esprit
Ecole Supérieure Privée
d'Ingénierie et de Technologies

REPORT:

PORT OF GOULETTE

4-ERP BI 3

2016-2017

Prepared by :

-Hedi Gaied

-Lina Ben Rhouma

-Majd Ben khalifa

-Nedia Nacef

-Sara Jaziri

-Sirine Dorgham

Table of contents

Table of figures

Introduction

Nowadays, we live in the age where information is power. In any business enterprise, it is imperative that everyone has the critical information they need to accurately and effectively fulfill their business obligations. To deal with frequent and quick market changes, and to enable quick decision making, every business needs to understand Business Intelligence and Business Analytics.

The term Business Intelligence (BI) represents tools and systems that play a key role in the strategic planning for a business. These systems allow a company to gather, store, access and analyze corporate data that aids in recognizing the strengths and weaknesses as well as the threats and opportunities surrounding the business.

As part of our academic projects, we are asked to implement a decision support system.

The present report is a quick presentation of our project during this semester, it will contain a small description of the key elements of our project and it contains four chapters:

- The first chapter: General Context: presents the project context, introduces the problematic. It also presents the needs and the offered solutions.
- The second chapter: Analysis: presents the functional requirement and non-functional requirement and finally our use case.
- The third chapter: modeling: describes the architecture of a data warehouse. It is also presenting the operations and the types of OLAP.
- The fourth chapter: Implementation: presents the tools we are going to use and project realization's phase: ETL, Reporting, Big Data, Datamining and Java application.
- Finally, we end our report with a general conclusion that summarizes the work we have achieved and presents our outlook.

Chapter1: General Context

Introduction

In the first section of this chapter, we will present the organization of the port of Goulette, its organization, its field of activity and the different projects that it carries out. The second section will first give a general description of the project as well as the related problems and will end by specifying its objectives. The final section will amplify the approach taken to carry out this work.

1. Presentation of the port Goulette:

1.1 Overview:

The port of Goulette is one of the most popular destinations in the western basin of the Mediterranean.

The port of Goulette is the point of convergence of the major road and rail networks of Tunisia. It is the outlet of the region richest historically, the most culturally diverse and the most populous including the city of Tunis and its suburbs.

In addition to its passing and cruising activities, the port La Goulette also receives vessels carrying homogeneous cargoes: cars, bulk grains, miscellaneous ... However, the port's development plan foresees its specialization in port reserved exclusively for passenger traffic and Cruise lines.

1.2 Areas of activity:

Ships	1 278 Number (Entry)
Passengers	-714 453 Represent 99% of This whole activity -268 143 Number Passenger cars
Cruise Touristic	-752 246 Number Represent 99 % of this whole activity -358 Ships number
Containers (TEU)	-6 324 Number
Trailers	-21002 Number

2. Context and objectives of the project:

2.1 Introduction:

In an increasingly uncertain and complex environment, the development of a strategy that better achieves objectives, development of action plans, verification of deviations from initial prediction, adaptation Cannot be imagined without the use of information technology to help managers make decisions.

2.2 Issues:

All the maritime agents of the Port Goulette would like to have a clearer view on the state of their depot throughout the year to better manage their activity.

3. Conduct of the project:

In order to properly carry out the mission, it is wise to follow an approach that embodies the classic cycle of a decision-making project, while opting for an iterative behavior concerning deliverables; And this for a better agility in terms of project management.

The approach will be consistent with the traditional decision-making cycle:

*Modeling

* ETL (Extraction / Transformation / Loading):

- Extraction of source data: Resolution of integration and data quality problems from the source to the target.
- Transformation: Application of filters, aggregations, processing of missing or aberrant data and control of discharges, integrity and consistency.
- Load data: Synchronize loads, transfer files or transfer from base to base.

* Feed the OLAP cubes.

* Preparation of reports and dashboards.

* Data mining

Conclusion

In this first chapter, we presented the context of our project in order to make things clear, and to help you more understand our project

Chapter2: Analysis

Introduction

All data warehouse must be able to meet the expectations of users. This may, of course, be done without a thorough study of their needs.
This chapter's main purpose is to present and describe the approach for the detection of needs as well as the presentation of the summary that will be made.

1. Functional Requirements

The official definition for a functional requirement specifies what the system should do: "A requirement specifies a function that a system or component must be able to perform."

Typical functional requirements are:

- Improved Solution Quality
- Veracity of prediction
- Longer Lasting Results
- Rapid Results

2. Non-Functional Requirements

The official definition for a non-functional requirement specifies how the system should behave: "A non-functional requirement is a statement of how a system must behave, it is a constraint upon the systems behavior."

Non-functional requirements specify the system’s quality characteristics or quality attributes.

Typical non-functional requirements are:

- Reliability
- Data Integrity
- Maintainability

Conclusion

In this second chapter, we presented the functional requirement and non-functional requirement.

Chapter3: Modeling

Introduction

Modeling in Business Intelligence project is the most important phase for a successful implementation of a decision support system. In this chapter, we will define and we will present the architecture of a data warehouse.

1. Definition and architecture of a data warehouse

1.1) Definition of a data warehouse:

Data warehouse is a system used for reporting and data analysis, and is considered as a core component of Business Intelligence environment. DWs are central repositories of integrated data from one or more disparate sources. They store current and historical data and are used for creating analytical reports for knowledge workers throughout the enterprise.

A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process:

✓ Subject-Oriented: a data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject.

✓ Integrated: a data warehouse integrates data from multiple data sources. For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product.

✓ Non-volatile: once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.

✓ Time-Variant: historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse. This contrasts with a transactions system, where often only the most recent data is kept. For example, a transaction system may hold the most recent address of a customer, where a data warehouse can hold all addresses associated with a customer.

1.2) Architecture of a data warehouse:

In this section, we present the architecture and the different functions of a data warehouse.

The following are the functions of a data warehouse tools and utilities:

✓ Data Extraction: Involves gathering data from multiple heterogeneous sources.

✓ Data Cleaning: Involves finding and correcting the errors in data.

✓ Data Transformation: Involves converting the data from legacy format to warehouse format.

✓ Data Loading: Involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions.

✓ Refreshing: Involves updating from data sources to warehouse.

A Data Warehouse contain data from various sources. While loading data from various data sources data can be cleansed, transformed as per analysis need and then loaded in data warehouse in a common format which then can be used for reporting & analysis purpose. Since data from all the data sources is now available in common data format and data issues like missing values, incorrect format is corrected while loading data in data warehouse it would be much easier for end user to perform the required data analysis on this data using automated data analysis tools available in the market.

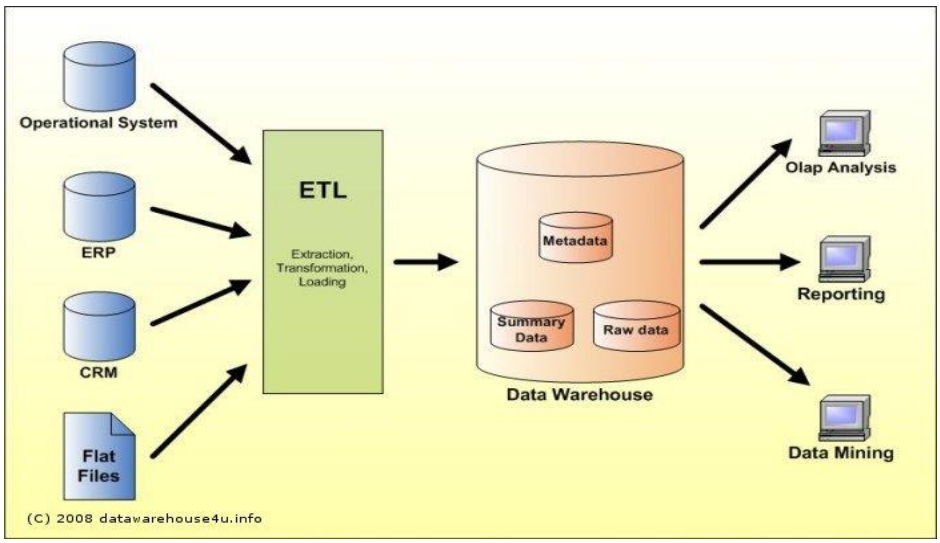

Figure 1 Data warehouse architecture

2. Data warehouse: OLAP

Online Analytical Processing Server (OLAP) is based on the multidimensional data model. It allows managers, and analysts to get an insight of the information through fast, consistent, and interactive access to information

3. Data warehouse design

The following page shows our data warehouse design in details.

