# Visualization of Big Data with the Map-Reduce program execution platform: Hadoop

[1] Sara Riahi, [2] Azzeddine Riahi

[1] *Department of Mathematics and Computer Science, Chouaib Doukkali University, Faculty of Sciences , Po Box 20 , Postcode, 24000, El Jadida, Morocco.*
*IMC Laboratory, Department of Physics Chouaib Doukkali University Faculty of Sciences, El JADIDA, Box 20 , Postcode, 24000, Morocco*

**Abstract :** *"Big data" is the fashionable term currently found in all professional conferences related to data science, predictive modeling, data mining, to name just a few areas literally electrified by the prospect of integrating larger datasets and data flows more quickly into their business processes and other organizational processes. As is often the case when new technologies begin to transform industries, new terminologies emerge, along with new approaches to conceptualize reality, solve problems, or improve processes. A few years ago, we limited ourselves to "segment" customers into groups that could acquire specific properties or services. It is now possible and common to build models for each customer in real time as they browse the Internet for specific properties: Instantly, prospects' interests are analyzed and it is possible to display highly targeted advertising, which is a level of personalization inconceivable only a few years ago. Inevitably, the disappointment may be up to expectations in many areas as technology around big data are promising. A limited number of data accurately describing a critical aspect of reality (vital to the business) is far more valuable than a deluge of data on less essential aspects of that reality. The purpose of this article is to clarify and highlight some interesting opportunities around big data, and illustrate how analytic platforms can leverage this wealth of data to optimize a process, solve problems, or improve customer knowledge.*

**Keywords:** *Massive data, Analytics, Hadoop, HDFS, Map Reduce*

## *1 Introduction:*

After each technological revolution, the question of the interest of the new technologies arises. However, whatever the answer, the result is always the same: technological advances will never stop to attract the gain a competitive advantage, big data projects are changing the way companies live and are starting to grow. First seen as a luxury, "big data" is now seen by most companies as a necessity, and is at the heart of corporate strategy to create value in different areas by developing information **[1]**. We have moved from relational database management systems (RDBMS) that follow a rigid model to ensure the needs of management computing towards a more flexible model that of the web, centered on a varied content (structured, semi-structured and unstructured),bulky and arrives at different speeds (Figure 1). With the arrival of the Web, new needs have emerged, such as scalability at a time when intensive computing could be done on large amounts of data using super computers such as computer grids and when necessary, it was enough to add storage disk, RAM and CPU to increase power (Scale up). Today's needs are firstly to ensure large power storage, RDBMS that can not be achieved through the horizontal scalability principle (Scale Out) of adding low-cost machines to parallelize the processing **[2]**. Secondly, ensure high computing power of varied data that can not be stored and processed in conventional RDBMS and are of different sources such as the flow of events, log files and social networking.

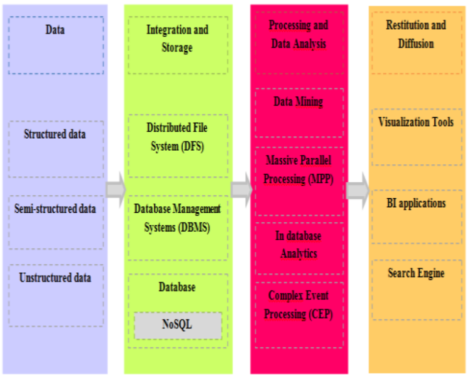
Data
Structured data
Semi-structured data
Unstructured data
Integration and Storage
Distributed File System (DFS)
Database Management Systems (DBMS)
Database
NoSQL
Processing and Data Analysis
Data Mining
Massive Parallel Processing (MPP)
In database Analytics
Complex Event Processing (CEP)
Restitution and Diffusion
Visualization Tools
BI applications
Search Engine

**Figure 1:** Big Data is based on several technologies, which are used to exploit large amounts of data.

These scalability requirements for the storage and processing of large volumes of data by major Web players gave birth to the term "big data". The big data is to use a controlled perspective which includes the new distributed storage technology and parallelized processing often unstructured massive data **[3]**. While taking advantage of the hardware evolutions that result in lower storage costs to add cheap machines for storage distribution and processing on multiple nodes, to ensure scalability at lower cost, which can generate a return on investment (ROI). Big data can be treated from different angles, but this article aims to address the issue of "How to choose a big data architecture".

## *2 .When can we talk of massive data or big data:*

Obviously, there is no universal definition, and the correct answer is "it depends". In fact, from a practical point of view, and in most discussions related to this theme, big data are characterized by very large datasets, of the order of several gigabytes to a few terabytes. These data can be easily stored and managed in the "traditional" databases with conventional hardware (database servers) **[4]**.

