Version 1.0 - April 2016

White Paper
BIG DATA

ILNAS
Institut luxembourgeois de la normalisation, de l'accréditation, de la sécurité et qualité des produits et services

ANEC
AGENCE POUR LA NORMALISATION ET L'ÉCONOMIE DE LA CONNAISSANCE

Acknowledgments

The *Institut Luxembourgeois de la Normalisation, de l'Accréditation, de la Sécurité et qualité des produits et services* (ILNAS) would like to acknowledge Mr. Robert Van Wessel (from ApexIS) for his huge contribution provided to this document, in the frame of the dedicated collaboration between ILNAS and ApexIS.

ILNAS would also like to thank the economic interest grouping *Agence pour la Normalisation et l'Économie de la Connaissance* (ANEC G.I.E.) - *Département Normalisation*: Ms. Anna Pochylska, Mr. Nicolas Domenjoud, Mr. Juan Luis Jiménez Laredo, Mr. Joseph Emeras and Mr. Johnatan Pecero Sánchez for their involvement in this document.

Executive Summary:

This document aims at surveying current advances in Big Data and Big Data Analytics from two complementary points of view: a technical analysis perspective and a business and economic prospective analysis. Therefore, the document is intended for those professionals seeking guidance in one or both domains and can be used in its whole as a compendium where technical and IT governance aspects of Big Data are equally treated. Standards and technical standardization are also presented as an essential tool to improve the interoperability between various applications and prevent vendor lock in. They also provide interfaces between relational and non-relational data stores and support the large diversity of current data types and structures. Finally, some conclusions on Big Data are presented with an outlook on how to integrate them in the business environment to create value.

Foreword

The "*Institut Luxembourgeois de la Normalisation, de l'Accréditation, de la Sécurité et qualité des produits et services*" (ILNAS) is an administration, under the supervision of the Minister of the Economy in Luxembourg. ILNAS is the national standards body, and, in this frame, has developed, in partnership with the University of Luxembourg, a certificate on "Smart ICT for Business Innovation" (lifelong learning framework) at the end of 2014. The current White Paper has been carried out in the context of this university certificate. It aims to support the development and to strengthen the standardization culture at the national level, specifically in an economically meaningful field like Smart Information and Communication Technology (Smart ICT). This initiative is *de facto* in line with the Luxembourg's Policy on ICT Technical Standardization 2015-2020 in the context of education about standardization.
The university certificate offers a broad view of cutting-edge Smart ICT concepts and provides various tools to the students in order to develop their sense of innovation. ILNAS commissioned ANEC GIE standardization department to implement yearly the university certificate, and to carry out its development. In this framework, ANEC GIE is actively contributing to the creation of pedagogical materials related to the Smart ICT topics addressed and covered by the university certificate, mainly from the standardization point of view.

Overall, the pedagogical program has been developed based on a common-thread that describes the role of technical standardization as one enabler of innovation. In this context, ICT is considered as a dynamic horizontal sector that supports the development of other economic sectors (vertical convergence). In the intersection between the horizontal and the vertical sectors, technical standardization can be considered as an enabler to allow the interoperability between them. All in all, technical standardization is not only a specific module in the academic program, but it is present in each module as a reference and as a key factor to trigger innovation. In other words, technical standardization represents the general keystone of the university certificate.

With the aim of providing the students with a reliable source of information and of recent breakthroughs, the different standardization committees involved in Smart ICT developments are considered like a basis for the certificate. They represent the unique ecosystem gathering both the public (Ministries, administrations, etc.) and private sectors (manufacturers, researchers, business innovators, and other stakeholders…), making them the beating heart of the ICT progress, and thus creating a conducive common technical "platform" for the students of the certificate. More in detail, the focus of the certificate relies on important aspects of Smart ICT and their applications, including development of Smart Cities, Smart Grid, Big Data and Analytics, Cloud Computing, Internet of Things and Digital Trust. Moreover, ICT Governance and environmental issues related to ICT are likewise addressed.

This document, which is used as a basis for the development of the Big Data lecture in the context of the university certificate, surveys current advances in Big Data and Big Data Analytics from two complementary points of view: a technical analysis perspective and a business and economic prospective analysis. From the technical analysis perspective, the document surveys technologies supporting the development of Big Data and used to analyze Big Data and acquire information from it. It presents some technological challenges related to Big Data. The document also surveys considerations when implementing Big Data to reap benefits from the business and economic prospective analysis. Standards and technical standardization are presented as an important tool to improve the interoperability between various applications and to share good practices. In this context, the document presents major efforts related to the development of Big Data standards.

Jean-Marie REIFF, Director
Jean-Philippe HUMBERT, Deputy Director
ILNAS

Table of Contents

FOREWORD 6

ABBREVIATIONS 3

1. BIG DATA – AN OVERVIEW 5

1.1 Characteristics and Business Relevance 5

1.1.1 The Vs of Big Data 5

1.1.2 The Business Importance of Big Data 7

1.2 How Big is Big Data? 8

1.3 Big Data Types 9

1.4 Operational and Analytical Big Data 12

1.5 No Single Definition for Big Data 14

2. PROCESSING BIG DATA: A TECHNICAL ANALYSIS 17

2.1 Introduction 17

2.2 Analytical Big Data: the MapReduce Paradigm 19

2.2.1 Google MapReduce 20

2.2.2 Apache Hadoop 21

I. Hadoop MapReduce 21

II. Hadoop Distributed File System (HDFS) 22

III. Hadoop Ecosystem 22

2.3 Operational Big Data: Non-Relational Databases 24

2.4 Other Big Data Tools 25

2.4.1 Batch processing 25

2.4.2 Stream processing 26

2.4.3 Interactive analysis 27

2.5 Challenges 28

3. BUSINESS AND ECONOMIC PROSPECTIVE ANALYSIS 30

3.1 Introduction 30

3.2 Implementation 31

3.3 Other Considerations 34

3.3.1 Data quality 34

3.3.2 Cloud Computing 34

3.3.3 Data security 35

4. BIG DATA STANDARDIZATION 36

4.1 Big Data Standardization from Formal Standards Body 36

4.2 Big Data Standardization from Fora and Consortia 44

5. CONCLUSION AND OUTLOOK 47

LITERATURE 50

Abbreviations

ACRONYM	DESCRIPTION
ACID	Atomicity, Consistency, Isolation, Durability
BLOBs	Binary large objects
CAPEX	Capital Expenditure
CCTV	Closed-circuit television
CEN	European Committee for Standardization
CENELEC	European Committee for Electrotechnical Standardization
CEO	Chief Executive Officer
CEP	Complex Event Processing
CERN	European Organization for Nuclear Research
CFO	Chief financial officer
CIO	Chief information officer
CRM	Customer relationship management
DBMS	Database management system
DVD	Digital versatile disc
EDI	Electronic Data Interchange
ETL	Extract, Transform and Load
ETSI	European Telecommunications Standards Institute
GAFA	Google, Apple, Facebook, Amazon
GPS	Global Positioning System
HDD	Hard disk drive
HDFS	Hadoop Distributed File System
ICT	Information and Communication Technology
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IF	Input Format
ISO	International Organization for Standardization
ITU	International Telecommunication Union
JSON	JavaScript Object Notation
JTC	Joint Technical Committee
LISP	List Processing (programming language)
MPEG	Moving Picture Experts Group
NBD-PWG	NIST Big Data Public Working Group
NIST	National Institute of Standards and Technology
NoSQL	Not only SQL
OASIS	Organization for the Advancement of Structured Information Standards
ODBMS	Object-oriented database management system
OGC	Open Geospatial Consortium
OLAP	Online Analytical Processing

OPEX	Operating Expense
PDF	Portable Document Format
RDF	Resource Description Framework
RFID	Radio-frequency identification
ROI	Return On Investment
RPC	Remote Procedure Call
RR	Record Reader
SC	Subcommittee
SOA	Service Oriented Architecture
SQL	Structured Query Language
SWIFT	Society for Worldwide Interbank Financial Telecommunication
UI	User Interface
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
XML	Extensible Markup Language
YARN	Yet Another Resource Negotiator

1. Big Data – an Overview

Big Data is a topic that has gained enormous attention in the last couple of years by industry, governments and academia. Although the term Big Data was coined in 1997 to refer to large volumes of scientific data for visualization (Cox and Ellsworth, 1997), the question raises: why all this sudden and recent interest?

Today, more data are generated in 10 minutes than all of humanity has ever created through to the year 2003, says Dave Turek from IBM. This stream of data is gigantic, fast-paced and very diverse. Our lives are becoming at this point soaked with data from connected objects, social networks, online purchase and behavioral activities that we have reached a point where it all can be correlated. From this correlation of data, we can discover, understand, learn and improve much valuable information. This is Big Data.

90% of data in the world has been generated in the last 2 years.

1.1 Characteristics and Business Relevance

The term Big Data is used in a variety of contexts, to better define the technology and its concepts, it is necessary to focus first on its key characteristics.

1.1.1 The Vs of Big Data

Big Data is characterized by a collection of huge data sets (**Volume**), generated very rapidly (**Velocity**) and with a great diversity of data types (**Variety**). The original **three Vs** (Volume, Velocity and Variety) were introduced in 2001 by Doug Laney from Metagroup. In those days, Laney did not use the term 'Big Data', but he envisioned that e-commerce accelerated data generation with incompatible formats and structures were pushing traditional data management principles to their limits (Laney, 2001). Because of the massive amount of data and the variety of its sources, another characteristic of Big Data is the inherent error, noise and induced bias of erratic data (**Veracity**) (Schroeck et al., 2012; Zikopoulos et al., 2013).

Such data are difficult to process by traditional data processing platforms, such as relational databases, and impossible to analyze with traditional techniques, such as data mining[1].

Big Data refers to technologies that involve data that are too massive, diverse and fast-changing to be processed efficiently with conventional techniques.

The combination of the original three Vs and this fourth characteristic of Big Data are generally adopted as the Big Data **four Vs**, depicted in Figure 1 and presented in Table 1.

[1] Data mining is the analysis of data for relationships that aims to discover previously unknown interrelations among apparently unrelated attributes of data sets by applying methods from several areas including database systems, statistics and artificial intelligence (such as machine learning).

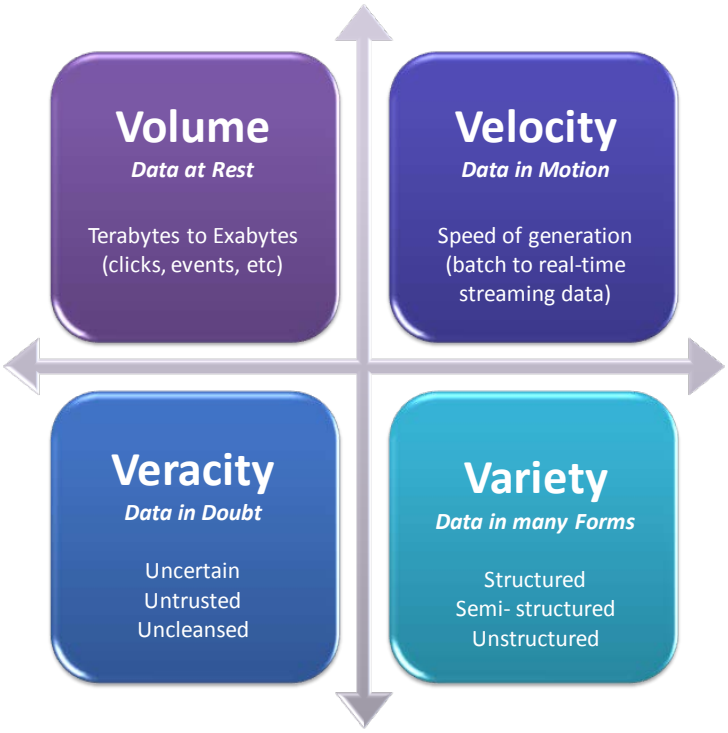

Figure 1 The four Vs of Big Data

Characteristic	Description
Volume	How much data: the amount of data that organizations try to harness to improve decision-making across the enterprise
Velocity	How fast data are created: the speed of incoming data and how quickly they can be made available for analysis (e.g. payment data from credit cards and location data from mobile phones)
Variety	The various types of data: the different types of structured and unstructured data that an organization can collect, such as transaction-level data, text and log files and audio or video
Veracity	How accurate are data: the trust into data might be impaired by the data being uncertain, imprecise or inherently unpredictable (e.g. trustworthiness, origin and reputation of the data source).

Table 1 The four characteristics of Big Data. Also referred as the four Vs

More recently, other Vs have been proposed such as **Variability**, which refers to data whose meaning is constantly changing (Hilbert, 2015). **Variability** gives the temporal dimension to the aforementioned four Vs and thus could also be viewed as a property of data itself.

One important aspect of Big Data is also its **Value** (Chen, 2014). The value of Big Data is multiple. It is first the inherent value of data itself in terms of information contained and its worth in money (e.g. for reselling purposes). It is also the value in terms of analytical possibilities offered by Big Data. Processing Big Data and looking for correlations, predicting consumer behaviors, financial risk, sentiment analysis, leads to huge potentials in terms of e.g. dollar savings, market penetration analysis or user satisfaction improvement.

Along time, other Vs have been suggested such as **Visualization, Volatility, etc.** and it is expected that new propositions will arise as the technology continues to mature.

1.1.2 The Business Importance of Big Data

The volume of data generated, stored, and mined for insights has now become economically relevant to businesses, government, and consumers.

The use of Big Data is now becoming a crucial way for leading companies to outperform their peers.

According to Wikibon Executive Summary 2015 (Wikibon, 2015), the growth of Big Data market started to slow down from 2014 (see Figure 2), although it still remains significantly faster than other IT markets. This means that this disruptive technology has already started maturing and has now fully reached the market. Still according to Wikibon, on the period 2011-2026, the Big Data market growth will reach up to 17%.

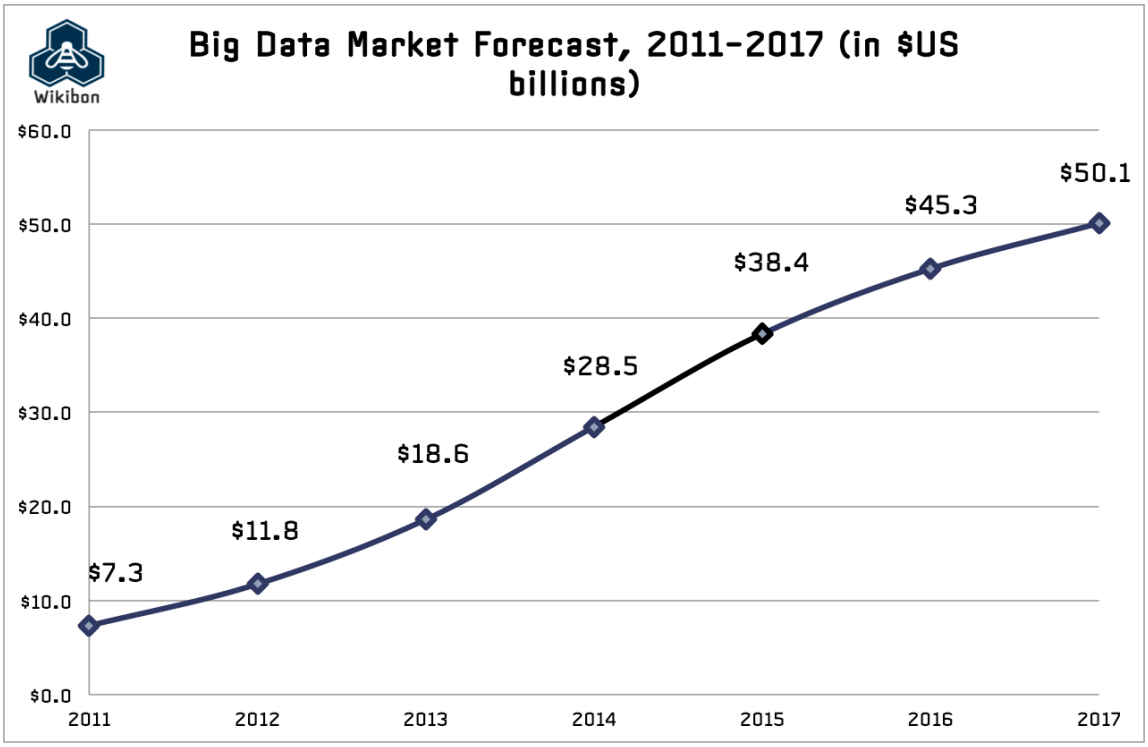

Figure 2 Big Data Market Growth Expectations (Wikibon, 2015)

In Gartner's Hype Cycle for Emerging Technologies[2], Big Data was even not anymore considered as an emerging technology because it *¨has become prevalent in our lives¨* says Betsy Burton, vice president and distinguished analyst at Gartner. On the contrary, Big Data is now mature enough to generate emerging trends and technologies. With the advent of Big Data, data are now becoming a valuable asset to the company seeking to exploit it.

