Big Data, indispensable today

Radu – Ioan ENACHE, Marian Adrian ENE
Academy of Economic Studies, Bucharest, Romania
radu.enache91@gmail.com, ene.marianadrian@outlook.com

Big data is and will be used more in the future as a tool for everything that happens both online and offline. Of course , online is a real hobbit, Big Data is found in this medium , offering many advantages , being a real help for all consumers. In this paper we talked about Big Data as being a plus in developing new applications, by gathering useful information about the users and their behaviour.We've also presented the key aspects of real-time monitoring and the architecture principles of this technology. The most important benefit brought to this paper is presented in the cloud section.

Keywords: *Big Data, Cloud Data, Financial Data, Data encryption*

1 About Big Data

From the beginning, data was always highly structured. All the data was divided into fields, those fields had a specific length, and the data that entered in every field was constrained by a set of rules. Today, most of the data entered by humans is unstructured, having the form of free text. This text comes email messages, tweets, documents and so on. [3]

Big Data is the latest trend emerged in data management. It is defined as any data source which has three features:

- Very large data volume;
- Very high data transfer rate;
- Huge variety of data.

Big data is very important because it enables organizations to collect, store, manage and manipulate large amounts of data at optimal speed, to acquire necessary information. This is not a single technology, but rather a combination of 50 years of technological evolution. Each wave in data management is born from the need to solve some problems like this. [1]

Stages of development for data management in the last 50 years have culminated to the point where we are today: the beginning of the big data age.

1. <u>Creating manageable data structure</u>

As computers have moved on the commercial market, in the late 60s, the data was stored in files that had no structure. When companies understood their consumers at a more detailed level, they had to apply brute force, including detailed models of programming to reach the required data.

Later, after the mid-70s, things have changed with the advent of the relational database management system and the relational database, requiring a structure and a method for improving performances. Finally, a relational database got to store objects of type BLOB (binary large objects).

2. <u>Content Management and the Web</u>

Most of the data available today aren't structured. Paradoxically, companies have turned their investments to structured data systems. Content management system,in the business field, has evolved in the '80s, so companies could have the ability to better manage unstructured data, mostly being documents.

In the '90s, with the rise of the Web, organizations wanted to move beyond documents to store and manage Web content, images, audio and video. The market has evolved from a set of disconnected solutions on a unified model that brought together these elements, in a platform that incorporates business process management. [1]

3. <u>Administrating big data</u>

Big data is a new technology that was derived from data management history. It is built on the evolution of 50 years of data management practices. With big data, it can be made a big virtualization of data that can be stored efficiently, and using cloud storage methods, it can do so at a lower cost.

Variety is one of the main principles of big data. Although data management exists form a long time, in the world big data, two factors have recently emerged:

- ✓ Some big data sources are new, such as data generated by sensors, mobile or tablets;
- ✓ Data that was created in the past, was not captured or stored and analyzed in a useful manner. The main reason for this is that the technology did not exist to do this. In other words, there is an effective way in terms of the cost to analyze all that data.

The term “structured data” generally refers to data that have a length and a well-defined format. Examples of structured data can be numbers and groups of words called strings. Most experts say that this type of data represents about 20% of all existing ones. Structured data is typically stored in a database and can be queried using programming languages such as SQL. [1]

The data sources are divided into two categories:

- ➢ The ones generated by computers or machines: machines generated data refer to those created without human intervention;
- ➢ The ones generated by humans: These types of data came from the interaction of people with PCs.

Some experts argue that there is a third category, which is a hybrid between the two above.

2. Automatic generated data

The data generated by computers or machines may include the following:

- ❖ The data collected by sensors: Examples are radio waves, medical devices or GPS;
- ❖ Captured data from the web: When servers, applications and networks operate on the internet, they capture a variety of data, following their activity. Their volume can be very high and useful for creating service level agreements or to predict future security breaches;
- ❖ Information about sales: When the cashier scans the barcode of a product purchased, that product and all of it’s related data are generated and recorded.
- ❖ Financial Data: Many systems are now scheduled. These are made based on predefined rules that automate processes.

