2015

Big Data Technologies: Additional Features or Replacement for Traditional Data Management Systems?

Dhouha Jemal
University of Tunis, Tunisia, dh.jemal@gmail.com

Rim Faiz
LARODEC, IHEC of Carthage, Tunisia, Rim.Faiz@ihec.rnu.tn

Sami Mahfoudhi
University of Tunis, Tunisia, smahfoudhi@yahoo.com

BIG DATA TECHNOLOGIES: ADDITIONAL FEATURES OR REPLACEMENT FOR TRADITIONAL DATA MANAGEMENT SYSTEMS?

Completed Research

Jemal, Dhouha, University of Tunis, Tunis, Tunisia, dh.jemal@gmail.com

Faiz, Rim, University of Carthage, Tunis, Tunisia, rim.faiz @ ihec.rnu.tn

Mahfoudhi, Sami, University of Tunis, Tunis, Tunisia, smahfoudhi@yahoo.com

Abstract

With the data volume that does not stop growing and the multitude of sources that led to diversity of structures, the classic tools of data management became unsuitable for processing and unable to offer effective tools for information retrieval and knowledge management. Thereby, a major challenge has become how to deal with the explosion of data to transform it into new useful and interesting knowledge. Despite the rapid development and change of the databases world, this data management systems diversity presents a difficulty in choosing the best solution to analyze, interpret and manage data according to the user's needs while preserving data availability. Hence, the imposition of the Big Data in our techno-logical landscape offers new solutions for data processing. In this work, we aim to present a brief of the current buzz research field called Big Data. Then, we provide a broad comparison of two data management technologies.

Keywords: Big Data, MapReduce, RDBMS.

1 Introduction

The data is growing at an alarming speed in both volume and structure. With the voluminous data which does not stop growing and the multitude of sources which led to diversity of structures, relational databases which have been proven for over 40 years have reached their limits (Ordonez, 2013), and the classic tools of data management became unsuitable for processing and unable to offer effective tools to find information within massive data and extract value from it. Thereby, as described in (Cuzzocrea *et al*, 2013), the problem of dealing with the explosion of data has become our major challenge.

As the world becomes more information-driven than ever before, a new technological field has emerged in order to cope with the new requirements for data analysis, information retrieval and knowledge management: the Big Data, as presented in (Sagiroglu and Sinanc, 2013) and (Narasimhan and Bhuvaneshwari, 2014), aims to provide an alternative to traditional solutions database and analysis. Thus, MapReduce (Dean and Ghemawat, 2008) is presented as one of the most efficient Big Data solutions. This framework has found great success in analyzing and processing large amounts of data on large clusters.

Several studies have been conducted to compare MapReduce and Relational DBMS. MapReduce has been presented as a replacement for the Parallel Database Management Systems. However, as proposed in (Stonebraker *et al*, 2010), MapReduce can be seen as a complement to a RDBMS for analytical applications, because different problems require complex analysis capabilities provided by both technologies.

In this environment of data explosion and diversity, the question arises what technology to use for data analysis and information retrieval, how to benefit the data management systems diversity?

The remainder of this paper is organized as follows. In section 2, we give an overview of the definition and characteristics of big data, presenting some research works that focuse on this actual research trend. In section 3, we present the Big Data challenges. Then, we aim to provide a broad comparison of two data management technologies, presenting for each one its strengths and its weaknesses. In section 4, we introduce MapReduce, then in section 5, we describe Relational DBMS. Finally, section 6 concludes this paper and outlines our future work.

2 The Big Data era

Big data presents the next frontier for innovation, competition, and productivity. In this context, we are going to present in this section, the main research work conducted in the Big Data field and to introduce the Big Data characteristics and applications.

2.1 Big Data research work

A strong interest towards the term Big Data is arising in the literature actually. Many research works focuse on this actual research trends in the field. In this context, Sagiroglu and Sinanc in (2013), present an overview of big data's content, scope, samples, methods, advantages and challenges, details big data's main component and discusses privacy concern on it. Then, Narasimhan and Bhuvaneshwari in (2014), provides a brief of the buzz-field called Big Data and cover the components of big data from a Hadoop perspective. This study aims to highlight the field's caracteristics with two additional dimensions, and to provide a thorough understanding of big data and its various components in the Hadoop framework.

