

2017 - 2018
GRADUATION PROJECT

NATIONAL ENGINEERING DEGREE

SPECIALTY : INFORMATION TECHNOLOGY

TITLE: Data warehouse augmentation,
Reporting process automation
&
Data analysis

By: *Majd Ben Khalifa*

Academic supervisor: *Wissal Neji*

Corporate Internship Supervisor: *David Pardo*

 **Homebell**

Supervisors ‘validation

Mr. David Pardo

--

Ms. Wissal Neji

--

Dedication

I want to express my sincere gratitude to my loving parents, Hedi Ben Khalifa and Salma Sarsout, whose without their support and their words of encouragement it would not have been possible.

My sisters Lina and Eya, and my brother Alaa also deserve some credits for being the best family a guy can ask for.

I also dedicate this dissertation to Homebell's Business Intelligence team whose have always supported me throughout the internship.

I dedicate this work and give a special thanks to whoever contributed in this work, in a way or another.

Acknowledgments

Words cannot express my gratitude towards Mr. David Pardo, my internship supervisors for his passion to share knowledge and expertise, and for his friendly state of mind.

A special thank goes to Mr. Michael O'toole, Head of Business Intelligence department at Homebell, for his supervision and assistance for my project, for the lessons, informations and valuable advices he shared with me and for his trust and his faith in my skills.

I would like to thank also Ms. Wissal Neji, my academic supervisor at Esprit for her amazing continuous support and advices during the project, and for being helpful and friendly.

Abstract

This present report is a part of the final project in order to obtain the National Diploma of Computer Engineer. Our solution must integrate all the data necessary for decision making and the follow-up of the of the company activities and of its customers. To meet these needs, we used several techniques and several technologies. Our project must allow Data Scientists to have a clear visibility and a flexible accessibility to Homebell data, allow also the decision makers to control and follow the efficiency of marketing campaigns through reports and data visualization tools. The current document present the various steps of the realization of our project.

Keywords: Marketing analysis, report automation, data warehouse, PostgreSQL, ETL, Power BI, restful API, Redshift, logistic regression, R.

Contents

General Introduction	1
CHAPTER I: PROJECT CONTEXT	2
1. Introduction	3
2. Company presentation	3
3. Company products	4
3.1. Website Event Tracker	5
3.2. In-house ETL application	6
4. Project presentation	7
5. Conclusion	8
CHAPTER II: BUSINESS UNDERSTANDING & REQUIREMENTS' SPECIFICATION	9
1. Introduction	10
2. Identification of actors	10
3. Requirements specification	10
3.1. Functional requirements	10
3.2. Non-functional requirements	11
4. Use case	12
5. Project plan	14
6. Methodology and software development life cycle	15
7. Conclusion	15
CHAPTER III: PROJECT ENVIRONMENT	16
1. Introduction	17
2. Logical system architecture	17
3. Tools	17
3.1. Node JS application for ETL	17
3.2. Database tools	18
3.2.1 Amazon simple storage service (S3)	18
3.2.2 Amazon Redshift	19
3.3. Reporting tools	19
3.3.1 Microsoft Power BI	19
3.3.2 Google Sheets	20

3.3.3 R	20
3.4. Project management tools	21
3.4.1 Atlassian Jira	21
3.4.2 Atlassian Confluence	22
4. Conclusion	22
CHAPTER IV: DATA MODELING, ETL & INTEGRITY TESTING	23
1. Introduction	24
2. Data modeling	24
3. ETL jobs	25
3.1. Raw data tables	26
3.2. ETL scripts	27
4. Integrity testing	31
5. Running the ETL jobs	33
6. Conclusion	35
CHAPTER V: DATA EXPLOARION	36
1. Introduction	37
2. Power BI reporting	37
2.1. Existing reporting system	38
2.2. New reporting system	39
2.3. Reports & dashboards	41
3. Google Sheets reporting	45
3.1. Existing reporting system	45
3.2. Reports	47
4. Conclusion	50
CHAPTER VI: DATA ANALYSIS	51
1. Introduction	52
2. Cancellation reasons analysis	52
2.1. Preface	52
2.2. Descriptive analysis	53
2.2.1 Cancellation by partners	54
2.2.2 Cancellation by agents	55
2.2.3 Cancellation by city	56
2.2.4 Cancellation by city & verticals	57

2.3. Features importance approximation	58
2.3.1 Pre-book cancellation dataset	59
2.3.2 Post-book cancellation dataset	61
3. Conclusion	62
General Conclusion	63
Webography	64

List of figure

- Figure 1: Homebell in numbers
- Figure 2: Choosing the job type
- Figure 3: Choosing the job surface 5
- Figure 4: Estimated price for the job 5
- Figure 5: Region partner availability 5
- Figure 6: Event tracker representation 6
- Figure 7: position of the BI system in Homebell 8
- Figure 8: use case diagram 12
- Figure 9: Project planning chart 14
- Figure 10: Logical system architecture 17
- Figure 11: Node JS logo 18
- Figure 12: Amazon S3 logo 18
- Figure 13: Amazon Redshift logo 19
- Figure 14: Power BI logo 19
- Figure 15: Google Sheets logo 20
- Figure 16: R logo 20
- Figure 17: TidyVerse packages 21
- Figure 18: Jira logo 21
- Figure 19: Confluence logo 22
- Figure 20: Marketing datamart model 24
- Figure 21: Data flow diagram 25
- Figure 22: lead dimension test script 31
- Figure 23: adgroup_utm_map test script 31
- Figure 24: Channel dimension test script 32
- Figure 25: event_session_map test script 32
- Figure 26: facts.marketing test script 33
- Figure 27: JSON manifeste script 34
- Figure 28: power BI data model 38
- Figure 29: publish data model from desktop to service 38
- Figure 30: Streaming dataset creation 40
- Figure 31: Power BI script 40
- Figure 32: Month over month marketing metrics 41
- Figure 33: Number of leads per marketing sub-channel per day 42

Figure 34: Distribution of number of leads per marketing sub-channel per day 43

Figure 35: Distribution of number of leads per marketing sub-channel per day 43

Figure 36: Sales dashboard 44

Figure 37: Customers reviews report 45

Figure 38: Conversion rates per verticals report 47

Figure 39: Sales overview report 47

Figure 40: Landing page analysis report 48

Figure 41: Top view marketing report 49

Figure 42: Lead age in the funnel report 49

Figure 43: Cancellation by partners results 54

Figure 44: Cancellation by partners results summary 55

Figure 45: Cancellation by agents results 55

Figure 46: Cancellation by agents results summary 56

Figure 47: Cancellation by city results 56

Figure 48: Cancellation by city results summary 57

Figure 49: Cancellation by city & verticals results 58

Figure 50: Cancellation by city & verticals results summary 58

Figure 51: Features importance on the pre-book dataset 59

Figure 52: Logistic regression on the pre-book dataset 60

Figure 53: Decision Tree's accuracy on the pre-book dataset 60

Figure 54: Features importance on the post-book dataset 61

Figure 55: Logistic regression on the post-book dataset 61

Figure 56: Decision tree accuracy on the post-book dataset 62

List of tables

TABLE 1: DETAILS OF ACTORS USE CASES 13

TABLE 2: RAW DATA TABLES 26

TABLE 3: JSON MANIFESTE SCRIPT DESCRIPTION 34

General Introduction

Business Intelligence served companies and organizations for the previous years to explore and mine the new world's oil: Data.

In today's customer-centric, digital-first world, many business owners and managers are bombarded with 'information overload' and are urgently seeking ways to derive greater control, understanding and intelligence from their organization's data.

Nowadays, almost all the fields require using Business intelligence in a way or another. Any company wants to make the right decisions at the right moment, cares about its business efficiency, risk and cost reductions, will have to deal with data. Therefore, data analytics is no longer a luxury, it became a must.

Over the past few years, more than 90% of all the world's data has been created. SQL is improving and turning faster and more efficient, NoSQL databases are becoming more and more popular and big data infrastructures such as Spark and Hadoop are chipping away large-scale data processing problems. Many fast and affordable solutions are emerging and consequently, we begin diving in the data oceans instead of drowning.

Hence, the Big question is no longer about how are we going to face and explore those endless flows of data, but instead it's becoming "How smart are we exploring those data flows?" Nowadays, new ways to intelligize how we are dealing with data emerged and imposed themselves, and here comes the Artificial Intelligence (AI) on the top of the most important trends in the big data world in particularly, and in the world generally.

No one can remain indifferent to the major changes happening across every segment of society. Actually, people are playing an important part in the intelligent transition we are facing and have a big responsibility and opportunity that needs to be shaped for the best.