Structured data that can be generated by man are:

- ❖ Input: These can include any type of data that a person can enter into a computer, such as names, ages, incomes, responses accumulated after some polls. These data can be used to understand consumer behavior.
- ❖ Data generated from clicks: These data is generated every time someone clicks on a link, within a certain website. These data can be analyzed to understand the behavior of consumers and their habits of purchase.
- ❖ The data generated from games: Every move you make in a game is recorded. These data can help understand the behavior of players in a certain application of gaming.

Real-time aspects of the big data can be revolutionary when companies need to resolve issues of significant difficulty. In general, this approach is only relevant in real time when the response to a question is urgent. It may be related to something important like a hospital detection equipment performance or anticipation of a security breach. Below is a list of some examples when some companies can benefit from this data in real time:

- ✓ Monitoring to identify some problems with certain information such as fraud and intelligence;
- ✓ News monitoring and social networks to identify events that could have a negative impact on financial markets

such as consumer reaction when a new product is announced;

- ✓ Change of location where ads are placed during a sporting event;
- ✓ Offering a discount coupon for a buyer, during a sale. [1]

In terms of data architecture, principles of good design are critical when creating or when migrating an environment to support big data, speaking of storage, reporting, analysis or applications. In the process of creating the environment, it should be considered the hardware, software infrastructure, well-defined APIs and even development programs. The architecture must be able to address all fundamental requirements:

- Capturing;
- Integrating
- Organizing;
- Analyzing;
- Action.

3. Technology layers

Big Data technology layers are:

A. Physical Infrastructure

At the lowest level of the architecture of a system, big data's physical architecture is being represented by hardware, networks and others. Big data implementations have specific requirements regarding the architecture's elements. It is important to bear in mind several principles that can be applied in this case, such as:

- ▪ Performance: This is also known as latency and is often measured with a single transaction or a single demand;
- ▪ Validity: This is a percentage and is computed based on service availability in a given period of time;
- ▪ Scalability: This is represented by the size of the infrastructure, storage space and processing power;
- ▪ Flexibility: This is represented by the time when there can be added new infrastructure resources, or how long the service can return to normal in the event of a system failure;
- ▪ Cost: This principle is laid down by the cost of equipment.

B. Infrastructure security

Privacy and security requirements of big data are similar to those of conventional media data. Security was aligned with business requirements. Sometimes unique challenges arise when the big data is part of a strategy, such as:

- ▪ Access to data: User access to the processed big data or not, has almost the same technical requirements as implementations which do not relate big data;
- ▪ Access to application: Most APIs provide protection against the use or access. The level of protection is probably adequate for most big data implementations.
- ▪ Data encryption: Data encryption is the biggest challenge related to the security of a big data system in one environment. In traditional media, encrypting and decrypting data require a lot of resources, but because of the volume, variety and velocity associated with big data technology, this issue is no longer valid.
- ▪ Threat Detection: The inclusion of mobile devices and social networks has grown exponentially the amount of data, but also the number of threats.

C. Operational databases

At the base of each environment more big data system databases contain data of all collections relevant to a particular business. These systems must be fast, scalable and resilient. These systems are not created the same, so one from an environment may be different from another in other environment. SQL is the most used programming language for querying databases, but other languages can provide solutions to some of the big data challenges. Also, you can use other alternatives such as Python or Java. It is very important to understand what types of data are handled by the database or if it supports transactional behavior. Designers describe this database by the acronym ACID. This means:

- Atomicity: A transaction represents "all or nothing" when it is atomic. If any part of the transaction fails, the entire transaction fails;
- Consistency: Only transactions that contain valid data will run on the database. If the data is in a damaged condition, the transaction will not be completed and no data will be placed in database;
- Isolation: multiple transactions, simultaneously, will not interfere with one another. A valid transaction will be executed until it is completed, and their order of execution is influenced by the time they were sent;
- Durability: After transactional data has entered in the database, it stays there forever.

