the main technology components and open issues for implementing big data management as a service in cloud by reviewing and structuring the existing literature. Over hundred publications were first extracted from searches made on three reference and citation-enhanced indexing databases, Google Scholar, Scopus, and Web of Science for the following keywords: big data, cloud computing, Hadoop, MapReduce, Spark, NoSQL, programming language. We paid a particular attention to publications of research results on digital libraries: ACM, IEEE Xplore, SpringerLink, and Elsevier. The time range for this search was limited from 2014 until November 2017. The challenges related to the cloud platforms and frameworks, techniques for big data storage, pre-processing and processing, databases, algorithms, and programming models were all within the scope of this review paper. We focused on reviewing researches of open source technologies but important development on commercial ones are analyzed as well. Papers that were purely focused on technical design were left out of this review. From the numerous research publications, at the end 40 researches and reports were selected and analyzed.

The selected publications are classified based on their main research focus; the categories are: Frameworks, Databases and Data Processing Techniques, and Programming Languages. We analyzed them from a quantitative and qualitative perspective.

3 Literature Review: Quantitative Results and Analysis

In this section we present the results of our work in a quantitative perspective. The selected publications are analyzed and evaluated based on their research contributions. They are noted by their type as Review, Survey, Technical Report, New Design Proposal, Comparative Study, and special attention is paid to real experiments, simulation/emulation and system implementations made by authors. Table 1 shows all the selected publications for this review.

Table 1. List of selected publications.

Ref.	Frameworks	Databases	Programming models	Type	Implement/ Experiments	Year
[1]	Main Topic		x	Survey		2017
[2]	Main Topic			Technical Report	x	2016
[3]	Main Topic		x	New Design	x	2014
[4]	Main Topic		x	New Design	x	2017
[5]	Main Topic		x	New Design	x	2015
[6]	Main Topic			New Design	x	2014
[7]	Main Topic			Review		2015
[8]	Main Topic		x	New Design	x	2016
[9]	Main Topic		x	New Design	x	2016
[10]	Main Topic		x	New Design	x	2015
[11]	Main Topic		x	Case Study	x	2015
[12]	Main Topic		x	New Design	x	2015
[13]	Main Topic			Comparative Study	x	2014
[14]	Main Topic			Survey		2015
[15]	Main Topic			New Design	x	2015
[16]	x	Main Topic	x	Review		2016
[17]		Main Topic		Review		2014
[18]		Main Topic		Comparative Study	x	2017
[19]	x	Main Topic	x	New Design	x	2016
[20]	x	Main Topic	x	Survey		2017
[21]		Main Topic		Survey		2015
[22]	x	Main Topic	x	Survey		2017
[23]		Main Topic		Survey		2017
[24]		Main Topic		New Design	x	2015
[25]		Main Topic		Review		2015
[26]	x	Main Topic		New Design	x	2015
[27]		Main Topic		Comparative Study		2016
[28]		Main Topic		New Design	x	2016
[29]	x	Main Topic	x	New Design	x	2015
[30]			Main Topic	Review		2017
[31]			Main Topic	Review		2017
[32]			Main Topic	Review		2014
[33]			Main Topic	Survey		2015
[34]	x		Main Topic	New Design	x	2017
[35]			Main Topic	Technical Report	x	2015
[36]	x		Main Topic	Comparative Study	x	2014
[37]			Main Topic	New Design	x	2017
[38]			Main Topic	Survey		2017
[39]			Main Topic	New Design	x	2014
[40]			Main Topic	New Design		2016

Based on the research contribution, we present the results on the total number of publications per category in Table 2. During that process we observed that framework component is mostly researched.

Table 2. Distribution of main technology aspects

Domain	Total no. of publications	Percentage Papers/Category (%)
1. Frameworks	15	37
2. Databases and data processing techniques	14	35
3. Programming languages	11	28
Total	40	100

In the Table 3, it can be noticed that the majority of the publications considered and analysed are original researches proposing new designs and improvements for big data management in cloud computing.

Table 3. Representation of the total number per publication type.

Publication type	Total no. of publications	Percentage Publication/Type (%)
Survey	8	20
Review	7	17
New Design	18	45
Technical Report	2	5
Comparative Study	4	10
Case Study	1	3
Total	40	100

4 Literature Review: Topics-Related Analysis

This section is an overview of each of the selected papers. The publications are mapped based on the main topic and their contributions on big data as a service implementation in cloud. Our work is focused more on reviewing researches of open source technologies, but important development on commercial ones are analyzed as well.

