International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 4 Issue 7, July-2015

REVIEW: Big Data on Cloud Computing

Akram Roshdi[1], Mahboubeh Shamsi[2]

[1]Departmentof Engineering, Khoy branch, Islamic Azad University, Khoy, IRAN

[2]Department of Engineering, Qom University of Technology, Qom, IRAN

ABSTRACT

Today, the world has become closer due to the development of Internet. More people communicate via Internet, and the volume of data to be handled also grows. Nowadays, we talk about peta- and zettabytes of data and this volume of data needs to be processed and analyzed further which had led to the research field of big data storage and analysis. Cloud computing is another emerging area in which the services such as infrastructure, storage, and software are provided to the consumers on demand basis. In this paper, we discuss about the big data, cloud computing, and how big data are handled in cloud computing environment. Furthermore, The definition, characteristics, and classification of big data along with some discussions on cloud computing are introduced. The relationship between big data and cloud computing, big data storage systems, and Hadoop technology are also discussed. Lastly, research challenges are investigated, with focus on scalability, availability, data integrity, data transformation, data quality, data heterogeneity, privacy, legal and regulatory issues, and governance. Lastly, open research issues that require substantial research efforts are summarized.

Keywords: Big data, Cloud computing, Classification, Hadoop, Scalability.

1. INTRODUCTION

In this Internet era, the volume of data is growing. For example, social networking sites such as Facebook, Twitter, and LinkedIn produce voluminous data every day. This paves the way to analyze and manage these large amount of data in the fields such as information network analysis, semantic Web analysis, bioinformatics data analysis, and multimedia data analysis. The management of these large amounts of data usually tera to petabytes of data leads to the research of big data analysis. In the year 2013, Gartner's Inc. Hype Cycle states that cloud computing and big data are the fastest growing technologies which dominate the research world for the next decade [1]. In this paper, the overview of big data management technologies and handling of the big data in cloud computing environment are discussed. In Sect. 2, definitions of big data given by various experts are discussed. Section 3 discusses about cloud computing definition, benefits, and challenges. Section 4 gives an overview of various big data management systems. section 5 discusses about relationship between Cloud computing and big data**.** Section 6 discusses about Research chalenge.

2. DEFINITION AND CHARACTERISTICS OF BIG DATA

Big data means bulky volume of data with different dimensions. Gartner defines big data as high volume, high velocity, and high variety information and value assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. The volume of data is very large, and it cannot be handled by traditional database systems. The data are growing at an exponential rate as we talk about tera and petabytes of data in today's Internet world. The Internet produces a variety of information, i.e., some data are structured, some are semi-structured, and some are unstructured [2]. IBM states that all unstructured data collected from various applications such as social media sites, digital pictures and videos, and climatic information are big data [3]. Chang et al. [4] define big data as datasets which cannot be handled and processed by the current technology within a tolerable elapsed time. This 4V (fig 1) definition is widely recognized because it highlights the meaning and necessity of big data.

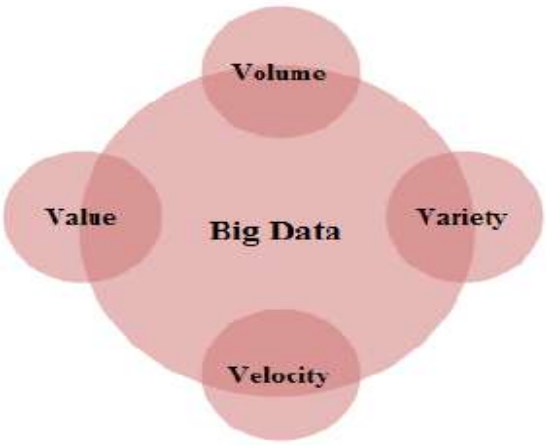

Fig1. Four vs of big data

2.1. Classification of big data

Big data are classified into different categories to better understand their characteristics. Fig. 2 shows the numer- ous categories of big data. The classification is important because of large-scale data in the cloud. The classification is based on five aspects: (i) data sources, (ii) content format, (iii) data stores, (iv) data staging, and (v) data processing. Data sources include internet data, sensing and all stores of transnational information, ranges from unstructured to highly structured are stored in various formats. Most popular is the relational database that come in a large number of varieties [8]. As the result of the wide variety of data srouces, the captured data differ in size with respect to redundancy, consistency and noise, etc.

