# Big Data Components for Business Process Optimization

Mircea Răducu TRIFU, Mihaela-Laura IVAN
Bucharest University of Economic Studies, Bucharest, Romania
trifumirceraradu@yahoo.com, ivanmihaela88@gmail.com

*In these days, more and more people talk about Big Data, Hadoop, noSQL and so on, but very few technical people have the necessary expertise and knowledge to work with those concepts and technologies. The present issue explains one of the concept that stand behind two of those keywords, and this is the map reduce concept. MapReduce model is the one that makes the Big Data and Hadoop so powerful, fast, and diverse for business process optimization. MapReduce is a programming model with an implementation built to process and generate large data sets. In addition, it is presented the benefits of integrating Hadoop in the context of Business Intelligence and Data Warehousing applications. The concepts and technologies behind big data let organizations to reach a variety of objectives. Like other new information technologies, the main important objective of big data technology is to bring dramatic cost reduction.*

***Keywords****: Big Data, Hadoop, MapReduce Model, Business Intelligence, Business Analytics*

## 1 Introduction

In this paper is explained one of the concepts and technologies behind the Big Data, the Map Reduce, the ways that Hadoop data can be interrogated and then used. In Business Intelligence there are some forever issues with data computation, data transformation and analysis speed. Once the explosion in amount of data appeared, predictions and data mining are not separate disciplines. Therefore, customers need to be able to go beyond simple reports and see new ways to understand the data and detect trends and opportunities. When thinking of the advantages big data can bring, there are a lot, but only two of them are the more important. One of them is regarding the financial benefits and outcome and the second one is about the entire process flow, starting from the organization part to delivering part.

Some organizations researching big data says that terabyte storage of structured data is currently most cheaply provided with big data technologies such as Hadoop clusters. For example for a company with the cost of storing one terabyte for a year was $37,000 for a traditional relational database, $5,000 for a database appliance, and $2,000 for a Hadoop cluster. Of course, these figures cannot be directly comparable, because the more traditional technologies may be somehow more safe and easily administrated [1].

Big Data is a concept that promise to help in all that areas, using the three V's, volume, velocity and variety.

- *Volume*: big data is that "Ocean of data" that we mentioned about in the rows above. It is represented by information that can came from every possible sensor, and some even say that we people are also sensors and data gatherers for *big data*. [9] The challenges of having such a big quantity of data is that is very hard to sustain it, to store it, to analyze it and ultimately to use it.
- *Velocity:* is all about the speed of data traveling from one point to another and the speed of processing it. Sometimes it is crucial for the manager to be able to decide in a very little time on a variety of issues [2]. The most important issue is that the resources that analyses data is limited compared to the *volume* of data, but the requests of information is unlimited and usually information gets through at least one bottleneck.
- *Variety*, the third characteristic is represented by the types of data that are stored. Because there are many types of sensors and sources, the data that came from them is vary very much in size and type. It is very complicated to analyze text, images and sounds in the same context and get a result that can be relied

DOI: 10.12948/issn14531305/20.1.2016.07

on. And then is the issue of dark data, data that sits in the organization and is unused and also is not free.

- There are one new dimension that were added to the existing ones: *Veracity*
  *Veracity* is the hardest thing to achieve with big data, because due to the *Volume* of information and the *variety* of its type is hard to identify the useful and accurate data form the “dirty data”. The biggest problem is that the “dirty data” can lead very easy to an avalanche of errors, incorrect results and can affect the *Velocity* attribute of Big Data. The main purpose of the Big Data can be corrupted and all the information can lead to a useless and very expensive Big Data environment if there is not a good cleaning team. The *Veracity* attribute is in its self also an objective for the Big Data developers. If the data cannot be accurate, is redundant or is unreliable, the whole Company can have a big problem, especial the companies that use big data to sell information like the marketing ones, or the ones that make market studies. Many social media responses to campaigns could be coming from a small number of disgruntled past employees or persons employed by competition to post negative comments.

