School_____ School of engineering and entrepreneurship _____
Direction of training (Specialty)_____ 31.04.02   Management_____

MASTER'S THESIS

Topic of the work
Big data in the public sector

UDC: 334.012.33: 004-65.047.44

Student

Group	Full name	Signature	Date
3АМ7Ф	Rime Belahcene		

Scientific Supervisor

Position	Full name	Academic degree, academic rank	Signature	Date
Associate professor	A. Dankov	PHD		

ADVISORS:

Section "Social Responsibility"

Position	Full name	Academic degree, academic rank	Signature	Date
Associate professor	N.A. Cherepanova	PHD		

ADMIT TO DEFENSE:

Head of the Program	Full name	Academic degree, academic rank	Signature	Date
Management	N. O Chistyakova	PHD		

Tomsk – 2018

School:_____School of engineering and entrepreneurship _____
Direction of training (Specialty) _____31.04.02 Management_____

APPROVED BY:
Head of the program

_____ _____ N.O Chistyakova

(Signature)   (Date)   (Full name)

ASSIGNMENT
for the Master's Thesis completion

In the form:

Master's Thesis

For a student

Group	Full name
3АМ7Ф	Rime Belahcene

Topic of the work:

Big data in the public sector	
Approved by the order of the Head	1982/c from 20.03.2018

Deadline for completion of the Master's Thesis:	09.06.2018

TERMS OF REFERENCE:

Initial data for work *(the name of the object of research or design; performance or load; mode of operation (continuous, periodic, cyclic, etc.); type of raw material or material of the product; requirements for the product, product or process; special requirements to the features of the operation of the object or product in terms of operational safety, environmental impact, energy costs; economic analysis, etc.).*	1. Legal documents 2. Dissertations on the topic 3. Periodicals 4. Internet sources 5. Information about SNCF (Société nationale des chemins de fers), social insurance organizations 6. CGI's intranet materials

<table>
<tr><td>List of the issues to be investigated, designed and developed

<i>(analytical review of literary sources in order to elucidate the achievements of world science and technology in the field under consideration, the formulation of the problem of research, design, construction, the content of the procedure of the research, design, construction, discussion of the performed work results, the name of additional sections to be developed; work conclusion).</i></td><td>Section 1: Big data main definition and main issues of the public sector
Section 2: Big data in public sector: examples
Section 3: The big data in public sector: future vision and recommendation</td></tr>
<tr><td>List of graphic material

<i>(with an exact indication of mandatory drawings)</i></td><td>No</td></tr>
</table>

Advisors on the sections of the Master's Thesis
(with indication of sections)

Chapter	Advisor
Social responsibility	N.V. Cherepanova

Date of issuance of the assignment for Master's Thesis completion according to a line schedule	15.11.2017

The task was issued by the Head:

Position	Full name	Academic degree, academic status	Signature	Date
Associate professor	A. Dankov	PHD		

The assignment was accepted for execution by the student:

Croup	Full name	Signature	Date
3АМ7Ф	Rime Belahcene		

3

Ministry of Education and Science of the Russian Federation
Federal Independent Educational Institution
"NATIONAL RESEARCH TOMSK POLYTECHNIC UNIVERSITY"

School   School of engineering and entrepreneurship

Direction of training (Specialty)   31.04.02   Management

Level of education   Master's degree program

Period of completion   fall/spring semester 2017 /2018

Form of presenting the work:

Master's thesis

SCHEDULED COURSE ASSESSMENT
CALENDAR
for the Master's Thesis completion

Deadline for completion of the Master's Thesis:	09.06.2018

Assessment date	Title of section (module) / type of work (research)	Maximum score of the section (module)
25.04	*Big data main definition and main issues in public sectors*	
15.05	*Big data in public sector: examples*	
01.06	*The big data in public sector: future vision and recommendation*	
05.06	*Social responsibility*	
06.06	*Practice report*	
07.06	*Final control*	
11.06	*Downloading in electronic system*	

Made by professor:

Position	Full name	Academic degree, academic status	Signature	Date
Associate professor	A. Dankov	PHD		

AGREED:

Head of the Program	Full name	Academic degree, academic status	Signature	Date
Management	N. O Chistyakova	PHD		

ABSTRACT

Master's Thesis 75 p., 29 fig., 3 tabl., 30 references, 1 appendix.

Key words: big data, public sector, administration, public organization, public companies, digital. The objects of the research are French public organizations. The purpose of the work is to made a survey and I tried to analyze what people think about the big data and how they measure the effectiveness of big data in the public sector. As a result of research I analyzed the methods of the implementation of big data analyses in public organizations. Application area: all the public organization can use my researched, IT companies which provide software for big data analyzes, governmental agencies.

Economic efficiency/significance of research outcomes they can help to improve the communication systems, the effectiveness management.

Table of content:

Acknowlegdments:

I would like to thank Natalia Chistyakova, who helps us during all this semester and she has been listening and available. Our professors for taking time to discuss and debate and give us a good education.

I also would like to thank all the personnel from Tomsk polytechnics university for developing this partnership between Tomsk and Orléans and permit us to discover this beautiful country, Russia.

Introduction

According to previsions from IBM, 1,7 MB of new information will be produced every seconds for each human on the world. From the retail business to health, from travel to automotive sector, each sector should adopt very fast the digital services adapted itself to this change.

The notion of big data sends back to mass data produced by administrations, companies, users and all other entity. The quantity of information made these different technologies. The big data era is a new way to manage data, not only like passive element, which permitted to handle record, but more like an active tool which can influence the situation management.

The beginning of this era was with new Internet services like Facebook, Google and Twitter. This type of services was free, so they had to find a business model, a way to earn money. The heart of the business is the user's data, their behaviours or the pages they visit for example.

Public organisation especially social insurance and social security have more difficulties to deal with big data and it takes more time to implement this process. That conduce us to ask how big data can be used in public sector?

Public sector includes public organisation, public companies, administration, and association with public interest, governed by the State.

First we will explain the big data main definition and particularities, how we can use big data, in general, in private companies and how we can use it in public sector the main issues, problems.

Then, we will give some examples with national and local impact. These examples will concern social area, health and cities and infrastructures development.

To conclude we will expose some preconisation about the use of big data in public sector, what should be done, what should be developed and how to overcome the gaps.

Section 1. Big data main definition and main issues of the public sector

1.1 Main definition of the big data.

In 2000, 3 V characterized the big data: Volume, Velocity, Variety. This characterisation was proposed by Doug Laney.

Volume: companies should deal with more and more information, nowadays, these information are counted by teraoctet. In the future information technologies will be capable of running on at least petaoctets and more. For example, 72 hours of video are uploaded to YouTube every minute. This amount of information introduces some other question about storage or methodology to analyse. Indeed, which type of storage should be used, at what price, and more important the data security.

Velocity: immediacy to send and receive data for every one in every activity force companies to improve their promptness to respond. They should adapt their analysis tools. The main challenges concern the modelling, the control of data and how to give the right information to the right person.

Variety: there are more and more information, more and more type of information and many forms, structured or not. That's one of the main purposes of big data. Information comes from many sources, big data tools should manage this forms, not to lose capital information.

Nowadays, we can see in addition veracity, validity, volatility. Veracity: this refers to biases and noise for example. We have a lot of information and all of them are not seen as necessary to do analysis. The main challenge for companies is to have clean data, they should be updated. If we keep dirty data, it's a time bomb. Analysis and prediction with dirty data can conduce to bankrupt. Validity: the validity means that the data should be correct. It's a huge challenge for companies, for example, Phil Francisco, VP of Product Management from IBM spoke about IBM's big data strategy and tools they offer to help with data veracity and validity.

Source : http://whatis.techtarget.com/definition/3Vs

Volatility: due to the important volume, economic actors ask themselves about the validity time of the data, how many days, month, and year we can store a data and use it for a good analysis.

Beyond this tridimentional aspect, the big data has been launched because of the traditional statistics tools obsolescence and their predictive dimension.

Traditional statistics used hypothetic-deductive model, it consists to reflect at the beginning of the process about what data to use and variable that can analyse it with some hypothesis.

Big data try to eliminate this first part of the process to let algorithm to search some regularities. This evolution permits to have less human biases.

1.2 Main challenges:

The main challenges for extracting business value from are:

-        Properly business challenges (like the potential ROI, design a viable business model, etc)

-        Operational challenges (new operating model: decentralized for example, internal culture, retain right talent to have a good data science team)

-        Technological challenges (define the good infrastructure, big data solution)

-        Legal challenges (anticipate new regulation impacts, user empowerment, trusting relationship with client, be transparent)

The decision-makers take more and more in account the recent data to make their quotidian and strategic decisions. Data are not only used to support decisions. They are an individual part of the decision. Time and old treatment methods can't do it, and they also can't manage all the data because of the data flow and complexity. For any reason the big data use degree, organisations have to reduce the time between the data collection and the implementation of concrete actions. More than 75% of decision-makers declare that they want to have necessary information to take decisions in 24 hours (before big data it was around 1 week).

1.3- The big data market:

There are many actors in this market. In this scheme, we have the main companies working divided by "sectors":

-        Vertical apps:

The sofftware application that supports a business process and target for a small group of user with specific skills. Example an ERP

-        Ad/Media apps:

That use bug data for advertising and communication

-        Business intelligence:

Collect data, prepare it for analysis, developp queries against data , create reports and data visualizations to make analytical results for corporate decision.

-        Analytics and Visualization

Analyse and modelisation of the data

- Log data apps

File of event logged by a software application, it contains error, informational event and warning.

- Data as a service that sell data

- Analytics infrastructure

- Operational infrastructure

- Indrastructure as a service

- Structured databases that help to structure the data base because of the different format

- Technologies which sell the technoligies to manage data like servers etc…
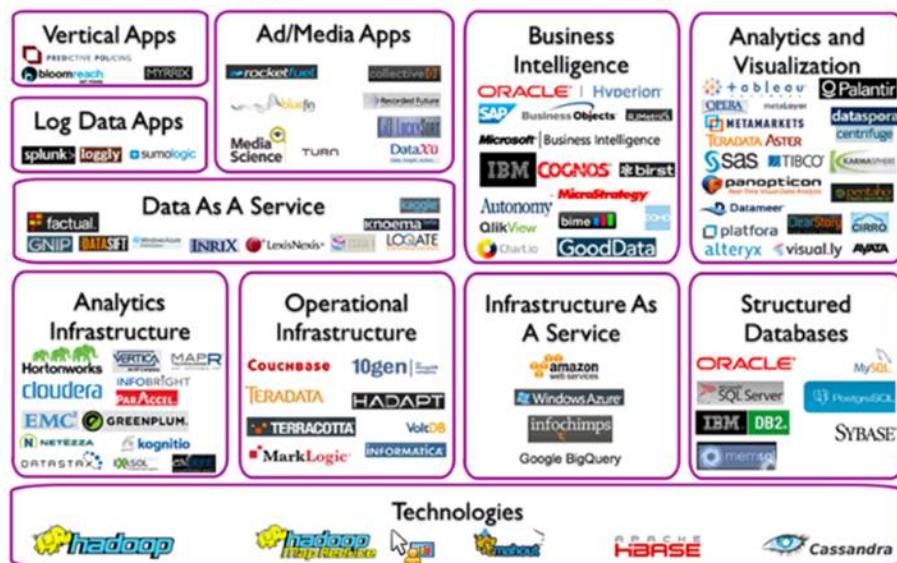
For different organisations big data is a goldmine. Actors should use it and understand it efficiently like a huge resource for the economy and society.

The using of big data to be efficient involved a strong ecosystem. In this schema, the different parts of the process are linked, and help to the evolution of the next part.

Big Data Value Ecosystem
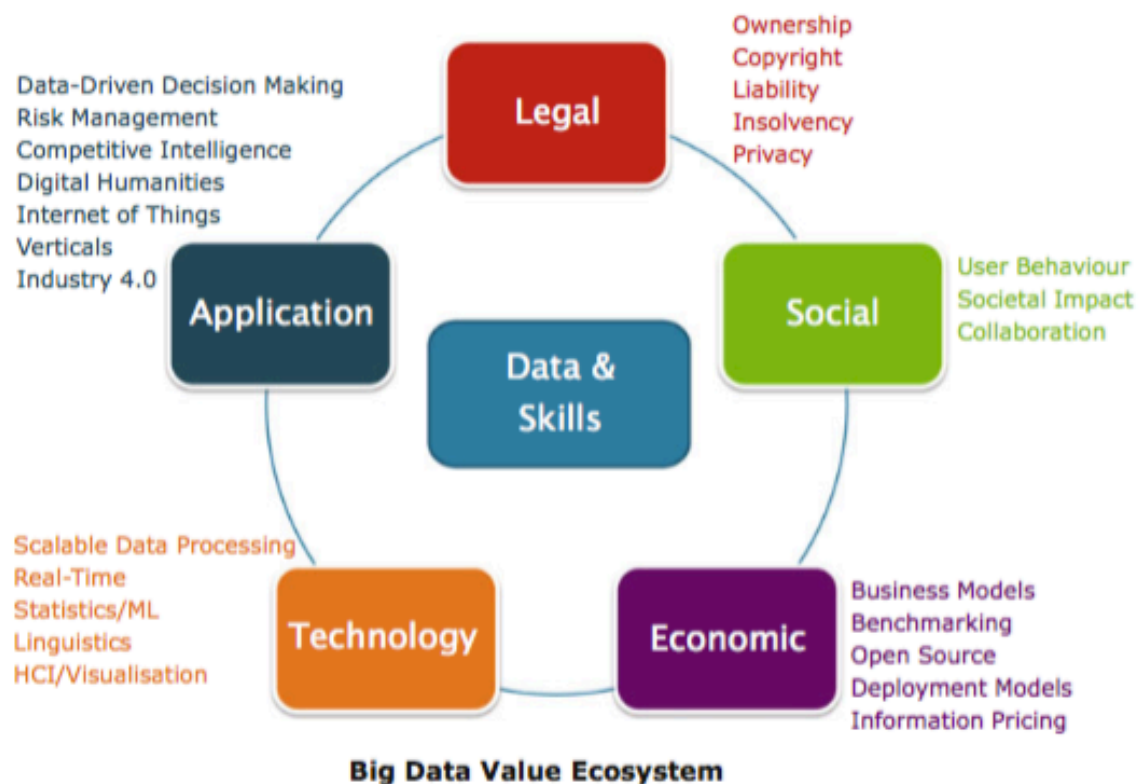
This ecosystem is focused on data, the availability and access to this resource is the fundament of this process. As we said previously, the data have many forms, are structured or not, from many sectors.

