Journal of Energy and Power Engineering 10 (2016) 405-410
doi: 10.17265/1934-8975/2016.07.004

Big Data; Definition and Challenges

Shirin Abbasi

Computer Department, Islamic Azad University-Tehran Center Branch, Tehran 13185-768, Iran

Received: April 26, 2016/ Accepted: May 09, 2016 / Published: July 31, 2016.

Abstract: In recent years, due to the widespread use of electronic services and the use of social network as well, large volumes of information are being made that this information contains various types of things such as videos, photos, texts etc. besides large volume. Due to the high volume and the lack of specificity of this information, covering them through traditional and relational databases is not possible and modern solutions should be used for processing them, so that processing speed is also covered. Data storage for processing and the way of accessing to them in memory, network communication, covering required features for distributed system in solutions that are in use for storing big data, are the items that should be covered. In this paper, a collection of advantages and challenges of big data, special features and characteristics of them has been provided and with the introduction of technologies in use, storage methods are studied and research opportunities to continue the way will be introduced.

Key words: Big data, cloud computing, hadoop, the analysis of big data.

1. Introduction

Fremont Rider, in an article about the future of the library of Yale University, predicted that on the basis of annual increase of research sources, in the year 2040, two million books will be available that if they want to be kept on paper , their shelves, it will cover a distance of six miles. The amount of data that are generated and processed is increasing day by day.

Big data refer to a set of data that are stored in the shape of structured or unstructured, and they are complex data that are composed of various aspects. The first feature of big data is their volume that returns to their quantity and due to high volume, managing and analyzing them is different and it can not be performed through traditional data bases. If a small number of processing nodes are used, due to the high volume, processing will be done with a lower speed. To increase processing speed, more nodes and also stronger process power are needed that demand higher cost. One of the primary ways that is recommended in this field in data compressing. This method does not work in big data, because one of the other features of big data is their diversity. Big data contain different types of videos, photos, texts etc., and because of not being structured their compression is difficult and in some cases the same time that is wasted for processing them through traditional methods is wasted for compression and in addition due to this type diversity, it has its own complexity. For this reason, compression is not applied in big data processing. The other thing that should be considered in data processing is that these data are used in applications that transform the data online or they should be analyzed in a situation which they response people in a defined time. So in the process, timing is such that accountability be done with no delay. Traditional methods of data management are not accountable for big data management. In big data management, all cases among different data structures, different data aspects and the lack of structures of them should be considered.

In big data management special cases should be considered:

(1) If parallel processing is used, the nodes in the cluster must act in such a way that in case of failure of

Corresponding author: Shirin Abbasi, M.Sc. student, research fields: software engineering.

any particular node, no disturbance happened in big data processing.

(2) Files of the systems that are used for big data should be such that they can cover the big volume of big data and their capisity should be in the GB or higher.

(3) Read and write operation in many applications are running in succession and they should be optimized to meet the real time mode and better response time.

(4) Some of the operations for better response should be transferred to applications and in order to escape from a compatible hardware and communication needs, it is better to use cloud environments for covering this case.

When it comes to big data, cloud computing also comes largely due to the same concerns in these both cases. In both of them hardware and software sources are in the same network and physical bases for users have been simulated by the aim of that on the base of the demand of the people, seemingly infinite space, be availble for them. Cloud databases are used for appropriate allocation of resources in big data, and on the other hand, parameters such as scalability and availability are also a challenge for cloud environments.

In this article, earlier works in the field of big data will be mentioned at first and then we will discuss about the main features in this data and challenges ahead and at the end, the assessment of the future will be mentioned.

2. Prior Works

During the 1970s, the basic concepts of database were made. At first, they were used for data storage and then with the spread of concepts of relational databases to data analysis, methods of searching and querying of data have been spread as well, as a result the possibility of data processing has provided. When over time, the volume of data increased, there was no possibility to use on computer for processing. In the early 1980s, the possibility of using multiple systems and sharing data provided and each system has its own processor, memory and independently performs processing. Considering the increase of volume of linked data and the need of processing them simultaneously, different solutions have been considered for big data. When Google faced the problem of lack of memory and the problems of analyzing large amount of web pages, they developed files of the systems in using and benefited distributed analysis model. Together with this parallel processing, Google designed a highly scalable database that was called Big Table. This type of database was able to index and tag information, because of this, access to information was more rapidly taking place and according to using in Google's products and services, it was introduced as a starting point for data processing and after development in cloud environments, it was used in big data processing and query field as a standard. Google technology was not an open source technology and because of this reason, Yahoo proceeded to develop these databases on the base of open source formats like hadoop that data bases like HBase and Hive have been proposed on this base.

In managing big data, there are various approaches like the use of Big Table, the use of non-relational data management or entity data management and character creation on the base of different aspects of data.