4.1 Frameworks

In this subsection, we will analyze and discuss recent researches of the most essential component of a big data system. To simplify the decision on choosing the best framework solution for big data implementation in cloud, we reviewed 15 publications that address development, improvements, experiments and open issues on various frameworks [1–15].

Firstly we analysed the publications found during our search for Hadoop framework. Hadoop is the backbone of several large scale applications in different domains for analyzing large scale data. In the first paper [1] Hadoop framework is used to analyze workload prediction of data from cloud computing. Hadoop was also used for analyzing the tweets on the large scale in paper [2]. Authors in [3] used Hadoop, and MapReduce parallel processing paradigm to propose a new solution - Keyword-Aware Service Recommendation method for analyzing data in service recommender systems. The results of these papers show that Hadoop framework is a good choice for batch processing that are not time-sensitive. Original researches proposing improvements and new designs for Hadoop are done. Authors in [4] propose an integrated Hadoop and MPI/OpenMP system for higher processing speed.

Open issues are raised in literature relating Hadoop security, due to missing encryption at the storage and network levels. According to this issue authors in [5] propose a new security model for G-Hadoop, which is based on public key cryptography and the SSL protocol.

In addition to Hadoop, there are several other frameworks studied and proposed in the following papers.

Authors in [6] investigated the performance of big data applications on Spark with different virtualization frameworks. Review paper [7] analyses the infrastructure of another open source framework, Apache Storm. In paper [8] authors discuss on the evolution of big data frameworks and propose a new design framework Scalation using the Scala programming language. Spark, Samza, Kafka and Scalation frameworks are evaluated through experiments. Authors in [9] propose a new framework design, High Performance Analytics Toolkit (HPAT). Their evaluation has demonstrated that HPAT is faster than Spark. Authors in [10] have implemented a cloud-based analytics service using Hadoop and Spark, and the results are being compared. In paper [11] authors investigated the Yahoo!S4 framework for processing of real time streams. Authors in [12] present a new cloud based framework for big data management in smart grids. In paper [13] a detailed comparative study of three different frameworks Apache Hadoop, Project Storm, Apache Drill is performed. Survey paper [14] also presents a theoretical comparison of the main frameworks for big data. Authors in [15] analyze DistributedWekaSpark, a distributed Spark framework for Weka workbench.

At the end of this subsection we outline our findings as following: Hadoop framework is suited for workload where time is not critical. It's a good choice for batch processing that are not time-sensitive. It is an open source and easier to implement than other solutions. For stream processing Storm and S4 are typical frameworks for real-time large scale streaming data. These frameworks have very low latency processing. Samza framework when is integrated with Kafka also is a good solution for stream processing. Flink frameworks support stream processing and also handle batch processing. It is heavily optimized, but it is still unstable. For interactive environment, Spark is very adequate. Apache drill is also best for interactive and ad-hoc analysis.

There are many options for processing data in cloud based big data system but the best fit for any implementation depends on the data to process and time requirements. Workbenches like the one analyzed in this section are available for testing the

frameworks. For researchers and developers, it's possible to simulate some situations prior the final implementation of any project. The biggest concerns that should be taken into consideration while using frameworks in cloud computing are data security and data privacy.

4.2 Databases and Data Processing Techniques

In this subsection 14 recent publications on data storage and processing techniques are reviewed [16–29].

The first article reviewed on this topic [16] outlines the appropriate technologies for implementing big data project. This article includes technologies such as in-memory databases, NoSQL and NewSQL systems, and Hadoop based solutions.

A comparative study about the performance of NoSQL databases: BigTable, Cassandra, HBase, MongoDB, CouchDB, CrowdDB is done in paper [17]. Authors in [18] also compare the performance of HBase and MongoDB. Authors in [19] present a new solution that integrates Cassandra with MapReduce. Paper [20] is a review of the researches done for incorporation of data warehouse with MapReduce for handling of big data. Paper [21] is a survey of in-memory big data management systems. Authors in [22], also presents a survey on recent technologies developed on big data, for Data Processing Layer, Data Querying Layer, Data Access Layer and Management Layer. Paper [23] studies the solution of various types of unstructured data storage, analyzes all the problems existing in the storage system, and summarizes the key issues to achieve the unified storage of unstructured data.