D. The organization of data and the instruments

The organization of data and tools for their capture, validation and assembly of big data components in a certain context is a relevant collection. Because big data is huge, techniques have evolved to process data more efficiently. One of these techniques is called MapReduce and is one of the most used. Organizing data services is in reality an ecosystem of tools and technology that captures and records data for further processing. Technologies in this category include the following:

- A distributed file system: It is necessary to provide scalability and storage;
- Serialization services: This is necessary for persistent storage systems and procedures to run some from a distance;
- Coordination of services: It is necessary for building distributed applications;
- Tools for extract, transform and load (ETL): They are necessary for loading structured and unstructured data in Hadoop. With YARN, Hadoop V2's Job Tracker has been split into a master Resource Manager and slave-based,ApplicationMaster processes. It separates the major tasks of the Job Tracker:resource management and monitoring/scheduling. The Job History server now has the function of providing information about completed jobs. [2]

E. Deposits of analytical data

Data warehouse and data mart are the two techniques that organizations use to optimize your time and help in making decisions. Because many data warehouse and data mart sites contain data accumulated from different sources within a tax refund companies, costs of data normalization are not ignored. With big data, there are some differences:

- Traditional data sources can produce detailed separate data;
- Lots of data sources exist, each needing a certain degree of manipulation before the data accumulated can be used in a business;
- Content sources also should be cleaned and they may need different techniques that are used in the structured data.

Existing tools and techniques for analyzing Big Data are invaluable. Since these algorithms do not work in normal parameters because of large data flows, they should be optimized.

Big data custom applications offer an alternative distribution and examination of data sources. Since all components are important in a big data system, this is the place where it is the most activity in terms of innovation and creativity. These applications are aligned horizontally, that addresses common problems in companies or vertically oriented, intended to solve one problem. [1]

Virtualization in a big data system is very important. The separate service resources, allow the creation of several virtual systems in a single physical system. One of the reasons why companies have implemented virtualization is to improve the performance and efficiency of processing.

Using a set of distributed resources, such as servers, in a more flexible and efficient it delivers significant benefits in terms of reducing costs and increasing productivity. This practice has several benefits and among these are:

- Virtualization can improve utilization of physical resources;
- Virtualization allows better control over the use and performance of IT resources;
- Virtualization can provide a level of automation and standardization to optimize the computing environment;
- Virtualization provides the foundation for a cloud computing system.

4. Cloud and Big Data

We all know the power of cloud is that users can access whenever required storage resources with little or no IT support or the need to purchase more hardware or software . One of the key characteristics of cloud is elastic scalability : Users can add or subtract resources almost in real time based on changing requirements . Cloud plays an important role in the world of Big Data. Dramatic changes happen when these infrastructure components are combined with advances in data management .

Bed and optimized infrastructure supports the implementation of Big Data.

Cloud computing is a method of providing a set of shared computing resources including applications, computing, storage, networking, development, and deployment platforms and business processes.

A popular example of the benefits of cloud is so big that support data can be observed both Google and Amazon.com. Both companies depend on the capacity to manage massive amounts of data in bulk to make things right direction. These providers need to come up with the infrastructure and technology that could support applications so massive a scale. Consider the millions and millions of Gmail messages that Google processes daily as part of this service. Google has been able to optimize the Linux operating system and software environment to support e-mail in the most efficient way; therefore, it can easily support hundreds of millions of users.

Even more important, Google is able to capture the massive amount of data about both: users of mail and search engine users to develop business.

Two key cloud patterns are important in the discussion of Big Data - public cloud and private cloud. For organizations that adopt and implement cloud delivery models, most will use a combination of private calculation made by an external company for the sharing of a variety of customers who pay a per-use fee. How these companies argue in the public and private balance depends on a number of hatches, including privacy, latency, and purpose.

Public Cloud

Public Cloud is a set of hardware, networking, storage, services, applications, and interfaces owned and operated by a third party for use by other companies and individuals.

These commercial providers create extreme scalability of data center infrastructure that hide underlying details of the consumer.

Public cloud is viable because typically manage workloads relative or simple repetitive. For example, Email is a very simple application.

Therefore, a cloud provider can optimize the environment so as to be best suited to support a large number of clients, even if many messages saved

In contrast, the data center (data center) to bear so many different applications and workloads that cannot be easily optimized. A public cloud can be very effective when an organization runs a complex data analysis and needs more computing cycles to handle the load. In addition, companies can choose to store data in a public cloud where the cost per gigabyte is relatively inexpensive compared with an acquiring the storage.

