Received 4 April 2016
Revised 4 April 2016
Accepted 5 April 2016

Big data projects: just jump right in!

Hajar Mousannif
Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakech, Morocco, and
Hasna Sabah, Yasmina Douiji and Younes Oulad Sayad
Faculty of Sciences and Techniques Gueliz, Cadi Ayyad University, Marrakech, Morocco

Abstract

Purpose – This paper aims to provide a roadmap for organizations to build big data projects and reap the most rewards out of their data. It covers all aspects of big data project implementation, from data collection to final project evaluation.

Design/methodology/approach – In each stage of the proposed roadmap, we introduce different sets of information and communications technology platforms and tools to assist IT professionals and managers in gaining a comprehensive understanding of the methods and technologies involved and in making the best use of them. The authors also complete the picture by illustrating the process through different real-world big data projects implementations.

Findings – By adopting the proposed roadmap, companies and organizations willing to establish an effective and rewarding big data solution can tackle all implementation challenges in each stage of their big data project setup: from strategy elaboration to final project evaluation. Their expectations of privacy and security are also baked, in advance, into the big data project design.

Originality/value – While technologies to build and run big data projects have started to mature and proliferate over the last couple of years, exploiting all potentials of big data is still at a relatively early stage. The value of this paper consists in providing a clear and systematic methodology to move businesses and organizations from an opinion-operated era where humans' skills are a necessity to a data-driven and smart era where big data analytics plays a major role in discovering unexpected insights in the oceans of data routinely generated or collected.

Keywords Business intelligence, Big data, Advanced analytics, Big data project, Big data technologies, Methodologies and tools

Paper type Research paper

1. Introduction

Every time we visit a website, "like" or "follow" a social page and share our experiences, thoughts, feelings and opinions on the internet, we make already "big" data even "bigger"! Every day we collectively generate mountains of data that is waiting to be processed and analyzed. As an example, almost 500 terabytes of data are uploaded each day to Facebook servers (Kern, 2012), while Youtubers upload 100 hours of video every minute (Youtube.com, 2016), and over 571 new websites are created every minute of the day (Wikibon Blog, 2016). Yet, using big data is not about collecting or generating massive amounts of data, but more about making sense of it. In fact, big data are absolutely worthless if it is not actionable. Hence, what companies and organizations gather about customers, suppliers, transactions and such may be of no use if no insights are smartly and timely extracted from it.

International Journal of Pervasive Computing and Communications
Vol. 12 No. 2, 2016
pp. 260-288

1742-7371
DOI 10.1108/IJPCC-04-2016-0023

How to establish an effective and rewarding big data solution is the major concern of any company or organization willing to embark on the big data adventure (Mousannif *et al.*, 2014). Throughout this paper, we will show how business leaders and directors can leverage their data in the most efficient way possible through a clear methodology where descriptive, inquisitive, predictive and prescriptive analytics enter into action to improve results, support mission-critical applications and drive better decision-making. While doing so, this paper attempts to provide satisfying answers to the following fundamental questions:

RQ1. Where does big data come from?

RQ2. What is/are the appropriate system(s) to capture, cure, store, explore, share, transfer, analyze and visualize data?

RQ3. What is the size range of data and its implications in term of storage and retrieval?

RQ4. How could big data be used to determine market opportunities and seize them? And how could it contribute in making forecasts?

RQ5. And finally, how to take into account people's expectations of privacy and bake it in advance into the big data project design?

The remainder of this paper will be organized as follows: Section 2 introduces some existing methodologies for implementing big data projects in today's enterprise, and highlights our contribution. In Section 3, we describe a clear roadmap for building smart and effective big data projects within organizations and illustrate the stages of the process through different sets of platforms and tools, as well as real-world big data projects use cases. Conclusions are given in Section 4.

2. Related work

A recent survey of Gartner showed that companies are now more aware of the opportunities offered by analyzing larger amounts of data and are increasingly investing or planning to invest in big data projects, from 58 per cent in 2012 to 64 per cent last year (Gartner.com, 2016). This trend is accompanied by an increase in the need of a global model or roadmap to assist IT departments not only in implementing a big data project but also in making the best use of it to meet business objectives. The Gartner approach in (Sicular, 2013) introduces a roadmap to succeed big data solutions adoption, starting from the stage of company unawareness of the necessity of big data in facing today's business objectives, to the final stage of data-driven enterprise.

The other example is that of the US Census Bureau, which has implemented a big data project to conduct a head count of all the people in the USA. The life cycle of the US census bureau big data project includes three fundamental steps: data collection using a multi-mode model, data analysis to explore technology solutions based on methodological techniques and data dissemination by implementing new platforms for integrating census and survey data with other big data (Bostic, 2013). In King (2013), ASE consulting provides a well-considered approach for building big data projects and which consists of six steps that are not committing to any particular technology or tool, ranging from understanding the scope of the project by identifying business problems and opportunities, to evaluating the big data project while providing insights into what worked well and what did not. IBM in their recent report (Desouza, 2014) introduced a three-phase approach for building big data projects, namely, planning, execution

and post-implementation, and which mainly consists in understanding the business and legal policies, communication between IT departments and the project stakeholders and conducting an impact analysis at the end of the implementation.

With respect to all related literature presented above, the existing efforts either fail to cover some fundamental aspects of big data project setup, or limit their approach to providing basic guidelines for big data projects implementation without further insights into the technologies and platforms involved. The present work comes to overcome such limitations by:

- providing a holistic approach to building big data projects, which tackles all implementation challenges a company or organization may face in each stage of their big data project setup – from strategy elaboration to final project evaluation;
- assisting companies and organizations, willing to establish an effective and rewarding big data solution, in gaining a comprehensive understanding of the technologies involved and in making the best use of them;
- baking in advance people's expectations of privacy and security into the big data project design; and
- illustrating the proposed process through different real-world big data projects implementations.

3. Roadmap for building smart big data projects

In this section, we explore the design of the proposed methodology and provide a set of useful tips to follow in each phase of the big data project setup. The suggested approach, as shown in Figure 1, consists of three major phases: elaboration of the global strategy, implementation of the project and post-implementation.

3.1 Global strategy elaboration

Starting a big data project requires different changes and new investments. The changes particularly include the establishment of a new technological infrastructure and a new way to process and harness data. Here are a few points (Figure 2) to consider before undertaking any changes:

3.1.1 Why a big data project? To answer this question, companies have to find the problems that need a solution and decide whether they could be solved using new technologies or just with available software and techniques. Those problems could be volume challenges, real time analytics, predictive analytics or customer-centric analytics, among others. In fact, the trigger point for big data adoption may vary from one organization to another. In the case of Tynt (2014), it was the rate of change of data, as reported by Cameron Befus, a former Vice President Engineering at Tynt. The volume of their data was not very important, but it was hard to deal with the growing rate of change using their available technology (Villars *et al.*, 2011), which drove them to switch to a big data technology that is Hadoop.

Defining business priorities is the second aspect to consider, by determining first the most important activities that form the greatest economic leverage in the business, and identifying what actions have got to be changed or improved. The key activities that the company has to focus on depend on the business itself. Here are a few use cases:

- *Services companies:* One of the top priorities is the improvement of customer centricity, to increase the customers' base and ensure their loyalty. British

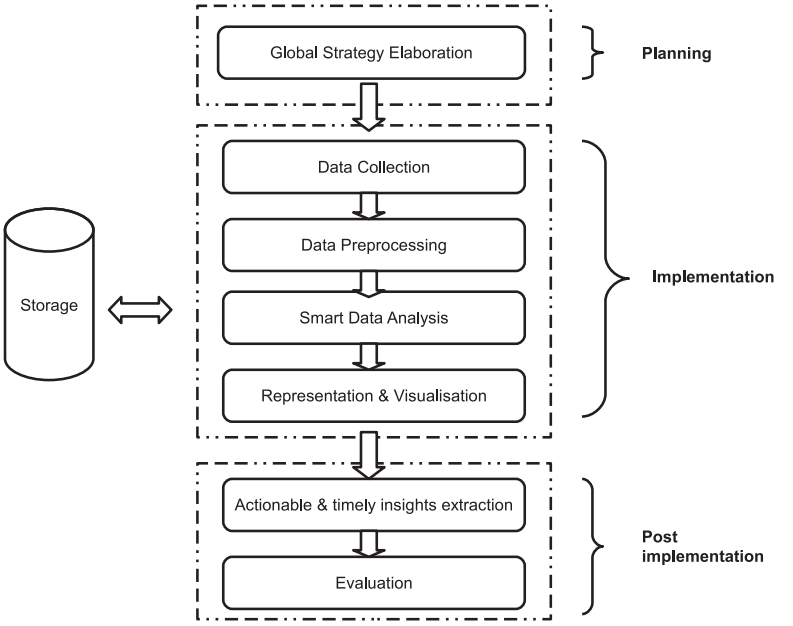

Figure 1.
Big data project workflow

Figure 2.
Global strategy elaboration

Airways (BA) (Doug, 2013) is one example of big data practitioners who got a better understanding of their customers thanks to a program called "Know Me", which mixed loyalty data with other data generated by the online behavior and purchasing habits of 20 million BA customers.