## 2 MapReduce Model

Map reduce is a programming model of a concept which is used for generating and processing large data sets. The computation takes a set of output key/value pairs and has two levels of processing. The first stage is Map, and is written by the user. It takes an input pair and produces a set of intermediate key and transfer them to the Reduce Function. The Reduce function, is also written by the user and it takes the intermediate key and a set of values for the key, after that it merges the values and create a form much smaller of values.

There are many different implementations of Map Reduce and the right one depends on the environments, the scope is to use parallel processing as much as possible. The model is most used on multiple computers with different configurations and needs a pretty good network environment to assure parallel processing on multiple cores of multiple computers [3].

An example of the way it works can be to find out what was the maximum temperature in each city in Romania in one year. Let’s presume that wave multiple input files containing temperature values (Figure 1).

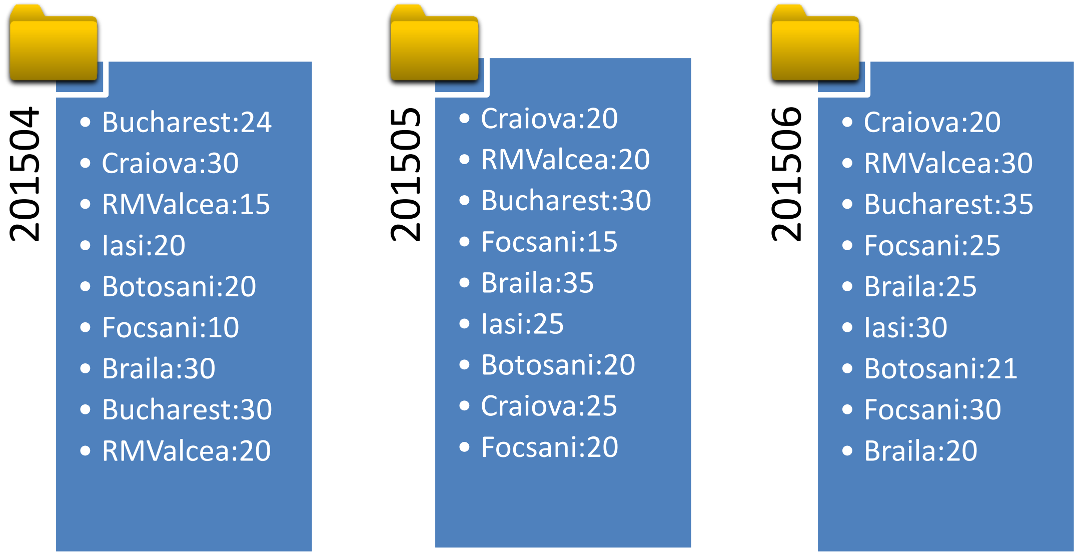


**Fig. 1.** Maximum temperature in each city in Romania

DOI: 10.12948/issn14531305/20.1.2016.07

There will be multiple mappers, and every mapper will use some input files, finds the maximum value in the files and takes the result to the reducer. Figure 2 presents the mappers for the maximum temperature values in each month.

