



NOVA

IMS

Information
Management
School

MAA

Mestrado em Métodos Analíticos Avançados

Master Program in Advanced Analytics

**French Government Transparency Policy on
Healthcare: Building and Monitoring a Public Data
Base**

An ETL and Business Intelligence Project

Maxence Boels

University Supervisor: Vítor Duarte dos Santos

Company Supervisor: Melaine Euzenat

Internship report presented as partial requirement for
obtaining the Master's degree in Data Science and Advanced
Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

2017

Title: Transparency on Healthcare: Building and Monitoring a
Public Data Base

Maxence
Boels

MAA

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

**FRENCH GOVERNMENT TRANSPARENCY POLICY ON HEALTHCARE:
BUILDING AND MONITORING A PUBLIC DATA BASE**

AN ETL AND BUSINESS INTELLIGENCE PROJECT

By

Maxence Boels

Internship report presented as partial requirement for obtaining the Master's degree in Data Science Advanced Analytics

Advisor: *Vítor Duarte dos Santos*

Advisor: Melaine Euzenat

ACKNOWLEDGMENTS

May 2019, Brussels

The internship I had with Deloitte was a great chance for learning and professional development. Therefore, I consider myself as a very lucky individual as I was provided with an opportunity to be a part of it. I am also grateful for having the chance to be part of this master program at university Nova IMS which has always been inspiring and welcoming their students like being part of a family.

I am using this opportunity to express my deepest gratitude and special thanks to Mariem Khairralah, my manager at Deloitte who in spite of being extraordinarily busy with her duties, took time out to hear, guide and keep me on the correct path at their esteemed organization.

I express my deepest thanks to Professor Vítor Duarte Dos Santos from Nova IMS for taking part in useful decision, giving necessary advices, and guidance throughout this internship report. I choose this moment to gratefully acknowledge his efforts to make his busy agenda flexible, his friendly attitude towards me and, it was a real pleasure to work with such a good-tempered person.

I perceive this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives. Hope to continue cooperation with all of you in the future.

Sincerely,

Maxence Boels

List of Abbreviations

Abbreviations

AI: Artificial Intelligence	6
API: Application Programming Interface.....	29
BI: Business Intelligence	10
CIO: Chief Information Officer.....	11
DFT: Data Flow Task	29
DTA: Deloitte Technology and Analytics	2
DW: Data Warehouse.....	8
EST: Execute SQL Task	30
ETL: Extract Transform Load.....	3
IDE: Integrated Development Environment.....	20
IMS: Information Management School	28
IT: Information Technology	11
KPI: Key Performance Indicator.....	33
OLAP: Online Analytical Processing.....	10
OLE DB: Object Linking and Embedding Database.....	29
RDC: Remote Desktop Connection.....	20
ROI: Return On Investment	11
SSDT: SQL Server Data Tools	20
SSIS: SQL Server Integration Services.....	20
SSMS: SQL Server Management Studio.....	20

List of Figures

Figure 1: A hierarchy of consulting purposes (Turner, 1982).....	6
Figure 2: ETL Process in BI (Subramani, 2018)	7
Figure 3: Interactive Dashboard (highcharts.com, 2019).....	11
Figure 4: Data Roles and Skills Sets (Data Science Central, 2016).....	17
Figure 5: Mission's Macro Planning (Deloitte, 2019)	23
Figure 6: SSIS in Microsoft Visual Studio (Deloitte, 2019).....	24
Figure 7: Flash Report on data controls - work in progress (Deloitte, 2019)	24
Figure 8: Flash Report on data controls - development grid (Deloitte, 2019)	25
Figure 9: Parameter file with data controls sources [A] and all sub-sources [E:] (Deloitte, 2019)	28
Figure 10: Conditional Import with Parameter table in SSMS (Deloitte, 2019)	28
Figure 11: SSIS's Control Flow (Deloitte, 2019)	Error! Bookmark not defined.
Figure 12: Sources counting in SSMS (Deloitte, 2019)	Error! Bookmark not defined.
Figure 13: Precedence Constraint Editor in SSIS (Deloitte, 2019).....	30
Figure 14: For Each Loop Editor (Deloitte, 2019)	31
Figure 15: SSIS's Data Flow (Deloitte, 2019)	31
Figure 16: Qlik Sense project (Deloitte, 2019)	32
Figure 17: Qlik Sense Welcome Page (Deloitte, 2019).....	33
Figure 18: Qlik Sense SAP dashboard (Deloitte, 2019).....	33

List of Tables

Table 1: Master Program modules in the 1st year (Deloitte, 2019).....	1
Table 2: Deloitte’s departments (Deloitte, 2019)	2
Table 3: Data Lake vs. Data Warehouses (talend.com, 2019).....	8
Table 4: Agile vs. Traditional Development (Linchpin SEO Team, 2019).....	15
Table 5: Agile or Traditional Project (Deloitte, 2019).....	18
Table 6: Data Controls number and complexity (Deloitte, 2019)	22
Table 7: Orphan Payments (Deloitte, 2019).....	26
Table 8: SAP stands controls (Deloitte, 2019)	27

Table of Contents

1. Introduction.....	1
1.1. Academic Context.....	1
1.2. Business Context	1
1.3. Missions and Objectives of the Internship.....	3
2. Theoretical background.....	5
2.1. Analytical Auditing.....	5
2.2. Consulting Services.....	5
2.3. ETL (Extract Transform Load)	7
2.4. Business Intelligence	10
3. Methodologies, Technologies and Tools.....	13
3.1. Methodologies	13
3.1.1. Deloitte's General Methodologies	13
3.1.2. Agile Methodologies.....	14
3.1.3. Transparency project's methodology.....	17
3.1.4. Is the Transparency project Agile?	18
3.2. Technologies and Tools	19
4. Developed Activities.....	21
4.1. Background.....	21
4.2. Transparency Project Detailed Report	21
4.3. Difficulties.....	34
4.4. Lesson learned.....	34
5. Conclusion	35
5.1. Assessment of Internship	35
5.2. Critical Appraisal of Work Developed	36
5.3. Future Perspectives	37
Bibliography.....	38
Attachments	39

1. Introduction

1.1. Academic Context

This internship Report is part of the second year of Master's in data science and advanced analytics from University of Nova IMS in Lisbon. Students are required to complete either a thesis, a working project or an internship and its written report. My internship started on the 3rd of September 2018 and finished on the 1st of March 2019. Thus, it is a 6 month long internship.

Students are free to go abroad and in any kind of structure as long as the context in which the internship takes place is related to the subject of the master. In my case, I had to find an internship related to the data science or advanced analytics.

The content of this master is defined by its courses. The following courses were given during the first year of the master.

First Semester	Second Semester
Big Data	Predictive Models
Data Mining I	Data Mining II
Computation Intelligence for Optimization	Inferential Analytics
Data Warehousing	Business Intelligence

Table 1: Master Program modules in the 1st year (Deloitte, 2019)

All those courses are theoretical and practical. Indeed, in each of those courses, a hands-on approach has been given to the students during practical classes to help them implement the theoretical knowledge taught daily. Students were asked to work in teams on every project which enhance their team working skills and sense of responsibility.

During the second year of the Master, students are asked to land an internship and find a report supervisor by themselves. Nova IMS is providing to the best students of the first year a chance to get their internship in one of its partner companies such as Accenture, Fidelidade, SAS, and so on. On the other hand, the student office is sharing to students some internship offers to who might need some help to find an internship.

At the end of the internship, students must complete a written report and present it to a jury composed by professors and professionals. After this final step, students are graduating from Nova IMS.

1.2. Business Context

The internship is taking place at Deloitte France in the business area of Paris, La Defense. Deloitte France is composed of more than 15 offices with 10,300 employees around the country. The French Headquarters is located in La Defense, in one of the biggest towers of the country (Majunga tower). Deloitte France has an annual turnover of 1.11 billion €.

Deloitte was created in 1845 in London by William Welch Deloitte and George Touche. Since its creation the company has become the biggest consultancy firm in the world with 244,000 employees worldwide and an annual turnover of \$36.8 billion US Dollar (deloitte.com, 2019).

Deloitte is offering services in almost every sector:

- Life Science and Healthcare
- Consumer
- Energy, Resources and Industrial
- Government and Public Services
- Financial Services
- Technology, Media and Telecommunications

And its services are regrouped in 5 different branches:

- Consulting
- Audit and Insurance
- Financial Advisory
- Legal
- Risk Advisory

The department I am working for is Risk Advisory. More specifically, the Deloitte Technology and Analytics (DTA) department. Which is structured like the fellows:

Department	Sub department
Risk Advisory	Strategic and Reputation Risk
	Regulatory Risk
	Financial Risk
	Cyber Risk
	Operational Risk: DTA, Software Asset Management

Table 2: Deloitte's departments (Deloitte, 2019)

1.3. Missions and Objectives of the Internship

In this section, I will briefly sum up and describe the different missions I worked on and the objectives I received or set. To do so, I will order them by months, starting from September until March.

First, from September to October I was assigned to financial audits:

- Data cleaning with Python using Pandas.
- Testing and checking for some gaps or any inconsistencies in client's Accounting (Invoice, General Balance and General Ledger) in Python.
- Presenting my results to clients with Data Visualization tools such as Tableau.

Then, during October I worked on SAP data extractions missions:

- Extracting clients' data at their places to gather it and send it to the UK's office for further investigation.

Finally, from November until March I was given my main business intelligence mission for a big pharmaceutical company. During this time, I had the opportunity to work on those different tasks:

- ETL solution implementation
 - SQL Server Management Studio (SSMS)
 - Developing SQL scripts tests to make sure the data is consistent and accurate.
 - SQL Server Integration Services (SSIS)
 - Implementing the Data Flows and Controls Flows.
- Data Visualization
 - Qlik Sense
 - Creating Dashboards to monitor the data and reporting.

The biggest part of my work during this 6-month internship will be put on this ETL project for the biggest French Pharmaceutical company.

The new French Transparency Regulation project imposes upon pharmaceutical companies the obligation to disclose all their data related to benefits, payment, and advantages given to doctors and medical representatives to make sure all of this business is done without any abuse and an unethical influence. The risks would be that pharmaceutical companies are paying unreasonable amounts or providing benefits to doctors in order to sell their products to French citizens. Therefore, an online platform has been launched on which the public is able to consult every cent and fringe benefits received from any doctor and medical practitioner (Gouvernement, 2019).

Thus, our role is to make sure the data is accurate before being disclosed to the public. My role is to develop the tests in SSMS, build the data flows and architecture in SSIS and finally design the data visualization tool with Qlik Sense.

I am staffed on fulltime for this mission and backed up by my manager. We are a 2 people team on this big projects, so a lot of pressure and responsibilities is on my shoulders. I went every Monday afternoons at the client's office to participate in a weekly meeting and had a

workshop with our client. Moreover, every Thursday afternoon, we had a Skype meeting to discuss the latest updates and set the next deadlines. Our client was very demanding regarding the deadlines and was always squeezing as much as possible for our planning.

To conclude, the subject of this project was completely on target with regards to my master's program and was very challenging. Thus, I learned a lot every day and had a great human and professional experience.

2. Theoretical background

The theoretical background focuses mainly on the Business Intelligence and ETL part as did the internship. This chapter aims to retrieve the theory behind what was implemented during the different tasks at Deloitte. Therefore, this section includes a theoretical approach of analytical auditing, consulting, ETL and data visualization.

2.1. Analytical Auditing

Deloitte is responsible for monitoring businesses correct financial compliance with its legal duties. For example, the company has to evaluate if its client has been running its business in a proper manner which is regulated by some strict standards. Deloitte is appointed by the local authorities to check if the accounts, the General Ledger and General Balance have been calculated in the correct legal way, and then write an official report for the stakeholders, shareholders and public authorities. All of this must be done with integrity, which means honesty, moral values, professional behavior, and so on.

Audit is the examination or inspection of various books of accounts by an auditor followed by physical checking of inventory to make sure that all departments are following documented system of recording transactions. It is done to ascertain the accuracy of financial statements provided by the organization.

Audit can be done internally by employees or heads of a particular department and externally by an outside firm or an independent auditor. The idea is to check and verify the accounts by an independent authority to ensure that all books of accounts are done in a fair manner and there is no misrepresentation or fraud that is being conducted. All the public listed firms have to get their accounts audited by an independent auditor before they declare their results for any quarter (The Economic Times, 2019). In this particular case, the audit is done by an independent auditor, namely Deloitte.