- *Manufacturers:* Manufacturing and products managers believe that the best way to use big data in this context is to locate product defects, support quality and enhance supply planning (TATA Consultancy Services, 2014).

- *Retailers:* Must take advantage of big data in getting a comprehensive understanding of the market, products, customers, competitors and locations distribution.

3.1.2 What data should the organization consider? Once priorities are defined, business leaders and IT practitioners must target the data that will yield most value, to zero in on the right technology investment that should be made; they could first start by evaluating data stores that have been prepared for analysis and think about how they could be extended or improved, they also have to define which ones of unstructured datasets should be converted to workable format. This important phase is defined by IBM as Data exploration (IBM Software, 2013), as it is about exploring both internal and external data available to the company, to ensure that it can be accessed to sustain decision-making and everyday operations. IBM InfoSphere Data Explorer (IBM, 2014) is one example of tools that allows performing such a task. It allows federated discovery, navigation and inquiry over a wide extend of data sources and types, both inside and outside the organization, to help companies start big data initiatives.

As an example, the marketing services company Tivo Research Analytics (Doug, 2013) found the right mix of data to analyze by combining supermarket transaction information with television-viewing data, to provide advertisers and agencies with deep insights into buying behavior.

3.1.3 Where to store and process data? Companies are increasingly adopting cloud-based big data solutions. In 2013, a survey conducted by Bridge Venture Partners (ATTUNITY, 2013) on more than 850 business employees and IT decision makers showed that 75 per cent of participants acknowledged using at least one cloud-based platform comparing to 67 per cent in 2012. Yet, outsourcing storage or processing to a cloud provider might not be the right solution in all situations. In fact, companies are different in terms of size, existing IT infrastructure, budget, business objectives and technical abilities. Hence, choosing in-house, cloud-based or hybrid solution must take into account considerations such as cost, technical requirements, project longetivity and security constraints among others.

3.1.3.1 In-house. Setting up the project completely on premise allows full customization and control over data, processing and security. This is particularly important when dealing with sensitive data, when fast data access or extended integration with existing infrastructure is needed.

This option is the most demanding regarding budget because of potential hardware acquisition, deployment and maintenance costs. Changing demands that need high elasticity might be costly as well. Finally, carrying out needed data processes requires technical expertise, which requires in turn new hiring/training/consulting if the existing staff are not qualified enough. As a consequence, the overall project setup is also time-consuming.

Opting for in-house is relevant when big data cross the heart of business, which would justify the needed investments. It is also a suitable option for large size enterprises with a significant budget and an in-house team of big data experts. Good examples of large in-house solutions are loyalty programs of retailers (e.g. Tesco), tailored marketing (e.g. Amazon) or vendor-managed inventories (e.g. Walmart).

3.1.3.2 Public cloud. As advanced technologies, skills and multiple data services are needed for a big data project, the cloud option is at first glance a cost-effective alternative

that reduces much of the in-house overhead. Thanks to fast deployment and better elasticity, companies can focus more on the value of data analysis instead of data management. This option also makes sense when large volumes of data are already located in the cloud, making it expensive to move to internal network. Moreover, the cloud solutions in general allow better collaboration and extended access to business application by supporting the mobile option (Intel IT Center, 2013).

Main drawbacks include limited customization, overall control and potentially long-term costs that may exceed the case of in-house option. Security in particular can pose serious problems regarding availability and confidentiality.

Thus, the following points should be carefully considered when choosing the cloud provider (Network Computing, 2010):

- What is the provided control over data and processes?
- What is the adopted security policy, to ensure availability, access control, confidentiality and accountability?
- What security audits are conducted? What are their results?
- What is the legal framework that regulates data transfers and disclosure? Companies should be aware that they may be responsible, and not cloud providers, for data breaches (Kshetri, 2010).

The cloud option is beneficial both for small and midmarket business with small or no IT infrastructure, and for larger corporations looking to lower operating costs. It is also the best fit for short-term projects that need high elasticity.

3.1.3.3 The hybrid approach. In the hybrid solution, the on-demand cloud resources are used to complement in-house deployments, thus combining the two delivery models. One typical use case is to ensure the security of sensitive corporate data by storing and processing it on-premise, while non sensitive data or data without personally identified information can be processed in the public cloud. For example, financial applications involving sensitive data would be deployed rather in-house, while those improving collaboration and productivity could be bequeathed to a cloud provider. Companies can also gain advantage from on-demand storage space and computing owing to cloud services, which is adequate for short-term projects or occasional workload increases. Concerning the project setup and in addition to all considerations relevant to the cloud option, working within the hybrid approach needs efforts for data integration and security policies maintenance (TERADATA, 2013).

Table I summarizes the pros/cons of each type of deployment.

3.1.4 How to protect data? Securing a huge amount of continuously evolving data can be very complicated, considering that firms' servers cannot store all the needed data. Moreover, the fact that big data are most of the time processed in real-time induces even more security challenges. The Cloud Security Alliance, is one of the few organizations that took care of this issue by providing, in their recent report (Cloud Security Alliance, 2013), a set of solutions and methods to win every privacy or security challenge. Similarly, the Enterprise Strategy Group (ESG) shows in Olstik (2013) the most significant obstacles facing the implementation of a security policy in a big data environment, and gives valuable tips for Chief Information Officers to enter the big data security analytics era, whereby companies would not only be able to monitor the traffic

Table I.
Pros and cons of in-house, cloud-based and hybrid deployment types

	In-house	Cloud-based	Hybrid cloud
When	Need for customization Storage of sensitive data Need to integrate the new solution with existing software Have the ultimate control on data and processes The presence of an in-house team	Enterprise application already hosted in the cloud High-volume external data sources that require considerable pre-processing Cost-effective, less overhead, fast deployment Need for a full collaboration in real-time	Need for a greater capacity and efficiency High-performance computing needs Need of an easily integration of next generation application platforms
Pros	It offers the most security and control of the data Data storage is instantly accessible Better knowledge of the business	Data storage is flexible and scalable Cloud storage is up to the demands of big data No need of a professional team Fast results Appropriate for smaller companies	Assist organizations in addressing security concerns in their private cloud Offers tremendous potential for business innovation Allows using the existing infrastructure More control over security
Cons	Hard limits on storage capacity Too expensive No elasticity Deploying Hadoop on-premise is difficult No flexibility/scalability on Data storage	Less secure and less control over data and processes Limited customization Problems with language Product quality control	The private resources are no longer large enough to handle demands Infrastructural dependency Complex networking

coming into and getting out of their systems to detect threats but also predict cyber-attacks even before they happen.

We consider that there are three main features to focus on when planning to implement a new security management solution:

(1) *Protection of sensitive data*: By controlling the access to the data or by providing encryption solutions or both. The chosen solution for this purpose must be easy to integrate within the current system and consider performance issues. In the case of cloud-based big data solutions, the company should discuss with different cloud vendors, their adopted policy for preserving data security and privacy, and for managing security incidents, and then make a decision based on its own requirements of security and privacy.

(2) *Network security management*: By monitoring the local network, analyzing data coming from security devices and network end-points, timely detecting intrusions and suspicious traffic and reacting to it, without impeding the main objective of the big data project.

(3) *Security intelligence*: By providing actionable and comprehensive insight that reduces risk and operational effort for any size organization using data generated by users, applications and infrastructure. Before implementing a security intelligence extension, companies must consider the following questions (IBM Software, 2013):

- Do we need to enrich our security or intelligence system with real-time data of unused or underleveraged data sources (videos, email, call data records, etc.)?
- Do we need to sub-second detection, identification and resolution of physical or online threats?
- Do we need to follow criminals, terrorists or people's activities on a watch-list?
- Do we need to make any big data forensics, or look for associations and patterns in the data we have?
- Do we need to enhance security information and event management systems with unstructured data to improve cyber-threat detection and remediation?

Table II classifies some security platforms according to the provided above features.