[2] http://www.gartner.com/newsroom/id/3114217

The real-time and sheer volume characteristics of Big Data provides the ability to estimate, model, simulate business metrics such as user satisfaction, product and service interest, or to drive risk analysis and decision making immediately whereas it could only be done retrospectively before.

Big Data can help business and industry in building new applications that were not possible before, develop competitiveness and improve customer satisfaction by reducing costs, better segmenting customers to precisely tailor products and services, designing next generation products along with reducing their Time to Market and much more.

Many successful stories of using Big Data exist such as the city of Chicago using Big Data to cut crime and improve municipal services. By collecting, aggregating and analyzing geospatial data in real-time from over 30 departments, and marrying structured and unstructured data, the city of Chicago drives online analysis to determine if an uptick in crime is more likely than usual. Other examples such as a photo-sharing website company that reduced their cumulated CAPEX and OPEX by 80% using an operational Big Data storage system or a leading newspaper analyzing their user behavior to increase the customer engagement and thus their global revenue (MongoDB White Paper, 2015).

1.2 How Big is Big Data?

The massive increase of data that are being generated is caused by a number of reasons (Fernández et al, 2014; Assunção et al., 2014):

1. billions of devices such as sensors of mobile devices, security cameras on highways, and GPS tracking systems are transmitting and collecting data continuously;
2. the widespread diffusion and adoption of social network websites including YouTube, Facebook and Twitter where users create records of their daily activities, events they attend, things they eat and drink, places they visit, pictures they take, etc.;
3. organizations are increasingly generating large volumes of data as a result of monitoring of user activity / web site tracking as part of their business processes;
4. applications in science experiments result in an increase of data sets at an exponential rate;
5. storage capacity become so cheap that it is often easier and less costly to buy more storage space rather than deciding what to delete;
6. machine learning[3] and information retrieval techniques have significantly improved in the last couple of years, thus enabling the acquisition of a higher degree of knowledge from data.

Examples show the massiveness of the amount of data generated every day in business:

- In 1 second: more than 2,100 Skype calls; 700 Instagram photos uploaded; 34,500 GB of Internet traffic; 53,900 Google searches; 121,400 YouTube videos viewed[4].
- 300 hours of video were uploaded to YouTube every minute (March 2015)[5].
- Twitter serves around 650 million active users, who produce 9100 tweets every second[6].
- Facebook creates 10 terabytes ($10x10^{12}$ bytes) data every day, and Google produces 24 terabytes of data every day just from its search operations (Chang et al., 2014).

[3] Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data. However, instead of extracting data for human comprehension - as is the case in data mining applications - machine learning uses that data to improve the program's own understanding.
(source: whatis.techtarget.com/definition/machine-learning)

[4] http://www.internetlivestats.com/one-second/ (Last accessed April 2016)

[5] http://www.youtube.com/yt/press/statistics.html

[6] http://www.statisticbrain.com/twitter-statistics/

- Each day 2.5 exabytes ($2.5x10^{18}$ bytes) is created, so that 90% of the data in the world today has been created in the last two years alone[7].

and in scientific research:

- CERN's Data Centre processes about one petabyte (10^{15} bytes) of data every day - the equivalent of around 210,000 DVDs. Its Large Hadron Collider, the largest particle accelerator, generates 40 terabytes per second[8].
- 32 petabytes of climate observations and simulations are conserved on the discovery supercomputing cluster in the NASA Center for Climate Simulation (NCCS) (Chen and Zhang, 2014).
- The Large Synoptic Survey Telescope (LSST) will record 30 exabytes ($30x10^{18}$ bytes) of image data in a single day (Chen and Zhang, 2014).

The possible applications of Big Data and its Analytics include aircrafts, audio, automobiles, call detail records, click streams, financial services, government, healthcare, military, mobile phones, resource management, photography, private sector, retail, RFID, search indexing, social networks, text, video, web logs and various kinds of sciences (e.g. astronomy, biology, chemistry, genomics, geology, medicine). Business consultancy firms claim that **retailers can achieve up to 15–20% increase in ROI** by putting Big Data into Analytics (Perrey et al., 2013), however, few detailed empirical studies have been conducted to assess the real potential of Big Data (Wamba, 2015).

1.3 Big Data Types

Big Data incorporates all kinds of data and from a content perspective one can make the distinction between structured data, semi-structured data and unstructured data (Kambatla et al, 2014). In practice mixed combinations of these three Big Data types occur which is referred to as Poly-structured data.

- *Structured data* are data that are part of a formal structure of data models associated with relational databases or any other form of data tables. They can be generated both by computer software or humans.
- *Semi-structured* data are data that are not part of a formal structure of data models. However, they contain markers (e.g. tags) to separate semantic elements and enforce hierarchies of records and fields within the data. Therefore, this term is also known as self-describing structure. Examples are EDI, SWIFT, and XML and JSON data.
- *Unstructured* data[9] are data that do not belong to a pre-defined data model and include data from e-mails, video, social media websites and text streams. They account for more than 80% of all data in organizations (Holzinger et al., 2013). Until recently, software technology did not effectively support doing much with them except storing or analyzing manually. Just as with structured data, unstructured data are either machine generated (by computer or software) or human generated. Machine generated unstructured data include radar or sonar data, satellite images and, security, surveillance, and traffic videos but also Scientific related data such as seismic imagery, atmospheric data, and high-energy physics. Human generated includes text messages, e-mails and social media data.

[7] http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

[8] http://public.web.cern.ch/public/en/LHC/Computing-en.html

[9] The term unstructured is misleading as such data are not really unstructured because each file contains its own specific formatting, needed for its specific use by humans or software. In fact, it is the content of the document that is unstructured. Basically, only white noise is really unstructured (a random signal with a constant power spectral density).

Often, data are generated by a combination of these three groups:

- Machine generated data from computers or devices, such as sensor or log data, audio, video and click statistics.
- Machine generated data with a specific Business propose, such as CRM data, master data and transaction data.
- Human generated data, such as social networking, audio, video, images, free text, forms, logs, and web content.

Big Data technologies assist companies to make sense of all these different types of data.

Hashem et al. (2015) also make the distinction between structured data, semi-structured data and unstructured data and puts them into a broader classification scheme (see Figure 3), that aside content format also contains four other aspects: data source, data store, data stage, and data processing. Each of these categories has its own characteristics and complexities as described in Table 2. For a description of some specific examples in this table, such as MongoDB and S4, please refer to Sections 2.3 and 2.4.

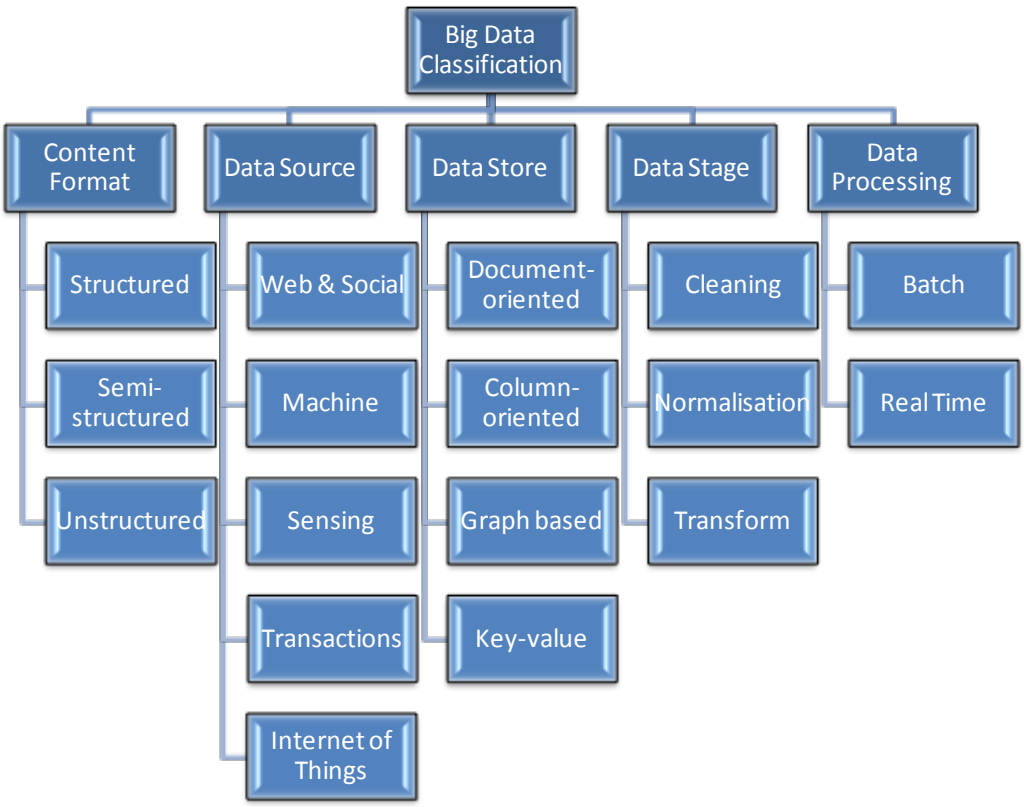

Figure 3 Big Data classification (Hashem et al., 2015)

Classification	Description
CONTENT FORMAT	
Structured	Structured data are often managed with SQL. Structured data are easy to input, query, store, and analyze. Examples of structured data include numbers, words, and dates.
Semi-structured	Semi-structured data are data that do not follow a conventional database system. Semi-structured data may be in the form of structured data that are not organized in relational database models, such as tables. Capturing semi-structured data for analysis is different from capturing a fixed file format. Therefore, capturing semi-structured data requires the use of complex rules that dynamically decide the next process after capturing the data.
Unstructured	Unstructured data, such as text messages, location information, videos, and social media data, are data that follow their own format. Considering that the size of this type of data continues to increase through the use of smartphones, the need to analyze and understand such data has become a challenge.
DATA SOURCE	
Social media	Social media is the source of information generated via URLs to share or exchange information and ideas in virtual communities and networks, such as collaborative projects, blogs and microblogs, Facebook, and Twitter.
Machine-generated data	Machine data are information automatically generated from hardware or software, such as computers, medical devices, or other machines, without human intervention.
Sensing	Several sensing devices exist to measure physical quantities and change them into signals.
Transactions	Transaction data, such as financial and work data, comprise an event that involves a time dimension to describe the data.
Internet of Things	Internet of Things represents a set of objects that are uniquely identifiable as a part of the Internet. These objects include, e.g. smartphones, digital cameras, tablets, smartwatches... When these devices connect with one another over the Internet, they enable more smart processes and services that support basic, economic, environmental, and health needs.
DATA STORE	
Document-oriented	Document-oriented data stores are mainly designed to store and retrieve collections of documents or information and support complex data forms in several standard formats, such as JSON, XML, and binary forms (e.g., PDF and MS Word). A document-oriented data store is similar to a record or row in a relational database but is more flexible and can retrieve documents based on their contents (e.g., MongoDB, SimpleDB, and CouchDB).
Column-oriented	A column-oriented database stores its content in columns aside from rows, with attribute values belonging to the same column stored contiguously. Column-oriented is different from classical database systems that store entire rows one after the other (e.g. BigTable).

Graph database	A graph database is designed to store and represent data that utilize a graph model with nodes, edges, and properties related to one another through relations (e.g. Neo4j).
Key-value	Key-value is an alternative to relational database system that stores and accesses data designed to scale to a very large size. Examples of key-value stores are Apache Hbase, Apache Cassandra, Dynamo and Voldemort.
DATA STAGE	
Cleaning	Cleaning is the process of identifying incomplete and not correct data.
Transform	Transform is the process of transforming data into a form suitable for analysis.
Normalization	Normalization is the method of structuring database schema to minimize redundancy.
DATA PROCESSING	
Batch	MapReduce-based systems have been adopted by many organizations in the past few years for long-running batch jobs. Such system allows for the scaling of applications across large clusters of machines comprising thousands of nodes.
Real time	To continuously collect data at the same rate they are generated and promptly react to critical information related to business and operations. An example is S4, a real time process-based, scalable, partially fault tolerant, general purpose, and pluggable platform.

Table 2 Various categories of Big Data (Hashem et al. 2015)

1.4 Operational and Analytical Big Data

Big Data systems may be operational or analytical oriented. Operational systems provide real-time capabilities for capturing and storing flows of interactive workloads while analytical systems provide retrospective and complex analysis capabilities. These two types of systems are thus complementary and have different constraints. Operational and analytical systems have evolved to address their particular demands separately and in very different ways. Each has driven the creation of new technology architectures. Operational systems focus on servicing highly concurrent requests while exhibiting low latency for responses operating on highly selective access criteria. Analytical systems, on the other hand, tend to focus on high throughput; queries can be very complex and touch most if not all of the data in the system at any time (see Table 3). Both systems tend to operate over many servers operating in a cluster, managing tens or hundreds of terabytes of data across billions of records.

	Operational	Analytical
Latency	1 ms - 100 ms	1 min - 100 min
Concurrency	1000 - 100,000	1 - 10
Access Pattern	Writes and Reads	Reads

Queries	Selective	Unselective
Data Scope	Operational	Retrospective
End User	Customer	Data Scientist
Technology	NoSQL	MapReduce, MPP Database

Table 3 Overview of Operational vs. Analytical Systems (MongoDB Big Data Explained, 2016)

Big Data Analytics (also known as Analytics, Business Analytics or even Advanced Analytics and Next Generation Analytics) refer to techniques and technologies that are used to analyze Big Data and acquire information from it. Big Data Analytics do not require Big Data per se, but can explain trends or events that are not discernible with existing techniques and technologies.

Kwon et al. (2014) define Big Data Analytics *"as technologies and techniques that a company can employ to analyze large scale, complex data for various applications intended to augment firm performance in various dimensions."* Big Data Analytics provide algorithms for complex analysis of structured and/or unstructured data and includes sophisticated statistical models, machine learning, neural networks, text analytics and advanced data-mining techniques. Assunção et al. (2014) distinguish three categories of analytical methods that can be classified as descriptive, predictive, or prescriptive (see Table 4).