Usually, the balance sheet and all operations information are thoroughly analyzed to prevent any errors or fraudulent reporting activity. The term “analytical” is present because the nature of the audit is driven by numbers and their relation between them. For instance, Analytical Audit typically explore the turn-over by computing all the sources of revenue and costs.

2.2. Consulting Services

According to the Oxford Dictionary, consulting means “engaged in the business of giving expert advice to people working in a specific field”. Comparing consulting 50 years ago from now and how it has changed, one can assert that the activity is still the same. However, many different types of consulting services have arisen. For example, digital consulting, management consulting, corporate consulting, software consulting etc.

The following image depicts the different purposes of consulting back in 1982 based on a Harvard Business Review article.

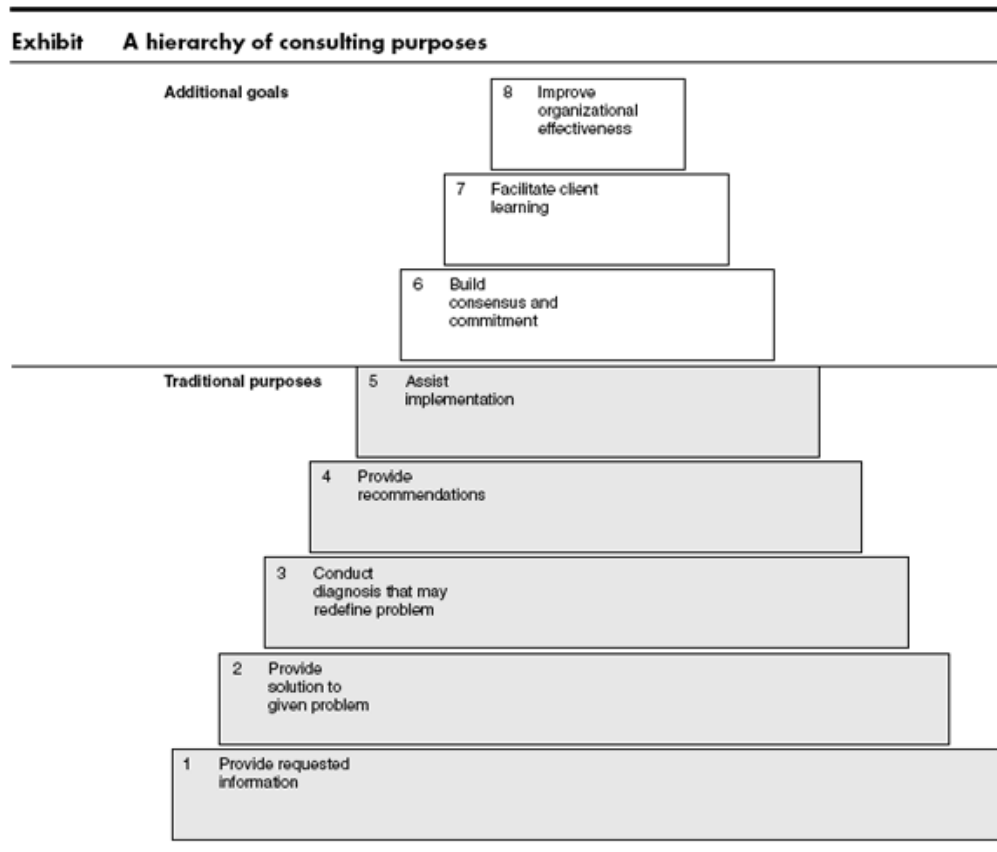


Figure 1: A hierarchy of consulting purposes (Turner, 1982)

Thus, one can conclude that a consultant is someone who has some level of expertise that a particular group of people find valuable, and people within that group are willing to pay the consultant to access their expertise (consulting.com, 2019)

Those are the different ways to deliver consulting services to companies:

- Done for you – Delivering the service yourself or using contractors to do the work.
- Done with you – Doing some of the work and the rest to the client.
- One-on-one coaching – Advising the client on how to execute the work.
- Group coaching – Advising a group of client on how to do it at the same time.
- Online Programs – Creating an online program to let the client learn themselves (consulting.com, 2019).

Consulting is a diverse industry and it's needed to raise the difference between the elite Big 3 in management consulting with McKinsey & Co as the biggest one in annual revenue followed by Boston Consulting Group (BCG) and then Bain & Company, and the Big 4 in financial auditing consulting which are PriceWaterhouseCoopers (PwC), Ernst & Young (E&Y), Deloitte, and KPMG. A third group which consist of technology firms with a strong consulting branch such as IBM and Hewlett-Packard (The Economist, 2018).

The Big four and Big three are facing new competitors providing cloud storage and AI technologies: Amazon, Google and Microsoft are all striving for the cloud computing market worth \$300bn which is fiercely competitive. For example, all 3 firms are offering pre-trained computer vision applications that corporate clients can use to improve their services and

create new ones. As a matter of fact, the corporate market for AI software, hardware and services is valued at \$58bn by 2021, compared with \$12bn in 2017.

Cloud provider companies are winning over management consultancies, which are charging fat fees helping their clients to take advantage from technological disruptions. “The Googles, Amazons and Microsofts of the world may take over from the McKinseys, Boston Consulting Groups and Bains,” says Roy Bahat of Bloomberg Beta, a venture-capital firm. “Consultancies are built for two-by-two matrices. AI’s matrices are a million by a million.” It seems everyone would agree on shifting from traditional consulting strategy to deep expertise in data and technology. To give an example, McKinsey bought up QuantumBlack, an advanced analytics firm, to strengthen its position as data-driven consultancy services in 2015. However, many clients are directly requesting services from tech firms since they developed themselves the latest cutting-edge AI applications, according to Morag Watson of BP, an oil giant.

Another example is the IBM case, during the last decades they have been bridging the gap between tech and conventional consulting. They might have chosen the right strategy to get the best of both merging markets (The Economist, 2018).

2.3. ETL (Extract Transform Load)

ETL is a type of data integration that refers to the three steps (extract, transform, and load) used to blend data from multiple sources. It's often used to build a data warehouse. During this process, data is taken (extracted) from a source system, converted (transformed) into a format that can be analyzed, and stored (loaded) into a data warehouse or other system.

ETL gained popularity in the 1970s when organizations began using multiple data repositories, or databases, to store different types of business information. The need to integrate data that was spread across these databases grew quickly. ETL became the standard method for taking data from disparate sources and transforming it before loading it to a target source, or destination (sas.com, 2019).

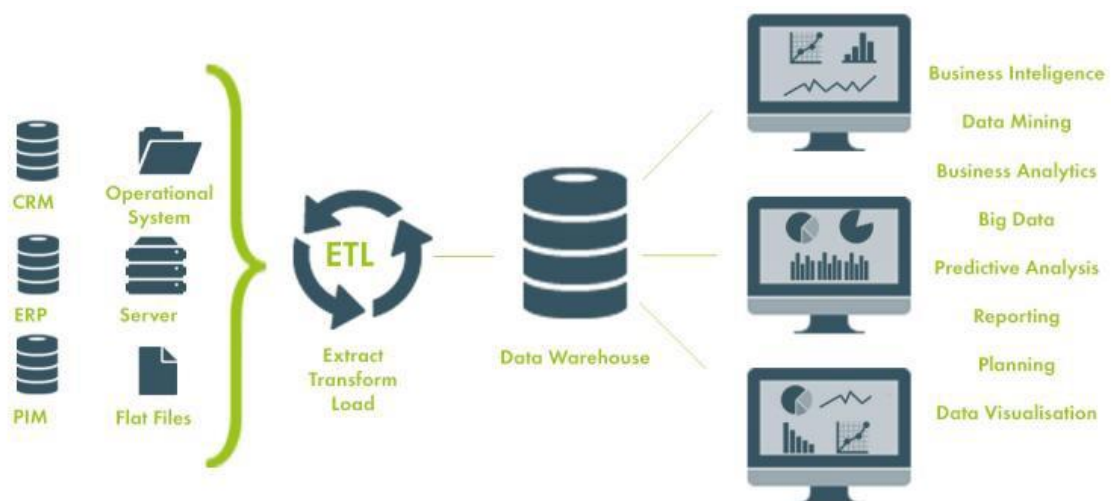


Figure 2: ETL Process in BI (Subramani, 2018)

First, the extract function reads data from a specified source and takes out a desired subset of the data. Those sources have usually different types. For example, extracting data from an access database or a flat file (csv, excel).

Second, the transform function works with the acquired data — using rules or lookup tables, or creating combinations with other data to convert to the desired state. Many other transformation can be done to the data before loading it into the final destination. For instance, making sure all the data types from all sources are homogeneous and amounts can be aggregated correctly. To give a concrete example, many stores might register customers' addresses or date and time format in different ways. The transform part will merge all those heterogeneous data source in a common format to enable uniform analysis through aggregated data from all stores.

Finally, the load function is used to write the resulting data (either the whole subset or overwrite only the changes) to a target database, which may already exist. If not, a new database is created and the data is written on it (Subramani, 2018).

On the above image, the ETL process loads the data into a Data Warehouse. To define it, a data warehouse is designed to support business decisions by allowing data consolidation, analysis and reporting at different aggregate levels. The structure is most of the time shaped as a star with a fact table in the center and its dimensions shaping the tails. Dimensions are providing information with different levels of granularity. For example, typical dimensions are time, location, suppliers, products, customers, etc. Thus, data is populated into the DW through the processes of extraction, transformation and loading (techopedia.com, 2019).

Many companies are now drifting from traditional data warehouses towards data lakes. One must understand their differences to determine which one suits best its organizational needs. Thus, we will compare both data repositories to have a better picture of the latest business intelligence solutions on the market.

First let's define a data lake as a central storage repository that holds big data from many sources in a raw, granular format. It can store structured, semi-structured, or unstructured data, which means data can be kept in a more flexible format for future use.

Data lakes and data warehouses are both widely used for storing big data, but they are not interchangeable terms. A data lake is a vast pool of raw data, the purpose for which is not yet defined. A data warehouse is a repository for structured, filtered data that has already been processed for a specific purpose. The two types of data storage are often confused, but are much more different than they are alike. In fact, the only real similarity between them is their high-level purpose of storing data. There are several differences between a data lake and a data warehouse. Data structure, ideal users, processing methods, and the overall purpose of the data are the key differentiators (talend.com, 2019).

	Data Lake	Data Warehouse
Data Structure	Raw	Processed
Purpose of Data	Not Yet Determined	Currently In Use
Users	Data Scientists	Business Professionals
Accessibility	Highly accessible and quick to update	More complicated and costly to make changes

Table 3: Data Lake vs. Data Warehouses (talend.com, 2019)

Data Structure: Raw vs. Processed

Raw data is data that has not yet been processed for a purpose. Perhaps the greatest difference between data lakes and data warehouses is the varying structure of raw vs. processed data. Data lakes primarily store raw, unprocessed data, while data warehouses store processed and refined data. Because of this, data lakes typically require much larger

storage capacity than data warehouses. Additionally, raw, unprocessed data is malleable, can be quickly analyzed for any purpose, and is ideal for machine learning. The risk of all that raw data, however, is that data lakes sometimes become data swamps without appropriate data quality and data governance measures in place. Data warehouses, by storing only processed data, save on pricey storage space by not maintaining data that may never be used. Additionally, processed data can be easily understood by a larger audience (talend.com, 2019).

Purpose: Undetermined vs. In-Use

The purpose of individual data pieces in a data lake is not fixed. Raw data flows into a data lake, sometimes with a specific future use in mind and sometimes just to have on hand. This means that data lakes have less organization and less filtration of data than their counterpart.

Processed data is raw data that has been put to a specific use. Since data warehouses only house processed data, all of the data in a data warehouse has been used for a specific purpose within the organization. This means that storage space is not wasted on data that may never be used (talend.com, 2019).

Users: Data Scientists vs Business Professionals

Data lakes are often difficult to navigate by those unfamiliar with unprocessed data. Raw, unstructured data usually requires a data scientist and specialized tools to understand and translate it for any specific business use. Alternatively, there is growing momentum behind data preparation tools that create self-service access to the information stored in data lakes. Processed data is used in charts, spreadsheets, tables, and more, so that most, if not all, of the employees at a company can read it. Processed data, like that stored in data warehouses, only requires that the user be familiar with the topic represented (talend.com, 2019).

Accessibility: Flexible vs Secure

Accessibility and ease of use refers to the use of data repository as a whole, not the data within them. Data lakes have no structure and are therefore easy to access and easy to change. Plus, any changes that are made to the data can be done quickly since data lakes have very few limitations.

