International Journal of Economics, Commerce and Management
United Kingdom Vol. VI, Issue 8, August 2018
http://ijecm.co.uk/ ISSN 2348 0386

BIG DATA IN GOVERNMENT: SPECIAL REFERENCE TO ALBANIA

Selma Kaçaniku
Lecturer at Faculty of Business, University of Durres, Albania
selma-kacaniku@hotmail.com

Ejona Duci
Lecturer at Faculty of Business, University of Durres, Albania
jonaduci@yahoo.com

Enida Istrefi
Lecturer at Faculty of Business, University of Durres, Albania
enida.istrefi@gmail.com

Abstract

Platforms for open data are increasingly used in various fields of science, society and business, including urban planning, cultural heritage protection, crime prevention and more. Open data is becoming more prevalent and necessary and the need for effective and rapid analysis is increasing. Different methods of gathering and processing of open data provides data that the users can store and analyze according to their needs. Some requirements to open data are imposed as to be readable and regularly updated. The aim of this paper is to give an idea about what big data is, different types of it, how it is used by government and an example from Albania. Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Challenges include capture, storage, analysis, search, sharing, transfer, visualization, querying, updating and information privacy.

Keywords: Big data, open data, open government, data types, data governance

INTRODUCTION

Big Data was born with the objective of covering needs not satisfied by existing technologies, such as the storage and processing of large volumes of data that have very specific characteristics defined as the three V's (there may be more):

Volume, refers to the size of the data that can come from multiple sources.

Velocity, defines the speed with which data arrives using units such as tera, peta or exa bytes

Variety, we talk about data (Justin Ellingwood 2016): Structured; Semi-structured and Unstructured.

An important feature about data, is that they are considered as the source of the truth, that is, they do not alter during their treatment. The underlying technology in Big Data is Apache Hadoop, it now has eight years of history, but what is Hadoop? **Hadoop** is a distributed operating system that allows large volumes of data to be processed in parallel on conventional hardware. It is a type of special operating system, since it works on another one like Linux or Windows using the implementation of Hortonworks (Anismaj 2015). It has the following characteristics:

- Scalable, allows to create clustered structures, to which is possible to add new nodes easily.
- Flexible, adapts to multiple data formats, can use or not schemes to treat information and allow users to use it at different levels.
- Reliable, it has been designed, considering that the hardware and software can fail.
- Fast and slow, it is very fast to handle large amounts of data, but it can be slow when working with little information.

Hadoop covers a very important area, as is the processing of unstructured data, usually those that are not stored in conventional databases, but that some studies cipher in 95% of the data produced by a company.

Probably where more doubts appear, it is when considering to whom this technology is directed. If we think of large volumes of data, there are a large number of companies that have withdrawn their data history, because it was unfeasible to deal with conventional technologies, such as banking, insurance, research centers, but new needs arise from data processing associated with social networks, since many companies have made a significant investment in them, this opens the Big Data market to a wide range of companies that currently do not perform a data analysis, and therefore, lose the opportunity to improve or generate new lines of business. What in principle looked like a technology aimed at companies with very high volumes of data, is based on the idea that it can also be applied to small and medium enterprises, with very specific needs.

BIG DATA WITH EXPONENTIAL GROWTH

Let's take into account that currently the data growth rate is exponential, so it is logical to think of new tools that help in the treatment of these silos of information, which can come from multiple and different channels, such as:

- Data history, data accumulated during years, that can shed very representative statistics and trends.
- Social networks, very useful if we can filter and analyze the feedback of customers and suppliers.
- Sensors, can generate real mountains of data to evaluate.
- Mobile devices.
- Internet, is a powerful tool if we are able to organize the information we need.

When considering the need to implement a big data implementation, it is important to take into account that on one hand we will have a solution for our structured data (conventional RDBMS) and unstructured or semi (Hadoop) and that we will need to respond to the analysis of data, for this the ecosystem Big data has multiple solutions.

OPEN DATA, BIG DATA AND OPEN GOVERNMENT: DIFFERENT TYPES OF DATA

What do open data and big data have in common? At what point do the open data and open government movements converge? In order to answer these theoretical questions, it would first be necessary to understand each of the three concepts independently. Thus, we can define big data as large datasets that cannot be treated in a conventional way, as they exceed the capabilities of the usual technological tools for capture, management and processing.