In (Cuzzocrea *et al*., 2011), open problems and actual research trends are highlighted with the aim of providing an overview of state-of-the-art research issues and achievements in the field of analytics over big data, and extend the discussion to analytics over big multidimensional data. This work presents several novel research directions a rising in this field, which plays a leading role in next-generation Data Warehousing and OLAP research. In (Cuzzocrea *et al*., 2013), open problems in the field of Data Warehousing and OLAP over Big Data are highlighted. This work aims to present challenges to adopt Data Warehousing and OLAP methodologies with the goal of collecting, extracting, transforming, loading,

warehousing and OLAPing such kinds of data sets, by adding significant add-ons supporting analytic over Big Data.

2.2 Definition, characteristics and applications

The term Big Data, explained by Narasimhan and Bhuvaneshwari, in (2014), was raised the first time by the Gartner office in 2008. It refers to the explosion of data volume and new technological capabilities offered to answer it. Big data can be defined as "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 strong interest towards the term Big Data is arising in the literature actually. Though scalable data management has been a vision for more than three decades and much research has focused on large scale data management in traditional enterprise setting, Big Data brings its own set of novel challenges that must be addressed to ensure the success of data management solutions. This field has received much attention from across the computing and research community, and a lot of work has been done in this context such as (Dean and Ghemawat, 2008; Cohen *et al*., 2009; Agrawal *et al*., 2011; Wang and Chan, 2013). The goals of the big data solutions are to meet the new challenges of treating very important volume of structured and unstructured data, located on various terminals.

As presented in (Cuzzocrea *et al*., 2013), Big Data repositories have two intrinsic factors: (i) size, which becomes really explosive in such data sets; (ii) complexity which can be very high in such data sets. Every day, the amount of data created and manipulated is increasing. All the sectors of activity are affected by this phenomenon. This exponential growth is due to several factors such as trends in the number of users of IT (Information Technology) solutions and data generation by machines. These masses of data bring larger and finer opportunities for analysis as well as new uses of the information. Data today comes from multiple sources such as business transactions and social networks, and in all types of formats: Structured numeric data in traditional databases, and unstructured such as text documents, email, video, audio and financial transactions. Managing, merging and governing both explosion amount and different varieties of data is something many organizations still grapple with.

The web and social networks, whether they are open to all or developed in a professional context, provide kind of opportunities for big data. As described in (Sagiroglu and Sinanc, 2013), McKinsey Global Institute in (Manyika *et al*., 2011), specified the potential of big data in five main topics:

- Healthcare (clinical decision support systems, analyze disease patterns, improve public health).
- Public sector (discover needs, decision making with automated systems to de-crease risks, innovating new products and services).
- Retail (in store behavior analysis, variety and price optimization, product placement design, web based markets).
- Manufacturing (developed production operations, supply chain planning).
- Personal location data (smart routing, geo targeted advertising or emergency response).

In this context, Big Data approach aims to provide an alternative to traditional solutions database and analysis. Big data solutions add some features to classics DBMSs in order to satisfy new data management needs in a new ecosystem of explosive volume of structured and unstructured data. Actual research trends in the field of Data Warehousing and OLAP over Big Data are rising, such as (Cuzzocrea *et al*., 2013). The next section presents the general architecture of a data-warehouse.

3 Big Data challenges

Nowadays, Big data and its analysis are at the center of modern science and business. It requires a revolutionary step forward from traditional data analysis. As mentioned in (Sagiroglu and Sinanc, 2013),

the term Big Data is for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. Big Data, as presented in (Sagiroglu and Sinanc, 2013), is characterized by three main components, the 3 V's: Volume, Velocity and Variety.

1. Volume. It is the first feature brought by the term "big". The size, which can be-come a real bottleneck from practical applications, refers to the vast amounts of data generated every second. The volume of data stored today is booming. This amount of data that is being collected daily presents immediate challenges for businesses. We can Just think of social media messages going viral in seconds, the speed at which credit card transactions are checked for fraudulent activities, or the milliseconds it takes trading systems to analyze social media networks to pick up signals that trigger decisions to buy or sell shares.

2. Velocity. Speed of data in and out; describes the frequency at which data are generated, captured and shared. Growing flows of data must be analyzed in real time to meet the needs of chrono-sensitive processes. Reacting fast enough and analyzing the streaming data is troubling to businesses, with speeds and peak periods of-ten inconsistent. Big Data approach opens the possibility to integrate data streams and generate results or data visualization in (almost) real time.

3. Variety. The volume of Big Data puts data centers in front of challenge: the variety of data. It's not traditional relational data, this data is raw, semi structured or un-structured. In fact, 80% of the world's data is now unstructured1, and therefore can't easily be put into tables (think of photos, video sequences or social media updates). Big Data is in the form of structured and unstructured data. The structured data types are ready for insertion into a database, while unstructured types have an implicit and irregular structure, and not a fixed pattern (non-relational). With big data technology we can now harness differed types of data (structured and unstructured) including messages, social media conversations, photos, video or voice recordings and bring them together with more traditional, structured data.

