Boarding to Big data

Oana Claudia BRATOSIN
University of Economic Studies, Bucharest, Romania
oc.bratosin@gmail.com

Today Big data is an emerging topic, as the quantity of the information grows exponentially, laying the foundation for its main challenge, the value of the information. The information value is not only defined by the value extraction from huge data sets, as fast and optimal as possible, but also by the value extraction from uncertain and inaccurate data, in an innovative manner using Big data analytics. At this point, the main challenge of the businesses that use Big data tools is to clearly define the scope and the necessary output of the business so that the real value can be gained.
This article aims to explain the Big data concept, its various classifications criteria, architecture, as well as the impact in the world wide processes.
Keywords: *Big data, Predictive Analytics, Data mining, Internet of Things*

1 Introduction

Doug Laney introduced the big data concept by using the 3V model, Volume, Velocity and Variety, concept extended a few years later to 7V.
Data volume is getting more and more wider, generating this way an inverse connection with the data value, which decreases as much as data volume increases.
The velocity that that the data is extracted and transformed to information turns out to be one of the most competitive attribute that makes the difference between the Big data tools.
The last big attribute, variety, stands on the multiple and various data sources, with incompatible data format and unsubstantial data format. [1]
Next to the three previously described attributes for Big data, it follows the value and veracity of data. Now the model is extended to 5V, but it lays on some elements described on the first attributes.
The veracity is defined by three main characteristics, the first one is represented by quality, consistency and data accuracy. Next come the source and data origins and the last one is represented by the purpose of using data. [2]
The last extension of the model comes from the variability of the data sense and from the visualization, as the information derived from data must be understood properly and as fast as possible. This attributes generate a lot of implications of Big data over all the companies as they define the last level of data accuracy. [3]

2 Big data classification

There are different criteria to classify Big data, but the most meaningful ones are the analysis type, processing methodology, content format and data sources. Besides these, Big data can be classified also by data frequency, data type, data consumers or even by the hardware. [4]
The analysis type defines the velocity attribute of Big data and it stands on the data processing system, divided in three categories grounding on time criteria.
The first one, real time processing, is defined as a synchronous process where an input request is sent, processed and an output response is received immediately after the processing is completely finished. Significant examples about this processing type are the bank ATMs, radar systems, fraud detection or customer care services, as here an instant processing is completely needful.
Near real-time processing stands right next to real time processing, with the difference

of the time unit. Here, the processing time is managed in minutes, while in real time it leads to seconds. An example of near real-time processing is the operational intelligence to run query analysis on real data.

The batch processing is managed by batch jobs that can run up to several hours, even days. The process starts first with data retrieval, then the data is processed and finally it is sent to output. Most of the payroll and billing operations are managed at the end of the month or year through batch processing.

