# Big Data Challenges

Alexandru Adrian TOLE
Romanian – American University, Bucharest, Romania
adrian.tole@yahoo.com

*The amount of data that is traveling across the internet today, not only that is large, but is complex as well. Companies, institutions, healthcare system etc., all of them use piles of data which are further used for creating reports in order to ensure continuity regarding the services that they have to offer. The process behind the results that these entities requests represents a challenge for software developers and companies that provide IT infrastructure. The challenge is how to manipulate an impressive volume of data that has to be securely delivered through the internet and reach its destination intact. This paper treats the challenges that Big Data creates.*

***Keywords***: *Big Data, 3V's, OLAP, security, privacy, sharing, value, infrastructure, technological solutions*

## 1 Introduction

Economic entities and not only, had developed over the years new and more complex methods that allows them to see market evolution, their position on the market, the efficiency of offering their services and/or products etc. For being able to accomplish that, a huge volume of data is needed in order to be mined so that can generate valuable insights.

Every year the data transmitted over the internet is growing exponentially. By the end of 2016, Cisco estimates that the annual global data traffic will reach 6.6 zettabytes[1]. The challenge will be not only to "speed up" the internet connections, but also to develop software systems that will be able to handle large data requests in optimal time.

To have a better understanding of what Big Data means, the table below represents a comparison between traditional data and Big Data (**Table 1.** Understanding Big Data).

**Table 1.** Understanding Big Data

| *Traditional Data* | *Big Data* |
|---|---|
| Documents | Photos |
| Finances | Audio and Video |
| Stock Records | 3D Models |
| Personnel files | Simulations |
| | Location data |

This example provides information about the volume and the variety of Big Data.

It is difficult to work with complex information on standard database systems or on personal computers. Usually it takes parallel software systems and infrastructure that can handle the process of sorting the amount of information that, for example, meteorologists need to analyze.

The request for more complex information is getting higher every year. Streaming information in real-time is becoming a challenge that must be overcome by those companies that provides such services, in order to maintain their position on the market.

By collecting data in a digital form, companies take their development to a new level. Analyzing digital data can speed the process of planning and also can reveal patterns that can be further used in order to improve strategies. Receiving information in real-time about customer needs is useful for seeing market trends and forecasting.

The expression "Big Data" also resides in the way that information is handled. For processing large quantities of data that is extremely complex and various there needs to be a set of tools that are able to navigate through it and sort it. The methods of sorting data differ from one type of data to

another. Regarding Big Data, where the type of data is not singular, sorting is a multi-level process.

Big Data can be used for predictive analytics, an element that many companies rely on when it comes to see where they are heading. For example, a telecommunication company can use data stored from length of call, average text messages sent, average bill amount to see which customers are likely to discard their services.

**2. Volume, Velocity, Variety**

The "3V's", how Doug Laney calls them in his article *3-D Data Management: Controlling Data Volume, Velocity and Variety*, published in 2001, represents key elements that are considered vital regarding the characteristics of Big Data systems.

The first characteristic of Big Data, which is "**Volume**", refers to the quantity of data that is being manipulated and analyzed in order to obtain the desired results. It represents a challenge because in order to manipulate and analyze a big volume of data requires a lot of resources that will eventually materialize in displaying the requested results. For example a computer system is limited by current technology regarding the speed of processing operations. The size of the data that is being processed can be unlimited, but the speed of processing operations is constant. To achieve higher processing speeds more computer power is needed and so, the infrastructure must be developed, but at higher costs.

By trying to compress huge volumes of data and then analyze it, is a tedious process which will ultimately prove more ineffective. To compress data it takes time, almost the same amount of time to decompress it in order to analyze it so it can be displayed, by doing this, displaying the results will be highly delayed. One of the methods of mining through large amount of data is with OLAP solutions (Online Analytical Processing) (**Fig.1.** Data warehouse -> OLAP). An OLAP solution consists of tools and multidimensional databases that allow users to easily navigate and extract data from different points of view. Therefore, it identifies relations between elements in the database so it can be reached in a more intuitive way. An example of how OLAP systems are rearranging the data imported from a data warehouse is below. For obtaining results various OLAP tools are used in order for the data to be mined and analyzed.