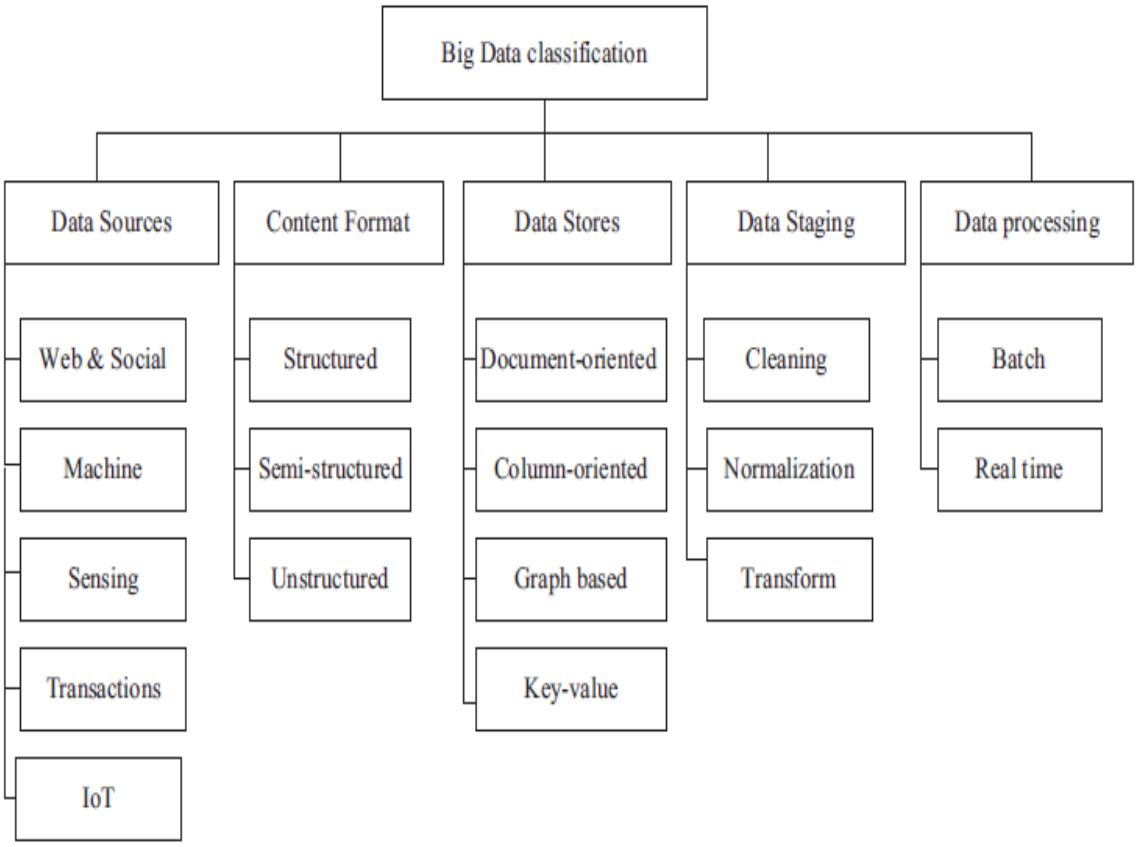

Fig. 2. Big data classification

3. CLOUD COMPUTING

Cloud computing is nothing but providing software and hardware services over the Internet and managed by third parties [5]. The cloud computing model allows access to information and computer resources from anywhere that a network connection is available. US National Institute of Standards and Technology (NIST) defines cloud computing as Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics (On demand self service, Broad network access, Resource pooling, Rapid Elasticity, Measured Service), three service models (Infrastructure as a Service, Platform as a Service, Software as a Service), and four deployment models (Private Cloud, Community Cloud, Public Cloud, Hybrid Cloud). [6]

A. Benefits and Challenges of Cloud Computing

According to Forbes, the benefits of cloud computing [7] are Cost saving, elasticity, scalability, load "bursting" and storage on demand and the drivers for cloud are Pay—as—you go, Multi-tenancy, Elasticity of resources—Allows Ease of Implementation, On-Demand self-service, Reduces carbon Footprint.