Chapter I : Project Context

1. Introduction

A few years ago almost all the application editors of Business Intelligence focused exclusively on large companies. Currently, Small and medium-sized enterprises are also investing in Business Intelligence solutions. This is the case of Homebell, which despite not being an IT company, has invested in having a good and reliable business intelligence system in order to get a place and keep it in a competitive market. Business intelligence and data management enable small and medium businesses such as Homebell to have control over the company's resources, maintain a channel of communication with clients, align business process with its strategy, determine the situation of the organization, and also to know in which direction the company should focus its efforts.

Generally SMEs make "Information Management" with the tools they have available. Most often simple Excel table is enough to satisfy their basic needs in "reporting" and statistics. Nevertheless, it becomes difficult to use this kind of technology when it comes to perform more complex analyzes. Business Intelligence comes in in the information system of a company in order to increase the control of process internally and at the same time its ability to predict variations in the market. [1]

2. Company presentation:

Homebell is a home improvement and decoration startup based in Berlin, founded late 2015 by Sasha Weiler and employs more than 200 people from more than 40 nations. Homebell is Germany's fastest-growing platform for home decoration. In less than 2 years, Homebell expanded to the Dutch market and has now over 200 craftsmen partner all over Germany and Netherlands. The key of success of Homebell is a strong belief from the higher management in information technology and marketing for the success of a business. Being a customer based business, Homebell backbone is the sales team. But the success of the sales team alone cannot ensure the success of the business as whole. Homebell rely on IT and online marketing to create better interaction channels with the customer as well as the partners and better understand its cash flow to preserve precious resources. [2]

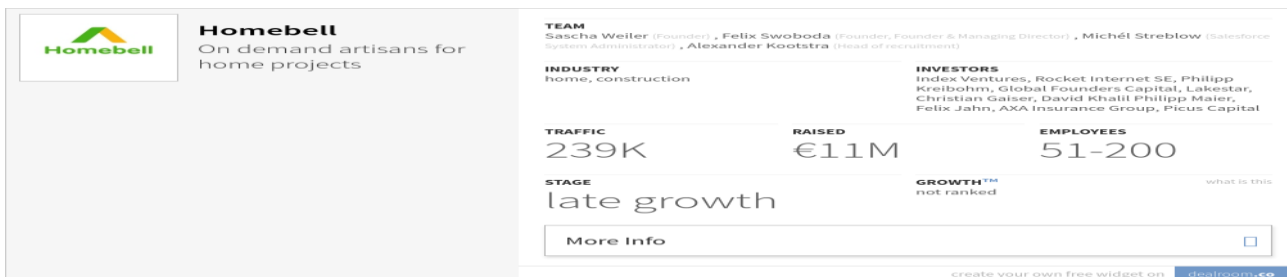


Figure 1: Homebell in numbers

3. Company Products:

As a technology-enabled company, Homebell is enabling its customers with home improvement and decoration reservation system online. The customer can put all the project information and gets a real-time costs calculation and craftsmen available, making it easily accessible from online and mobile channels. In the same Platform, Craftsmen can become partner as well with a few clicks. New Partners can leverage the reach of Homebell's reservation platform to boost their customer base, drive utilization in low-traffic times and generate additional business on-demand. Homebell provides its customers the easiest way to explore the different choices and options for home improvements and decorations as painting, floor papering, parquet, laminate, wall tiles and other different services. With just a few clicks, the home improvement project can be initiated and customer can get the price and the timeframe for of the project. A list of close by partners is also notified so they can apply for the project. A partner is chosen based on a score associated to each partner based on previous project and customers feedbacks. Headquartered in Berlin/Germany, with international footprints in the Netherlands and Luxembourg, the founder's vision is to grow Homebell to the world's leading marketplace for home improvements. [2]