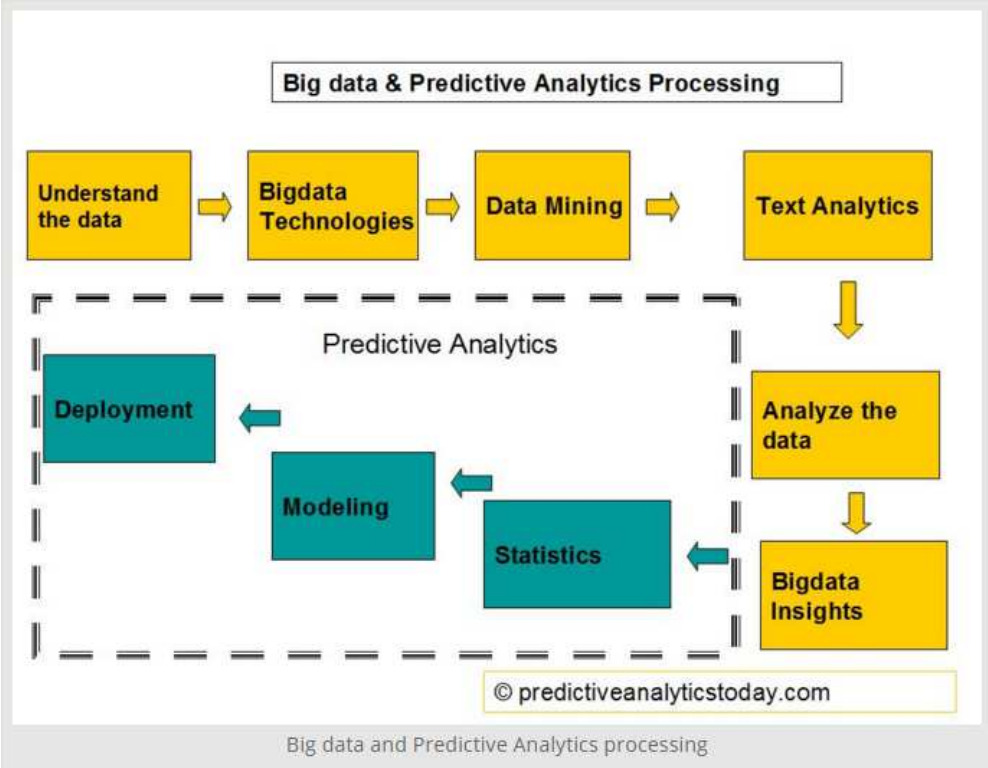

Fig. 1 Big data Predicative Analytics processing [5]

Big data Processing methodology is classified in three categories: Predictive Analytics, Analytical, Query and Reporting.

Predictive Analytics is used to forecast the future processes, so that they can be managed in an optimized manner. It can use historical, current or future data and apply predictive models over them as data mining, statistical modelling or machine learning. The entire process is depicted in Fig. 1.

The predictive analytics can be used in all the industries, starting with utilities industry, to forecast the necessary quantity of energy necessary for a month, to the retail industry, to forecast the consumer's behavior. [5]

The analytical technique is used to manage the business overview using social network analysis, statistical techniques, speech analytics or face recognition.

Besides query and reporting, techniques used mostly in Business Intelligence, there are also the 3D reconstruction and the translation techniques.

The content format refers to the input data format and it is the starting point for how it should be processed and what technology should be applied on it.

Thereby we can have structured data as data that can be stored in a relational database with a relational key, **semi structured** data for data that it is not stored in relational database, but does have some rules that allows it to be quickly analyzed, and unstructured data as data that cannot be systemized in a database. The structured data represents about 5 to 10% from the data, as much as the semi

structured data. Examples of semi-structured data are the CSV files, XML files, NoSQL databases. The rest of the data, about 80% is own by the unstructured data, for example satellite images, photographs, social media data or web site content.

Various data sources for Big data define the scope from the business perspective: web and social media, machine generated, human generated, internal data sources, transaction data, biometric data, via data providers and via data originator.

Each business can definitely have more than one data sources, for example a simple healthcare business can have machine generated data, transactional data, web data and human data. [4]

3 Big data architecture

The main component of the Big data architecture is the data pipelines, as they pick the raw data and transform it to value. Over this process a very important role has the Big data engineer, as he is the one that takes the most important decisions regarding the methods to process the data and how to build it in a proper form so that it can be easily understood and used by the final end user.