The added value can be on matching different data, on acquiring data and exchange it. It permits to improve the data quality, assure the data validity.

It also focused on skills; it's the big challenge, having skilled workers to follow the needs expressed by the big data. This ecosystem will need data scientist and engineers who are expert on machine learning, data mining, statistics and analytics.

Around this central area, we have the legal, technical, application, business and social environment.

The legal framework must be clear concerning the big data, to help its evolution. Some details need to be specified about use, privacy, protection, data ownership, security, intellectual property rights and cybercrime.

Concerning the technical part, the scientific research must overcome these challenges. The technology will evolve, very fast, and the companies, which will invest on research and development, will be victorious.

Big data have the power to change the situation in many sector so innovative value-driven applications and solutions must be developed, validated, and delivered in the big data ecosystems. It can also help the economy and the business to face to the environment change. Indeed, this process help the business to evolve from an « old » form to a more flexible companies. More and more start-ups are setting up, with new business models.

In another part, big data will help companies, cities, and states to manage new social challenges such as healthcare efficiency or transparency in government.

The different stakeholders involved and/or impacted by the big data technology are:

Start-up

Technologie providers

Data suppliers

Data marketplace
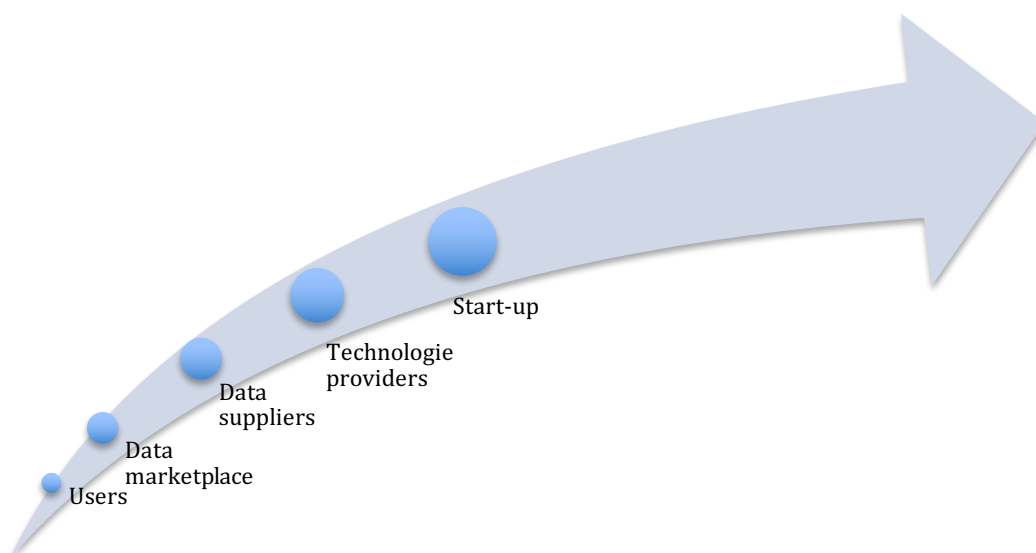
Users

• Users: person or organizations from different sectors that use big data technology and services to do business or to help them to improve their

process.

- • Data Marketplace: organization that host data and offer it to users.
- • Data Suppliers: person or organization that create, collect and transform data from public and private sources
- • Technology Providers: organizations as providers of tools, platforms, services, and know-how to manage the data
- • Entrepreneurs and start-up: they try to develop innovative data-driven technology, and products and services related to data.
- • Researchers and Academics: investigate new algorithms, technologies, methodologies, business models, and societal aspects needed to advance big data.
- • Investors and Incubators: person or organization that provides resources to develop the potential of big data in commercial term. These resources can be financial, services or infrastructures.
- • Regulators: like states, public organization to protect data privacy and build legal environment.
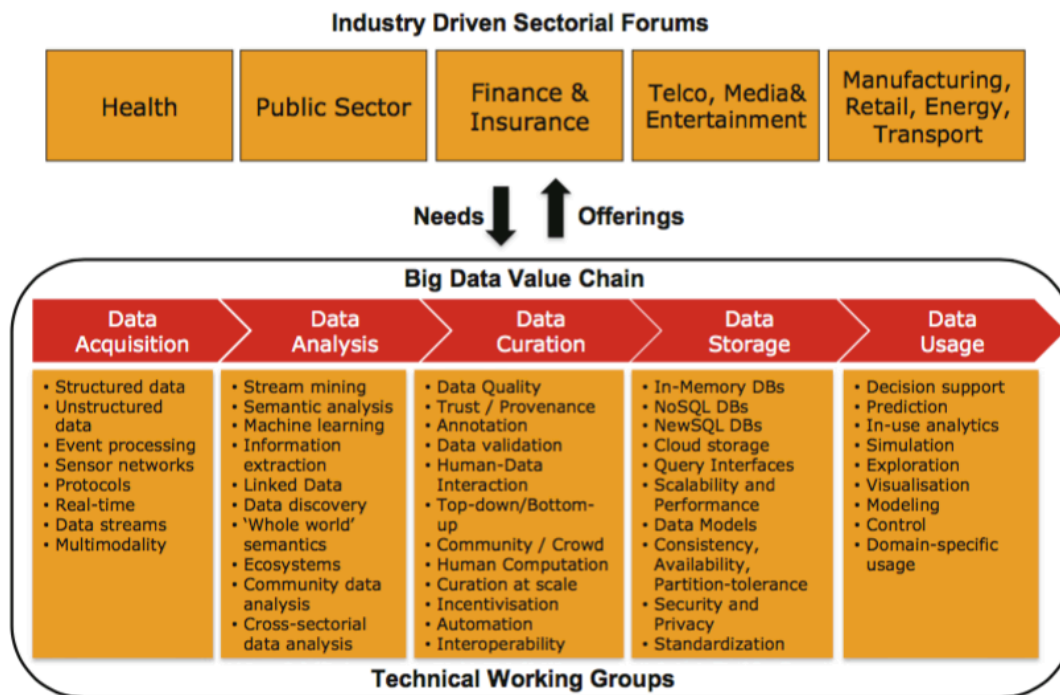
In this scheme we see the big data ecosystem with all the stakeholders. Main of the business actors and data provider are from the public sector like SNCF, Orange and Meteo France. It shows that public companies have a big place on the big data providing because all the people use this services.
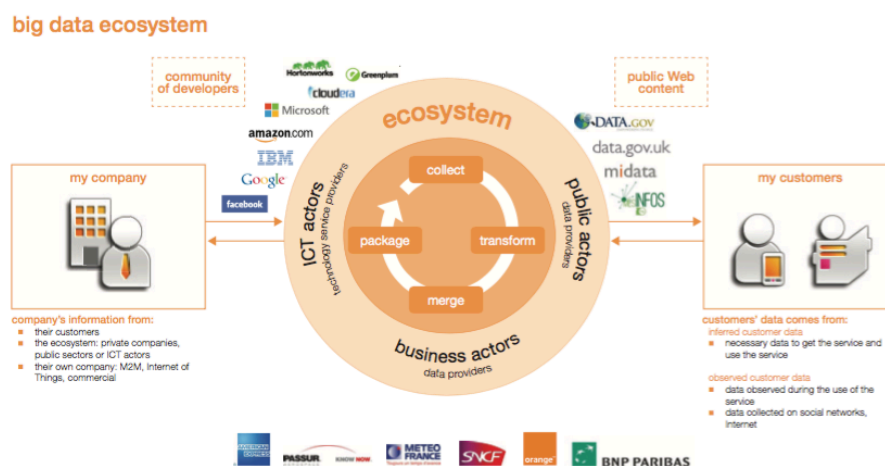
source: Orange Consulting

1.4- The big data extracting value:

Now, we will see the process of the creation of value with the big data value chain. On the top, we have the different sectors, public and private and the relation with the steps of the value creation, this value can be quantifiable in terms of money or not (for example for public organisation).

Industry Driven Sectorial Forums

Health	Public Sector	Finance & Insurance	Telco, Media& Entertainment	Manufacturing, Retail, Energy, Transport

Needs ↓ ↑ Offerings

Big Data Value Chain

Data Acquisition	Data Analysis	Data Curation	Data Storage	Data Usage
• Structured data • Unstructured data • Event processing • Sensor networks • Protocols • Real-time • Data streams • Multimodality	• Stream mining • Semantic analysis • Machine learning • Information extraction • Linked Data • Data discovery • 'Whole world' semantics • Ecosystems • Community data analysis • Cross-sectorial data analysis	• Data Quality • Trust / Provenance • Annotation • Data validation • Human-Data Interaction • Top-down/Bottom-up • Community / Crowd • Human Computation • Curation at scale • Incentivisation • Automation • Interoperability	• In-Memory DBs • NoSQL DBs • NewSQL DBs • Cloud storage • Query Interfaces • Scalability and Performance • Data Models • Consistency, Availability, Partition-tolerance • Security and Privacy • Standardization	• Decision support • Prediction • In-use analytics • Simulation • Exploration • Visualisation • Modeling • Control • Domain-specific usage

Technical Working Groups

This modelization was made by a group of work during the BIG: Big Data Public Private forum for the European strategy for the big data.

5 parts divide the value chain. First part, the acquisition, basic step, buy or collect information from clients, citizen, and users.

After, we have the data analysis, companies will analyse this data with data mining (explore data) try to see the big tendencies, link the data between them.

Then, the data curating, we had our big tendencies, now the main purpose is to evaluate the quality of data, is it trustable or not.

The collecting and analysis part finished, the question of storage is put forward. Which server should we use, which one is the most efficient, has the best value for money, and above all the most secure.

The end of the big data value chain is the data usage, the most important part. It's the moment to predict future, modelize, and make decision with this data. Using big data permit organisation to have a larger and longer view of their actual and future strategy.

There is a two-way exchange, on the one hand organisations express their needs and the technology try to follow this needs creating new tools to manage data and analyse it.

On the hand, technology evolves and organisations follow evolutions and use these methods in the company management.

Companies for many issues use big data, we will see how small, big or e-commerce companies apprehend this data management.

To make a successful strategy that creates value, organisations should know how to:

- Know the data kind the company have and could have, the ecosystem and the potential partnerships => observe and assess data

- Try to understand the way to analyze data to create value and opportunities => identify potential revenue

- Be aware of the operational, business, legal and technological challenges => recognize and mitigate risks.

The using in big or small companies is quite the same.

One survey based on 1100 answers from workers from 26 different sectors in 95 countries by IBM Institute for Business Value and the Business School of Oxford university analysed how companies use big data to develop their activities and increase their competitiveness.

1.5- Big data in marketing and sales:

We can ask ourselves if big data is only used in marketing and sales. Certainly these tools represent on priority a sales issue. Organisations (public or

private) try to improve the user experience and user interface, to know the client preferences and behaviours.

The relation with clients began with mass marketing, companies didn't really care about the type of customer. Then there was the direct marketing, it conduced to the customer relation management. Nowadays companies should deal with the user experience and user interface.

We have pure players companies, only on Internet and brick and mortar, traditional companies.

Even if traditional companies don't really use Internet for sales, they use new technologies and big data, with fidelity card for example. They analyse the basket of each consumer, count the number of passage and match all those information to analyse it and give predictions. And sometimes, companies combine the Internet interface and physical shop.

Companies give discount with its app, because it's interesting for them to have all the phone information from user.

Today, everyone have all his information on his phone, like bank app, administrative documents, transport application etc… We do more and more think with our smartphone, calling obviously but we also pay, have all our fidelity card, traffic apps and more.

For each application the smartphone is a goldmine, they can find the localisation, the social media habits and they can give to us a "better" advertising. It's very surprising when you enter in a supermarket or a mall and you receive a message on your phone with a new discount.

The analysis and the decision of sending a message is not based on your age, sex etc, the central part of using big data is your place, all the information about you and how companies can give the better advertising on the right place and right time.

For e-commerce website, the change is that the owner don't really choose his target. Obviously you have a target at the beginning, but you will analyse the

traffic on your web page and try to have a fine analyse to know your target. The target is not like the old pattern. It's not about age, sex, socio-professional category (SPC), etc...

As of now, websites try to emerge the type of visitors and customer's behaviours to focus on.

For instance, they will analyse:

Elements of analysis	Why?
Number of visit	Fix objective of visit per days and month
How many time the user stay in one page	It permits to highlight which page keep the attention of the user and the time of reflexion of our future client
More popular page (more visited)	Can help the website to know which product are most popular and the most attracting
Consulted product	The company should know what is the client path on the website
Purchased product	To know what is the average basket
Hours of visit	To adapt the advertising
Provenance of the user/client (website)	To adapt the advertising and focus this advertising in some websites
Provenance of the user/client (advertising)	It permits to measure the efficiency of advertising campaigns
Key word on request (search engine)	To know the main work that conduce to the website
Type and place of delivery	Adapt the offer

If we had 1000 visit per day, how can the company raise this level? Combined with the number of order, it also can help to have the transformation rate, how many visit conduce to an order.

The time that user stay in a page to know what keep the attention, and try to improve it, for example as a magazine, we can see which article is the most read and there we can improve the user experience to stay longer on other pages. But, warning, the most visited page is not the page where user stay longer, the popular page can be the loss leader.