The interrelation of open data, big data and open government has given rise to six different subtypes of data with common traits coming from each of the parent categories, as shown in the above chart, but depending on the intersection of these have different characteristics:

1. **Big data, but not open.** Purchase data, clinical information, economic transaction records. A large amount of macro data falls within this category, most of which have a significant commercial value. Their treatment and reuse is of great use both for the public and private sectors since, thanks to them, commercial patterns, demographic trends or epidemiological outbreaks can be predicted.
2. **Open government, without opening of data.** The open government movement advocates for citizen participation in the decisions of local, regional or national governments. However, this current does not always include open data policies and sometimes does not entail the opening of public sector information.

3. **Big data, open but not government**. Currently, there is a large amount of data that does not come from public bodies, but from academic, business or private sources, with an open and reusable approach.
4. **Government data but no big data.** Government data does not have to be massive to be valuable. Publication of a modest amount of public information, such as public transport schedules, can have a positive impact on society through the development of new value added products or services.
5. **Open and private data.** This category includes those private sector data that companies choose to open for their own purposes, for example to satisfy potential investors or improve their corporate reputation.
6. **Open data, big data and open government.** The perfect trinomial. The opening of these datasets can have a great socioeconomic impact on their environment. In fact, according to statistics from the European Data Portal, reusing open data could save 7,000 lives a year or save up to 629 billion hours on the roads.

This diagram is only one of the possible representations of the data ecosystem, a dynamic scenario that is constantly evolving as information technologies advance and governments invest more resources in the storage and opening of data; a sector that, in the European Union alone, translates into a market valued at 75.7 billion euros and savings of 1.7 billion in public spending by 2020 (European Data Portal).

DATA TYPES IN BIG DATA: CLASSIFICATION BY CATEGORY AND BY SOURCE

When creating Big Data projects that detect, consume, manage, organize and present such data in an optimized way and so that they contribute something to our business we generally face the following questions:

- Where do we get the data?
- What data gives more information to business?
- What data is available outside of organization to help them out?
- What volume of data do we have to handle?
- What format do they have?
- How often do we use them?
- How to integrate them into the management system?

Although all these questions are important, the most important of all is:

- What problem do we want to solve?

If we are not clear about the problem, we cannot consider starting to work with data to find a solution. When we have located the problem that we want to solve we can ask the initial questions and extract information. The process of obtaining it from the data is reflected in the famous pyramid DIKW (https://en.wikipedia.org/wiki/DIKW_pyramid) or pyramid of knowledge, which relates four components: Data, Information, Knowledge and Wisdom (Data, Information, Knowledge and Wisdom).

BIG DATA TYPES

The categorization of data is important for any project, and especially when working with large volumes (Big Data). Two of the categorizations most used in Big Data are usually those that relate the structure of the data and those that depend on the origin of the same:

Types of data by categories

Data types are usually organized into 2 main categories (guru99.com/what-is-big-data.html):

Structured:

- Created: data generated by the systems in a predefined way (records in tables, XML files associated with a schema)
- Triggers: Indirectly created data from a previous action (reviews of restaurants, movies, businesses (Yelp, TripAdvisor)
- Transaction-driven: data that results in completing a prior action correctly (self-generated invoices when making a purchase, receipt of an ATM when performing a withdrawal of cash)
- Compiled: summaries of company data, public services of group interest. Among them we find the electoral census, registered vehicles, public housing)
- Experimental: data generated as part of tests or simulations that will allow validate if there is a business opportunity.

Unstructured:

- Captured: data created from a user's behavior (biometric information of movement wristbands, activities tracking applications (running, cycling, swimming, GPS position)
- Generated by users: data that a user specifies (publications in social networks, videos reproduced in Youtube, searches in Google)

Multi-structured or hybrid:

- Emerging Market Data
- E-commerce
- Weather data

Data types by source

Although there is no single criterion for categorizing data types, the most widespread is to divide them into 5 groups:

Web and Social Networks

- Information about clicks on links and elements
- Google Search
- RRSS (Twitter data sources, Facebook postings, other RRSS)
- Web content (pages, images, links, etc.)

Communication between machines

- RFID Readings
- GPS Signals
- Other sensors (parking meters, vending machines, cash machines, etc.)

Transactions

- Communications records (calls, messaging, etc.)
- Billing records (card payments, online payment, etc.)

Biometric

- Face Recognition
- Genetic information (DNA)

Generated by people

- Recordings to customer service operators
- E-mail
- Electronic Medical Records

GOVERNMENT SEEKS TO IMPROVE MANAGEMENT WITH BIG DATA

The Government created a National Observatory of Big Data, within the scope of the Ministry of Information and Communication Technologies, with the objective of implementing new technologies to improve management and other aspects relevant to citizenship.