| NAME | AGE | LOCATION |
|---|---|---|
| John | 21 | England |
| Mary | 32 | USA |
| Ray | 44 | Canada |

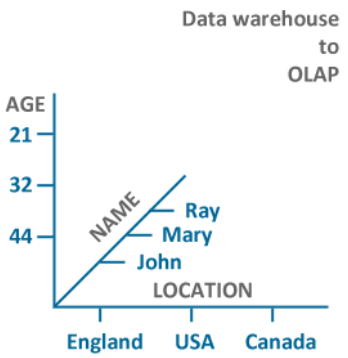


**Fig. 1.** Data warehouse -> OLAP

"**Velocity**" is all about the speed that data travels from point A, which can be an end user interface or a server, to point B, which can have the same characteristics as point A is described. This is a key issue as well due to high requests that end users have for streamed data over numerous devices (laptops, mobile phones, tablets etc.). For companies this is a challenge that most of them can't keep up to. Usually data transfer is done at less than the capacity of the systems. Transfer rates are limited but requests are unlimited, so streaming data in real-time or close to real-time is a big challenge. The only solution at this point is to shrink the data that is being sent. A good example is Twitter. Interaction on Twitter consists of text, which can be easily compressed at high rates. But, as in the case of "Volume" challenge, this operation

is still time-consuming and there will still be delay in sending-receiving data. The only solution to this right now is to invest in infrastructure.

"**Variety**" is the third characteristic of Big Data. It represents the type of data that is stored, analyzed and used. The type of data stored and analyzed varies and it can consist of location coordinates, video files, data sent from browsers, simulations etc. The challenge is how to sort all this data so it can be "readable" by all users that access it and does not create ambiguous results. The mechanics of sorting has two key variables at the beginning: the system that transmits data and the system that receives it and interpret it so that can be later displayed (**Fig. 2.** Send-Receive).

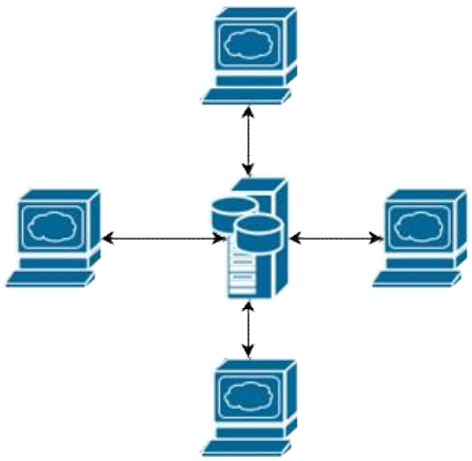

**Fig. 2.** Send-Receive

The issue of these two key aspects is that they might not be compatible regarding the content of the data transferred between them. For example, a browser can send data that consists of user's location, favorite search terms and so on. Meanwhile, the Big Data system receives all this information unsorted, so it's difficult to for it to understand whether this user is from "London" or from "orange". To avoid this "mess" created in Big Data solutions, all systems that send data should be standardized so that can send data in a logical array that, afterwards, it can be easily analyzed and displayed in a proper manner.

After Laney's "3V's" another two "V's" where added as key aspects of Big Data systems.

The fourth "V" is "**Value**" and is all about the quality of data that is stored and the further use of it. Large quantity of data is being stored from mobile phones call records to TCP/IP logs. The question is if all together can have any commercial value. There is no point in storing large amount of that if it can't be properly managed and the outcome can't offer insights for a good development.

"**Veracity**" is the fifth characteristic of Big Data and came from the idea that the possible consistency of data is good enough for Big Data. For example, if A is sending an email to B, B will have the exact content that A sent it, if else, the email service will not be reliable and people will not use it. In Big Data, if there is a loss regarding the data stored from one geo-location, is not an issue, because there a hundreds more that can cover that information.

Current technologies software technologies try to overcome the challenges that "V's" raises. One of these is Apache Hadoop, which is open source software that its main goal is to handle large amounts of data in a reasonable time. What Hadoop does is dividing data across a multiple systems infrastructure in order to be processed. Also, Hadoop creates a map of the content that is scattered so it can be easily found and accessed.