Data warehouses are, by design, more structured. One major benefit of data warehouses is that the processing and structure of data makes the data itself easier to decipher, the limitations of structure make data warehouses difficult and costly to manipulate (talend.com, 2019).

To conclude, data lakes are not a better version of data warehouses but instead it provides a more flexible structure for unstructured raw data with yet undetermined purpose except for the sake of storing the data for further utilities and with a better accessibility over a data warehouse. However, data lakes require users to be data scientist rather than being at ease with spreadsheets.

Note also the existence of data marts which are a subset to data warehouse environments used to store data relative to a specific business department whereas data warehouses are integrating all business units with an enterprise-width dept. Data marts are easier to access

when frequently using the same data and are created at a lower cost. However, silos as architecture can bring data dependencies issues.

2.4. Business Intelligence

Business intelligence (BI) is the collection of processes, technologies, skills, and applications used to make informed, data-driven business decisions. BI includes data collection, data aggregation, analysis, and meaningful presentation that facilitates decision-making. As mentioned before, data repositories for BI applications include: data warehouses (centralized or decentralized), production databases, operational data stores, and data marts (talend.com, 2019).

The following features are key characteristics for BI tools (talend.com, 2019).

Efficient business intelligence requires the right tools. Different types of BI tools perform various pieces of the overall BI process, and function according to different standards:

- **Online analytical processing (OLAP)** — BI tools that are used to analyze large volumes of historical data with drill-down functionality. Information is stored in OLAP cubes, and provides a multidimensional view of data.
- **Ad hoc analysis** — BI tools that allow any user to make queries and generate a report to answer a specific question, often by using an OLAP “point and click” dashboard.
- **Reporting** — BI tools that provide a visual representation of data that is extracted in a query such as charts, maps and graphs. Benefits of using BI reporting tools include increased speed, efficiency, and accuracy of reports used for analysis.
- **Advanced analytics** — BI tools that are used by data scientists when constructing predictive and prescriptive analytical models. These autonomous or semi-autonomous tools have sophisticated capacities to predict future outcomes and make recommendations.
- **Operational BI** — BI tools that process incoming data in real-time, giving visibility and faster access to information for decision-making. With real-time data and insights, a company can respond rapidly to market trends and events.
- **Open source BI** — BI tools developed with open source code that can be modified as needed. These tools typically come as a suite of products with reporting and analysis capabilities included.
- **Self-service BI** — BI tools that do not need any training in statistical analysis or data mining to use. Self-service systems are configured to allow any user to make queries, design reports, and gain insights using interactive dashboards.

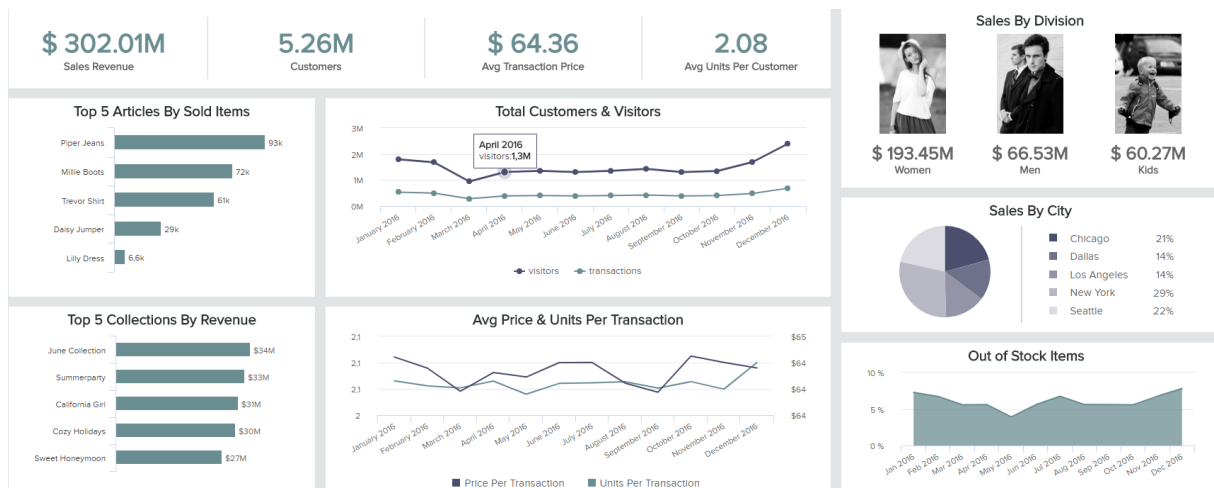


Figure 3: Interactive Dashboard (highcharts.com, 2019)

Here are a few pros for using Business Intelligence (talend.com, 2019):

- **Accelerated time-to-answer** — In-memory analytics with cloud-based data warehouse solutions can analyze data in real-time, providing fact-based information in minutes.
- **Better business decisions** — BI extracts facts and transforms data into actionable information that can be trusted.
- **Improved operational efficiency** — BI makes the interconnections between different components of the business more visible, so problems and inefficiencies can be identified and dealt with more quickly.
- **Increased ROI** — BI helps identify resources needed to reach goals, increases productivity by making data analysis quicker, and aids in the discovery of new revenue streams.
- **Faster reporting** — BI provides real-time reporting of up-to-the-minute, accurate data sets giving organizations a competitive edge in solving complex business problems.
- **Accurate strategies** — BI helps identify important trends and patterns in data that can be utilized to set priorities and allocate resources to meet desired goals.

According to Talend, the future of big data BI is the cloud and artificial intelligence. Pulling data from a production database and dumping it into a spreadsheet for BI reporting is becoming less common. As enterprises move to the cloud, the BI system is automated using cloud-native BI applications that extract insights, make suggestions, and create visual representations of the data.

Today, the needs to be data-driven and to address informational complexities and data modernization, are the driving forces behind businesses cloud strategies. The top three reasons CIO's give for adopting cloud computing information technologies are to:

- Improve agility and responsiveness.
- Accelerate product development and innovation.
- Save money.

Cloud Computing

Compared with older IT systems, cloud computing is often much cheaper. It adds tremendous flexibility: firms that need more computing capacity no longer have to spend

weeks adding new servers and installing software. In the cloud they can get hold of it in minutes. Their applications can be updated continually, rather than just every few months. Individual users can reach their emails, files and photos from any device. And cloud services also tend to be more secure, since providers know better than their customers how to protect their computing systems against hackers (The Economist, 2015). Cloud computing clearly appears as the only ubiquitous and scaled solution able to cope with the big data gathered by companies (The Economist, 2018). With Amazon Web Services, Microsoft Azure and Google Cloud Platform as the biggest cloud companies. Start-ups and giant corporations rent the core resources, along with related software, instead of owning and running their own machines.

What's next? As innovations like artificial intelligence and connected devices become popular, customers are putting cloud components in mobile computing, home games and email marketing campaigns. In other words, the big clouds aim to be everywhere (Quentin Hardy, 2016).

This core cloud business is a \$60 billion-a-year market, which grew by 50 percent in the first quarter of this year, according to Synergy Research Group. In that fast-growing market, Amazon holds a 33 percent share, unchanged since the end of 2015. Over the same span, Microsoft's share climbed from 7 percent to 13 percent, and Google's doubled to 6 percent (Steve Lohr, 2018).

3. Methodologies, Technologies and Tools

In this chapter, we will first present the methodologies used at Deloitte regarding the business intelligence and other missions. Then, we will cover the technologies and tools used during the internship.

3.1. Methodologies

In this subchapter, we start with discussing the general methodologies that are applied in Deloitte's business organization regarding the different business levels. Then, we will cover in more detail the different methodologies that have been applied for the development of the Transparency project. And finally, we will assess Deloitte's ability to use the Agile methodology.

3.1.1. Deloitte's General Methodologies

In this section we will go over a few methodologies, not all of them, used by the Deloitte Technology and Analytics team.

Missions

One of the main characteristics of consulting is its diversity of missions. Those are in most cases done by employees who are willing to work on it, can handle the mission by maintaining Deloitte's high standard of quality and obviously are available to work on it. Those 3 features are not restricting motivated employees to take part in new challenging types of missions for which they can learn or strengthen many skills. Another sought aspect by applicants for consulting is its changing environment and people. Thus, the general methodology to deal with this aspect, is to send Deloitte's consultants to its client's office. This methodology ensure a better following of the project's evolution by both actors since they are working in team close to each other. Because, clients can easily change their minds on some aspect of the solution they want to implement in their business, consultant and clients must work side by side.

Work space organization

The way companies decide to organize their working space can highly affect people and organizational performances. During the last decade, we have seen companies opening their work environment to make them more collaborative and to stimulate synergies between people and departments. Going from closed personal offices to open spaces, and then to open desks. At Deloitte Paris, desks are open which means everyone is free to sit wherever they want (except a few privileged places). But let's come back to the open office pros and cons (Samantha Pena, 2017).

Pros:

- Better Communication between workers and Team working
- Cost-effectiveness (less overheads per office)
- Flexibility of the space
- Better esthetics

- Trendiness of tearing down walls

Cons:

- Distraction
- Lack of Privacy
- Stress and Germs

Security Compliance

Deloitte has recently earned a new security certificate and is putting a lot of efforts to stay in the few top companies which are able to offer such high standards in terms of Security for their clients. Because Deloitte is aware of the new risks companies are facing regarding security breaches and data leaks. Therefore, one can affirm that Deloitte's strategy is to ensure it is always going to be a leader when it comes to integrate market latest security methodologies to its core practices.

Integrity and Ethics

Values are crucial for every company and being able to stick to it can be challenging. Deloitte's methodology regarding the values it wants to represent, is based on integrity. Applying values to its whole business requires to make those values part of the daily practices and thus, to the company's methodology. Therefore, Deloitte's first of all making sure employees fit to those values. Generally speaking, people must be trustworthy, honest, possess moral values, a professional behavior, rightness and respect. Those values are part of Deloitte's methodologies.

3.1.2. Agile Methodologies

General Principles of Agile in Software Development

An Agile Method in software development is a particular approach to project management. This method assists teams in responding to the unpredictability of constructing software. It uses incremental, iterative work sequences that are commonly known as sprints. These methodologies all use the concept of iteration and constant feedback in order to refine a system under development. So, the general principles of the agile methodology are (Linchpin SEO Team, 2019):

- Satisfy the client and continually develop software.
- Changing requirements are embraced for the client's competitive advantage.
- Concentrate on delivering working software frequently. Delivery preference will be placed on the shortest possible time span.
- Developers and business people must work together throughout the entire project.
- Face-to-face communication is the best way to transfer information to and from a team.
- Working software is the primary measurement of progress.
- Constant attention to technical excellence and good design will enhance agility.
- Simplicity is considered to be the art of maximizing the work that is not done, and it is essential.
- Self-organized teams usually create the best designs.
- At regular intervals, the team will reflect on how to become more effective, and they will tune and adjust their behavior accordingly.

Brief History

In 1970, Dr. William Royce published a paper that discussed the managing and developing of large software systems. The paper outlined his specific ideas about sequential development. His presentation stated that a project could be developed much like a product on an assembly line. Each phase of the development had to be complete before the next phase could begin.

Benefits

This Agile Method grew out of the experience with the real-life projects of leading software professionals from the past. Subsequently, this Agile Method has been accepted by the industry as a better solution to project development. The use of iterative planning and feedback results in teams that can continuously align a delivered product that reflects the desired needs of a client. It easily adapts to changing requirements throughout the process by measuring and evaluating the status of a project (Linchpin SEO Team, 2019).

Criticism of Agile Development

- It is developer-centric rather than user-centric.
- Agile focuses on processes for getting requirements and developing code and does not focus on product design.
- Agile methodologies can also be inefficient in large organizations and certain types of projects.

Difference between Agile and Traditional (Waterfall or Spiral) Development

FEATURES / TYPES	Traditional	Agile
Fundamental Assumptions	Systems are fully specifiable, predictable.	High-quality, adaptive software can be developed by small teams using the principles of continuous design improvement and testing based on rapid feedback and change.
Control	Process-centric	People-centric
Management Style	Command-and-control	Leadership-and-collaboration
Knowledge Management	Explicit	Tacit (implicit)
Role Assignment	Individual	Self-organizing teams
Customer's Role	Important	Critical
Project Cycle	Guided by tasks or activities	Guided by product features
Technology	No restriction	Favors object-oriented

Table 4: Agile vs. Traditional Development (Linchpin SEO Team, 2019)

Another comparison has been done in the “Manifesto for Agile Software Development”. This research has uncovered better ways of developing software and values (Cecil, 2001):

- **Individuals and interactions** over processes and tools.
- **Working software** over comprehensive documentation.
- **Customer collaboration** over contract negotiation.
- **Responding to change** over following a plan.