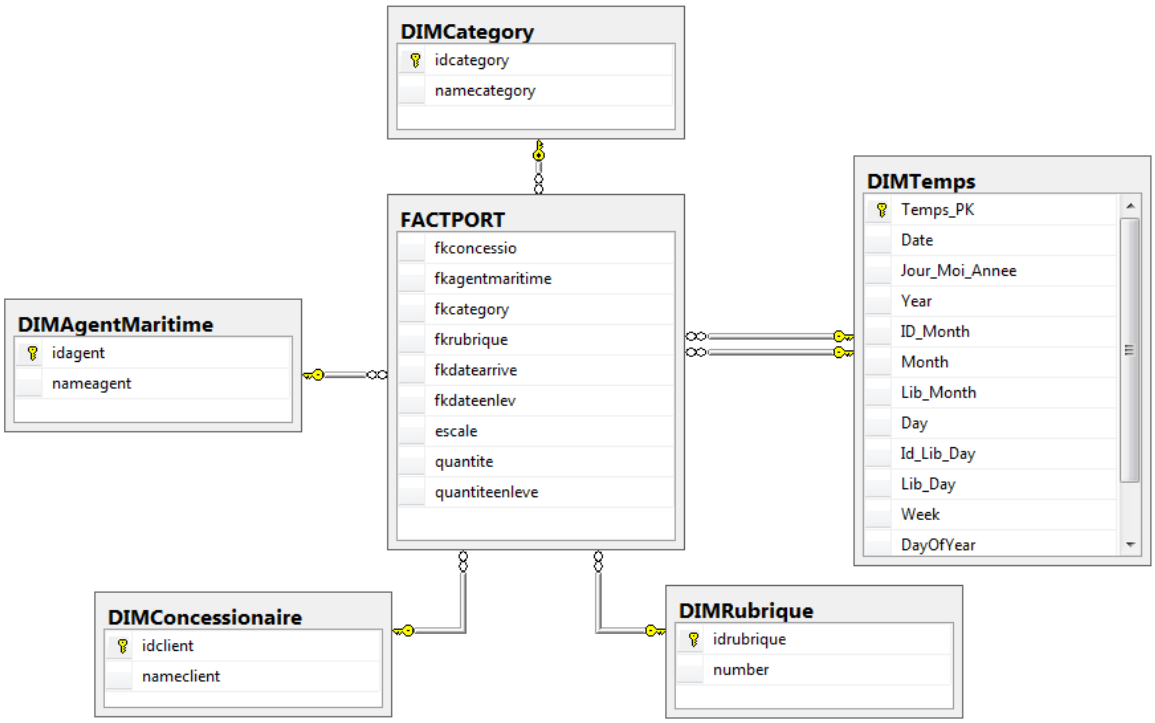

Figure 2 Data warehouse model

Fact_escale	Contains the foreign keys of the different dimensions and measures them We will perform the reports
Dim_temp	Contains the month, day as well as year and another useful field
Dim_Agentmaritime	Contains the primary key and the name of the shipping agent
Dim_Voiture	Contains the primary key and the name of the car
Dim_Client	Contains the primary key and the name of the client
Dim_Rubrique	Contains the primary key and the section number

Conclusion

Throughout this chapter, we presented the architecture of a data warehouse, then we presented the operations and the type of OLAP and finally we presented our data warehouse design.

Chapter4: Implementation

Introduction:

In this last chapter, we will move to the process of project implementation by shedding light on the various tools, and technologies, as well as the different steps of our solutions.

1. Tools and technologies

1.1.1 Pentaho

Pentaho is an open source BI solution developed entirely in java.

1.1.2 Talend Open Studio

Talend Open Studio is the most open, innovative and powerful data integration solution on the market today.

1.2 PostgreSQL

PostgreSQL, is an object-relational database management system (ORDBMS) with an emphasis on extensibility and standards-compliance. As a database server, its primary function is to store data securely, supporting best practices, and to allow for retrieval at the request of other software applications.

1.3 Mondrian Schema Workbench

The Mondrian Schema Workbench allows you to visually create and test Mondrian OLAP cube.

1.4 Tableau

Tableau visualization and analytics products aim to help business managers, analysts and executives see the relationships between different data points, regardless of users' technical skill levels.

Tableau provides reporting, dashboarding and scorecards, BI search, ad hoc analysis and queries, online analytical processing, data discovery, spreadsheet integration, and other data analytics and analysis functions. This ad hoc analysis and data visualization make up the core of Tableau's products.

1.5 Qlik Sense

Qlik Sense gives users a smart analytics tools that can generate personalized reports and very detailed dashboards. It’s designed for Businesses that are looking for a way to fully explore huge amounts of data and high-quality it’s perfect for small to huge organizations and even professionals who work individually, Qlik Sense helps users make sense out of their data.

1.6 Jasper Reports ® Server

Jasper Reports Server is a stand-alone and embeddable reporting server. It provides reporting and analytics that can be embedded into a web or mobile application, JRS can offer a lot to our reporting solution.

1.7 Hadoop

Hadoop framework is capable enough to develop applications capable of running on clusters of computers and they could perform complete statistical analysis for a huge amount of data.

1.8 Apache Hive

Apache Hive - a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis.

Hive provides a mechanism to query the data using a SQL-like language called HiveQL that interacts with the HDFS files.

1.9 R language

R is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data.

2. Realization

2.1. Data Integration:

During this phase, we faced a lot of challenges due to the poor and corrupted data, unusable fields, Lack of documentation and the Lack of analysis axes…

At the end using this model, we've been able to organize the row data that we have and transform it into meaningful homogenous and centralized data.

Now that we have useful, organized data we continued to analysis and reporting phase.

*In this section, we will present the different steps of feeding the data warehouse using both Pentaho and Talend:

Staging Area:

As shown in the figure above, we have five steps:

In the first step:

We extracted the data from the database and started by sorting it and eliminating redundancies.

In the second, third and fourth step:

We began filtering and importing the correct data to replace it with data that contains the same data.

In the fifth step:

We have converted the fields to insert them into the database.