### 3. Useless to useful

The quantity of data that is being stored is not always one hundred percent useful. On the other hand the data stored is, in most cases, not already sorted and it represents piles of data that can consist of location information, web traffic log, financial data etc. To become useful, specialists must sort so that can be later analyzed and can be of any value (**Fig. 3.** Sort - Analyze). IT specialists say that they spend more time trying to "clean up" the data than they are analyzing it. Sorting and cleaning up data is a challenge that is hardly overcome. To do that, companies usually hires trained

people that are able to manipulate the type of data that executive employees or higher management will further use. This process is time consuming and the costs are proportional to the volume of data.
A lot of companies try to sort and analyze the data that they stored with their own employees that have minimum skills or don't have them at all. The lack of skills in sorting Big Data will most certainly conclude into faulty results and/or truncated data that cannot serve its purpose.

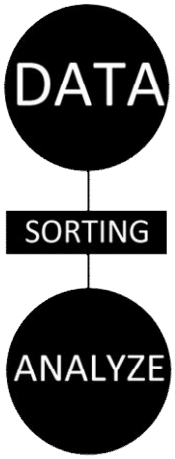


**Fig.3.** Sort - Analyze

A solution to eliminate the "need" for specialists is to implement software solutions which do not require special skills to understand how to put it at work. To be able to do that there is one more obstacle to overcome: quality of data. To achieve this, the architecture of the source that collects the data must be able to already sort it logical. For this to be accomplished, the data that is collected must be received in an understandable manner for the software that sorts it. These are the obstacles that will not be easy overcome, because there is no such thing as the idea of "controlled environment" when it comes to describing the World Wide Web. This problem can only be solved if the sets of data come, for example, from the financial department to the higher management of a company. In this case, data is already manipulated and is easy to understand and analyze to create projections.
A set of data, in Big Data environment, can be processed with OLAP tools. By doing so, there are connections between information that can be made. This set of tools have the purpose to rearrange the data provided into "cubes", which represents an IT architectural design that has the meaning of creating sets of data assembled into a logical and easy way to access it. By doing this, specialists achieved a higher speed, therefore a smaller waiting time, in processing large amount of data. The usefulness of the data that is being processed in an OLAP environment can still be questionable because all the data that was provided is being analyzed and sorted.

**4. Data privacy. Data security.**
This has many implications and it concerns individuals and companies as well. Individuals have the right, according to International Telecommunications Union, to control the information that may be disclosed regarding them. Information posted by users on their online profiles is likely to be used in creating a "users profile" so that can be further used by companies to develop their marketing strategies and to extend their services. Individual's privacy is still a delicate problem that can only be solved with drastic solutions. Allowing persons to choose whether they post or not information about them is a more secure way to achieve privacy, but will also cause software to "malfunction". For example in a social network, if a person is allowed to choose whether he/she wants to complete the fields regarding personal information and, in the same time, allow them to choose if the social network can store information about their IP address, location etc., this could be a possible threat to everyone else that is using the same social network.
For companies the privacy issue is more related to the sensitive data that they work with. Whether is financial data, clients list,

perspective projects, all represents valuable data that may or may not be disclosed. Companies have multiple choices regarding where to store their information. They can either store it on cloud systems (**Fig. 4.** Cloud computing), "in-house" systems or a hybrid solution.

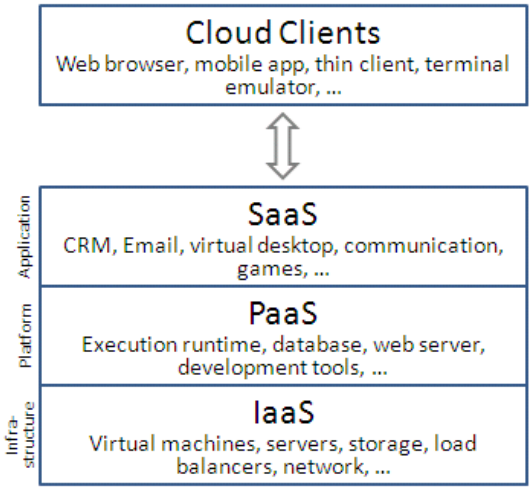


**Fig. 4.** Cloud computing

By storing data on cloud systems is more convenient for companies in terms of cost. Also, a cloud system is not only characterized by storage space, but as well for the speed of processing requested operations. The data security still remains a contentious issue in this case.