Google launched Adwords in 2000, these advertising sales analyse the customer behaviours on Internet and outside Internet. When we go on Internet, we visit many website and search for some information or some goods; we have some application on our computers.

Google match all this information to buy some advertising insert. Then when we go to another website we have advertising linked to our profile and our previous visits on Internet.

The big data has also a big role in the transformation process of companies. With the mass data, the process of production, logistics evolves. Economic, social, technological, legal environment is unstable, it changes quicker and quicker so companies must to be flexible and adaptable to have a sustainable growth.

This data can be collected on different way:

data depending on the collection source

collection source	description	examples
Web	• user-generated content: social networks, file storage, pictures, videos and audio • e-commerce: transactional history and data from customers	• Facebook, Twitter, YouTube, LinkedIn, FlickR • Amazon, Ebay, etc.
machine to machine (M2M) and Internet of Things	• RFID, sensor networks, etc. • log (i.e., mobile, Internet, fixed) and call data records • NFC, QR codes, etc.	• product localization • location-based data, real-time traffic analysis • points of sale
commercial data	• CRM: customer's profile, behaviors, audio, log file data[2], etc.	• transactional history • loyalty programs: customer preferences
personal data	• legal identity, health records, legal information	• passport, medical file, criminal records
open data	• administration files, statistics, etc.	• data.gouv.fr • data-publica.com

source: Orange Consulting

Currently, two thirds of data sets are unstructured data, (dark data "contraire" to clean data that we can use directly). They are mostly present in CRM, in the Web (audio, videos, pictures, social media) or M2M (machine to machine)

As we saw previously, on the web, users generate content with social networks, files, comments, photos, etc also in e-commerce with the process before, in and after their buying. But data could be generated also machine to machine.

They are commercial data, it could be linked with a-commerce and traditional shops with the CRM (customer relation management), companies can analyze the customer's behaviors and give a better service with right offer in the right time and have a better strategy to predict futures behaviors and needs.
The difficulty is to in convince customers that they would get more specific offers or customized services by giving information to the company.

1.6- Big data and legislation:

The personal data are more sensible, in France we had CNIL, this organism gives instruction for the data protection. It's for example the name, surname, address and all information about legal identity, health.

CNIL (national commission for informatics and freedom) is the regulator of personal data. This organization helps professional actors to follow the rules and citizens to manage their personal data. It's an independent administrative authority. This commission is composed by 18 members 4 parliamentarians ,6 members of high jurisdiction, 2 members of the economic and social council, etc.

The CNIL control and sanction companies. 3 types of control exist: on place, on line, on hearing.

Controls are made after plaints and also because of the big contemporary thematic.

During the control the CNIL can:

- Access to all the professional building

- Ask for any document and take copies

- Take all the useful information and hear any person

- Access to the informatics program and data

After the control, the CNIL president decides the formal notice. The sanction are pecuniary for an amount of maximum 3 million euros and the commission can also "imposer" the data treatment stop by the company.

Organizations must follow these rules from the informatics and freedom law:

- Define a purpose of this data collection to limit how organization can use the data

- Data relevance, if an organization really need this information or not

- Data preservation, how the data will be preserved by whom

- Data rights, user must be informed conforming to their right that their data will be collected.

- Data security, organization must guarantee the data security

The RGPD (general rules for data protection) will be launched the 25 may 2018. These rules will organize the data protection. It obliges all organization European or not to follow this legislation. Indeed, even if an organization is not based in Europe, from the moment when it exchanges with European citizen, it must follow some disposition.

Before the set-up of the RGPD, companies followed the CNIL disposition. Now the RGPD take most of the CNIL disposition and add more dispositions, the RGPD is harder.

If the public sector wants to take the big data road, public organizations must follow these dispositions. It's one of the big issues of using big data.

All these companies must follow these principles:

-        Have a responsible of data. All companies should have a referent for the data protection and data keeping plan.

-        Data protection and keeping plan. All companies must have a document where all the data protection and data keeping, how the data is stored, by whom etc…

-        Client right. All the client data or supplier data can be modified if the user asks for it. The data should be kept only for a defined period, for a defined reason. The user can also have access to his data.

The first step of the sanctions is to inform by writing the company, and there are different sanctions:

-        To 10 million euros or 2% of the turnover, for the default like the non conformity of the DPO (data protection officier) designation rules, non conformity to the register of handling or lake of notification in case of data hacking.

-        To 20 million euros or 4% of the world turnover of the company in case of non-conformity to the respect of personal data protection principles, to the rules of data transfer outside European union and to the consent rules.

<u>1.7- Public sector: how to use big data?</u>

What is the public sector? The public sector includes, government, administration, health insurance, public organization, semipublic organization, it also includes some private actors with public interest like transport, health, etc…. "The term "public sector" refers to all economic or social activities carried out under the total or partial control of the State and local authorities." (Definition from latoupie.com an official website for economics definition).

A survey of Eurostat (an European organism for statistics inside Europe) shows that French people use a lot internet for their administrative "demarches" for example to contact public administration 63% of French use internet, it's more than Germany and Grand Britain. They are 42% to send completed forms

to the administration.

	UE 28	France	Allemagne	GBR	Suède
Avoir des contacts avec les pouvoirs publics	46	63	53	49	73
Obtenir des informations à partir des sites web publics	40	44	52	37	69
Télécharger des formulaires officiels	28	35	32	27	42
Renvoyer des formulaires remplis	26	42	17	32	45

Champ : individus âgés de 16 à 74 ans.
Source : Eurostat, enquête communautaire sur les TIC (particuliers) 2015.

The Society is changing; the millennials and digital natives are more important in the society and have more power of decision. The sharing economy take more place, with Blablacar, Airbnb etc… we use more our smartphones and computer with this sharing economy. The baby boomers use more and more new technologies, because they use smartphones, laptop but also in their flat with the automation in houses, it uses big data to analyze the babyboomer behavior.

An organization public or private must deal with a new generation of user, today a large part of citizen does most of their administrative think on internet,

they are digital customer and digital citizen. They also communicate more and more they found a digital community, it's the open data, people share a lot of information. Obviously cities, organization and regions become digital, to follow the "customer" needs and evaluate.

It's in two ways, in one part organizations, cities, states build the new use and way to use digital.

In another part, the customer want to have a better user experience and user interface. Big data help to do it because it analyzes the behavior on internet without age, sex, localization central consideration. With this digital policy, we have digital employees, they want to do their tax declaration on Internet, public organization must be more flexible because of their work time.

It's an evolution of the way of using Internet in public sector but it's also the consequence and the origin of the money saving by public organization. The big data can help to save money, because we know where the money will go, if it's efficient, if we have the right worker on the right place. It's also a willingness of the European public organization.

They have less and less budget, they should save money and dematerialize the main part of their process help them to do it.

This transformation was in parallel with the structure changes. The NOTRe law in France for the local mutation helps this change, it was a rejuvenation of old region to become more competitive and attractive. This sharing wants to save money and follow better the citizen needs.

The digitalisation was a regulatory recommendation. Indeed, some procedures must be dematerialized, with the positive implementation of two project PES and ACTE. PSE for de dematerialization of the public tender and ACTE for the payroll management. The Chorus project imposes the electronic "facture".

<div align="center">1.8- Big data include open data?</div>

Matched with the big data, the open data, one of the main issue of the modern government it permits to be more transparent, have more proximity with citizens. The French administration to do it with some websites like *data.gouv.fr* and all the public websites with the raw *.gouv.fr.*

The open data in public sector has 4 features, the degree of access, the machine readability, the cost and the rights.

The degree of access is who has access to the information, in open data, everyone has access. The machine readability, in open data, the information can be easily reached by everyone. The information is free, and concerning the rights, there is unlimited right to use this data.

	Completely open	**More liquid**	**Completely closed**
Degree of access	Everyone has access		Access to data is to a subset of individuals or organizations
Machine readability	Available in formats that can be easily retrieved and processed by computers		Data in formats not easily retrieved and processed by computers
Cost	No cost to obtain		Offered only at a significant fee
Rights	Unlimited rights to reuse and redistribute data		Re-use, republishing, or distribution of data is forbidden

SOURCE: McKinsey Global Institute analysis

But it could be difficult for organizations to collect data, we divide data on different degree of protection and data characteristics in the schema bellow.

1.9- Data characteristics, data collection:

The data characteristics depend of the identification, the timing, the sensitivity, the purpose, the process and the source. The identification is if we know to whom belong the information or not, if it's anonymous, identifiable (we can know to whom it belongs, but we don't know it automatically) and identified (each data is linked to someone a customer, a user etc…). Then we have the timing between the data creation and data collection. The data sensitivity, if the data are sensitive like personal data, credit card number for example. The purpose of this data collection. The process, how is done the collection, voluntary, the customer give himself the data, observed, it's collected thanks to the client behavior, inferred it's between voluntary and observed way, the organization get data with data voluntary given and the customer behavior. And the source from where data are collected.

The degree of protection is low, medium or high. If the information are anonymous, not in real time, non sensitive, neutral, the user give this information by himself and if the source is the ecosystem, this type of data don't need a lot of investment on the protection. In contrast if it's identified, in real time, highly sensitive, transferable to a third party (it can be sold or companies can make an exchange of data), inferred and from the customers organization must protect this data and invest on this protection.

Glossary :
- Identification : capacity to identify to whom the information belong
- Timing : time between data creation and collection
- Purpose : reason for which data is collected
- Process : collection method used
- Source : place where data is collected

degree of protection for key data characteristics			
data characteristics			degree of protection
1) identification	anonymous	identifiable	identified
2) timing	non real time	near real time	real time
3) sensitivity	non sensitive	sensitive	highly sensitive
4) purpose	neutral	commercial	transferable to a third party
5) process	voluntary	observed	inferred
6) source	ecosystem	CRM	customers
	low	medium	high

Different government tries to be leader on the big data market. In Europe, many forums are taking place, and they try to have a European coordination. In November 2017, the European Big Data Value Forum take place in Paris, this event is cooperation between European Data Forum and Big Data Value Association Summit.

The European Data Forum is an event where European researchers, scientist, politics, industrial meet to talk about economic opportunities to use big data and the innovation with data in Europe. The Big Data Value Association Summit is more for private sector, many companies take part to this event to talk about the big data and its economic issues.

The European Big Data Value Forum permit to sign contract and elaborate program on research and innovation and elaborate new politics in Europe.

The reaching of leadership is for two raisons. Firstly, to help companies and have more competitive companies, secondly to implement this research on the public organizations.

1.10- Where to use big data in public sector:

A research of the McKinsey Global Institute2 shows that European public organizations in can earn 300 billion euros using big data technologies for administrative cost, it represents a decrease of 20%. The saving is due to a

higher efficiency and a better tax revenue collection. There are many way of using the big data in public sector:

- Know the citizen sentiment:

Public organizations can with big data have a large score of the public opinion. They are not able to analyze all the blogs, websites, social networks (twitter, facebook) with an human participation. It needs to use technology to manage a big amount of information, modelize it and analyze it. It permits to be informed of the social climate in the country.

- Open government:

This point could be related with the last one. The open data helps to have a better social climate. The citizens feel more involved in the public decision. The free-flow of information between public organizations and population make the relation better and people trust more government.

- Security:

To identify threats organization could look for financial information, actualities, social medias, blogs, social networks can help organizations to discover relation between illegal activities and sources of funding and transport of dangerous materials. This point take more and more place in the using of big data by the public sector because of the increasing of terrorism in Europe and in the world.

- Cyber Security:

Big data technologies can collect, organize by modeling and analyze a big amount of data from informatics networks to give to cyber defenders a bigger capacity to detect and fight malicious attacks.

- Fraud detection and prevention:

Social service and revenue collector can analyze and track the citizen activities, like income change or situation change. This tracking will help to highlight some abnormal behavioral patterns, they could show mistake or frauds.

- Economic analysis:

Making links between different data sources, analyst can help government to face the actual volatility and predict the future financial situation.

- Tax collections:

Public organization can use different type of information, formal or informal, structured or not to validate some information and if they have some doubt make audits.

The big data can also help public organizations for a better pilotage. Public organizations collect always data from their agent but they don't use it to take decisions and anticipate.

One of the most important is the human resources, the main expense is human resources, salaries in some organization it represent 60% of the total expenses. Big data can help to have a vision of needs in the future. With this vision, managers can predict how many people they need. This aspect concerns the training of the public sector workers, the evolution inside the public organizations, and the recruitment.

For example, for the public transport, if the company counts the number of user, their behaviors inside the station, the different travel, what kind of transport users take before, the company can develop his business and hire new workers. If we count the number of user in a bus by hours and how many people take a ticket. We can hire an amount of inspector, this information matched with the population growth can give the future number of metro/train/bus conductors. It also permits to help the worker mobility, if we have to much worker in one area we can propose to them to move in another service or another city.

The big data give a strategic dimension to the human resources.