In addition to the points mentioned above, there are various myths surrounding the concept of "big data" as shown in the Gartner paper (Sicular, 2013) that companies willing to embark on projects should be aware of and not fall victim to, below are some facts to consider before undertaking any change:

- Technology is not the goal of a big data project; it is rather a mean to be seriously thought about once business objectives are set.
- There is no ever-lasting technological solution for implementing the whole cycle of a big data project. As big data solutions proliferate, it becomes difficult to predict which platforms, applications or methods will better work in the future. Hence, companies should stay open to any new big data solution.
- Avoid the warehouse-or-Hadoop trick, it is imperative to use both of them, as they work well alongside and complement each other.

Table II.
Security platforms classification

Applications	Protection of sensitive data	Network security management	Security intelligence	Additional features
Zettaset's Security Data Warehouse (SDW) (Business Wire, 2011)		✓	✓	Mine security information from website traffic, business processes and other day-to-day transactions
Vortmetric Encryption (Vortmetric, 2014)	✓		✓	Data encryption in addition to a firewall
LogRhythm security information and event management 2.0 (LogRythm, 2014)		✓	✓	Uniquely combines enterprise-class security information and event management, log management, file integrity monitoring and machine analytics, with host and network forensics
Fortscale (2014)		✓	✓	Hadoop cluster that can be integrated with various big data repositories or security information and event management systems
RSA (EMC^2, 2014)		✓	✓	A toolbox for cyber analysts
InfoSphere Streams and InfoSphere Big Insight (IBM Software, 2013)	✓	✓	✓	Use big data to help customers narrow down the data to the incidents and offenses they need to address
Voltage (2013)	✓			Secure sensitive data *entering* Hadoop, then control access. Assure global regulatory compliance. Optimize performance and scalability

3.2 Data collection

Data collection comes as the first step of the implementation process, and as data-intense technologies are widely used nowadays, big data sources are pervasive and extremely diverse. While human-generated data represent an important part today, machine-generated data are growing at an incredible rate, driven by Machine to Machine communication. The mobile data alone are expected to reach a volume of 130 exa-bytes per year by 2016 (Cisco, 2013). This section will shed light on some major sources including internet of things (IoT), open data, social media and crowd-sourcing.

3.2.1 Internet of things. Sensors are a major source of big data. They are increasingly deployed everywhere: smart phones and other daily life devices, commercial buildings or transportation systems. With an expected population of 1 trillion by 2015 (Khanzode, 2012), sensors allow collecting various types of data including body related metrics, location, movement, temperature and sounds. Coupled with ubiquitous wireless networks (Mousannif *et al.*, 2011), sensors are driving myriad of smart innovations in the context of IoT, for example, smart buildings where lightening and air-conditioning are optimized, smart transportation and traffic management system that monitor both vehicular and pedestrian traffic for better flow and better evacuation in emergencies (Mousannif *et al.*, 2012) and smart phones that automatically recognize our emotional states and appropriately respond to them (Mousannif and Khalil, 2014).

3.2.2 Open data. Public institutions, organizations and a growing number of private companies are making some of their data sets available for public. A major contributor to open data initiative is the WorldBank (2014a). The catalog includes macro, financial and sector databases. In addition to searching data sets on the World Bank site and downloading tables, users can also access the data via different application programming interface (WorldBank, 2014b).

The open data catalog (Datacatalogs, 2014) maintains a list of worldwide open data and shows an increase in governments' contribution. Prominent examples are the USA (DATA.GOV, 2014), the European Union (Open_data.Europa, 2014) and the United Nations (Data.un.org, 2014). A number of companies are also making parts of their data available through download or via application programming interface like Yelp (Yelp.com, 2014) which gives academics access to its business rating database. Finally, there are companies specialized in collecting and pre-processing all publically available data sets to offer "data supermarkets" such as Infochimps and DataMarket.

3.2.3 Social networks. Social networks are storing huge amounts of data comprising users' posts, messages, images, relationships and preferences. These data are generally accessible via proposed application programming interface or through special grants. For example, Twitter proposes its own application programming interface (Twitter developers, 2014) that give access to fractions from all tweets. The company also established a certified partners program in 2012: partners, such as Gnip (Gnip.com, 2014) are given a deeper access to twitter data, which they process to offer custom data sets and services. Finally, Twitter announced in 2014 a data grant project that will give selected research institutions access to its public and historical data (Blog.twitter.com, 2014). Facebook is less generous with data that can only be collected through custom Facebook applications, provided that users explicitly authorize access.

3.2.4 Crowdsourcing. Collecting massive amounts of data can be quite challenging, especially when it has to be done on a large scale. Crowdsourcing is a great solution for data collection and emerged in the past decade as an efficient way to harness the

creativity and intelligence of crowds. Recently, researchers sought to apply crowdsourcing to human subject research (Schmidt, 2010). Technical University Munich's ProteomicsDB and the International Barcode of Life projects are two good examples of collecting and gathering data using crowdsourcing (Schitka, 2014).

One of the best crowdsourcing platforms is Amazon Mechanical Turk, a crowdsourcing framework where assignments are distributed to a population of many unknown workers for fulfillment. This framework is getting to be progressively prominent with researchers and engineers (Ross *et al.*, 2010).

Another example of using crowdsourcing is a book, website and application called "The Human Face of Big Data" (Smolan and Erwitt, 2012); it is a crowdsourced media project focusing on humanity's new capacity to collect, analyzes, triangulate and visualize vast amounts of data in real time. The Human Face project details some of the projects that have begun crowdsourcing mass data "from accelerometers in computers and mobile devices to detect and warn of earthquakes" to safe-cast monitoring radiation levels in Japan (Schitka, 2014).

3.3 Data preprocessing

After data are collected, it is important to lay the ground for data analysis by applying various preprocessing operations to address potential imperfections in the raw data. For example, different sources may be involved, namely, sensors, social networks, internet-based applications and many other sources, thus implying different formats. Data collection methods and conditions are not flawless: faulty equipment or human's inattention can lead to a noisy dataset containing errors, redundancies and outliers. There may also be missing values and inconsistencies in codes (Gehrke, 2013; Singhal and Jena, 2013). Finally, data may simply need to fit requirements of analysis algorithms. As a consequence, collected data must be inspected, fused and all the above problems corrected during a pre-processing phase that covers a wide range of operations (Kotsiantis *et al.*, 2006):

- data cleaning eliminates the incorrect values and checks for data inconsistency;
- data integration – combines data from databases, files and different sources;
- data transformation – converts collected data formats to the format of the destination data system, and it can be divided into two steps: data mapping which links data elements from the source data system to the destination data system and data reduction which converts the data into a structure that is smaller but still generates the same analytical results; and
- data discretization is the process of converting continuous values to discrete ones and it can be done with different methods, and it has an important role as many learning algorithms require or work better with discrete values (Liu *et al.*, 2002).

Most data mining and business intelligence platforms include data preprocessing tools such as:

- WEKA which is a comprehensive open source toolset of machine learning and data visualization written in Java, created at the University of Waikato, New Zealand; and

- data cleaner (Datacleaner.org, 2013), another popular open source package for data pre-processing that offers a comprehensive toolset of analyzers, data transformers and writers.

3.4 Smart data analysis

Extracting value from a huge set of data is the ultimate goal of many firms. One efficient approach to achieve this is to use advanced analytics, which provides algorithms to perform various analytics on either structured or unstructured data. There are four types of advanced analytics (Figure 3):

(1) *Descriptive analytics*: Answer the question: *what happened* in the past? Knowing that in this context, the past could mean one minute ago or a few years back (Bigdata-startups.com, 2013). It is the most used class of big data analytics; last year over 80 per cent of the business analytics were descriptive analytics (community.lithium.com, 2013), and also it is the simplest as it only uses descriptive statistics such as sums, counts, averages, min and max to provide meaningful results about the analyzed data set. Descriptive analytics are typically used in social analytics and recommendation engines, such as Netflix recommendation system (Analytics.northwestern.edu, 2016), another common example is management reports that give information about sales, customers, operations, finance and to find correlations between the various variables.

(2) *Inquisitive analytics*: Also called diagnostic analytics, they answer the question *why something is happening*? By validating or rejecting business hypotheses. Data analysis techniques used here include analytical drill downs into data, factor analysis and conjoint analysis, among others (Mu-sigma.com, 2016; Bihani and Patil, 2014).