In (Narasimhan and Bhuvaneshwari, 2014), we consider two additional dimensions when thinking about big data:

4. Veracity. Accuracy of collected data is a key feature. As mentioned in (Cuzzocrea *et al*., 2011), very often, data sources, storing data of interest for the target analytic processes, such as web and social networks are strongly heterogeneous and incongruent. Big Data becomes bigger and the multiple sources of big data are ever increasing. So, build confidence in the Big Data represents a significant challenge due to the possibility of inconsistency and abnormality in the Data. Very large data volumes and multiple heterogeneous sources amplify the need for rigor in the collection and crossing data to remove data uncertainty to build confidence and ensure the secu-rity and integrity of data.

5. Value. Big Data is gradually transforming organizations around the valuation of information. Big Data approach is designed to achieve the strategic objectives of value creation f or the company. With the Big Data approach, the non interesting data when it is taken apart,, can take a meaning when considered globally. A big data strategy gives businesses the capability to better analyze data with a goal of accelerating portable growth. Having access to big data, companies generate value from data.

Big data and its analysis are at the center of modern science and business. Inspired by this main motivation, (Cuzzocrea *et al*., 2011) present a number of open problems and actual research trends related to big data analytics, such as: The data Source Heterogeneity and Incongruence, Filtering-Out Uncorrelated Data, Strongly Unstructured Nature of Data Sources, High Scalability. In this context, a main challenge that has interested the research community and has been the subject of several works such as (Abouzeid *et al*., 2009; Gruska and Martin, 2010) is: Combining the Benefits of RDBMS and NoSQL Database Systems. It is one of the more relevant features to be achieved by big data analytic systems. As discussed in (Cattell, 2011), it is necessary to combine the benefits of traditional RDBMS database systems and those of the new generation of NoSQL database systems in order to obtain the critical flexibility feature which refers to the property of covering a large collection of analytic scenarios over the same big data partition.

The question is not any more " can Big Data become a relevant competitive ad-vantage? ", but "How can we exploit the opportunities offered by these solutions to optimize our analysis and decision making process? "

4 MapReduce

MapReduce is a programming model developed by Google, which was introduced by Dean and Ghemawat (2008). It was designed for processing large data sets with a parallel, distributed algorithm on a cluster.

MapReduce was created in order to simplify parallel processing and distributed da-ta on a large number of machines with an abstraction that hides the details of the hardware layer to programmers: it hides the details of parallelization, fault-tolerance, locality optimization, and load balancing. Google uses the MapReduce model to deploy large variety of problems such as: generation of data for Google's production web search service, data mining, machine learning, etc.

The MapReduce programming model has been successfully used for many different purposes. These included: parallelizing the effort; distributing the data; handling node failures.

The term MapReduce actually refers to two separate and distinct tasks: Map and Reduce. The mapper is responsible for reading the data stored on disk and process them; it takes a set of data and converts it into another set of data: reads the input block and converts each record into a Key/Value pair. The reducer is responsible for consolidating the results from the map and then write them to disk; it takes the out-put from a map as input and combines those data tuples into a smaller set of tuples.

At first, Google developed their own DFS: the Google File System (GFS). As de-scribed in (McClean *et al*., 2013), MapReduce tasks run on top of Distributed File Systems (DFS). The distributed storage infrastructure store very large volumes of data on a large number of machines, and manipulate a distributed file system as if it were a single hard drive. The DFS deals with data in blocks. In order to prevent data loss, each block will be replicated across several machines to overcome a possible problem of a single machine failure. So, this model allows the user to focus on solving and implementing his problem.

Nevertheless, the lack is that the MapReduce is independent of the storage system, it can not take into account all the input data for an available index. This explains the critics mainly from the database community. As described in (Gruska and Martin, 2010), the data-base community sees the MapReduce as a step backwards from modern database systems, in view of the MapReduce is a very brute force approach and it lacks the optimizing and indexing capabilities of modern database systems.

MapReduce, the powerful tool characterized by its performance for heavy pro-cessing to be performed on a large volume of data that it can be a solution to have the best performance hence makes it very popular with companies that have large data processing centers such as Amazon and Facebook, and implemented in a num-ber of places. However, Hadoop, the Apache Software Foundation open source and Java-based implementation of the MapReduce framework, has attracted the most interest. Firstly, this is due to the open source nature of the project, additionally to the strong support from Yahoo. Hadoop has its own extensible, and portable file system: Hadoop Distributed File System (HDFS) that provides high-throughput access to application data.