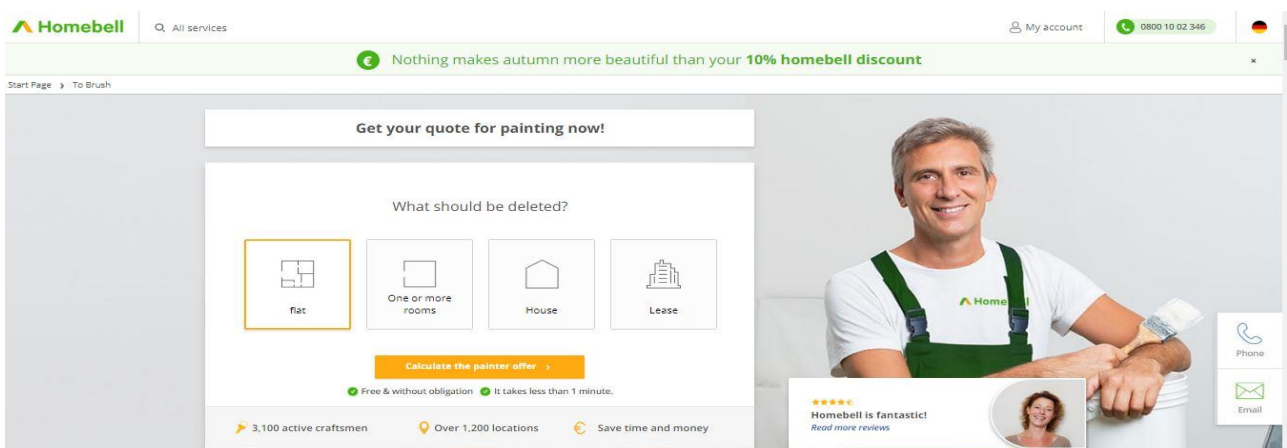


Figure 2: Choosing the job type

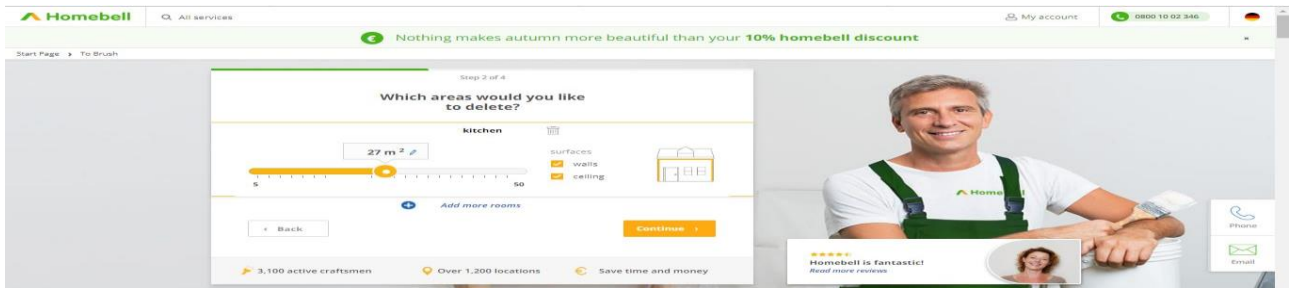


Figure 3: Choosing the job surface

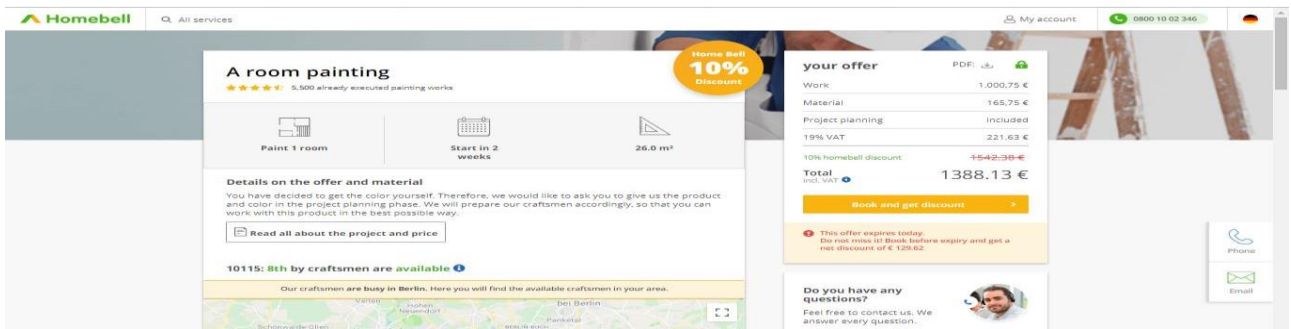


Figure 4: Estimated price for the job

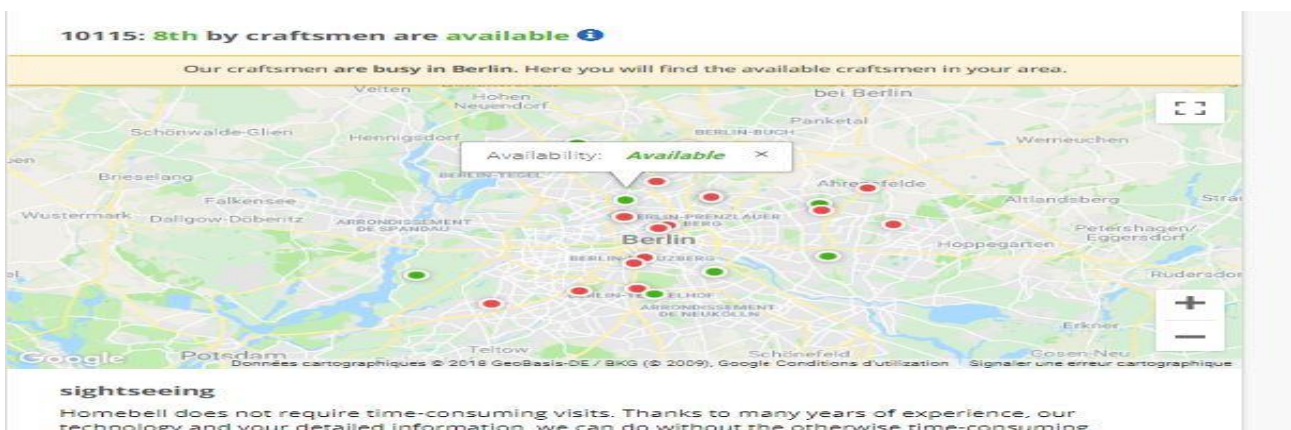


Figure 5: Region partner availability

3.1 Website Event tracker:

Homebell has its own event tracking tool which is build and maintain by a developers team in India. The event tracker is a node JS application based on the event data modeling. This Homebell web site event tracker tracks user interactions with content that can be tracked independently from a web page or a screen load. Downloads, mobile ad clicks, gadgets, Flash elements, AJAX embedded elements, and video plays are all examples of actions you might want to track as Events.

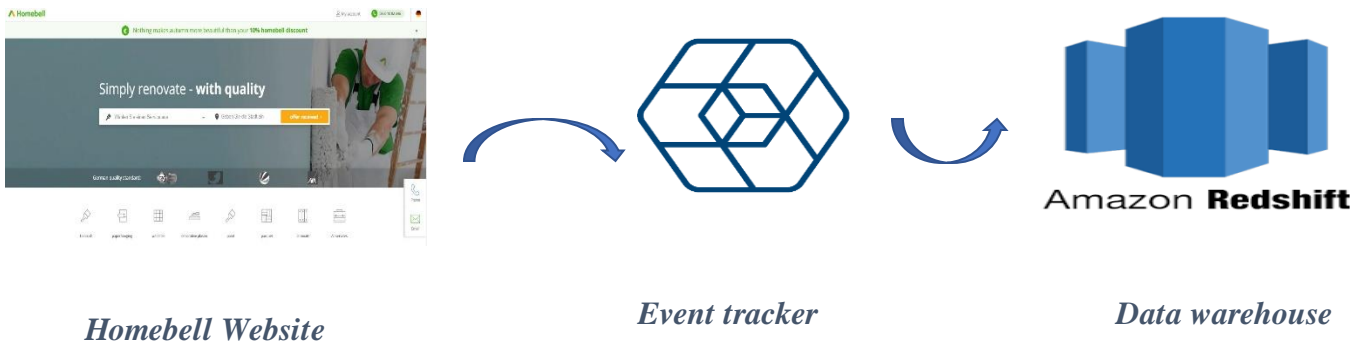


Figure 6: Event tracker representation

The Event Tracker is built over the event data modeling which is the process of using business logic to aggregate over event-level data to produce modeled data that is simpler for querying.

The data produced by the event tracker goes to the atomic database in redshift, which should become the basis for the marketing datamart in the next chapter. [3]

3.2 ETL in-house Application:

Homebell has its own ETL in-house real time ETL tool written in Node.js. Node is an excellent tool for ETL pipelines, because it's ability to handle data in streams instead of batch processing. It is the right tool to use especially when working with distributed database systems, like in our case AWS Redshift and event-driven applications, where businesses process data in real time and at scale. Therefore, using Node.js application for ETL enable us to embed stream processing directly into each service, and core business applications can rely on a streaming platform to distribute and act on events. The library that enables this stream process of data in Node.js is socket.io, which allows a very simple synchronous communication in the application, that is to say real-time communication. [4]

The application offers a front end application which enables us to query the redshift database, search the result set of the query, download the result set as a CSV file, saving the query in a MongoDB database and loading the saved queries from the mongo dB database.

In the backend, the application does the real time ETL in the following way:

- **Extract - Load:** The application extracts data as it is and load it into our redshift database. The app extracts these data from the following sources: Salesforce, Google Sheets, Google analytics and data from the event tracker.
- **Transform – Load:** the app then enables us to do the transformation needed to go from the raw data loaded in the previous steps into our desired data warehouse schema. This is done by running manifest JSON files which defines the order of running our SQL file which define the structure of our dimensions and facts tables.

4. Project presentation:

Homebell is a sales startup. Which mean it primarily relies on the sales team to make benefits. At the beginning, how Homebell operates is by getting a contact list of leads from leads provider and getting it to sales team to call the potentials leads and get deals. As Homebell starts growing, they start getting the own leads by marketing campaign, essentially TV ads and social media ads, and this is when this projects starts.

Marketers rely heavily on data to determine where to place their campaigns, whom to target, and how to best allocate their resources. To be effective, they need to properly turn raw numbers into actionable insights that will enhance their strategies. [5]

Now, Homebell is investing in enhancing more and more a data-driven culture within the company, especially the new marketing team by collecting more data, analyzing it and share it across departments, with as many employees as possible integrating data into their daily business running to make more evidence-based decisions.

As part as the Business Intelligence team and working closely with the new marketing team, my main role is to work on the following tasks:

- ❖ Designing the new marketing datamart.
- ❖ Creating new SQL scripts and manifest JSON files and integrating them within the already existing real time ETL application.
- ❖ Creating tests to check the integrity of the new facts and dims tables
- ❖ Automating the use of Google sheets for reporting purposes.

- ❖ Creating new marketing reports dashboard based on the team request using Google sheets and power BI.
- ❖ Creating some deep dive analysis Using R to dig more into business questions.

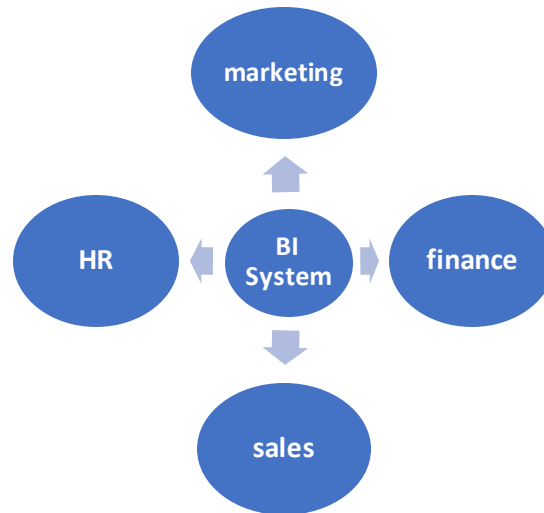


Figure 7: position of the BI system in Homebell

5. Conclusion

In this chapter, we have presented the project context by de defining the company products and business goals and we have presented the motivation of Homebell to rely on data to create its own leads by creating their own marketing team. In the next chapter, we are going to present in more details the project requirements and the business understanding behind it.

Chapter II : Business Understanding & requirements' specification

1. Introduction

The requirements analysis and specification phase presents a fundamental step in the development cycle of a project. Indeed, it makes it possible to study the feasibility of the project and produces the contract between the future users and the designers. It consists of the expression, the collection and the formalization of the needs of the potential customers and all the constraints. To do this, we have adopted a specific approach to achieve a well-organized, robust and scalable solution.

In this chapter, we will present the specification of the functional and non-functional requirements as well as the detailed description of the expected use cases.

2. Identification of actors

This project involves two major actors:

- ❖ The Business intelligence developer : His main tasks are the design of the datamarts, creating ETL jobs, creating integrity test scripts for the data warehouse, creating and automate the process of reporting, creating dashboard on request, creating deep dive analysis and data models.
- ❖ Decision markers: This is the final user who will use Google-sheets, Power BI and R to visualize the reports created by the BI developer.

3. Requirements specification

3.1 Functional requirements:

The functional requirements means to express the services offered to the user and the functionalities in response to the customer's request.