Inquisitive analytics can be used at different levels in the organization and in different fields, especially in marketing and retailing analytics for detecting

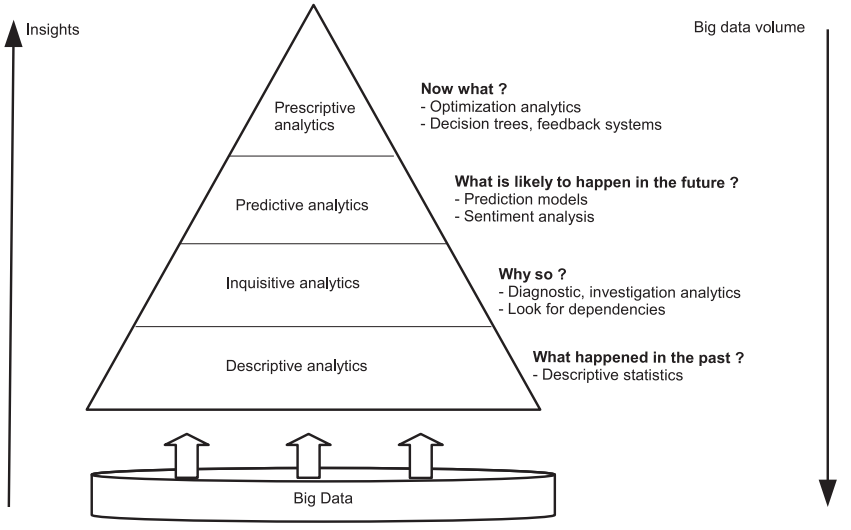

Figure 3.
Descriptive inquisitive predictive prescriptive analytics

anomalies in a manufacturing process or for diagnosing competitive strengths and weaknesses. They also could be used in healthcare area, as an example: Explorys, a leader in healthcare big data, used inquisitive analytics to find out that an unexplained variation in the evaluation of patients' weight was mainly due to some documentation gap (Explorys, 2013).

(3) *Predictive analytics*: Consists in studying the data we have, to predict data we do not have, such as future outcomes, in a probabilistic way (MikeW, 2013a), answering thereby the question "what is likely to happen?", they mainly use discriminate analysis to predict for example market behavior based on demographic and psychographic variables.

Predictive analytics are not all about time, there are also non-temporal predictive analytics, used for example in sentiment analysis to determine the affective state of social media users, or their judgment to a particular topic, by applying some techniques like natural language processing or computational linguistics to their posts and comments, another example is determining the influence of a social media user on the others, through data retrieved from his/her activity in the social application.

(4) *Prescriptive analytics*: Or optimization analytics consists in guiding decision making by answering the question "*so what?*" or "*what we must do now?*" It can be used by companies to optimize their scheduling, production, inventory and supply chain design 0. This class of analytics is relatively young, only around since 2003, and just 3 per cent of companies use this techniques and still with too many errors in it (Bigdata-startups.com, 2013; Tarantola, 2013). A prescriptive method includes a predictive model but with two more added component (MikeW, 2013b):

- First, the model must come up with an actionable outcome, based on which data users could make a decision.
- Then it must provide a feedback system that is able to take in the complex relationship between the user's actions and the adjusted result through the feedback data.

Conjoint analysis or choice modeling is an example of involved technique: it allows for example to simulate possible response when modifying some key features.

To guide companies in their software analytics choice, Gartner published the first magic quadrant (MQ) for advanced analytics (Herschel *et al.*, 2014), which presents 16 analytics platforms divided into four areas: leaders, challengers, visionaries and niche players, based

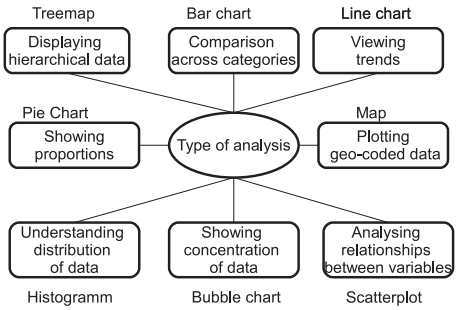

Figure 4.
Charts and graphs choice

Advanced analytics platforms	Descriptive	Inquisitive	Predictive	Prescriptive
SAS Visual analytics (Sas.com, 2016a, 2016b; MarketWatch, 2014)	✓	✓	✓	✓
IBM SPSS (IBM SPSS, 2013; Woodie, 2014)				
SPSS text analytics for surveys	✓			
SPSS data collection surveys reporter				
IBM SPSS statistics family	✓	✓	✓	✓
SPSS modeler professional			✓	
SPSS modeler premium				
IBM SPSS decision management				✓
Knime desktop (Knime.org, 2016)	✓	✓	✓	✓

Table III.
Advanced analytics platforms

on two criteria, which are the completeness of vision and the ability to execute. Table III presents three of advanced analytics platforms leaders in Gartner (MQ):

3.5 Representation and visualization

Visualization guides the analysis process and presents results in a meaningful way. For the simple depiction of data, most software packages support classical charts and dashboards. The choice is generally dictated by the type of desired analytics (Herschel *et al.*, 2014). Figure 4 shows some examples.

Working at big data scale brings multiple technical issues (sas.com, 2016a, 2016b). First, there are challenges related to the volume such as processing time, memory limitations and the need to fit different display types. Different approaches are explored to scale to big data, for example, at Intel Science & Technology Center for Big Data, projects work on various techniques such as visual summaries using data reduction principles, caching and pre-fetching to hide data store latency, query steering and large-scale parallelism (Istc-bigdata.org, 2013).

Second, deriving meaning from semi-structured and unstructured data requires adequate visualizations that are variably supported by software. Examples are:

- *Word clouds*: Express the occurrence of words in text format, where the size of the text corresponds to the frequency of the word. They can be used to analyze the main focus or topics of discussion for social media content or online feedback mechanism such as a survey.
- *Association trees*: Express the similarities of words meanings based on latent semantics analysis. They can be used to understand words associations in large quantities of text.
- *Network diagrams*: Show and quantify relationships between groups of subjects. They are commonly used to analyze leaders and followers roles in twitter social network.

Concerning the market of data visualization platforms, a study by Forrester Inc. (Evelson and Yuhanna, 2012) highlights market's diversity and the importance of both technology and visual design quality. It also shows that main technical differentiation factors are the performance of the in-memory engine, the quality of the graphical user interface and the comprehensiveness of data exploration and discovery tools. Leaders' board includes Tableau Software, TibcoSpotfire and SAS BI. A white paper by Tableau Inc. (Hanrahan *et al.*, 2009) identifies, as shown in Table IV, seven key features to assess a visual analytics application.

3.6 Actionable and timely insights extraction

Many sectors have already grasped big data regardless of whether the information comes from private or open sources. Big data can be used in almost any sector, thereafter are examples of cross-category use cases of Big Data that bring common benefits such as:

- Optimizing IT operations by tuning overall performance, allowing for better capacity planning and helping incidents prevention and management.
- Fine-tuning cloud-based application with better capacity management, prediction of user behavior and design of better targeting offerings.

Key elements	Description and importance	
	Description	Importance
Comprehensive visual exploration process	Integrates data exploration and querying	Ease of use and better focus
Augmented human perception	Effective visual properties and well-designed graphics	Fostering visual thinking
Visual expressiveness	Depth, flexibility and multi-dimensional expressiveness	Simple displaying of complex problems
Automatic visualization	Automatic suggestion of effective visualizations	Assisting the analysis process
Visual perspective shifting	Easily shifting between alternative visualizations of data	Assisting the analysis process
Visual perspective linking	Visual correlation of information in different visualizations	Assisting the analysis process
Collaborative visualization	Ease of sharing and distributed collaboration	Enhancing productivity

Table IV.
Assessment of visual analytics application

- Getting great insights into customers and public opinion more generally through sentiment analysis that explore all possible media sources including the trendy social networks.

Below are examples of industry-specific extracted insights, illustrated through successful big data projects implementations:

- *Manufacturing*: Manufacturing is one of the hardest hit sectors of the economy and also one of the sectors to use big data the least. There are numerous ways that a manufacturer can use big data to get an edge on their rival. By using a big data solution, manufacturers can not only deal with the growing data volume but also better analyze and share data, so issues could be tended quickly and significant pro-active insights can be gained (Nemschoff, 2014).

  Companies specialized in consumer products or retail organizations are using social networks such as Facebook and Twitter to get an exceptional perspective into customer behavior, preferences and product perception; by analyzing data uploaded by consumers, companies can know what kind of products they prefer. Other manufacturers are using big data to predict the optimal time to replace or maintain the products. Trading it excessively wastes money; replacing it too late provokes an unreasonable work stoppage (Nemschoff, 2014), a successful story in this field is that of Duke Energy (Acquia, 2014).

- *Finance*: This is a sector that massively benefits from big data to get various insights in risk management, fraud detection, marketing and customer retention.