Category	Description
Descriptive	Describe models past behavior using historical data and statistical analysis to identify patterns
Predictive	Predicts future outcomes based on historical and current data
Prescriptive	Assists decision making by determining actions and assessing its impact concerning business objectives, requirements, and constraints

Table 4 Categories of Analytics (Assunção et al., 2014)

- Descriptive analytics is used, for example, to discover patterns in temperature fluctuations or traffic congestion.
- Predictive analytics is one of the most popular Big Data Analytics use cases and becomes more and more important to Business. It is based on statistical or data-mining solutions that can be used on both structured and unstructured data. Examples are to predict customers' next moves based on what they buy and when they buy it or to predict fraud with credit cards. Another example is recommendation systems whose goal is e.g. to encourage a customer to discover new products based on its preferences and historical purchases.
- Prescriptive analytics is used in so-called Complex Event Processing (CEP) that deals with a few variables in order to correlate it with a specific business process. It is based on event processing from a business process that collects and combines data from different sources to discover events that may result into action. An example is a loyalty real-time response when a client makes an on-line purchase.

To get information out of Big Data sources, a number of phases can be identified. Assunção et al. (2014) describe the four phases of an analytics workflow for Big Data (Figure 4):

1. Data from various sources, including databases, streams, data mart and data warehouses, are used to build models.
2. The large volume and different types of the data can demand pre-processing tasks for integrating, cleaning and filtering. The preparation of data for analysis is a time-consuming and labor-intensive task.
3. The prepared data are used to train a model and to estimate its parameters. Once the model is estimated, it should be validated. Normally, this phase requires the use of the original input data and specific methods to validate the created model. Finally, the model is applied to arriving data. This phase, called model scoring, is used to generate predictions, prescriptions, and recommendations.
4. The results are interpreted and evaluated, used to generate new models or calibrate existing ones, or are integrated to pre-processed data.

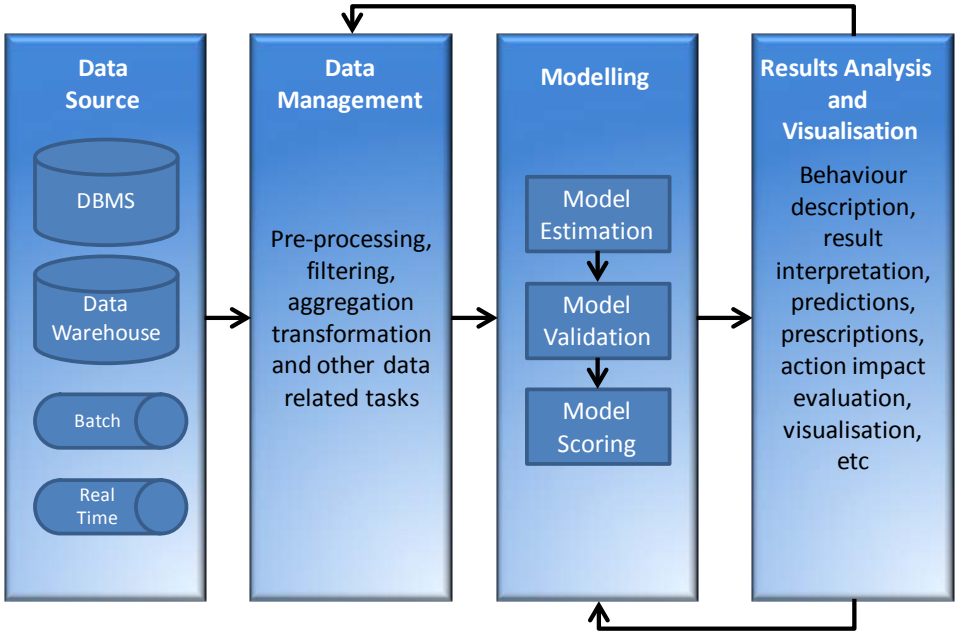

Figure 4 Overview of the Analytics workflow for Big Data (Assunção et al., 2014)

In the modeling phase, various statistical and data-mining related algorithms can be used such as:

- Neural networks: statistical learning algorithms inspired by biological neural networks;
- Logistic regression: statistical technique based on standard regression and extends it to deal with classifications;
- Classification trees: classifies dependent categorical variables based on measurements of one or more predictor variables.

With the ever increasing amount of data that Big Data Analytics need to cope with, good visualization and reporting tools are crucial. These have to assist in the three major types of analytics (see Table 4). However, dedicated hardware for visualization is becoming more and more important for Big Data Analytics as well (Assunção et al, 2014).

1.5 No Single Definition for Big Data

Despite, or maybe because of, the widespread interest for Big Data no unanimously accepted definition exists for Big Data (Mayer-Schönberger and Cukier, 2013). Based on Laney's (from

Metagroup) three Vs[10], several authors made additions to the Volume, Velocity and Variety characteristics of Big Data, which is reflected in the various definitions. Some relate to the data itself whereas others include technology aspects or aspects of data analysis.

Gartner for example, that acquired Metagroup in 2005, added information processing capabilities and now defines Big Data as follows:

> *Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making*[11].

Similarly, the TechAmerica Foundation has the following definition:

> *Big data describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information*[12].

Others refer to the potential issues with Big Data and Analytics. McKinsey defines Big Data as:

> *Data sets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze* (Manyika, 2011).

Or similarly:

> *Data sets whose size, type, and speed-of-creation make them impractical to process and analyze with traditional database technologies and related tools in a cost- or time-effective way*[13].

A definition from Berkeley University by Steve Todd[14]:

> *Big data is when the normal application of current technology does not enable users to obtain timely, cost-effective, and quality answers to data-driven questions*.

And similarly:

> *Big data is what happened when the cost of storing information became less than the cost of making the decision to throw it away*[15].

IDC argues that Big Data has three main characteristics: 1) the data itself, 2) the analytics of the data, and 3) the presentation of the results of the analytics. Therefore, they define it as

> *A new generation of technologies and architectures designed to extract value economically from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis.*

The IDC definition encompasses hardware and software that integrates, organizes, manages, analyses, and presents data. IDC, Oracle and Forrester all included in 2012 another Big Data characteristic being 'Value' (Chen, et al., 2012; Wamba, 2015) which is defined by the perceived value of the data and the technology to any given organization. However, 'Value' is an outcome of the analytics and not an ex-ante property of data. Others added yet different Vs such 'Volatility': "How long do you need to store this data?" and Visualization: "how to effectively show the derived

---

[10] http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[11] http://www.gartner.com/it-glossary/big-data/

[12] http://www.techamericafoundation.org/bigdata

[13] http://wikibon.org/wiki/v/Big_Data

[14] VP Strategy and Innovation at EMC, see http://stevetodd.typepad.com/my_weblog/2011/08/amped-at-uc-berkeley.html

[15] Tim O'Reilly quoting personal correspondence via email from George Dyson, 20 March 2013 (source: http://www.opentracker.net/article/definitions-big-data)

information?" (Sagiroglu and Sinanc, 2013). Adding such attributes that are not directly related to the technical characteristics of data, does not make it clearer since the distinction between the data itself and what one can do with it, such as capture, curation, storage, search, sharing, transfer, analysis, and visualization are separate things.

IBM, on the other hand, has added another fourth V "Veracity" that specifically relates to the data itself (Schroeck et al., 2012; Zikopoulos et al., 2013). Some data are inherently uncertain, such as weather conditions, sentiment of buyers of products, or statements in social networks. Despite the relevance of this data, the uncertainty can, in principle, not be eliminated through just any analytical method. Yet regardless of this uncertainty, the data may still contain valuable information. If this is acknowledged, organizations can embrace it and determine how to use it to their advantage.

The ISO/IEC JTC 1, (2014) [16] definition of Big Data introduces "Variability", which relates to a change in one or more of the other Big Data characteristics.

> *Big data is a data set(s) with characteristics (e.g. volume, velocity, variety, variability, veracity, etc.) that for a particular problem domain at a given point in time cannot be efficiently processed using current/existing/established/traditional technologies and techniques in order to extract value.*

[16] http://www.iso.org/iso/big_data_report-jtc1.pdf

2. Processing Big Data: a Technical Analysis

2.1 Introduction

Big Data technologies are the result of several decades of technological evolution, which includes advances in relational databases for managing structured data, technologies such as Object Database Management System (ODBMS) to store unstructured data (such as BLOBs, binary large objects) and natural language–based analysis tools to analyze unstructured data. Developments in Big Data technologies were triggered in the late 1990s by leading Internet companies that wanted to monetize on the value from their data. This required innovative approaches to store, access, and analyze massive amounts of data in near real time. For that purpose, research on tools and technologies such as in-memory databases, virtualization and Cloud Computing were driven.

In 2004, employees at Google began using Big Data algorithms to support distributed processing (Tambe, 2014). In particular, software related innovations for large-scale analytics based on the MapReduce parallel programming model, with Apache Hadoop being the most relevant implementation for Big Data Analytics, proved to be successful to process immense amounts of data efficiently and timely.

Big Data technologies allow organizations to store, modify and analyze vast amounts of data in order to leverage it, which would have been inconceivable only five years ago. Big Data has the potential to significantly change the way organizations do business, such as analyzing buying patterns to enhance product and service offerings, analyzing transaction data in real time to upsell music concert tickets or assessing on-line customers in real time to provide additional offers to that customer. In other words, businesses want to gain insights and actionable results from any kind of data at the right time, no matter how much data are involved.

Big Data technologies also bring major advantages for areas other than business, including science and governments. The incomprehensible amount of data generated in Large Hadron Collider research experiments of the European particle physics laboratory CERN, or the efficient analysis of the human genome can only be properly processed with such technologies. The same is true for governmental organizations that tap data from the Internet for antiterrorist activities or investigate fraudulent events based on voice data worldwide. So all this data looks like a gold mine, but as with a gold mine, one has to delve it first. Therefore, a better understanding of the technology around Big Data Analytics is necessary.

To get a better understanding on Big Data and Analytics, Pääkkönen and Pakkala (2015) created a technology agnostic reference architecture by integrating several approaches of companies that successfully applied the latest technologies. It provides an overall picture by containing typical functional components, data stores, and connecting data flows in a Big Data Analytics system. It is based on an analysis of published implementation architectures of Big Data use cases of companies, such as Facebook, LinkedIn, Twitter and Netflix. The reference architecture and associated classification of related products/services can be used for architecture design and the selection of technologies or commercial solutions, when creating Big Data systems.

Figure 5 presents the design of the high-level reference architecture in which data stores are presented as ellipsis, functionality as rectangles, and data flows as arrows. Similar functionalities are grouped into functional areas which are detailed in Table 5. The authors map the companies' Big Data infrastructures to their reference architecture and a key position is taken in by Apache Hadoop (the Apache implementation of the MapReduce paradigm and the storage system HDFS). They also map

other functionalities, such as streaming Analytics platforms and complex data analysis processes to their architecture showing the general applicability of this architecture.

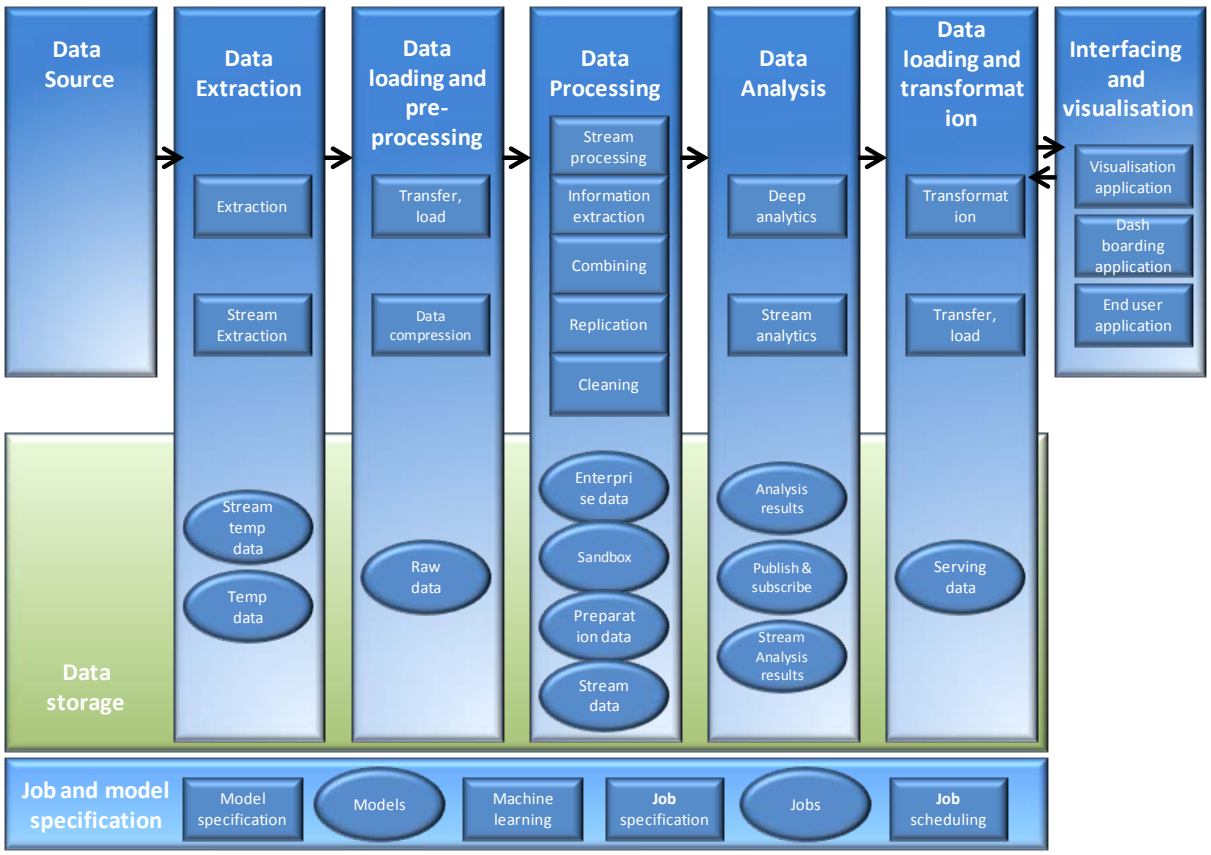

Figure 5 A technology independent Big Data Analytics reference architecture (Pääkkönen and Pakkala, 2015)

Stage	Description
1. Data extraction	Data extracted from data sources may be stored temporarily into a *temporary data store* or directly transferred, and loaded into a *Raw data store*. Streaming data may also be extracted, and stored temporarily.
2. Data loading and pre-processing	Data are transferred loaded and processed, such as data compression. The Raw data store contains unprocessed data.
3. Data processing	Data from the *Raw data store* may be *cleaned* or *combined*, and saved into a new *Preparation data store*, which temporarily holds processed data. *Cleaning* and *combining* refer to quality improvement of the raw unprocessed data. Raw and prepared data may be *replicated* between data stores. Also, new *information* may be *extracted* from the *Raw data store* for *Deep Analytics*. *Information extraction* refers to storing of raw data in a structured format. The *Enterprise data store* is used for holding of cleaned and processed data. The *Sand-box store* is used for containing data for experimental purposes of data analysis.
4. Data analysis	*Deep Analytics* refers to execution of batch-processing jobs for *in situ* data. Results of the analysis may be stored back into the original data stores, into a separate *Analysis results store* or into a *Publish & subscribe* store. *Publish & subscribe* store enables storage and retrieval of analysis results indirectly between subscribers and publishers in the system. Stream processing refers to processing of extracted streaming data, which may be saved temporarily

	before analysis. *Stream analysis* refers to analysis of *streaming* data, to be saved into *Stream analysis results*.
5. Data loading and transformation	Results of the data analysis may also be *transformed* into a *Serving data store*, which serve *interfacing and visualization applications*. A typical application for *transformation* and *Serving data store* is servicing of Online Analytical Processing (OLAP) queries.
6. Interfacing and visualization	Analyzed data may be visualized in several ways. *Dashboarding application* refers to a simple UI, where typically key information is visualized without user control. *Visualization application* provides detailed visualization and control functions, and is realized with a Business Intelligence tool in the enterprise domain. *End user application* has a limited set of control functions, and could be realized as a mobile application for end users.
7. Job and model specification	Batch-processing *jobs* may be specified in the user interface. The jobs may be saved and scheduled with *job scheduling* tools. Models/algorithms may also be specified in the user interface (*Model specification*). *Machine learning* tools may be utilized for training of the models based on new extracted data.