The challenges in the cloud computing environment are assurance of privacy and security, reliability and availability, transition and execution risk, cultural resistance, regulatory ambiguity, especially internationally, costs associated with a migration from legacy infrastructure to the cloud, and issues of taxation [7].

4. BIG DATA MANAGEMENT IN CLOUD

Traditionally, to store data, relational databases are used. As big data involve voluminous, heterogeneous, and continuously varying data, the traditional relational databases are not suitable to handle the big data. Various data management systems are available to handle the big data in cloud. Table 1 provides the list of big data management systems.

4.1 Programming Models

MapReduce. [10] is a simplified programming model for processing large numbers of datasets pioneered by Google for dataintensive applications. The MapReduce model was developed based on GFS [14] and is adopted through open-source Hadoop implementation, which was popularized by Yahoo. Apart from the MapReduce framework, several other current open-source Apache projects are related to the Hadoop ecosystem, including Hive, Hbase, Mahout, Pig, Zookeeper, Spark, and Avro. Twister [15] provides support for efficient and iterative MapReduce computations. MapReduce allows an unexperienced programmer to develop parallel programs and create a program capable of using computers in a cloud. In most cases, programmers are required to specify two functions only: the map function (mapper) and the reduce function (reducer) commonly utilized in functional programming. The mapper regards the key/value pair as input and generates intermediate key/value pairs. The reducer merges all the pairs associated with the same (intermediate) key and then generates an output. Table 8 summarizes the process of the map/reduce function.

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

HDFS [13] is a distributed file system designed to run on top of the local file systems of the cluster nodes and store extremely large files suitable for streaming data access. HDFS is highly fault tolerant and can scale up from a single server to thousands of machines, each offering local computation and storage. HDFS consists of two types of nodes, namely, a namenode called "master" and several datanodes called "slaves." HDFS can also include secondary namenodes. The namenode manages the hierarchy of file systems and director namespace (i.e., metadata). File systems are presented in a form of namenode that registers attributes, such as access time, modification, permission, and disk space quotas. The file content is split into large blocks, and each block of the file is independently replicated across datanodes for redundancy and to periodically send a report of all existing blocks to the namenode.

HaLoop. HaLoop is developed by making certain modifications in the Hadoop MapReduce framework. HaLoop programming model [11] is developed to include iterative processing in the applications as HadoopDB lacks support

for iterative computations. HaLoop uses the distributed file system and task queue structure of HadoopDB but has introduced a new application programming interface, new loop control module, a task scheduler and caches, and indexes of application data on the slave node. The task scheduler and task tracker are modified to incorporate iterative programming applications.

Table 1: List of big data management systems

Big data management system	Description
Google: Bigtable	Bigtable is a distributed storage system used by Google products and projects such as Google search engine, Orkut, Google docs, and Google Maps. Bigtable can store petabytes of data across thousands of servers [8]
AppEngine Datastore	It is a data storage system provided by Google for Web applications. AppEngine uses a language called GQL. It provides a Python interface [9]
Yahoo!: PNUTS/ Sherpa	PNUTS is a database system for yahoo Web applications. It is renamed to Sherpa. It servers the data for Web applications [10]
Amazon: Dynamo	This is a data storage system built to support Amazon's internal applications. It highly available reliable, scalable, and distributed key-/value-based database system [11]
Amazon S3	Amazon simple storage service (S3) is an object-based online public storage service
Amazon simple DB	Simple DB is designed for structured data storage
Amazon RDS	Amazon relational database service is provided to work with MySql database
Microsoft: DRYAD	DRYAD is designed for data parallel applications used internally by Microsoft. The high-level language used is DryadLINQ. Supports T-SQL, ODBC, and ADO.NET data access [12]
SQL Azure	It is a cloud-based relational database service built on Microsoft SQL technologies. Highly available, scalable, and multi-tenant database service hosted in the cloud [13]
Open source project Cassandra	The Apache Cassandra is an decentralized open source database which provides high availability, scalability, and fault tolerance [14]
Hypertable	Hypertable is an open source database system which runs on the top of the distributed file systems [15]
CouchDB	CouchDB is a document-based database storage system. Documents can be accessed using Web browser. JavaScript is used to query and manipulate the documents [16]

5. THE DIALECTICAL RELATIONSHIP BETWEEN CLOUD COMPUTING AND BIG DATA

Cloud computing has development greatly since 2007. Cloud computing's core model is large-scale distributed computing, providing computing, storage, networking, and other resources to many users in service mode, and users can use them whenever they need them [17]. Cloud computing offers enterprises and users high scalability, high availability, and high reliability. It can improve resource utilization efficiency and can reduce the cost of business information construction, investment, and maintenance. As the public Cloud services from Amazon, Google, and Microsoft become more sophisticated and better developed, more and more companies are migrating toward the Cloud computing platform. Cloud computing and big data are complementary, forming a dialectical relationship.