Also in the current cloud storage services such as Amazon, they benefit of the ways that on the base of them, the reliability of stored data is very high and these services are used for distributed systems. One of these services is Dyamo. By considering three main features-data consistency, accessibility and flexibility and on the base of CAP (consistency, availability, partition tolerance) theory that believes software systems would cover two of these three features in any situations, relational and comparative models only support integrity and accessibility. Key-based models only cover accessibility and division, Big Table-based models cover integration and division and document-based models cover accessibility and division.

3. The Definition of Big Data

Big data are some data with high volume that cover a combination of structured and unstructured data and processing them with traditional databases methods is not possible, and because of this reason for managing them, special techniques are used. Three special features of big data that are as a benchmark to identify big data are used. These three features are known to 3V and contain quantity, type and speed of process that we will define them in big data features.

4. Characteristics of Big Data

The volume of data: the amount of data in the data collection of big data is high. This amount is one of the features that is identified as a main feature for big data. As it was mentioned in the previous sections, in today's world the amount of data is increasing and this should be considered in data processing. Because in many cases refining and filtering the information is needed and also the ways of accessing and storing the information should be personalized on the base of this amount.

The speed of process: speed of creation, stream, processing and aggregation of information should be appropriate to the characteristics of today's data groups. According to the speed of information generation in today's world and need of proactive response in many of applications and social networks the speed of processing should be in a way that it could perform appropriate to this features. On the other hand, because of big data is usually kept distributed, communication and memory access should also be taken in consideration.

Various types of data: data that place in big data groups, including different type of data like photos, texts, videos and so on, that are obtained from various sources. They have different formats and it is difficult to categorize them, also because of this big data are called non-structured, they can not be defined with a particular format or structure.

Data value: due to the huge volume of data, several examples are provided to assess that considering this high volume if a problem happens, all the data can be re-refined and chose, and it works when there is a problem in results of the evaluation. You should also consider that the amount of data in data sets are trusted for how long and its results are correct. This is very important in big data, because many of decisions and programs, in the field of various industries, are done based on this data's processing.

Due to big data distribution becomes very important, different parts of a data set that are placed on different servers, the accuracy and integrity of the data should be integrated and the latest updated version of that also be exist in the parts that may be copied on different servers.

5. Useful Tools in Analysis of Big Data

The tools that are used for data analysis are developing day by day and be able to help data collection and analysis on the base of different kind of data, and in big data special tools are used because of their special features that are different from traditional tools based on relationship of the components.

5.1 Parallel Tools of Hadoop

There are many technologies that have been raised in the context of massive data processing, but hadoop is one of the most famous. Hadoop is an open source platform that is used for different types of data processing and storage that helps data base industries for fast access to their hidden values and proccessing and exploring in them. The main features of these tools are as follows:

- This tool is an open source and therefore its resources, libraries and functions are easily accessible.
- Its layers and components act independently and are not integrated.
- Access to external files is supported.
- When system is high-load, hadoop breaks the operation of one task into several groups. And therefore planning for works that needs several operation groups to be done easier.

• Establishment of automatic balance of load in each distributed groups increases when data traffic increases.

• Support for replacement of cars and nodes when a problem exists in this tool has a layered architecture. There is a record-based memorial layer in the lowest layer and this data set is managed by rows and columns and in each machine in distributed parts there is a memory manager that manages system's internal memory. The middle lower is an implementational workflow layer that includes relational operators for doing operations on data set. In the lowest level of hadoop software, there is a distributed file that briefly called HDFS (hadoop distributed file system). Each file is departmentalized and located in sequence of addressable and continuous memory and processed in batches. In the middle software layer, files have been devided and each part of process is working for on node and finally results are also collected from nodes and converted to the final output. This devision and sum up is done on the base of map-reduce function.

5.1.1 Hadoop Distributed File System

As the way in the hadoop architecture part this file system was mentioned, we call it HDFS briefly and this part includes the main memory system in hadoop. When information is interred, these file systems will devide them into smaller parts and distribute them in different servers that are located in system as a node, each server keeps just a small part of the main information and processes them and each of these parts has been copied on several servers for being bearable against error. Central node that has the responsibilty of supervision and devision and collection of results is called Name and different nodes that maintain the information are called Informational node or Data. The middle layer that has been mentioned in architecture part is used for HDFS's performance optimization for better data communication.

5.1.2 Mapping Function-Decrease

An architectural model has been distributed in computational systems for parallel processing. In this model, data have been devided into smaller parts and each process is also broken to smaller instructions and different nodes in distributed systems manage part of operations on the base of these devisions. This issue causes the increasment of processing power. The first part of this algorithm, mapping, is used for data devision and allocates a subset of instructions and tasks to each of the computational nodes. At this step of algorithm, the input data are read at first and a set of middle records are provided and labeled for computations and these records are distributed among computational nodes on the base of the use of hash functions and each of the nodes starts their own process separately and each of them transfers their own provided result as the output to the central node. At this time the second step which is the decreasement will start. This function collects the results and will provide the main result on the base of the correct format of output.