The bid data can also win the worker's loyalty. Human resources director know what kind of job advertising is most popular, have a better vision of the future needs

<u>1.11- Research and public sector which relation?</u>

The following schema shows the process of the research for the BIG Big Data Public Private forum for the European strategy for the big data but it could be used to understand the general situation.
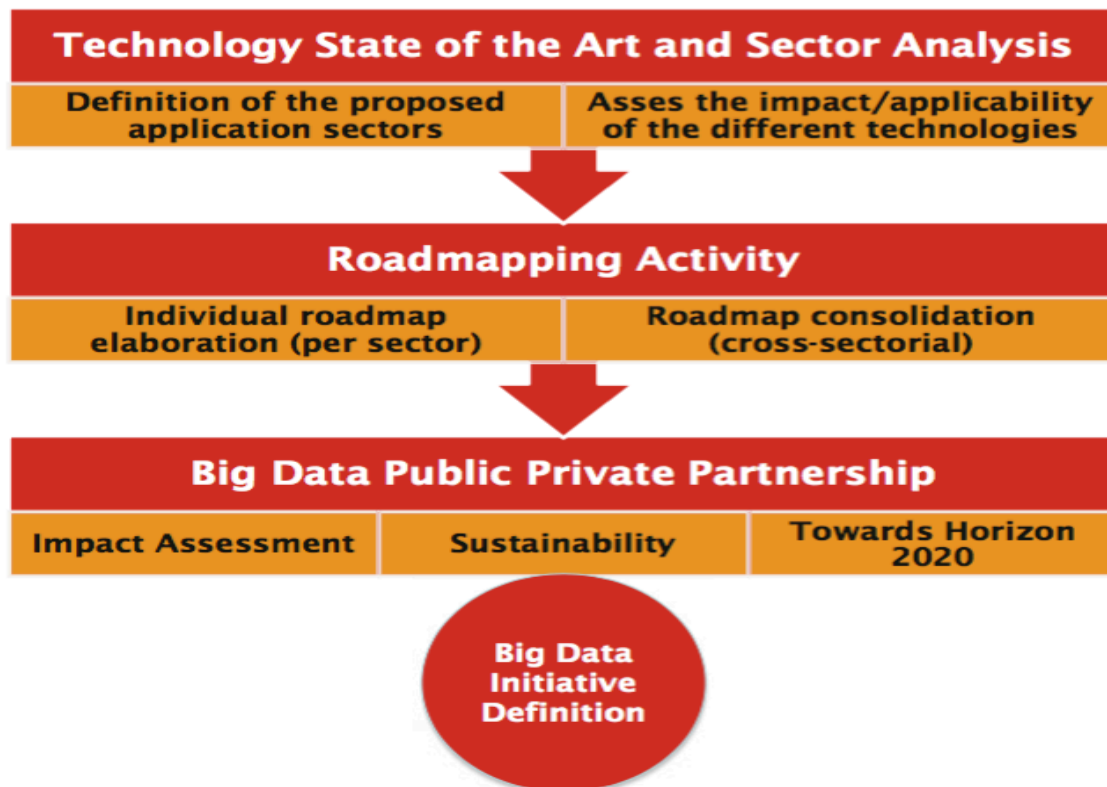
The technology state of the art tries to explain what is actually on the market, the different technologies in place, the actual researches, the researches planned, the budget voted concerning the big data. These explanations are made with different stakeholders this sample represent what is in place actually. The sector analysis part is an analysis of each sectors to highlight how can impact the using of this technologies and their applicability.

The result of this examination is to know if technology is ready for the sectors needs and the matches between needs and available technology it will be a base to know was is possible in the future.

Then the next part was the road mapping activity per sector, what is needed by companies in term of technology and help. Then these needs were consolidated to have a big map including all the sectors. During this part many

interviews and survey with stakeholders will be conducted, it will help to have an homogenize demand.

Private sector and public sector have many differences but they can help each other. The public sector can help the private sector by their investment in research for example, and companies from public sector can help the research in big data like IBM.

The partnership concerning big data technologies between public and private sector should take into account the impact of this partnership, the sustainability it should have a long-term impact.

The purpose of this type of forum and research is to conclude by a clear definition of the future initiative in big data to help companies to be more competitive and to have a better long-term vision.

The public sector including administration and public organisation can be pioneer on this "market". Indeed, they have a huge database with many-detailed information. They have the raw materials to implement and improve the big data use.

<u>1.12- What is the people thinking?</u>

We did a survey to know how people think the big data and their perspective of the bid data using in public sector. For 93,2% of the sponsor the big data means only a huge amount of data.

Qu'est ce que vous évoque le terme big data?

45 réponses

Surveillance	7 (15,6 %)
Gros volume de données	42 (93,3 %)
Nouvelles technologies	18 (40 %)
Segmentation, enjeux commerciaux et gou…	1 (2,2 %)
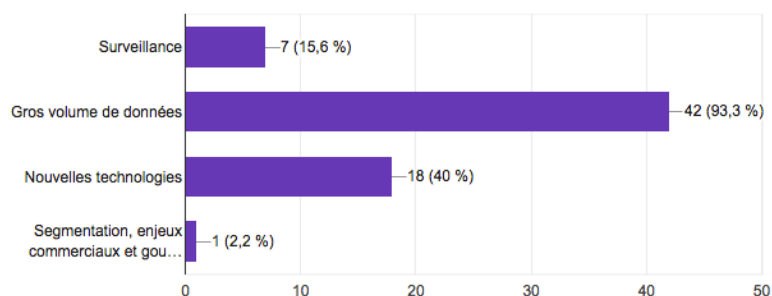
They meet big data on internet on e-shops websites and social medias.77,3% of them know that they don't have power on their data (on internet or in real life). For most of them big data is not a good think and it's not a very bad think, it's not a threat for them, but it doesn't represent an opportunity for them.

Le big data est-il une bonne chose pour vous?
45 réponses
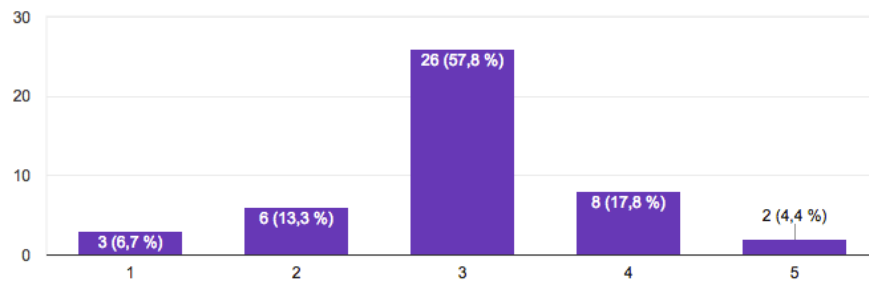
40,09% of the sponsor think that the main role of State is to ensure freedom of citizens and for 43,2% it's to ensure the security. For Only 9% the state has a social role. When we ask them about the budget optimization, they think to the right budget on the right place. 81,8% of the sponsor answer that the big data can help this budget optimization and the public sector transformation.

Q7- Do you think that big data can help this optimization ?
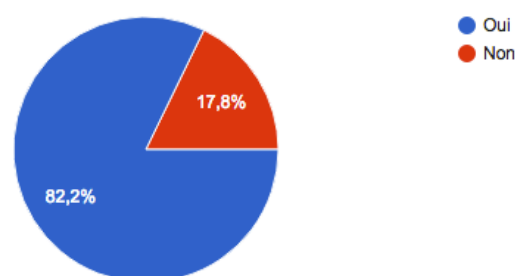
The main idea is that the big data can help this optimization mostly for the fraud fighting and criminality fighting. Then there are to help citizen in their quotidian life. A paradox is relevant, because the main idea in public organization is to fight frauds but mostly to help citizen to find their right, for

example in France, each year, 12 billion euros of social allocation are not reclaimed.

The other main idea is to fight criminality (in real life and cyber criminality) with big data.

| Fight frauds | Help citizen to find their rights | Fight criminality | Have a better user experience |
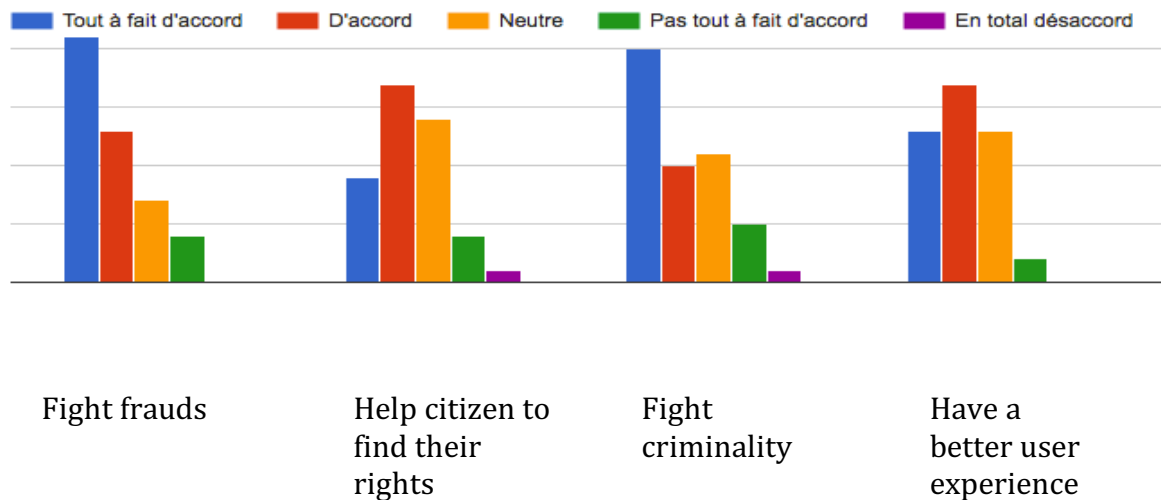
For a big part of people (77,3%) in the survey the States must invest in big data, because they have a big amount of data and should deal with to have a better policy and also because like in medical research, the state should invest money on the research to protect the citizens and the companies.

Q10- Do you think that states shoulg invest in big data research?

If Government doesn't help companies to manage the digital transformation and big data using, companies will lose their comparative advantage comparing to foreign companies. The data protection takes more and

more space in the citizen concern. The RGPD is one of the consequences of this concern.

But the most popular answer is that the public sector should invest in big data research for an internal use.

Section 2. Big data in public sector : examples :

1- Big data in social insurance :

At a national level, we have many using of big data, in social organization, public health insurance or transport. We will explain some example and try to understand the big issues of this big data using.

The 138 article of the 2006 law and the article 114 of the social security code launched the creation of huge files between all the person which are registered to the social security the national protection social register. It regroups the number of registration in this register, the data from civil status, which allocation the person have, addresses.

It's a available portal by the social security agents but with some conditions.

It's to give a better service, to help people in there administrative procedures, have a better control in the social payment, simplify and rationalize the data sharing, help new social beneficiaries to find their rights.

For a better security all these data (very delicate) the consultation is in real time.

It permits to inject directly the information, documents, official documents, to avoid to do the same action many times.

In Belgium, the BCSS (Banque Carrefour de la sécurité sociale; bank for social security) was established in January 1990. It permits to share all the information from the social security institution, the user and companies to simplify procedures and reduce cost. They launched an unique identification, for all the social security organizations, the NISS.

Each organization continue to manage its information, the only change is the sharing and the standardization of these information.

Unemployed and person who can't work receive automatically their allocation without any demand. Some other categories have automatics allocation with a reduction because of taxes without any documentary proof. It's

easy for them and for the public organization because they gain time to manage these data.

It also helps to know the population and the population needs for politics. To begin our analysis, we will analyze the CNAF (Caisse Nationale d'Allocation Familiale), it's a French public organization created in 1945 to help families and people to get the legal minimum to leave, this organization has a social responsibility. The CNAF has many sub-offices, it's named CAF.

In 2016, the CNAF discovered 43 000 frauds, 8% more than the previous year, it represents 275 million euros. But this increase is not due to a fraud rise but the CAF has a better system to find frauds. They use more technology in particular the data mining. The data mining refers to the different methods that permit to analyze a big amount of data and give significant in order to analyze similar data through these models.

Between 2011 and 2013, the on place controls performances made by the CAF increase of 46% thanks to targeting methods based on data mining. In 2011, on place controls permitted to detect 150 million euros of fraud and 263 million in 2013. This increase is explained by the deployment of the data mining tools since 2004 in the CNAF network.

From 2004, Dijon CAF, then, Bordeaux CAF interest themselves to the data mining methods through their accounting agency to fill some fraud fighting deficiencies.

These controls try to discover frauds concerning housing grants, so it must be fast and efficient. Housing grants is based on user resources, heritage, familial situation, etc… these aspects can change very fast the speed of these methods is essential. These CAF took few thousand of user files to base the models, it permits to highlight the first fraudulent profiles. These profiles were

More than an efficient control, it reduces the unfocused control number and the users feeling that they are controlled for nothing or for bad reason.

The CAF don't target a priori risked population but profiles based on risked files on more objective criteria. It reduces the human impact present in traditional controls.

But some right defender, highlight the fact that one of the risks of this method is that data mining can conduce to discrimination, stigmatization of these risked profiles.

The target is to fight frauds but also to manage all the human risks like mistakes. The first work group in 2004 shows that some technician because the legal environment changes a lot and is more and more complex. More than the quantitative aspect, the CAF try to have a quality target, give a better service to the users.

The big data technologies like machine learning in the future can help the predictive logics. The more the risks identification is known early the more the risks gone happen.
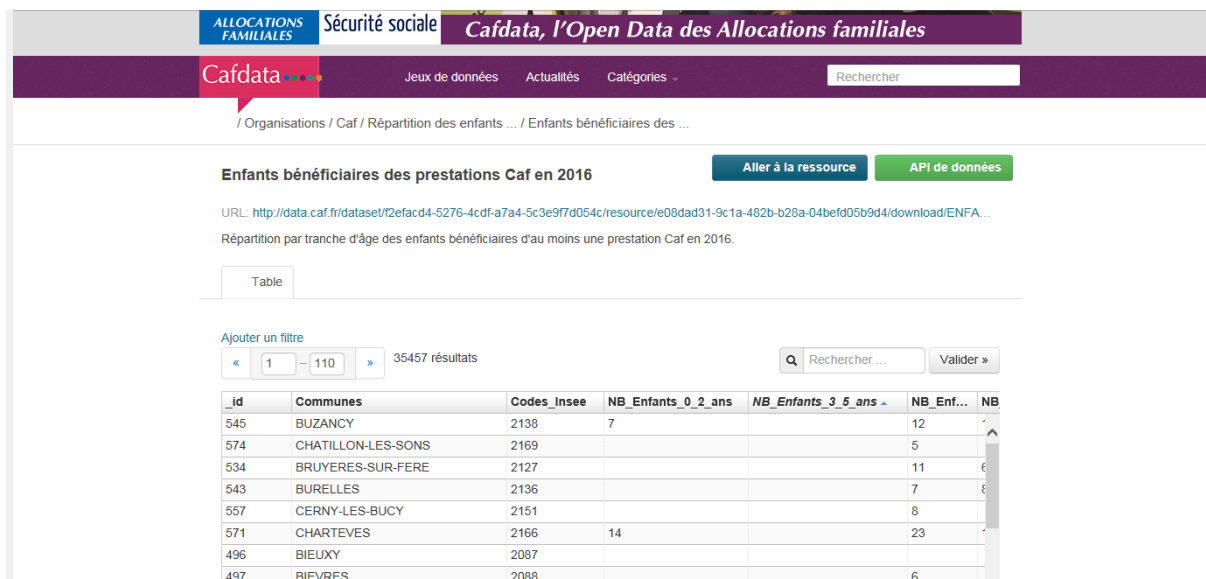
Another part of the big data using is to help the communication targeting and the right targeting, the good allocation to the good user, as we said previously, most of the allocation are not taken by users.