  Financial Services associations are using data mined from customer interactions to slice and dice their clients into finely tuned fragments. This empowers these financial institutions to make progressively important and complex offers.

  The Oversea-Chinese Banking Corporation, as an example, analyzes customer data to determine customers' preferences. It designed an event-based advertising procedure that focused on a large volume of coordinated and customized marketing communications over numerous channels (IBM and Said Business School, 2013).

- *Health care*: Big data can bring great insights to the healthcare industry, provided that necessary changes are made to be able to capture the full value of big data. In addition to sentiment analysis, the sector can get various insights using predictive analytics. An example of use of big data in health care is a drug company in Seattle which uses data related to the genetics of cells to test cancer drug effectiveness (Horowitz, 2013).

  To identify undiscovered uses for drugs, the company decided to implement a data analysis system that incorporates cancer and non-cancer cell data with data from research published on Medline. Instead of having to conduct their own experiments, scientists use this combined database to test their theories against all known data that exist.

  While health care costs may be vital in big data's increase, clinical trends also play a role. When making treatment decisions, doctors have generally been using their judgment; however, in the past few years, there has been a move toward evidence-based medicine, which includes efficiently evaluating clinical data and making treatment decisions based on the best available information (Kayyali *et al.*, 2013).

- *Advertising and marketing*: Big data analytics have become critical for advertising and marketing. The capacity to collect and harness big volumes of consumer data and adapt it is vital. The accessibility of so much data offers big opportunities.

  Combining data with modern technology allows marketers and advertisers to aim ads to individual consumer at the most perfect time and spot (McLellan, 2014).

  Amazon has decided to expand its usage of big data to gain competitive advantage; they included remote computing services, via Amazon Web Services (AWS), to their massive product and advertising services. AWS started in 2002, but only recently they added Big Data services including data collection, data storage, data computation and data collaboration (AWS, 2014).
- *Agro-food industry*: Already in the early 1990s, food retailers started investing in data-mining systems to scan, analyze and exploit the purchasing behavior of consumers. It was at this time that has also developed loyalty cards, to track long-term purchases homes. With the explosion of internet, there will be a proliferation of digital traces, this statistical approach of "data" treatment is back on the front through "Big Data" (Rannou, 2014).

The use of "data" provide competitive advantage to overcome barriers to entry. Amazon for example has positioned itself in the market of products' traditional distribution in the USA. On a smaller scale, we are already seeing the emergence of original initiatives that are mixing the concepts of "data" and food. Here is a non-exhaustive list of innovative projects in food industry:

- *Open food system*: This research project, supported by SEB should enable the development of smart, connected devices capable of cooking food preparation in accordance with the nutritional properties (Reinhart, 2012).
- *Food genious*: This American company collects information on menus of 350,000 restaurants and retail food sales, to provide quantitative analysis (Food Genius, 2014).
- *Food pairing*: Food-pairing is based on the following hypothesis: The ingredients and drinks can be combined when they have many aromas in common. The principle is to combine well, foods that contain the same principal aromatic components (Foodpairing, 2012).
- *Openfoodfacts*: This collaborative Web platform offers to collect barcodes and all information on food products to bring more transparency and collectively participate in improving the supply process (Openfoodfacts.org, 2014).
- *Web mapping of coffee*: Mychefcom (MYCHEFCOM, 2014) operates analysis of hyperlinks between websites to represent the positioning of the actors from the world of coffee.

3.7 Evaluation of big data projects

To develop procedures for Big Data evaluation, we first need to study an integral model that identifies the factors driving successful projects and identifying what is missing in failed projects.

The integral model, made by Ken Wilber (Metcalf and Brenza, 2013), offers a successful assessment framework to enhance the initiative outcome. It looks at the intersection of four key components that, when aligned, promote successful transformation and when not

aligned contribute to transformation failure. The key components of the integral model are illustrated in Figure 5 and incorporate the following:

- "*Individual self*" is the leader's qualities including his values, objectives and convictions.
- "*Action*" is where the innovative leader acts using the abilities referenced above: Analytic expertise set (science, space learning and engineering), communication aptitudes, collaboration, customer-driven, strategic skills and problem solving.
- "*Society*" reflects the organization's society and the leader's understanding of it to make arrangement between him and the society.
- "*Systems*" incorporates the organizational and specialized systems and methodologies that dictate how the association achieves its work.

The leader needs to understand the current systems and guarantee they are updated to reflect the new actions needed to be successful. When implementing change, the leader must monitor the four elements of the model and make sure they are changing in ways that are well aligned and help each other.

Once we have finished evaluating the project from a leadership and management perspective, now we must proceed with the technical evaluation of the project. We need to consider a range of diverse data inputs, their quality and expected results. These aspects can span an extensive range of topics. To obtain a successful Big Data project, we need to have the right method for evaluating each one as a key effort for business productivity. To start developing procedures for Big Data evaluation, we first need to answer the following questions.

3.7.1 Return on investment (ROI). What is the ROI of the project? One of the most important questions to ask during a project evaluation process is, of course, whether companies are seeing a return on their investments. In fact, even if most of big data solutions use open source software, there are still costs included for designing, developing, deploying and keeping up the solution. Given this, what are the gains

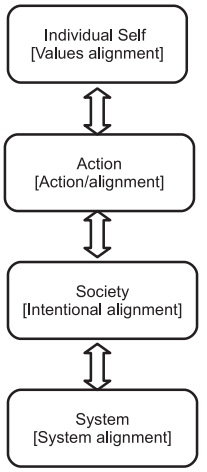

Figure 5.
The key elements of the integral model

harvested from such an investment? Are customers more satisfied? Are there any reduced costs?

First, the company should notice the minimal cost of the new solution, as it basically consists of open source software and commodity hardware, which provide more effective and faster results with less efforts. For example, Hadoop now enables cost effective parallel processing (Open Data Center Alliance, 2012).

Second, it should notice one or more changes within the business which improve the outcomes or reduce unnecessary costs expended before, those changes are going to be specific to the field or the activity concerned by the project, for example: what's that worth when a telecommunication firm is able to reduce customer churn by 10 per cent owing to a big data initiative? When a manufacturer can reduce the amount of monthly defective stock through a better inventory and orders management? Or when an organization can respond to business requests twice as fast?

Of course, a successful big data project will deliver a substantially bigger ROI than the business used to, moreover, it must give more value than simple incremental improvements of existing business models. To give an idea about how big could be the ROI of a big data project: a recent study conducted by TATA Consultancy Services which covered 643 companies, showed that the average ROI for companies using big data is 73 per cent; for energy and Resources Company, it was 61 per cent, high-tech companies estimated that the ROI was also higher than average, at 52 per cent, while banks and financial services organizations were lower than average (44 per cent). Companies in the heavy manufacturing (29 per cent ROI), life sciences (35 per cent), retailing (36 per cent), travel, hospitality and airlines (38 per cent) and telecom sectors (38 per cent) had the lowest expected returns on Big Data in 2012 (TATA Consultancy Services, 2013).

3.7.2 Technical accuracy

- *Does the project allow stream processing and incremental computation of statistics?* To get answers in real time about what is happening in the business, data streams are required. Examples of technologies for queuing data streams include TIBCO (2014), Zeromq (2014) and Esper (2014).
- *Does the project parallelize processing and exploit distributed computing?* Working with distributed data requires distributed processing, so that data will be processed in a reasonable amount of time.
- *Does the project easily integrate with visualization tools?* Once the implementation of the project is done, it must be able to integrate multiple visualization tools.
- *Does the project perform summary indexing to accelerate queries on big data sets?* Summary indexing is the procedure of making a pre-calculated summary of data to accelerate running queries.

3.7.3 Paradigm shift. How well are the new technologies adopted by internal staff? A survey by Talend shows that 52 per cent of respondents consider that in-house expertise is a real challenge for big data project success (Talend, 2012). The paradigm shift concerns not only technical staff who are required to understand and use the new technology but also decision makers who should think about business and ask questions in different ways.

3.8 Storage

Companies handling large amounts of data, such as Google and Amazon, were first to experience the limitations of traditional database management systems. The accelerated growth in data size requires horizontal scaling, which is the ability to extend the database over additional servers. But it turns out that managing sharding and replication with SQL databases is difficult and slow, as separate applications are required to handle these tasks. Furthermore, managing rapidly changing data needs greater flexibility in schema definition that is not available in classical databases, where structure and data types must be defined in advance. Finally, unstructured data are poorly supported, and the transaction mode can penalize performance for some big data analysis.