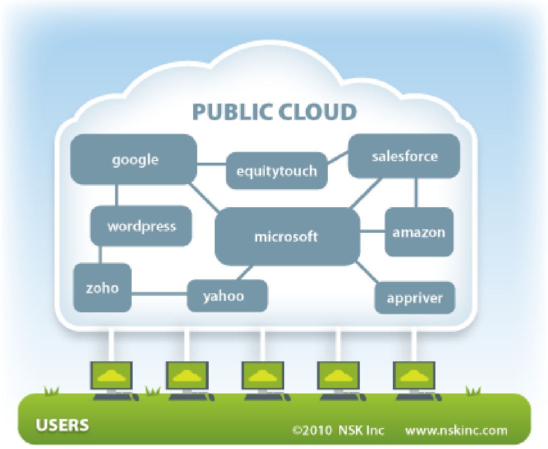

Fig. 1. Public Cloud

Therefore, all Public cloud environments are not the same. Some are scalable managed services with high security and high level service management. Other are less robust and less secure, but they are much less expensive to use.

Your choice will depend on the nature of big data projects and the amount of risk it can be assumed.

Private cloud

A private cloud is a set of hardware, network, storage, services, application and interfaces owned and operated by an organization for use by employees, partners and customers. A private cloud can be created and managed by a third party for the exclusive use of a company.

It is a highly controlled environment, not open for public consumption. The private cloud behind the firewall. Iit is highly automated with a focus on governance, security and compliance. Automation replaced manual processes of IT service management to customer support. In this way, business rules and processes can be implemented internal software so that the environment becomes more predictable and manageable.

If organizations are focused on managing a project of Big Data and applications that are processing massive amounts of data, private cloud might be the best choice in terms of latency and security.

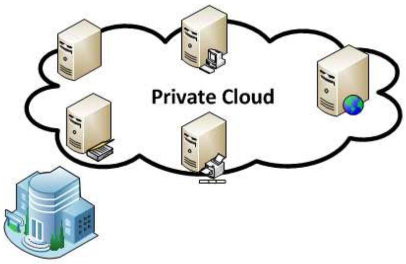

Fig. 2. Private Cloud

A Hybrid Cloud is a combination of a private cloud combined with the use of public cloud services with one or more touch points between environments .

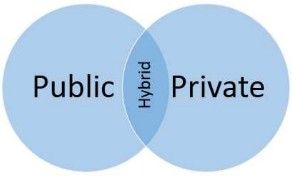

Fig. 3. Hybrid Cloud

The goal is to create a well-managed cloud environment that can combine services and data from a variety of cloud models to create a unified, automated, and well-managed computing environment.

Web companies are among the early adopters of big data, largely because of the volume of unstructured information that they must deal with on a regular basis. However, even traditional industries such as telecommunications, retail, financial services, and healthcare are launching pilots and testing the waters to see what big data has to offer. [5]

In fact, a large number of cloud features, define an ecosystem of Big Data. Here are some of them:

Scalability

Scalability in terms of hardware refers to the ability to go from small to large amounts of processing power with the same architecture. In terms of software, it refers to the consistency of performance per unit of energy and increase hardware resources. Cloud can scale to large volumes of data. Distributed computing system, an integral part of the cloud, really working on a plan to "divide and conquer". So if you have large volumes of data, they may be divided on "cloud servers".

Elasticity

Elasticity refers to the ability to expand or shrink computing resources in real time based on need. One advantage is that cloud community, customers have the potential to access a service as much as they need when they need it. This can be useful for Big Data projects which could expand the value of the resources to cope with the volume and speed data.

Of course, this feature is very attractive to end customers means that a service provider needs to design a platform architecture that is optimized for this type of service.

Pooling resources

Cloud architecture enables efficient creation of shared resource groups that make the cloud economically viable.

Pay as You Go:

A typical billing option for a cloud provider is Pay As You Go (PAYG), which means you are charged for the resources used by the court based pricing. This can be useful if you are unsure of what resources you need for your Big Data project (unless you are under budget).