Authors in [24] propose new solutions for implementation of big data warehouses under the column oriented NoSQL DBMS. In paper [25], a detailed classification for modern big data models is done. Paper [26] introduces the combination of NoSQL database HBase and enterprise platform Solr. Authors in [27] have compared SQL with NoSQL databases and the four NoSQL data models (document-oriented, key-value pairs, column-oriented or graphs). In paper [28], a new design allowing the automatic transformation of a multidimensional schema into a tabular schema is implemented in Hive. Last paper [29] evaluates the performance of Spark SQL.

At the end of this subsection we summarize our findings. Traditional relational database management systems are not suitable for big data. NoSQL database management systems are designed for use in high data volume applications in cloud environments. Several open source NoSQL databases exist, MongoDB, Cassandra, CouchDB, CrowdDB, Hypertable, HBASE, Couchbase, etc. NoSQL databases are highly scale able, flexible and good for big data storage and processing. The current issue with NoSQL databases is that they do not offer a declarative query language similar to SQL and there is no single, unified model of NoSQL databases. Integrated solutions are researched and proposed. Applications that combine relational and procedural queries run faster. We noticed that many researches are still going forward to optimize data storage and processing techniques. Heterogeneity of data is also a problem that is currently under study. We have review NoSQL data models that can process big data up to the petabyte range. Exabyte range data processing is still an open problem under-researched.

4.3 Programming Languages

We selected 11 publications from our search results that address latest researches on programming languages [30–40].

Firstly we have reviewed all existing literatures that summarize existing programming languages and paradigms available in cloud-based big data area. Review papers [30–32] discuss and compare various programming models, analyzing how they fit into the big data projects.

Authors in [33], presents a survey of programming models for big data implementations in grid and cloud. Nystrom in [34] describes a Scala framework that can be used for experimenting with supercompilation techniques. In this paper a supercompiler for JavaScript is implemented. Authors in [35] analyze the new programming language Julia which is appropriate for parallel computing. In paper [36] programming language R is integrated with Hadoop framework and they are used together for big data statistical programming. Authors in [37] propose a new design in order to manage language runtimes. Survey paper [38] analyzes the MapReduce-based algorithms for handling big RDF graphs. Paper [39] investigates the program transformations for Pig Latin. Authors in [40] propose a new approach JAVA2SDG, for stateful big data processing.

In the reviewed publications we found many programming languages like R, Python, Java, Scala, Hadoop languages (Pig Latin, Hive), Julia and new programming language proposals. We analyzed all of them in order to have a full understanding of the actual research process in the field of programming paradigm for cloud-based big data. Programming language R is adequate for executing large numbers of calculations. Python is a good choice for advanced analytics. Scala is also a good choice for large streaming since it is a hybrid programming language, which combine both Functional Programming with Object Oriented Paradigms. Java is used as a basic code for many frameworks but it has a very high learning curve. The new programming language Julia fits well for real-time streams applications.

We conclude that Functional Programming (FP) is considered to be the most adequate for big data implementations. Its difficult syntax has pushed researchers to try new hybrid solutions. During our search as it is shown in Sect. 3, we didn't find quite many recent researches on programming languages. As stated at IEEE Spectrum "The 2017 Top Programming Languages", it seems a period of consolidation in coding as programmers digest the tools created for cloud and big data applications [41].

5 Challenges and Open Research Issues

Although big data management in cloud computing is widely used, it still faces challenges and open issues. This section presents an overview of the open issues identified in the literature that affect big data deployment in cloud computing. We address the following open research issues for further improvement:

- Security and privacy: Data security and data privacy are critical issues when migrating big data to cloud environment. Many technical solutions using data

encryption are applied: Data encryption affects the performance of big data processing thus researchers are still working for further improvements to this issue.
- Data transmission: While transferring large-scale data to the cloud, the capacity of the network bandwidth is the main obstacle. Over the years many algorithmic proposals and improvements are being applied to minimize cloud upload time, however, this process still remains a major research challenge.
- Data volume: The exponential growth of data to the exabyte range raises lots of concerns for the big data storage and processing in the cloud environment. Due to limited network bandwidth, cloud computing is not suitable for exabyte data processing. Exascale computing is a major research challenge.
- Data storage and processing: The current issue with NoSQL databases is that they do not offer a declarative query language similar to SQL and there is no single, unified model of NoSQL databases. Integrated solutions that combine relational and procedural queries for better performance are still being researched.

6 Conclusion

In this paper we identified the key technologies and open research issues of big data management via cloud computing. We reviewed 40 publications in order to address the recent researches with respect to frameworks, databases, data processing techniques, and programming languages. We found that framework component is mostly researched. There are various framework solutions for big data in cloud but the best fit depends on the data to process and time requirements. Regarding databases many NoSQL integrated solutions are researched and proposed. Functional Programming (FP) is considered to be the most adequate programming paradigm for big data implementations.