The analytical platforms are multitasking for all the fundamental operations of data access (reading), and for all its algorithms of transformation and predictive modeling (and scoring), which makes it possible to analyze these datasets (actually very bulky ) without having to use new specialized tools.

### *2.1 Big Data Volumes:*

Most IT administrators set up data protection measures which they have knowledge. But in many cases their businesses contain data, terminals, and systems that they do not even know existed. Indeed, the 21st companies, extended and distributed, shares valuable data with its suppliers, partners and customers, via various platforms. A part of unknown that evolves rapidly, as the proliferation of terminals and platforms, exposing companies to new risks. Fortunately, simple and thorough evaluation of the security environment usually enough to identify gaps to be filled **[5]**.

However, many companies do not even take the trouble. In the best cases they are wrong about the savings at worst, an offense occurs costing them much more than would have required a solid security program and mobilizing valuable resources, and since IT staff will have to devote time to find punctual solutions and other tinkering or to prevent future disasters**[3]**. Hence the importance of detecting and identifying all data available in the company, they are simply stored or actively used. Small office servers and laptops are a prime example of the data sources the company does not know about. By identifying sensitive data, their storage location, and their mode of transit, security requirements can be met, data loss can be minimized, and processes and applications that manage this data can be optimized. No company can claim a comprehensive security program as long as it does not ignore the risks and the origin of the threats.

To handle large volumes of data, companies need to know**[6]**:

- Where are localized data: the goal is to determine if this data is inactive (stored in shared files, on workstations, etc.) or in transit (in e-mails, transferred files, instant messages, etc.)
- What type of data is it and who it belongs to: the nature of the data and the identity of the owner will determine the security procedures / standards necessary for their protection (eg customer listings, social security numbers, credit cards, medical information, resumes, financial reports, etc.).
- How are these data used and how important are they: companies must know the connection between their data and business processes, know what use is made (very common or rare use, only for retention purposes, etc.) and be able to assess the sensitivity of (security rules).

Increase flexibility through cloud services **[7]**:

Data classification has another interest that better manage the development of the company. Conversely, the uncontrolled increase in data volumes is gradually straining the agility of companies, whose infrastructure ends up no longer satisfying expectations and which are forced to integrate new servers with disparate configurations **[4]**. The result is slightly changing environment over the long term, with data scattered everywhere. In addition, this non-extensible architecture prevents companies to exploit their data service satisfaction strategic objectives.

Of course, in reality, IT departments do not develop at the same pace as Big Data. And still more investment strategy is anyway not wisely to manage Big Data, especially in the current context of reduced cost of ownership and optimize ROI **[8]**.

Finding answers to previous questions can already help companies determine how the IT infrastructure can truly meet their strategic needs. But a computer model appears increasingly as the solution to big data management problems. This is cloud computing (public, private or hybrid, as needed), promising economies of scale and gains in flexibility **[9]**.

Per-use, cloud computing reduces capital expenditure in hardware, software and services. Generally accessible without initial investments or for a minimal investment, Cloud resources are available as cancellable at any time, eliminating the risk of non business. In most cases, only the operating costs are thus paid by companies.

Current cloud services must also ensure high levels of availability to comply with the strictest Quality of Service Commitments. They are accompanied from some powerful options enabling companies to easily move workloads between their data center and the cloud according to their capacity constraints and strategy of the moment **[6]**. The best Cloud Service Providers go further in the transparency process by enabling corporate customers to physically access their data in the cloud provider's datacenter. This sense of control reassures information systems managers that they know exactly where their data is and how it is managed **[10]**.

Against all expectations, the main advantage of cloud computing model probably lies in its security, often strengthened compared to what today's companies can guarantee. Cloud providers have specially trained security teams, and a robust security protocols environment systematically update the latest threats, strengths with which most IT companies can not compete.

### *2.2 Large Data Volumes and Big Data:*

In general, discussions around big data focus on data warehouses (and their analysis) exceeding several terabytes. Specifically, some data warehouses can exceed several thousand terabytes, reaching several petabytes .Beyond petabytes, data storage capacities are measured in exabytes. In some applications, the data accumulates very quickly. For example, for industrial applications or automated production lines, such as for power generation, continuous data flows are generated every minute or second, sometimes for tens of thousands of parameters **[11]**. In the same way, we have seen in recent years the emergence of smart-grid technology for "smart" electricity distribution networks, which make it possible to measure the electricity consumption of each household minute by minute, or even second per second.

For this type of application that requires data storage over several years, it is not uncommon to see very quickly accumulate huge data (Figure 2). There are more and more applications in the administration and the commercial sector where the amount of data and the speed at which this data is accumulated requires several hundred terabytes or petabytes dedicated to storing and analyzing data **[9]**.