Several alternatives have been developed and are loosely grouped within the NoSQL family (NoSQL-database, 2014). These products are dissimilar as they fit different needs but most natively support horizontal scaling owing to automatic replication and auto-sharding. They also support dynamic schemas allowing for transparent real-time application changes. But with those distributed systems, only two features out of availability, consistency and partition tolerance can be provided (CAP theorem) (Brewer and Berkeley, 2000; Gilbert and Lynch, 2002); other forms of consistency are possible, like eventual consistency (CouchDB, 2014). Figure 6 shows the distribution of main databases according to provided CAP features.

We can also group NoSQL databases according to other shared characteristics including data model, horizontal scaling management and typical use cases. The four main categories are (CouchDB, 2014; MongoDB, 2014; Neo4j, 2014):

(1) *Document databases*: Designed to manage and store documents, they tolerate incomplete data and are programmer friendly. Their query performance is rather low and there is no standard query syntax. Examples are MangoDB and CouchDB.

(2) *Graph stores*: Store data in graphs that are the most generic of data structures. They are used for recommendation engines, storing network information and calculating their characteristic parameters. The downside is they are not easy to cluster and require traversing the entire graph for definitive answers. Examples are Neo4j and InfoGrid.

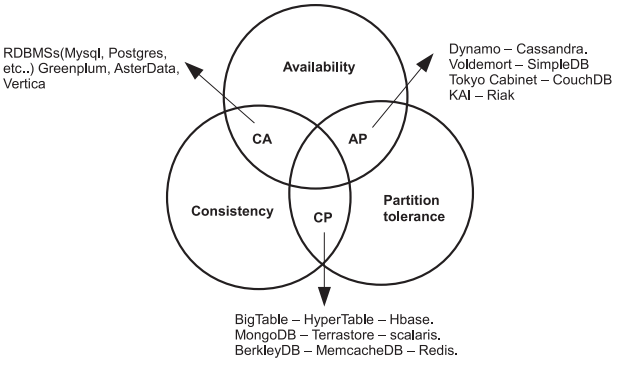

Figure 6.
Databases and CAP features

(3) *Key-value stores*: Simply store every single item as an attribute name (or key) with its corresponding value. Being programmer friendly and handling size well, they are used for managing users' sessions, shopping carts, fast lookups and content-caching. Examples are Dynamo (Amazon) and Voldemort (Linkedin).

(4) *Wide-column stores*: Use a column-oriented data model and are most patterned after Google's Bigtable. They are good for distributed data storage, versioning data, large scale data processing, and predictive analytics. Examples are HBase and Cassandra (Facebook).

On another aspect, distributed massive storage naturally raises computational problems. Google MapReduce paradigm (Dean and Ghemawat, 2008) is an efficient solution that distributes extremely large problem across extremely large computing cluster, allowing programs to automatically parallelize. Most NoSQL databases adopted the MapReduce model and one of the most popular open source implementation of MapReduce is Hadoop.

Hadoop was originally created at Yahoo, by building upon Google MapReduce and Google File System papers. It is now a comprehensive framework for big data projects, maintained by the Apache Software foundation. Hadoop is essentially a batch processing system but can integrate with other projects for real-time analysis such as Apache Storm. Its architecture can be logically split into three main layers, as illustrated in Figure 7 (Apache Hadoop, 2014).

The application layer provides high-level frameworks for distributed programming, such as Pig for data analysis and Hive for data warehouse software. In the middle layer, the non-relational database Hbase offers key features such as automatic-sharding, automatic failover and workload management across different resources using the MapReduce model. HBase runs on the top of Hadoop Distributed file system, the component that manages the actual storage at the data layer. In addition to distribution and replication, hadoop distributed file system offers high fault-tolerance, high throughput data access and is designed for use on commodity hardware.

Many vendors offer customized and extended Hadoop distributions for specific needs, according to an assessment published by Forrester research (Gualtieri and Yuhanna, 2014) that covers nine Hadoop solutions. Top three leaders are AWS, Cloudera and HortonWorks. There are also solutions outside Hadoop ecosystem, such as Spark and Disco which are appreciated by engineers for their speed and real-time support (Byte-mining, 2011).

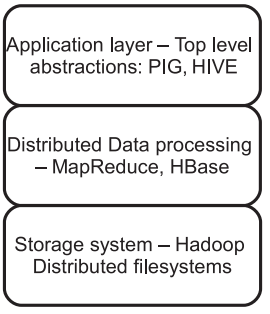

Figure 7.
Simplified Hadoop architecture

Choosing the right solution must take into consideration company's needs, required technical competence and the characteristics of existing solutions. Decisions then should be made regarding capacity planning, infrastructure, tradeoffs and complexity of various processes (Open Data Center Alliance, 2012):

- *Impact on existing infrastructure*: As companies generally have an existing storage infrastructure, it must be decided whether the big data solution will replace it or complement it.
- *Capacity planning*: It is important to decide on potential required hardware, storage types and network configurations.
- *Tradeoffs*: The company must align its project priorities with the characteristics of big data solutions by deciding on tradeoffs between availability, consistency and partition tolerance, and between scalability, elasticity and availability.
- *Latency*: Whether the company needs batch processing of data or real-time analysis has different impact in term of infrastructure.
- *Complexity*: There should be a clear vision on the flow and potential difficulties of the deployment, maintenance, monitoring and management processes.

Regarding solution sources, any chosen solution fells into one of the following categories:

- *Open source solutions*: These offer great flexibility and portability with no license fees. This category is best suited for companies with technically qualified staff as the implementation and maintenance are rather difficult in the absence of official training and support. Apache Hadoop is a case in point.
- *Third party distributions*: These offer better support through training and certification programs, provided upgrades, bug-fixes and patches. Costs include license fees, limited flexibility and dependence on vendor's expertise. This category best fits beginners and large companies with diverse information system. Examples are Cloudera and Hortonworks.
- *Proprietary solutions*: These are the easiest to deploy as vendors offer complete support and can even be partners in developing and maintaining the platform. In addition to price, costs are the same as the third party distributions. This category suits well for companies who need turnkey solutions, and top providers include Oracle and IBM.

4. Conclusion

In this paper, we presented a clear and step-by-step roadmap that covers the whole life cycle of a big data project setup, from data collection to implementation and then evaluation. We tried to provide answers to some fundamental questions any company or organization, willing to embark on the big data adventure and reap the most rewards out of its data, would ask. The staged methodology aims at covering all aspects and issues related to big data projects implementation and giving hints for ICT platforms and tools to assist in big data processing/managing. The paper also pointed to several real-world big data projects. The paper might be used as a kind of checklist by business leaders (Chief Information Officers and Chief Executive Officers) for a systematic view of the development process of big data projects.

References

Acquia (2014), "Examples of big data projects", available at: www.acquia.com/fr/examples-big-data-projects (accessed 29 April 2014).

Analytics.northwestern.edu (2016), "North Western engineering", available at: www.analytics.northwestern.edu/program-overview/analytics-exampleshtml (accessed 13 March 2014).

MYCHEFCOM (2014), available at: www.mychefcom.com/ (accessed 29 April 2014).

Apache Hadoop (2014), "Welcome to apache Hadoop!", available at: http://hadoop.apache.org/ (accessed 20 March 2014).

ATTUNITY (2013), "Cloud adoption rates reaching 75 per cent in 2013", available at: www.attunity.com/learning/articles/cloud-adoption-rates-reaching-75-percent-2013 (accessed 16 May 2014).

AWS (2014), "Amazon web services (Français)", available at: http://aws.amazon.com/fr/ (accessed 28 April 2014).

Bigdata-startups.com (2013), "Datafloq – the one-stop shop for Big data", available at: www.bigdata-startups.com/understanding-business-descriptive-predictive-prescriptive-analytics (accessed 12 March 2014).

Bihani, P. and Patil, S. (2014), "A comparative study of data analysis techniques", *International Journal of Emerging Trends & Technology in Computer Science*, Vol. 3 No. 2, pp. 95-101.

Blog.twitter.com (2014), "Introducing twitter data grants | twitter blogs", available at: https://blog.twitter.com/2014/introducing-twitter-data-grants (accessed 16 March 2014).

Bostic, W.G. (2013), "Big data for policy, development, and official statistics", available at: http://unstats.un.org/unsd/statcom/statcom_2013/seminars/Big_Data/UN_BigData_Bostic_022213.pdf

Brewer, E.A. and Berkeley, U.C. (2000), *Towards Robust Distributed Systems*.

Business Wire (2011), "Zettaset's new security data warehouse enables big data mining for forensic analysis", available at: www.businesswire.com/news/home/20110802005827/en/Zettaset?s-Security-Data-Warehouse-Enables-Big-Data#.U58M9JR5NDW (accessed 7 March 2014).