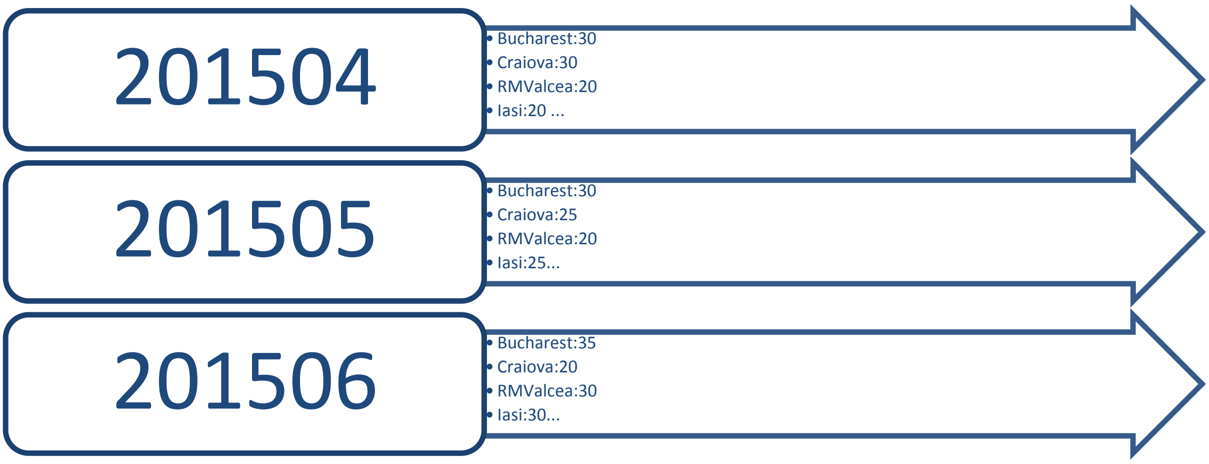


**Fig. 2.** Mappers of maximum temperature values

The reducer takes all the input from the mapper and calculates the maximum values from all the inputs from the mappers. This process is shown in Figure 3.

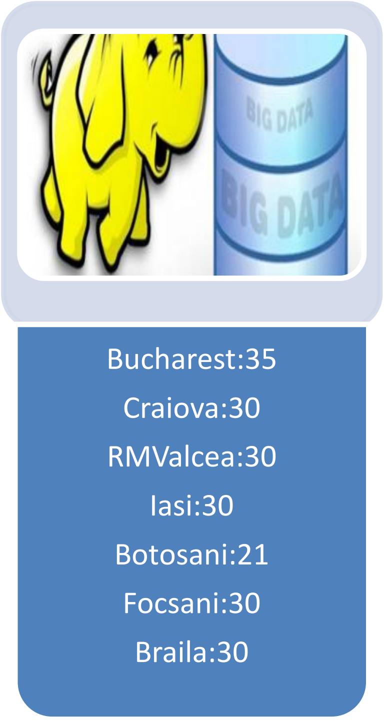


**Fig. 3.** The reducer component

Every mapper can utilize one ore multiple processing cores ore computers, and so everything is calculated in parallel. Unlike traditional databases, MapReduce does not need structured data and well-defined schemas, it uses more a semi-structured or unstructured data. The only demand is that data must be provided to the map function as a series of key value pairs, resulting in also a set of key value pairs that is taken by the reduce function that performs aggregation to collect the final set of results [4].

### 3 HDFS (Hadoop Distributed File System)

The Data in Hadoop is scatted across multiple blocks on a cluster, that way the map and reduce function are executed on smaller subsets of data, and this assure scalability of the big data.

A good example is where you have a phone book spared on multiple servers in a cluster, *Hadoop Distributed File System* can reconstruct the phone book and then replicates the smaller pieces on different servers, and in this way higher availability is achieved [5].

Here are some HDFS features:

- the HDFS blocks are much larger than the most file system files, the HDFS block is typically around 64 MB in size;
- is optimized for throughput over latency and so is very efficient at streaming read requests for large files, but is poor in seeking request on small ones;
- optimized for workloads that are write one and read multiple times type.

Both HDFS and MapReduce exhibit several architectural principles:

DOI: 10.12948/issn14531305/20.1.2016.07

- are design to run on clusters of servers;
- can scale there capacity by adding more servers;
- use mechanisms for identifying and working around failures;
- both provides many of their services transparently, so that the user can detect and fix issues;
- have a architecture where a software cluster sits on the physical servers and controls all aspects of executions [4].

HDFS and MapReduce are software clusters and share some common characteristics:

- both have an architecture where a worker note is managed by a special master note;
- the master node (NameNode in HDFS and JobTraker in MapReduce) monitors the health of the cluster and handles failures by moving data blocks or by rescheduling failed work;
- processes on each server (DataNode and TaskTracker) are responsible for performing work on the physical host, receiving instructions from the master and the reports back the health and progress status.

## 4 Integrating Hadoop into Business Intelligence and Data Warehousing

In the last 20 years different data structures and technologies have been developed to increase performance or enable a Business Intelligence capability. Hadoop has gained popularity with data storage mode in this technology evolution. It is presumed that the integration of Hadoop with Business Intelligence and Data Warehousing applications will become very used in few years.