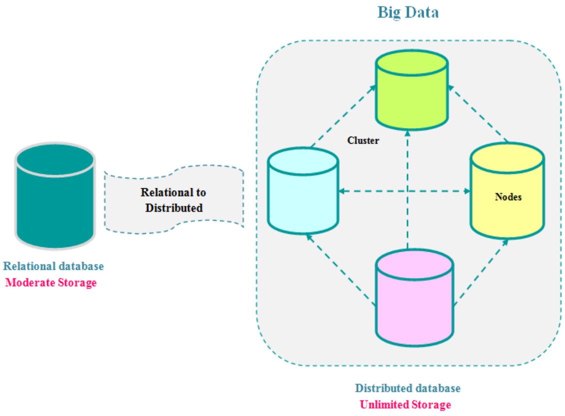

**Figure 2:** The nodes can be physically heterogeneous, Big Data management systems manage this heterogeneity

Modern technology now can track individuals and their behavior in various ways, for example, when we surf the Internet, we buy products on the Internet or in supermarkets, or that we leave our phone activated leaving information about where we have been and where we are going. The various modes of communication, the phone call to the information shared on social networks like Facebook, or video sharing sites like You Tube, which generate massive amounts of new data daily. Similarly, modern health technologies generate massive amounts of data for the delivery of care (pictures, movies, real-time monitoring) and reimbursement of healthcare organizations **[12]**.

### *3. Technical Challenges of Big Data:*

There are basically three types of challenges around big data **[7]**:

- The storage and management of massive data, of the order of a hundred terabytes or petabyte which exceed the current limitations of conventional relational databases from the viewpoint of storage and data management.

- The management of unstructured data (which often constitute the bulk of data in the big data scenarios), that is to say how to organize text, videos, images, etc ...
- The analysis of this massive data, both for reporting and advanced predictive modeling, but also for deployment.

### *3.1 Storage of Big Data:*

The big data are generally stored and organized in distributed file systems. If there are different approaches and implementation strategies, the information is stored on several hard disks (sometimes several thousand) and conventional computers. An index ( "map") allows to know where (on which computer / disk) is specific information. In fact, for reasons of security and robustness, each piece of information is generally stored several times, for example in the form of triplets.

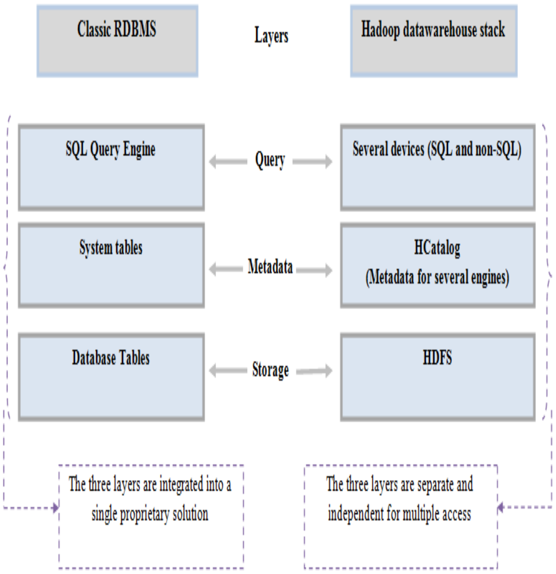
Classic RDBMS
Layers
Hadoop datawarehouse stack
SQL Query Engine
Query
Several devices (SQL and non-SQL)
System tables
Metadata
HCatalog
(Metadata for several engines)
Database Tables
Storage
HDFS
The three layers are integrated into a single proprietary solution
The three layers are separate and independent for multiple access

**Figure 3:** Comparison of classical data management architecture with that based on hadoop

Through the use of standard hardware and open source software for the management of the distributed file system (such as Hadoop) (Figure 3) **[13]**, it is possible to create easily enough reliable data warehouses in the order of petabytes, and systems storage is becoming more common.

### *3.2 Non-Structured Information:*

Most of the information collected in distributed file systems is unstructured information such as text, images or videos. This has advantages and disadvantages. The advantage is that businesses and administrations can store "all data" regardless which are relevant and useful in a decision-making perspective **[8]**. The disadvantage is that it is necessary to set up massive data processing in order to extract interesting information. While some of these operations may be relatively simple (for example, calculating simple numbers, etc.), others require more complex algorithms that need to be developed specifically to function effectively on the distributed file system.

Unstructured data and information: As was the case during the generalization of relational databases, the main challenge today is that even if we store large amounts of data, only the information we can extract makes them useful **[14]**. In more general terms, while the amount of data grows exponentially, our ability to extract information and act on this information remains limited and tends asymptotically towards a limit (regardless of how the data is stored).

Essential consideration that we will develop: methods and procedures for extracting and updating models, with a view to automating decisions and decision-making processes, must be designed in conjunction with data storage systems to ensure the best interest and utility of these systems for the company **[5]**.