The big data also conduce to open data (the public information, data, it can be used freely and redistributed by anyone). The CAF have a special website for open data. It gives information and access to many data to the users.
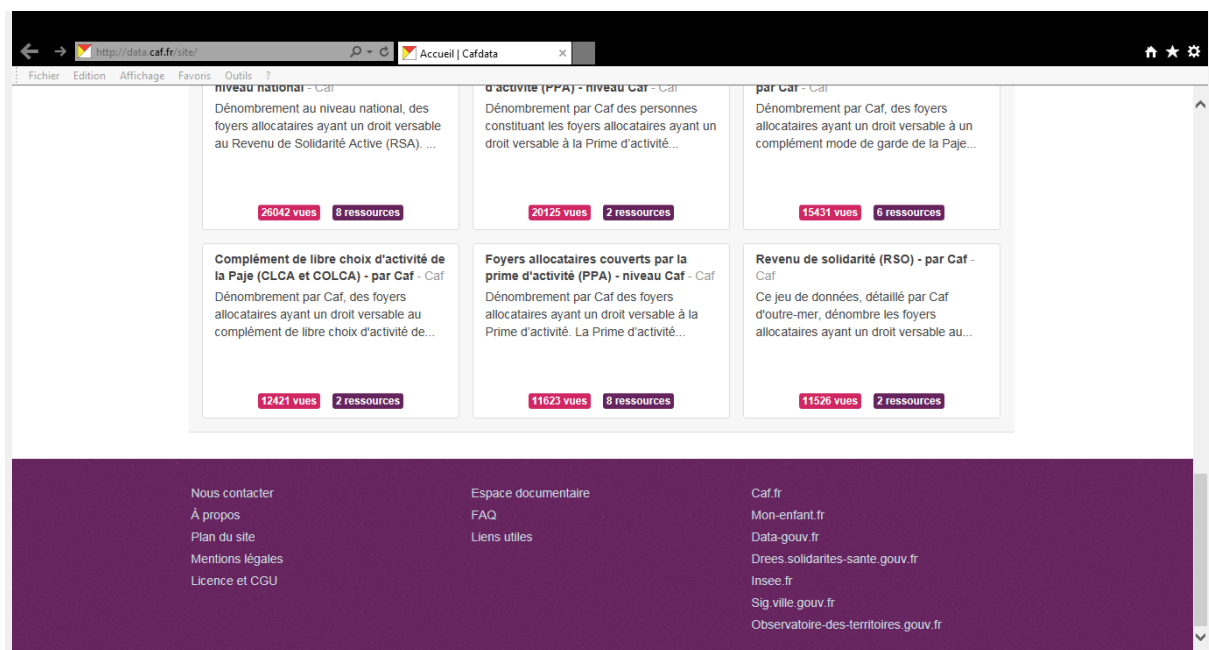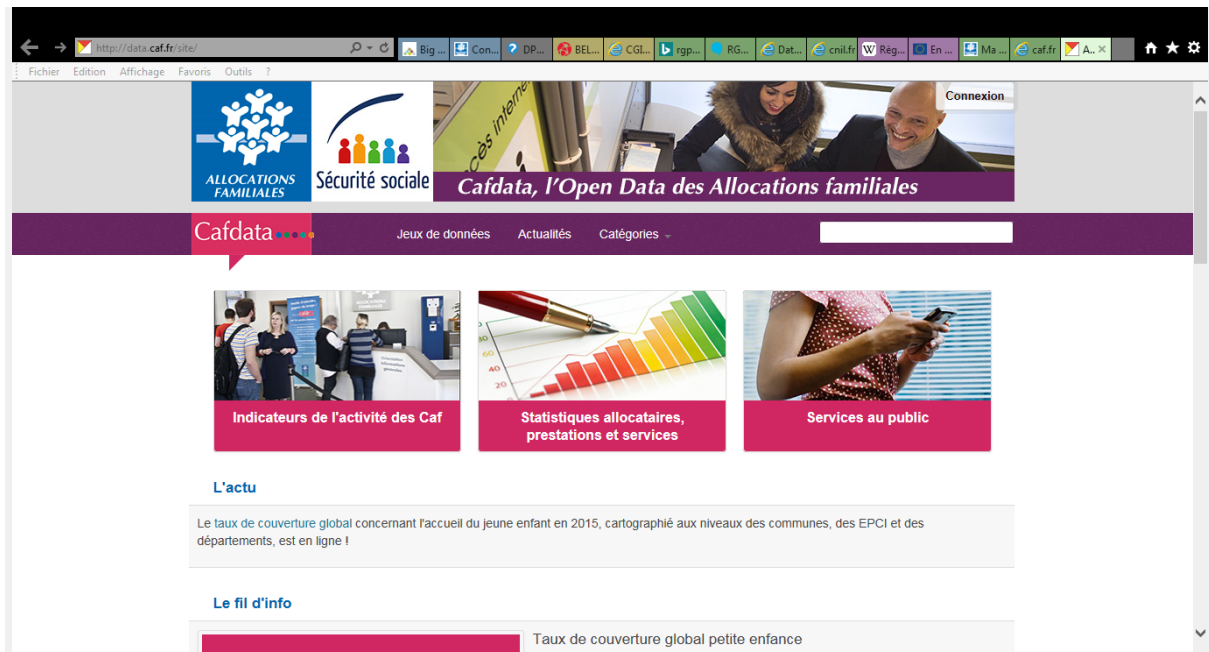
In the screen shots we have all the data and some analysis about the user's rights. For example, how many children have a minimum of one allocation, the matching of different allocations like the number of users with a minimum allocation in overseas departments, we can see mapping of the different type of allocation in different region.

There, we have the number of children with at minimum one allocation by age and by localization.

In France like the CNAF, the CPAM (caisse primaire d'assurance maladie), the public health insurance. Has also an open data strategy. We can see in the screen the website. Different survey about medicines, rights, allocations are public.
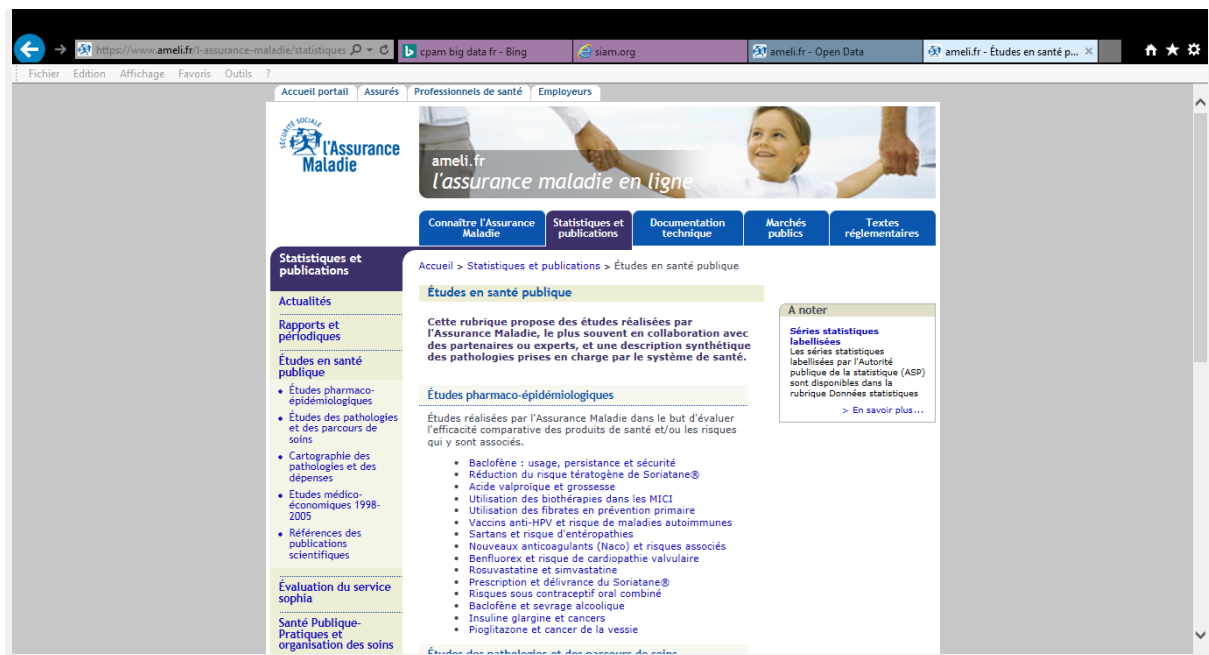
In another part, polytechnics school in France and the national health insurance for employees work together to develop new trail to exploit the data of the SNIRA (Système National Inter-Régimes de l'Assurance Maladie, it's the system which regroup all the information concerning the health insurance). This database is constituted of 65 billion people with all their relative data.

It's the biggest database in the world, and specifically in health sector. In 2014 this 2 organism signed a convention of research and development of 4 years to promote the technologic development of big data in the health sector.

This database includes all the data of health refund and hospitalization of all the French people who are registered in the obligatory health insurance system. The insurance organism in France takes these database for public health and pharmacovigilance in partnership with health authority. For example the first research about the thromboembolic risks of third generation of contraceptive pills was made with the SNIIRAM database. This partnership will made algorithms to help the CNAMPTS to improve the quality of its work and also help to deal with the public health issues.

This algorithm have tree research ways:

- To detect "anomalies" in pharmacoepidemiology
- To identify useful factors to have a better health path
- To fight frauds

<u>2.2- Big data in health sector:</u>

The big data can help to reduce health industry expense to 450 billion dollars.

But it's also a quality question:

- The right living: the patient takes more active act to improve health

- The right care: a coordinated developing approach with caregivers, they should have the same information

- The right provider: all professional must have strong performance records and have the best outcomes

- The right value: improving the care quality and the value simultaneously

- The right innovation: innovate on the health-care delivery, how to it, with which tools.

Today there are many reasons of the data growing in healthcare.

The standard medical practice is moving from relatively ad-hoc and subjective decision making to evidence-based healthcare, in the past, if a patient goes to see different doctors, he can have different diagnostics.

The professionals and hospitals are prompted to use EHR (electronic health record,) technologies. (EHR stores data in a structured format, this structured data allows patient information to be easily retrieved and transferred, and it allows the provider to use the EHR in ways that can help patient care).

They are more and more data:

- With the development of new technologies like sensors and mobile applications

- The genomic information became cheaper.

- The social communications of patient in digital increase

- The collection of knowledge and discoveries increase

In this scheme, we have the goals of big data. The big data analytics helps to manage the database with electronics health records, genomic, public health, behavioral to have a lowest cost and to improve outcomes, to make smarter decisions.
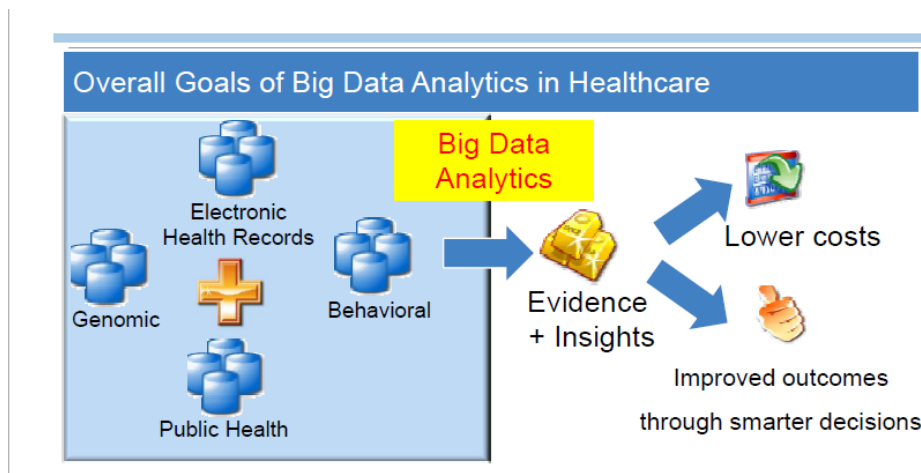
The massive data amounts helps to provide the right intervention to the right patient at the right time. (Personalize care to the patient and give the right information to the right stakeholder)

The issues are:

- Inferring knowledge from heterogeneous and complex patient sources (sometimes, patient says wrong information to doctors).

- Medical personnel and big data technologies should understand a clinical description in the right context.

- Big data technologies should be efficient when it is useful.

The big data health sector should capture the patient's behavioral data with sensors, taking into account some interactions and communications.

There we have an example of the process of data collection and analysis.