The Hadoop Distributed File System (HDFS) is designed to store transactional data. The map-reduce processing supported by Hadoop frameworks can deliver great performance, but a week point is that cannot be run the same proficient query optimization as mature relational database technologies do. To achieve this query optimization implies to use new methods, such as query accelerators or writing code. Which means could be simple and fast to get a specific list of transactions for some dates, geography and so on, but complex calculations like aggregation of oriented calculations will require to interfere with programming skills to acquire de desired performance. Hadoop tends to be used mostly for reporting capabilities in Business Intelligence applications because of its batch oriented processing. Actually, an acceptable performance for interactive capabilities of some areas can be reached, but for ad hoc queries would not be so satisfying due to the need to interfere with settings for job processing [6].

The figure below illustrates the correlations between Hadoop and BI architecture. As can be noticed Hadoop data is situated in the data store. Also, it is represented the BI modelling effort for an operational data store (ODS), enterprise data warehouses (EDW) and online analytical processing (OLAP) data store. In fact, when is intended to increase the performance then the process to extract, transform and load (ETL) transactional data to ODS or to EDW or to some variation of OLAP is started. Actually the diagram shows the 'pieces' together.

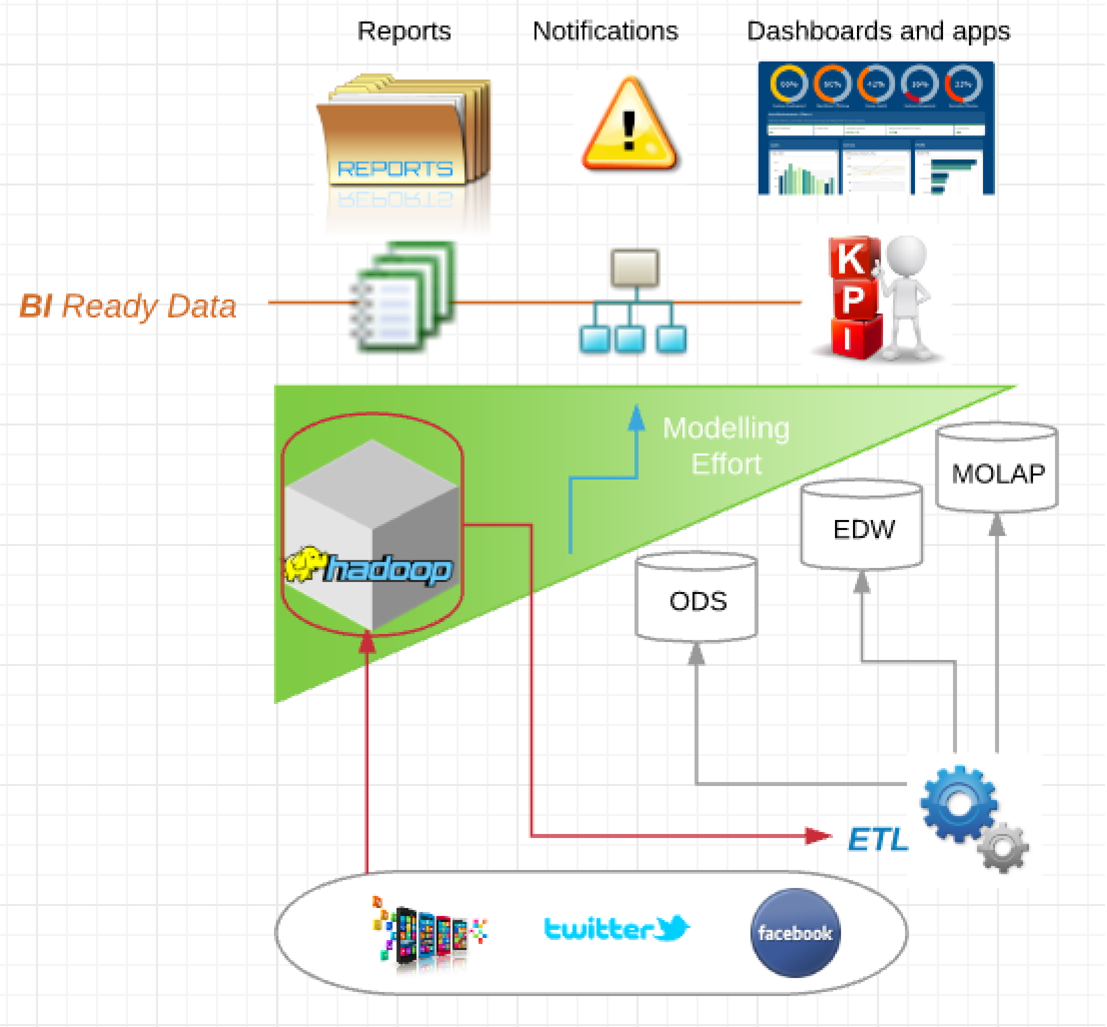


**Fig. 4.** Inclusion of Hadoop in a Business Intelligence application [6]

The above figure shows how the data warehouse can serve as a source for the big data ecosystem. Likewise, Hadoop can come with key data output that can consolidate the data warehouse for further analytics.

In reality, many organizations continue to use the already known data warehouses for traditional BI and analytics reporting. In this new environment, the data warehouse may continue with its usual workload, using data from operational systems and storing historical data to provision traditional Business Intelligence and analytics results.

The old analytics environment contains the operational systems that serve as source for data; a data warehouse and has integrated the data for a collection of analysis functions; and a set of business intelligence and analytics tools that enable decisions from the use of ad hoc queries, dashboards, and data mining.

Figure 5 presents the standard big company data warehouse system’s architecture.

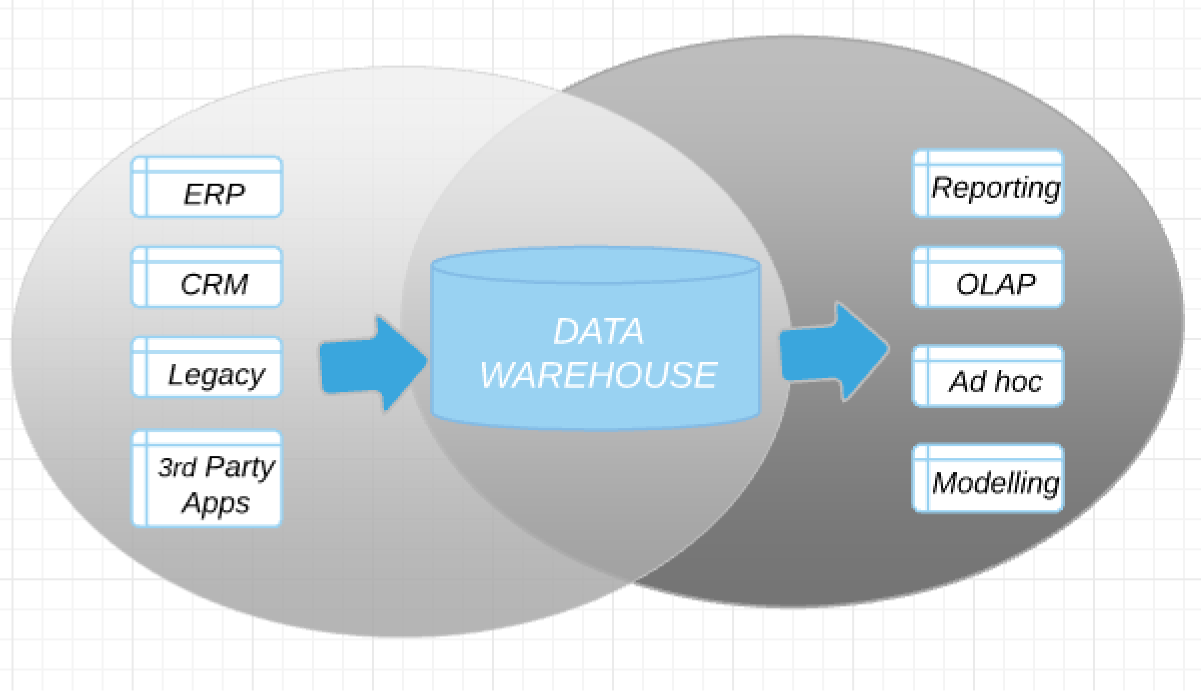


**Fig. 5.** A standard Data Warehouse environment [1]

Big companies which invested time and money for their data warehouses, are not willing to change the current environment which works as expected. Mainly choose to combine what they already have, in special analytics environments, with the new big data technologies.