### *3.3 Big Data Analysis:*

This is undoubtedly the main challenge when working on massive data, often unstructured: how to analyze effectively. In fact, there are many papers on this topic than on technologies and storage solutions to manage big data. There are, however, a number of things to consider.

#### *3.3.1 Map-Reduce :*

Generally, when we analyze hundreds of terabytes or even petabytes of data, it is not realistic to extract the data to another location in order to analyze it. The process of moving data across a cable to a server or multiple separate servers (for parallel processing) requires too much time and bandwidth **[15]**. Instead, the analytical calculations must be made in physical proximity to where the data is stored. It is much easier to bring analytics to the data, than to bring the data to analytics. This is exactly what map-reduce algorithms do, that is, analytic algorithms designed for this purpose (Figure 4). A central component of the algorithm will delegate the sub-calculations in different parts of the distributed file system and then combine the results calculated by the individual nodes of the file system (the reduction step).

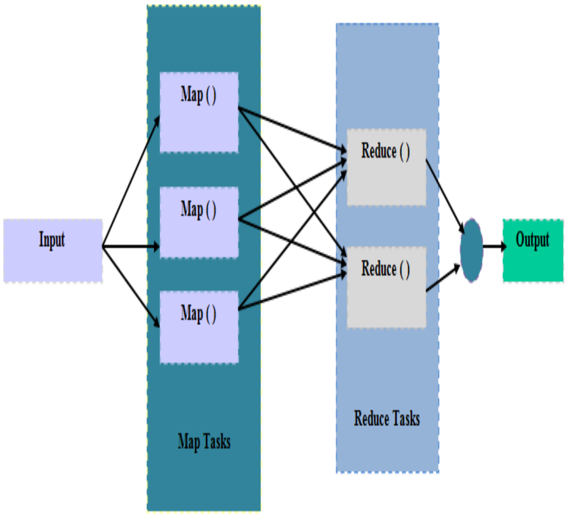

**Figure 4:** Map Reduce process

In summary, to calculate an effective, the algorithm will compute in parallel in the distributed file system, subtotals in each node, and then return to the master component these subtotals which are then summed. There is much information on the Internet about the various calculations can be made with the owner of map-reduce architecture, including predictive modeling.

### *3.3.2 Basic Statistics, Business Intelligence (BI)*

For reporting BI needs, there are various open-source solutions for calculating totals, averages, proportions, etc. using map-reduce. It is therefore quite simple to obtain accurate numbers and other basic statistics for reporting **[2]**.

### *3.3.3 Predictive Modeling, Advanced Statistics:*

At first glance, it may seem more complex to build predictive models using a distributed file system; but in practice, this is not the case for various reasons.

#### *3.3.3.a Data preparation:*

Remember that most of the data we find in the distributed file systems for big data is often unstructured information (eg, text). In fact, it is rare to find applications where the measurements or collected values generate petabytes of data **[16]**. If datasets can actually become very large, as with all continuous process data, the information they contain can be synthesized. Therefore, a "smart" aggregation of data upstream (where the data is stored) is required and can be performed simply, which provides the data files containing all the necessary information on changes dynamics impacting efficiency for modeling and optimization.

#### *3.3.3.b Customers experienced analysis and data preparation:*

This example illustrates that large data sets often contain information that we can synthesize. The data collected by electric meters using "smart-grid" technology will have characteristics similar to the feelings expressed by the same person on a particular subject, or even by a group of individuals on a larger number of subjects. If we can extract a large number of relevant tweets on an hourly or daily basis, the complexity of the feelings they express is actually quite limited (and small size). Most tweets are complaints and short sentences describing "bad experience"**[17].** In addition, the number and intensity of these feelings is quite stable over time and for specific types of grievances (eg, lost luggage, canceled flight, poor quality food, etc ...).

Therefore, in this example, a simple compression tweets feelings scores using methods of text mining allows for much smaller size of data sets it is possible to align more easily with existing structured data to better understand the stratification of specific customer groups and their complaints. There are many tools to achieve this type of data aggregation (eg scoring feelings) in distributed file systems, and analytical process can be easily implemented.

### *3.3.4 Construction of Models:*

Sometimes we have to quickly build accurate models of big data stored in a distributed file system. In fact, it is usually more useful to construct a large number of models on finer segments of data in a distributed file system, but we will return to this point later. In fact, we can deploy map-reduce for a variety of common data mining and predictive modeling algorithms suitable for mass parallel data processing for distributed file systems (Figure 5). But even if we can increase the volume of data significantly, would our final prediction model be better or more accurate.