The electronics health records, records different type of data: Pharmacy, laboratory, radiology and narrative. These information are stored in HER database. This database will be used for patient, for national statistics and research.
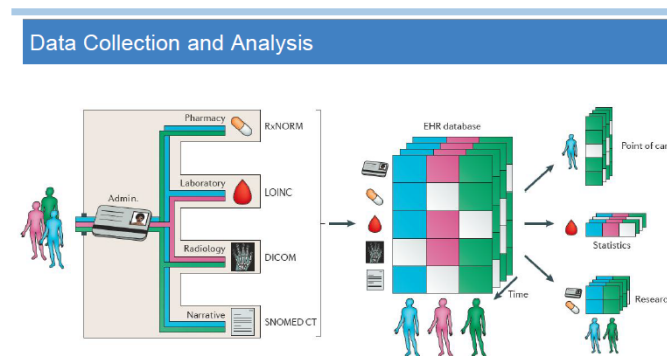
In France and Europe, the big data try to be implemented in the health sector.

Since 70' the hospitals used new information and communication technologies. Some survey highlight the fact that those type of technologies help to improve the healthcare quality and decreasing of medical error. For the hospital personnel nowadays it's impossible to take decision without this hospital information system. However for many actors it's a big investment in a budgetary restriction context. But they don't really prove their efficiency.

2.3- Big data in media sector:

The big data includes the mass media, this type of media gives a new issue for democraty.

The « observatoire transmédia » (OTmedia) have a research project maneged by l'INA (institut nationaal de l'audiovisuel, national institut of « audiovisual »). It's an opportunity for media but « surtout » for the public sector. The OTmedia between 2012 and 2015 captured an astronomical mass of data from 5 billions documents et sous différente forme, écrite, audio, video to analyse it. From 8 TV chanel, 9 radio chanel ans 1800 news websites, 20 000

tweeter account and the AFP (agence France presse, french press agency »

<p style="text-align:center">2.4- Smartcities :</p>

<p style="text-align:center">New cities thanks to big data : the case of transport :</p>

The rate of population in London « explose », 12% in the pas decade combined with the growing stain on the city's infrastructure, the Mayor in 2012 launched the 2020 vision report. It laid out the municipality strategy focused on crating Smart City initiatives with the technological expertise in the private sector. (with human and social capital and the municipal ability to coordinate with local firms. Different big axes was developped the need to transform the cities data ecosystem into a centralized and open data ecosystem contrary to the old patern decentralized and close, it was not easy to get some information. And take benefit from the data collection daily and analysis of huge quantity of data.

For many years transport for london (TFl) worked to incorpore the data strategy. It collects data from people using multiple services in this organisation. It did a partnership with Massachusetts Institute of Technology (mIT) to find new ways to use data and how to personalize information on services disturbances.

More recently, TfL has made big investments to improve its data management to allow its 517 full-time IT staff to take advantage of the high volume of information they are collecting.

In 2014, TfL selected services from the analytics Tibco to bring together data resources across the organization's multiple directorates. The goal was to create a centralized data infrastructure and improve the data collection and analytics process. Concurrently, it invested in SAP's in-memory analytics software HAnA to manage and improve decision-making in real time, allowing TfL to go from overnight processing to having the data processed almost immediately.

The SNCF include a lot of think, it's a travel organism, it sells hotels nights, fly tickets, train tickets, it's also the big actor of transport in France.

In the travel part, SNCF use the big data like a basic company, with :

- Google Adwords

- Microsoft AdCenter (Bing)

- Yahoo Search Marketing

They have banners like this :

They capt as we see before all the information of your way in internet, your localisation, etc… And give the « better » communication for the futur client.

But the main issues of big data for SNCF is to regulate the trafic and avoid the accidents.

For this SNCF takes information from user smartphone, has captors in train station. To give the better user experience and to know in which infrastructure it should invest in.

The SNCF with the captors count how many people pass in each area, how many paid their tickets, and which services they use. The sncf try to make a commercial place in its train station to attract new user and also to keep old users.

For the accidents, the sncf have captors in tain station and tout au long of the rails but also on social media. They are informed of an accident by the captors all along of the rail but also by social media. They have keywords and a robot analyse a majority of the social media content to detect some problems in rail transport.

To manage the flows SNCF have many axes :

-       Analyse the visitors flows : in big train station, the exploitation of the wifi data permit to know the visitors flow. Its possible to have analyze how many time people stay, the density during all day, the entrance and exit flows in this area, travelers or not.

This information helps to analyze the comportment of the visitors during some periods like train delay for example or events like sport competitions.

In term, it should naturaly help to optimize the planning of the rail station, with the safety and help to compare the differents train station.

-       Optimize the display in train station : this second point concern all the information given to travellers. The data will permit to see where sncf is performant or not, depending on the hour, the train type, the station type. For example the dock display is displayed 20 minutes befor the departure . We can ask why it's not before, the big data using will help to see this evolution and the consequences in term of travellers flows in real time.

-       Test and learn logic: The analytic make possible the logic test and learn. We can experiment new planning in station and observe quickly the results to correct them, refine and potentially generalize what we have learned.

This method what implement for the security gantries installation. Different formulas were tested to see which one will had the better effect on the flows. It also permits to evaluate the information dispositive during works, to see if information are quickly understandable by the travelers, etc…

For the tickets selling, SNCF use big data to give better way for a better price. 1st step: SNCF technologies's teams first set up a route search engine based on an intelligent price and availability cache system for offers distributed by Voyages-sncf.com, built on the basis of consultations performed by web and mobile clients.

This cache system is based on an analysis of the responses to the quotes requested by the customers (logs, constituting a sufficiently representative

sample to draw a statistical representation) and the application of predictive rules in addition (Machine Learning, rules designed and maintained by the Data Scientist).

Step 2: Once this cache system stabilized, the teams were able to work on other applications such as alternative routes.

For these paths, the setting of the algorithm is different; these are destinations where the first selection criterion is the price in front of the waiting criterion in the matches.

This allows the creation of new combinations and the appearance of trips at lower prices with a little longer transport time. Today, this concerns 10 destinations, usually very popular for the summer and often complete at this time:

-        Already available: Paris <> Arcachon, Lille <> Marseille, Lille <> Avignon, Lille <> Aix en Provence

- In the coming days: Paris <> Biarritz (via Bordeaux), Paris <> La Baule (via Nantes), Paris <> La Rochelle (via Poitiers), Paris <> Vannes (via Rennes), Paris <> Marseille (via Lyon), Paris <> Montpellier (via Lyon).

This dozen proposed destinations should at least triple in the coming months.

Alternative routes are indicated by an orange pictogram, here on the example of a Paris> Arcachon trip on July 15, 2017. The fare is much lower than the usual fares on this date:

Ministry of interior wants to install new radars in France, they try to find better "emplacement" to decrease the number of accidents. The database was on the statistics based on the accident and big data analysis from CGI where I did my trainee.
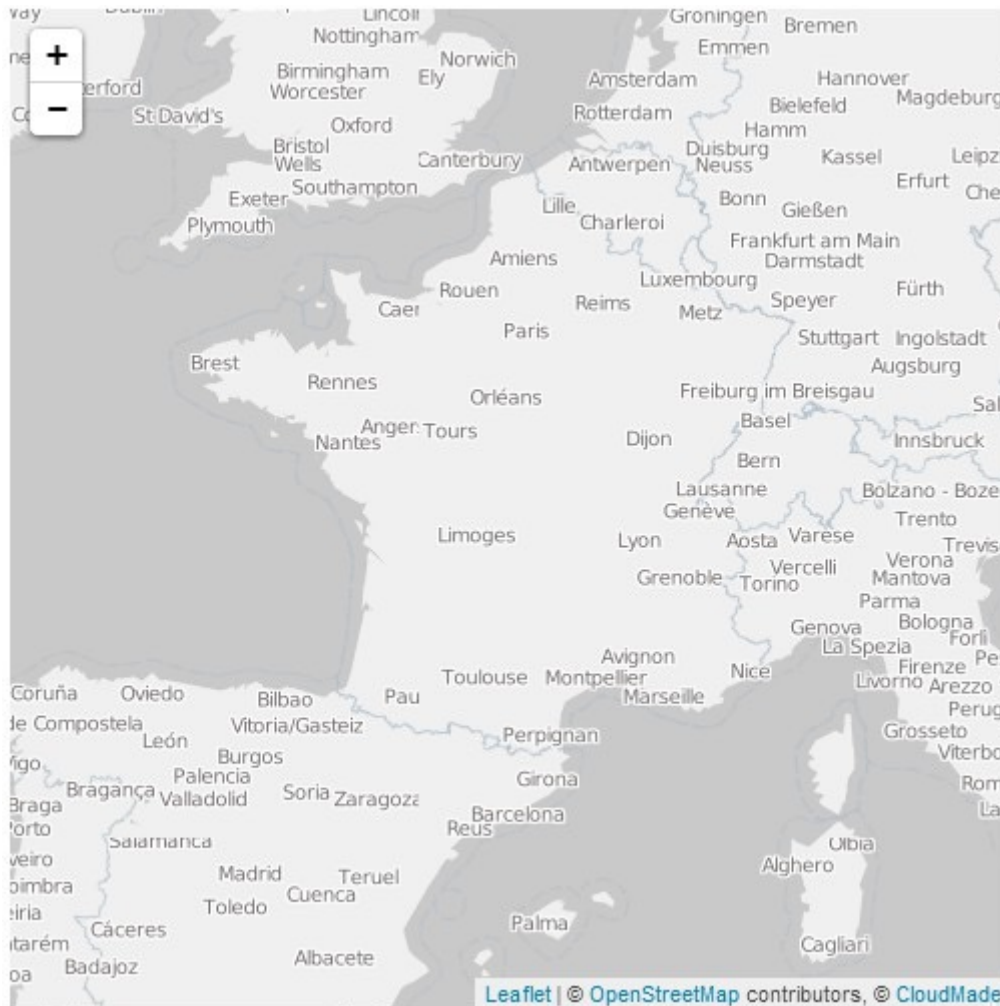
They took accident data in France during the last 5 years, the speed radars data and addresses in the national base addresses (BAN). For the accident it's on csv format (Exel, table), for the speed camera data it's data from open street map (OSM) and for the addresses data base, it's from the government website for each regions.

These data will be stored in databases, which is easily reachable by the future programs. It's stored in the search engine ElasticSearch. It indexes the different data, even if the format it's not the same.

After the indexation by ElasticSearch, this search engine is used and another tool named Kibana. Kibana permit to modelize on a map the data and give "visual" and graph.

ElasticSearch is a search engine open souce with an architecture REST (REpresentational State Transfer, it defines all the restraint on HTTP format).

The big principles of ElasticSearch are:

Data and analyses in real time, as soon as one data is included in this search engine, it did the analyses directly. All the data are stored in JSON document structures (it will be the interface between the different software). All the fields are indexed.

Free document structuration, any document is imposed, when a JSON document is added, ElasticSearch detect the data type, index it and this document will be searchable. Then, depending of the using domain, it could be interesting to structure these documents.

The research with usual methods, for example for a phone book, we will have:

- Name

- First name

- Number

They also use kibana, it's a plugin of data visualisation (opensource) for elasticsearch. It permit to make visuals with data in the indexation. We can also build a new plugin, for example in this case its « accidentologie »

Kibana has different features :

📽Discover : this part display all the documents indexed.

📽 Visualize : we can create or modify the visualizations (a graph or a map)

🎬Dashboard : regroup many visualization which can help to do analysis.
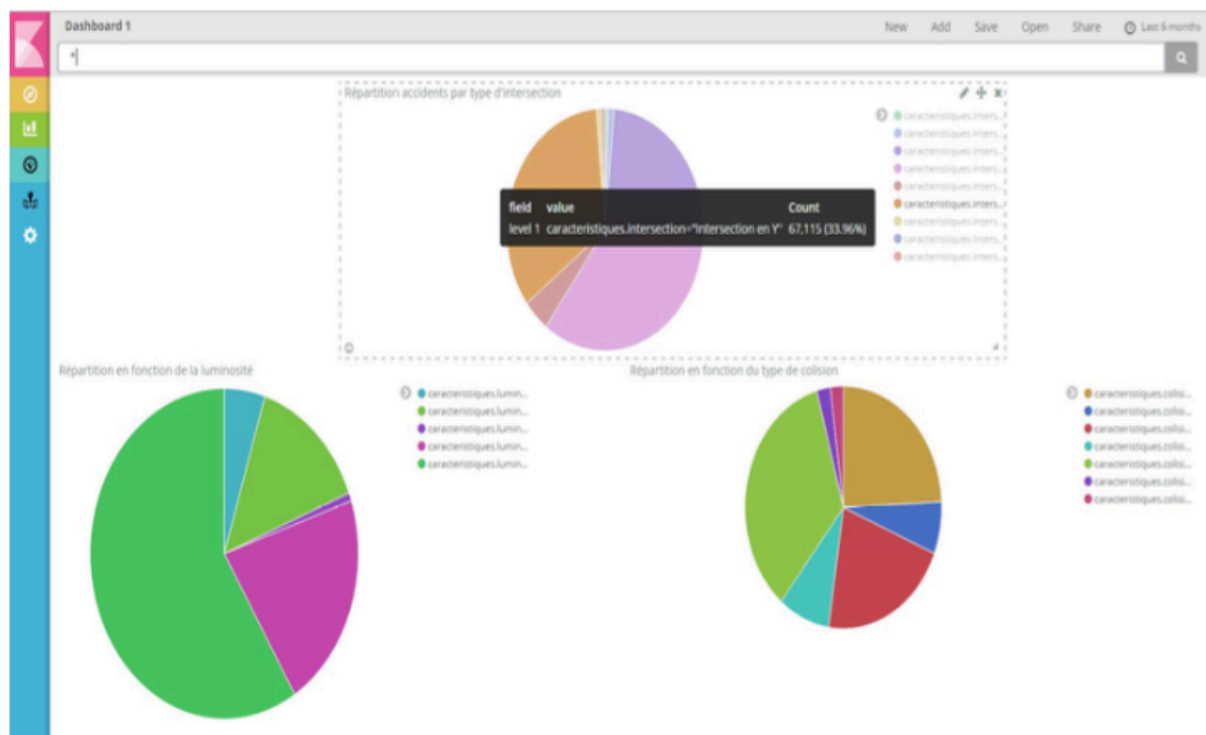
🎬Settings : it permit to acess to the parameters, there we can add a new index

The match between the road mapping and the accident with these big data technologies give to the minister to build "radar" to reduce accidents.