To solve that, some companies choose to build their own infrastructure for storing and manipulate the data that they have. For smaller companies this can be a solution, but, in most cases, to implement such a system the costs are high. Also, to maintain this type of system it requires trained personnel and the more the company grows, the more it will be needed an add-on to the infrastructure. After all, this solution will prove redundant. The only gain of this solution is privacy.

Manipulating data, collecting it and store it in a proper manner that is in the advantage of the beneficiary and as well for the user that provides the data, will remain an important issue to be solved by IT security specialists. One solution for this matter, besides keeping all the data stored on an "in-house" Big Data system, is to encrypt it (**Fig. 5.** Encryption). By encrypting the data with a personal key it makes it unreadable for persons that don't have the clearance to see it. The downside of using encryption is that you have to use the same software that encrypted it to read it and analyze it, or, in worst case scenario, if you want to make it available for every software that is on the market the process implies more steps which take time. First step is to encrypt it using special encryption software, after that, each time the data is used for manipulation or analysis it must be decrypted and after work is finished, encrypt it again.

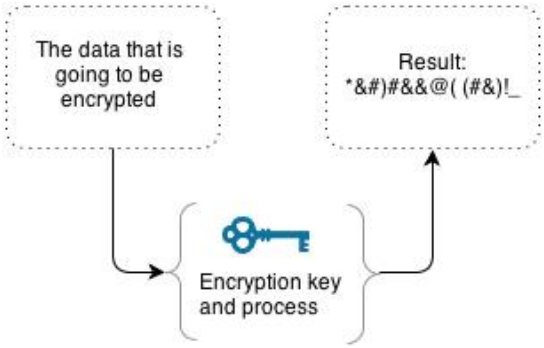


**Fig. 5.** Encryption

To achieve performance from this process there has to be an OLAP system that is capable of doing encryption at the same time with reading data. By doing this the process will be much faster and data can be managed almost in real-time.

## 5. Sharing data

Sharing the information proves to be one of the most valuable characteristics of development. Information about almost anything can be found by simply doing a "Google search". Every person and company has at their disposal large amount of information that can use it to serve their purposes. Everything is available only if everyone shares it. Regarding persons, there is a difference between what is personal and what can be made public. The issue of what is personal and what is public mostly resides in the point of view of the services that they use.

Regarding companies, this is a challenge that most refuse to overcome. The reason that companies don't want to share their

own "Big Data" warehouse is more related to competitiveness and sensitive data that they have. Otherwise, if this line is crossed, each company will have more data that they can analyze so that more accurate results can be obtained. With better results, comes better planning. If companies share the information that they hold about current market situation and/or possible clients and strategies to approach them, the grade of development will be drastically reduced and they will start focusing on how to hold to their current clients.
Sharing "Big data" at a level where each entity will show all the information that they hold is impossible. The framework of displaying data should be wider. A more transparent representation of current information that a company holds will be in the advantage of everyone. By doing this, the type of information and the way it is structured can help further development of software systems that can be standardized and can work with all types of data imported from various sources.

**6. Infrastructure faults**

Storing and analyzing large volumes of data that is crucial for a company to work requires a vast and complex hardware infrastructure. If more and complex data is stored, more hardware systems will be needed.
A hardware system can only be reliable over a certain period of time. Intensive use and, rarely, production faults will most certainly result in a system malfunction. Companies can't afford to lose data that they gathered in the past years, neither to lose their clients. For avoiding such catastrophic events they use a backup system that does the simple operation of storing all data. By doing this, companies obtain continuity, even if they are drawn back temporary. The challenge is to maintain the level of services that they provide when, for example, a server malfunction occurs right when a client is uploading files on it. To achieve continuity, hardware systems are backed by software solutions that respond in order to maintain fluency by redirecting traffic to another system. When a fault occurs, usually a user is not affected and he/she continues work without even noticing that something has happened. (**Fig. 6.** System failure)