Our solution must offer the following features:

❖ **Datamart design:**

- A new design of the marketing datamart integrated with the already existing datamarts.

❖ **Extract – Transform - Load :**

- Extract the missing data into the row data in the redshift database.
- Create SQL script and ETL jobs (JSON manifest files) to extract the data from the raw data into the fact and dims tables.

❖ **Integrity testing :**

- Create SQL scripts to test the integrity of the data with each ETL jobs running.

❖ **Create Reports and dashboard:**

- Create reports for the marketing team based on their daily requests.
- Automate the process of updating the Google sheets reports
- Create dashboard on Power BI to use them instead of relying on Google analytics dashboards.
- Create company KPI for the higher management.

3.2 Non-functional requirements:

The non-functional requirements describe the features and quality criteria of the solution in order to make the functional requirements operational because they act as constraints on the solutions, but taking them into consideration avoids several inconsistencies in the system.

- ❖ **Performance:** ETL Jobs must work in a consistent, scheduled and error-free way, through the use of a job scheduler.

- ❖ **Response time and optimization:** The large volume of data storage involves a significant response time. We opted for the elimination of useless data, as well as the commands that make the loading of the data more cumbersome.

- ❖ **Security:** Our solution must respect especially the confidentiality of reports (each report should be accessible only to whom it concerns).

4. Use Case:

In this part, we expose the global use case diagram that represents the functionalities of an actor.

The interest of the global use case diagram is to give a global vision on the various functionalities provided by the system attributed to the different actors.

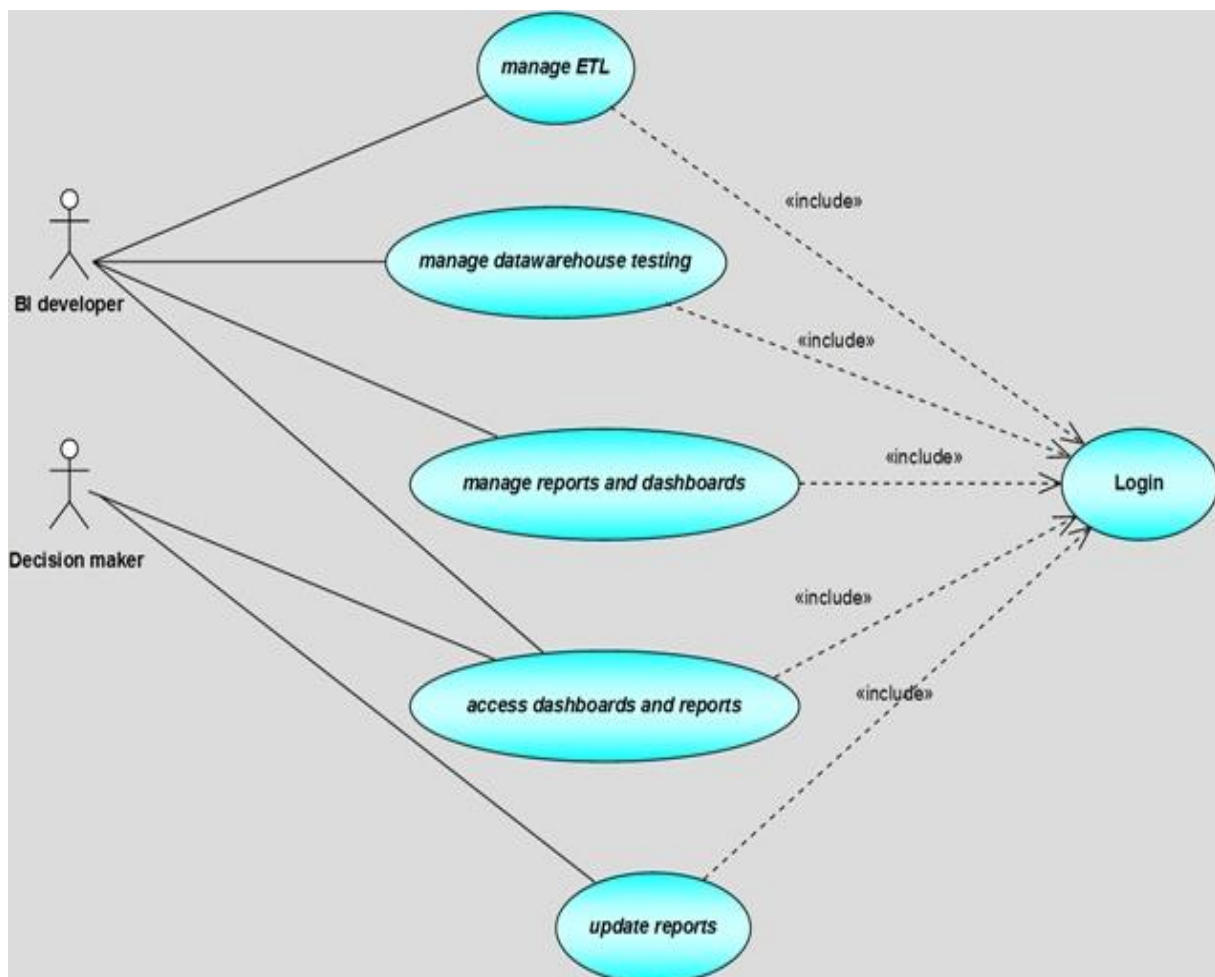


Figure 8: use case diagram

Use case	Actor	Description
Manage ETL	BI developer	Creating and updating ETL jobs (SQL scripts, Manifest JSON files).
Manage Datawarehouse Testing	BI developer	Creating and updating Test scripts.
Manage Reports and dashboards	BI developer	Creating and updating dashboard and reports (Google Sheets, Power BI).
Access reports and dashboards	Decision makers	Allows managers and decisions makers to access the reports and dashboards created.
Update reports	Decision makers	Allows managers and reports users to update the reports without the need to a BI developer intervention
Login	BI developer / decision makers	All the above mentioned use cases require an authentication using the Homebell mail account

Table 1: Details of actors use cases

5. Project plan :

In the part, we list the phases to be executed in the project in a time scale with outputs and dependencies between the phases.

The project plan is dynamic due to the startup nature of the company. Therefore, it is compulsory to review the achievements and the progress at the end of each phase and update the project plan accordingly.