Cloud computing and the Internet of Things' widespread application is people's ultimate vision, and the rapid increase in big data is a thorny problem that is encountered during development. The former is a dream of humanity's pursuit of civilization, the latter is the bottleneck to be solved in social development. Cloud computing is a trend in technology development, while big data is an inevitable phenomenon of the rapid development of a modern information society.

To solve big data problems, we need modern means and Cloud computing technologies. The breakthrough of big data technologies can not only solve the practical problems, but can also make Cloud computing and the Internet of Things' technologies land on the ground and be promoted and applied in in-depth ways.[9]

6. RESEARCH CHALLENGES

Although cloud computing has been broadly accepted by many organizations, research on big data in the cloud remains in its early stages. Several existing issues have not been fully addressed. Moreover, new challenges continue to emerge from applications by organization. In the subsequent sections, some of the key research challenges, such as scalability, availability, data integrity, data transformation, data quality, data heterogeneity, privacy and legal issues, and regulatory governance, are discussed.

6.1 Scalability

Scalability is the ability of the storage to handle increasing amounts of data in an appropriate manner. Scalable distributed data storage systems have been a critical part of cloud computing infrastructures [19]. The lack of cloud computing features to support RDBMSs associated with enterprise solutions has made RDBMSs less attractive for the deployment of large-scale applications in the cloud. This drawback has resulted in the popularity of NoSQL [18].
A NoSQL database, also called ''Not Only SQL'', provides the mechanism to store and retrieve large volumes of distributed data. The features of NoSQL databases include schema-free, easy replication support, simple API, and consistent and flexible modes. Different types of NoSQL databases, such as key-value [20], column-oriented, and document-oriented, provide support for big data. Table 2 shows a comparison of various NoSQL database technologies that provide support for large datasets[9].

Table 2: Comparison of NoSQL Databases

Feature/ Capability	NoSQL database name									
	DynamoDB	Redis	Voldemort	Cassandra	Hbase	MangoDB	SimpleDB	CouchDB	BigTable	Apache Jackrabbit
Storage type	KV	KV	KV	KV	KV	Doc	Doc & KV	Doc	CO	Doc
Initial release	2012	2009	2009	2008	2010	2009	2007	2005	2005	2010
Consistency	N/A	✔	N/A	N/A	✔	✔	N/A	N/A	✔	✔
Partition Tolerance	N/A	✔	✔	✔	✔	✔	N/A	✔	✔	N/A
Persistence	✔	✔	✔	✔	✔	✔	✔	N/A	✔	✔
High Availability	✔	✔	✔	✔	✔	✔	✔	N/A	✔	✔
Durability	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Scalability	High	High	High	High	High	High	High	High	High	High
Performance	High	High	High	High	High	High	High	High	High	High
Schema-free	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Programming Language	Java	Ansi-C	Java	Java	Java	C++	Erlang	Erlang	C, C++	Java

6.2 Availability

Availability refers to the resources of the system accessible on demand by an authorized individual [8]. In a cloud environment, one of the main issues concerning cloud service providers is the availability of the data stored in the cloud. For example, one of the pressing demands on cloud service providers is to effectively serve the needs of the mobile user who requires single or multiple data within a short amount of time. Therefore, services must remain operational even in the case of a security breach [8]. In addition, with the increasing number of cloud users, cloud

service providers must address the issue of making the requested data available to users to deliver high-quality services. Lee et al. [8] introduced a multi cloud model called "rain clouds" to support bigdata exploitation. "Rain clouds" involves cooperation among single clouds to provide accessible resources in an emergency. Schroeck et al. [8] predicted that the demand for more real time access to data may continue to increase as business models evolve and organizations invest in technologies required for streaming data and smartphones.