GOVERNMENT APPLYING BIG DATA, ALBANIA (Portaliqeveritar e-Albania)

The government applying big data?

It sounds like a mission impossible to achieve because of the high costs, disorder in the administrative areas and data that are lost over time without any control. But the generation of a strategy is not very far for governments, there are approaches in data analytics.

The region has the necessary capacity to apply data analysis but there are different challenges that the governments must assume in order to perform these functions. Representatives are proposing as a challenge a change of general attitude in the administrative part of the TICs.

There are several principles that should be part of the 'personality' of officials, not just ICT, but at all levels. The first principle is that 'what is not measured cannot be corrected or improved'; if we only focus on budget execution we lose sight of many other indicators that can allow us to make better decisions and manage more efficiently.

On the other hand, Milena Harito, minister of Information and Communication Technologies, pointed out three key points in which the IT area of the government should be supported to face a state-of-the-art analysis in the government sector.

- The first is to have the digital devices to generate data in the place and time in which it occurs.
- The second is to process that data that is stored in the cloud or another place for a manager to have to process them statistically to generate some type of information.
- The third step is to make the decision, since you have the information. What will be done with it?

For Harito, with these three barriers there are equal number of areas in which it could be definitively applied:

- Health: Wearables can be the key. "All environmental computing technologies that generate data on human health from blood pressure and heart rate to issues such as the digital medical record of each person, that would generate a lot of health data from people who could feed large databases.

- Public transport: "If we had a good system of payments in buses, to be paid with a cell phone or card of some kind, the data would flow when making decisions." From data related to income tax issues on what citizens pay on buses or trains, how people move from one place to another, what their flow is, one could control where they got into where they went down.

Finally, it was necessary to turn the orientation of the way analytical problems are addressed today, going from a more focused environment in performance management analysis of what happened with the data (past), to a more scientific approach trying to find "hidden behavior" in the data (future), and with this to improve the processes at different levels in the institutions of the country.

Is the volume, speed, variety and truthfulness of the data, characteristics of the Big Data, playing against citizens or can it make life easier for us?

It is the time of the "data scientists", the "Internet of the whole" and a new era. The faith in this "new oil", as has already been called the Big Data, is based on the power of information.

But like oil, you have to refine and refine the data so that the analyzes are correct and contain no noise that could distort the conclusions drawn from them.

Today we not only have an unimaginable amount of data of all kinds, (1) from ourselves (wearables, trace of our activities in the network) (2) of the machines and objects with which we relate (from the house to the connected car thanks to the telematics) and (3) of the relation that is established between all of them to each other; but we also have the ability to analyze them in an intelligent way, in real time, and to establish patterns of behavior, which gives us the possibility of making predictions. Now it is possible to perform accurate diagnostics and reliable forecasts and to anticipate the future, improve.

There are numerous examples of the practical application of Big Data to the business world. Once we can analyze in addition to traditional structured data (databases, spreadsheets, etc.) also new unstructured data (emails, interactions in social networks, opinions and feelings, etc.) companies are available to locate tendencies, thus being able to adjust its offer in advance to the demand. We already talk about analysis of feelings or mining opinions, if we attend to the tastes, desires and needs of consumers and we work to satisfy them before they manifest (eg Google Now).

Some examples that the US Government is already developing

- Predicting earthquakes before they are detected

In August 2012, the US Geological Survey tracked thousands of tweets looking for the word "earthquake". Using data on the time and position contained in these tweets, he managed to locate a major earthquake in the Philippine Islands before the seismographs recorded it (Konkel, 2013a).

- Early warning on epidemic outbreaks

Google discovered in 2009 that there was a close relationship between the number of people conducting searches related to the flu and people who actually suffered from flu-like symptoms. At present, this tool, known as Google Flu Trends, offers data on influenza activity in different countries and regions around the world (ncbi.nlm.nih.gov/pmc/articles/PMC4215636/).

THE GOVERNMENT OF THE DATA, LOOKING FOR THE EXCELLENCE OF THE BIG DATA

Large companies, including some banking entities, have embarked on their journey towards digitalization and thereby have increased both the volume of data and the variety of the same and their sources. In recent years it has been observed the growing power of social networks, as well as the emergence of open data projects of the official subjects, increasingly demanded as external sources (Edmund Ingham 2014). The union of all this information offers these

companies invaluable to improve the benefits: anticipation to potential customer leakage through behavioral analysis and predictive models, fraud detection, performance improvement in their processes and detection of new opportunities are just a few examples of using the data as an asset of a company.