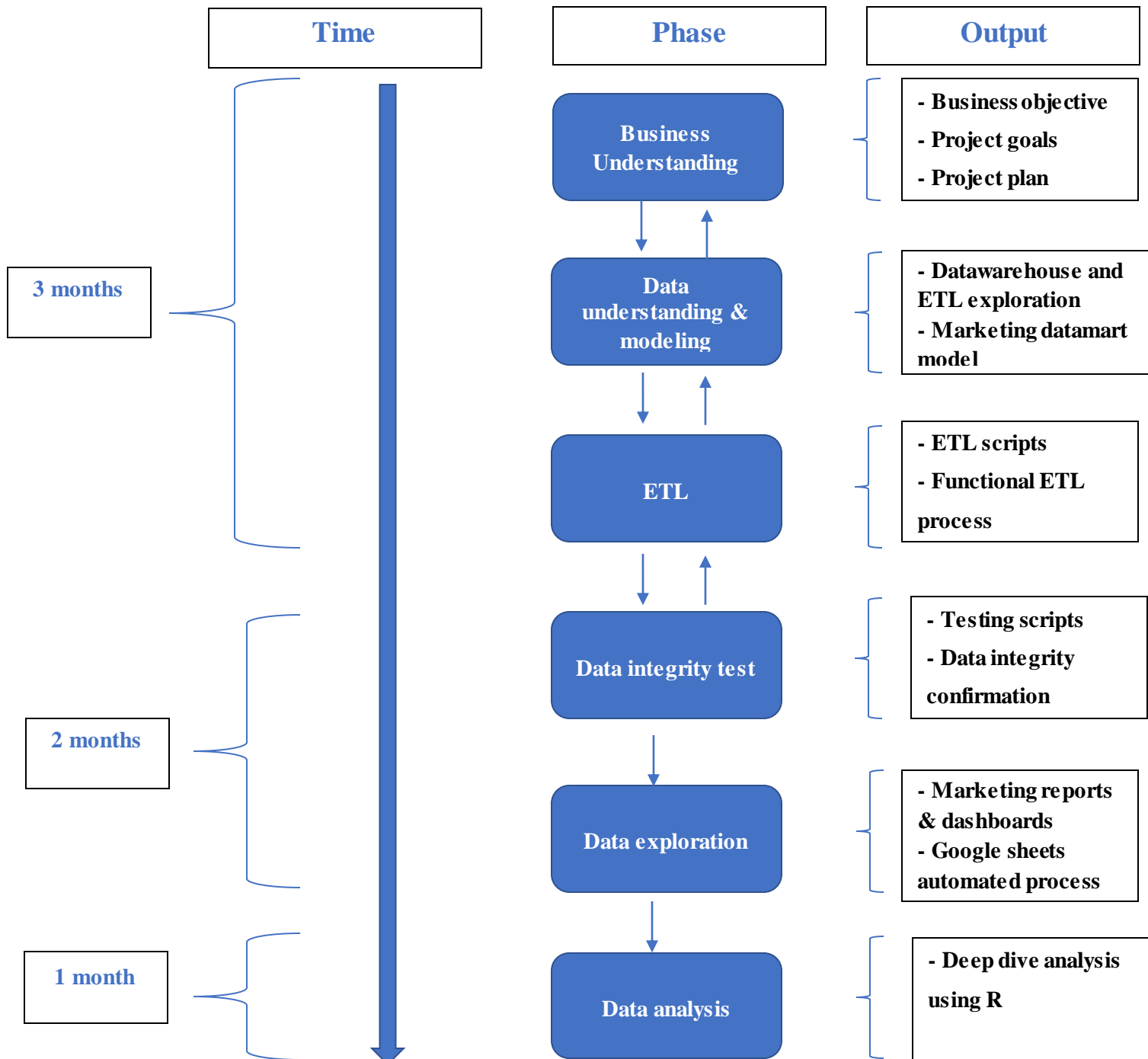


Figure 9: Project planning chart

6. Methodology and software development life cycle

For the development of a marketing business intelligence system, we are going to follow the bottom-up approach of Ralph Kimball. The Data Warehouse can be seen, according to him, as the union of datamarts coherent with each other thanks to the shared dimensions between them. In the bottom-up approach, datamarts are therefore created to provide reports and analytical capacity dedicated to specific business processes (marketing in our case). This choice is understandable since we have an existing data warehouse and we are going to integrate a marketing datamart within this data warehouse.

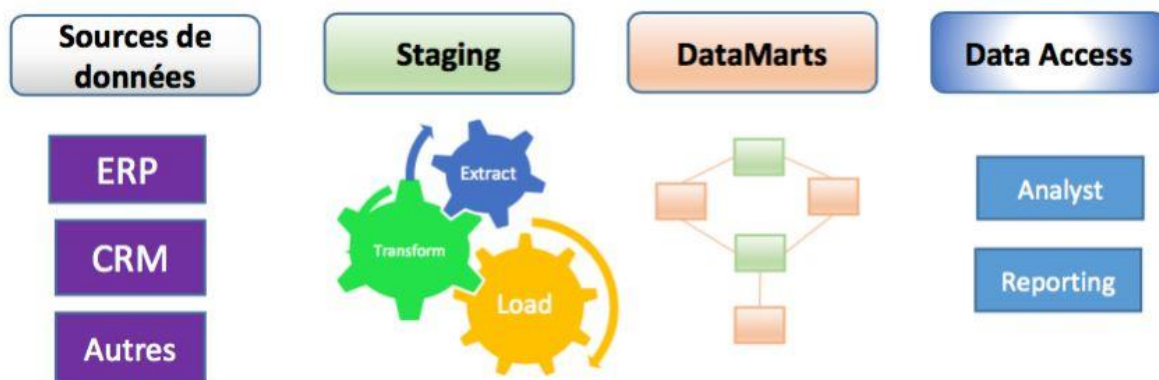


Figure 9: Ralph Kimball approach

7. Conclusion

In this chapter, we defined the functional and non-functional requirements of our project along with identifying the actors of the system and the project road map. Therefore, the implementation of the decision solution can be initiated.

The next Chapter will be about the project environment and the different tools and platforms used during the realization of this project.

Chapter III: Project Environment

1. Introduction

After having defining the project context and the business requirements that this project has to deliver, we will present in the following chapter the methodology adopted for the development. Thereafter we present the different tools and frameworks used to achieve our business goals.

2. Logical system architecture

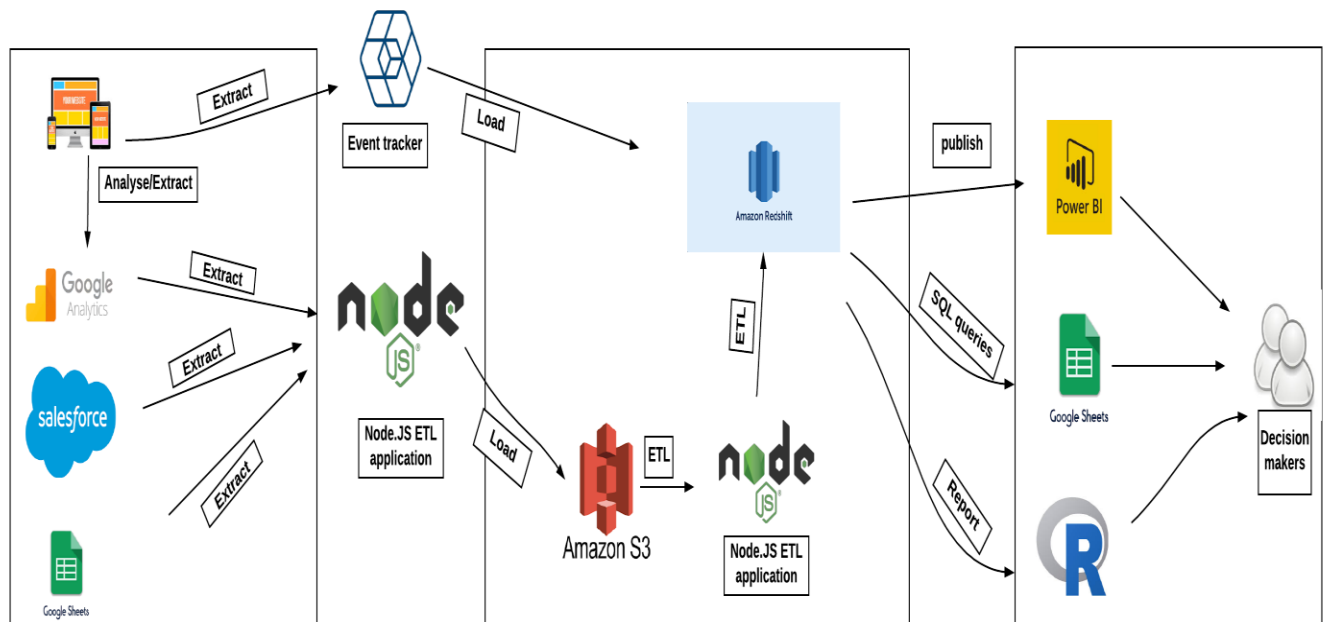


Figure 10: Logical system architecture

3. Tools:

3.1 Node JS application for ETL:

Node.js is a JavaScript runtime. It is basically a JavaScript language with an enormous package libraries to perform actions such as writing to standard output, open / close network connections or create a file.

It is built on top of the Chrome V8 Engine, which uses an asynchronous event-driven model that can be used for creating scalable web applications.

One of the big reasons to use Node.js for ETL (Extract, Transform, Load) is because of its asynchronous nature. If you have hundreds of rows of data that need to be transformed and transmitted, Node.js can quickly process each one with non-blocking calls. This way your scripts are doing more processing and less waiting around. Another reason to use Node JS is that it works natively with JSON which is a common data-interchange format used by many APIs. As soon as you load a JSON data source into Node.js, it's already formatted in JavaScript Object Notation, ready to be read and manipulated as needed. [6]



Figure 11: Node JS logo

3.2 Databases tools:

3.2.1 Amazon Simple Storage Service (Amazon S3)

Amazon Simple Storage Service (Amazon S3) is a scalable storage web service designed for online backup and archiving of data and application programs. S3 lets you store virtually any file or object up to five gigabytes (5 GB) in size. Amazon.com does not limit the number of items a subscriber can store. The data is stored in redundant servers distributed in different data centers. S3 has a web interface and encrypts user authentication. [7]

In our project, S3 is our production database, the first storage system of data coming from the different data sources.



Figure 12: Amazon S3 logo

3.2.2 Amazon Redshift

Amazon Redshift is a fast and fully managed data warehouse service on the cloud based on PostgreSQL. It allows you to analyze data in a simple and economical way using existing business intelligence tools and standard SQL syntax.

Data is loaded into our Redshift data warehouse by our Node.js ETL application. All of our reporting tools such as Power BI and Google Sheets connects to this data warehouse in order to build reports and dashboards. [8]



Figure 13: Amazon Redshift logo

3.3 Reporting Tools:

3.3.1 Power BI:

Power BI is a data analysis solution developed by Microsoft to enable organizations to aggregate, analyze, and visualize data from multiple sources. The solution consists of a software named Power BI Desktop and a system on the cloud.