We conclude that big data management via cloud computing has still open research issues related to the data transfer, data volume, data storage, data security and data privacy. All of these aspects makes cloud-based big data management a viable research field.

Within this paper we provide a reference guide for researchers and developers, to determine the best emerging technologies for implementing big data as a service in cloud computing.

References

1. Mallika, C., Selvamuthukumaran, S.: Hadoop framework: analyzes workload predicition of data from cloud computing. In: 2017 International Conference on IoT and Application (ICIOT), pp. 1–6. IEEE (2017)
2. Nodarakis, N., Sioutas, S., Tsakalidis, A., Tzima, G.: Using Hadoop for Large Scale Analysis on Twitter: A Technical Report. arXiv preprint arXiv:1602.01248 (2016)
3. Meng, S., Dou, W., Zhang, X., Chen, J.: KASR: a keyword-aware service recommendation method on MapReduce for big data applications. IEEE Trans. Parallel Distrib. Syst. **25**(12), 3221–3231 (2014)

4. Bhimani, J., Yang, Z., Leeser, M., Mi, N.: Accelerating big data applications using lightweight virtualization framework on enterprise cloud. In: High Performance Extreme Computing Conference (HPEC), pp. 1–7. IEEE (2017)
5. Ortiz, J.L.R., Oneto, L., Anguita, D.: Big data analytics in the cloud: spark on hadoop vs MPI/OpenMP on Beowulf. Procedia Comput. Sci. **53**, 121–130 (2015)
6. Zhaoa, J., Wang, L., Tao, J., Chen, J.: A security framework in G-Hadoop for big data computing across distributed Cloud data centres. J. Comput. Syst. Sci. **80**(5), 994–1007 (2014)
7. Huang, T., Lan, L., Fang, X., An, P., Min, J., Wang, F.: Promises and challenges of big data computing in health sciences. Big Data Res. **2**(1), 2–11 (2015)
8. Miller, J., Bowman, C., Harish, V., Quinn, S.: Open source big data analytics frameworks written in scala. In: 2016 IEEE International Congress on Big Data (BigData Congress), pp. 389–393 (2016)
9. Totoni, E., Anderson, T., Shpeisman, T.: HPAT: High Performance Analytics with Scripting Ease-of-Use. arXiv preprint arXiv:1611.04934 (2016)
10. Khan, Z., Anjum, A., Soomro, K., Tahir, M.A.: Towards cloud based big data analytics for smart future cities. J. Cloud Comput. **4**(1), 2 (2015)
11. Xhafa, F., Naranjo, V., Caballé, S.: Processing and analytics of big data streams with Yahoo! S4. In: 2015 IEEE 29th International Conference on Advanced Information Networking and Applications (AINA), pp. 263–270 (2015)
12. Baek, J., Vu, Q., Liu, J., Huang, X., Xiang, Y.: A secure cloud computing based framework for big data information management of smart grid. IEEE Trans. Cloud Comput. **3**(2), 233–244 (2015)
13. Chandarana, P., Vijayalakshmi, M.: Big data analytics frameworks. In: Proceedings of the International Conference on Circuits, pp. 430–434. IEEE (2014). ISBN: 978-1-4799-2494-3
14. Singh, D., Reddy, C.K.: A survey on platforms for big data analytics. J. Big Data **2**(1), 8 (2015)
15. Koliopoulos, A., Yiapanis, P., Tekiner, F., Nenadic, G., Keane, J.: A parallel distributed weka framework for big data mining using spark. In: 2015 IEEE International Congress Big Data (BigData Congress), pp. 9–16 (2015)
16. Zicari, R., Rosselli, M., Korfiatis, N.: Setting up a big data project: challenges, opportunities, technologies and optimization. In: Studies in Big Data, vol. 18, pp. 17–47. Springer (2016)
17. Sharma, S., Tim, U.S., Wong, J., Gadia, S.: A brief review on leading big data models. Data Sci. J. **13**, 138–157 (2014)
18. Matallah, H., Belalem, G.: Experimental comparative study of NoSQL databases: HBASE versus MongoDB by YCSB. Comput. Syst. Sci. Eng. **32**(4), 307–317 (2017)
19. Dede, E., Sendir, B., Kuzlu, P., Weachock, J., Govindaraju, M., Ramakrishan, L.: Processing Cassandra datasets with Hadoop-streaming based approaches. IEEE Trans. Serv. Comput. **9** (1), 46–58 (2016)
20. Ptiček, M., Vrdoljak, B.: MapReduce research on warehousing of big data. In: Mipro 2017 (2017)
21. Zhang, H., Chen, G., Ooi, B.C., Tan, K.L.: In-memory big data management and processing: a survey. IEEE Trans. Knowl. Data Eng. **27**(7), 1920–1948 (2015)
22. Oussous, A., Benjelloun, F.Z., Lahcen, A.A., Belfkih, S.: Big data technologies: a survey. J. King Saud Univ.-Comput. Inf. Sci. (2017)
23. Peng, S., Liu, R., Wang, F.: New Research on Key Technologies of Unstructured Data Cloud Storage. Francis Academic Press, UK (2017)