The reality is what Hasso Platner said about the needs of the new applications, these will have to handle with big data. We have to analyze on the go, so we need an analytical and transactional system in the same time. We can’t have a multi-level system. This is very slow for modern applications [7].

From our point of view big data is part of our life [8]. The society makes the data creation through the new devices and their intelligent applications. Collecting data from device sources is the way of finding the consumer behaviors. In the future without new technologies designed for big data we will not survive. In this moment are more mobile devices than people on the planet, which produce a big quantity of data.

## 5 Conclusions

Every software solution has some particularities, strong and week points, and every solution is built to solve a couple of issues. Hadoop was design to manage, interrogate and use large data sets, and because of that is not the best solution to handle a data warehouse or a database.

The architecture of the Hadoop is created to be very flexible and scalable in order to process large sets of data, but there are consequences of this type of architecture. Hadoop is a batch processing system, it handles huge data sets relatively quickly but is not the type of system that is built to generate data in real-time. Hadoop is not well suited to low-latency queries or real-time system. When a Hadoop job is running, some settings must be done such as determining which tasks are run on each note or other housekeeping activities, and this affects the overall execution time. On small data sets the same tasks must be performed and this impact the execution time. In conclusion, to integrate a Hadoop infrastructure with a Business Intelligence tool it has to be found for what will be used the application. In order to find further how the new information technologies for Business Intelligence evolve, through the future researches will be analyzed other proficient components for business process optimization.

DOI: 10.12948/issn14531305/20.1.2016.07

## References

[1] T. H. Davenport, J. Dyché, “Big Data in Big Companies,” *SAS Institute Inc.*, May 2013, Available at: http://www.sas.com/resources/asset/Big-Data-in-Big-Companies.pdf

[2] L. Arthur, “What Is Big Data?,” *Forbes / CMO Network*, August 2013, Available at: http://www.forbes.com/sites/lisaarthur/2013/08/15/what-is-big-data/

[3] J. Dean, S. Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters,” *Communications of the ACM - 50th anniversary issue: 1958 – 2008*, Vol. 51, No. 1, 2008, New York, USA, pp. 107-113.

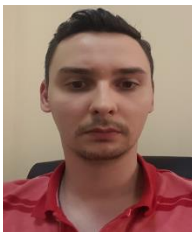
[4] G. Turkington, *Hadoop Beginner's Guide*, Packt Publishing Ltd, 2013, 398 pg.

[5] IBM, *What is the Hadoop Distributed File System (HDFS)?*, 2016, Available at: https://www-01.ibm.com/software/data/infosphere/hadoop/hdfs/

[6] T. Groves, “Where Does Hadoop Fit in a Business Intelligence Data Strategy?,” *IBM Big Data & Analytics Hub*, January 2013, Available at: http://www.ibmbigdatahub.com/blog/where-does-hadoop-fit-business-intelligence-data-strategy

[7] H. Plattner, Top 20 quotes from Hasso Plattner, SlideShare, 2016, Available at: http://www.slideshare.net/bestskills/top-20-quotes-from-hasso-plattner

[8] M. R. Trifu, M. L. Ivan, “Big Data: present and future,” Database Systems Journal, Vol. 5, No. 1, 2014, pp. 32-41.



**Mircea Răducu TRIFU** graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest University of Economic Studies in 2011 and the Informatics Security Master in 2013. He also finished the Faculty of Management in 2009, and the Master in Business Administration in 2011, both at the Bucharest University of Economic Studies. At present he is a System Support at Data warehouse team in the department of the Application Development of the ING Bucharest.

**Mihaela Laura IVAN** graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest University of Economic Studies in 2011. She also finished the Master's degree in Economic Informatics in 2013, at the Bucharest University of Economic Studies. Starting with 2013, Mihaela is a PhD candidate at Bucharest University of Economic Studies in the field of Economic Informatics. At the present, she is a SAP Development Consultant at SAP Near Shore Centre Romania.

DOI: 10.12948/issn14531305/20.1.2016.07