DATA ANALYSIS IN THE BIG DATA ERA

Stefan IOVAN, *Universitatea de Vest, Timişoara, ROMÂNIA*
Ramona MARGE, *Universitatea din Oradea, ROMÂNIA*

ABSTRACT: *This article is intended to be a very explicit guide to the Big Data approach. The material is organized in three parts: the importance of 'humanizing' Big Data, the strategic role of business analysts in extracting the context or 'storyline' they say, and a guide to innovative BI tools for extracting data from value and acquiring of competitive intelligence. Digital information is growing at a rapid pace and in such a large amount that the traditional storage systems used for storing and managing them are outdated. An analytical workflow is an important part of a company's intellectual property and increases its value as it is popularized, re-used and completed. Current technology offers progress in terms of the speed, scalability and financial savings that organizations need to work in a world where data is the basis for competitive advantage. Numerous organizations experiencing a dramatic increase in information volumes appeal to Apache Hadoop, an open-source data processing technology to address the need to store and manage a large amount of information.*

KEY WORDS: big data, data analysis, data warehouse, digital information, sensors, competitive intelligence.

1. INTRODUCTION

Most organizations are sitting on mountains of data that they could use in making decisions. Opportunities to respond to business questions are boundless when companies use their accumulated data in the Data Warehouse and, equally, in unstructured databases or unformatted text generated in Social Media. But the answers are hard to find.

1.1. What we understand from Big Data

Discussions that occur more and more frequently around Big Data often start from incorrect premises. Most conversations are conducted on the basis of Big Data processing platforms that focus on volume, variety, speed, leaving aside any reference to value.
In order to get value from Big Data it is necessary to add contextual information, which can only be achieved by placing analytical capabilities in the hands of those who need to extract value from them. In other words, Big Data needs '*humanization*'.
To humanize is to transform something inaccessible into something usable, making it difficult - easy, complex - simple, abstract - concrete. It means the process of putting information in context to *'tell the story'* about whom or what generates that information.
Big Data must therefore be extracted from the world of *'bits'* and *'bytes'* where only experts are available and made accessible, useful and capable of *'telling the story'*; Only in this way can they be converted into real insight for real business people.
Big Data has to be brought down to a plain state, where people who know the business can use them in making decisions, thus revealing their true value. Making Big Data depends on two critical elements:

- *Easy access to data:* Ability to access, integrate, Big Data analysis should be available to data and business analysts involved in strategic decisions of the organization.
- *Access to context-aware tools to help Big Data reveal reality*: Big Data can only provide a business-convertible image if it is accompanied by the complete context of all available data and if sophisticated analyzes can be performed without the need for advanced data or statistical knowledge skills.

Big Data's current workloads consume too much resources and people to generate value. With most existing BI tools, reports and analyzes are created based on previously structured data. These are typically internal data only, or without market insight, competitive intelligence, or location data. For this reason, they say only part of reality.

Current Data streams of Big Data have several parts. Data must be purchased from countless sources and cleaned. Then the data must be sorted and linked to allow interrogation.

Subsequently, file types that support unstructured formats must be stored in the system. Analysts and programmers should then work together in a statistical environment like R, SAS, or SPSS to query the data.

In the end, the data can be viewed in several formats - as a statistical report or sometimes as a 2D or 3D view. As a result, Big Data workflows - as they are today - look more like leaps than fluent streams.

The problem is that all this work is not driven by a business user working closely with data analysts. It is largely executed by a team of IT specialists behind the scenes, and any step in this process requires the involvement of someone else who usually has substantial backlogs with other projects.

Big Data workflows involve a number of changes or reprocessing, resulting in delays due to high staff demanding advanced analytical skills [1].

The person closest to the business user, that is, the business or data analyst, can not perform much of his own work, and so the time that elapses from the query to the insight is a lot of delays. In fact, it often happens that decisions are made on the basis of limited information, long before the answers come back from Big Data's workflow.

Thus, the rapid and independent access of analysts to data and their endowment with powerful analytical capabilities and tools is the essence of the humanization process of data [1, 5].

We will continue to discuss how the emancipation of existing analysts' role in companies by raising them to the ranks of artisans of data, able to extract relevant insight, empowers decision-makers to make quick and fully informed decisions.

1.2. Big Data in the sensor Era

The Big Data concept is seen by many as referring to structured and unstructured data coming from classical sources and new sources such as social media but all related to human activity. But the data generated by sensors (the so-called M2M phenomenon) has recently begun to generate the largest amount of information that needs analysis and interpretation [2].