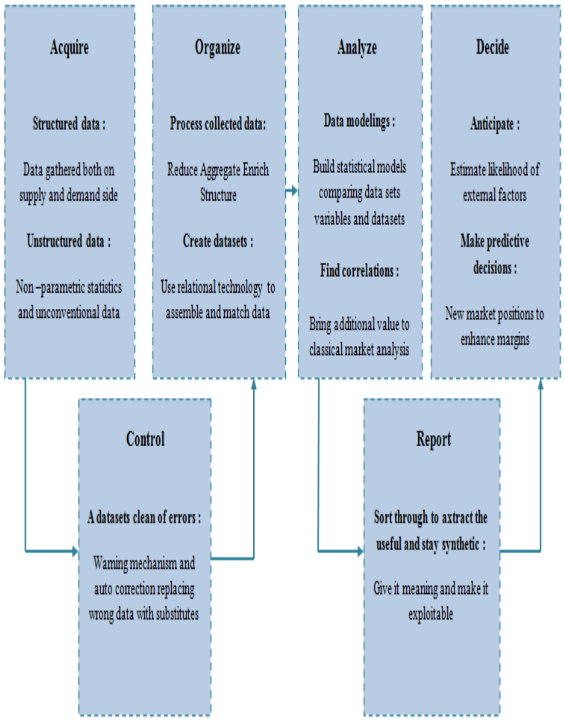
Acquire
Structured data :
Data gathered both on supply and demand side
Unstructured data :
Non –parametric statistics and unconventional data
Organize
Process collected data:
Reduce Aggregate Enrich Structure
Create datasets :
Use relational technology to assemble and match data
Analyze
Data modelings :
Build statistical models comparing data sets variables and datasets
Find correlations :
Bring additional value to classical market analysis
Decide
Anticipate :
Estimate likelihood of external factors
Make predictive decisions :
New market positions to enhance margins
Control
A datasets clean of errors :
Warning mechanism and auto correction replacing wrong data with substitutes
Report
Sort through to axtract the useful and stay synthetic :
Give it meaning and make it exploitable

**Figure 5:** Implementation of successful Big Data strategy requires a structured and robust framework ,from acquisition of info to decision making process

The statistical and mathematical reality follows this logic: a linear regression model using, for example, 10 predictors on a probability sample correctly derived from 100,000 observations will be as accurate as a model constructed from 100 million observations **[18]**. Contrary to the claims of some players in the wake of big data claiming that "all data must be processed", the truth is that the accuracy of a model is based on the sample quality (each observation the population to have a known probability of being fired) and its size relative to the complexity of the model. Regardless of the size of the population. This is why, for example, are generally obtained remarkably accurate results from the closure of polling stations the election night even though these estimates are based on only a few thousand voters nationally.

### *3.3.4.a Mapreduce sampling, data compression, data selection:*

There are different sampling algorithms (MapReduce) very effective for distributed file systems. These algorithms can be an excellent approach to exploit these massive data in a perspective of simple and effective predictive modeling, and get a quick return on investment vis-à-vis the storage infrastructure **[19]**. For most practical applications, this is an excellent strategy, for example, by deploying analytic platforms and Data Mining as an analytical tool in addition to interfaces to the distributed file system (big data) that perform the steps data preparation / aggregation and / or probability sampling using map-reduce algorithms. Further aggregation and data sampling, the system can also achieve the necessary detailed selection data (e.g., based on a micro-segmentation of specific groups of customers) to send data to the flat analytical Platform which will then build accurate models for specific segments (for example, financial service offerings for high value homes) **[14]**.

### *3.3.4.b Integrating analytics platform with free software:*

A unique feature of the platform and data mining is that it was developed from the outset as a platform calculation for the company, using universal programming languages and standard data interfaces. So we can easily integrate this platform, in addition to high-performance tools for the platform, the emerging free tools for management and data preparation, or specialized analytical procedures using the map-reduce technology. These procedures are handled across the platform, just like any other analytic node in the analytic processes. For example, the open-source R platform is frequently used to implement procedures and highly specialized statistical calculations and analytics platform is compatible with the platform R for many years with the basic integration R scripts in analytical processes **[12]**. The analysis and use of big data is in full emergence but is also evolving very quickly. It is important that the analytical platform organized around the distributed file system can easily integrate new methods of preparation and aggregation of data, sampling and layering to monetize made the investment as quickly as possible in the distributed file system.

### *3.3.4.c Implementation of specialized procedures via map-reduce:*

In addition to easy integration with platforms and other open-source tools, it is important that the chosen analytical platform provides the ability to customize the analytical process to meet specific analytical requirements based distributed file system the big data. Practical applications and best practices in the field of big data analytics are emerging and evolving rapidly; there is no universal consensus opposed to the analytical and predictive analytics "traditional" that are well documented otherwise **[8]**. However, this situation can change quickly since all major database vendors and BI tools (Microsoft, Oracle, Teradata and others) offer interfaces and tools for accessing and processing data effectively.