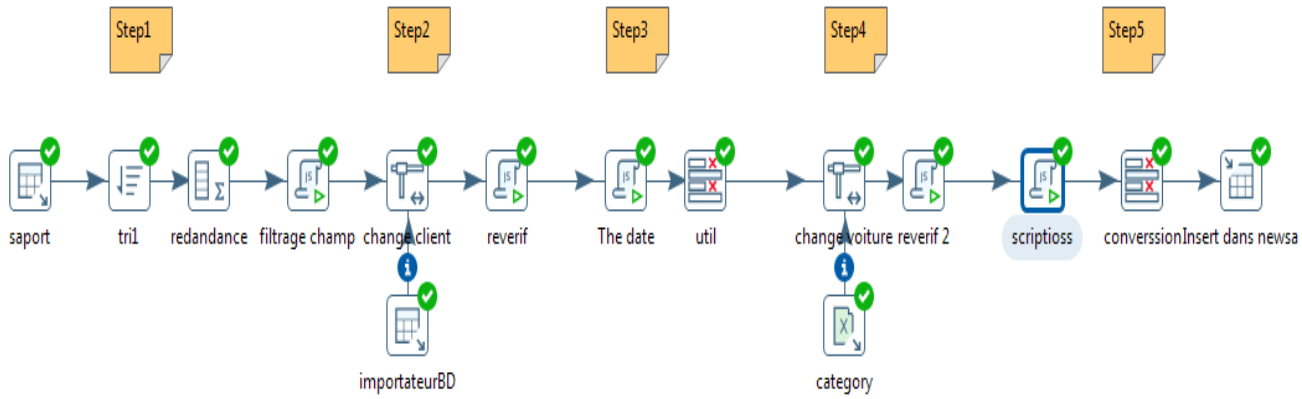

Figure 3 Staging area

Dimension:
-We filled the dimensions.

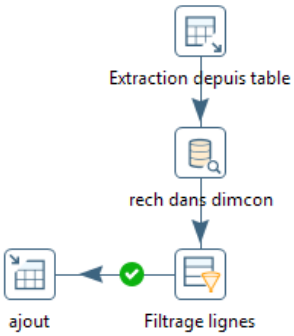

Figure 4 Feeding of Dimension

Fact table

-As shown in the figure above, we have fed our FactPort as follows:

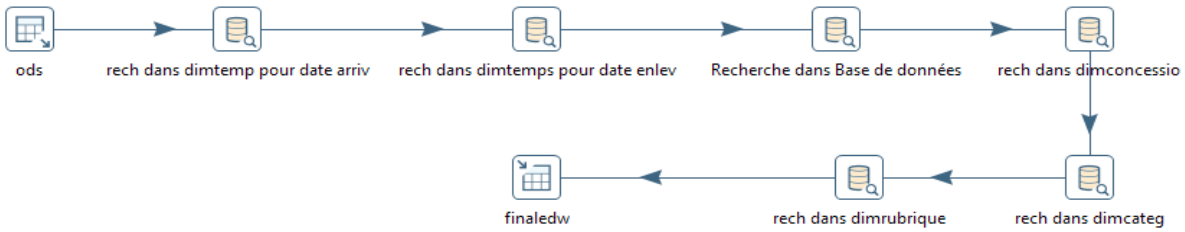

Figure 5 Feeding of fact table

-Here is an illustration of our fact table in our data warehouse:

fkconces... integer	fkagent... integer	fkcategory integer	fkrubrique integer	fkdatear... integer	fkdateen... integer	escale integer	quantite integer	quantite... integer
706	171	520	8735	20100101	20140131	0	0	1
680	171	512	8735	20140331	20140414	0	123	3
680	171	512	8735	20140331	20140405	0	169	100
680	171	512	8735	20140331	20140407	0	0	65
680	171	512	8735	20140331	20140414	0	0	4
680	171	512	8735	20140331	20140408	0	3	3
680	171	512	8735	20140331	20140408	0	31	31
680	171	512	8735	20140331	20140414	0	9	29
690	168	516	8735	20140331	20140403	0	295	252
690	168	516	8735	20140331	20140404	0	0	43
696	166	520	8735	20140508	20140512	0	7	5
696	166	520	8735	20140508	20140523	0	0	2

Figure 6 Fact table

Job:

-We have regrouped the different tasks in a single job to facilitate the execution.

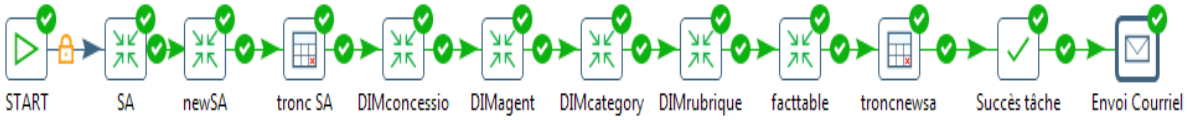

Figure 7 Job

2.2. Analysis

We used the "Schema Workbench" tool to achieve our OLAP cube, first, we created our fact table, and then we added our dimensions.

The measures represent the final elements of our cube: quantitee, quntiteenleve.

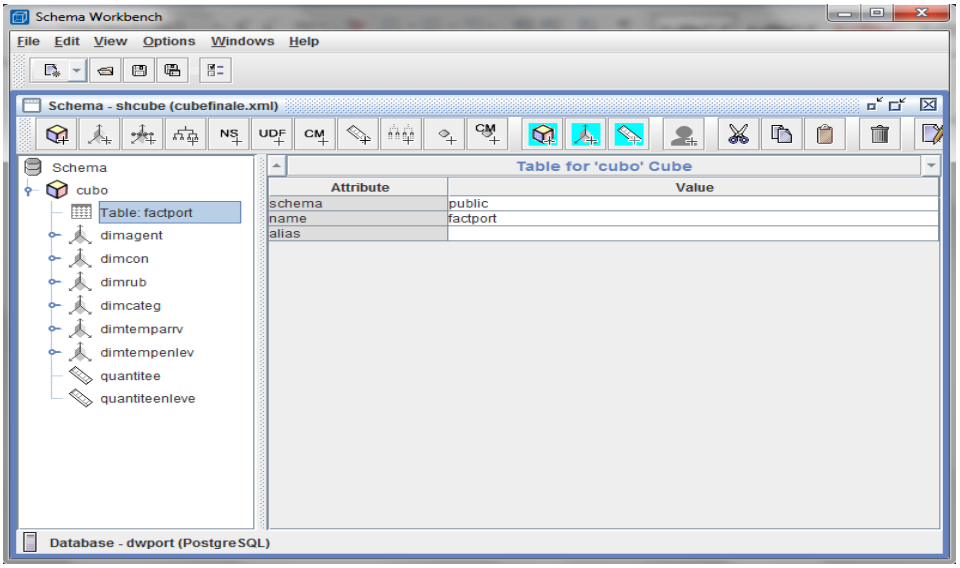

Figure 8 OLAP cube using schema workbench

2.3 Reporting

Using analysis and reporting tools we were able to create these dashboards that responds to the client different requirements and will help them eventually in the decision making.

- **Using Qlick sens :**

-The line chart expresses the rate of cars remaining throughout the year for each shipping agent. And the pie chart besides, shows us for each month the distribution of the cars for the different dealers.

We see an important peak in March for the shipping agent Genmar.