Table 5 Functional areas of a Big Data Analytics infrastructure (Pääkkönen and Pakkala, 2015)

2.2 Analytical Big Data: the MapReduce Paradigm

MapReduce is a programming paradigm for processing parallelizable problems across huge amount of data using a large number of computers. In the MapReduce model, the dataset to process is decomposed into smaller sub-datasets that will be processed separately on different computers (nodes), interconnected either locally (in this case we will refer as Cluster) or geographically distributed (in this case we will refer as Grid or Cloud). Then, each result of the processing is gathered to provide the final result. The data splitting process is called the **Map** step, the data gathering and delivering step is called the **Reduce** step. By this means, the time to process the large problem could be divided by the number of nodes used if the problem is perfectly parallelizable. This enable to leverage the complexity of a very large and time consuming computation by using many computers in parallel.

Initially designed in 2004 by Google, the MapReduce concept and its eponym implementation became rapidly successful. By the end of 2011, Apache released an open-source implementation of this model named Hadoop. At the time of writing, more implementations of the MapReduce paradigm are also available, in particular the next generation Apache Spark designed for performance (10 to 100 times faster than Hadoop) and portability (possibility to run on different platforms).

Although being a very effective multi-purpose model for batch processing, the MapReduce approach has a number of drawbacks (Fernández et al, 2014; Chen and Zhang, 2014):

- Apart from the fact that the MapReduce algorithm can only be applied for batch processes, it does not provide any significant improvement in performance when the work cannot be parallelized.
- Processing many small files or performing calculations with small data sizes incurs performance issues such as long startup time.
- Not all algorithms can be efficiently formulated in terms of Map and Reduce functions since these have a number of restrictions, such as implementation of iterative jobs or processing networked data such as graph structures.
- Because of the high latency characteristic of MapReduce, it is almost impossible to be applied for real-time Analytics.

- In order to fully benefit from the MapReduce features, specific expertise of programmer is required that should carefully tune several factors such as providing a file-system to store data, exploiting indexes, and the use of an efficient grouping and scheduling algorithm.

In the following sections, we present the two reference implementations of the MapReduce concept, which are the Google MapReduce framework and Apache Hadoop. Google's MapReduce framework was the first released implementation of the MapReduce paradigm while Apache Hadoop is currently *'de facto'* the most widely used implementation of this paradigm.

2.2.1 Google MapReduce

MapReduce is originally a programming model that was designed in 2004 by Google for processing and generating vast data sets using a parallel distributed algorithm on a cluster. In the meantime, Google released the MapReduce framework, the implementation of this model. At this time, the term MapReduce was referring to both the concept and its only implementation available: the proprietary Google technology. It is only later with the arrival of other implementations of the MapReduce concept (such as Hadoop) that MapReduce became genericized. In the following of this section, we consider that the term *MapReduce* refers to Google's implementation of the MapReduce paradigm.

The MapReduce framework orchestrates the processing by distributing computing tasks in parallel across systems, manages communications and data transfers, handles load balancing and provides techniques for redundancy and fault tolerance. The framework is inherently robust as it is designed based on the assumption that hardware failures should be automatically handled in software. It enables software developers to write programs that can process massive amounts of unstructured data in parallel. The MapReduce framework combines two existing capabilities, Map and Reduce, from functional computer programming languages, such as LISP, and has added scalability and fault-tolerance capabilities (Sagiroglu and Sinanc, 2013).

MapReduce is divided into two stages: the *Map* procedure carries out filtering and sorting, such as e.g. sorting staff by last name into queues, one queue for each name. It applies its computation to each data element of a list (defined as a key-value pair) and produces a new list. It uses input data and each element of the output list maps to a corresponding element of the input list. The *Reduce* procedure analyses and merges input data from the Map steps. It aggregates and summarizes the results, such as e.g. counting the number of staff in each queue, yielding name frequencies.

MapReduce has the following key behaviors:

- Scheduling: jobs are broken down into individual tasks for the map and the reduce elements.
- Synchronization: an execution framework keeps track of all mapping and reducing operations establishing a synchronization between them.
- Fault/error handling: if some of the mapping tasks are not executed correctly, the execution framework will assign the tasks to a different node to finish the job.
- Code/data co-location: keeping the data and the mapping functions on the same machine is strongly advised to get optimal MapReduce performance and high degree of fault tolerance, because this results in the smallest latency.

MapReduce is capable to take a very large dataset and break it into smaller, more manageable chunks, operate on each chunk independently, and then pull it all together at the end. It is able to run on inexpensive clusters of commodity hardware and standard networks and gets the same result as if all the work was done on a single machine. Over the years, a number of implementations of the MapReduce framework have been created and are available as both open source and commercial products. Hadoop MapReduce is part of the first category and is the most well-established software

platform that support data-intensive distributed applications (Chen and Zhang, 2014; Fernández et al., 2014).

Since June 2014, Google announced during the Google I/O conference in San Francisco that they were now replacing their internal usage of MapReduce to process Big Data by a newer version of their Big Data Analytics technology named DataFlow [17].

2.2.2 Apache Hadoop

Hadoop is a set of open-source software modules written in Java from the Apache Software Foundation[18], and consists of a set of algorithms for distributed storage and processing of massive data sets. It allows applications to run on large computer clusters of commodity hardware.

Hadoop parallelizes data processing across computing nodes, splits files into large blocks (default 64MB or 128MB) and distributes those across the nodes in the cluster for parallel processing to speed computations. Hadoop is able to process huge amounts of structured and unstructured data (terabytes to petabytes) and can be implemented on commodity servers as a Hadoop cluster. These servers can be added or removed from the cluster dynamically because Hadoop is designed to be highly fault tolerant. The core of Hadoop consists of two primary components: Hadoop MapReduce and Hadoop Distributed File System (Sagiroglu and Sinanc, 2013).

- Hadoop MapReduce**:** a scalable *processing engine* that computes each results in batch. It is a high-performance parallel/distributed data processing implementation of the MapReduce framework.
- Hadoop Distributed File System (HDFS): a scalable, fault tolerant, high-performance, low-cost, *data storage cluster* that is responsible for storing data on the clusters.

Hadoop's MapReduce and HDFS components were inspired by the Google papers on MapReduce and the Google File System.

I. Hadoop MapReduce

Hadoop MapReduce[19] can be seen as the engine of the Hadoop system, in which the main working flow can be described as follows: when a software client requests a MapReduce operation, the first step is to locate and read the input file containing the raw data. Although the file format is completely arbitrary, the data must be converted first into something the program can process. This is performed by the functions InputFormat (IF) and RecordReader (RR). IF decides how the file is going to be broken into pieces for processing and subsequently it assigns a RR to transform the raw data for processing by the map. Because the map and reduce operations work together, the program has to collect the output from independent mappers and pass it to reducers. This task is performed by the function OutputCollector. The function Reporter provides information gathered from map tasks so that it is known when the map tasks are complete. Hadoop MapReduce uses Hadoop YARN[20], which is a platform for scheduling of users' applications and managing computational resources.

[17] https://cloud.google.com/dataflow/

[18] The Apache Software Foundation, consisting of a decentralized community of developers, is an American non-profit corporation to support a broad range of open source software projects and provides financial, legal and organizational support. See https://www.apache.org/ and http://hadoop.apache.org

[19] http://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

[20] Yet Another Resource Negotiator (YARN) is a core Hadoop service next to MapReduce and HDFS.

II. Hadoop Distributed File System (HDFS)

The Hadoop Distributed File System (HDFS)[21] is the distributed, scalable, and portable file-system of the Apache Hadoop framework. To effectively support Big Data Analytics, data are written once and then read many times thereafter, instead of the constant read-writes of other file systems. It stores large files, typically from gigabytes to terabytes, across a commodity hardware cluster and provides the best performance when the entire cluster is in the same physical rack in a data center. The main difference with respect to other file systems is that HDFS is not implemented in the kernel of the operating system, but lies on top of the operating system stack and works as a process in the user space.

A Hadoop cluster has nominally a single NameNode and number of DataNodes. HDFS works by breaking large files into data blocks, which are stored on DataNodes. The NameNode keeps track of what blocks on which DataNodes make up the complete file. Within a HDFS cluster the NameNode manages the file operation, including creates, reads, updates and deletes. HDFS metadata (such as location of data blocks and rights to view or modify data) are stored in the NameNode. Furthermore, it is the NameNode's job to manage the complete collection of all the files in the cluster, which is referred to as the file system namespace.

DataNodes store the actual data and request the NameNode whether there is anything to do. The data nodes also communicate among themselves so that they can cooperate during normal file system operations. Data blocks are replicated across several DataNodes, so failure of a server not necessarily corrupts a file. DataNodes provide messages to ensure connectivity between the NameNode and the DataNodes. When a connection is longer present, the NameNode un-maps the DataNode from the cluster and assigns a replica to proceed with the MapReduce calculations as if nothing happened. The degree of replication, the number of DataNodes, and the HDFS namespace are established when the cluster is implemented. All such parameters can be adjusted during the operation of the cluster.

The term "Hadoop" often refers not only to the Apache base modules described above but also to a collection of additional software packages that can be installed on top of or alongside these base modules. Such software packages make up the so-called Hadoop Ecosystem that will be presented in the next sub-section.

III. Hadoop Ecosystem

The Hadoop ecosystem[22] is a large, growing set of tools and technologies designed specifically to smooth the development, deployment, and support of Big Data solutions. The current set of Apache Big Data applications consists of around a dozen of software tools with various functionalities (see Table 6).

Software tool	Brief description
Ambari	A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters. It also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually along with features to diagnose their performance characteristics.

[21] http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsUserGuide.html
[22] http://hadoop.apache.org/

Avro	A system of data serialization. The tasks performed by Avro include data serialization, remote procedure calls, and data passing from one program or language to another. In the Avro framework, data are self-describing and are always stored with their own schema. The software is suitable for application to scripting language, such as Pig.
Cassandra	A scalable multi-master database with no single points of failure. It supports replicating across multiple datacenters and provides low latency. Cassandra's data model (a popular NoSQL database) offers column indexes with the performance of log-structured updates, support for de-normalization and materialized views, and built-in caching.
Chukwa	A data collection system for managing large distributed systems. It is incorporated with MapReduce and HDFS; the workflow of Chukwa allows for data collection from distributed systems, data processing, and data storage in Hadoop.
HBase	A scalable, distributed storage system for structured data that allows random, real time read/write access. It is an open-source, distributed, non-relational database modeled after Google's BigTable and leverages the distributed data storage provided by the Google File System on top of Hadoop and HDFS. HBase is one of the NoSQL data base implementations (see Section 2.3.)
Hive	A data warehouse infrastructure for HDFS, originally developed by Facebook, which provides data summarization, ad hoc querying, and managing large datasets residing in distributed storage. Hive provides a mechanism to query the data using a SQL-like language called HiveQL. At the same time, this language also allows traditional MapReduce programmers to plug in their custom mappers and reducers when it is inconvenient or inefficient to express this logic in HiveQL.
Mahout	A scalable machine learning and data mining library. It has four main groups: collective filtering, categorization, clustering, and parallel frequent pattern mining; compared with other pre-existing algorithms, the Mahout library belongs to the subset that can be executed in a distributed mode and is executable by MapReduce.
Pig	A high-level data-flow language and execution framework for parallel computation originally developed by Yahoo! to process data on Hadoop. It involves a high-level scripting language (Pig Latin) and offers a run-time platform that allows users to execute MapReduce on Hadoop
Spark	A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
Tez	A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute tasks to process data for both batch and interactive use-cases. Tez is being adopted by Hive, Pig and other frameworks in the Hadoop ecosystem and by other commercial software (e.g. ETL tools), to replace Hadoop MapReduce as the underlying execution engine.
ZooKeeper	A high-performance coordination service for distributed applications. It allows distributed processes to manage and contribute to one another through a shared hierarchical namespace of data registers (z-nodes) similar to a file system; Zoo

	Keeper is a distributed service with master and slave nodes and stores configuration information.

Table 6 Hadoop-related projects at Apache (Hashem et al., 2015)

The Hadoop ecosystem and the supported commercial distributions are changing at a very rapid pace. New tools and technologies are introduced, existing technologies are improved, and some technologies are retired by improved replacements. In the next sections, an important ingredient of any operational Big Data environment is presented: the so-called NoSQL databases, followed by Big Data tools for analysis, visualization, decision making and reporting.

2.3 Operational Big Data: Non-Relational Databases

At the core of any Big Data platform are databases that contain collections of data elements relevant to the organization. Traditional relational databases cannot be utilized when dealing with huge amounts of disparate data types because the nature of Big Data does not conform to the relational model. To allow storage of Big Data non-relational databases have been designed, which are collectively called NoSQL (Not only SQL)[23] databases. NoSQL databases provide a mechanism to store and retrieve data that are modeled in other means than the tabular relations used in relational databases. Because no relational scheme is required, the data can be inserted in a NoSQL database without defining any table structures. Furthermore, these databases provide great flexibility, since the data format may change at any time, without stopping the application. The particular suitability of a given NoSQL database depends on the problem to solve. Requirements for common data definitions and requirements to extract and transform data sources hold for both relational and non-relational data bases. However, NoSQL databases will be primary used for Big Data analysis efforts, while the traditional relational databases will be used for transactional and history purposes.

As data resides in several redundant servers, NoSQL databases have a distributed and fault-tolerant architecture. Therefore, such a system is easily scalable by adding or removing servers, whereas the failure of one single server can be managed without difficulties. Because NoSQL databases do not conform to the relational model, the well-known ACID[24] properties of a database management system no longer hold. Usually, a final consistency is given only, or the transactions are limited to unique data items. Given a sufficiently long period of time in which no changes are made, the updates are propagated over the system (Fernández et al., 2014).

Hbase (part of the Hadoop ecosystem) is a well-known NoSQL database and probably the most popular one is Apache Cassandra (Chen and Zhang, 2014). HBase uses the Hadoop file system and MapReduce engine for its core data storage needs. Cassandra was a Facebook proprietary database, but has been released as open source in 2008. Other NoSQL implementations include Google BigTable (one of the first implementations of this type of database), MongoDB, and Voldemort. Fernández et al. (2014) provide an overview of categories of NoSQL databases (Table 7).

Category	Description	Examples
key-value pair database	In key-value pair databases, the data are considered to be a single opaque collection which may have different fields for every record.	Redis http://redis.io Riak http://basho.com/riak

[23] Its name easily leads to the misunderstanding that NoSQL means 'no SQL'; however, a query language similar to SQL can be used although it is not fully conformant.

[24] ACID (Atomicity, Consistency, Isolation, Durability) is a set of properties which guarantee database transactions are processed reliably.

	This offers great flexibility and scalability. Most of the data are stored as strings.	Voldemort http://www.project-voldemort.com/voldemort DynamoDB http://aws.amazon.com/dynamodb
document database	Document-oriented databases are inherently a subclass of the key-value store but the data processing relies on the internal structure of documents in order to extract metadata that the database engine uses for further optimization.	MongoDB http://www.mongodb.org CouchDB http://couchdb.apache.org
columnar database	In columnar databases data are stored across rows. It is very easy to add columns that can be added row by row, thereby offering great flexibility, scalability and performance.	HBase http://hbase.apache.org Cassandra http://cassandra.apache.org
graph database	Graph databases are based on a so-called "node-relationship." This structure is most useful when dealing with highly interconnected data and are often used to model and optimize networks (e.g. telecommunications) or to manage geographic data (e.g. oil exploration).	Neo4J http://www.neo4j.org HyperGraphDB www.hypergraphdb.org/
spatial database	A spatial database (or geodatabase) is optimized to store and query data that represents objects defined in a geometric space. It allows representing geometric objects such as lines, points, and polygons.	PostGIS www.postgis.org

Table 7 Categories of NoSQL databases and some examples (based on Fernández et al., 2014)

2.4 Other Big Data Tools

Next to the Apache Hadoop related Big Data software tools, many other applications exist that can be divided into three categories (Chen and Zhang, 2014): those for batch processing, stream processing and interactive analysis. In these fields, most of the innovation takes place nowadays and new tools emerge around the clock.