Byte-mining (2011), "Hadoop fatigue: alternatives to hadoop byte mining", available at: www.bytemining.com/2011/08/hadoop-fatigue-alternatives-to-hadoop/ (accessed 10 March 2014).

Cisco (2013), "Cisco visual networking index forecast projects 18-fold growth in global mobile internet data traffic from 2011 to 2016", available at: http://newsroom.cisco.com/press-release-content?articleId=668380 (accessed 11 May 2014).

Cloud Security Alliance (2013), "Expanded top ten big data security and privacy challenges, community.lithium.com", *Big Data Reduction 1: Descriptive Analytics*, available at: http://community.lithium.com/t5/Science-of-Social-blog/Big-Data-Reduction-1-Descriptive-Analytics/ba-p/77766 (accessed 12 March 2014).

CouchDB (2014), "Eventual consistency", available at: http://guide.couchdb.org/draft/consistency.html (accessed 20 March 2014).

Datacatalogs (2014), available at: http://datacatalogs.org/ (accessed 18 March 2014).

Datacleaner.org (2013), "Reference documentation", available at: http://datacleaner.org/resources/docs/3.5.10/html/ (accessed 19 March 2014).

DATA.GOV (2014), "Open data in the United States", available at: www.data.gov/open-gov/ (accessed 18 March 2014).

Data.un.org (2014), available at: http://data.un.org/ (accessed 18 March 2014).

Dean, J. and Ghemawat, S. (2008), "Mapreduce: simplified data processing on large clusters", *Communications of the ACM*, Vol. 51 No. 1, pp. 107-113.

Desouza, K.C. (2014), "Realizing the promise of big data: implementing big data projects", available at: http://icma.org/Documents/Document/Document/305991

Doug, H. (2013), "Big data success: 3 companies share secrets", available at: www.informationweek.com/big-data/big-data-analytics/big-data-success-3-companies-share-secrets/d/d-id/1111815? (accessed 2 May 2014).

EMC2 (2014), "RSA – information security, governance, risk, and compliance", available at: www.emc.com/domains/rsa/index.htm (accessed 8 March 2014).

Esper (2014), "Complex event processing", available at: http://esper.codehaus.org/ (accessed 6 April 2014).

Evelson, B. and Yuhanna, N. (2012), *The Forrester Wave™: Advanced Data Visualization (ADV) Platforms*, Q3 2012, Forrester.

Explorys (2013), "Explorys categorizes analytics as descriptive, diagnostic, predictive and prescriptive", available at: www.explorys.com/results/news-results/2013/12/09/explorys-categorizes-analytics-as-descriptive-diagnostic-predictive-and-prescriptive (accessed 15 March 2014).

Food Genius (2014), "Welcome to food genius", available at: http://getfoodgenius.com/ (accessed 29 April 2014).

Foodpairing (2012), "Foodpairing – homepage", available at: www.foodpairing.com/fr/home (accessed 29 April 2014).

Fortscale (2014), "Big data security analytics", available at: www.fortscale.com/ (accessed 8 March 2014).

Gartner.com (2016), "Gartner survey reveals that 64 per cent of organizations have invested or plan to invest in big data in 2013", available at: www.gartner.com/newsroom/id/2593815 (accessed 23 September 2013).

Gehrke, J. (2013), "Real data, real problems: pre-processing for big data analytics", available at: http://blog.ecornell.com/real-data-real-problems-pre-processing-for-big-data-analytics/ (accessed 2 May 2014).

Gilbert, S. and Lynch, N. (2002), "Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services", *ACM SIGACT News*, Vol. 33 No. 2, p. 51, available at: http://portal.acm.org/citation.cfm?doid=,564585.564601

Gnip.com (2014), available at: http://gnip.com/ (accessed 16 March 2014).

Gualtieri, M. and Yuhanna, N. (2014), *The Forrester Wave: Big Data Hadoop Solutions*.

Herschel, G., Linden, A. and Kart, L. (2014), "Magic quadrant for advanced analytics platforms", available at: www.gartner.com/doc/2667527 (accessed 16 March 2014).

Horowitz, A.S. (2013), "Let's play moneyball: 5 industries that should bet on data analytics", available at: http://plotting-success.softwareadvice.com/moneyball-5-industries-bet-on-analytics-1113/ (accessed 30 April 2014).

IBM and Said Business School (2013), *Analytics: The Real-world Use of Big Data in Financial Services*.

IBM (2014), "IBM – infosphere data explorer", available at: www-03.ibm.com/software/products/en/dataexplorer (accessed 2 March 2014).

IBM Software (2013), "The top five ways to get started with big data", available at: http://files.netcommunities.com/ibm-swg/Sep/the_top_five_ways_imw14710usen[1].pdf

IBM SPSS (2013), "Get to know the IBM SPSS product portfolio", IBM Software, Business Analytics.

Intel IT Center (2013), *Big Data in the Cloud: Converging Technologies Big Data in the Cloud: Converging Technologies*, Intel IT Center.

Istc-bigdata.org. (2013), "Making big data visualization more accessible | Intel Science and Technology Center for Big Data", available at: http://istc-bigdata.org/index.php/making-big-data-visualization-more-accessible/ (accessed 24 March 2013).

Kayyali, B., Knott, D. and Steve, V.K. (2013), "The big-data revolution in US health care: accelerating value and innovation", available at: www.mckinsey.com/insights/health_systems_and_services/the_big-data_revolution_in_us_health_care (accessed 29 April 2014).

Kern, E. (2012), "Facebook is collecting your data – 500 terabytes a day | Gigaom", available at: http://gigaom.com/2012/08/22/facebook-is-collecting-your-data-500-terabytes-a-day/ (accessed 10 February 2014).

Khanzode, G.P. (2012), *Insights Internet of Things: Endless Opportunities*.

King, C. (2013), "Navigating through the big data space", available at: http://ase.co.uk/wp-content/uploads/2014/02/Navigating-Through-The-Big-Data-Space.pdf

Knime.org (2016), "KNIME | open for innovation", available at: www.knime.org (accessed 17 March 2014).

Kotsiantis, S.B., Kanellopoulos, D. and Pintelas, P. (2006), "Data preprocessing for supervised leaning", *International Journal of Computer Science*, Vol. 1 No. 1, pp. 111-117.

Liu, H., Hussain, F., Tan, C. and Dash, M. (2002), "Discretization: an enabling technique", *Data Mining and Knowledge Discovery*, Vol. 6 No. 4, pp. 393-423.

LogRythm (2014), "LogRythm's security intelligence platform", available at: www.logrhythm.com/ (accessed 8 March 2014).

McLellan, M. (2014), "Big data powers digital advertising and marketing programs", *Bizmology*, available at: http://bizmology.hoovers.com/2014/02/28/big-data-advertising-marketing/ (accessed 30 April 2014).

Metcalf, M. and Brenza, J. (2013), "Evaluating big data projects – success and failure using an integral lens", available at: http://integralleadershipreview.com/10945-evaluating-big-data-projects-success-failure-using-integral-lens/ (accessed 29 April 2014).

MarketWatch (2014), "Sas a leader in Gartner's 2014 BI and analytics platforms magic quadrant", available at: www.marketwatch.com/story/sas-a-leader-in-gartners-2014-bi-and-analytics-platforms-magic-quadrant-2014-02-27 (accessed 17 March 2014).

MikeW (2013a), "Big data reduction 2: understanding predictive analytics", available at: http://community.lithium.com/t5/Science-of-Social-blog/Big-Data-Reduction-2-Understanding-Predictive-Analytics/ba-p/79616 (accessed 12 March 2014).

MikeW (2013), "Big data reduction 3: from descriptive to prescriptive", available at: http://community.lithium.com/t5/Science-of-Social-blog/Big-Data-Reduction-3-From-Descriptive-to-Prescriptive/ba-p/81556 (accessed 12 March 2014).

MongoDB (2014), "NoSQL databases explained", available at: www.mongodb.com/nosql-explained (accessed 21 February 2014).

Mousannif, H. and Khalil, I. (2014), "The human face of mobile", Linawati, M.S., Neuhold, E.J., Tjoa, A.M. and You, I. (Eds), *ICT-Eur Asia 2014*, Springer Berlin Heidelberg.

Mousannif, H., Al Moatassime, H. and Rakrak, S. (2011), "An energy-efficient scheme for reporting events over WSNs", *International Journal of Pervasive Computing and Communications*, Vol. 7, pp. 44-59.

Mousannif, H., Khalil, I. and Olarju, S. (2012), "Cooperation as a service in VANET: implementation and simulation results", *Mobile Information Systems Journal*, Vol. 8, pp. 153-172.