That is, while there is value in the items on the right, the agile methodology values the items on the left more.

Scrums and Sprints

Scrum, the most popular agile framework in software development, is an iterative approach that has at its core the sprint — the scrum term for iteration (Figure 1-3).

The scrum approach includes assembling the project’s requirements and using them to define the project. You then plan the necessary sprints, and divide each sprint into its own list of requirements. At the end of every sprint, you hold a sprint retrospective to look for ways to improve the next sprint.

Within each sprint, the development team builds and tests a functional part of the product until the product owner accepts it and the functionality becomes a potentially shippable product. When one sprint finishes, another sprint starts. Scrum teams deliver product features in increments at the end of each sprint. A product release occurs at the end of a sprint or after several sprints (Layton, 2015).

We saw that the Agile Methodology arose from the need of software developers to change the way they were carrying out projects to a more suited methodology. This process has been now largely adopted by big software services companies like IBM and Microsoft.

Nowadays, data science is a steadily growing field and newer developments keep occurring as well. It is responsible for multiple benefits for varying business industries. Small and large businesses alike are catching up; discovering high potential for growth using data analytics (Nellutla, 2018). Therefore, companies across all industries willing to deploy data science projects must also adapt their methodology to avoid struggling when implementing such projects.

In this subchapter, we will try to identify some of the main features for data science projects to be Agile and try to answer the following question: “how can we create a rigorous methodology to apply agility to the practice of data science?”

Applying methods from agile software development to data science projects (Jurney, 2017).

Agile Data Science is organized around the following principles:

1. Iterate, iterate, and iterate: tables, charts, reports, predictions.
 - a. Building accurate predictive models can take many iterations of feature engineering and hyper parameter tuning. In data science, iteration is the

essential element to the extraction, visualization, and productization of insight.

2. Ship intermediate output. Even failed experiments have output.
 - a. If we didn't ship incomplete or intermediate output by the end of a sprint, we would often end up shipping nothing at all. And that isn't agile.
3. Prototype experiments over implementing tasks.
 - a. In any given task, we must iterate to achieve insight, and these iterations can best be summarized as experiments (tables, charts, reports).
4. Integrate the tyrannical opinion of data in product management.
 - a. Without understanding what the data "has to say" about any feature, the product owner can't do a good job.
5. Climb up and down the data-value pyramid as we work (Figure 1-2).
 - a. The pyramid expresses the increasing amount of value created when refining raw data into tables and charts, followed by reports, then predictions, all of which is intended to enable new actions or improve existing ones.
6. Discover and pursue the critical path to a killer product.
 - a. Once a goal is determined, for instance a prediction to be made, then we must find the critical path to its implementation and, if it proves valuable, to its improvement.
7. Get Meta. Describe the process, not just the end state.
 - a. So where does the product come from? From the palette we create by documenting our exploratory data analysis and by shipping intermediate content.

3.1.3. Transparency project's methodology

One of the reason I decided to cover both the agile methodology for software engineering and data science is because of the nature of the transparency project. The software technology (Microsoft SQL Server) lies between those two fields.

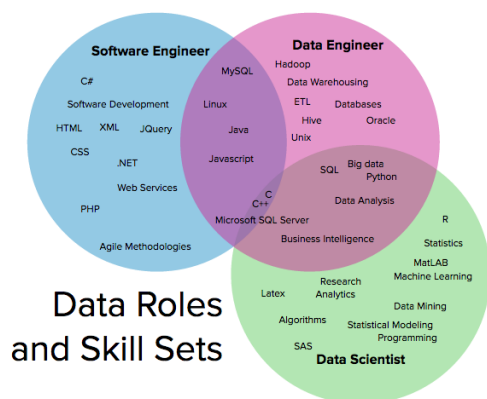


Figure 4: Data Roles and Skills Sets (Data Science Central, 2016)

Getting in touch with clients

Twice a week, we met on Mondays afternoon at the client's office during 2 hours and on Thursdays with a skype meeting of 1 hour. Additionally, emails were exchanged every day to send latest updates and results to the client.

Work in progress

Since the very beginning of the project both teams agreed on sharing weekly results on Thursdays to enable the client to send his feedback on Fridays. Thus, every week we managed to send a flash report (pdf) and the latest updates of our business intelligence solution. Our client asked to send them our updates only if the solution was running perfectly, otherwise the deadline had to be pushed back (dead loop).

Team working and collaboration

Many aspects of the project had to be adapted through the project development. For this reason, communication and team working was crucial. Generally speaking, the communication between both sides was good.

Flexibility

The budget between the client and our company had been signed. Thus, it was not possible to add more content to the agreed solution. Hopefully, both sides had fully documented the project's features and what would be the final product, which helped a lot to stick to the plan. However, any changes had to be negotiated and quit difficult to obtain.

3.1.4. Is the Transparency project Agile?

To evaluate the "Agility" of the project, I decided to select criteria from software and data science agile development based on their ability to discriminate both whether it is Agile or Traditional. However, since the Transparency project has more characteristics related to software development, I decided to retain more features belonging to this area. Therefore, to be classified as "Agile" this project must possess agile characteristics from both sides.

Features	Agile or Traditional
Delivery frequency	High, Weekly updates and delivery
Co-working (client's role)	Critical, weekly meetings and daily emails
Project Flexibility	No, contract negotiation rather than enabling changes
Project cycle	Product features rather than activities
Sprints	Yes, each sprint is well defined in the planning
Data opinion	Yes, understanding the data and check if it makes sense
Get meta	Yes, describe the process for legacy not just the end state

Table 5: Agile or Traditional Project (Deloitte, 2019)

In a nutshell, we can qualify the transparency project's methodology as a mostly agile project although all the features are not identical to the agile standard.

3.2. Technologies and Tools

This subchapter deals with all the softwares and IT tools used during the internship, especially for the business intelligence mission.

Client Laptop

We were asked to develop the ETL tool on the client's server using a laptop they provided us with. For this, we received all the access to connect remotely to the client's server from Deloitte's office. Once they were done at checking, purchasing and installing all the required softwares on their server, we received all the login authorization from their IT department to start developing the ETL solution.

Project Team Communication

On the one hand, chatting among colleagues using Skype for Business is the fastest and most convenient way. On the other hand, using outlook to send emails for formal communications to keep historical information is an imperative. Thus, both of those tools were used on a daily basis.

Remote Desktop Connection to Client's Server

"Remote Desktop Connection (RDC) is a Microsoft technology that allows a local computer to connect to and control a remote PC over a network or the Internet. It is done through a Remote Desktop Service or a terminal service that uses the company's proprietary Remote Desktop Protocol" (techopedia.com, 2019).

SAP (spotlight)

Spotlight is a software developed by Deloitte UK to extract data from SAP. Using Spotlight helps Deloitte's French consultants to easily extract the data and send it to Deloitte UK, which is appointed to verify the extractions of data.

Spyder (Python)

Deloitte Technology and Analytics team has chosen Spyder as its open source integrated development environment (IDE) for analytical programming in the Python language because of the variable explorer feature, anaconda cross-platform and the integrated common data science libraries such like SciPy, Numpy, and Matplotlib.

SQL Server Management Studio (SSMS)

This software developed by Microsoft within Microsoft SQL Server has been broadly used for querying databases and building data warehouses using T-SQL (transact SQL language). For the sake of the business intelligence mission, SSMS was used to develop all the data control scripts. The purpose of those scripts is to check if there are any mistake in the data stored in the database. For each script an output table was created to detect the mistakes and highlight them into a column which is used as filter.

SQL Server Integration Services (SSIS)

SSIS uses the Microsoft Visual Studio user interface to design data integration and workflow applications. This Integration Services application is part of the SQL Server Data Tools (SSDT) package which is used for Business Intelligence projects. During the project, SSIS was tasked to automatically run the ETL data flow from extracting the source files, transforming the data with SQL scripts, to loading the tables in the server database.

Data Visualization Tools (Qlik and Tableau)

During this internship, I had the opportunity to work with interactive data visualization softwares focusing on business intelligence and specializing in visualization techniques for exploring data cubes and relational databases. Those two software applications are almost identical in term of performance, except Tableau's interface is more user-friendly and intuitive. According to the Gartner data visualization magic quadrant 2019, Tableau has a better ability to execute than Qlik, but a slightly less completeness of vision. During the internship, Tableau was chosen by the department for being user-friendly. However, our client picked Qlik for its more advanced features.

Deloitte's other tools

- Safebox: platform to share files and folders in a secure manner internally. It works like a Dropbox.
- Jooxter: for booking meeting rooms.
- WorkDay: managing your obligation and notification as an employee. For instance, asking for a day off or taking compulsory e-learning courses.
- Sigma: All employees are required to fill in their time sheets by entering the daily hours spent on internal or external missions. This must be done at the end of each month to optimize budgets and business performances. Additionally, this platform is used to add all claims to reimbursed expenses.

4. Developed Activities

This chapter is divided into 4 steps. All related to the developed activities during the business intelligence mission for Deloitte's client, namely the pharmaceutical company. First, we will understand the background of this project before I arrived in the company. Then, we will focus on the detailed report. Next, I will present the difficulties that I faced. Finally, every professional experience whether good or bad, has lessons to learn from.

4.1. Background

Before I landed this mission at Deloitte Technology and Analytics, my manager had already started some work regarding the client's needs. She discussed with the client and wrote down all the required data sources and data controls that needed to be developed for the transparency project. Therefore, I had to quickly become familiar with all the data sources and controls. For this, I was given the Word files with all the project specifications ("Data Controls for the French Transparency Disclosure" and "Functional specifications of the Automatic Data controls solution set-up for the French Transparency disclosure"). Thus, during the first week I figured out the connections between all the data sources and the controls. For this, I read all the files that were stored in the mission's folder on the team's server.

4.2. Transparency Project Detailed Report

Project Technical Specifications

Based on the previous work, my first task was to fill in a client's form to describe the requirements, functional and detailed designs, architecture specifications and operational activities. The purpose of this form was purely for the client's internal procedure compliance. I suggested to bring clarity in the project workflow by designing an image of the data flow between the data sources, ssms, ssis, and the final dashboard. Thanks to the data warehousing and business intelligence classes at university NOVA IMS, I was already at ease with the architecture of the business intelligence data flow. Indeed, during those courses we had to create an ETL project with a data warehouse and then, build a dashboard with Microsoft Power BI using the metrics in the fact table and the dimensions shaping the "star schema". However, for this project there was no need to design a data warehouse in a star schema since all the sources were different and the number and granularity of the dimensions were not sufficient. Thus, each source table had to be independently depicted in the data visualization tool, as requested by our client.

Data Controls Design

The main part of this project was to develop the data control code in SQL (Structured Query Language). In total there were 58 data controls to be performed on 15 data sources. The following table illustrates all the controls with their respective complexity, based on the

queries to be coded. For this project, the big pharmaceutical company had already started to code some data controls with Access or macros in Excel. My goal was to either write the sql code based on the existing controls or based on the briefings we had during the face-to-face meetings with the client. Although the data controls had been discussed beforehand with our client, some of them had to be edited later on due to the possibility of logical inconsistencies or enhancements. For instance, removing tests that are redundant or adding missing test to the data controls. During those meetings, I was able to bring some pertinent arguments to add data controls and to raise alerts regarding suggested controls.