Market players Cloud

Cloud Players come in all shapes and sizes to provide more differentiated products. Some are household names while others are newly emerging. Providers offering cloud services for Big Data projects are:

Amazon.com, AT & T, GoGrid, Joyent, Rackspace, IBM, and Verizon / Terremark.

However, companies cloud and cloud service providers are also suppliers of software solutions specifically targeting Big Data projects.

5. Practices Big Data

While we are in an early stage in the evolution of big data, it is never too early to start with good practice, so you can set up what is learning and experience is gained. As every important emerging technology, it is important to understand why you need to leverage technology and a concrete plan.

A. Understand your goals

Many organizations begin their journey by experimenting Big date with one project that could provide some tangible benefits. By selecting a project, test freedom without risking capital expenditure. However, if you get to do a number of specific projects, you probably will not have a good plan when you begin to understand the value of leveraging Big Data within the company. Therefore, after ending some experiments and have a good initial understanding of what might be possible, you need to set some goals - both short and long term. What you want to achieve with the Big Data? It is important to have collaboration between IT and business units to better define your goals.[1]

B. Establish a trajectory

After setting goals, Amazon is a way you could define later. The company expert in exploiting analytical data to create customer intimacy as a competitive advantage. Recommendations are built on the data type "Might you interesting and ..." "Those who bought this item also bought ...". It is an example of using exploited getting better and ecommerce players in Romania.

C. Discover your data

No company ever complains you hold too little data. In reality, swim in date. The problem is that some companies do not know to use these data to predict future pragmatic, execute important business processes, or simply gain new insights.

Big time strategy aim must be to find a way to leverage data for business results more predictable. But it must go forward, and start by embarking on a process of discovery. This process will provide a lot of perspectives.

For example, let you know how many sources of data you have and how much overlap there. This process will also help you understand who those sources goals.

D. Understanding of technological options

Now, you understand your company's objectives, have an understanding of what data you have, and you know what data are missing. But as take steps to execute the strategy? You must know what technologies are available and how they could help the company produce better results. Therefore, do your homework.

Begin to understand the value of technologies like Hadoop, offers products and complex data processing streaming events. You should look at the different types of databases, such as memory databases, spatial databases, and so on. You should familiarize yourself with the tools and techniques that are being developed as part of the big data ecosystem. It is important that your team has an understanding of the technology available to make informed choices.

Public health

The use of big data can improve public health surveillance and response. By using a nationwide patient and treatment database, public health officials can ensure the rapid, coordinated detection of infectious diseases and a comprehensive outbreak surveillance and response through an Integrated Disease Surveillance and Response program. [4]

6. Conclusions

From our point of view, Big Data is being indispensable for everything related to online content . At the moment, speaking as a whole, Big Data is used in large capacities, but it's usage will increase in the future. The content consists of information gathered so far, and research does not stop there, which leads us to think about Big Data.

Do not overlook the need to manage the performance of your data!

Big Data demonstrates that we are able to use more data than ever before in a faster rate of speed than was possible in the past. This ability to get multiple perspectives is a huge benefit. However, if the data is not managed in an efficient manner, it will cause serious problems for the enterprise. Therefore, we need to build flexibility on our path to build our Big Data plan.

References

[1] Brand, W., 2013. *Big Data For Dummies*, New Jersey: John Wiley & Sons;

[2] Frampton, M., 2014. *Big Data made easy*, New York: Apress;

[3] Berman, J., 2013. *Principles of BigData*, Waltham: Elsevier;

[4] McKinsey Global Institute, 2011.*Big data: The next frontier for innovation,competition, and productivity*;

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

Radu Ioan ENACHE graduated in 2013 from the *Economic Marketing Faculty* at the *Academy of Economic Studies* in Bucharest. At the moment he is pursuing the *Database for Business Support* master program. His area of interest are: Mobile Apps, Web Development and website optimization.

Marian Adrian ENE graduated in 2013 from the *Economic Marketing Faculty* at the *Academy of Economic Studies* in Bucharest. At the moment he is pursuing the *Database for Business Support* master program. His area of interest are: Databases, Web Development and multimedia applications.