6.3 Data integrity

A key aspect of big data security is integrity. Integrity means that data can be modified only by authorized parties or the data owner to prevent misuse. The proliferation of cloud-based applications provides users the opportunity to store and manage their data in cloud data centers. Such applications must ensure data integrity. However, one of the main challenges that must be addressed is to ensure the correctness of user data in the cloud. Given that users may not be physically able to access the data, the cloud should provide a mechanism for the user to check whether the data is maintained [8].

6.4 Transformation

Transforming data into a form suitable for analysis is an obstacle in the adoption of big data [8]. Owing to the variety of data formats, big data can be transformed into an analysis workflow in two ways as shown in Fig. 4. In the case of structured data, the data is pre-processed before they are stored in relational databases to meet the constraints of schema-on-write. The data can then be retrieved for analysis. However, in unstructured data, the data must first be stored in distributed databases, such as HBase, before they are processed for analysis. Unstructured data are retrieved from distributed databases after meeting the schema-on-read constraints.

6.5 Data quality

In the past, data processing was typically performed on clean datasets from well-known and limited sources. Therefore, the results were accurate [8]. However, with the emergence of big data, data originate from many different sources; not all of these sources are well-known or verifiable. Poor data quality has become a serious problem for many cloud service providers because data are often collected from different sources. For example, huge amounts of data are generated from smartphones, where inconsistent data formats can be produced as a result of heterogeneous sources. The data quality problem is usually defined as "any difficulty encountered along one or more quality dimensions that render data completely or largely unfit for use" [8]. Therefore, obtaining high-quality data from vast collections of data sources is a challenge. High-quality data in the cloud is characterized by data consistency. If data from new sources are consistent with data from other sources, then the new data are of high quality [8].

6.6 Heterogeneity

Variety, one of the major aspects of big data characterization, is the result of the growth of virtually unlimited different sources of data. This growth leads to the heterogeneous nature of big data. Data from multiple sources are generally of different types and representation forms and significantly interconnected; they have incompatible for- mats and are inconsistently represented [8]. In a cloud environment, users can store data in structured, semi-structured, or unstructured format. Structured data formats are appropriate for today's database systems, whereas semi-structured data formats are appropriate only to some extent. Unstructured data are inappropriate [9] because they have a complex format that is difficult to represent in rows and columns. According to Kocarev and Jakimoski [9], the challenge is how to handle multiple data sources and types.

6.7 Privacy

Privacy concerns continue to hamper users who out- source their private data into the cloud storage. This concern has become serious with the development of big data mining and analytics, which require personal information to produce relevant results, such as personalized and location-based services [9]. Information on individuals is exposed to scrutiny, a condition that gives rise to concerns on profiling, stealing, and loss of control [9].

6.8 Legal/regulatory issues

Specific laws and regulations must be established to preserve the personal and sensitive information of users. Different countries have different laws and regulations to achieve data privacy and protection. In several countries, monitoring of company staff communications is not allowed. However, electronic monitoring is permitted under special circumstances [9]. Therefore, the question is whether such laws and regulations offer adequate protection for individuals' data while enjoying the many benefits of big data in the society at large [8].

6.9 Governance

Data governance embodies the exercise of control and authority over data-related rules of law, transparency, and accountabilities of individuals and information systems to achieve business objectives [9]. The key issues of big data in cloud governance pertain to applications that consume massive amounts of data streamed from external sources [8]. Therefore, a clear and acceptable data policy with regard to the type of data that need to be stored, how quickly an

individual needs to access the data, and how to access the data must be defined [9]. Big data governance involves leveraging information by aligning the objectives of multiple functions, such as telecommunication carriers having access to vast troves of customer information in the form of call detail records and marketing seeking to monetize this information by selling it to third parties [9]. Moreover, big data provides significant opportunities to service providers by making information more valuable. However, policies, principles, and frameworks that strike a stability between risk and value in the face of increasing data size and deliver better and faster data management technology can create huge challenges [8]. Cloud governance recommends the use of various policies together with different models of constraints that limit access to underlying resources. Therefore, adopting governance practices that maintain a balance between risk expo- sure and value creation is a new organizational imperative to unlock competitive advantages and maximize value from the application of big data in the cloud [8].