Data Sources	Number of Controls	Complexity	Existing Controls
Source 1	8	High	Access
Source 2	11	Low	Macro Excel
Source 3	4	High	No
Source 4	3	High	No
Source 5	2	High	No
Source 6	4	Low	No
Source 7	2	Low	No
Source 8	1	High	Access
Source 9	4	High	Access
Source 10	3	Low	Access
Source 12	2	Low	Access
Source 13	1	Low	Access
Source 14	3	Low	Access
Source 15	1	Medium	Macro Excel
Source 16	8	High	Access

Table 6: Data Controls number and complexity (Deloitte, 2019)

Macro Planning and Roadmap

As a consultant, I had to provide my client with a clear planning of the mission's milestones. I decided to design this agenda under the supervision of my manager with five key steps:

1. Obtaining the client's data sources
2. Setting up the IT environment
3. Data controls development and tests
4. Data visualization dashboards

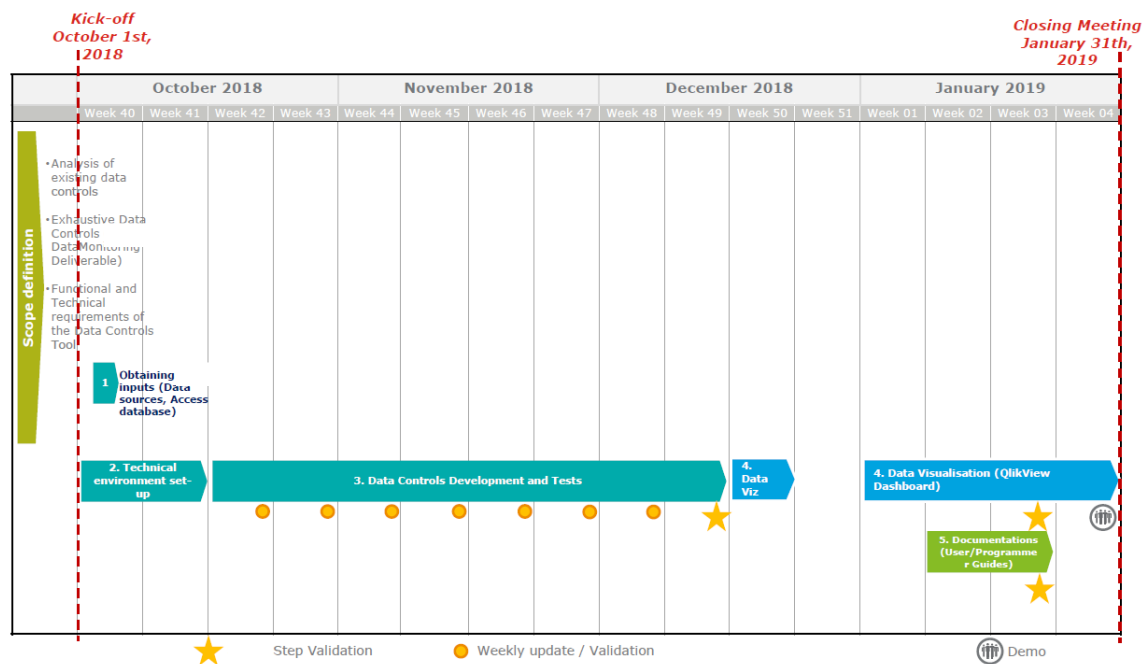


Figure 5: Mission's Macro Planning (Deloitte, 2019)

I quickly realized that our job as consultant is to shrink the amount of hours spent on a project once the budget has been approved and signed by our client while allocating enough time on our schedule to be realistic and make the company's clients satisfied because we are able to deal with deadlines and meet our engagement towards them.

Obtaining our client's data sources

In order to receive the data sources from our client, they shared links to our team with the data sources and existing code (macros and access). Then, I had to store all those datasets on my local machine in folders and go through the data to get familiar with it. I had a lot of sensitive information on my computer and thus, I was not allowed to use Skype, Google Drive, Gmail or any other tool to send the data to the client's local machine we were given to work on it. Instead, Deloitte has its own secured channel to store, send and receive data, which is called SafeBox. Hence, Deloitte's competitive offer on the data and cyber security consulting services. Hopefully, since my first year in my master's in data science and advanced analytics I was used to work with transfer links, zipped files and storing large amount of data on local machine in an organized manner. I was also aware of the need to avoid certain messaging sites or clouds such as WeTransfer, Gmail, Drive, etc.

Setting up the IT environment

This task has been highly underestimated and caused the largest delay in the project's delivery. It turned out that setting up the remote desktop access to the pharmaceutical company's server on the local machine, was more complicated than expected. Moreover, Deloitte was not allowed to install the Microsoft Sql Server Software on their locale machine due to security reasons. I initially wrote 2 weeks on the planning for having the server and applications running properly but had instead a 15 weeks delay on the - agenda pushing everything until February. Fortunately, in the meantime my manager and I were able to carry on the code development on our computers at Deloitte. Nonetheless, I had to install the softwares on my pc which often leads to issues when doing it. But, I still remembered

the steps I took to install SSMS and SSIS during the data warehousing course in Lisbon and therefore, I was able to do it properly on my machine and also advice my manager on the steps to follow to avoid pitfalls. To illustrate with an example, prior SSIS and SSMS can run properly, one must verify the correct installation of “AccessDataBaseEngine.exe” on the machine or check if SSIS runs on a 64 or 32 Bit machine then, change the parameter to 64Bit or 32Bit.

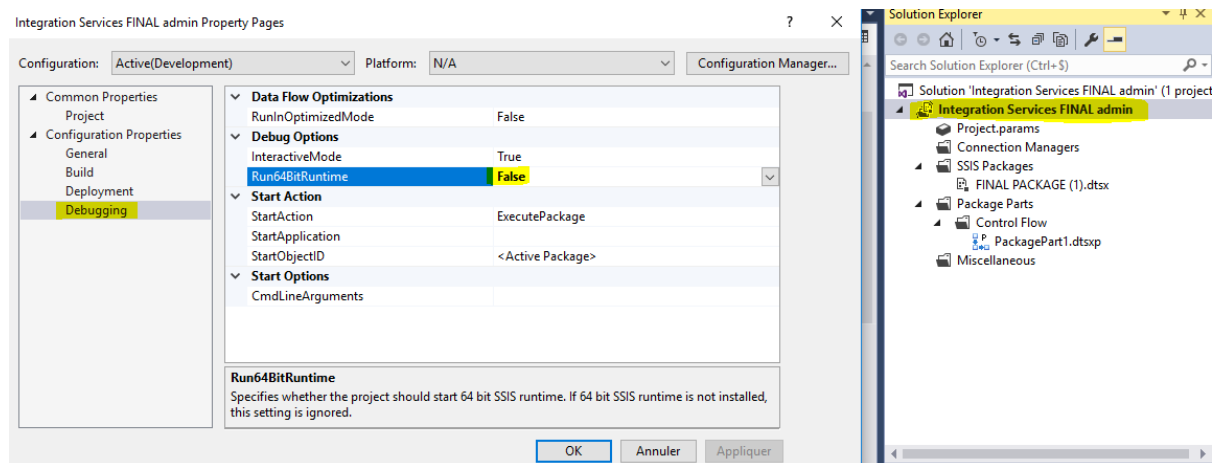


Figure 6: SSIS in Microsoft Visual Studio (Deloitte, 2019)

Flash Reports meetings and Project Management

Our team decided to present once a week our latest updates to our client through a flash report. This was used as support document during the weekly skype meeting that we held every Thursday afternoon. Having a clear visualization of the achievements and objective is critical for long term projects. The following image depicts four points. First, we displayed our latest achievements since our last meeting. Secondly, we presented our work in progress. Next, we raised all the upcoming short term objectives and finally, we raised the attention to the lurking risks regarding the mission’s good progress.

Work performed / finalized since the last Committee - Workshop with Thanh to set up controls for [redacted] – Details in page 3. - SQL developement for Contracts Vs Beneficiary – Details in page 4. - SAP testing with new data sources.	Work in progress - SQL developement for [redacted] - SQL developement for [redacted]
Next steps - Share with Thanh the results of SAP testing with new data sources – Details in page 4. - Obtain new data sources for [redacted] testing. - Workshops with Thanh to set up controls for the data sources listed below (Monday November, 19th): ✓ [redacted] Individual Invitation ✓ [redacted] Orator	Risks identified, alerts ➢ Set up of [redacted] environment is still in progress. ➢ Complexity of [redacted] (if emp data source is confirmed into scope).

Figure 7: Flash Report on data controls - work in progress (Deloitte, 2019)

I would definitely recommend and would use again this tool and methodology for mid- and long-term projects. It was very effective at guiding us through the meetings. To sum up, those 4 windows turned out to be very convenient to avoid getting side-tracked and tackling all the topics on the agenda.

The next figure gathers all the data controls to be developed and then validated. Every week, just before our meeting I quickly updated the table by replacing the red crosses with green circles where the data sources had been worked on. The second column was completed when a data control had been discussed during our weekly face-to-face meeting at the client's office. During those meetings, my role was to write down the logic of each data control per data source and when it was needed I asked pertinent questions to make my notes more understandable for my manager and me. As result, the objective was to provide me with enough information to let me develop the sql script on my own at Deloitte, which I successfully accomplished. Next, with regards to the development and validation columns, they were put as green when on the one hand, a data control's code was fully written. This means that, with a given input data set, I was able to obtain the same expected result as the one given by our client after running my script on the data. For instance, the pharmaceutical company has a table with 1 million rows where a data control should extract 5 names of healthcare professionals that declared more than 60 euro for their business lunch during an event organized by the company. If my sql script extracted those same 5 names, then the development was marked as green. On the other hand, the validation column was marked as green when a data control was validated after having tested the output of the sql script with a new dataset given by our client. Once we were sure the sql script performed as good on a new data set, it was marked as green. The same methodology was applied on the visualization development. First, it was marked as developed when all the dashboards of each source had been designed and then, validated by the client if they were satisfied about the final dashboards.

Data source/ Item Controlled	Updates on existing controls	Deloitte Step 3: Development	validation for Step 3	Deloitte Step 4: Qlik Visualisations	validation for Step 4
	✓	✓	✗	✗	✗
	✓	In progress	✗	✗	✗
	No existing controls	In progress	✗	✗	✗
	No existing controls	In progress	✗	✗	✗
	✓	✓	✗	✗	✗
	✓	✓	✗	✗	✗
	✓	✓	✗	✗	✗
	✓	✓	✗	✗	✗
	✓	✓	✗	✗	✗
	✓	✓	✗	✗	✗
	✓	✓	✗	✗	✗
	✓	In progress	✗	✗	✗
	No existing controls	✗	✗	✗	✗
	No existing controls	✗	✗	✗	✗
	No existing controls	✗	✗	✗	✗
	No existing controls	✗	✗	✗	✗

Figure 8: Flash Report on data controls - development grid (Deloitte, 2019)

To give a concrete example of 2 different sources and their data controls, let's take the "Orphan Payments" and "SAP Stands" sources:

1. Data controls on Orphan Payments:

Here, the purpose of this control is to identify orphan payments from SAP source and provide maximum information for treatment.

- | |
|--|
| <ul style="list-style-type: none">i. Identify Orphan Payments from source in NALA.ii. Use ID number to find related contract and status in historical data.iii. Use Reference Platform to find duplicate data with Individual Invitation Event, the match is done with ID and Amount.iv. Use Expenses workflow (column Assignment) to find duplicate data with Individual Invitation Event, the match is done with ID and Amount. |
|--|

Table 7: Orphan Payments (Deloitte, 2019)

To conclude, this simple example of control is executed using a 'WHERE' clause to filter the source type equal to 'SAP' and also selecting the duplicated rows.

2. Data controls on SAP Stands:

The purpose of this control is to make sure the automatically filling of “SAP stands” in NALA is correctly filled.

i.	Extract all lines in SAP correctly filled Stand Rules and name it STAND_SAP list.
ii.	Extract all lines in NALA NDC correctly filled Stand Rules: STAND_NDC.
iii.	Compare the two lists (STAND_SAP and STAND_NDC) and identify deviations:
○	List of all lines in SAP and NALA NDC correctly filled Stand Rules (no deviations)
○	List of all lines in NALA NDC correctly filled Stand Rules but not in SAP (deviations)
○	List of all lines in SAP correctly filled Stand Rules but not in NALA NDC (deviations)

Table 8: SAP stands controls (Deloitte, 2019)

Thus, in this data control, we are using a “full outer join” to find the missing matches between tables.

Those controls can more or less be easily transformed into sql queries when working with a Microsoft server database. Usually, using ‘SELECT column FROM table’ is not enough and it must be completed with more advanced sql languages features. For example, joins or string functions to select the right part of the data in a particular column. Hopefully, I already had some programming knowledge with sql. For instance, I got hands-on experience during the big data course at university NOVA IMS. This enabled me to quickly jump into action and get my hands dirty on code development. Furthermore, I also did some exercises on Khan Academy’s website to brush up my sql skills.

Finally, I pride myself on being now able to write advanced sql queries and also use dynamic sql which was needed for controls on data sources that don’t follow specific pattern in the column naming or table name.