5.2 Nosql Database

It is one of the data management systems that can be used in cloud environments and for non-relational data. In this system, data are not recognized on the base of their relations. Various types of database can be different but the common point of them is that they can have a big effect on big data storage. Also in these systems, there is a solution that in the case of increase in some data, data be stored in just one place and in the other situation, processing on them through accessing to them be available .

6. The Challenges in Big Data

Due to the lack of structure of data, data filtering and setting different access levels for users are very difficult and expensive. To solve this problem, usually information labeling is used.

Production of information about the data in big data needs the use of special technologies, because the data must be interpreted and due to the high volume of data

and also various structures and different databases, determination of criteria, value and the way of information obtaining is difficult.

Heterogeneity in hardware resources: Due to big data are maintain distributed, different hardware formats in the side of servers could cause problem that by the use of common supply resources, this level of compatibility is established to a high level.

Because these data are not structured, benchmarking to measure them, and appropriate extraction of different categories that work in decisions that are related to business is difficult and on the base of different points of view, different factors could be considered.

Big data management is done in various forms that issues such as determining the level of access and sources management should be considered due to sources difference and system distribution.

Various algorithms are used in this field. Information is located on various systems after blocking therefore hardware and software resources are more available.

Due to the large volume of data, refining in a particular subject, is time consuming. This issue should be measured that if the data are the same data that will be used for decision or not and also we should be sure that data amount of them is still available and reliability for this amount exists. In addition, it should be considered in data refining that how much data are needed to predict the possibilities ahead.

The speed of data processing also should be considered in big data. Because as it is mentioned in big data features, processing speed is one of the factors that big data is known and measured through that.

Because of big data aggregation from multiple sources, the determination of that who is responsible for accuracy and correctness of information and also the determination of how should the levels of access be determined and who is the owner of information, is challenging.

Reduce of redundancy and compression: In general, there is a high level of data redundancy in big data collections and by the aim of reduce indirect cost, it is necessary that, this redundancy, decline relatively without damage to the main information.

Analysis mechanism in big data should be in the way that analysis non-structured data and also be able to response in a limited time. The solutions that are used in traditional and relational data bases can not be developed, on the other hand in non-relational databases the topic of function is raised. For this reason combinatorial solutions of these two issues should be used.

7. Conclusions

Considering the increasing of data volume and the importance of them in different fields, big data management requires its own specific issues. The only challenge of big data management is not its high volume, but different types of data and the speed of data processing are very important issues that are even considered in big data definition. In this article, the other raised challenges in big data management have been taken into account and communicational, networking and storing issues have been investigated and features, advantages and disadvantages of big data have been extracted as the paper's output.

Big data management is an important issue that through proper function in it, needed knowledge in different areas can be extracted. This article can be a starting point for extensive research in the field of big data.

References

[1] Hristidis, V., Chen, S., Li, T., Luis, S., and Deng, Y. 2010. "Survey of Data Management and Analysis in Disaster Situations." *Journal of Systems and Software* 83 (10): 1701-14.

[2] Grolinger, K. 2013. "Disaster Data Management in Cloud Environments." Ph.D. thesis, The University of Western Ontario.

[3] Kossmann, D., and Kraska, T. 2010. "Data Management in the Cloud: Promises, State-of the Art, and Open

Questions." *Datenbank-Spektrum* 10 (3): 121-9.

[4] Dewan, H., and Hansdah, R. 2011. "A Survey of Cloud Storage Facilities." In *Proceedings of the IEEE World Congress,* 224-31.

[5] Wolf Halton. 2010. "Security Issues and Solutions in Cloud Computing." Wolf Halton. Accessed August 2010. http://wolfhalton.info/2010/06/25/security-issues-and-solutions-in-cloud-computing.

[6] Chung, F., and Dean, J. 2008. "Big Table: A Distributed Storage System For Structured Data." *ACM Transactions on Computer Systems* 26 (2): 4-24.

[7] Guo, W., Qiao, C., and Jin, Y. 2010. "Demonstration of Joint Resource Scheduling in an Optical Network Integrated Computing Environment." *IEEE Communications Magazine* 48 (5): 76-83.

[8] Vogels, W. 2009. "Eventually Consistent." *Communications of the ACM Rural Engineering Development* 52 (1): 40-4.

[9] Zadrozny, P., and Kodali, R. 2013. *Big Data Analytics Using Splunk*. CA: Berkeley.

[10] Decandia, G., Hastorun, D., and Jampani, M. 2007. "Dynamo: Amazon Highly Available Ley-Value Store." In *Proceedings of the 21st ACM Symposium on Operating Systems Principle*s, 205-20.

[11] Grolinger, K., Higashino, W., and Tiwari1, A. 2013. "Data Management in Cloud Environments: NoSQL and NewSQL Data Stores." *Journal of Cloud Computing* 2 (December): 2-24.

[12] Minelli, M., Chambers, M., and Dhiraj, A. 2013. *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses.* Chichester: John Wiley & Sons.