The paid version of Power BI allows you to plan the update of the data. It also includes Power BI Gateway which is a gateway for connecting live to, for example, a database located on a remote server. This feature can be very useful in our case, especially to automate the process of updating the reports. [9]



Figure 14: Power BI logo

3.3.2 Google Sheets

The Google Sheets app is an online spreadsheet that allows you to create, format, and collaborate on spreadsheets. As a reporting tool, Google sheets is important especially for sales reports, when the decision makers needs to see the exact numbers and go deeper into the details. A part of this project is to automate the process of updating the Google sheets reports.



Figure 15: Google Sheets logo

3.3.3 R

R is a programming language and open source software for statistics and data science supported by the R Foundation for Statistical Computing.

The R language is widely used by statisticians, data miners, data scientists for statistical software development and data analysis.

One of the most used R package in our project is Tidyverse.

The tidyverse is a coherent system of packages for data manipulation, exploration and visualization that share a common design philosophy.

The following figure illustrates a canonical data science workflow, and shows how the individual packages fit in. [10]



Figure 16: R logo

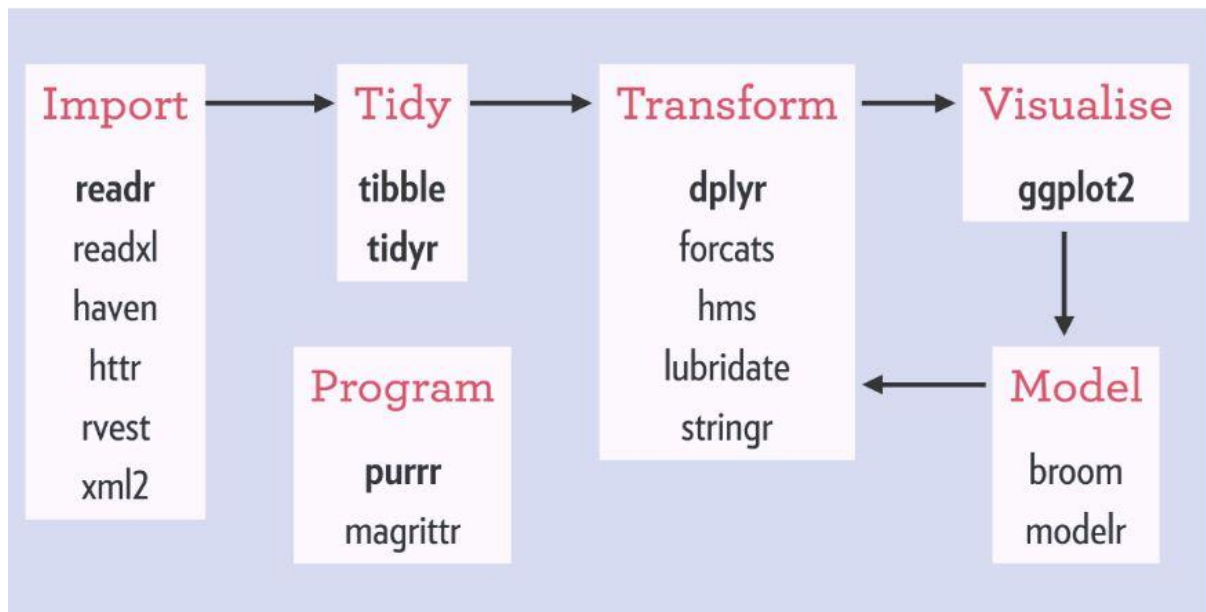


Figure 17: TidyVerse packages

3.4 Project management tools

3.4.1 Atlassian JIRA

Jira is Atlassian's product. It is a cross platform issue and bug tracking software with advanced project management capabilities and features. It is suitable for any company size and is extremely valuable to all collaborating teams, stakeholders and project managers.

JIRA helps your team share information and engage others seamlessly, manage sprints, display issues on agile boards with custom agile workflows, display work in progress limits and check their efficiency planning and assign members with certain tasks. It also allows the ability to work together with colleagues using joint-editing tools and monitor the team's progress and updates of each task. [11]



Figure 18: Jira logo

3.4.2 Atlassian Confluence

Like Jira, Confluence is developed and marked by Atlassian Company, Confluence is a wiki for team collaboration. Team members can create, share, and collaborate within their team and the other members of the company. Practically, we use Confluence for a wide variety of purposes. Some of the most popular uses include:

- Intranet: internal company portal
- Extranet: we can make some documentation public to our partners, if needed
- Project documentation: we use Confluence internally to document and collaborate for projects we work on, document our business intelligence processes and the abbreviation of the different KPIs. [12]



Figure 19: Confluence logo

4. Conclusion

During this chapter, we presented the development tools that contributed to the realization of our project, we also presented the logical architecture of the information system in Homebell. This architecture is the basis on which we will realize the design of the system.

Chapter IV :

Data modeling, ETL & Integrity Testing

1. Introduction

The purpose of this ETL phase is to determine precisely our marketing data model according to the business requirements described above, to identify and define the different data sources that we already have and present the different steps that enables us to move from raw data into structured and integrated data ready for analysis.