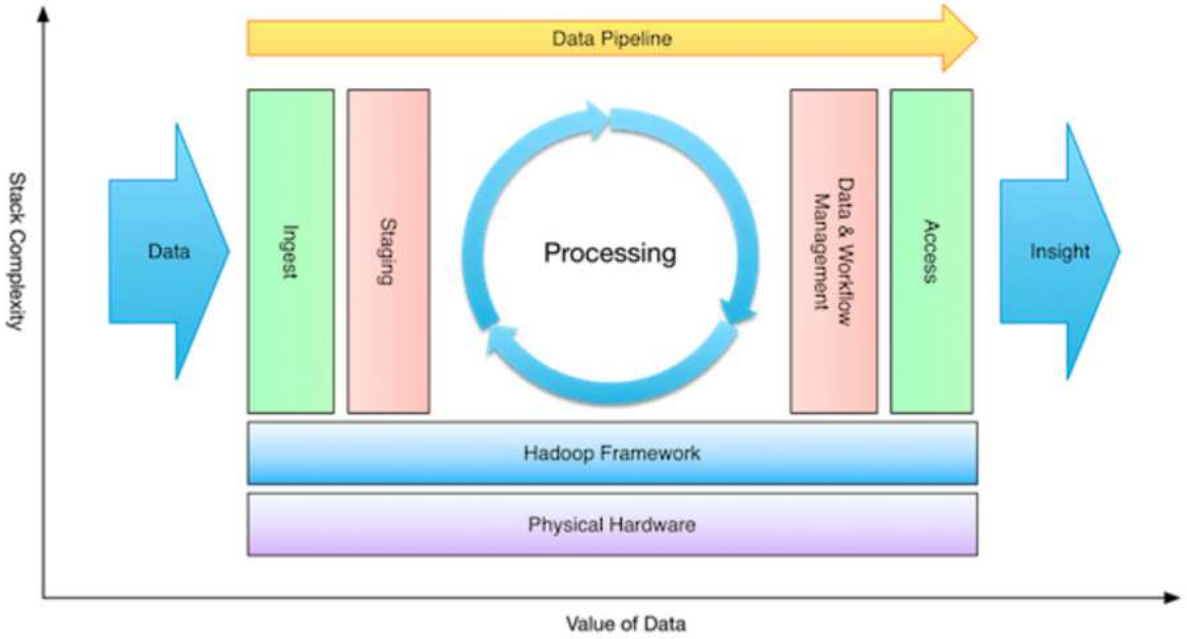

Fig. 2 Big data architecture [6]

In **Fig. 2** it is pictured the Big data architecture over the Hadoop framework; it is the most representative architecture model for Big data.

The first architecture's component is the ingest, so how the data will be loaded. There are two options to load the data, batch load and event driven load.

The batch ingest can be applied for structured data sources as for data from a relational database using Apache Sqoop, but the most complex case is for the file loading due to their location. It is recommended to use the event ingest for file loading whenever it is possible, so that it can skipped the bulk load of files.

The event ingest relies on agents configured to buffer data and to exchange data between them. One main aspect here is the tuning of the events that leap from one agent to another, as it will ensure smaller latencies and faster responses.

The staging follows data loading. At this point, the data is standardized to the right format, encoded and compressed.

Next to the staging comes the data processing step. First data is transformed and then analytic models are applied over it. Once the results are gathered, they must be represented in a proper form to the user.

This means to provide them proper access, for each user category. [6]

4 Big data market forecasts

Big data development process continues to increase over the time, leading to the market growth from the vendor revenue by selling Big data products and large business that are willing to use the Big data tools.

The Big data market is measured by the revenue obtained from the selling of Big data related hardware, software and services.

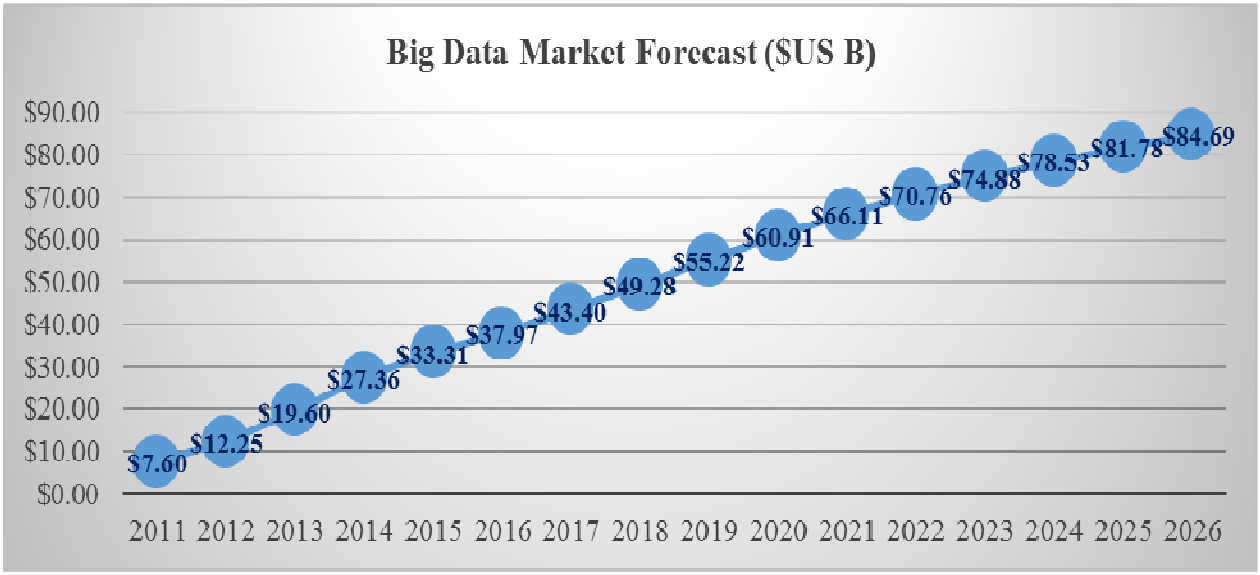

Fig. 3 Big data market forecast in US billion $ [7]

In **Fig. 3** it is depicted the growth evolution for Big data marked for the time span starting with 2011 to 2026.