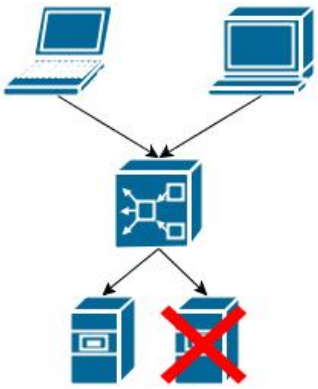

**Fig. 6.** System failure

The flow of data must not be interrupted in order to obtain accurate information. For example, Google is sending one search request to multiple servers, rather than sending it to only one. By doing this, the response time is shortened and also there is no inconsistency in the data that users sends – receives.
System failure affects the process of storing data and is making more difficult to work with. There can be created a permanent connection between the device, that is sending data, and the system that is receiving it, as a solution to this problem. By creating a loop, the "sender" will make sure that the "receiver" has no gaps regarding the data that should be stored. This loop should work as long as the system that is receiving data tells the system that sends it to stop because the data that is stored is identical to the one sent. So, is a simple comparison process that can prevent loosing data. This process can also slow down the whole process. To avoid this from happening, for any content that is transmitted, the sender must generate a "key". This key is then transferred to the receiver to compare it with the key that it generated regarding the data that was received. If both keys are

identical than the "send-receive" process was successfully completed. For better understanding, this solution is similar with the MD5 Hash that is generated over a compressed content. But, in this case, the keys are compared automatically.

Loosing data is not always a hardware problem. Software can as well malfunction and cause irreparable and more dangerous data loss. If one hard drive fails, there is usually another one to back it up, so there is no harm done to data, but when software fails due to programming "bug" or a flaw in the design, data is lost forever. To overcome this problem, programmers developed series of tools that will reduce the impact of a software failure. A simple example is Microsoft Word, which saves from time to time the work that a user is doing in order to prevent the loss of it in case of hardware or software failure. This is the basic idea of preventing complete data loss.

## 7. Technologies for Big Data

Once realized the amplitude of information that is crossing the internet, specialists started to question how to handle this amount of data. To obtain good insights and mine this information they had to develop tools capable of creating the expected results. A common implementation that handles Big Data is **MapReduce** (**Fig. 7.** MapReduce).

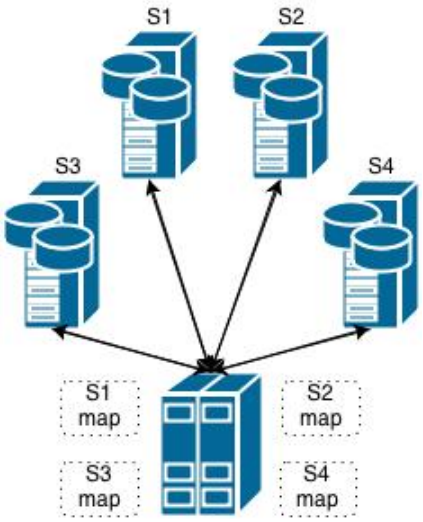


**Fig. 7.** MapReduce

This is more of a technique that programmers use when they are confronted with large amount of data.

MapReduce consists of two things: mapping and reducing. By mapping a certain dataset is restructured into a different set of values. Reducing is a process that takes several "mapped" outputs and forms a smaller set of tuples.

The most popular technology that is able to mine and sort data is **Hadoop**. Being open source software, Hadoop is the most implemented solution for handling Big Data. It has enough flexibility to work with multiple data sources, or even assemble multiple systems to be able to do large scale processing. Hadoop is used by large companies such as Facebook and Google. Hadoop also use HDFS (**Hadoop Distributed File System**) that has the role to split data into smaller blocks and distribute it throughout the cluster. In order to assist Hadoop, Facebook developed a software system called **Hive**. Hive is basically a "SQL-like" bridge that connects with Hadoop in order to allow conventional applications to run queries. The advantage is that is simple to use and understand. It combines the simplicity and utility of a standard relational database with the complexity of a Hadoop system. The downside of using Hive is latency. Because it is built on Hadoop, Hive can have high latencies on executed queries, compared to IBM's DB2. Large companies use Hadoop as a starting point in order to deploy other solutions.

DB2 is a fast and solid data manipulating system. It has feature that reduces the cost of administration by doing an automated process that increases storage efficiency and improves performance.

Oracle, on the other hand, comes with a complete system solution for companies (**Fig. 8.** Oracle solution).