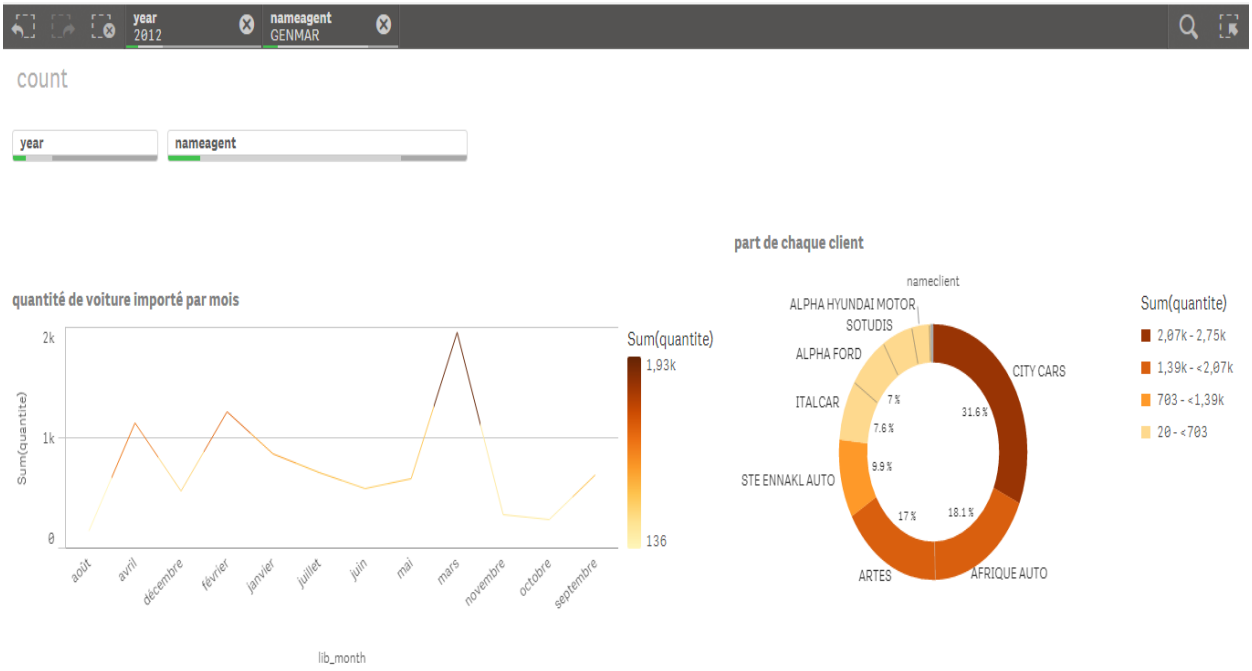

Figure 9 Number of cars imported per agent

-This figure shows us the saturation threshold of the warehouse by each month.

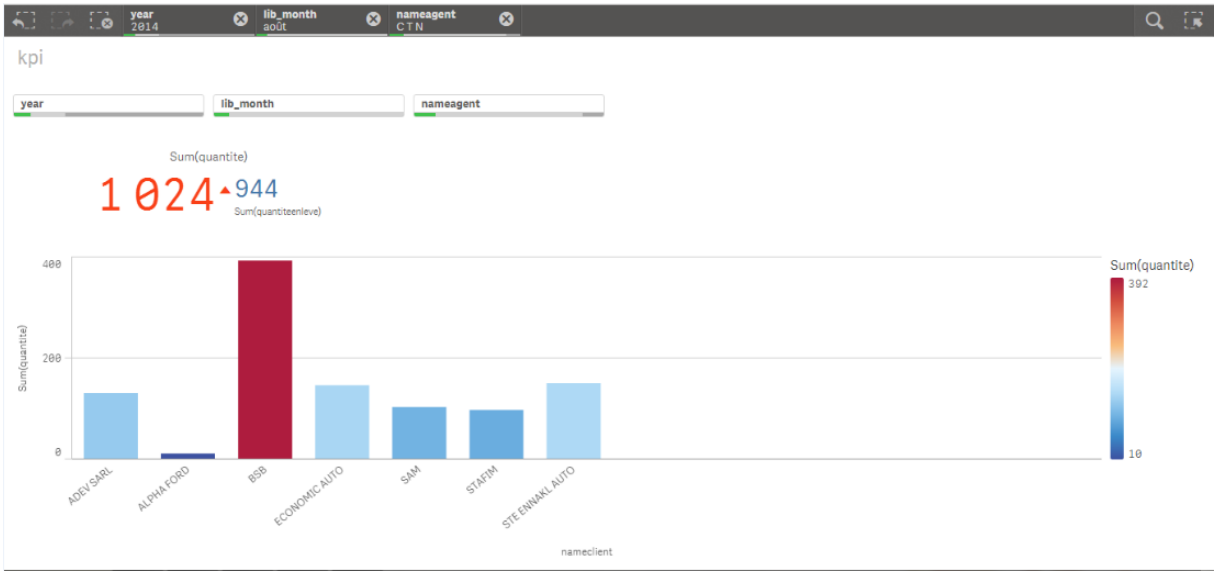

Figure 10 KPI

- **Using tableau:**

This horizontal bar chart shows the classification of dealers by the import rate for each shipping agent. Some of the data was not available but we can see that the highest number of import is BSB.

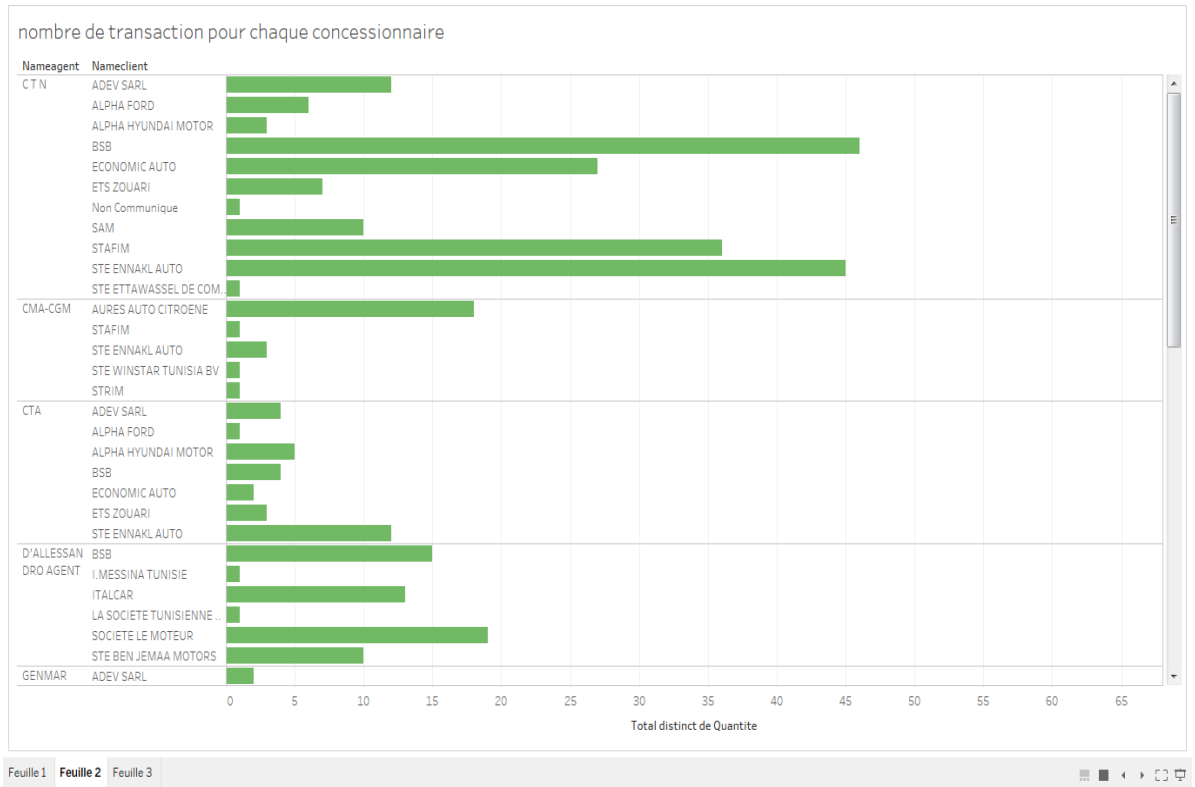

Figure 11 Dealers classification by the import rate

- **Using Jasper report:**

This figure shows us the share of car import of each shipping agent for the year 2013 and one finds that the two agents Hibmar and x are the most dominant.