From 2011 to 2014 the annual growth rate is very high as a consequence of the following factors:

- The data warehouse has been optimized
- Big data products have been expanded over business segments.
- The decision making process of Big data has been maturated, stabilized and maximized over the Big data vertical market

In 2014 the Big data marked got to $27.36 billion, with an increase of 40% of the growth rate towards 2013. Compared to 2013, the 2014 market growth rate slowed with 20%, and the forecast until 2026 is that the annual growth rate will slow down, around 17% for the depicted 15 years.

The sub-segments of Big data market consist by:

- Hardware
- Data management software
- Hadoop software
- SQL software
- NoSQL software
- Applications and analytics tooling
- Professional services
- Cloud platforms [7]

Regarding the Big data market share in 2015, the company that owns the biggest part is IBM with 9.3% from the total share. SAP follows IBM on the second place with 3.9% and Oracle is on the third place with 3.3% from Big data market share. [8]

5 Big data security and privacy challenges

In order to process huge amounts of data in a timely manner, the parallel processing can be applied. First, the input data is divided in chunks, and then the chunks are processed in parallel. For example the MapReduce framework divides an input file into several chunks. The process starts with the Mappers, one for each chunk, that read the data, manipulate it and outputs a list of pair keys. Next, a Reducer makes the associations between the outputs from each mapper and display the final, aggregated result.

At this point, two main issues are raised, one to secure the mappers and another one to secure the data for untrusted mappers. These issues can lead to inconsistent results and information leaks.

NoSQL databases are not yet matured enough from the security level point of view, as the security was not part of the NoSQL database design model. To ensure the security for this kind of database, the developers often include the security in the middleware.

Data is usually automatically stored in multi-tier storage media. The security levels are directly related to the tires, so that the lower tiers have a lower level of security, comparing to the higher tiers where the security level is increased. In this case an attack can appear over the wrongly tiered data, for example bank account information or money transfers that are wrongly tiered on a lower level, therefore a lower level of security is applied on very sensitive data.

Another security challenge lays on trusting the input data. All the big businesses collect data from many sources represented by hardware devices and software solutions. This data needs to be analyzed to define if it is trustful and filtered by malicious intruders.

Real time security requires real-time anomaly detection to identify properly the false positives, so that the correct behavior can be further applied.

Data anonymity is again one main attribute for the Big data security. Untrusted analysts or partners can reveal anonymous information and it is very important to set rules and recommendations to put a stop for this kind of attacks.

Sensitive data should be managed with a cryptographically secure communication framework. It manages access control policies and data encryption.

The granular access provided for data offers a high level of security over it. If the granularity level is properly managed, it will ensure the visibility of data only for who it should be visible and the chance that sensitive data arrive in the wrong hands will decrease substantially.

Security audits held data with respect to the real time attacks. This kind of information is necessary to understand what went wrong, but as well to be compliant with the legal requirements for each geographical region at a time. [9]