SSMS: Data Controls Development and Tests

As I explained before, SQL Server Management Studio is an application developed by Microsoft to query tables in a transactional database using Transact-SQL programming language.

One of my achievements was to build a parameter file (PARAMS.xlsx), define it with variables in SSMS and integrate it into SSIS to decide whether a data control would be executed or not, rather than launching all of them in SSIS, when our client wishes to verify only one data source. Another, motivation for this methodology came because of the redundant sub-data sources required to run many data controls on the main sources. Indeed, starting from the 5th column, sliding to the right, one will find all the data sub-sources listed. All those sources are required in the database to run properly one or more data controls. Values have been hard coded as “YES” and should not be considered as a parameter to be changed. A “YES” in a row means that a particular data control needs that sub-data source to execute the control (Figure 8).

The parameter file has been imported in the sql server database using SSIS. This excel file is structured as following:

Source	Start date	End date	FileName					
	01/01/2018	30/06/2018		YES	YES	YES		YES
	01/01/2017	30/06/2018					YES	YES
	01/01/2018	31/12/2018					YES	YES
	01/01/2018	30/06/2018						
	01/01/2018	30/06/2018						
	01/01/2018	30/06/2018						
	01/01/2017	30/06/2018					YES	YES
	01/01/2017	30/06/2018					YES	YES
	01/01/2017	30/06/2018					YES	YES
	01/01/2017	30/06/2018					YES	YES
	01/01/2017	30/06/2018				YES	YES	YES
			Matrices S1 2018_Version 18 OCTOBRE 2018_PROD V1					
			liste transparence S1_2018					

Figure 9: Parameter file with data controls sources [A] and all sub-sources [E:] (Deloitte, 2019)

The user decides which data control (listed as rows) has to be run by filling either the columns [Start date] and [End date] or the column [FileName].

On the following figure, the objective of this SQL script is to read the parameter table, [DB_CLIENT_Test].[dbo].[PARAMS], which exists in the server database, in a way that helps SSIS to determine which data source needs to be imported to successfully run a chosen data control by giving to every sub-sources a particular variable [@S1;@S23].

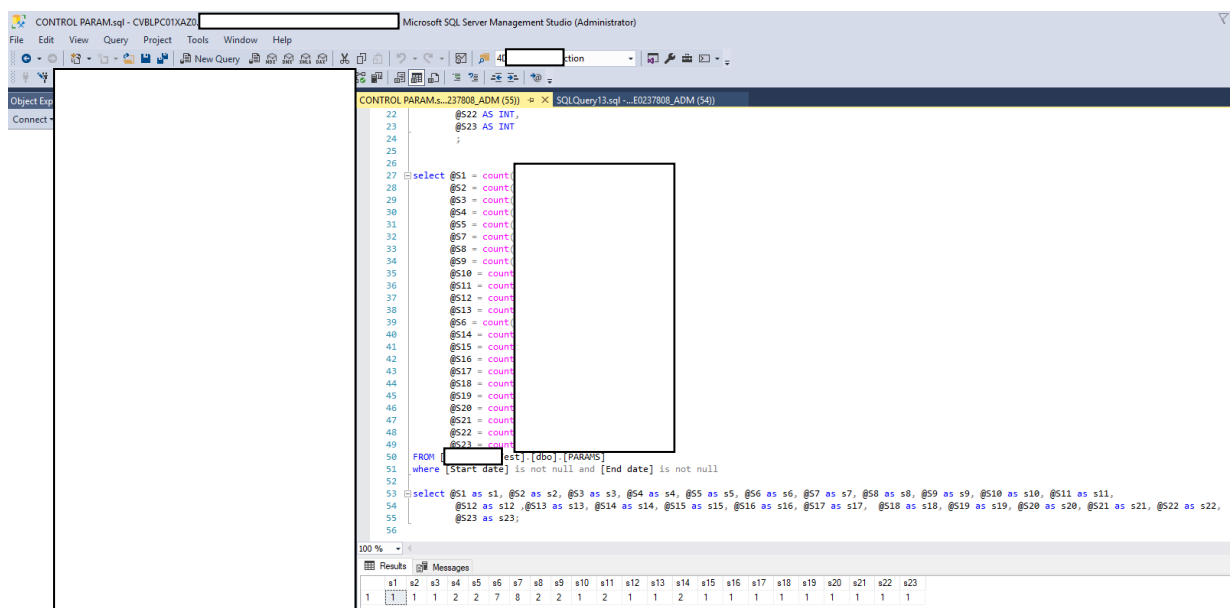


Figure 10: Conditional Import with Parameter table in SSMS (Deloitte, 2019)

In a nutshell, this method has been implemented to optimize the time execution and resources allocation. Because, it avoids importing multiple times the same data sources when data controls share the same input files. Further explanations will be presented in the next section.

SSIS: Control Flow and Data Flow

As mentioned in the previous chapter (Methodologies, Technologies and Tools), Microsoft SQL Server Integration Services is an ETL (Extract, Transform, Load) application used to automate data manipulation and data flows from a source to a destination. Usually, from flat files to an OLE DB (Object Linking and Embedding Database) Destination. The user interface

is Visual Studio designed by Microsoft. The SSIS package belongs to the SQL Server Data Tools which is an extension of the Visual Studio API.

I instantly felt at ease when starting the SSIS project because of my first project during my data warehousing course in Lisbon. So I kicked off the project by creating a new Integration Services project within the Business Intelligence tab. Then, I already knew the Control Flow as being the main page of the project and that it could be compared as the package's macro-view. Usually, one will find the following two types of nodes in the control flow:

- A Data Flow Task (DFT): usually used to connect a source file to a destination table, or from a source table to a destination file.
- An Execute SQL Task (EST): used to perform a query on a database. For instance, truncate a table.

For the sake of understandability of this project, I decided to only represent a snapshot of the Control Flow I built for this project rather than the finale project which includes too many items to define and explain. Thus, here we can observe the main architecture of the Control Flow despite the image's low quality. It has an "Event watcher Task", a "File System Task", "truncate table", "conditional import query", "Data Control scripts" and other EST, also several DFT which are importing all the flat files into the database and exporting the running information (logging).

Now, I will briefly explain the most critical items in this Control Flow in order of execution (top-down). First, the Event Watcher Task is constantly looking for an incoming file in a given folder matching the expression ".xlsx" (Excel files). When this event occurs, SSIS fires a predetermined event, which is returning the node with success which then move the Excel file to another directory to avoid looping this task indefinitely. Here, the objective is to be able to launch the package without having access to the Visual Studio application (client's request). Secondly, I truncated tables to avoid overlapping data every time I run the package and thus, tables are not deleted but emptied before importing the data into their tables. Next, I decided to use a "conditional import query" to select which data source would be imported (time and resource optimization). For this, I had to first create integer variables for each sub-data source that would count and store the number of time this source is required to enable the different data controls to be successfully executed (*Figure 1-4*) and then, use them in an SQL query developed in SSMS.

After, I used the Precedence Constraint Editor to set a Boolean expression on the green arrows (if True=Execute next node) (*Figure 13*).

Precedence Constraint Editor

A precedence constraint defines the workflow between two executables. The precedence constraint can be based on a combination of the execution results and the evaluation of expressions.

Constraint options

Evaluation operation: Expression and Constraint

Value: Success

Expression: @[User::s8] > 0 Test

Multiple constraints

If the constrained task has multiple constraints, you can choose how the constraints interoperate to control the execution of the constrained task.

☒ Logical AND. All constraints must evaluate to True

☐ Logical OR. One constraint must evaluate to True

OK Cancel Help

Figure 13: Precedence Constraint Editor in SSIS (Deloitte, 2019)

Following that, I implemented a ForEach Loop to load Excel files and CSV files dynamically to their sql server tables. Sources files don't have a consistent name through time. Because the files' name is defined based on the current date, such as "NDC_ALL_OPERATIONS_20190126.csv" or "RECORD_NALA_20190115.xlsx". The Connection Manager properties must include some parameters to enable SSIS to recognize source files based on a certain pattern. This pattern could for instance be, "NDC_ALL_OPERATIONS_*.csv". It ensures this source is being imported whatever comes between "OPERATIONS_" and ".csv". Thus, the date won't affect the file name when importing the NDC_ALL_OPERATIONS source file (*Figure 1-7*).

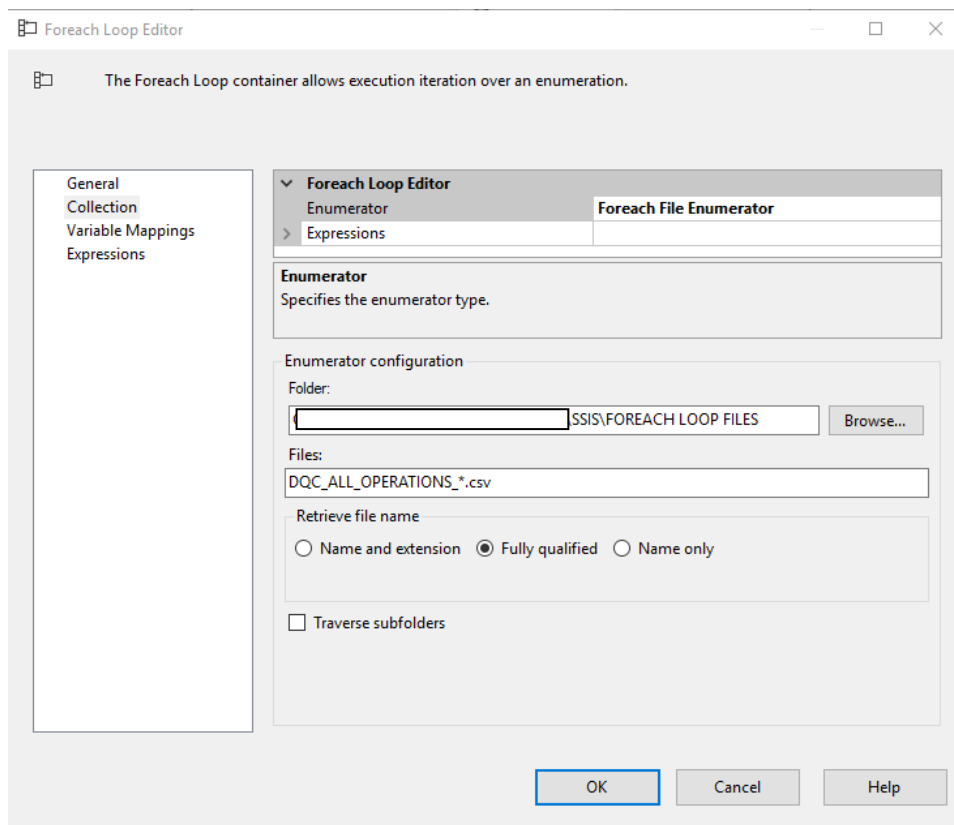


Figure 14: For Each Loop Editor (Deloitte, 2019)

Another idea that I brought up, was to build an output table with the logging information and add a timestamp to monitor the running time of each data control. This table could then be used in the dashboard to view performances without having access to SSIS (concern raised by our client) (Figure 1-5).

On the other hand, the Data Flow is the micro-view. It is the next level of depth in the package where the Source node and Destination node are linked. To create a basic data flow, I dragged and dropped a Data Flow Task from the SSIS Toolbox on the Control Flow and double-clicked on it to access the Data Flow tab. Then, to automatically import a data source into the database created in sql server, one must select a Source Assistant and an OLE DB Destination. The OLE DB Destination loads data into a variety of OLE DB-compliant databases using a database table (Microsoft.com, 2017).

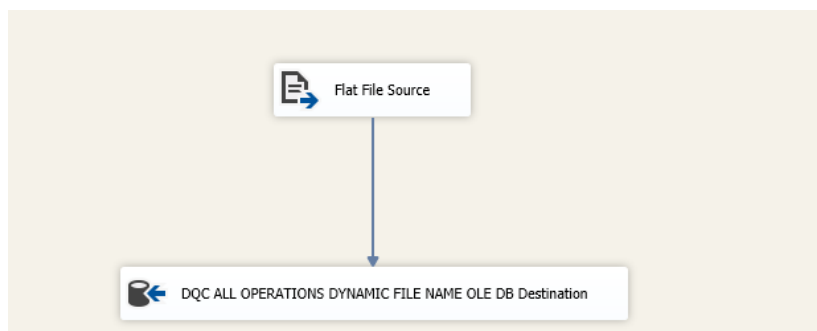


Figure 15: SSIS's Data Flow (Deloitte, 2019)

Finally, the most unexpected and yet challenging task in the Data Flow was to cope with special characters within the source files. I must raise the fact that none of the source files could be changed in any way since they were automatically produced by another software. This led to significant issues for which I had trouble-shooting skills to build on the job.