2.4.1 Batch processing

As discussed earlier, Apache Hadoop (including Mahout) is specially designed for batch processing. Other tools in this category come from various other sources and examples are listed in Table 8.

Software tool	Brief description
Jaspersoft BI suite	The Jaspersoft package is a commercial open source software vendor, owned by TIBCO that produce reports from database columns. The software has already been installed in many business information systems. It is a scalable Big Data analytical platform and fast to get started with no need for ETL.

	https://www.jaspersoft.com/
Pentaho business analytics	Pentaho is another software platform for Big Data. It also generates reports from both structured and unstructured large volume of data. http://www.pentaho.com/5.0
Skytree server	Skytree Server is the first general purpose machine learning and advanced Analytics system, designed to accurately process massive datasets at high speeds. It offers many sophisticated machine learning algorithms. http://www.skytree.net/
Tableau	Tableau has three main products to process large-scale data set, including Tableau Desktop, Tableau Sever, and Tableau Public. http://www.tableau.com/
Talend Open Studio	Talend Open Studio is open source software for Big Data applications that provides users graphical environment to conduct their analysis visually. https://www.talend.com/products/talend-open-studio

Table 8 Big Data tools for batch processing (Chen and Zhang, 2014)

2.4.2 Stream processing

For certain streaming data applications, real-time response is required for processing large amount of stream data. This includes data from Smart Meters, industry sensors, machine-to-machine communications and telematics. Several Big Data tools based on stream processing are designed specifically for real-time stream Big Data Analytics. Next to the well-known Storm platform, a number of other software tools in this category are listed in Table 9.

Software tool	Brief description
Kafka	Kafka is a high-throughput messaging system written in Scala that was developed by LinkedIn and is now an open-source message broker project developed by the Apache Software Foundation. It works as a tool to manage streaming and operational data via in-memory analytical techniques for obtaining real-time decision making. http://kafka.apache.org/
SAP Hana	SAP Hana is an in-memory Analytics platform developed by SAP that aims to provide real-time analysis on business processes, predictive analysis, and sentiment data processing. http://hana.sap.com/abouthana.html
Splunk	Splunk is a real-time Big Data platform for exploiting information from machine-generated Big Data. http://www.splunk.com/

s-Server	SQLstream s-Server is another Big Data platform that is designed for processing large-scale streaming data in real-time. It focuses on intelligent and automatic operations of streaming Big Data. SQLstream is appropriate to discover patterns from large amounts of unstructured log file, sensor, network and other machine-generated data. SQLstream works very fast, as it uses in-memory processing. The standard SQL language is still adopted in the underlying operations. http://www.sqlstream.com/blaze/s-server/
Storm	Apache Storm is a free and open source distributed, fault-tolerant real-time computation system, specifically designed for processing limitless streaming data. Storm makes it easy to reliably process unbounded streams of data, doing for real time processing what Hadoop does for batch processing. Storm can be used with any programming language, and has many use cases such as real-time Analytics, online machine learning, continuous computation, distributed RPC and ETL. https://storm.apache.org/
S4	S4 is a general-purpose, distributed, scalable, fault-tolerant, pluggable computing platform for processing continuous unbounded streams of data. It was initially released by Yahoo! in 2010 and has become an Apache Incubator project since 2011. http://incubator.apache.org/s4/

Table 9 Big Data tools for stream processing (Chen and Zhang, 2014)

2.4.3 Interactive analysis

In an interactive Big Data Analytics environment users can undertake their own analysis of data. The data can be reviewed, compared and analyzed in tabular and/or graphic format at the same time. Table 10 specifies some of those tools with this capability.

Software tool	Brief description
Dremel	Dremel is a distributed system developed by Google for interactively querying large datasets and powers Google's BigQuery service (a web service that enables interactive analysis of massively large datasets working). It is scalable for processing nested data. Dremel has a very different architecture compared to Apache Hadoop, and acts as a successful complement of Map/Reduce-based computations. http://research.google.com/pubs/pub36632.html
Apache Drill	Apache Drill is an open source, low latency SQL query engine for Hadoop and NoSQL. Drill is the open source version of Google's Dremel system. Drill supports various query languages, data formats and data sources. http://drill.apache.org/

Table 10 Big Data tools for interactive processing (Chen and Zhang, 2014)

2.5 Challenges

Although there are nowadays many powerful technologies designed to address Big Data questions, there are still many data structuring and exploitation challenges that lie in areas such as data capture/staging, storage, analysis, security and visualization (Chen and Zhang, 2014; Fernández et al., 2014; Hashem et al, 2015). If those challenges were not addressed effectively, Big Data would become the gold mine that could not be delved.

- **Data capturing and staging** - Adequate accessibility of Big Data is essential for Analytics and knowledge discovery. However, how to handle the ever increasing Big Data Vs (volume, velocity, variety and veracity of data) especially when the data are poly-structured? Moreover, how to aggregate and correlate streaming data from multiple sources? Transforming and cleaning such data before loading it into databases for analysis are yet other challenging tasks. New protocols and interfaces are necessary to manage data of heterogeneous nature (structured, unstructured, semi-structured) and sources.
- **Data storage -** Projections suggest that the growth of data will outpace [25] foreseeable improvements in costs and density of storage technologies, the available computational power for processing it, and the associated energy footprint (Kambatla et al., 2014). Current storage technologies (hard disk drives, HDD) cannot possess the same high performance for both sequential and random I/O simultaneously. This requires a rethinking of how to design storage subsystems for Big Data processing systems. So, how to store large volumes of data in a way it can be timely retrieved? New developments such as solid-state drives that replace HDDs and phase-change memory[26] could provide relieve but are probably far from enough. In addition, current Cloud technologies do not provide the necessary high performance. Furthermore, how to store data in a way that it can be easily migrated between data centers and Cloud providers?
- **Data analysis** - The selection of an appropriate model for large-scale data analysis is critical. However, current algorithms are inefficient in terms of Big Data analysis. Furthermore, the Big Data Vs are increasing at exponential rate, but the improvement of information processing methods is relatively slower. Another challenge is scalability when dealing with Big Data analysis. For real-time Big Data applications, like financial marks, social networks, traffic navigation, transport systems, timeliness is a top priority. The need to process continually increasing amounts of disparate data is one of the key factors driving the adoption of Cloud services. But there are still many unanswered questions, such as how to optimize resource usage and energy consumption when executing Analytics applications?
- **Data security** - Security challenges have to deal with the confidentiality, integrity, and availability of the Big Data. Once Big Data environments are outsourced to Cloud service providers, several threats and issues become even more important, such as, intellectual property protection, privacy protection, commercial secrets and financial information protection. At the very least, Cloud vendors must ensure that all service level information security agreements are met. And with Cloud data storage becoming more and more popular, the network bandwidth capacity is the main bottleneck in Cloud and distributed systems that impacts the availability of the Big Data environment. Furthermore, in (ENISA, 2015), the most prominent Big Data security challenges highlighted are: access control and authentication, secure data management, source validation and filtering and software and infrastructure security.
- **Data visualization** - The key objective of data visualization is to represent knowledge as intuitively and effectively as possible, for example using real-time adjustable 3D graphs. For Big

[25] Data traffic grew 56-fold between 2002 and 2009, compared to a corresponding 16-fold increase in computing power

[26] A type of non-volatile random-access memory that is 500 to 1,000 times faster than conventional (flash) memory and it also uses up to half the power.

Data applications, it is especially hard to conduct data visualization because of the Vs of Big Data. Current Big Data visualization tools, on the whole, have issues regarding functionality and performances in scalability and response times.

Moreover, in addition to these technical challenges, other regulatory challenges arise from the application of Big Data in the society at large and threaten to slow this momentum. Legal and IT governance issues must be expressly addressed, in particular through the normative spectrum.

- **Legal aspects** – Specific laws, regulations and standards must be established precisely to preserve the private and/or sensitive information of the users. This aspect goes beyond data security offered by the service level information security agreements of Big Data operators. Different countries have different laws and regulations to address privacy protection and the use of standardization is promoted to ensure that this protection is adequate and consistent.
- **Governance** – Data governance embodies the control and authority exercised by the governments on data. To respect the data-related rules on transparency, accountability and information retention, one major Big Data challenge is how to address this governance compliance with massive amounts of data from multiple external sources. **Data provenance** is also one key aspect, as the complexity to deal with data under different regulations may rapidly become a challenge impossible to address when striving to reconcile legal aspects, governance rules of law and data privacy. Therefore, a clear, acceptable and standardized data policy with regards to the type of data stored must be defined.

Over time, Big Data infrastructures should emerge where users can assemble a solution from a palette of components that can be mixed and matched. This requires at least standardized interfaces for rapid rollout of Big Data appliances. And, finally yet importantly, organizations need well educated programmers and "data scientists" to be able to get the most out of these appliances.

3. Business and Economic Prospective Analysis

3.1 Introduction

The Big Data market, although still in an immature shape, continues to grow rapidly. According to estimates from Wikibon[27], the Big Data market will double between 2014 and 2017. In fact, there is ample evidence that investments in Big Data capabilities can unleash major positive impacts on reducing business costs, improving business insights, and unraveling strategic information, and subsequently boosting quality and effectiveness of corporate decision making (Kwon et al., 2014). Executives from companies that reported positive ROIs stated that the biggest benefits from Big Data capabilities included obtaining a better understanding of their consumer base, being able to better predict consumer loyalty, and the ability to effectively evaluate product performance. In addition, some firms that combined their own data with additional third party Big Data had advantage over competitors to provide an innovative service (Villas-Boas, 2014).

From 2006 to 2011, Hadoop investments were linked to a 3% faster productivity growth, for firms with significant data assets and adequately technical skilled staff (Tambe, 2014). Research in industry shows that retailers can achieve up to a 20% ROI increase by putting Big Data into Analytics (Perrey, Spillecke, Umblijs, 2013). Manyika et al. (2011) specified the transformative potential of Big Data in five domains:

1. Retail: behavior analysis, performance improvements, product placement design, optimization in distribution and logistics, optimization in price and variety, optimization of work, web based markets;
2. Manufacturing: demand forecasting, production operations, sales support, supply chain planning, web search based applications;
3. Public sector: automated decision making, customization of suitable products and services, discover needs, improve performance, innovate new products and services, risk management;
4. Healthcare: analyze disease patterns, clinical decision support systems, individual analytics for patient profiles, improve public health, personalized medicines;
5. Personal location data: geo targeted emergency response, smart routing or urban planning.

With respect to the *retail* domain, Kambatla et al. (2014) provide examples of the vast amounts of behavior data, including customer transactions, inventory management, store-based video feeds, advertising and customer relations, customer preferences and sentiments, sales management infrastructure, and financial data, among others. Given the soaring popularity of smartphones, retailers will soon have to cope with numerous streaming data sources that demand real-time analytics (Gandomi and Haider, 2015).

Wrobel (2012) gives examples of Smart product development that are made possible with Big Data appliances in the *manufacturing* domain:

- Smart homes and appliances without programming efforts
- Virtual assistants for various issues
- Machines and equipment with self-maintenance functions
- Electric vehicles as part of a Smart Grid
- Service robotics in complex environments
- Autonomous vehicles

[27] http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017

Concerning the *public* sector, Chen and Zhang (2014) discuss that people of different age groups need different public services and its related data. For example, teenagers need more education related information whereas elders require higher level of healthcare services. Fraud detection is a significant Big Data Analytics application in the public sector and mobile electronic voting will become viable in the near future (Kambatla et al., 2014). Furthermore, a traffic management system using Big Data can reduce traffic jams on highways to decrease accidents, save fuel, and reduce air pollution.

Kambatla et al. (2014) also mention *healthcare* informatics as one of the largest integrated distributed environments covering individual medicine prescriptions to global disease patterns. Another Big Data application in healthcare is to monitor premature infants to determine when data indicates an intervention is required.

Regarding *personal* location data, they provide an example of the speed and impact of information flow in social networks: ¨*When a 5.9 Richter earthquake hit near Richmond, VA, on August 23rd, 2011, residents in New York City read about the quake on Twitter feeds 30 s before they experienced the quake themselves*¨.

Kambatla et al. (2014) expand the potential of Big Data to other domains:

6. Nature and natural processes: environmental footprint from satellite imagery, weather radar, and terrestrial monitoring and sensing devices that is stored in various datasets (land use, deforestation, carbon emissions, terrestrial and satellite-based imaging, and computational modeling)
7. Computational and experimental processes in science: from quantum-mechanical modeling to astro-physical simulations.

3.2 Implementation

When creating a Big Data Implementation Road Map, at least the following five factors need to be considered (Hurwitz et al., 2013).

- **Business impact** - because the adoption of a Big Data program has broad implications for the company's overall strategy, the time and effort required to design Big Data solutions must be clearly noted on the project portfolio road map.
- **Projected capacity** – it must be very clear how much data are required and how fast it needs to be analyzed, since this provides the context for the phases of the road map.
- **Preferred software development method** – because Big Data projects are well suited for an agile development process, iterative methodologies should be considered. Short time cycles with rapid results and continuous user involvement will incrementally deliver a proper business solution.
- **Available budgets and skill sets** - understand the expected investments and required knowledge for the Big Data implementation and ensure appropriate budget and sponsorship.
- **Risk appetite** - depending on the kind of business, one may be forced to take less or more risk with the Big Data project, for example regarding the scope and expected benefits.

In order to reap the full benefits from Big Data investments, managers need to align existing organizational culture and capabilities across the organization (Wamba, 2015). A key challenge is to make Big Data trustworthy and understandable to employees. Return on Investment in Big Data will be achieved unless employees at all levels are not able to understand and include data in their decision making (Shah, et al., 2012).

Furthermore, firms with satisfying experience in utilizing internal data sources may not be too enthusiastic about Big Data Analytics as this requires a costly investment in its deployment and has a steep learning curve (Kwon et al., 2014). These firms may be hesitant to take risks and to adopt an innovative IT capability that is not directly linked with their existing skills and experience. Therefore, as a pre-condition of successful Big Data adoption, any company that considers successful deployment of Big Data has to put much effort that relevant new IT competencies are made available.

Accenture (2014) looked into the main challenges of implementing Big Data capabilities in a company. Respondents answered that challenges from security, budget and lack of talent, were the biggest hurdles to overcome (Figure 6). Furthermore, integrating existing information systems and limitations of Big Data vendors were reported by one third of the respondents. Bringing together the required expertise becomes critical for successful Big Data projects. Especially in the areas of security and integration, Big Data related standards are welcome and some initiatives have been started in these areas (see Section 4).