Mousannif, H., Sabah, H., Douiji, Y. and Sayad, Y. (2014), "From big data to big projects: a step-by-step roadmap", *2014 International Conference on Future Internet of Things and Cloud*. doi: 10.1109/FiCloud.2014.66.

Mu-sigma.com (2016), *Descriptive and Predictive Analytics |Prescriptive Analytics | DIPP™ Framework | Mu Sigma*, available at: www.mu-sigma.com/analytics/ecosystem/dipp.html (accessed 10 April 2014).

Nemschoff, M. (2014), "Big data and manufacturing quality", available at: https://smartdatacollective.com/michelenemschoff/176666/how-big-data-can-improve-manufacturing-quality (accessed 30 April 2014).

Neo4j (2014), "What is a graph database?", available at: www.neo4j.org/learn/graphdatabase (accessed 21 February 2014).

Network Computing (2010), "Negotiating cloud computing contracts – network computing", available at: www.networkcomputing.com/cloud-infrastructure/negotiating-cloud-computing-contracts/d/d-id/1231794 (accessed 15 May 2014).

Kshetri, N. (2010), "Privacy and security issues in cloud computing", *33rd Annual Pacific Telecommunications Conference*, pp. 1-711.

NoSQL-database (2014), "List of NoSQL databases", available at: http://nosql-database.org/ (accessed 20 March 2014).

Olstik, J. (2013), "The big data security analytics era is here", available at: www.emc.com/collateral/analyst-reports/security-analytics-esg-ar.pdf

Open Data Center Alliance (2012), *Big Data Consumer Guide*.

Open_data.Europa (2014), "European Union open data portal", available at: https://open-data.europa.eu/en/data/ (accessed 18 March 2014).

Openfoodfacts.org (2014), "Open food facts – France", available at: http://fr.openfoodfacts.org/ (accessed 2 May 2014).

Hanrahan, P., Stolle, C. and Mackinlay, J. (2009), "Selecting a visual analytics application", Tableau Software.

Rannou, A. (2014), "Big data and insight solutions for the food industry", available at: www.mychefcom.com/blog/quelles-applications-pour-le-big-data-dans-lagroalimentaire- (accessed 29 April 2014).

Reinhart, L. (2012), "Investissements d'avenir: 9,1 millions d'euros pour le project Open food system Project".

Ross, J., Irani, L., Silberman, M., Zaldivar, A. and Tomlinson, B. (2010), "Who are the crowdworkers?: Shifting demographics in mechanical turk", *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pp. 2863-2872.

Sas.com (2016a), "Data visualization techniques from basics to big data with SAS® visual analytics", available at: www.sas.com/offices/NA/canada/downloads/IT-World2013/Data-Visualization-Techniques.pdf (accessed 12 March 2013).

Sas.com (2016b), "SAS visual analytics", available at: www.sas.com/en_us/software/business-intelligence/visual-analytics.html (accessed 17 March 2014).

Schitka, J. (2014), "Collecting massive data via crowdsourcing – strata 2014", available at: http://strataconf.com/strata2014/public/schedule/detail/33541 (accessed 18 March 2014).

Schmidt, L.A. (2010), "Crowdsourcing for human subjects research", *CrowdConf 2010*.

Sicular, S. (2013), *The Roadmap for Successful Big Data Adoption*.

Singhal, S. and Jena, M. (2013), "A study on WEKA tool for data preprocessing, classification and clustering", available at: http://ijitee.org/attachments/File/v2i6/F0843052613.pdf (accessed 16 March 2014).

Smolan, R. and Erwitt, J. (2012), *The Human Face of Big Data*, available at: http://humanfaceofbigdata.com (accessed 2 May 2014).

Talend (2012), "How big is big data adoption?", available at: http://info.talend.com/rs/talend/images/WP_EN_BD_Talend_SurveyResults_BigDataAdoption.pdf

Tarantola, A. (2013), "How prescriptive analytics could harness big data to see the future", available at: http://gizmodo.com/how-prescriptive-analytics-could-harness-big-data-to-se-512396683 (accessed 12 March 2014).

TATA Consultancy Services (2013), "The emerging big returns on big data", available at: www.tcs.com/SiteCollectionDocuments/Trends_Study/TCS-Big-Data-Global-Trend-Study-2013.pdf (accessed 23 March 2014).

TATA Consultancy Services (2014), "Manufacturing: big data benefits and challenges", available at: http://sites.tcs.com/big-data-study/manufacturing-big-data-benefits-challenges/ (accessed 2 May 2014).

TERADATA (2013), *The Cloud: An Abundance of Choices*.

TIBCO (2014), "Global leader in infrastructure and business intelligence software", available at: www.tibco.com/ (accessed 6 April 2014).

Twitter Developers (2014), "History of the REST and search API", available at: https://dev.twitter.com/docs/history-rest-search-api (accessed 16 March 2014).

Tynt (2014), available at: www.tynt.com/ (accessed 2 May 2014).

Villars, R.L., Olofson, C.W. and Eastwood, M. (2011), *Big Data: What It Is and Why You Should Care?*

Voltage (2013), "Voltage enterprise security for big data", available at: www.voltage.com/wp-content/uploads/Voltage_SB_Corporate.pdf

Vortmetric (2014), "Big data security intelligence solution", available at: www.vormetric.com/data-security-solutions/applications/big-data-security (accessed 8 March 2014).

Wikibon Blog (2016), "The big list of big data infographics", available at: http://wikibon.org/blog/big-data-infographics/ (accessed 24 February 2014).

Woodie, A. (2014), "SAS and IBM king of analytics hill, but for how long?", available at: www.datanami.com/datanami/2014-03-03/sas_and_ibm_king_of_analytics_hill_but_for_how_long_.html (accessed 17 March 2014).

WorldBank (2014a), "Open data catalog", available at: http://datacatalog.worldbank.org/ (accessed 18 March 2014).

WorldBank (2014b), "The world bank for developers", available at: http://data.worldbank.org/developers (accessed 18 March 2014).

Yelp.com (2014), *Yelp Dataset Challenge | Yelp*, available at: www.yelp.com/academic_dataset (accessed 18 March 2014).

Youtube.com (2016), *Statistics – YouTube*, available at: www.youtube.com/yt/press/statistics.html (accessed 10 March 2016).

Zeromq (2014), "Code connected", available at: http://zeromq.org/ (accessed 6 April 2014).

Further reading

Abadi, D., Boncz, P., Harizopoulos, S., Idreos, S. and Madden, S. (2012), "The design and implementation of modern column-oriented database systems", *Foundations and Trends® in Databases*, Vol. 5 No. 3, pp. 197-280, available at: www.nowpublishers.com/articles/foundations-and-trends-in-databases/DBS-024 (accessed 16 June 2014).

Abadi, D.J., Madden, S.R. and Hachem, N. (2008), "Column-stores vs row-stores: how different are they really? Categories and subject descriptors", *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data in SIGMOD '08, ACM, New York, NY*, pp. 967-980.

Hardin, M., Hom, D., Perez, R. and Williams, L. (2016), "Which chart or graph is right for you?", Tableau Software.

About the authors

Hajar Mousannif is an Assistant Professor in the department of computer science at the Faculty of Sciences Semlalia (Cadi Ayyad University, Morocco). She holds a PhD degree in computer sciences and an engineering degree in telecommunications. Her primary research interests include big data cloud computing, human computer interaction and next-generation internet technologies. In addition to her academic experience, she was in the Technical Program Committee of many international conferences. Hajar Mousannif is the corresponding author and can be contacted at: hajar.mousannif@gmail.com

Hasna SABAH is a PhD Candidate at Cadi Ayyad University of Morocco. Her topics of interests include data mining, big data and open source. She obtained her engineering degree from the National Institute of Posts and Telecommunications and has a working experience in teaching and system/network administration.

Yasmina Douiji is a PhD Student in engineering sciences at the University of CADI AYAD, Marrakesh, Morocco. Her research interests primarily in artificial intelligence. Her work focuses on developing a multimodal emotional recognition system, using machine learning and applied to mobile technology. Yasmina graduated from the same university at the faculty of sciences and techniques with an engineer diploma in informatics network and information system in 2013.

Younes Oulad Sayad is a PhD Candidate in computer sciences at the University of Cadi Ayyad, Marrakesh, Morocco, with research interests primarily in big data, cloud computing and e-health. His work focuses on developing a health care system using big data and cloud computing. Younes graduated from the same university at the faculty of sciences Semlalia with a master's degree in engineering information systems in 2013.