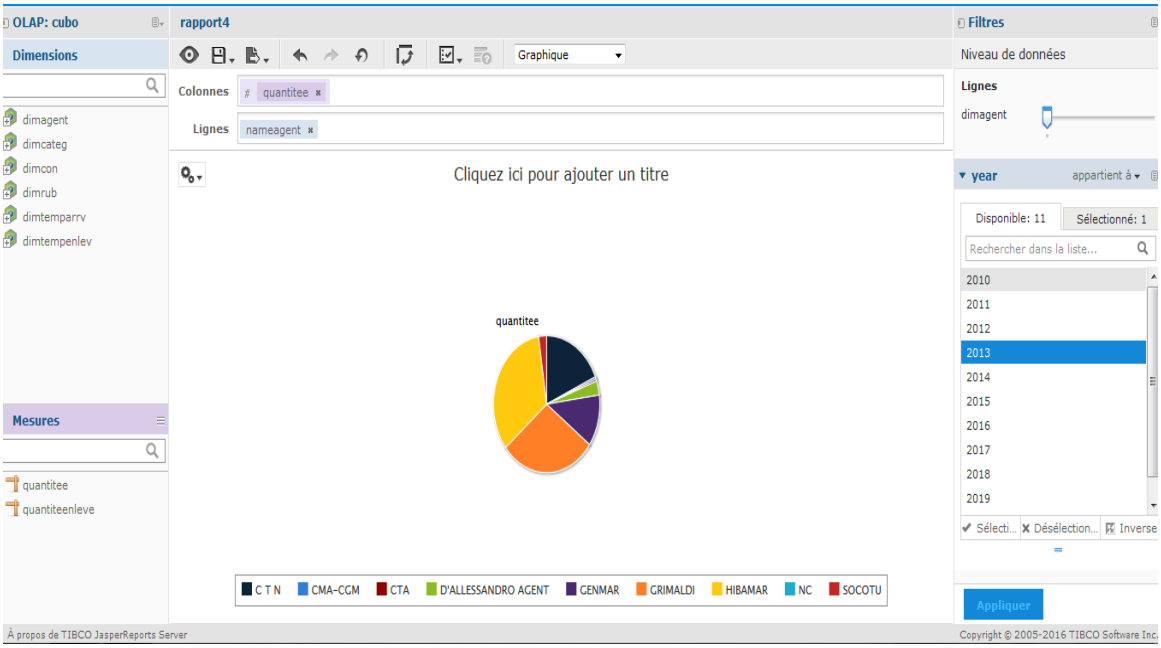

Figure 12 Import rate for each agent

2.4 Datamining:

Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve.

In this part, we applied Data Mining algorithms to analysis our data and then make decision.

- **Principal component Analysis :**

The main goal of a PCA analysis is to identify patterns in data also detect the correlation between variables. If a strong correlation between variables exists, the attempt to reduce the dimensionality only makes sense.

Therefore, in our case we want to know what are the major factors that influent on the real estate market, and what is the relation between the individuals and variables, variables and variables.

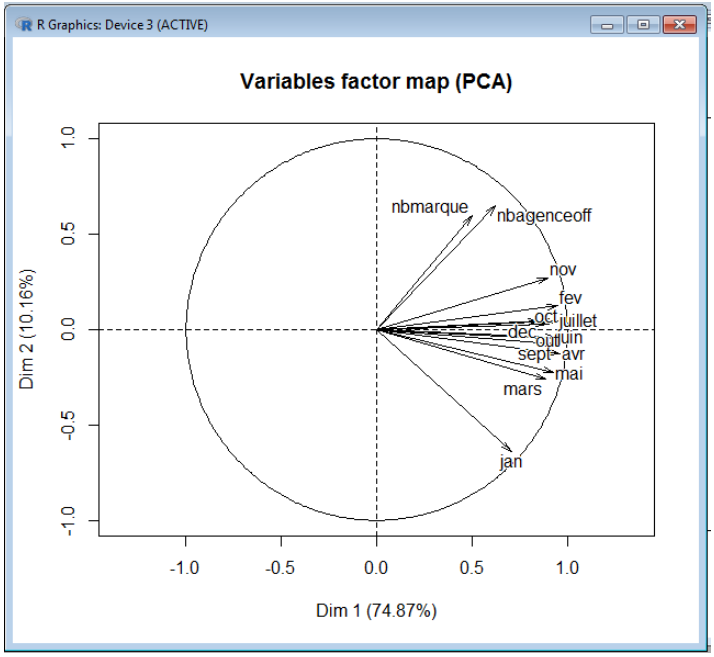

Figure 13 ACP: Variables factor map

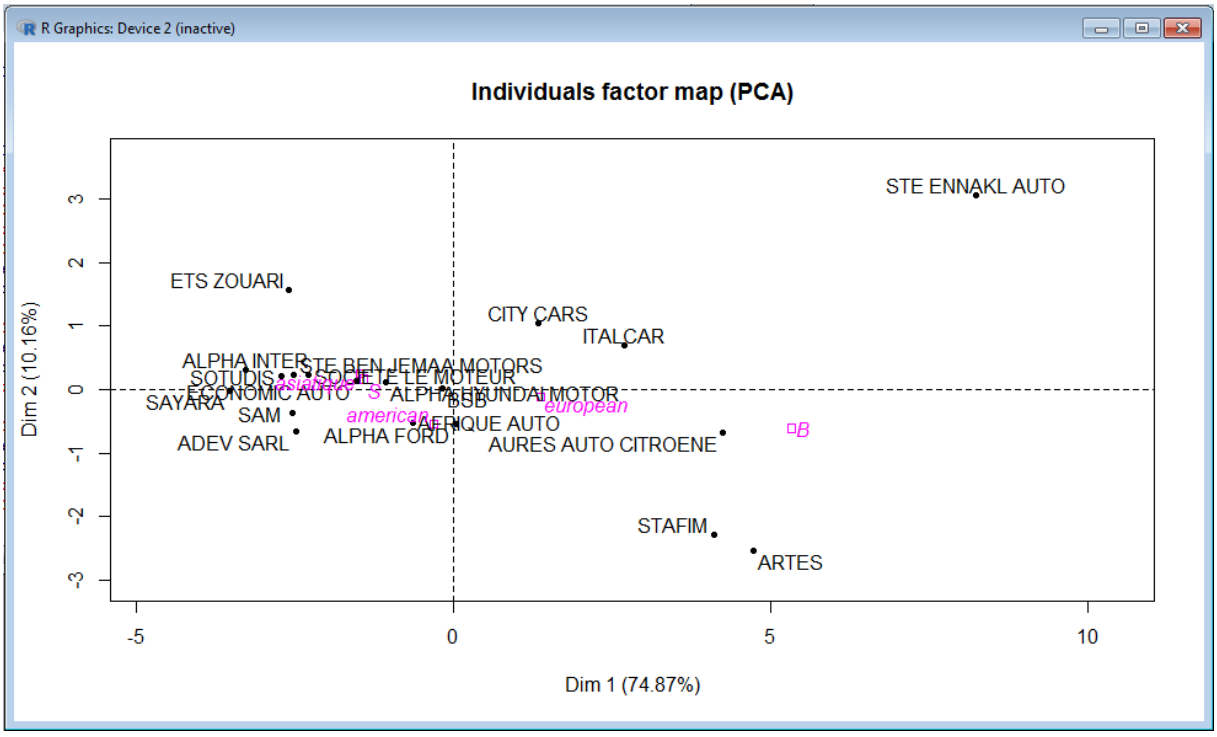

Figure 14 ACP: Individuals factor ma

For the varieties of the months that represent the quantity of cars imported are all positively correlated with axis 1.