It is already considered by analysts that these data provided by automatic sensors will generate the next wave of productivity gains and technological innovation. The statistics show, for example, that the Boeing engines that fit global airplanes have generated 10 TB of operational data and information for every half hour of operation.

A four-engine line engine can generate 640 TB of data across a single Atlantic crossing. If the number is multiplied by the more than 25,000 flights made daily, the volume of data thus obtained becomes dizzying.

Until now, almost all of this data was lost after the flights, but the situation is about to change. Why? Thanks to Big Data tools and technology, data can now be identified, stored, withdrawn and analyzed in a cost-effective and time-efficient manner.

And this opens up immense opportunities, for example in the area of preventive maintenance and aircraft failure - with direct effects on reducing flight delays and cancellations due to technical defects of the aircraft.

The example given above is not singular. Virtually the data provided by the sensors can now be effectively analyzed, in real time, in a wide range of areas of activity. For example, intelligent sensors used in electrical networks can generate large volumes of data that can be used to increase productivity. Thanks to such measuring instruments, electricity suppliers can read network parameters at different consumption points every quarter of a quarter instead of once a month.

This eliminates not only the need to send a person who reads the instruments on the

ground, but also allows differentiated electricity pricing according to peak hours.
Also differentiated charging can be used to flatten the consumer curve during load peaks, eliminating the need to create additional electrical capacities to cover these peaks and thereby generating substantial savings for energy suppliers in terms of capacity generation and maintenance costs of these capacities [2 - 3].
The above examples were not chosen by chance. Clearly, such examples can be found in many other areas. But the two have a high applicability in Romania, a country that has a national air operator and several power supply networks.
It remains to be seen whether Romania will become a country where simple economic considerations will also lead to a change of mentality in investments related to modern IT solutions for profit growth [4].

2. DATA ANALYSIS BECOME DATA ARTISTS

Today's top analysts are more artisans than reporters, as they add creativity and insight to strategic analyzes. Data craftsmen use their skills in what they do by infusing their knowledge deeply into the analytical process and associated applications.
They not only understand the problems and needs of those who run the organization, but also know where to find the right data to support strategic decisions [5]. However, data craftsmen are often not qualified to extract data for analysis or to strengthen company data structure. This is where some tools can help them perfect their art.
Making the Big Data make it accessible to business analysts, giving them skills that are usually available only to IT. This enables data to be rendered in the form of information, easily accessible and of high relevance.
It's about doing Big Data analysis effortlessly and in a natural way. Instead of strictly relying on specialized programming and statistical skills, data can be humanized by adding the right context and by providing simple and direct tools for building analytical applications [6].
Making Big Data is working directly with the data, so that they can tell the story. The full picture thus leads to insight of business. It also means a new opportunity for data analysts to improve their art and to extend their analytical capability independently. They become, consequently, artisans of data.

2.1. Emancipation of data craftsmen

In the past, data analysts needed advanced knowledge of statistics or business. At the moment, having access to analytical tools and contextual data that typically were only accessible to IT or IT experts, analysts can become true data craftsmen [5].
A craftsman is a person who takes raw materials (data in this case) and uses his skills, knowledge and vision to embed them into something of unique value.
Data craftsmen have a better understanding not only of the data they see, but also of the business and its problems. Because data craftsmen understand both data and business, by providing them with the right tools, they can really improve their results and apply them in a repetitive way to business issues [7].
The work of data craftsmen is the essence of Big Data humanization. Data craftsmen are the 'masters of humanization' when it comes to Big Data. Data craftsmen create data and analytical workflows that make data tell the story, regardless of whether they answer a specific question or a new application for business users at the end of the process.

2.2. Designing Principles for Big Data Humanization

Making Big Data involves designing some basic principles when it comes to creating solutions that deliver real insight:

- *Capturing and integrating data from any source***:** data collection systems, social media and 'sensitive' data are all to be taken into account, along with data from the Data Warehouse [8].
- *Finding patterns***:** Patterns hold the key to estimating future results. No perfect patterns in the smallest detail need to be

identified, but more general elements when it comes to unstructured data.

- *Communicating insight to decision points:* insight is more valuable when it is available to everyone. A store or warehouse manager is aware of the market because it has a tangent to it every day. Armed with powerful analytical tools and data that have been previously centralized, he can make informed and informed decisions.
- *Reusing the analytical IP:* A data craftsman can create a set of data to make known to a wide range of decision makers who can, in turn, chase, adapt, and build on its basis. Every time the story expands and also wins focus.

The purpose of Big Data humanization is to provide these capabilities to the analysts of the various departments of the business, enabling them to create reusable analytical workflows.