24. Dehdouh, K., Bentayeb, F., Boussaid, O., Kabachi, N.: Using the column oriented NoSQL model for implementing big data warehouses. In: Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA), The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), p. 469 (2015)
25. Sharma, S.: An extended classification and comparison of NoSQL big data models. arXiv preprint arXiv:1509.08035 (2015)
26. Chang, B.R., Tsai, H.F., Chen, C.Y., Huang, C.F., Hsu, H.T.: Implementation of secondary index on cloud computing NoSQL database in big data environment. Sci. Program. 19 (2015)
27. Sitalakshmi Venkatraman, K.F., Kaspi, S., Venkatraman, R.: SQL versus NoSQL Movement with Big Data Analytics (2016)
28. Santos, M.Y., Costa, C.: Data warehousing in big data: from multidimensional to tabular data models. In: Proceedings of the Ninth International C* Conference on Computer Science and Software Engineering, pp. 51–60. ACM (2016)
29. Armbrust, M., Xin, R.S., Lian, C., Huai, Y., Liu, D., Bradley, J.K., Meng, X., Kaftan, T., Franklin, M.J., Ghodsi, A., Zaharia, M.: Spark SQL: relational data processing in spark. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pp. 1383–1394 (2015)
30. Siddiqui, T., Alkadri, M., Khan, N.A.: Review of programming languages and tools for big data analytics. Int. J. Adv. Res. Comput. Sci. **8**(5) (2017)
31. Wu, D., Sakr, S., Zhu, L.: Big data programming models. In: Handbook of Big Data Technologies, pp. 31–63. Springer (2017)
32. Dobre, C., Xhafa, F.: Parallel programming paradigms and frameworks in big data era. Int. J. Parallel Prog. **42**(5), 710–738 (2014)
33. Jackson, J.C., Vijayakumar, V., Quadir, M.A., Bharathi, C.: Survey on programming models and environments for cluster, cloud, and grid computing that defends big data. Procedia Comput. Sci. **50**, 517–523 (2015)
34. Nystrom, N.: A scala framework for supercompilation. In: Proceedings of the 8th ACM SIGPLAN International Symposium on Scala, pp. 18–28, October 2017
35. Edelman, A.: Julia: a fresh approach to parallel programming. In: 2015 IEEE International Conference on Parallel and Distributed Processing Symposium (IPDPS), p. 517 (2015)
36. Oancea, B., Dragoescu, R.M.: Integrating R and hadoop for big data analysis. arXiv preprint arXiv:1407.4908 (2014)
37. Maas, M., Asanović, K., Kubiatowicz, J.: Return of the runtimes: rethinking the language runtime system for the cloud 3.0 era. In: Proceedings of the 16th Workshop on Hot Topics in Operating Systems, pp. 138–143, May 2017
38. Cuzzocrea, A., Buyya, R., Passanisi, V., Pilato, G.: MapReduce-based algorithms for managing big RDF graphs: state-of-the-art analysis, paradigms, and future directions. In: Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, pp. 898–905 (2017)
39. James Stephen, J., Savvides, S., Seidel, R., Eugster, P.: Program analysis for secure big data processing. In: Proceedings of the 29th ACM/IEEE International Conference on Automated Software Engineering, pp. 277–288 (2014)
40. Fernandez, R.C., Garefalakis, P., Pietzuch, P.: Java2SDG: stateful big data processing for the masses. In: 2016 IEEE 32nd International Conference Data Engineering (ICDE), pp. 1390–1393 (2016)
41. The 2017 Top Programming Languages, IEEE Spectrum ranking. https://spectrum.ieee.org/computing/software/the-2017-top-programming-languages. Accessed 27 Oct 2017