2. Data modeling

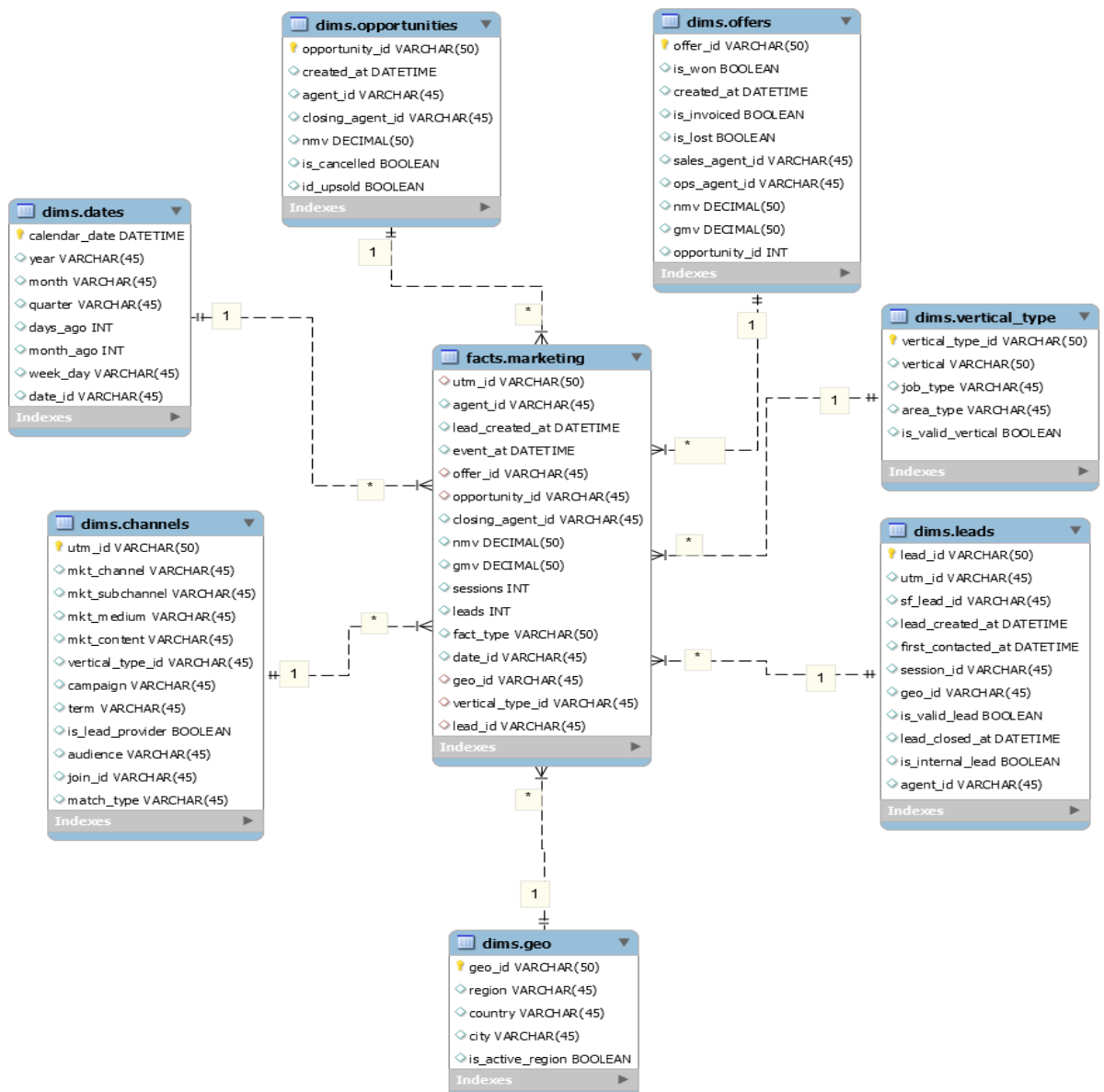


Figure 20: Marketing datamart model

3. ETL jobs:

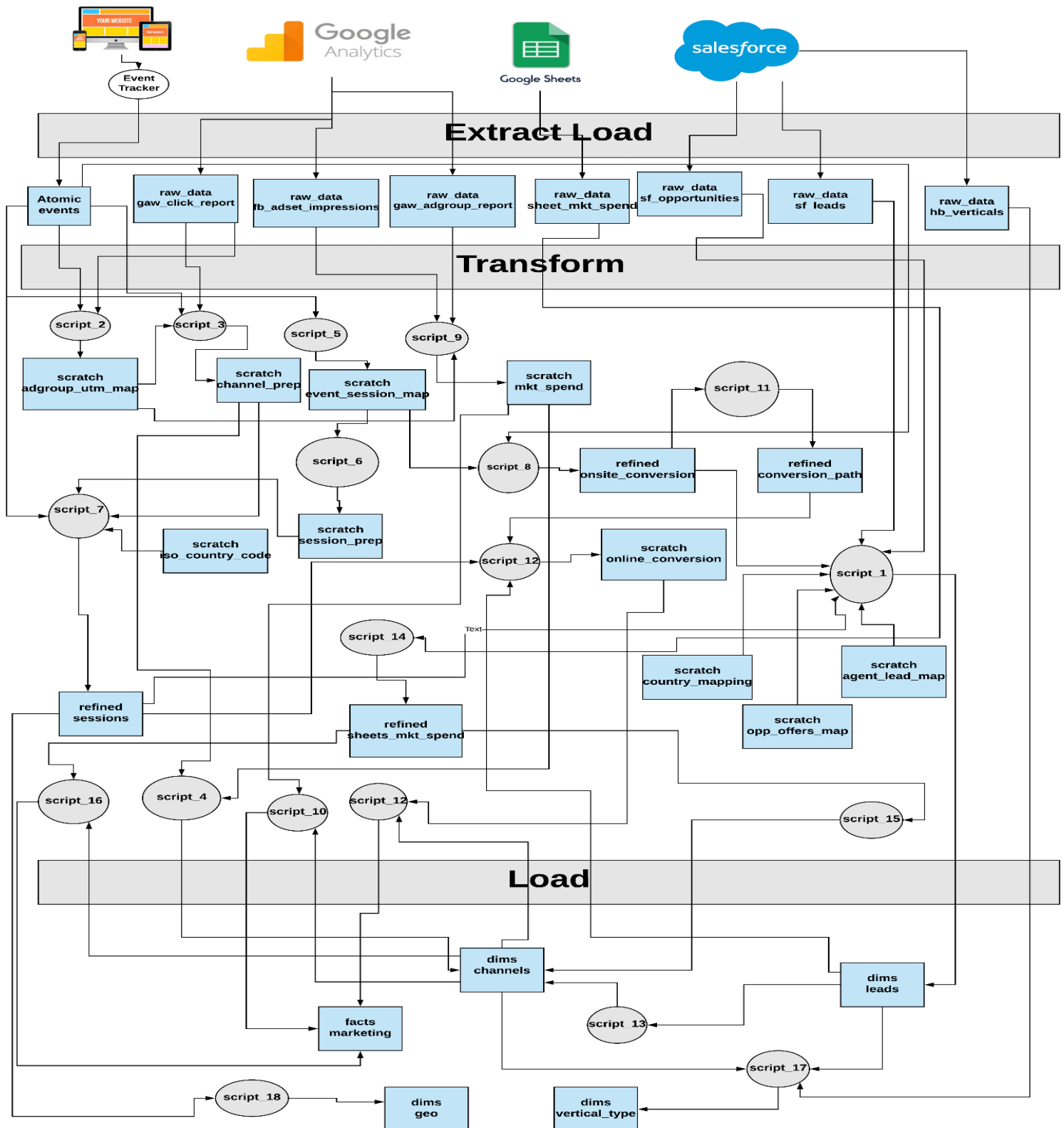


Figure 21: Data flow diagram

3.1: Raw data tables:

Table	Description
Atomic_event	The first stage of storage of data coming from our website event tracker.
Raw_data.gaw_click_report	Contains data coming from Google Analytics Website traffic about the different clicks.
Raw_data.fb_adset_impressions	Contains information coming from facebook API about the different facebook campaigns.
Raw_data.gaw_adgroup_report	Contains information coming from Google analytics about the different google campaigns.
Raw_data.sheet_mkt_spend	Data coming from a google sheet report manually updated by the marketing team which contains some channels costs informations (yahoo, taboola..)
Raw_data.sf_opportunities	Contains data coming through our SaleForce API about all the opportunities.
Raw_data.sf_leads	Contains data coming through our SalesForce API about all the leads.
Raw_data.hb_verticals	Contains all the active job verticals of Homebell (lackering, flooring, painting..)
Scratch.opp_offer_map	Contains informations that maps offers to opportunities.
Scratch_agent_lead_map	Contains informations that maps sales agents to leads.
Scratch_iso_country_codes	Contains informations that maps countries and cities to their corresponding ISO code.

Table 2: raw data tables

3.2: ETL scripts:

Our ETL application has JavaScript files called runners. Those files takes into an argument a JSON file which define the order of the ETL SQL scripts.

The runner then post the query of the script to our redshift database. The scripts that we are going to mention are the content of our RebuildMarketing.json file, which when called by the runner, rebuild our marketing datamart.

➤ Script_1:

We start by running the first script which is responsible for filling the dims.leads. The reason why start with the dims.leads is that the leads we have in Homebell are not only marketing leads but also internal and lead providers (as discussed in a previous chapter). The marketing fact table is going to have all the leads and not only the marketing leads and we need this dimension to fill other marketing dimension (channels and verticals in next scripts).

➤ Script_2:

This script loads data into scratch.adgroup_utm_map which map the data we have from atomic.events and raw_data.gaw_click_report using gaw.gcl_id = a.mkt_clickid as a join columns between the two tables.

➤ Script_3:

This script loads data into the scratch.channel_prep table which prepares data for the dims.channels. In the tables we create the utm_id which we are going to use later to join to the dims.channels and facts.marketing. we also prepare marketing data such as campaign, content, city_style..

➤ Script_4:

This scripts creates the dims.channels table. Two table are used in the creation of this dimension. The first one is scratch.channel_prep. And then we call the script

another time using `scratch.mkt_spend` which is GAW data. The second call is just in case some information was missing from the first call.

➤ **Script_5:**

This is the script responsible for turning our events based table (`atomic.events`) into a session based table (`scratch.event_session_map`).

We can aggregate events into sessions using `domain_sessionid`, which is generated by our event tracker and increments when a user was not active for 30 minutes. The output of this table is a session based table. Each session is defined with a `session_id`.

➤ **Script_6:**

In this script we take the table created in the previous step (`scratch.event_session_map`) as input and we select the `start_event` and the `end_event` for each session. The output is `scratch.sessions_prep`.

➤ **Script_7:**

This is the last step of defining our sessions. The output of this script is the `refined.sessions` table which at this step contains the `utm_id` (information about the channels which each session comes from).

➤ **Script_8:**

The output of this script is the `refined.onsite_conversions` table which contains the sessions and the events related to each session which turned into a potential lead (`tr_orderid` IS NOT NULL). The `tr_orderid` information that we have in the `atomic.events` table means that a session actually made an order in the website which means there is a potential lead.

➤ **Script_9:**

In this script we create the `scratch.mkt_spend` table which extracts data from the `GAW` and `facebook raw_data`. The created table contains informations about the marketing cost for each channel and also building the `utm_id` for each channel the same way we built in our `dims.channels` table.

➤ **Script_10:**

In this script we create the first part of our `facts.marketing` table which is the marketing cost part. The script extracts the data from the previously created table (`scratch.mkt_spend`) and load it in the `facts.marketing` with `fact_type` is "cost". The sum of cost per marketing channel (joining with the `dims.channels` using `utm_id`) gives us how much money did we spent for each marketing channel.

➤ **Script_11:**

This script extracts data from `refined.onsite_conversions` and load it with additional columns to `refined.conversion_path` without duplicates. It defines a `conversion_path_id` for each leads converted.