It starts from the basic ideas of Big Data sources, which can be traditional data generated by ERP systems, sensor data and social data, defined by the feedback that the company receives from customers and

other sources. The solution given by Oracle is to create a system from top to bottom, based on NoSQL. A NoSQL database is capable of handling various types of data that traditional relational databases are unable to handle and lose data consistency. NoSQL derives from "Not only SQL", which means that it allows regular SQL queries to be executed. Oracle's solution is presented in 4 steps: ACQUIRE > ORGANIZE > ANALYZE > DECIDE [2]. All the steps combine different solutions like HDFS, NoSQL, Oracle Big Data connectors and Cloudera CDH.

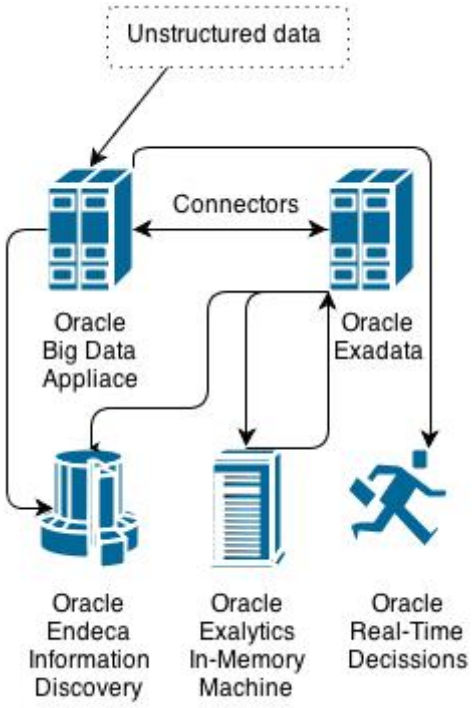


**Fig. 8.** Oracle solution

CDH or *Cloudera's Distribution Including Apache Hadoop*, offers batch processing (uses MapReduce and HDFS), interactive SQL (allows users to execute SQL queries) and interactive search [3]. All these key features that Cloudera offers are solutions that allow users to navigate through clusters and retrieve data that they need.

SAS offers multiple solutions to overcome Big Data mining and analysis. It also tries to cover all that is necessary for a company to create value from stored data. One solution is SAS DataFlux which is a data management solution that can provide users the right tools for integrating data, mastering data and data quality. It also allows access and use of data across company and also provides a unified set of policies in order to maintain data quality. SAS also provides high-performance analytics solution that is providing the company good insights from analyzing data in a structured, easy to read, report. This is basically one of the main goals when working with Big Data, to get best insights from quality data. Also SAS provides analytics solution that is based on a drag-and-drop system which can provide easy to understand and customized reports and charts.

SAS is more oriented in providing software solutions to help companies benefit from data that they have stored.

The problem of handling Big Data doesn't always resides in analysis and mining software solutions. It has a great impact over hardware systems and their capability of processing. The two of them, software and hardware solutions, create a complete Big Data system that can be viable and will produce the expected outcome. In order to handle large and complex data sets, a solution must be divided according to job process. For example, a basic data storage and mining solution should have a system that will store brief information about the data that exists on clusters. By doing this, the data mining process is drastically reduced because the role of this system is to orientate the user where to look and what to look for. Also, this system should be able to evenly "spread" the data among clusters. By performing this, clusters can be monitored so there will be no overloading and, therefore, slowing down the outcome.

Another solution to treat Big Data is to design a system that is capable of making differences between various types of data. Searching through one type of data is easier than to search through different types of information. As an example, searching through a full text file is easier to find answers than searching through a text

file that has images as well that can provide answers to questions. This is a "divide et impera" technique. By doing this, one cluster with special software design will handle video file, another will handle 3-D models, another plain text files etc. (**Fig. 9.** Divide et impera)
This system will allow data fragmentation which will be faster to process. The software solution that can handle data fragmentation can be achieved through MapReduce.