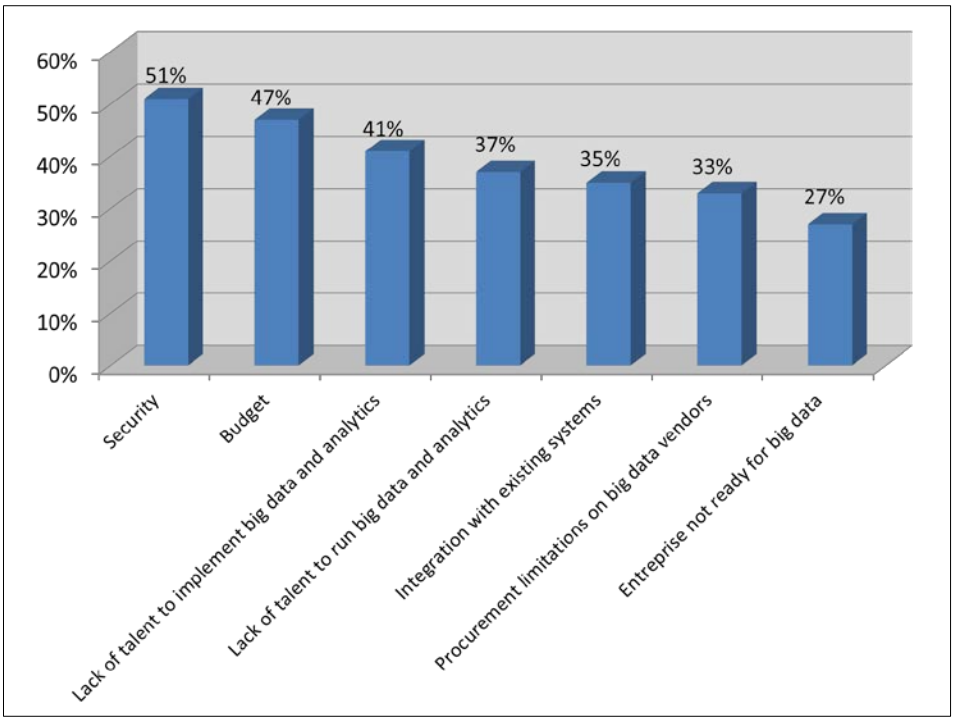

Figure 6 Main challenges with Big Data projects (Accenture, 2014)

Hurwitz et al. (2013) specified 10 Big Data good practices that companies may use when implementing a Big Data capability (Table 11). Next to technological options, requirement gathering, stakeholder management and setting realistic expectations are equally important to the success of any Big Data endeavor.

Good practice	Description
1. Understand your goals	After some experiments have been carried out and have a good understanding of what might be possible, set both short- and long-term goals. Collaboration between IT and Business Units are essential and set realistic expectations. As a Big Data capability affects every aspect of the organization the stakeholders in the Business need to be involved. This includes historical data and the information sources managed by different Business Units.

2. Establish a road map	Start with a one- to two-year road map with well-defined goals and outcomes. This road map sets foundational services that help the company get started. Part of the road map should include existing data services. Include both business and technical goals as part of the road map.
3. Discover your data	Start with a discovery process to understand what data are already available, where it is, who controls it, and how it is currently used.
4. Figure out what data you don't have	Determine what data are lacking for the Big Data capability. Involve business leaders, as these understand best what is keeping them from making even better decisions.
5. Understand the technology options	Understand the value of technologies such as Hadoop, NoSQL and streaming data offerings. Look at different types of databases and the integration possibilities.
6. Plan for security in context with Big Data	Data must be protected against internal and external risk. For example, some of this data needs to be treated as private and must be coded so that no one can access it without proper authorization.
7. Plan a data governance strategy	Accountability for managing data, including metadata, in the right way is fundamental to any adequate data governance strategy and effective data governance is a prerequisite for proper data quality.
8. Plan for Data Stewardship	Management and oversight of the organization's data assets by dedicated staff, both the content and metadata, will assist Business users with high-quality data that is easily accessible in a consistent manner.
9. Continually test your assumptions	Even if all processes are in place, e.g. to ensure that the right controls are in place and the correct metadata is being used, do not just assume that the data are always right. It is still important to check the data quality continuously.
10. Study best practices and leverage patterns	It is always better to find ways to learn from others rather than to repeat mistakes that others have made. Ensure proper education of staff and do not be averse in hiring experts.

Table 11 Ten Big Data Good Practices (Hurwitz et al., 2013)

After implementation Hurwitz et al. (2013) proposes three more stages: monitoring, adjusting, and experimenting.

- Monitoring in real time - Big Data Analytics allows monitoring data in (near) real time. This can have a profound impact on the organization.
- Adjusting the impact - it may be required to adjust business processes based on Big Data Analytics.
- Experimentation – the combination of experimentation, real-time monitoring and adjustment may transform the organization strategy in the end.

3.3 Other Considerations

3.3.1 Data quality

Kwon et al. (2014) discuss the importance of the quality of corporate data for effective Big Data Analytics adoption. They use the aspects of data completeness and data consistency as primary indicators of quality management of corporate data resources. Data completeness represents the extent to which all necessary data are successfully stored and managed. Data governance and metadata management are key elements for ensuring data consistency (Wamba, et al., 2015).

Although Big Data will introduce a new level of integration complexity, some basic principles still apply. Delivering quality and trusted data to the organization at the right time and in the right context remains a fundamental capability. Therefore, if not already in place, common rules for data quality must be established with emphasis on completeness and consistency. Furthermore, a comprehensive approach to assure the data life cycle throughout the organization needs to be in place to support integration of all types of data. This includes the data's origins and where it moves over time (i.e. data lineage).

In order to deliver information to the business in a trusted, controlled, consistent, and flexible way across the organization (Hurwitz et al., 2013):

- One must have a common understanding of data definitions.
- One must establish of a set of data services to qualify the data and make it consistent and ultimately trustworthy.
- One needs a streamlined way to integrate Big Data sources and systems of record.

Poor data quality and/or ineffective data governance are key challenges for Big Data (Wamba, et al., 2015). Data governance ensures ownership and an accountability framework for data management (DAMA, 2013). Adequate metadata[28] management ensures, among others, that common definitions of data elements are applied throughout corporate data sources. Lack of consistency in data is an obstacle in the successful processing if corporate data and external-source data are joined for Big Data Analytics and subsequent decision making. In addition, data profiling[29] tools should be used in the data quality process to improve content, structure and condition of the data.

3.3.2 Cloud Computing

Setting up and using a Big Data capability is still a challenging and time demanding task that requires much effort, expensive computational software and a large storage infrastructure. Furthermore, the amount of data currently generated by various activities in society is growing at an ever increasing speed. Start-up costs can be contained by finding cost effective environments.

Assunção et al. (2014) argue that Cloud computing helps in alleviating storage problems as this allows companies to provide resources on-demand and with costs proportional to the actual usage. It also allows companies to scale storage infrastructures up and down rapidly based on the current demand. Fernández et al. (2014) discuss two main advantages of Big Data with Cloud Computing: 1) the transparency for the programmer, who just needs to focus on the development of execution engines (e.g. Map and Reduce functions) and not in the inner operation of the storage system, and 2) the robust fault tolerance that allows a higher scalability and reliability for long running jobs.

[28] Data about data, such as allowed lower and upper bounds and definitions of terms.

[29] The process of examining the data available in an existing data source (e.g. a database or a file) and collecting statistics and information about that data. The purpose of these statistics may be to find out whether existing data can easily be used for other purposes.

Although Cloud infrastructures offer such elastic capacity to supply computational resources on demand, Assunção et al. (2014) also acknowledge that the area of Cloud supported Analytics is still in its infancy.

3.3.3 Data security

When companies start with Big Data analysis, they often overlook to maintain the same level of data security that is maintained in traditional data management environments (Hurwitz et al., 2013).

Once the Big Data has been acquired, the organization will be subject to compliance issues if it is not managed securely. The nature of the Big Data sources needs to be assessed in order to determine how much it can be trusted. For example, incorporating a source that includes sensitive personally identifiable information could put an organization reputation and its customers at risk. Masking private information when doing analysis on terabytes of data is typically disregarded at the outset, although this just needs to be done to meet privacy requirements. Alternatively, results of the Analytics may become corporate intellectual property or are essential to determine the next best action in a new product strategy. Either way, such information has to be secured properly so it does not put the organization at risk.

Furthermore, the acquired Big Data must be securely stored and checked against intrusion. Some of the data are probably not required and should be properly disposed of. In addition, some of the data sources may come from third-parties that require licenses. In order to make sure the organization does not violate rules and regulations, it must check whether one is allowed to use the data in the first place. Data and Information security awareness is essential to ensuring that everyone in the organization has an understanding of his roles and responsibilities with regard to security. Many international standards can provide guidance in this matter. The ISO/IEC 27000-series[30] consisting of various information security standards are a good reference in this context.

[30] http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=66435

4. Big Data Standardization

Standards for Big Data technologies are essential to improve, for example, the interoperability between the various applications and prevent vendor lock in. Standards can also help to prevent over fitting in Big Data. This happens when analytics designers tweak a model repeatedly to fit the data and being to interpret noise or randomness as truth. Another potential interest for standardization for Big Data is to support integration of multiple data sources. Security and Privacy are of paramount importance for both data quality and for protection. Some of the large volume of data come from social media and medical records and inherently contain private information. Analysis of such data, particularly in presence of context, must protect privacy. Big Data systems should be designed with security in mind. If there is no global perspective on security then fragmented solutions to address security may not offer full security, instead a partial sense of safety. Standards will play an important role for data quality and data governance to address veracity and value of data. The standards community has set up several initiatives and working groups on Big Data. Standards are being developed by a variety of Standards Developing Organizations (SDOs):

- Formal standards bodies, meaning an organization benefiting from a broad recognition and complying with the principles set out in annex 3 of the World Trade Organization (WTO) Technical Barriers to Trade (TBT) Committee agreement *Code of Good Practice for the preparation, adoption and application of standards*. Their primary activities are to develop, coordinate, promulgate, and produce *de jure* (formal) standards. Three are three formal standards bodies recognized broadly: the International Organization for Standardization (ISO), the International Electrotechnical Commission (IEC) and the International Telecommunication Union (ITU). There are also different regional formal standards bodies. For the European Union, the three recognized standards bodies are: The European Committee for Standardization (CEN), the European Committee for Electrotechnical Standardization (CENELEC) and the European Telecommunications Standards Institute (ETSI).
- Private standards bodies, commonly called "*consortia*" or "*fora*" for the ICT sector. In the standardization context, they are organizations regrouping individuals, companies, associations or governments with a common objective of participating in the creation of technical specifications or *de facto* standards, meaning these standards benefit from a widely recognition of the market. Some very well established consortia are the Institute of Electrical and Electronics Engineers (IEEE), the W3C, the Open Geospatial Consortium, the Organization for the Advancement of Structured Information Standards (OASIS).

4.1 Big Data Standardization from Formal Standards Body

ISO and IEC created a joint technical committee (JTC), ISO/IEC JTC 1[31], to provide a single comprehensive standardization committee in which to address international Information Technology and Information and Communications Technology standardization for business and consumer applications. The JTC 1 Committee is one of the largest and most prolific technical committees in international standardization. It is globally recognized as the focal point of formal standardization in ICT, which encompasses all technologies for the capture, storage, retrieval, processing, display, representation, organization, management, security, transfer, and interchange of data information.

[31] The reader is invited to visit the ISO/IEC JTC 1 webpage for additional information at http://www.iso.org/iso/home/standards_development/list_of_iso_technical_committees/jtc1_home.htm

ISO/IEC JTC 1 (2014)[32] specified a number of specific Big Data standard initiatives. It has identified current gaps in Big Data standardization and describes broad areas that are of interest to further investigate whether potentials standards should be made to close these gaps. The identified gaps in standardization activities related to Big Data are in the following areas:

1. Big Data use cases, definitions, vocabulary and reference architectures (e.g. system data, platforms, online/offline, etc.);
2. Specifications and standardization of metadata including data provenance;
3. Application models (e.g. batch, streaming, etc.);
4. Query languages including non-relational queries to support diverse data types (XML, RDF, JSON, multimedia, etc.) and Big Data operations (e.g. matrix operations);
5. Domain-specific languages;
6. Semantics of eventual consistency;
7. Advanced network protocols for efficient data transfer;
8. General and domain specific ontologies and taxonomies for describing data semantics including interoperation between ontologies;
9. Big Data security and privacy access controls;
10. Remote, distributed, and federated Analytics (taking the Analytics to the data) including data and processing resource discovery and data mining;
11. Data sharing and exchange;
12. Data storage, e.g. memory storage system, distributed file system, data warehouse, etc.;
13. Human consumption of the results of Big Data analysis (e.g. visualization);
14. Energy measurement for Big Data;
15. Interface between relational (SQL) and non-relational (NoSQL) data stores;
16. Big Data Quality and Veracity description and management.

A) ISO/IEC JTC 1/WG 9 Big Data

The ISO/IEC JTC 1/WG 9 is a working group on Big Data created in November 2014. The standardization focal point of WG 9 is to serve as the focus of and proponent for JTC 1's standardization program; to develop foundational standards for Big Data including reference architecture and vocabulary standards for guiding Big Data efforts throughout JTC 1 upon which other standards can be developed. WG 9 will work on the gap analysis to develop other Big Data standards that build on the functional standards when relevant to JTC1 subgroups that would address these standards do not exist or are unable to develop them. Clearly, to identify gaps in Big Data standardization. Regarding the industry engagement and promotion, this WG 9 will develop and maintain liaisons with all relevant JTC 1 entities as well as with any other JTC 1 subgroup that may propose work related to Big Data in the future.

The current WG 9 work program includes the development of the International Standard ISO/IEC 20546 "Big Data – Overview and Vocabulary". This international standard will provide an appropriate definition of Big Data based on foundations from the previous work of different authors and based references (including ISO 704:2009, Terminology work – Principles and methods, ISO 860:2007 Terminology work – Harmonization of concepts and terms, ISO/IEC 17788 Information Technology – Cloud Computing – Overview and Vocabulary). It will gather the main related vocabulary to support the definition of the ISO/IEC 20547 "Big Data – Reference architecture", which specifies the Big Data Reference Architecture and includes the Big Data roles, activities, and functional components and their relationships. The scope of the ISO/IEC 20546 includes Big Data Definition and Taxonomies, Big Data Elements, and Big Data Patterns. The standardization work is not intended to address legal,

[32] http://www.iso.org/iso/big_data_report-jtc1.pdf

regulatory, and cross-border trade discussions. It will help with, and provide a framework for, and better and deeper understanding of, issues involving such topics so that others in the international community can address them.

The ISO/IEC NP 20547 international standard will provide a technical reference to facilitate a common understanding of the concept. The standard gets the common features from the use cases as basis to create the Big Data Reference Architecture with the basic modules and their relationships and is intended to support in creating specific-domain Big Data Architectures, as a way to create proper Big Data solutions for companies worldwide. This standard will describe a generic high-level conceptual model that is an effective tool for discussing the requirements, structures, and operations inherent to Big Data, for example, to illustrate and understand the various Big Data components, processes, and systems, in the context of an overall Big Data conceptual model. The standard will provide a technical reference for government departments, agencies and other consumers to understand, discuss, categorize and compare Big Data solutions. It will also facilitate the analysis of candidate standards for interoperability, portability, reusability, and extendibility.

The standardization work of the ISO/IEC 20547 is composed of five parts:

1) ISO/IEC NP TR 20547-1, Information technology – Big Data Reference Architecture – Part 1: Framework and Application Process; this technical report (TR) will describe the framework of Big Data Reference Architecture and the process for how a user of the standard can apply it to their particular problem domain;

2) ISO/IEC NP TR 20547-2, Information technology – Big Data Reference Architecture – Part 2: Use Cases and Derived Requirements; this technical report would decompose a set of contributed use cases into general Big Data Reference Architecture requirements;

3) ISO/IEC NP 20547-3, Information technology – Big Data Reference Architecture – Part 3: Reference Architecture; this international standard specifies the Big Data Reference Architecture. The Reference Architecture includes the Big Data roles, activities, and functional components and their relationships. Some roles are the Data Provider, Data Consumer, Big Data Application Provider, and Big Data Framework Provider;

4) ISO/IEC NP 20547-4, Information technology – Big Data Reference Architecture – Part 4: Security and Privacy Fabric; this international standard specifies the underlying security and privacy fabric that applies to all aspects of the Big Data Reference Architecture including the Big Data roles, activities, and functional components. The international standard will provide a security and privacy taxonomy and requirements that supplements the general Big Data Reference Architecture;

5) ISO/IEC NP TR 20547-5, Information technology – Big Data Reference Architecture – Part 5: Standards Roadmap; the technical report will document Big Data relevant standards, both in existence and under development, along with priorities for future Big Data standards development based on gap analysis.