Section 3. The big data in public sector: future vision and preconisation

To conclude we will expose some preconisation about the use of big data in public sector, what should be done, what should be developed and how to overcome the gaps.

One of the main problems with using big data in public sector is that the return on investment is not easy to calculate. The results are more subjective and qualitative, like the user satisfaction, user experience.

In health it will be difficult to know the real difference for one patient between a decision took by an human and a decision by a machine. Public sector must reflect about quality indicators and financial indicators. The calculation of this return on investment integrating all the situation context will be more efficient and could give some evidence to continue to use big data.

The public sector must pay attention to public opinion, do not do the big brother, for the people, the state must guarantee them the security of their data and their freedom, that's why in France the organisation that control the Internet and numeric data is called CNIL (commission national informatique et des libertés, national commission informatics and freedom) the freedom aspect is very important. People care about their data and how they will be used (even if some times they don't really know where are going their data).

People want to use Internet in their administrative procedure but they are afraid. It's a huge challenge for public sector to change this population minding, to reassure them and show to them that's big data is a positive solution for them. It passes by teaching population and be more open.

Nowadays, the European cooperation is in a good way and they should work together. As we saw, in research this cooperation is very helpful. But the cooperation should be on the data sharing. Because of terrorism questions, Europe must share the data and use big data to analyze it.

Before an European cooperation, each country should improve the cooperation between all the public organization, for this it must have common way to store data and a common data management.

Sharing information in public administration is a very good idea but a big work should be made on the common norms, because for example what is the primary key, what king of data will identify one person.

More than this cooperation question, one of the barriers of big data using is the formation. Organizations must have the right person and retain it.

The public agent should also be trained to this big data, it cost a lot and take a lot of time. Public sector must deal with and try to have internal universities, of internal trainer to improve the value of using big data.

Training internal groups is expensive, actual methods of data mining involves high costs. It needs learning database from for example survey on person, which receive allocation.

The security is one of the big challenges of the public sector big data using.

Due to the interconnecting of the systems an informatics attack is more dangerous.

For example, in May 2017, the Britain system NHS (National Health Service) shows his vulnerability, because the information system was to old and not updated. 16 hospitals were attacked, a ransom was requested.

The telephone network, X-ray imaging services and the patient administrative management were touched. We can easily imagine if it happen in other organization like organization against terrorism or ministry of interior/

The information system, many of the computers, servers and so one must be changed and secure, but it represents a big cost and some public administration don't have this money to invest in new infrastructure. In a part, public administration must allocate a bigger budget to this information system

transformation. But it's also a challenge for suppliers to propose new infrastructures less expensive.

More than an infrastructure reform, public organization should change their internal culture. In Europe and particularly in France, administration becomes older and older. They should have a big reform to change this slow culture to a new culture more flexible, more open.

Even if the government invest in big data technologies, if the personnel doesn't want to us it's useless.

In addition, with this culture change, the head of this public administration must be flexible in their act but also in their way of thinking. If a data is available but nobody uses it, the return on investment will be null.

Keeping talent is very hard for public organizations because of budget, because off difficulties to manage in intern digital project. The reflection about salaries should be evaluated. Especially when people working in digital find very easily work.

For these reasons public sector prefers to outsource a big part of the big data management, to of informatics services providers it's a goldmine for many reason.

- Formation : they can help to form agents

- Change management: if the public sector need new information system, they also can "accompany" the change, help to express the needs, write the "specification"

- Provide new technologies: public sector will need more and more technologies and new infrastructure

- Provide "expert": to help this change, public sector need experts from these informatics services companies, because it's too expensive to hire them. These expert help to manage the big data project.

These companies should be very attentive to this market because it could represent a big amount of money. For example, if you win a public tender for the CAF you will be responsible of all the CAF in France.

Today, public administrations are linked with the government mandate and the president mandate. Even if the president and government don't have the power to manage the big data policy, they have indirectly power because they can allow public budget and the strategy is to reduce public expense. Public administration should be more independent and for example can have specific budget for each IT project.

European countries must also be careful of what is done in other world part. As we saw, the European union tries to have a positive dynamic in term of research. But saw what is done in other part, from other public organization can help the European public organization to be more competitive, to give a better user experience and a better service.

Section 4. Social responsibility:

I did my trainee in CGI, a Canadian company of informatics services. This company is divided in many business units. My trainee was in the business unit public sector. It permits me to see how new information systems and big data are used in this sector and the particularities of this sector.

Stakeholders:

Stakeholders	Comment
Shareholders	People who have shares
Employees	People who world in the company
Environmental organizations	Impact on the environment, action for sustainable development
Government	Relation with the legal part and the government
Citizens	Relation with the citizens
Health research	Investment on research

The CGI's internal social responsibilities:

The CGI culture is very friendly they have proximity between managers and employees. Employees are called members to highlight their implication.

They have an interesting company comity, who organize games like this year for the world cup, employees can bet on their team and win a telephone, a power bank or cinema tickets. Employees also have some subventions from CGI for holidays.

CGI helps women to have access to work. They are only 31% but CGI have a real policy to help women.

CGI try to give a good working condition, every year the air quality is tested and the filters for air conditioning are changed.

The CGI's external social responsibilities:

CGI try to be involved in its social responsibility by many ways.

In term of education, CGI is involves. CGI volunteers in France provide career advice to young job seekers.

This program is connected with CGI's Dream Connectors program, a company-wide initiative to support the involvement of CGI professionals in their communities.

CGI volunteers in France partnered with Rézo City to help advance its mission of providing career development tools and guidance to young job seekers. Rézo City is a program within France's Éduquer pour Réussir (Educate to Succeed) association.

CGI started by donating 10,000 euros, which was used to renovate the Rézo City office and provide much needed equipment such as computers. CGI volunteers were then trained as job coaches by Rézo City members and began coaching candidates.

In the first semester of CGI's participation in the program, more than half of the candidates found employment.

In addition to donating money and providing job coaching, CGI participated in "job meeting" events held by Rézo City with organizations that promote social integration and access to employment. During these meetings, CGI job coaches answered candidates' questions about jobs in IT and presented CGI's Developer School to Rézo City candidates.

CGI's Developer School is an intensive five-month program that helps budding professionals hone their IT skills and learn new ones that will help them succeed at CGI. Program graduates receive a salary during the training program, and full-time jobs with CGI upon completion (with a required two-year commitment).

CGI give a special attention to the shareholder's communication, this illustration is an extract of it.

It shows the client satisfaction, the turnover but also one of the huge social responsibility: the women work.

| Revenue | CA**$10.5** BILLION | Backlog | CA**$18.2** BILLION | **68,000** Members |

MEMBERS

| Member Satisfaction Assessment Program score **7.4/10** 2012: 7.7 // 2013: 7.4 | Members who are CGI shareholders 2012: 24,905 // 2013: 37,611 | **44,667** | Women in senior management 2012: 18% // 2013: 17% | **18**% |
| | Women in total workforce 2012: 30% // 2013: 30% | **31**% | Voluntary turnover 2012: 12.7% // 2013: 12.2% | **12.6**% |

Members who participated in the Project Leadership Program	**18,023**
SHAREHOLDERS	
8/10 Shareholder Satisfaction Assessment Program score	

CLIENTS

| **9/10** Client Satisfaction Assessment Program score | **820** In-person client interviews (49% INCREASE) |

Alberta Health receive the 2014 ingenious award from ITAC for an organ and tissue donation registry, CGI was one of the helper to this project.

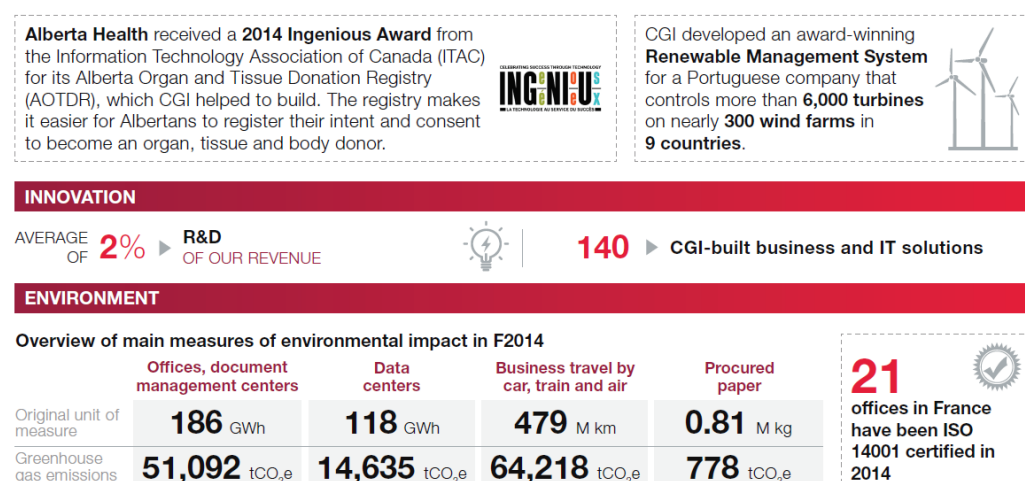CGI have also a social responsibility on protect the environment; it developed an award–winning Renewable Management System for a Portuguese company that control turbines and wind farms. In France, CGI impact a lot the public sector and the health sector. We have many projects with cancer fighting organism.

Alberta Health received a **2014 Ingenious Award** from the Information Technology Association of Canada (ITAC) for its Alberta Organ and Tissue Donation Registry (AOTDR), which CGI helped to build. The registry makes it easier for Albertans to register their intent and consent to become an organ, tissue and body donor.

CGI developed an award-winning **Renewable Management System** for a Portuguese company that controls more than **6,000 turbines** on nearly **300 wind farms** in **9 countries**.

INNOVATION

AVERAGE OF **2**% ▸ R&D OF OUR REVENUE   **140** ▸ CGI-built business and IT solutions

ENVIRONMENT

Overview of main measures of environmental impact in F2014

	Offices, document management centers	Data centers	Business travel by car, train and air	Procured paper	
Original unit of measure	**186** GWh	**118** GWh	**479** M km	**0.81** M kg	**21** offices in France have been ISO 14001 certified in 2014
Greenhouse gas emissions	**51,092** tCO$_2$e	**14,635** tCO$_2$e	**64,218** tCO$_2$e	**778** tCO$_2$e	

With global warming high on the world's agenda, energy efficiency is of growing importance, especially for activities that consume large quantities of

energy. Since data centers are major energy consumers, there is a growing demand from our clients to improve the energy efficiency of these facilities, and CGI is responding.

As CGI grow, CGI continue to measure and improve energy efficiency and reduce our carbon emissions by combining energy-focused methods, processes and solutions to promote power and cooling efficiency.

With these initiatives, the group strive to :

- Ensure environmental monitoring systems are in place to continuously measure power utilization at the rack/server level to manage and adjust power consumption and heat emissions

- Assess virtualization rates across all physical infrastructure, not only to reduce energy consumption, but also to reduce the operational risks and costs of maintenance

- Adopt good practices around hot/cold aisles and the use of row (or even rack) level containment solutions, ensuring appropriate power and cooling solutions are deployed in line with processing density

- Leverage free air cooling technology that enable our data centers, where applicable, to use outside air in the winter months to provide cooling instead of air conditioning units that use additional power.

We saw that CGI try to have a good social responsibility, to manage its relation with the stakeholders. Shareholders are satisfied, the stock exchange price for CGI is stable. The employees are also satisfied, people stay long time in the company. The relation with the government is good because it's also a big client of CGI, CGI is very attentive to follow the laws and have a positive influence on the state.

CGI try also to have in internal and external way to help the sustainable development. In conclusion, CGI is in harmony with his social responsibility vision. The only point is the low level of women working but it's due to the sector. The informatics sector is composed by a lot of men.

ASSIGNMENT FOR THE SECTION "SOCIAL RESPONSIBILITY"

For student

Group	Full name
3АМ7Ф	Rime Belahcene

School	School of engineering and entrepreneurship	**Division**	
Level of education	Master's degree	**Direction / specialty**	Management

Initial data to the section "Social responsibility"	
1. Description of the workplace (working area, technological process, equipment used) for the case of occurrence of: *- harmful manifestations of factors in the production environment (meteorological conditions, harmful substances, lighting, noise, vibration, electromagnetic fields, ionizing radiation)* *- dangerous manifestations of factors in the production environment (mechanical nature, thermal nature, electrical, fire nature)* *- negative impact on the environment (atmosphere, hydrosphere, lithosphere)* *- emergency situations (man-made, spontaneous, ecological and social)*	No issues have been reviewed
2. List of legislative and normative documents on the topic	No issues have been reviewed
List of issues to be investigated, designed and developed:	
1. Analysis of factors of internal social responsibility: *- the principles of the organization corporate culture;* *- the system of labor organization and its security;* *- development of human resources through learning programs and training and development programs;* *- system of social guarantees of the organization;* *- assistance to workers in critical situations.*	

2. Analysis of external social responsibility factors: *- assistance in environmental protection;* *- interaction with the local community and local authorities;* *- sponsorship and corporate charity;* *- preparedness to participate in crisis situations, etc.*	
3. Legal and organizational issues of ensuring social responsibility: *- Analysis of legal norms of labor legislation;* *- analysis of special (characteristic for the investigated field of activity) legal and regulatory legislative acts;* *- Analysis of internal regulatory documents and regulations of the organization in the field of research activities.*	No issues have been reviewed
List of graphic material:	No issues have been reviewed

Date of issuance of the assignment according to a line schedule	

The task was issued by the Advisor:

Position	Full name	cademic degree, cademic status	Signature	Date
Associate professor	N.V. Cherepanova	PHD		

The assignment was accepted for execution by the student:

Croup	Full name	Signature	Date
3AM7Ф	Rime Belahcene		

Conclusion:

In this thesis, we developed the theoretical part of using big data in organization, publics and private. The definition of big data, involving the 3V and the diversity of data. The different using of big data in general, then the particularities of this using in public sector and the reaction and the feeling of people. Then, in the second part we demonstrate with examples how is using today big data in public organization, firstly in social insurance with the CAF (Caisse d'allocation familliale) and the CPAM (Caisse primaire d'assurance maladie) focused on the data matching and the open data. The smartcities, with the transport. London is one of the first smart cities and his transport network use big data early in cooperation with MIT. The SNCF (Société nationale des Chemins de fer) use many captors, data collecting on mobile and computers to make the strategy and develop itself in some new market. And the ministry of interior that use big data for implementing new radars.