7. CONCLUSION

In this data age, the volume of data is growing at a faster pace because of the increase in the computational capability. The Internet has become the source of information for the whole world. The social networking sites, for example, Facebook, twitter, and LinkedIn, have users in millions and more and new users gets added frequently and the data go on accumulated. The traditional relational database management systems are not suitable for handling this voluminous data. This has led to the research of handling and management of this big data. Cloud computing is another emerging filed and is widely used because of its pay-as-you-go model and economic advantage. In this paper, we discussed about big data, cloud computing, cloud computing models, benefits, challenges, and the big data management in cloud computing environments. The challenges of big data management in cloud are providing privacy, security, reliability, and availability which necessitate further research in big data management in cloud computing environments.

8. ACKNOWLEDGMENT

I'd like to thank Dr. Mahboubeh Shamsi for her support during this research paper. Any errors are my own, however!

REFERENCES

[1]. D. Laney, "The importance of Big Data: A definition", (2008).
[2]. P. Devi, T. Gaba, "Cloud computing", Int. J. Adv. Res. Comput. Sci. Softw. Eng. 3(5) (2013).
[3]. P. Mell, T. Grance, "Definition of cloud computing", Technical report(NIST) (2009).
[4]. F. Chang, J. Dean, S. Ghemawat, W. Hsieh, D. Wallach, "Bigtable: A distributed storage system for structured data", ACM Trans. Comput. Syst. 26(2), 4 (2008).
[5]. B. Cooper, R. Ramakrishnan, U. Srivastava, "Pnuts: Yahoo!'s hosted data serving platform", PVLDB 1(2), (2008), 1277–1288.
[6]. G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati. "Dynamo: Amazon's highly available key-value store", in SOSP (2007), pp. 205–220.
[7]. M. Isard, M. Budiu, Y. Yu, A. Birrell, "Dryad: distributed data-parallel programs from sequential building blocks", in EuroSys (2007), pp. 59–72.
[8]. I. Abaker, T. Hashem." The rise of big data" on cloud computing: Review and open research issues", Information Systems 47(2015),98–115.
[9]. N. Badrul Anuar, A. Gani," The rise of "Big Data" on cloud computing", 24 July 2014, Available: www.elsevier.com/locate/infosys.
[10]. J. Dean, S. Ghemawat, "MapReduce: simplified data processing on large clusters", Communications of the ACM, 51(2008), 107-113.
[11]. K.S. Sangeetha , P. Prakash ,"Big Data and Cloud: A Survey", Artificial Intelligence and Evolutionary Algorithms in Engineering Systems, Advances in Intelligent Systems and Computing 325, DOI 10.1007/978-81-322-2135-7_81, Springer India 2015.
[12]. T. White, "Hadoop: The Definitive Guide: The Definitive Guide", O'Reilly Media, 2009.
[13]. K. Shvachko, K. Hairong, S. Radia, R. Chansler," The Hadoop Distributed File System, in: Mass Storage Systems and Technologies (MSST)", IEEE 26th Symposium on, 2010, pp. 1-10.
[14]. S. Ghemawat, H. Gobioff, "S.-T. Leung, The Google file system, in: ACM SIGOPS Operating Systems Review", ACM,2003, pp. 29-43.
[15]. J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S.-H. Bae, J. Qiu, G. Fox, "Twister: a runtime for iterative mapreduce, in: Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing", ACM, 2010, pp. 810-818.
[16]. S. Fiedler "Big Data Technologies and Cloud Computing", Optimized Cloud Resource Management and Scheduling. DOI: http://dx.doi.org/10.1016/B978-0-12-801476-9.00002-1, Elsevier, 2015 .

[17]. [17]. I .Foster, Y .Zhao, I .Raicu, "Cloud computing and grid computing 360-degree compared. Grid computing environments workshop", IEEE, 2008, p. 1_10.
[18]. [18]. R. Cattell, "Scalable SQL and NoSQL data stores", ACM SIGMOD Record, 39 (2011) 12-27.
[19]. [19]. P. Mell, T. Grance," The NIST definition of cloud computing (draft)", NIST special publication, 800 (2011) .
[20]. [20]. S. Das, D. Agrawal, A. El Abbadi, "G-store: a scalable data store for transactional multi key access in the cloud, in: Proceedings of the 1st ACM symposium on Cloud computing", ACM, 2010, pp. 163-174.