6 Big data and Internet of Things (IoT)

IoT is a five stages process, pictured in the **Fig. 4.**

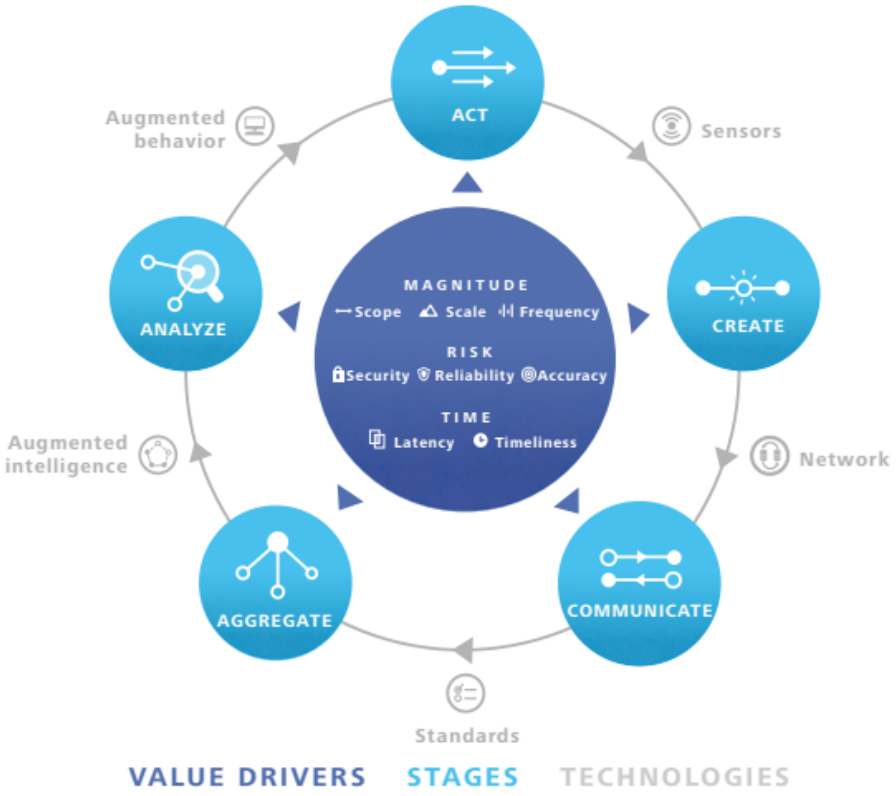

Fig. 4 The Information Value loop [10]

The first stage is the creation one, where sensors are used to generate data about indicator or physical events. Next, in the communication stage, previously generated data is passed between several sources.

Following comes the aggregation stage, where data coming from various sources and different time units is grouped together for the analysis stage. Here, data is transformed to information that it is used in the last stage, acting, where a behavior is applied accordingly.

Each stage from the information value loop is correlated to its preceding stage, using one of the following technologies:

- Sensors – devices that produce an electronic signal as a result of an event or physical condition
- Networks – transmission environment for the electronic signals
- Standards – instructions for actions
- Augmented intelligence – related to the analytical tools to improve the capability to predict, describe and manage relationships between the events
- Augmented behavior – improve the agreements with the corresponding actions. [10]

All this interconnected devices will generate a lot of data and here the Big data makes its entrance. Combining Big data with IoT will open new doors for the top companies to evolve and innovate the business processes, to perform faster and to met the clients values.

The connection of Big data with IoT starts with the Big data capability to organize and analyze different types of data, structured, semi structured or unstructured, even if it comes from various data sources in a various format. Once data collected, Big data analytics manages huge volume of data and extract the value of information from it. Here comes again the Big data's challenge to ensure data privacy and security over the entire process. [11]

7 Conclusions

Big data is an emerging matter, laying on

the quantity of information that is growing faster and faster over the time. It is very important to understand the Big data attributes, so that the proper tools and consulting services can be applied for each business, individually.
Nowadays it is very important to ensure the security of data, custom for each model and enterprise.

References

[1] S. M. Y. Z. V. C. L. Min Chen, Big Data: Related Technologies, Challenges and Future Prospects, Springer, 2014.

[2] C. Ballard, Information Governance Principles and Practices for a Big Data Landscape, 2014: Information Governance Principles and Practices for a Big Data Landscape.

[3] M. V. Rijmenam, Think Bigger: Developing a Successful Big Data Strategy for Your Business, AMACOM, 2014.

[4] S. K. S. J. Divakar Mysore, "How to classify big data into categories," IBM, September 2013. [Online]. Available: http://www.ibm.com/developerworks/library/bd-archpatterns1/.

[5] N. Merolla, "Big Data Analytics and Predicitve Analytics," Predicitive analytics today, December 2013. [Online]. Available: http://www.predictiveanalyticstoday.com/big-data-analytics-and-predictive-analytics/#content-anchor.

[6] L. George, "Getting Started with Big Data Architecture," Cloudera, September 2014. [Online]. Available: http://blog.cloudera.com/blog/2014/09/getting-started-with-big-data-architecture/.

[7] J. Kelly, "Big Data Vendor Revenue and Market Forecast, 2016-2026," Wikibon, 2015.

[8] R. Finos, "2015 Big Data Market Shares," Wikibon, 2016.

[9] C. s. alliance, "Top Ten Big Data Security and Privacy Challenges," 2012.

[10] M. M. M. E. R. M. C. Jonathan Holdowsky, "Inside the Internet of Things (IoT)," Deloitte university Press, 2015.

[11] Datameer, "Big Data Analytics and the internet of Things," 2015. [Online]. Available: http://www.datameer.com/pdf/eBook-Internet-of-Things.pdf.

Oana Claudia BRATOSIN graduated the Faculty of Cybernetics, Statistics and Economic Informatics of the University of Economic Studies in 2011 and gained the Master title in Cybernetics and Quantitative Economy at the same university in 2013.
Currently she is studying to earn her PhD Diploma in Economic Informatics.