Therefore, the first axis opposes the dealers who have a high import rate throughout the year compared to the other dealers who have a low import rate compared to the others.

- **Kmeans :**

Through the classification by k means it has been found that the origin which is closest to the European origin is the Asian origin

	class		
	1	2	3
american	0	2	0
asiatique	5	3	0
european	4	0	5

Figure 15 Segmentation Kmeans

- **Semantic and Sentimental Recognition**

we analyzed texts in the field of designation, we found that the word that repeats the most is bmw

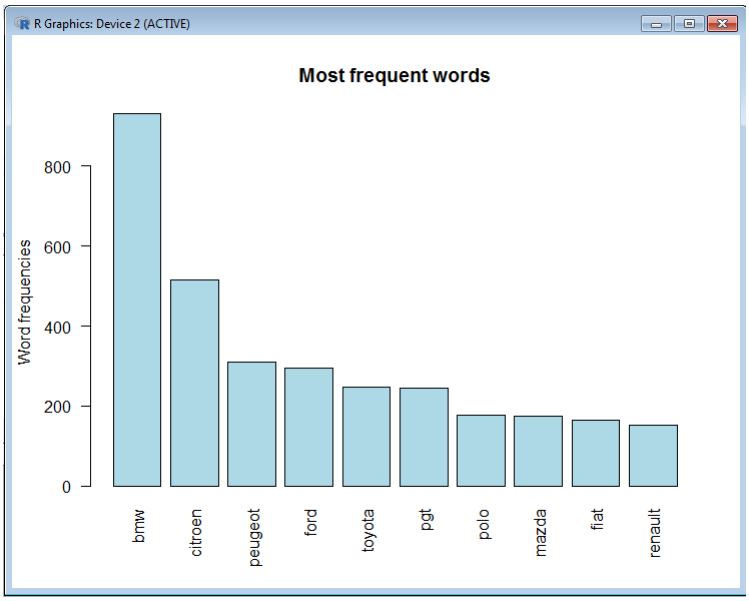

Figure 16 Text mining

It is the same results found by the word cloud using R:

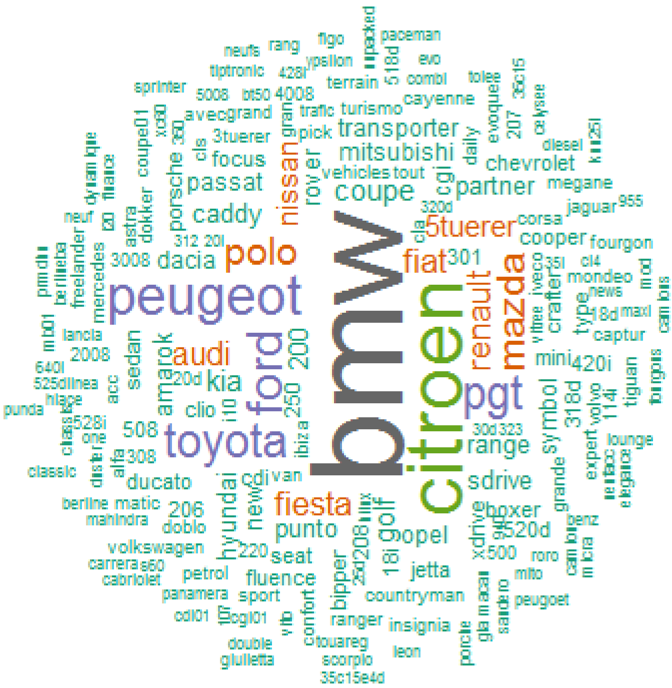

Figure 17 Text mining

- **Evolution of Import rates per month:**

This graph gives an overall view of the evolution of the series,

It shows the growth in the number of cars imported in each first half of the year and its pronounced seasonality.

For the seasonal aspect, we note that January, February and March shows the strongest activity given the arrival of the new models in this period
October, November and December are the months of lower activity, and this is repeated almost every year.

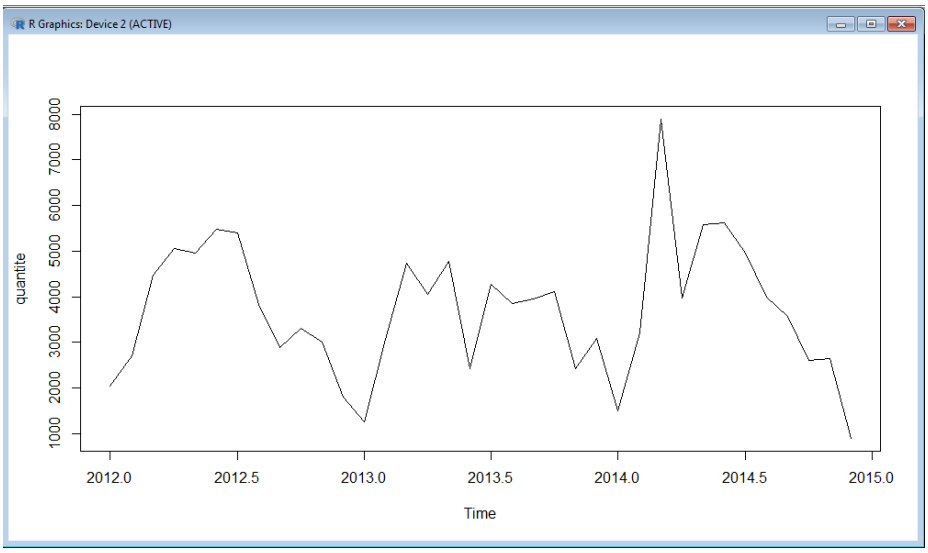

Figure 18 Plot

Data available about import rates allowed us, using some temporal series techniques, to create a model, with a certain degree of reliability, to predict the evolution of these two measures.

Import rates evolution function:

⇨ $Y(t) = 2.799 \times 10^{7} - 2.814 \times 10^{4}\, t + 7.072\, t2 - 172.4\, COS1 - 154.4\, COS2 + 101.3\, COS3 - 200.4\, SIN3 - 144.8\, SIN5$

Thanks to this modeling function, we could predict the evolution of the Import rates. This model is reliable with a rate of 82.11%.

2.5 Big Data:

Big data is a term that describes the large volume of data both structured and unstructured that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

While the term "big data" is relatively new, the act of gathering and storing large amounts of information for eventual analysis is ages old. The concept gained momentum in the early 2000s when industry analyst Doug Laney articulated the now-mainstream definition of big data as the three Vs:

✓ **Volume** : Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. In the past, storing it would've been a problem – but new technologies (such as Hadoop) have eased the burden.

✓ **Velocity** : Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time.

✓ **Variety :** Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.

✓ **Veracity :** Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

***Analyze Twitter Data:**

Twitter data in a traditional RDBMS is inconvenient, since the Twitter Streaming API outputs tweets in a JSON format which can be arbitrarily complex. In the Hadoop ecosystem, the Hive project provides a query interface which can be used to query data that resides in HDFS.

First, we need to get Twitter data into HDFS, and then we'll be able to tell Hive where the data resides and how to read it.