In any case, analytic platforms allow us to build our own implementation of specific analytic approaches using data from distributed file systems, but also recognize interfaces and turnkey tools accessible through the custom interfaces of major publishers. The latter approach is probably the approach most effective and most "natural" to bring analytics to big data.

## *4. Practical considerations for implementation:*

To summarize, big data essentially refers to unstructured information stored in a distributed file system whose individual data is scattered over hundreds or even thousands of hard disks and servers. The size of these distributed file systems can easily exceed several petabytes (several thousand terabytes) with current technologies. For basic data preparation, cleaning, and data extraction operations, it is more efficient to perform the respective analyzes on the site (the specific server) where the data are stored (to reduce and aggregate the data under form of summary statistics) using a map-reduce approach **[9]**.

### *4.1 Take advantage of Data Profusion to Build a Large Number of Models:*

As discussed earlier in this paper, the real interest of big data in distributed file systems is not to compute global (predictive) models using all available data, but rather valid data samples; in both cases, the results and the accuracy of the models will be the same. However, it is much more appropriate to use this wealth of data and tools available to segment effectively, and build a large number of models with smaller classes. For example, we can expect that upselling models built on larger segmentations will produce less accurate results than a large number of models built on smaller segmentations. Therefore, one way to take advantage of big data and use the information available is to build a large number of models, with a large number of segments, and to use these models to score (predict) observations using the most suitable model **[11]**. By pushing to the extreme, we could have a separate model for each individual of a giant data warehouse, to predict future purchases. The analytical platform connected to the data warehouse must therefore be able to manage several hundred or several thousand models, while offering the possibility to recalibrate these models at will, if necessary.

### *4.2 Deployment models for the scoring in real time:*

One of the fundamental components of analytical platforms used to score new data in real time using Web Services. In this environment, it is possible for external programs to call managed models (with version control) through the platform to score new data transmitted either directly through a remote call to the system, either by intermediary of an identifier designating a particular observation or a specific group of observations to score**[6]**. Regarding the massive data and distributed file systems, the scoring process is the same, whether data stored in relational databases or data stored in distributed file systems. The main challenge to maintain acceptable performance lies in the management and preparation of data, but these steps can be performed using MapReduce tools for preparation and extraction of data or considering other architectures. There may be specific data warehouses dedicated to analytical and scoring based on relational databases and ETL routines supplied by map-reduce, or one of the emerging technologies based on RAM clouds where the distributed file system itself is stored on the "disks" very fast memory for very fast access time to data. There are also a number of available commercial solutions such as Oracle Extreme, etc ..., all of which were designed with the same objective: access to huge data warehouse very quickly.

#### *4.2.1 Reviews of Big Data Strategies, Strategies for Implementation:*

Several platforms exist for several years and advise clients on best analytical practices in order to guarantee a rapid return on investment. Some platforms are limited only to their role as publisher of analytical solutions, and sell any hardware or dedicated storage solution **[15]**. Over the years, we have experienced several new technologies that are going through the usual cycle of a great initial enthusiasm, success for the pioneers and maturation through standard solutions and processes to maximize ROI. Inevitably in this scheme, there may be setbacks and disappointments when the initial promise of new technologies does not materialize. So there are a number of things to keep in mind:

### *4.2.1.a The Massive data do not necessarily provide better knowledge (Customer):*

Suppose that we have access to the price of all shares of the stock market of Morocco, and that we recorded these courses, second by second, with finally a mass of pharaonic data. Suppose further that we have access to a number of sensitive and targeted information concerning certain indicators and financial results of the company. What strategy should be favored to become a good trader. No doubt the second. Storing massive amounts of data describing easily observable phenomena of reality does not necessarily translate into a better knowledge of this reality. It's the same thing if we analyze stock quotes, tweets, medical data or marketing, or complex machineries data to perform predictive maintenance **[16]**. For example, it can be much more relevant to a furniture store and reliable design to have a list of prospects interested in furniture and home furnishings, with demographic information and household income indicators, rather than a massive amount of data on the browsing online courses on various furnishing sites. As is the case with any project to optimize organizational performance and business, it is important to ask a number of questions such as: "What should look ideal results", "How do I measure the success of my project (ie when I finished, and I won)," or "What information will I need to stretch towards the ideal results ". The answers to these questions may well lead to the establishment of a big data warehouse with analytical platform; but often this is not the case.

### *4.2.1.b Velocity data and response time:*

Another aspect to consider regarding the velocity of data or the speed with which the data are updated. The real issue is the "reaction time" needed. In other words, we can build models in a production environment to predict impending problems with one second ahead on the continuously collected database of thousands of parameters (Figure 6). However, if an engineer needs two hours to understand and take the necessary corrective measures, the system is absolutely useless.