Such tools are made available to BI analysts, a desktop Web platform that offers the fastest and most comprehensive insight of the consumer, business and market to either large and medium sized companies, government agencies, or the academic environment [9].

3. TECHNOLOGIES FOR THE BIG DATA ERA

3.1. Elastic storage solution proposed by IBM

IBM (***I**nternational **B**usiness **M**achines*) has developed a portfolio of storage software products that offer improved efficiency by enabling organizations to access and process any type of data from any storage device anywhere in the world.

Named Elastic Storage, the technology delivers unprecedented performance and infinite scalability, being able to reduce storage costs by up to 90% by automatically moving data to the most efficient storage and performance storage device.

Designed in IBM Research Laboratories, state-of-the-art technologies enable businesses to exploit, not just manage, the explosive growth of data from a variety of sources: devices, sensors, business processes, and social networks. The new storage software is designed for data intensive applications that require rapid access to large volumes of information - from seismic data processing, risk management, financial analysis, and scientific research - to deliver the best response to critical situations.

The main elements of Elastic Storage have been used in the Jeopardy competition IBM Watson and two previous winners. In this confrontation, IBM Watson has accessed 200 million structured and unstructured data pages - including the entire Wikipedia encyclopedia. By using Elastic Storage capabilities, approximately five terabytes of information (or 200 million data pages) were loaded into the computer memory in just a few minutes [9].

The main reason why these capabilities have been integrated into the IBM Watson system is the scalability of the system, the architectural limits that are capable of crossing thousands of "yottabytes." A yottabyte is equivalent to one billion petabytes or a one-million-piece data center.

IBM Research has demonstrated that Elastic Storage can successfully scan 10 billion files per cluster in just 43 minutes - a demonstration of technology that translates into unparalleled performance for organizations that analyze large data volumes to extract prospects from business.

For the National Center for Atmospheric Research's Computational and Information Services Laboratory (CISL), rising data volumes are part of its DNA. The organization, which stores and manages more than 50 petabytes of information between Wyoming and Colorado centers, is based on Elastic Storage to provide researchers with fast access to large volumes of different data.

An important component of Elastic Storage is the ability to automatically and intelligently move data to the most strategic and economical storage system available.

In addition, the software has native encryption and secure deletion, which ensures that the data has been permanently removed - to comply with the HIPAA and Sarbanes-Oxley regulations.

With the support of the OpenStack cloud management software, Elastic Storage enables customers to store, manage and access data from private, public cloud and hybrid cloud environments in order to share them globally.

3.2. The potential of large volumes of data put on spot by HP

HP (**H**ewlett-**P**ackard) has completed its portfolio of information optimization solutions designed to help companies exploit the information explosion, including data on operations, applications, and equipment.

The volume, variety and speed of information are now an unprecedented burden for organizations. According to research, only 2% of the operational and IT managers say their organizations can deliver the right information at the right time to get the best business results.

These deficiencies are particularly noticeable in the context of an era where consumer perceptions change is done through Twitter, YouTube, the Internet, phone calls and emails, many of which take place outside of the organization. The level of perception can also be recorded in the form of pedestrian traffic detected by sensors installed in commercial spaces [9].

HP has invested in innovation to make the most complete portfolio of information optimization solutions with patents and technologies that can handle the problems organizations face in large data volumes. Current technology allows organizations to manage, understand and act on the full amount of information they hold.

This is possible with HP converged infrastructure solutions as well as with Autonomy and Vertica technology, both HP companies as well as HP data management services.

Apache Hadoop is the first complete tool for organizations in the industry that simplifies and accelerates implementation in parallel with Hadoop's extensive workload performance analysis and analysis. Through Vertica 6, the solution combines HP Converged Infrastructure, custom management and advanced integration to deliver bulk data processing and real-time analytical indices. Organizations have the opportunity to find the appropriate solutions to their own information optimization challenges with the help of new services.

With the introduction of Vertica 6, the latest version of the HP Vertica Analytics platform, companies now have the ability to connect, analyze, and manage any type of information from any location using any interface. Architecture provides a flexible workflow for large data volume analytics, including advanced integration or collection through Hadoop and Autonomy technologies, or for any structured, unstructured or semi-structured data source.

As part of the Vertica 6 release, the work framework extends to include native support for parallel execution of advanced analytic language R. Benefiting from improved support for Cloud and SaaS implementations as well as advanced functions for mixed workload environments, Vertica 6 offers the industries most rugged and comprehensive industry platform for high data volume analytics.