Since it is introduced by Google, a strong interest towards the MapReduce model is arising. Many research works aim to apply the ideas from multi-query optimization to optimize the processing of multiple jobs on the MapReduce paradigm by avoiding redundant computation in the MapReduce framework. In this direction, MRShare (Nykiel *et al*., 2010) has proposed two sharing techniques for a batch of jobs. The key idea behind this work is a grouping technique to merge multiple jobs that can benefit from the sharing opportunities into a single job. However, MRShare incurs a higher sorting cost compared to the naive technique . In (Wang and Chan, 2013), two new job sharing techniques are proposed: The generalize d grouping technique (GGT) that relaxes MRShare's requirement for sharing map output. The second technique is a materialization tech-nique (MT) that partially materializes the map output of jobs in the map and reduce phase.

On the other hand, the Pig project at Yahoo (Olston *et al*., 2008), the SCOPE project at Microsoft (Chaiken *et al*., 2008), and the open source Hive project 2 introduce SQL-style declarative languages over the standard MapReduce model, aim to integrate declarative query constructs from the database community into MapReduce to allow greater data independence.

5 Relational DBMS

Since it was developed by Edgar Codd in 1970, as presented in (Shuxin and Indrakshi, 2005), the relational database (RDBMS) has been the dominant model for database management. RDBMS is the basis for SQL, and is a type of database management system (DBMS) that is based on the relational model which stores data in the form of related tables, and manages and queries structured data. Since the RDBMSs focuse on extending the database system's capabilities and its processing abilities, RDBMSs have become a predominant powerful choice for the storage of information in new databases be-cause they are easier to understand and use. What makes it powerful, is that it is based on relation between data; because the possibility of viewing the database in many different ways since the RDBMS require few assumptions about how data is related or how it will be extracted from the database. So, an important feature of relational systems is that a single database can be spread across several tables which might be related by common database table columns. RDBMS also provide relation-al operators to manipulate the data stored into the database tables. However, as discussed in (Hammes *et al*., 2014), the lack of the RDBMS model resides in the complexity and the time spent to design and normalize an efficient database. This is due to the sever-al design steps and rules, which must be properly applied such as Primary Keys, Foreign Keys, Normal Forms, Data Types, etc. Relational Databases have about forty years of production experience, so the main strength to point out is the maturity of RDBMSs. That ensure that most trails have been explored and functionality optimized. For the user side, he must have the competence of a database designer to effectively normalize and organize the database, plus a database administrator to maintain the inevitable technical issues that will arise after deployment.

A lot of work has been done to compare the MapReduce model with parallel relational databases, such as (Pavlo *et al*., 2009), where experiments are conducted to compare Hadoop MapReduce with two parallel DBMSs in order to evaluate both parallel DBMS and the MapReduce model in terms of performance and development complexity. The study showed that both databases did not outperformed Hadoop for user-defined function. Many applications are difficult to express in SQL, hence the remedy of the user-defined function. Thus, the efficiency of the RDBMSs is in regular database tasks, but the user-defined function presents the main ability lack of this DBMS type.

A proof of improvement of the RDBMS model comes with the introduction of the Object-Oriented Database Relational Model (ORDBMS). It aims to utilize the benefits of object oriented theory in order to satisfy the need for a more programmatic flexibility. The basic goal presented in (Sabàu, 2007), for the Object-relational database is to bridge the gap between relational databases and the object-oriented modeling techniques used in programming languages. The most notable research project in this field is Postgres (Berkeley University, Californie); Illustra and PostgreSQL are the two products tracing this research.

6 Conclusion

Day after day, zillions of data are generated all over the universe. Many factors contribute to the increase in data volume. The implications of the rise of data generation challenge the needs of data processing, which explains the technological development and diversity of the proposed data management solutions and explosive growth, both in the number of products and services offered and in the adoption of data analysis technologies.

There has been a significant amount of work during the last two decades related to the needs of new supporting technologies for data processing and knowledge management, challenged by the rise of data generation and data structure diversity.

In this paper, we try to detail challenges of the Big Data field. We have presented an overview of big data's characteristics and applications. Although this work has not resolved the entire subject about this trend topic. For this reason, we plan to involve a detailed study on challenges and issues with big data, and to focus on the use cases of Big Data solutions.

References

Abouzeid, A. Pawlikowski, K. Abadi, D. Silberschatz, A. Rasin, A. (2009). "Hadoopdb: An architectural hybrid of mapreduce and dbms technologies for analytical workloads." In: *Proceedings of the VLDB Endowment*, p. 922-933.