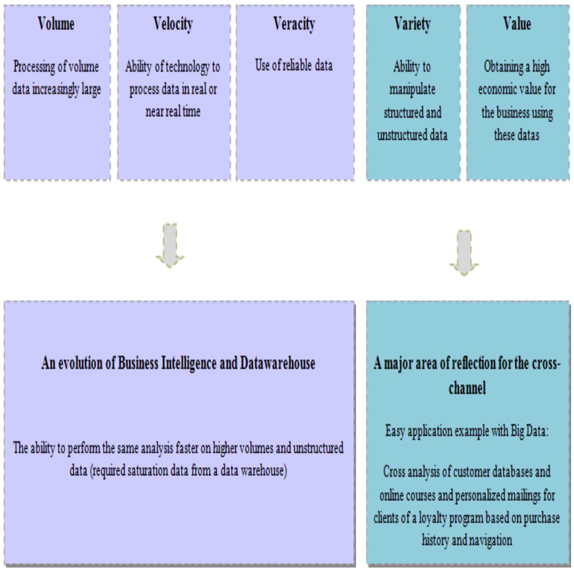
Volume
Processing of volume data increasingly large
Velocity
Ability of technology to process data in real or near real time
Veracity
Use of reliable data
Variety
Ability to manipulate structured and unstructured data
Value
Obtaining a high economic value for the business using these datas
An evolution of Business Intelligence and Datawarehouse
The ability to perform the same analysis faster on higher volumes and unstructured data (required saturation data from a data warehouse)
A major area of reflection for the cross-channel
Easy application example with Big Data:
Cross analysis of customer databases and online courses and personalized mailings for clients of a loyalty program based on purchase history and navigation

**Figure 6:** Big Data dimensions and main attributes

Similarly for a furniture store, it is important to receive an "alert" a month or two before a real estate transaction, rather than real-time information after completion of this transaction, when the prospect has already begun prospecting on the Internet to buy the furniture. An early warning would enable a professional to undertake various steps at this prospect before it begins its process of buying furniture, to propose special offers or convince him to go into a store and build a relationship personal privileged with the sign. Again, a real-time platform of Internet click stream may not be the ideal data warehouse to drive traffic and build a loyal customer base **[13]**. In general, the right approach is to define carefully from the beginning, what will be the final use and strategies for success. At this point, the required reaction times become evident, which sets the need, which obviously derives an optimal system for collecting and storing data and an analytical strategy.

## *5. Performance Analysis:*

The purpose of this article is to provide a brief overview of the specific challenges posed by big data, that is, terabyte data warehouses up to several petabytes of data (or more), and technologies and approaches to address these challenges in order to extract value from the massive data. The technology of distributed file systems deployed on popular servers and storage systems, made possible but also economically viable, the creation and maintenance of these warehouses **[17]**. On these systems, instead of storing data on a single file system, data is stored and indexed on several (sometimes thousands) hard drives and servers, with an index (the map part) redundant to know the specific place where specific information is located. Hadoop is probably the most common system to date, using this approach. To process data in a distributed file system, it is necessary for simple calculations such as calculating workforce, preparation and elementary aggregation, etc ...is performed at the physical location where the data are located in the distributed file system, rather than moving the data to the analytical calculation engine. The map portion of the respective calculation algorithms then control the individual results then the aggregates (the reduce part); this scheme for the implementation of computational algorithms is known as Map-Reduce **[19]**. In practice, the true value of big data rarely lies in calculating statistical results achieved on the completeness of data; in fact, there are statistics that show that bases these calculations produce more accurate results in most cases. However, the true value of big data, especially data mining and predictive modeling, lies in our ability to "micro segmenting" the available information in small groups and build a large number of specific models these small groups of observations. We also discussed other general considerations concerning the value of big data in this paper from the perspective of implementation; the big data analysis platform must be able to integrate emerging technologies algorithms that are often free projects in the public domain.

## *6. Conclusion*

Big Data techniques provide exciting new opportunities for organizations and new challenges for statisticians and data managers. While the framework presented in this paper tries to show an approach that addresses the scaling problem of statistical algorithms. There remains a number of constraints and to develop improvements. On one hand, the size and volume as well as high data dimensionality introduce statistics still to be resolved such as the accumulation of noise, false correlations, heterogeneity and measurement errors. On the other side, along with Hadoop, has developed another much more efficient framework: the Spark project that will be the subject of our next work. Although the idea of distributing the computations on a set of data nodes, has been retained as Hadoop, the Spark framework performs distributed computations directly into memory, which significantly accelerates the iterative calculations that are often encountered in the statistical algorithms.