To cut a long story short, the SSIS features have been built based on constraints and requirements expressed by our client. Because most of the pharmaceutical company's employees would not have access to SSIS and thus, every process had to be fully automated and dynamically configured. This was the most challenging aspect of this project. Hopefully, my manager and I were able to implement all those tasks to the ETL software and had the opportunity to develop our data engineer skills. Moreover, I strongly believe this was possible thanks to the skills I quickly built up during this full time master program in Lisbon and the data warehousing course project.

Data Visualization Dashboards

The client has been working with Qlik Sense for a while now. Since, they have the knowledge and the licenses, they did not decide to change their BI tool for the transparency mission. Building dashboards with this software was not really challenging for me since I already had practical experience working with Microsoft Power BI and Tableau. Qlik Sense is directly connected to the Microsoft Server database and is therefore interactive with the data sources. In this section, I will briefly present the dashboards and a few features we depicted on the data visualizations.

First of all, on to the following image we can directly observe the different highlighted dashboards for each data control when entering the Qlik Sense desktop. This makes it easier for users to choose their point of interest. Then, we suggested to create a welcome page based on the parameter table "[DB_CLIENT_Test].[dbo].[PARAMS]" (figure 14).

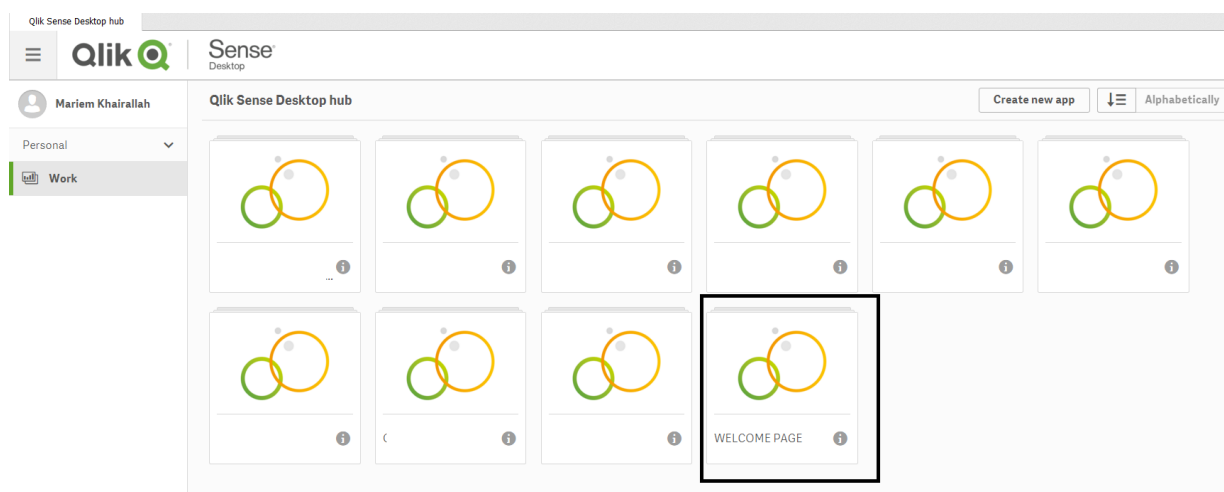
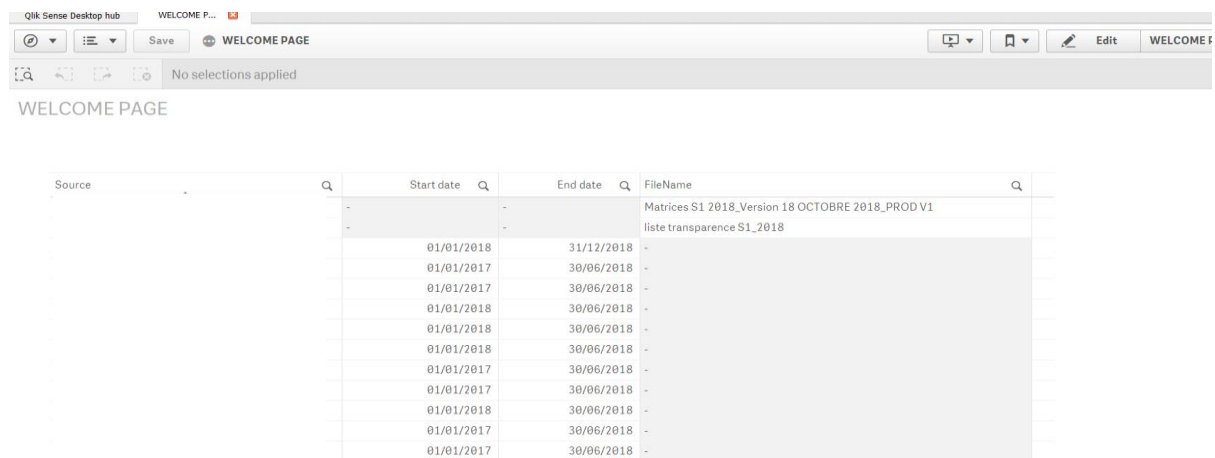


Figure 16: Qlik Sense project (Deloitte, 2019)

Let's consider the next image, we designed the solution with a welcome page to visualize all the data controls and the dates filters. For instance, any user can quickly understand which data controls are filtered on which date and then move on to a specific data control's dashboard. This welcome page enable the user to have a complete view of all the latest data controls that have been performed since we know that the start and end date, or the file

name columns' must be filled in order to execute a data control. So in this case all the controls have been run because the "Agency" sources have their filename column filled and all the other sources have a start and end date.



The screenshot shows the Qlik Sense Desktop hub interface. At the top, there's a header bar with 'Qlik Sense Desktop hub' and 'WELCOME P...'. Below it, a toolbar contains icons for search, filter, and other functions, along with a 'Save' button and a 'WELCOME PAGE' tab. The main content area displays a table with the following columns: Source, Start date, End date, and FileName. The table contains several rows of data, including dates like 01/01/2018 and 31/12/2018, and file names like 'Matrices S1 2018_Version 18 OCTOBRE 2018_PROD V1' and 'liste transparence S1_2018'.

Source	Start date	End date	FileName
-	-	-	Matrices S1 2018_Version 18 OCTOBRE 2018_PROD V1
-	-	-	liste transparence S1_2018
01/01/2018	31/12/2018	-	-
01/01/2017	30/06/2018	-	-
01/01/2017	30/06/2018	-	-
01/01/2018	30/06/2018	-	-
01/01/2018	30/06/2018	-	-
01/01/2018	30/06/2018	-	-
01/01/2017	30/06/2018	-	-
01/01/2017	30/06/2018	-	-
01/01/2018	30/06/2018	-	-
01/01/2017	30/06/2018	-	-
01/01/2017	30/06/2018	-	-

Figure 17: Qlik Sense Welcome Page (Deloitte, 2019)

Let's move on to a concrete example of dashboard. They all have their own metrics and business intelligence since all data controls serve different purposes. For example, on the following image, we can observe some KPI's related to the source SAP. We can see that the number of payments reaches 6.047 with 1.303 cancel payments to verify and 24.274 with 15 anomalies, for SAP Cancel Payment and SAP GLYV, respectively. They have been thought since the very beginning during our meetings with our client and computed by my sql script at Deloitte.

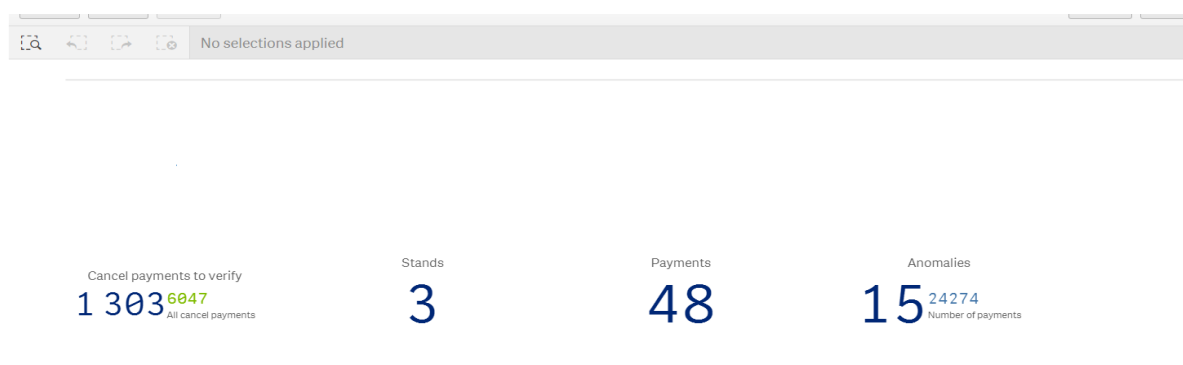


Figure 18: Qlik Sense Dashboard (Deloitte, 2019)

To conclude, the whole purpose of this business intelligence project is to provide clear, accurate, automated and, near real-time information to the decision makers. Our client is now able to detect outliers and anomalies within their previously untapped millions of records and, have it presented on a user friendly interface. All of this would not have been possible without those into depth data controls and a pertinent data visualization. I pride myself on being at the core of this solution and moreover, on having contributed to a better transparency for French citizens between their medical practitioners and large pharmaceutical companies. This is an example of technology initiatives for a better and more equitable society (Tech For Good).

4.3. Difficulties

The first advice I received after having started my internship was to better cope with stress. I learned that being serious and taking only a few short breaks during the day was the model to follow. But apparently, this is the best way to get a burn out. Hopefully, we had a great cohesion within our team and everyone was looking after each other. The second difficulty I faced, was becoming proficient in sql programming language on a short notice in spite of my basic skills. I tackled this issue with time and hands-on experience. This difficulty was directly related to the next one, which was enhancing my problem solving skills with regards to the softwares. Although I already had practical experience with those softwares during my data warehousing classes, it is not an easy task to master Microsoft Management Studio and Microsoft Integration Services. Thanks to YouTube and stackoverflow, I was able to solve most of these pitfalls. Despite, Deloitte France is a leading-edge company, I was surprised by the few amount of people able to reach out with sql expertise, which in my point of view led to considerable time-wasting during this mission.

4.4. Lesson learned

Working on this project taught me many lessons. In particular, I learned to be more resilient. Indeed, developing a script in any new programming language or getting used with any new software requires time and lots of debugging. But as long as, you fail fast and hack your way out to reach the final expected result, the job is done. Additionally, I would also say that I learned to work more rigorously and autonomously on long term projects and thus, I can better take self-initiatives regarding the way I choose to solve a case. And finally, I would add the prestige of working like a consultant for a big reputable company like Deloitte. It includes, having a professional behavior at Deloitte and with its clients. Moreover, always having a stewardship attitude regarding the firm's values and vision.

In a nutshell, all the developed activities I carried throughout this internship regarding the Transparency mission for the client have been described in the above chapter. From preparing the field before starting writing the sql script, then developing the lines of code, followed by monitoring the project's progression while giving a weekly feedback during our flash report on skype, and finally sending all our results on a weekly basis to guarantee we are working on an Agile method to avoid being side tracked during this 4 months long mission. In the end, I believe this project was a success because of our capacity to communicate our needs and issues to the client in spite of the many skills we had to build on the job.

5. Conclusion

To conclude, this internship has been a great experience for me in my personal and professional development. Here, I would like to structure my conclusion in 3 parts. Firstly, I will assess my internship in a general manner and thus, also considering its motivations and impact on my career. Secondly, I will deal with a critical appraisal of my work developed at Deloitte. Finally, I will explain my future perspectives.

5.1. Assessment of Internship

One of the reasons I decided to target a big consulting company is because of the fit with my academic path. On the one hand, my bachelor's degree in business management prepared me perfectly with consulting services in auditing and more broadly, in business consulting. I directly felt at ease when my manager asked me questions about accounting, corporate governance, economic theories and management concepts. Furthermore, it is largely accepted that top consulting companies are a great school to learn from, especially when starting your career and also that they can easily boost your employability on the job market. Companies are always keen to hire people coming from such international institutions. On the other hand, landing an internship in the Deloitte Technology and Analytics team in Paris was a good fit with my master's degree program, namely data science and advanced analytics. Indeed, after having thoroughly discussed the contract position and its responsibilities with different managers and a partner, I knew this internship role would enable me to put into practice all the theory I had seen during my master's courses. In the end, this is the purpose of an academic internship.