The third part is for the future of this using and what can be improved. Indeed, all these example show some problems, I give some recommendation about the internal culture, the management, the outsourcing, formation and potential market for IT companies.

The last part is the social responsibility of the company where I did my trainee, it's seems that CGI have good social responsibility action, and the SR vision feets the SR actions. Nevertheless, they are some areas for improvement like the women work.

This these permits me to discover new information about big data in public sector and to know some change concerning the administrative procedure for French citizen. I also matched theoretical part with my internship in a IT company in public sector and it permits me to have a real reflexion of the good and bad aspect of the big data using.

Bibliography:

1. John M.M. Rumborld, « what are data ? » A categorization of data sensitivity spectrum », https://www-sciencedirect-com.ezproxy.univ-orleans.fr/journal/big-data-research/vol/10/suppl/C

2. Margaret Rousse and Ivy Wigmore, 3Vs (Volume, Varety, Velocity) , http://whatis.techtarget.com/definition/3Vs

3. François Bourdoncle ,Peut-on créer un écosystème français du Big Data ?, https://www.cairn.info/revue-le-journal-de-l-ecole-de-paris-du-management-2014-4-page-8.htm?1=1&DocId=73528&hits=6813+6801+6800+6768+6767+6716+6715+6637+6636+6265+6264+5815+5809+5808+5793+5792+5758+5755+5754+5691+5634+5372+5371+5273+5272+5250+5249+5191+5190+5174+5128+4995+4994+4737+4736+4680+4679+4661+4660+4545+4544+4265+4264+4233+4148+3991+3990+3902+3901+3766+3727+3726+3706+3705+3627+3450+3358+3357+3282+3269+3268+3236+3175+3174+3166+3140+3066+2748+2747+2578+2477+2469+2141+2140+2138+2137+1196+1082+981+980+920+919+904+903+497+496+469+468+439+421+420+394+393+335+334+246+245+92+91+20+19+13+12

4. Chung An Chen and Chih-Wet Hsieh , Les motifs du partage des connaissances dans le secteur public : Le rôle de la motivation de service public, Revue international des sciences administratives, https://www.cairn.info/revue-internationale-des-sciences-administratives-2015-4-p-863.htm?1=1&DocId=165977&hits=8267+8243+8223+8214+8199+8113+8090+8052+8035+7982+7547+7546+7333+7292+7257+6686+6167+6123+6054+6043+6012+5979+5978+5942+5914+5875+5849+5832+2896+2158+2157+2124+2036+2035+2024+1444+1313+1263+1262+1240+1239+1230+1229+1211+1206+1205+1185+516+444+443+427+408+388+307+287+267+222+117+116+94+93+78+48+29+18+17+10

5. Jacques Priol , Le big data des territoires : open data, protection des données, smart city, civic tech, services publics... : les nouvelles stratégies de la donnée au service de l'intérêt general, https://buweb.univ-orleans.fr/ipac20/ipac.jsp?session=152I91T05685C.13019&menu=search&aspect=subtab66&npp=10&ipp=25&spp=20&profile=scd&ri=1&source=~%21la_source&index=.GK&term=big+data+service+public&x=0&y=0&aspect=subtab66

6. Guy Chesnot, Big data et cloud ; stockage et traitement de données du futur, Dalloz http://dallozbndpro-pvgpsla5.dalloz-bibliotheque.fr.ezproxy.univ-orleans.fr/fr/pvPageH5B.asp?puc=005442&nu=160&pa=1#8

7. Ange Chevallier et Géraldine Tauber, Big data et protection sociale : au-delà de la lutte contre la fraude, des opportunités à saisir pour améliorer l'accès aux droits, Regards, https://www.cairn.info/revue-regards-2017-2-page-205.htm?1=1&DocId=484182&hits=1149+1148+1138+1137+1089+1088+1079+1078+1069+1068+567+566+563+562+551+550

8. Oracle, http://www.oracle.com/us/industries/public-sector/public-sector-big-data-br-1676649.pdf

9. Data.gouv.fr, https://www.data.gouv.fr/fr/topics/culture/

10. Isabelle Falque-Pierrotin ,Protection des données: changer les règles pour le meilleur, Le Monde, http://lesclesdedemain.lemonde.fr/dossiers/protection-des-donnees-changer-les-regles-pour-le-meilleur_f-204.html#pdvInfluenceur

11. Ridha Loukil,Et si c'était les services publics qui tiraient le big data en France…, Usine digitale, https://www.usine-digitale.fr/editorial/et-si-c-etait-les-services-publics-qui-tiraient-le-big-data-en-france.N330953

12. Phillipe Jacqué,La SNCF veut developper des alliances avec ses rivaux

pour contrer les GAFA, Le monde,
https://www.lemonde.fr/economie/article/2017/10/04/la-sncf-veut-
developper-des-alliances-avec-ses-rivaux-pour-contrer-les-
gafa_5195961_3234.html

13. Jean-Pierre Blettner, Voyages SNCF Technologies : « nous sommes la
Big Data Fab de la SNCF, La revue du digital,
http://www.larevuedudigital.com/voyages-sncf-technologies-nous-
sommes-la-big-data-fab-de-la-sncf/

14. Jean-Charles Guézel, Les voies du big data chez la SNCF, Alliancy,
https://www.alliancy.fr/a-laffiche/bigdata/2017/10/03/les-voies-du-big-
data-chez-sncf

15. Jérôme Bassaler and Sarah Zaïm, What can business do to capture the full
potential of big data, Orange business https://www.orange-
business.com/en

16. Lisette Calderan, Pascale Laurent, Hélène Lowinger, Jacques Millet,
Nouvelles partitions de l'information. Actes du séminaire IST Inria,
https://www.cairn.info/big-data--9782804189150.htm

17. De Boeck Supérieur, Reflets et perspectives de la vie économique,
https://www.cairn.info/revue-reflets-et-perspectives-de-la-vie-
economique.htm

18. La documentation française, Revue française des affaires sociales,
https://www.cairn.info/revue-francaise-des-affaires-sociales.htm

19. Quelles possibilités offertes par l'analyse des Big Data pour améliorer les
téléservices publics ?, Marie-Pierre Hamel, https://www.cairn.info/revue-
francaise-d-administration-publique-2013-2-page-
437.htm?1=1&DocId=364525&hits=6111+6110+6077+6063+6062+6042
+6041+6040+6037+6016+6010+6009+6002+5992+5991+5988+5980+59
78+5977+5974+5953+5944+5943+5779+5778+5762+5761+5760+5743+
5742+5670+5669+5643+5642+5546+5545+5517+5427+5426+5182+500

9+4970+4920+4374+4342+3930+3917+3586+3585+3243+3228+3227+3
128+1746+1251+1250+1217+1157+994+715+714+305+304+279+278+2
58+257+91+58+57+14+9+8

20. Yves Grandmontagne, Les DRH sont-ils dépassés par le big data ?, Les échos, http://technologies.lesechos.fr/decryptage/les-drh-sont-ils-depasses-par-le-big-data-_f-70.html

21. Yves Calvi, Le big data a besoin de rêveurs, Les échos, https://www.lesechos.fr/02/06/2017/lesechos.fr/030365506765_le-big-data-a-besoin-de-reveurs.htm

22. Alex Howard, 3 big ideas for big data in the public sector, Radar, http://radar.oreilly.com/2012/11/3-big-ideas-for-big-data-in-the-public-sector.html

23. SNCF, https://www.sncf.com/fr

24. CPAM website, https://www.ameli.fr/

25. C-radar, un leader européen de l'innovation autour de la donnée, C-radar, https://www.c-radar.com/societe/c-radar-projets-recherche-developpement/

26. Big data paris 2018 ; SNCF : un train d'avance sur la maintenance prédictive,Emarketing.fr,http://www.e-marketing.fr/Thematique/data-1091/Breves/Big-Data-Paris-2018-SnCF-train-avance-maintenance-predictive-328400.htm#c5jtBi6sHUxo5LHy.97

27. Gilles de Richemond, Big data : vous proposer des meilleurs prix avec des combinaisons inédites de trajet,Ouisncf.frhttps://ouitalk.oui.sncf/blog/techno/bigdata-vous-proposer-des-meilleurs-prix-avec-des-combinaisons-inedites-de-trajet

28. Le big data pour réinventer l'expérience en gare,Atos,https://atos.net/fr/blog/le-big-data-pour-reinventer-l-experience-en-gare

29. La CNIL à l'international, CNIL, https://www.cnil.fr/fr/la-cnil-

linternational

Appendix 1. Big data and public sector (survey)

Q1- What do you think when you think on big data ?

- Overseeing

- Big amount of data

- New technologie

- Other

Q2- You are confronted to big data

- On social medias

- On e-shops

- On public organisation websites

- In real life

- Other

Q3- Do you think that you controle your data

- Yes

- No

Q4- For you big data is a good think ?

- Yes

- No

Q5- What is the main states role ?

- Ensure ecurity

- Ensure freedom

- Social role

- Other

Q6- what do you think when you think on state budget optimisation

(open question)

Q7- Do you think that big data can help this optimisation ?

- Yes
- No

Q8- If yes on which way ?

	Totaly desagree	Desagree	Neutral	Agree	Totaly agree
Fight fraud					
Fight criminality					
Help citizen to find their rights					
Give a better user experience					
Helps public workers					
Helps citizen in their life					

Q9- Do you think that big data can help this optimisation in an other way ?

(Open question)

Q10- Do you think that states shoulg invest in big data research?

- Yes

- No

Q11- Why ?

- To use it in public organisation

- To help companies to be more competitive

- No

Q12- Should the big data be framed ?

- Yes

- No

Q13- Why ?

(Open question)

Few questions about you :

Q14- What age group do you belong to ?

- 16-18

- 18-25

- 25-30

- 30-40

- 40-60

- 60 and more

Q15- In which domain did you work?

(Open question)

Q16- Which professional category do you belong to?

- Farmers

- Craftsman, entrepreneurs

- Managerial staff, superior intermediary professions

- Intermediary professions

- Employees

- Labor

- Without

- Other

Q17- Your postal code?

(Open question)

Big data et secteur public

Le big data désigne l'analyse de la masse de données créées par les appareils numériques (smartphones, ordinateurs, tablettes, automates,etc).
De plus en plus utilisée par les entreprises du secteur privé le but de ce questionnaire est de percevoir la possibilité de l'utilisation de telles technologie dans le secteur public et le ressentis des utilisateurs.

Qu'est ce que vous évoque le terme big data? *

- [ ] Surveillance
- [ ] Gros volume de données
- [ ] Nouvelles technologies
- [ ] Autre...

Vous y êtes confrontés dans votre vie de tous les jours ... *

- [ ] Via les réseaux sociaux
- [ ] Sur les sites d'e-commerce
- [ ] Sur les sites des organisations publiques
- [ ] Dans le monde réel (dans la rue par exemple)
- [ ] Autre...

Avez-vous le sentiment d'avoir le contrôle sur vos données?

- ( ) Oui
- ( ) Non
- ( ) Autre...

73

Le big data est-il une bonne chose pour vous? *

	1	2	3	4	5	
Pas d"accord	◯	◯	◯	◯	◯	Tout à fait d'accord

Quel est le rôle principal de l'État? *

◯ Assurer la sécurité

◯ Assurer les libertés

◯ Rôle social

◯ Autre…

Que vous évoque l'optimisation des dépenses de l'État *

Réponse courte

Pensez vous que le big data permettrait d'appuyer cette optimisation? *

◯ Oui

◯ Non

Si oui de qu'elles manières? *

	Tout à fait d'acco...	D'accord	Neutre	Pas tout à fait d'a...	En total désaccord
Lutter contre les f...	☐	☐	☐	☐	☐
Aider les citoyens...	☐	☐	☐	☐	☐
Lutter contre la cr...	☐	☐	☐	☐	☐
Apporter une meil...	☐	☐	☐	☐	☐
Faciliter le travail ...	☐	☐	☐	☐	☐
Faciliter la vie quo...	☐	☐	☐	☐	☐

Pour vous le big data peut-il appuyer cette optimisation d'une autre façon?

Réponse longue

D'après vous l'État doit il investir dans la recherche concernant le big data? *

○ Oui

○ Non

Pourquoi doit-il y investir? *

○ Pour l'utiliser au sein des organisations publiques

○ Pour aider les entreprises à être plus compétitives

○ Il ne doit pas y investir

Doit-il encadrer l'utilisation du big data et assurer la protection des données? *

○ Oui

○ Non

Dans quel domaine travaillez vous? (exemple: bâtiment, IT, confection, banque etc...) *

Réponse courte

Votre catégorie socio-professionnelle

○ Agriculteurs exploitants

○ Artisans, commerçants et chefs d'entreprise

○ Cadres et professions intellectuelles supérieures

○ Professions Intermédiaires

○ Employés

○ Ouvriers

○ Sans activité

○ Autre...