B) ISO/IEC JTC 1/SC 32 Data Management and Interchange

The interest area of SC 32[33] are data management and interchange, including database languages, multimedia object management, metadata management, and e-Business. Specifically, SC 32 standards include: reference models and frameworks for the coordination of existing and emerging standards; Definition of data domains, data types and data structures, and their associated semantics; Languages, services and protocols for persisting storage, concurrent access, concurrent update and interchange of data; Methods, languages, services, and protocols to structure, organize, and register metadata and other information resources associated with sharing and interoperability, including electronic commerce. This technical subcommittee currently works in several related areas of Big Data Technology. Some examples:

- SQL is already adding new feature to support Big Data. Moreover, SQL has been supporting bi-temporal data, two forms of semi-structured data (XML and JSON), and multidimensional arrays. SQL implementations are known to exist, which utilize storage engines that are built using several of the NoSQL technologies, including name-value pairs, big table, and document.
- Metadata efforts have focused on two major areas:
    - The specification and standardization of data elements, including the registration of those data elements;
    - The definition of metamodels (to describe data and application models) and definitions of those models themselves.

The SC32 launched in 2013 a study group on Next Generation Analytics and Big Data. The study group have identified[34] some opportunities for standards enhancement:

- Review the metadata standards to ensure the required support exists for Analytical and Big data projects and tools;
- Review the data storage standards to ensure the required support exist for storing and retrieving the volume and diversity of data required by Analytical and Big data projects and tools.

Some standards relevant in the context of Big Data that have been published by SC 32 are: ISO/IEC 9075-* Information Technology – Database languages – SQL. This international standard defines SQL. The scope of SQL is the definition of data structure and the operations on data stored in that structure. ISO/IEC 9075-1:2011 Information Technology – Data languages – SQL – Part 1: Framework (SQL/Framework), ISO/IEC 9075-2:2011 Information technology – Database languages – SQL – Part 2: Foundation (SQL/Foundation), ISO/IEC 9075-11:2011 Information technology – Database languages – SQL – Part 11: Information and Definition Schemas (SQL/Schemata) encompass the minimum requirements of the language. Other parts define extensions.

Other relevant standards related to Big Data from SC 32 are ISO/IEC 11179 and 19763 families[35]. ISO/IEC 11179 specifies a Metadata Registry and associated procedures, and specifies metadata required to describe specific types of metadata item. ISO/IEC 19763 specifies metadata models for registering various types of models and model mappings. The relevance of ISO/IEC 11179 and 19763 in the context of Big Data, Metadata standards help to address variety issues as for example, variety

[33] http://www.iso.org/iso/standards_development/technical_committees/list_of_iso_technical_committees/iso_technical_committee.htm?commid=45342

[34] http://www.jtc1sc32.org/doc/N2351-2400/32N2388b-report_SG_big_data_analytics.pdf

[35] http://metadata-standards.org/Document-library/Documents-by-number/WG2-N1901-N1950/WG2N1943_MetadataRegistries&BigData.pptx

of data represented in different formats, variety between database structures for representing the data for similar universe of discourse, variety different terminology used in the data, variety of data structure: structured, unstructured, semi-structured. Moreover, understanding Big Data still requires an understanding of data elements, concepts, conceptual domains and value domains, though these may be specified after the data has been captured instead of before. The ISO/IEC 11179 and 19763 registries could be extended to capture additional constructs.

C) ISO/IEC JTC 1/SC 38 Cloud Computing and Distributed Platforms

The SC 38[36] is responsible for the development of standards to support distributed computing paradigms – especially in the area of Cloud Computing. With the progression of service oriented architecture specification and the publication of ISO/IEC 17788 and 17789, standards presenting taxonomy, terminology and vocabulary, from the Cloud Computing collaboration with the ITU-T/SG 13, SC 38 is turning its focus to identifying other standardization initiatives in these rapidly developing areas. SC 38 standardization work deals with service-oriented architecture, service level agreement, interoperability and portability and data and their flow across devices and Cloud services. ISO/IEC JTC 1/SC 38 currently works areas related to Big Data Paradigm:

- Cloud Data Management Interfaces. The ISO/IEC 17826:2012 Information technology – Cloud Data Management Interface specifies the interface to access Cloud storage and to manage the data stored therein.
- Open Virtualization Format. The ISO/IEC 17203:2011 Information Technology – Open Virtualization Format (OVF) specification specifies open, secure, portable, efficient and extensible format for the packaging and distribution of software to be run in virtual machines.

D) Other ISO and ISO/IEC Related Technical Committees, Subcommittees and Working Groups

Related ISO/IEC and ISO technical committees, subcommittees and working groups developing standards linked to Big Data and its applications are ISO/IEC JTC 1/SC 29/WG 11 (MPEG) on Big Media, ISO/TC 69 – Applications of Statistical Methods, ISO/TC 204 – Intelligent Transportation, ISO/IEC JTC 1/SC 39 – Sustainability for and by Information Technology. ISO/IEC JTC 1/SC 29/WG 11 (MPEG) supports by identifying and characterizing existing multimedia Big Data deployment, identifying Big Media use cases, and identifying MPEG tools relevant for Big Media. ISO/TC 69 could explore new Big Data statistical methods, identify use cases (e.g. healthcare fraud, live twitter feeds), implement use cases using best practices Big Data computing ecosystem, document findings, and standardize new Big Data statistical methodologies. ISO/TC 204 focuses on information, communication and control systems for transportation. The interest in and motivation for Big Data is at a high level by the notion that each one of the working groups can provide at least one use case. Moreover, each working group has one or more applications that act as a data source for Big Data. The interest by ISO/IEC JTC 1/SC 39 is to examine Big Data Reference Architecture when ready to see if the Key Performance Indicators developed in the context of datacenters and datacenter equipment need amending, enhancement and provide feedback on architecture for inefficiencies. Two applicable international standards related to security in the context of Big Data are the ISO 27001, ISO/IEC 27018 and ISO 29100.

[36] http://www.iso.org/iso/standards_development/technical_committees/list_of_iso_technical_committees/iso_technical_committee.htm?commid=601355

E) ITU-T Study Group 13

The Study Group 13 from ITU-T published the first Big Data related standard[37]: **ITU-T Y.3600 Big Data – Cloud computing based requirements and capabilities.** This standard details the requirements, capabilities and use cases of Cloud-based Big Data as well as high-level system context view and its relationships with other entities. The Big Data paradigm provides an effective, scalable solution to deal with growing volumes of data and uncover patterns or other information capable of making data manageable and profitable. Cloud computing based Big Data provides the capabilities to collect, store, analyze, visualize and manage varieties of large volume datasets, which cannot be rapidly transferred and analyzed using traditional technologies. ITU-T Y.3600 outlines recommendations and requirements for data collection, visualization, analysis and storage, among other areas, along with security considerations. It addresses the following subjects:

- Overview of Big Data:
  - Introduction to Big data;
  - Big Data ecosystems and roles;
  - Relationship between cloud computing and Big Data;
- Cloud Computing based Big Data system context and benefits;
- Cloud Computing based Big Data requirements;
- Cloud Computing based Big Data capabilities.

The recommendation describes the Big Data ecosystem through roles and sub-roles. It also defines necessary activities for roles providing and consuming Big Data services as well as relationships between roles. This Big Data ecosystem includes data provider, Big Data service provider and Big Data service customer.

The ITU-T SG 13 study group is developing a recommendation related to the functional architecture of Big Data as a service. **ITU-T Y.BDasS-arch – Cloud computing – Functional architecture of Big Data as a Service.** This recommendation specifies the functional components, functional architecture, and reference points of Big Data as a Service (BDasS). The scope of this recommendation includes: overview of Big Data as a service functional architecture, the functional components of Big Data as a Service, the functional architecture of Big Data as a Service, and the reference points between functional components of Big Data as a Service.

A Big Data and Internet of Things (IoT) recommendation is under development by the SG 13 **(Y.IoT-BigData-reqts).** The purpose of this recommendation is to specify requirements and capabilities of the IoT for Big Data. This recommendation complements the developments on common requirements of the Internet of Things (ITU-T Y.2066) and functional framework of the IoT (ITU-T Y.2068) in terms of the specific requirements and capabilities that the IoT is expected to support in order to address the challenges related to Big Data. In addition, it constitutes a basis for further standardization work concerning Big Data in the IoT.

F) NIST Big Data Public Working Group for Big Data

The focus of the National Institute of Standards and Technology Big Data Public Working Group (NBD-PWG) is to form a community of interest from industry, academia, and government, with the goal of developing consensus definitions, taxonomies, reference architectures, and technology roadmaps. The aim is to create vendor-neutral, technology and infrastructure agnostic deliverables to enable Big Data stakeholders to pick-and-choose best analytics tools for their processing and visualization requirements on the most suitable computing platforms and clusters while allowing value-added

[37] http://www.itu.int/itu-t/recommendations/rec.aspx?rec=12584

from Big Data service providers and flow of data between the stakeholders in a cohesive and secure manner. NIST[38] is one of the organizations that have issued a number of draft specifications (dated April 23, 2014) as part of a seven-part Big Data Interoperability Framework:

- Volume 1: NIST Big Data Definitions, Version 1.0. The aim of this volume is to provide a common vocabulary for those involved with Big Data. For managers, the terms in this volume will distinguish the concepts needed to understand this changing field. For procurement officers, this document will provide the framework for discussing organizational needs, and distinguishing among offered approaches. For marketers, this document will provide the means to promote solutions and innovations. For the technical community, this volume will provide a common language to better differentiate the specific offerings;

- Volume 2: NIST Big Data Taxonomies, Version 1.0. The NBD-PWG Definitions and Taxonomy Subgroup focused on identifying Big Data concepts, defining terms needed to describe this new paradigm, and defining reference architecture terms. This taxonomy provides a hierarchy of the components of the reference architecture. It is designed to meet the needs of specific user groups, as follows:
    - For managers, the terms will distinguish the categorization of techniques needed to understand this changing field;
    - For procurement officers, it will provide the framework for discussing organizational needs and distinguishing among offered approaches;
    - For marketers, it will provide the means to promote Big Data solutions and innovations;
    - For the technical community, it will provide a common language to better differentiate Big Data's specific offerings;

- Volume 3: NIST Big Data Use Case & Requirements, Version 1.0. This volume was prepared by the NBD-PWG Use Cases and Requirements Subgroup. The effort focused on forming a community of interest from industry, academia, and government, with the goal of developing a consensus list of Big Data requirements across all stakeholders. This included gathering and understanding various use cases from nine diversified areas (i.e., application domains). To achieve this goal, the Subgroup completed the following tasks:
    - Gathered input from all stakeholders regarding Big Data requirements;
    - Analyzed and prioritized a list of challenging use case specific requirements that may delay or prevent adoption of Big Data deployment;
    - Developed a comprehensive list of generalized Big Data requirements;
    - Collaborated with the NBD-PWG Reference Architecture Subgroup to provide input for the NIST Big Data Reference Architecture (NBDRA);
    - Documented the findings in this report.

- Volume 4: NIST Big Data Security and Privacy Requirements, Version 1.0. The focus of the NBD-PWG Security and Privacy Subgroup is to form a community of interest from industry, academia, and government with the goal of developing consensus on a reference architecture to handle security and privacy issues across all stakeholders. This includes understanding what standards are available or under development, as well as identifying which key organizations are working on these standards.

[38] The National Institute of Standards and Technology (NIST) is a U.S. government agency. For the framework, please check: http://bigdatawg.nist.gov/V1_output_docs.php

- The scope of the Subgroup's work includes the following topics, some of which will be addressed in future versions of this Volume:
- Provide a context from which to begin Big Data-specific security and privacy discussions;
- Gather input from all stakeholders regarding security and privacy concerns in Big Data processing, storage, and services;
- Analyze/prioritize a list of challenging security and privacy requirements that may delay or prevent adoption of Big Data deployment;
- Develop a Security and Privacy Reference Architecture that supplements the NIST Big Data Reference Architecture (NBDRA);
- Produce a working draft of this Big Data Security and Privacy document;
- Develop Big Data security and privacy taxonomies;
- Explore mapping between the Big Data security and privacy taxonomies and the NBDRA;
- Explore mapping between the use cases and the NBDRA;
- While there are many issues surrounding Big Data security and privacy, the focus of this Subgroup is on the technology aspects of security and privacy with respect to Big Data.

- Volume 5: NIST Big Data Architectures White Paper Survey, Version 1.0. This volume was prepared by the NBD-PWG Reference Architecture Subgroup. The effort focused on developing an open reference Big Data architecture that achieves the following objectives:
  - Provides a common language for the various stakeholders;
  - Encourages adherence to common standards, specifications, and patterns;
  - Provides consistent methods for implementation of technology to solve similar problem sets;
  - Illustrates and improve understanding of the various Big Data components, processes, and systems, in the context of vendor- and technology-agnostic Big Data conceptual model;
  - Provides a technical reference for U.S. government departments, agencies, and other consumers to understand, discuss, categorize, and compare Big Data solutions;
  - Facilitates the analysis of candidate standards for interoperability, portability, reusability, and extendibility.

  The reference architecture is intended to facilitate the understanding of the operational intricacies in Big Data. It does not represent the system architecture of a specific Big Data system, but rather is a tool for describing, discussing, and developing system-specific architectures using a common framework. The reference architecture achieves this by providing a generic, high-level conceptual model, which serves as an effective tool for discussing the requirements, structures, and operations inherent to Big Data. The model is not tied to any specific vendor products, services, or reference implementation, nor does it define prescriptive solutions for advancing innovation.

- Volume 6: NIST Big Data Reference Architecture, Version 1.0. The goal of the NBD-PWG Reference Architecture Subgroup is to develop an open reference architecture for Big Data that achieves the following objectives:
  - Provides a common language for the various stakeholders;
  - Encourages adherence to common standards, specifications, and patterns;
  - Provides consistent methods for implementation of technology to solve similar problem sets;

- Illustrates and improves understanding of the various Big Data components, processes, and systems, in the context of a vendor- and technology- agnostic Big Data conceptual model;
- Provides a technical reference for U.S. government departments, agencies, and other consumers to understand, discuss, categorize, and compare Big Data solutions;
- Facilitates analysis of candidate standards for interoperability, portability, reusability, and extendibility.

- Volume 7: NIST Big Data Technology Roadmap, Version 1.0. The NBD-PWG Technology Roadmap Subgroup focused on forming a community of interest from industry, academia, and government, with the goal of developing a consensus vision with recommendations on how Big Data should move forward. The Subgroup's approach was to perform a gap analysis through the materials gathered from all other subgroups. This included setting standardization and adoption priorities through an understanding of what standards are available or under development as part of the recommendations. The goals of the Subgroup will be realized throughout the three-planed phases of the NBD-PWG work, as outlined in Section 1.1. The primary tasks of the NBD-PWG Technology Roadmap Subgroup include the following:
    - Gather input from NBD-PWG subgroups and study the taxonomies for the actors' roles and responsibility, use cases and general requirements, and secure reference architecture;
    - Gain understanding of what standards are available or under development for Big Data;
    - Perform a gap analysis and document the findings;
    - Identify what possible barriers may delay or prevent adoption of Big Data;
    - Document vision and recommendations.