The combination of Autonomy IDOL, Vertica 6 and the HP Apache Hadoop App System allows clusters to access a unique platform for processing and understanding various massive data sets.

4. CONCLUSIONS

Current technologies allow organizations to reduce risks and speed up decision-making processes, facilitating a deep understanding of the challenges that large data volumes and available solutions generate.

Customers learn how to align corporate IT infrastructure policies and organization goals to identify both critical success factors and methods to develop their own IT infrastructure so they can manage a large amount of data.

Previous information management approaches, based on overarching architectural, infrastructure and analytical indexes, fail to discover the concepts and value found in the content of any form of information. They are also unable to scale efficiently and process real time information

oceans collected in unstructured, structured data volumes and into data machines [10].

Today, large volumes of data are great opportunities - and challenges - for organizations. Strong information optimization solutions deliver the technologies and expertise needed to help organizations succeed in this new era - covering any type, source, and data environment.

Whether it's deployment in the company's data center, in the cloud environment or in a hybrid environment, solutions enable organizations to transform large volumes of data into competitive and developmental opportunities, but also in opportunities.

5. REFERENCES

[1] Ştefan Iovan, Alina-Anabela Iovan, (2015): *The Role and Importance of Data Analyst in Using Large Volumes of Data*, Tîrgu Jiu: "Academica Brâncuşi" Publisher, *Annals of the "Constantin Brâncuşi" University, Engineering Series,* Issue **3**/2015, (***CONFERENG 2015***), pag. 160 - 165, Disponibil: http://www.utgjiu.ro/revista/ing/pdf/2015-3/28_Stefan%20IOVAN.pdf;

[2] Cristian Ivănuş, Ştefan Iovan, (2016): *Internet of Things and Business Process Management*, Tîrgu Jiu: "Academica Brâncuşi" Publisher, *Annals of the "Constantin Brâncuşi" University, Fiability & Durability Series,* Issue: Supplement **1**/2016, (***SYMECH 2016***), pag. 199 - 205;

[3] Ştefan Iovan**,** Gheorghe Dăian, (2012): *New Challenges: "Big Data" and "Consumer Intelligence" / Noi provocări: "Big Data" şi "Consumer Intelligence"*, Tîrgu-Jiu: "Academica Brâncuşi" Publisher, România, *Annals of the "Constantin Brâncuşi" University of Tîrgu Jiu, Engineering Series,* Issue **4**/2012, (***CONFERENG 2012***), pag. 318 - 329;

[4] Ştefan. Iovan, Alina-Anabela Iovan (2016): *Using Business Intelligence in all Activities*, Tîrgu Jiu: "Academica Brâncuşi" Publisher, *Annals of the "Constantin Brâncuşi" University, Engineering Series,* Issue **4**/2016, (***CONFERENG 2016***), pag. 13 – 19;

[5] Ştefan Iovan, Alina-Anabela Iovan, (2016): *Avantajul cunoaşterii şi abordarea proactivă*, Cluj-Napoca: Editura Eikon, România, **EDUCAŢIA DIN PERSPECTIVA VALORILOR**, (Coordonatori: Octavian Moşin, Ioan Scheau, Dorin Opriş), Tom IX: SUMMA THEOLOGIAE, pag. 197 – 202;

[6] Ştefan Iovan, Marcel Litră, (2013): *Decision Making in Big Data Era*, Proc. of 14th European Conference (***E_COMM_LINE 2013***), Bucureşti, România, ISBN-10: 973-1704-23-X;

[7] Ştefan Iovan, Cristian Ivănuş, (2014): *Business Intelligence and the Transition to Business Analytics*, Tîrgu Jiu: "Academica Brâncuşi" Publisher, *Annals of the "Constantin Brâncuşi" University, Engineering Series*, Issue **4**/2014, (**CONFERENG 2014**), pag. 150 - 156;

[8] Mihai Dinu, Ştefan Iovan, (2014), *Harnessing Big Data Volumes*, Proc. of 7th Symposium "*Durability and Reliability of Mechanical Systems*", (***SYMECH 2014***), Polovragi – Gorj, pag. 250 - 256;

[9] Ştefan Iovan (2015): *Big Data Security Problems*, Proc. of 16th European Conference (***E_COMM_LINE 2015***), Bucureşti, România, ISSN: 2392-7240;

[10] Ramona Marge, Ştefan Iovan, (2018), *The impact of economic company data theft*, Tîrgu Jiu: "Academica Brâncuşi" Publisher, *Annals of the "Constantin Brâncuşi" University, Fiability & Durability Series*, Issue: **1(21)**/2018, (***SYMECH 2018***), pag. 275 - 280;