Now, let's move on to the gains and drawbacks of my internship. We will first consider the advantages and learnings and then, tackle the downsides during my 6-month long experience within the firm.

In my opinion, my greatest achievement during this journey was being able to jump into an ongoing mission and to manage to carry it beyond expectations. For instance, developing all those sql scripts and working on new softwares were challenges I successfully overcame in spite of my lack of programming experience. Therefore, I pride myself on having strengthened my problem-solving skills regarding programming and my skills regarding project management in IT. Another advantage of this internship, was being able to work directly with the clients as a consultant, but also working in teams and being responsible for my planning. Junior Consultants are quickly trusted and being sent alone to the firm's valuable clients to act as a true member of the company.

Last but not least, being well regarded among my department and colleagues was something I wanted to be worthy of. It takes discipline and self-control to go along with everyone without raising friction and disputes among a team. Thanks to everyone's flexibility and good temper along with Deloitte's incredible faculty to create this comfortable work environment where everyone feel included, respected and listened to.

Despite all the benefits of this internship, in my opinion there are still a few drawbacks I must bring forward. First and foremost, I was not able to participate to mission related to data science and artificial intelligence. Although, they were a many internal initiatives aiming

at fostering those kind of projects, I could not manage my time effectively to work on data science or machine learning missions. This being said, I would assess this internship as a rewarding and enriching experience and would take this decision over again.

5.2. Critical Appraisal of Work Developed

To assess myself, I will first consider the feedback I received from my manager I was working with on the French Transparency mission. Then, I will try to have an objective point of view with regard to my evaluation.

Let's start with the critical appraisal delivered by my manager and also teammate, Mariem Khairralah. On the whole, she was quite satisfied with my overall performance as to the work we delivered to our client. The reason for it is because we managed to overcome all the pitfalls and risks together; only the two of us, given that we had to build up the necessary skills on the job on a short notice. However, due to my lack of technical skills about the softwares and programming languages in the beginning, she realized I needed a few weeks to become familiar with the tools. Once I was at ease with it, developing complex data controls became really easy, except for a few special cases for which I had to do some deeper research on the web. Another comment she made, was my professional behavior when dealing with the clients no matter the circumstances and especially when it was tense. One of the shortcoming she mentioned, was the difficulty I had to follow her advice and sometimes express some doubts regarding it. All in all, according to my manager and our client, they were satisfied with my work since we managed to deliver the final solution on schedule and with all the requested features. Thus, we stuck to the budget, to our initial plan and reached our goals.

As far as I am concerned, it is largely accepted that it is not an easy task to assess oneself, especially when the work being assessed is subject to an evaluation by a third party (nova IMS). For this reason, I won't develop this section thoroughly and will only comment the above paragraph. First of all, I pride myself on sharing Deloitte's morale values and code of conduct. In addition, the fact that our team was able to carry out this challenging task till the end, was for me a great success. Now, I am confident that for my next challenges, I will kick-off into IT projects faster and be able to contribute more to the common burden and team effort.

5.3. Future Perspectives

Undoubtedly, this internship has had a critical impact over my professional career and future perspectives. Therefore, this last part will focus on mapping the different roads that present itself to ensure I possess a clear and plain view of this crossroad. I will start by presenting all the steps I took to create my own future opportunities and then, explain my reasoning based on my rationality and emotions to make the best decision for myself. Thus, it is a twofold approach, first I open as many doors as possible and then, I choose the one that fits best my plan.

So, to open the doors, I decided to first ensure I would receive an offer at Deloitte Paris for September, which I got orally. Next, I applied to Accenture Benelux and Deloitte Belgium in their Data Analytics department. After a tough selection process in both companies, I was made two offers. At that time, I already knew I would rather join Accenture Benelux than Deloitte France or Belgium for two reasons, the first one is because of the core business of Accenture which is focusing on digital and technology, and the second one is because I wanted to enter a new corporate environment and embracing change. Hence, I accepted Accenture's breakable employment contract. Still, I always had this bad feeling of going back to Brussels and missing out this opportunity to get out of my comfort zone in a foreign country. As a consequence, I applied to a French company's office in London, namely Artefact UK. Rapidly, I had once again received a contract offer for a data scientist position starting in September. However, this time in London, which was in my eyes more exciting and challenging.

Now, putting all those job offers aside and going back to my business intelligence mission, I can firmly say that working on a BI project instead of machine learning and data science initiatives has driven my decision to carry on my hunt for the dream job as a data scientist. Because, I knew the chances of encountering those kind of tasks in other top companies was really likely. Eventually, I am now convinced I want to deepen my machine learning skills and gain practical experience in computer vision projects to start my career in an ambitious AI company focusing in machine vision. For all those reasons, I have chosen to start a one year post master degree in computer vision and machine learning in the UK in September. Thanks to this internship which has led me to explore further my tastes and passion for data science and definitively, guided me to another promising experience.

Bibliography

- Cecil, R. (2001). *Agile Manifesto : Agile Software Development*.
- consulting.com. (2019). *What is consulting?* Récupéré sur <https://www.consulting.com/>:
<https://www.consulting.com/what-is-consulting-definition>
- Deloitte. (2019). Paris.
- highcharts.com. (2019). Récupéré sur highcharts.com: <https://www.highcharts.com>
- Jurney, R. (2017). *A manifesto for agile data science: Applying methods from agile software development to data science projects*.
- Layton, M. C. (2015). *Scrum for Dummies*. For Dummies.
- Linchpin SEO Team. (2019). *The Agile method*. Récupéré sur linchpinseo.com:
<https://linchpinseo.com/the-agile-method/>
- Microsoft.com. (2017). *OLE DB Destination*. Récupéré sur [microsoft.com](https://docs.microsoft.com/en-us/sql/integration-services/data-flow/ole-db-destination?view=sql-server-2017):
<https://docs.microsoft.com/en-us/sql/integration-services/data-flow/ole-db-destination?view=sql-server-2017>
- Nellutla, V. (2018). *Applying Agile IT Methodology to Data Science Projects*. Récupéré sur LinkedIn.com.
- Quentin Hardy. (2016). Why the computing cloud will keep growing and growing. *The New York Times*.
- sas.com. (2019). *What is etl?* Récupéré sur [sas.com](https://www.sas.com/en_us/insights/data-management/what-is-etl.html): https://www.sas.com/en_us/insights/data-management/what-is-etl.html
- Steve Lohr. (2018). Microsoft emerges as clear number 2 in cloud computing. *The New York Times*.
- Subramani, V. (2018). *Data preparation & ETL in business performance*. Récupéré sur [medium.com](https://medium.com/@vishwan/data-preparation-etl-in-business-performance-37de0e8ef632):
<https://medium.com/@vishwan/data-preparation-etl-in-business-performance-37de0e8ef632>
- talend.com. (2019). *What is a data lake?* Récupéré sur [talend.com](https://www.talend.com/resources/what-is-data-lake/):
<https://www.talend.com/resources/what-is-data-lake/>
- talend.com. (2019). *What's business intelligence?* Récupéré sur [talend.com](https://www.talend.com/resources/what-is-business-intelligence/):
<https://www.talend.com/resources/what-is-business-intelligence/>
- techopedia.com. (2019). *Extract Transform Load*. Récupéré sur [techopedia.com](https://www.techopedia.com/definition/24170/extract-transform-load-etl):
<https://www.techopedia.com/definition/24170/extract-transform-load-etl>
- techopedia.com. (2019). *Remote Desktop*. Récupéré sur [techopedia.com](https://www.techopedia.com/definition/3421/remote-desktop):
<https://www.techopedia.com/definition/3421/remote-desktop>
- The Economic Times. (2019). *Definition Audit*. Récupéré sur economictimes.indiatimes.com.
- The Economist. (2015). Sky is the limit. *The Economist*.

The Economist. (2018). *Management consulting services*. Récupéré sur economist.com:
<https://www.economist.com/topics/management-consulting-services>

The Economist. (2018). The Era of the cloud's total dominance is drawing to a close. *The Economist*.

Turner, A. N. (1982). Consulting is more than giving advice. *Harvard Business Review*.

Attachments

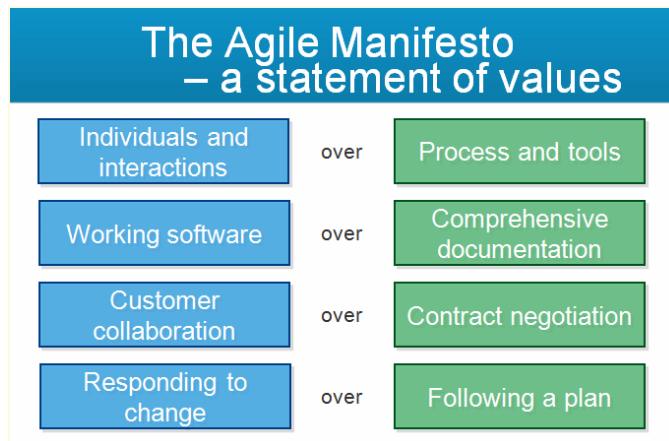


Figure 1-1. The Agile Manifesto (Robert Cecil Martin, 2001)

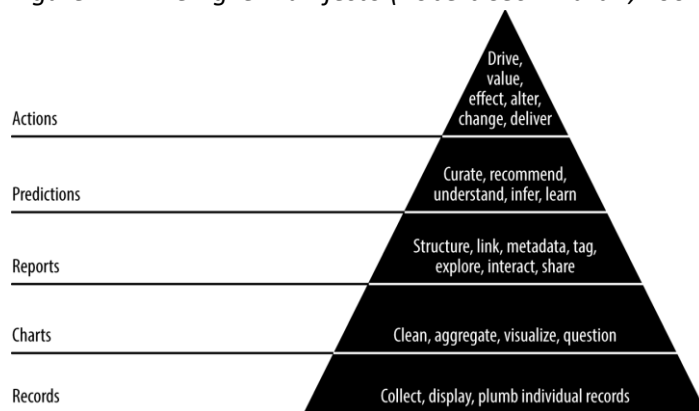


Figure 1-2. The data-value pyramid (Russell Journey, 2017)

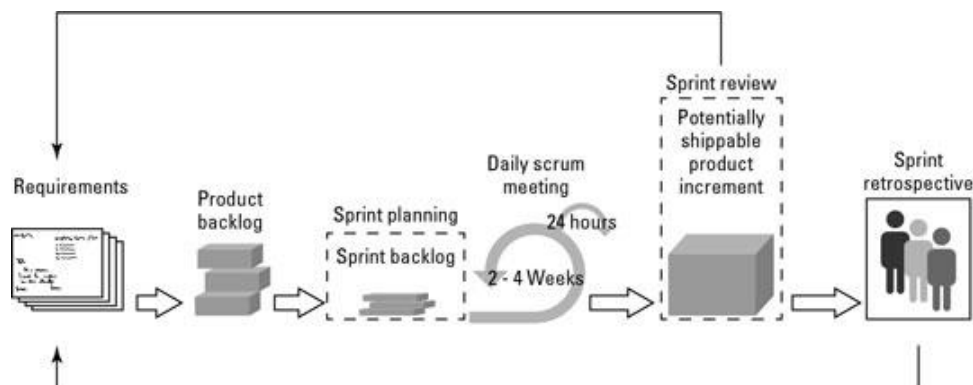


Figure 1-3. Scrum and Sprints (Mark C. Layton, 2012)

s1	MainPackage	Int32	0
s10	MainPackage	Int32	0
s11	MainPackage	Int32	0
s12	MainPackage	Int32	0
s13	MainPackage	Int32	0
s14	MainPackage	Int32	0
s15	MainPackage	Int32	0
s16	MainPackage	Int32	0
s17	MainPackage	Int32	0
s18	MainPackage	Int32	0
s19	MainPackage	Int32	0
s2	MainPackage	Int32	0
s20	MainPackage	Int32	0
s21	MainPackage	Int32	0
s22	MainPackage	Int32	0
s23	MainPackage	Int32	0
s3	MainPackage	Int32	0
s4	MainPackage	Int32	0
s5	MainPackage	Int32	0
s6	MainPackage	Int32	0
s7	MainPackage	Int32	0
s8	MainPackage	Int32	0
s9	MainPackage	Int32	0

Figure 1-4. Source Variables in SSIS (Deloitte, 2019)