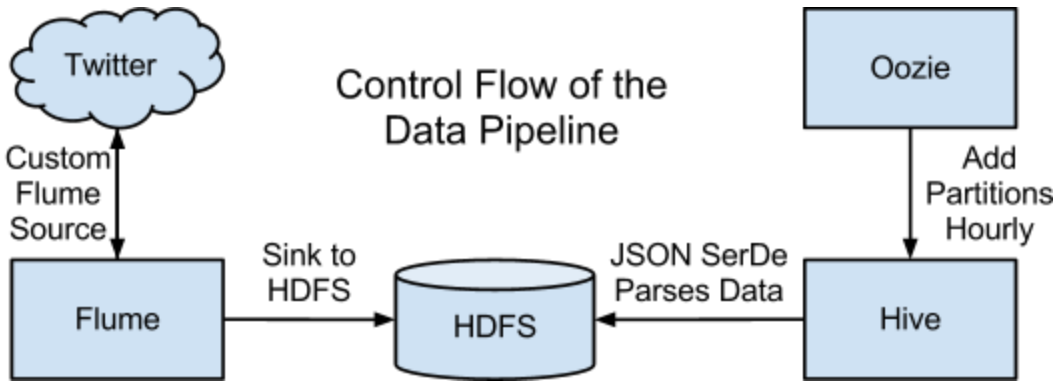

Figure 19 Control flow of data pipeline

1.Gathering Data with Apache Flume

Apache Flume is a data ingestion system that is configured by defining endpoints in a data flow called sources and sinks. In Flume, each individual piece of data (tweets, in our case) is called an event; sources produce events, and send the events through a channel, which connects the source to the sink. The sink then writes the events out to a predefined location. Flume supports some standard data sources, such as syslog or netcat. For this use case, we'll need to design a custom source that accesses the Twitter Streaming API, and sends the tweets through a channel to a sink that writes to HDFS files. Additionally, we can use the custom source to filter the tweets on a set of search keywords to help identify relevant tweets, rather than a pure sample of the entire Twitter firehose.

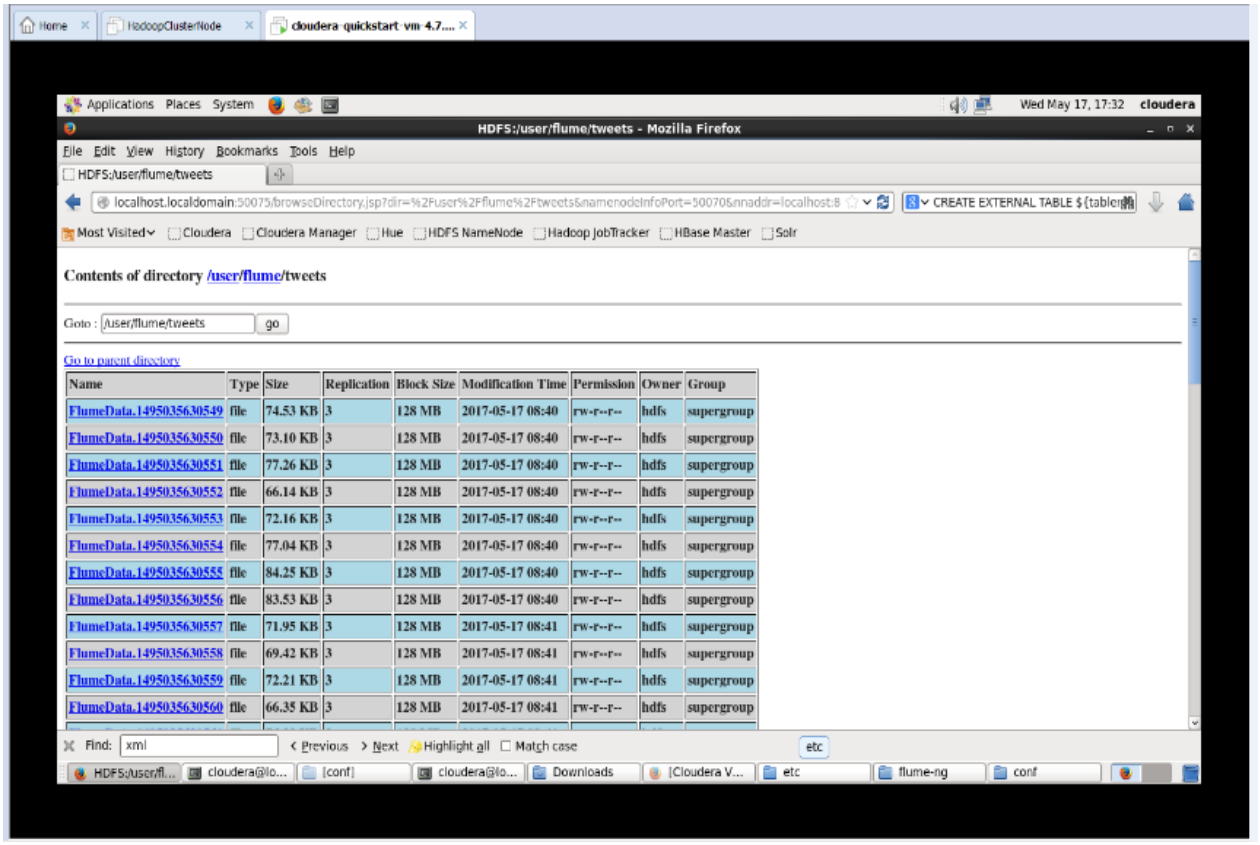

Figure 20 Data extraction with flume

2. Partition Management with Oozie

Once we have the Twitter data loaded into HDFS, we can stage it for querying by creating an external table in Hive. Using an external table will allow us to query the table without moving the data from the location where it ends up in HDFS.

To ensure scalability, as we add more and more data, we'll need to also partition the table. A partitioned table allows us to prune the files that we read when querying, which results in better performance when dealing with large data sets. However, the Twitter API will continue to stream tweets and Flume will perpetually create new files. We can automate the periodic process of adding partitions to our table as the new data comes in.

Apache Oozie is a workflow coordination system that can be used to solve this problem. Oozie is an extremely flexible system for designing job workflows, which can be scheduled to run based on a set of criteria. We can configure the workflow to run an ALTER TABLE command that adds a partition containing the last hour's worth of data into Hive, and we can instruct the workflow to occur every hour. This will ensure that we're always looking at up-to-date data.

3. Querying Complex Data with Hive

Hive allows us to flexibly define, and redefine, how the data is represented on disk. The schema is only really enforced when we read the data, and we can use the Hive SerDe interface to specify how to interpret what we've loaded.

2.6 Java application:

We have developed a java application to automate the execution of the various tasks and to manage the concessionaires (add, update, delete)

-This is the login interface Which allows access to the application

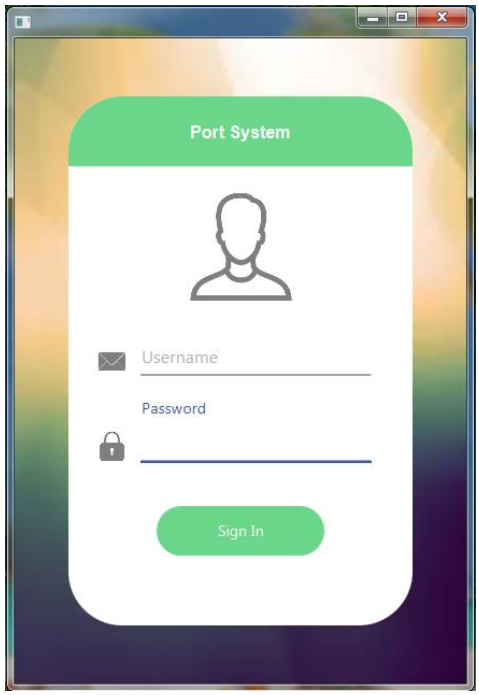

Figure 21 Login Interface

-This is the home page which allows to execute job and access the dealer management interface