If to all this we add another of the great advantages that exist today, as is the possibility of reducing the time to market leveraging the capabilities of the cloud (avoiding, at least in the first instance, the provisioning time machines), this is the perfect situation for the Big Data projects.

Which company is not, or should be already involved in one of them? Now, when entering the vortex of projects, it is necessary to speed up execution over mechanisms of information control. The large distributions on-premise software for mass data processing incorporate control and audit modules. Nevertheless, when it comes to talk about developments in the cloud, it can be found a lot of deficiencies in the centralized management of data step through each of the different stages of the treatment within an application. It is time to give importance which deserves to the Government of the data.

WHAT IS DATA GOVERNANCE?

The Master Data Management Institute (MDM Institute) defines as the formal orchestration of people, processes and technology to enable an organization to leverage data as an asset of the company (information-management.com/mdm). The data governance establishes a framework to regulate processes related to the processing of data, where value is given to the data and to all operations related to it. Since entering the system, it is important to establish which roles can perform certain actions about what data set and under what conditions. Power without control is useless; The value is not only found in the technology or in the advantages previously described. In a Big Data project, data control should be the main priority, since it is what will give value to the results obtained satisfying the demand of the customers.

CONCLUSIONS

In addition to being able to trace the data through the system and to audit access to them, to know the information and to perform a centralized treatment of it avoids possible duplications, which in turn helps mitigate consistency errors or data coherence, avoiding the harmful information parts. The data governance is the icing on the data management within the organization (datagovernance.com/defining-data-governance/). From document management and content management, through master data, metadata and security of the up to the transactions carried out with them, all these processes must be orchestrated and centralized to

fulfill the Government's high priority of data: quality of the data. That is, they are accurate, complete, consistent, accessible, reliable, consistent and unique. To ensure this, the defined framework should include, among other:

• The generation of a data dictionary.
• Validation of any data present in the system.
• Traceability and lineage.
• The ownership of the data.

In order to be able to measure and evaluate compliance with these aspects metrics should be established for each of them. In addition, depending on the sensitivity of the data, comply with a set of rules laid down by the regulatory bodies, which should be measured and audited.

Establishing a control methodology is indispensable. For this it is important to involve all areas of the company: from the technical area, all be aware of the importance of data governance. The ideal is to get the involvement of a stable and centralized team ensure compliance with established standards and carry out the necessary transformation. Thus, the results will soon be seen. The operational efficiency of processes (which work with quality data and do not need multiple validations) versus hours dedicated to data analysis of provenance or format, or decision-making based on reliable data (hindawi.com/journals/tswj/2014/712826/).

REFERENCES

Anismaj June 8 2015, How To Setup Apache Hadoop On CentOS, from https://www.unixmen.com/setup-apache-hadoop-centos/

Big data analytics, Big Data Value Association , from https://www.nuromedia.com/big-data-analytics/

Big Data: Survey, Technologies, Opportunities, and Challenges, July17 2014, from https://www.hindawi.com/journals/tswj/2014/712826/

DIKW pyramid, from https://en.wikipedia.org/wiki/DIKW_pyramid

Edmund Ingham August 20 2014, We're All Marketers Now: The Growing Power Of Social Media And Search Marketing, fromhttps://www.forbes.com/sites/edmundingham/2014/08/20/were-all-marketers-now-the-growing-power-of-social-media-and-search-marketing/#c98d11f5097a

European Data Portal, from https://www.europeandataportal.eu/

European Data Portal, fromhttps://www.europeandataportal.eu/

First steps to get ready for product data initiative, from https://www.information-management.com/mdm

Gwen Thomas, Data governance: The basic information, from http://www.datagovernance.com/defining-data-governance/

Introduction to BIG DATA: Types, Characteristics & Benefits, from https://www.guru99.com/what-is-big-data.html

Justin Ellingwood 2016, An Introduction to Big Data Concepts and Terminology, from https://www.digitalocean.com/community/tutorials/an-introduction-to-big-data-concepts-and-terminology

Portaliqeveritar e-Albania, from https://e-albania.al/

The Use of Google Trends in Health Care Research: A Systematic Review October 22 2014, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4215636/