## *Acknowledgment:*

We would like to thank the CNRST of Morocco (I 012/004) for support.

## *References:*

**[1]** Azzeddine RIAHI, Sara RIAHI,’’ The Big Data Revolution, Issues and Applications’’, International Journal of Advanced Research in Computer Science and Software Engineering’’, Volume 5, Issue 8,ISSN: 2277 128X, August- 2015, pp. 167- 173.

**[2]** Sara Riahi, Azzeddine Riahi, ’’ Innovation and opportunities of Big Data: promote business Intelligence’’, ISSN: 2395-1303 ,International Journal of Engineering and Techniques - Volume 3 Issue 6, Nov-Dec 2017,pp 471- 481

**[3]** *Zhang* Tingting, Qu Haipeng,’’ Range Query on Big Data Based on Map Reduce’’, IJMER | ISSN: 2249–6645 | www.ijmer.com | Vol. 4 | Iss. 2 | Feb. 2014.

**[4]** C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, “Secure ranked keyword search over encrypted cloud data,” in Proceedings of ICDC .IEEE, 2010, pp. 253 – 262.

**[5]** Vahid Ashktorab1 , Seyed Reza Taghizadeh2 and Dr. Kamran Zamanifar3 ,” A Survey on Cloud Computing and Current Solution Providers”, International Journal of Application or Innovation in Engineering & Management (IJAIEM), October 2012.

**[6]** S. M. K. R. Sahal and F. A. Omara, "GPSO: An improved search algorithm for resource allocation in cloud databases", in Computer Systems and Applications (AICCSA), 2013 ACS International Conference on, (2013), pp. 1-8.

**[7]** Rodriguez MA, Buyya R,"Deadline based resource provisioning and scheduling algorithm for scientific workflows on clouds" ,IEEE Transactions on Cloud Computing. 2014; 2(2):222–35

**[8]** Mrs. Premalatha P, Mrs. Marrynal S. Eastaff, " Big Data and Cloud Computing" , International Journal of Engineering and Applied Sciences (IJEAS)ISSN: 2394-3661, Volume-2, Issue-11, November 2015

**[9]** Pandey S, Voorsluys W, Niu S, Khandoker A, Buyya R (2012) An autonomic Cloud environment for hosting ECG data analysis services. Future Generation Computer Systems 28: 147-154.

**[10]**Ei Ei Mon, Thinn Thu Naing“THE PRIVACY-AWARE ACCESS CONTROL SYSTEM USINGATTRIBUTE-AND ROLE-BASEDACCESSCONTROL IN PRIVATE CLOUD”978 -1-61284-159-5/11,2011IEEE

**[11]**Navjot Sekhon, Richa Mahajan ,’’Data Security in Cloud Computing Using HDFS’’, International Journal of Computer Science Trends and Technology IJCST) Volume 5 Issue 2,ISSN: 2347-8578 ,Mar Apr 2017

**[12]** Prashant V. Dhakad, Krishnakant Kishore**,** ’’ Processing of Real Time Big Data for Using High Availability of HadoopNameNode’’, International Journal of Computer Systems, ISSN-(2394-1065), Vol. 03, Issue 05, May, 2016

**[13]** E. B. K. Manash and T. U. Rani, "Cloud computing- A potential area for research", International Journal of Computer Trends and Technology (IJCTT), Volume: 25, no.1, pp.10-11, 2015.

**[14]** Gandomi Amir and Haidar Murtaza (2015) Beyond the hype: Big data concepts, methods,Analytics. International Journal of Information Management, 35, 137-144.

**[15]** R.Devankuruchi "Analysis of Big Data Over the Years" International Journal of Scientific and Research Publications, Volume 4, Issue 1, January 2014 1 ISSN 2250-3153

**[16]** Furqan Alam, Rashid Mehmood, Iyad Katib, Nasser N. Albogami, Aiiad Albeshri, "Data Fusion and IoT for Smart Ubiquitous Environments: A Survey", *Access IEEE*, vol. 5, pp. 9533-9554, 2017, ISSN 2169-3536.

**[17]** Dobre, C., and F. Xhafa. 2014. "Parallel Programming Paradigms and Frameworks in Big Data Era." *International Journal of Parallel Programming* 42 (5): 710–738.

**[18]** García, S., J. Luengo, and F. Herrera. 2015. " Data Preprocessing in Data Mining." *Intelligent Systems Reference Library* 72. doi:10.1007/978-3-319-10247-4

**[19]** Zaheer Khan, Ashiq Anjum, Saad Liaquat Kiani, "Cloud based Big Data Analytics for Smart Future Cities", IEEE/ACM 6 th International Conference on Utility and Cloud Computing, 9-12 Dec 2013, Dresden, pp 381- 386, DOI: 10.1109/UCC.2013.77