4.2 Big Data Standardization from Fora and Consortia

Fora and Consortia, in the standardization context, are associations regrouping individuals, companies, organizations or governments with a common objective of participating in the creation of *de facto* standards or technical specifications. Many of them are involved in the Big Data standardization domain.

A) IEEE Standards Association

The Institute of Electrical and Electronics Engineers (IEEE)[39] standards association has introduced a number of Big Data applications related standards[40].

The **IEEE 2200-2012 Standard Protocol for Stream Management in Media Client Devices** defines the interfaces for intelligently distributing and replicating content over heterogeneous networks to portable and intermediate devices with local storage.

The **IEEE 42010-2011 ISO/IEC/IEEE Systems and Software Engineering – Architecture Description** addresses the creation, analysis, and maintenance of system architectures through the use of descriptions. The contents of an architecture description are specified, as well as architecture viewpoints, frameworks, and description languages for codifying conventions and common practices.

[39] https://www.ieee.org/index.html
[40] http://theinstitute.ieee.org/benefits/standards/standards-that-support-big-data

B) W3C World Wide Web Consortium

The World Wide Web Consortium (W3C) is an international community that develops open standards to ensure long-term growth of the Web. Given that one of the primary contributors to the growth of Big Data has been the growth of the Internet and World Wide Web, many of the developing standards around web technologies must deal with the challenges inherent in Big Data. Some examples of W3C standard efforts related to Big Data technologies and interest are:

- Model for Tabular Data and Metadata on the Web;
- Delivery Context Ontology;
- Efficient XML Interchange;
- Linked Data;
- Relational Database to resource description framework;
- Service Modelling Language;
- Sparse Query Language;
- Semantic Web standards;

C) Open Geospatial Consortium (OGC)

The OGC[41] consortium is an international industry consortium of companies, government agencies, research institutes and universities participating to develop standards for interoperable "geo-enable" solutions on the Web, wireless and location-based services. Geospatial data represents a common Big Data problem due to the volume and number of records of the data involved. OCG established Big Data Domain Working Group in 2014 aiming to clarify:

- Foundational terminologies in the context of data analytics understanding differences/overlaps with terms like data analysis, data mining, etc.;
- A systematic classification of analysis algorithms, analytics tools, data and resource characteristics, and scientific queries.

The following examples of OGC standard efforts relate to Big Data technologies and interfaces:

- Data Model Extension standards;
- Registry services;
- Metadata profiles;
- GeoAPI implementation standard (joint effort with ISO Technical Committee 211).

D) Organization for the Advancement of Structured Information Standards (OASIS)

OASIS [42] promotes industry consensus and produces worldwide standards for security, Cloud Computing, Web services, service oriented architecture, and other areas. Some OASIS technical committees and activities relevant to Big Data are:

- OASIS Advanced Message Queuing Protocol technical committee: defining a ubiquitous, secure, reliable and open internet protocol for handling business messaging;
- OASIS Key-Value Database Application Interface technical committee: defining an open application programming interface for managing and accessing data from database systems based on a key-value model;
- Cross-enterprise security and privacy authorization.

[41] http://www.opengeospatial.org/docs/is
[42] https://www.oasis-open.org/

Although these current initiatives to create Big Data related standards are established, the activities are at an early stage.

5. Conclusion and Outlook

The bigness of Big Data is growing faster than many businesses can cope with. In fact, Big Data is less about "data which is big" than it is about a capability to aggregate, search, cross-reference and visualize large data sets. In this context, the opportunities for business, academia and governments are manifold: they range from customer service and operational improvements, productivity and sales increases, logistic chains optimizations, environmental impact and cost reductions, better fraud detection and prevention, up to better scientific modeling for natural disasters and social unrest, health epidemics projections and revealing hidden correlations to treat rare diseases.

The Big Data trend is being seen by many businesses as a way to gain advantage over their competitors: if one is able to make sense of the information contained in the data reasonably faster than a competitor is, it will be able to obtain an advantage in the market. Therefore, since 2011 many businesses are experimenting and implementing Big Data capabilities. For an organization to get the greatest benefit from a Big Data capability, the data must be validated first in order to ensure that the quality of the information fits its purpose. Then, data must be turned into understandable consumable information and integrated with the intuitive knowledge of the operators to create valuable information.

In order to achieve previous purposes, most organizations will employ hybrid solution architectures, where structured operational data remains in existing relational data warehouses, while semi-structured and unstructured data are managed in new database forms (e.g. NoSQL databases).

Hurwitz et al. (2013) specify ten Big Data Do's and Don'ts (Table 12) of how an organization should embark on the Big Data journey.

1. Do Involve All Business Units in Your Big Data Strategy	Business Units can gain significant value when they are brought into the process early.
2. Do Evaluate All Delivery Models for Big Data	Evaluate the type of services that are Cloud based and determine which ones have the performance that you will need for certain tasks.
3. Do Think about Your Traditional Data Sources as Part of Your Big Data journey	Plan to use the results of Big Data Analytics in conjunction with the traditional data warehouse. This allows comparing the Big Data results against the benchmarks of the enterprise data, which is critical for decision making.
4. Do Plan for Consistent Metadata	Before results from Big Data results can be trusted make sure there is a consistent set of metadata. Only then, it can be analyzed in concert with the enterprise data from the systems of record.
5. Do Distribute Your Data	When dealing with Big Data, it will be impractical to manage it on a single server. Make use of distributed computing techniques (e.g. Hadoop) to effectively manage size, variety, velocity and veracity.

6. Don't Rely on a Single Approach to Big Data Analytics	A variety of analytic approaches are available, therefore, spend time to experiment and investigate the variety of technologies that will meet the organizations requirements best.
7. Don't Go Big Before You Are Ready	Start with pilot projects to gain some experience and work with experts can prevent common mistakes because of inexperience.
8. Don't Overlook the Need to Integrate Data	Big Data sources will not be effective if they live in isolation from each other. Be prepared not just to analyze but also to integrate those.
9. Don't Forget to Manage Data Securely	Make sure to have the same level of data security for Big Data that is maintained in the traditional data management environments. This includes compliance, regulatory and privacy requirements so it does not put the organization at risk.
10. Don't Overlook the Need to Manage the Performance of Your Data	Big Data needs to be managed effectively as any other type of data. Therefore, manageability of data has to be part of the Big Data journey.

Table 12 10 Big Data Do's and Don'ts (Hurwitz et al., 2013)

As stated earlier, there are many positives sides in Big Data technologies for individuals, companies, governments, academia, and organizations in general. However, the excessive use of surveillance by governments and intelligence agencies, search queries that are collated and analyzed by big Internet companies to create customer profiles even if they are not signed into search engines, and the mass-assembly conducted by retailers of publicly available metadata (that is also to be considered as potential private data) can put a strain on its further development. This implies that the inadequate and outdated data protection regulations need to be modernized as a matter of urgency.

Overall, there is no doubt that Big Data technologies are still requiring further developments. Therefore, more capital investments from governments and commercial enterprises should be made to further develop the practice and science of Big Data. And certainly not least of all, by setting new standards from both a technology as well as a privacy perspective. Given the potential of Big Data counterbalanced by a lack of best practices and both technical and cultural challenges, it is necessary to provide a new guidance to the adoption of the technology (British Standards Institution, 2016). This is the role of standards, developed in collaboration with the actors of the market and in the goal to support business.

Standards are essential not only to develop ICT technologies, but also to support their interoperability with other technologies and in other economic sectors. Moreover, standards contribute to promote and share good practices and techniques available in the ICT sector. They ensure the quality and performance of products, systems and services. They also facilitate dialogue and exchange between various stakeholders. In this sense, standardization represents an important economic lever to improve business productivity. In a nutshell, standards play a key role by facilitating trades and guaranteeing some fundamental characteristics such as interoperability, quality, security and risk management.

Technical standardization plays an important role not only giving a first-hand insight into latest developments, thus supporting innovation, but also contributing to harmonization of systems and procedures, opening access to external markets and ensuring constant progress.

Literature

Accenture (2014). *Big Success with Big Data*.

Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A., & Buyya, R. (2014). *Big Data computing and clouds: Trends and future directions*. Journal of Parallel and Distributed Computing.

Boyd, D., & Crawford, K. (2012). *Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon.* Information, communication & society, 15(5), 662-679.

British Standards Institution (2016) report on *Big Data and standards market research*.

Chen, H., Chiang, R. H. L., Storey, V. C. (2012). *Business intelligence and analytics: Form big data to big impact.* MIS Quarterly, 36(4), 1165–1188.

Chen, C. P., Zhang, C. Y. (2014). *Data-intensive applications, challenges, techniques and technologies: A survey on Big Data.* Information Sciences, 275, 314-347.

Chen, M., Mao, S., Liu, Y (2014). *Big Data: a Survey*. In Mobile Networks and Applications April 2014, Volume 19, Issue 2, pp 171-209.

Chang, R. M., Kauffman, R. J., & Kwon, Y. (2014). *Understanding the paradigm shift to computational social science in the presence of big data.* Decision Support Systems, 63, 67-80.

Cox, M., Ellsworth, D (1997). *Managing big data for scientific visualization.* ACM Siggraph, NASA Ames Research Center.

DMBOK (2013). *DAMA-DMBOK Guide Version 2 (Draft), the DAMA Guide to the Data Management Body of Knowledge.* DAMA International, USA.

Davenport, T. H., Barth, P., & Bean, R. (2013). *How 'big data' is different.* MIT Sloan Management Review, 54(1).

Dapp, T.F., Heine, V. (2014). *Big data. The untamed force*. Deutsche Bank Research, Frankfurt am Main, Germany

Deodhar, G., (2015). *From Hype to Action: Getting What's Needed from Big Data Analytics.* NTT Innovation Institute Inc.

ENISA (2015) report on *Big Data Security: Good Practices and Recommendations on the Security of Big Data Systems.*

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., Sugimoto, C. R. (2015). *Big data, bigger dilemmas: A critical review.* Journal of the Association for Information Science and Technology.

Fernández, A., del Río, S., López, V., Bawakid, A., del Jesus, M. J., Benítez, J. M., & Herrera, F. (2014). *Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4(5), 380-409.

Gandomi, A., Haider, M, (2015). *Beyond the hype: Big data concepts, methods, and analytics.* International Journal of Information Management, 35(2), 137-144.

Gobble, M. M. (2013). Big data: The next big thing in innovation. *Research-Technology Management*, *56*(1), 64.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., Khan, S. U. (2015). *The rise of "big data" on cloud computing: Review and open research issues.* Information Systems, 47, 98-115.

Hilbert M. (2015). *Big Data for Development: A Review of Promises and Challenges.* In Development Policy Review Volume 34, Issue 1, pages 135–174, January 2016. DOI: 10.1111/dpr.12142

Holzinger, A., Stocker, C., Ofner, B., Prohaska, G., Brabenetz, A., Hofmann-Wellenhof, R. (2013). *Combining HCI, Natural Language Processing, and Knowledge Discovery - Potential of IBM Content Analytics as an Assistive Technology in the Biomedical Field*. In Holzinger, Andreas; Pasi, Gabriella. Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data. Lecture Notes in Computer Science. Springer. pp. 13–24.

Hurwitz, J., Nugent, A., Hapler, F., Kaufman, M. (2013). *Big Data for Dummies.* John Wiley & Sons, Hoboken, NY.

IBM (2012). *Analytics: The real-world use of big data How innovative enterprises extract value from uncertain data.* IBM Institute for Business Value

ISO/IEC JTC 1 (2014). *Big data, Preliminary Report 2014.* International Organization for Standardization, Geneva, Switzerland.

Kambatla, K., Kollias, G., Kumar, V., Grama, A. (2014). *Trends in big data analytics.* Journal of Parallel and Distributed Computing, 74(7), 2561-2573.

Kraska, T. (2013). *Finding the Needle in the Big Data Systems Haystack.* IEEE Internet Computing, vol. 17, no. 1, pp.84-86, 2013.

Kshetri, N. (2014). *Big data's impact on privacy, security and consumer welfare*. Telecommunications Policy, 38(11), 1134-1145.

Kosinski, M., Stillwell, D., Graepel, T. (2013). *Private traits and attributes are predictable from digital records of human behavior.* Proceedings of the National Academy of Sciences, 110(15), 5802-5805.

Kwon, O., Lee, N., & Shin, B. (2014). *Data quality management, data usage experience and acquisition intention of big data analytics.* International Journal of Information Management, 34(3), 387-394.

Laney, D. (2001). *3D Data management: controlling data volume, velocity and variety*. Application Delivery Strategies, Meta Group (6 Feb 01.949 Addendum).

Manyika, J., Chui, M., Brown, B., Bughin J., Dobbs R., Roxburgh C., Byers A.H. (2011). *Big data: the next frontier for innovation, competition, and productivity.* McKinsey Global Institute Report, New York, NY.

Mayer-Schönberger, V., Cukier, K. (2013). *Big Data: A revolution that will transform how we live, work and think.* Boston: Houghton Mifflin Harcourt.

McAfee, A., & Brynjolfsson, E. (2012). *Big data: the management revolution.* Harvard business review, (90), 60-6.

MongoDB White Paper (2015). *Big Data: Examples and Guidelines for the Enterprise Decision Maker.* MongoDB White Paper.

MongoDB Big Data Explained (2016). https://www.mongodb.com/big-data-explained (last accessed April 2016).

Pääkkönen, P., & Pakkala, D. (2015). *Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems.* Big Data Research.

Perrey, J., Spillecke, D., Umblijs, A. (2013). *Smart analytics: How marketing drives short-term and long-term growth.* McKinsey Quarterly.

Sagiroglu, S., Sinanc, D. (2013). *Big data: A review", In Collaboration Technologies and Systems (CTS),* 2013 International Conference on (pp. 42-47). IEEE.

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., Tufano, P. (2012). *Analytics: The real-world use of big data. How innovative enterprises extract value from uncertain data.* IBM Institute for Business Value.

Shah, S., Horne, A., & Capellá, J. (2012). *Good data won't guarantee good decisions.* Harvard Business Review, 90(4), 23-25.

Tambe, P. (2014). *Big data investment, skills, and firm value.* Management Science, 60(6), 1452-1469.

Taylor, L., Schroeder, R., Meyer, E. (2014). *Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same?* Big Data & Society, 1(2).

Thompson, K., Kadiyala, R. (2014). *Leveraging Big Data to Improve Water System Operations.* Procedia Engineering, 89, 467-472.

Villas-Boas, S. B. (2014). *Big Data in Firms and Economic Research.* Applied Economics and Finance, 1(1), 65-70.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). *How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study.* International Journal of Production Economics.

Wikibon (2015). *Executive Summary 2015.* http://wikibon.com/executive-summary-big-data-vendor-revenue-and-market-forecast-2011-2026/ (last accessed April 2016).

Wrobel, S. (2012). *Big Data – Vorsprung durch Wissen.* Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Sankt Augustin, Germany

Zikopoulos, P.C., de Roos D., Parasuraman K., Deutsch T, Corrigan, D., Giles, J. (2013). *Harness the Power of Big Data: the IBM Big Data Platform.* McGraw Hill.

ILNAS

Institut luxembourgeois de la normalisation,
de l'accréditation, de la sécurité et qualité
des produits et services

ANEC

AGENCE POUR LA NORMALISATION ET
L'ÉCONOMIE DE LA CONNAISSANCE

CONTACT:
ILNAS / ANEC
Southlane Tower I
1, avenue du Swing
L-4367 Belvaux
Phone: (+352) 24 77 43 70
Email: anec@ilnas.etat.lu
www.portail-qualite.lu