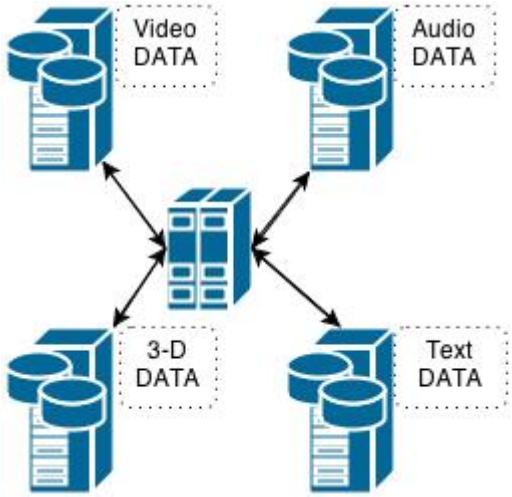


**Fig. 9.** Divide et impera

Besides the basic function of dividing data, MapReduce can be configured to recognize the type of data that is mapping. Special designed software for each cluster will have the job to sort data by various key elements which are set by the user. Clusters can also analyze in a basic manner the information stored so that the "mainframe" will act as a control center for the system. This will help analysts achieve fast and accurate results and will also allow real-time updates about ongoing information.
Regarding Big Data "Volume" is a characteristic, as well as a challenge. Trying to deal in a fast manner with large data sets will be difficult for a system. Trying to lower the volume might help solve this problem. For a Big Data solution is faster to process 1GB of data instead of 1TB. When it comes to Big Data, valuable information should be the subject of analysis. Spending time on "cleaning" information can lead to a result that is no longer valid or is too late to be applied. To speed up this process, an automated data "clean-up" process should be implemented. In most cases, not all the information collected is needed for the analysis. To do this, a data filtering solution can be created. For example, a mainframe can decide which data is needed and which is not. The one that is needed will be transferred to a main cluster that provides the information needed for immediate analysis. The rest of the information will be transferred to a secondary cluster that will hold only data that can be later analyzed or even deleted.
So, "looping", "divide et impera" and "filtering", these can all form a Big Data solution that can be helpful. This solution covers the data loss that can occur from a hardware malfunction or a software error. It will also manage a data distribution among clusters by data type and previously set aspects that will ensure better and faster analysis of the information stored. Least but not last, will provide an automatic filtering process that will facilitate the evaluation of valuable data. To achieve this, the solution must have a coordination center, a processing system and special designed software for each cluster system.
The only challenge that needs a new approach is the "Velocity" issue. In order to obtain higher processing and transferring speed, the volume of data that is manipulated must be reduced. This cannot be possible without slowing the analysis process and, therefore, "Volume" stays untouched.

## 8. Conclusions

Building a viable solution for large and complex data is a challenge that companies in this field are continuously learning and implementing new ways to handle it. One of the biggest problems regarding Big Data is the infrastructure's high costs. Hardware equipment is very expensive for most of the companies, even if Cloud solutions are

available. Each Big data system requires massive processing power and stable and complex network configurations that are made by specialists. Besides hardware infrastructure, software solutions tend to have high costs if the beneficiary doesn't opt for open source software. Even if they chose open source, to configure there is still needed specialists with the required skills to do that. The downside of open source is that maintenance and support is not provided as is the case of paid software. So, all that is necessary to maintain a Big Data solution working correctly needs, in most cases, an outside maintenance team.

Software solutions are limited by hardware capabilities. Hardware can only be as fast as current technologies can offer. A software solution just sends the tasks in order to be processed. The software architecture tries to compensate the lack of processing speed by sorting and ordering the requests, so that processes can be optimized in order to achieve best performance.

To sort through data, so that valuable information will be extracted for further use, requires human analysis skills. A computer program can only do what is programmed to do, it cannot see grey areas and cannot learn or adapt to new types of information unless is programmed to handle it. Therefore, human capabilities are used to sort data with a set of tools which speed up the process. All this will only increase the time that results will be displayed and so, the analysis of the results, in order to evaluate current position or forecast, will decrease the beneficiary's time for taking measures or plan properly.

## References

[1] Global data center traffic – Cisco Forecast Overview - http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns1175/Cloud_Index_White_Paper.html

[2] Oracle Big Data strategy guide, http://www.oracle.com/us/technologies/big-data/big-data-strategy-guide-1536569.pdf

[3] Cloudera's 100% Open Source Distribution of Hadoop, http://www.cloudera.com/content/cloudera/en/products/cdh.html

**Alexandru Adrian TOLE** (born 1986 in Romania) graduated from the Faculty of Domestic and International Commercial and Financial Banking Relations of the Romanian – American University in 2009. He also graduated the Scientific Master in Finance, Banking, Insurances. He works at the Ministry for Information Society. He is pursuing a Ph. D. in the area of Business Intelligence systems.