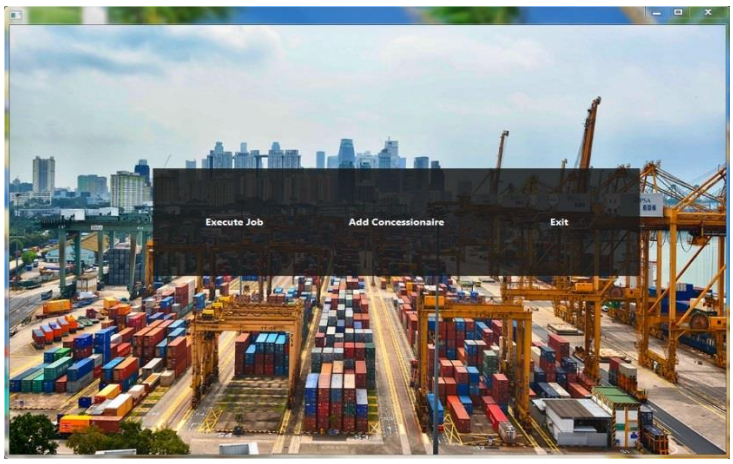

Figure 22 Home Interface

-If the task has been successfully executed, a notification is sent indicating its execution time

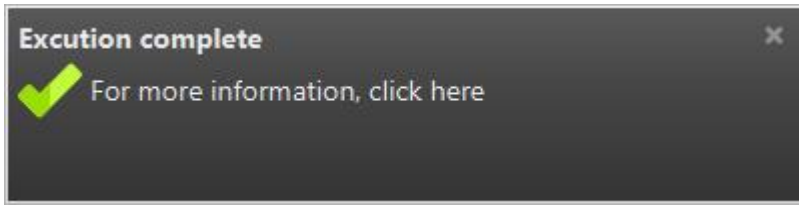

Figure 23 Notification Interface

-This interface allows us to manage dealers (add,update,delete)

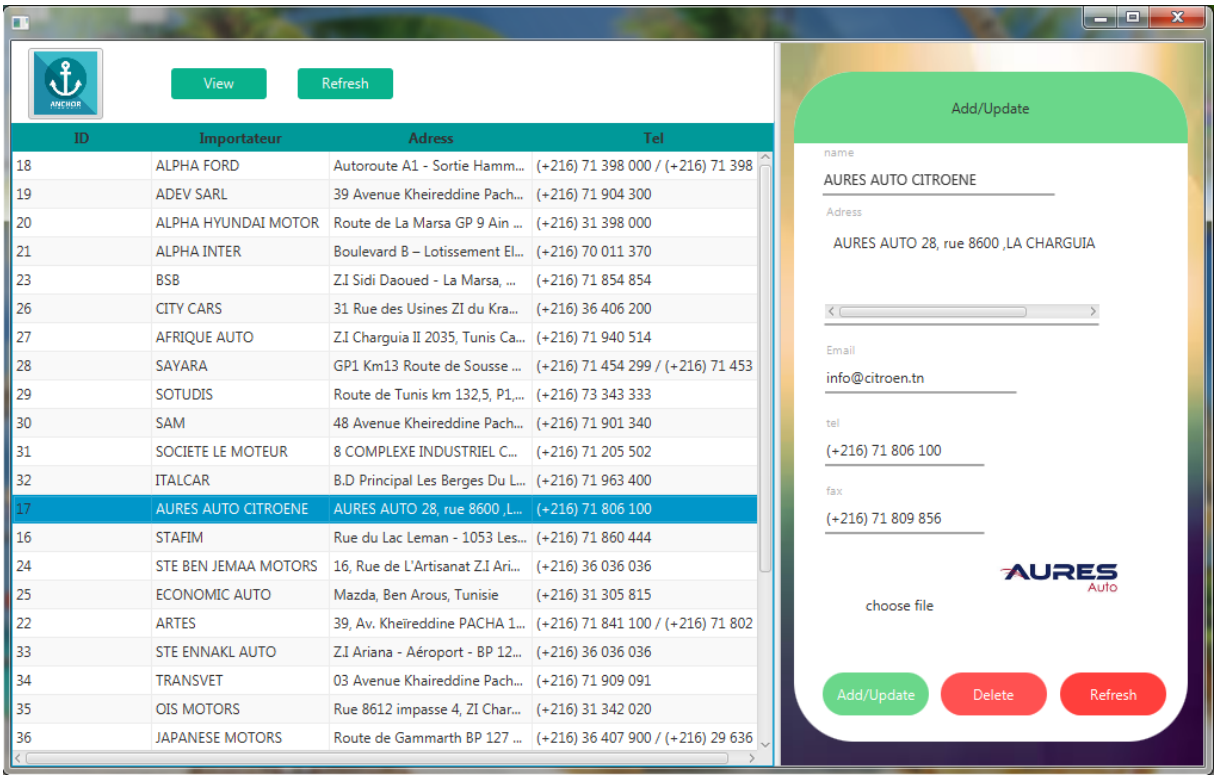

Figure 24 Dealer management

-This interface shows us the number of cars imported for dealers

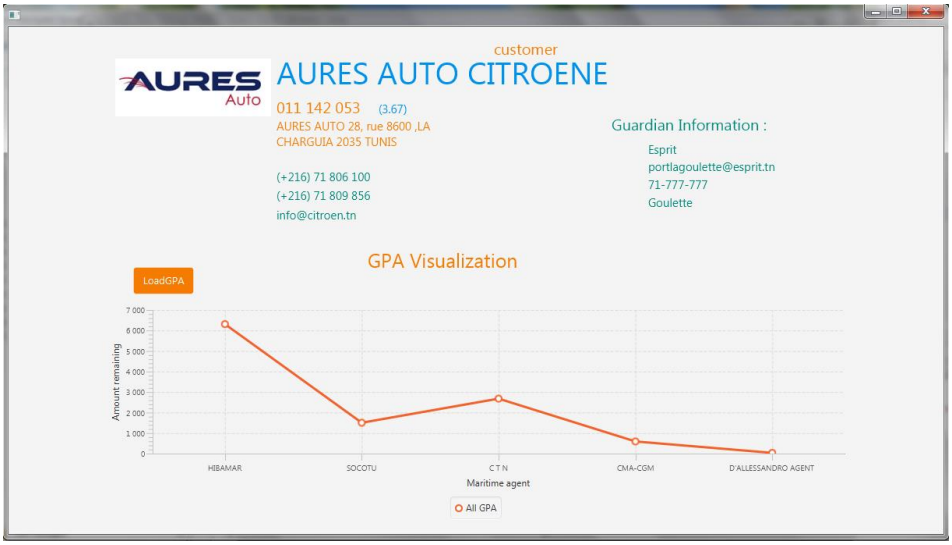

Figure 25 Quantity of imported car

Conclusion

In this last chapter, we presented the tools we used and the different steps of our solutions.

Conclusion

By the end of fourteen weeks of hard work and perseverance, we managed to finish our academic project in time in which we applied what we theoretically know in order to implement a decision support system.

During this period, we faced some technical problems due to unstable environment and lack of experience. We documented and were inspired by various tutorials. This experience was very beneficial for us on both the professional and personal level. We did not only apply the theoretical knowledge we gained, but we also improved our acquaintance in many domains.

This project gave us the possibility to interact with specialist teachers and to hear their feedback which enabled us to develop our communication skills, our ability to work in a team and above all else, to gain real world experience that we would never acquire through classes.