Agrawal, D. Das, S. and El Abbadi, A. (2011). "Big Data and Cloud Computing: Current State and Future Opportunities." In: *Proceedings of the 14th International Conference on Extending Database Technology*, ACM, p. 530-533.

Cattell, R. (2011). "Scalable SQL and NoSQL Data Stores." *ACM SIGMOD Record 39 (4), 12-27*.

Chaiken, R. Jenkins, B. Larson, P.A. Ramsey, B. Shakib, D. Weaver, S. and Zhou, J. (2008). "Scope: Easy and efficient parallel processing of massive data sets." In: *Proceedings of the VLDB Endowment 1 (2)*, p. 1265-1276.

Cohen, J. Dolan, B. Dunlap, M. Hellerstein, J.M. and Welton, C. (2009). "MAD Skills: New Analysis Practices for Big Data." In: *Proceedings of the VLDB Endowment 2 (2)*, p. 1481-1492.

Cuzzocrea, A. Bellatreche, L. and Song, I.Y. (2013). "Data Warehousing and OLAP over Big Data: Current Challenges and Future Research Directions." In: *Proceedings of the sixteenth international workshop on Data warehousing and OLAP* , ACM, p. 67-70.

Cuzzocrea, A. Song, I.Y. and Davis, K.C. (2011). "Analytics over Large-Scale Multidimensional Data: The Big Data Revolution!." In: *Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP*, ACM, p. 101-104.

Dean, J. and Ghemawat, S. (2008). "Mapreduce : Simplified data processing on large clusters." *Communications of the ACM 51 (1), 107-113*.

Gruska, N. and Martin, P. (2010). "Integrating MapReduce and RDBMSs." In: *Proceedings of the 2010 Conference of the Center for Advanced Studies on Collaborative Research*, IBM Corp., p. 212-223.

Hammes, D. Medero, H. and Mitchell, H. (2014). "Comparison of NoSQL and SQL Databases in the Cloud." *Southern Association for Information Systems (SAIS) Proceedings, 12*.

Manyika, J. Chui, M. Brown, B. Bughin, J. Dobbs, R. Roxburgh, C. and Byers, A.H. (2011). "Big data: The next frontier for innovation, competition, and productivity." McKinsey Global Institute.

McClean, A. Conceição, R.C. and O'Halloran, M. (2013). "A Comparison of MapReduce and Parallel Database Management Systems." In: *ICONS 2013, The Eighth International Conference on Systems*, p. 64-68.

Narasimhan, R. and Bhuvaneshwari, T. (2014). "Big Data A Brief Study." *International Journal of Scienti c & Engineering Research 5 (9)*.

Nykiel, T. Potamias, M. Mishra, C. Kollios, G. and Koudas, N. (2010). "Mrshare: sharing across multiple queries in mapreduce." In: *Proceedings of the VLDB Endowment, vol. 3 (12)*, p. 494-505.

Olston, C. Reed, B. Srivastava, U. Kumar, R. and Tomkins, A. (2008). "Pig latin: a not-so-foreign language for data processing." In: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, ACM, p. 1099-1110.

Ordonez, C. (2013). "Can we analyze big data inside a DBMS?." In: *Proceedings of the sixteenth international workshop on Data warehousing and OLAP*, ACM, p. 85-92.

Pavlo, A. Rasin, A. Madden, S. Stonebraker, M. DeWitt, D. Paulson, E. Shrinivas, L. and Abadi, D.J. (2009). "A comparison of approaches to large scale data analysis." In: *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, ACM, p. 165-178.

Sabàu, G. (2007). "Comparison of RDBMS, OODBMS and ORDBMS." *Informatica Economic*.

Sagiroglu, S. and Sinanc, D. (2013). "Big Data: A Review." In: *Collaboration Technologies and Systems (CTS), 2013 International Conference on. IEEE*, p. 42-47.

Shuxin, Y. and Indrakshi, R. 2005. "Relational data-base operations modeling with UML." In: *Proceedings of the 19th International Conference on Advanced Information Networking and Applications*, p. 927-932.

Stonebraker, M. Abadi, D. DeWitt, D.J. Madden, S. Paulson, E. Pavlo, A. and Rasin, A. (2010). "Mapreduce and parallel dbmss : friends or foes ?." *Communications of the ACM 53 (1), 64-71*.

Wang, G. and Chan, C.Y. (2013). "Multi-Query Optimization in MapReduce Framework." In: *Proceedings of the VLDB Endowment, 40th International Conference on Very Large Data Bases 7 (3)*.