➤ **Script_12:**

This script creates the `scratch.online_conversions` table which takes information about the lead converted from `refined.conversion_path` and associate it with the `scratch.lead_summary`. This way for each lead created we have all the metrics that we are going to use for analysis later. (`nmv`, `offers`, `opportunities`, `gmV`.) And then it loads this information into the `facts.marketing` table with "conversion/lead" as a `fact_type`. In this step we also associate the informations with the `dims.channels` table. This way for each lead converted we have all the information about the marketing channels and also all the metrics concerning the opportunities and offers created for this lead.

➤ **Script_13:**

This script loads data into the `dims.channels` table. The difference between this script and `script_4` is that this one extracts data from `dims.leads`. This way we have information about all the non-marketing channels (internal, lead providers.) It also acts as a backup to the `script_4` as it catches the marketing channels that we failed to catch in that script.

➤ **Script_14:**

Some of the marketing channels (yahoo, taboola, bing..) has different costs (CPL: cost per click) each month. This costs are manually added by the marketing team. This script extracts this costs from `scratch.sheets_mkt_spend` and load them into `refined.sheets_mkt_spend` with the appropriate `utm_id`.

➤ **Script_15:**

This script update the `dims.channels` with the new manual marketing costs from the `refined.sheets_mkt_spend` table.

➤ **Script_16:**

This script extracts the new marketing manual costs data from the `refined.sheets_mkt_spend` and `dims.channels` tables and loads it into the `facts.marketing` with “`mkt_spend_manual`” as fact type.

➤ **Script_17:**

This script extracts data from `raw_data.hb_verticals`, `dims.leads` and `dims.channels` to create the `dims.vertical_type` table. It contains the different type of job vertical that Homebell operates in (lacquering, wallpapering, flooring, painting).

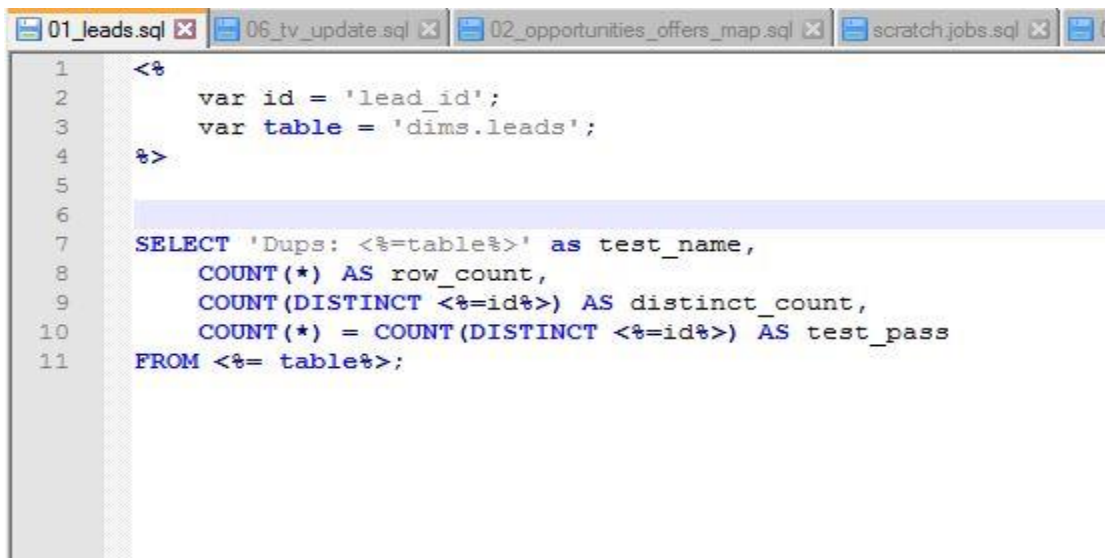
➤ **Script_18:**

This script creates `dims.geo` by extracting data from the `refined.session` table.

4. Integrity tests:

The integrity test consists of SQL scripts that runs after the ETL jobs. Some of SQL ETL script has their corresponding SQL Test scripts which has the same name. Since with don't have no null constraints when it comes to ids (every field is nullable), the integrity test are there to make sure we have and id for each row, and of coure only for the tables that needs an id for each row.

➤ Test_script_1:

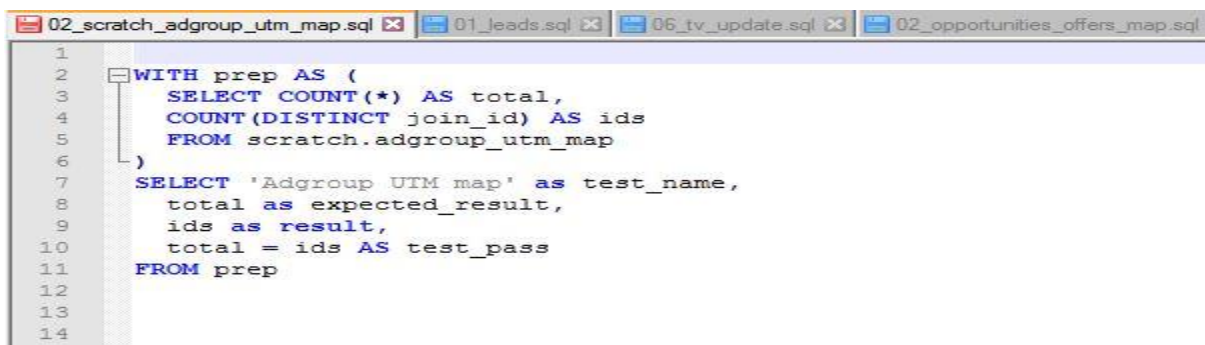


```
1  <%
2      var id = 'lead_id';
3      var table = 'dims.leads';
4  %>
5
6
7  SELECT 'Dups: <%=table%>' as test_name,
8      COUNT(*) AS row_count,
9      COUNT(DISTINCT <%=id%>) AS distinct_count,
10     COUNT(*) = COUNT(DISTINCT <%=id%>) AS test_pass
11  FROM <%= table%>;
```

Figure 22: lead dimension test script

This script above counts the number of rows and the number of lead_id and it returns false if the numbers are not the same (some rows missing lead_ids)

➤ Test_script_2:

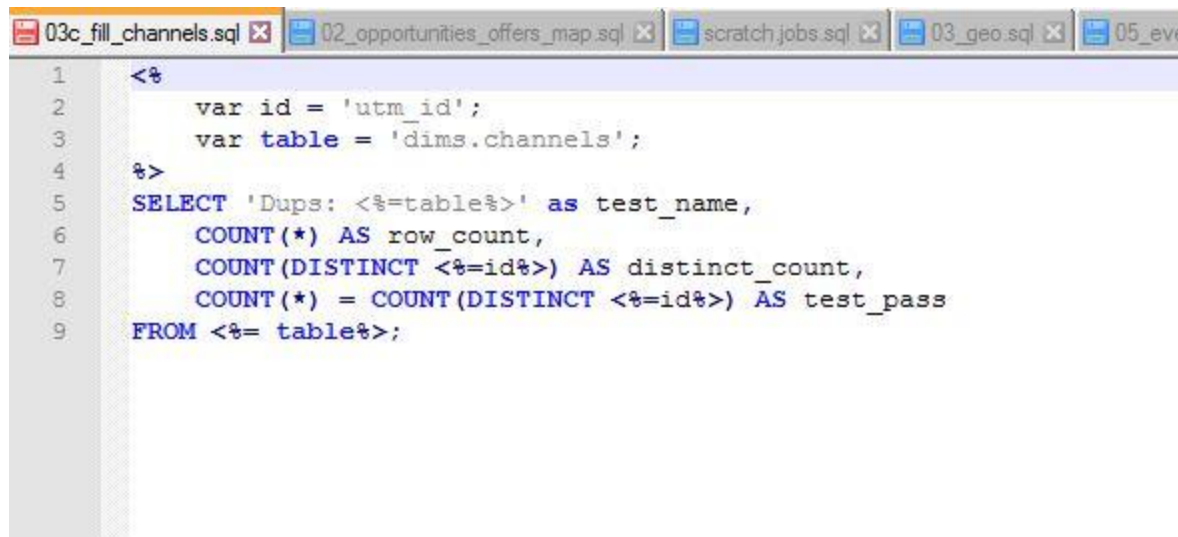


```
1
2  WITH prep AS (
3      SELECT COUNT(*) AS total,
4      COUNT(DISTINCT join_id) AS ids
5      FROM scratch.adgroup_utm_map
6  )
7  SELECT 'Adgroup UTM map' as test_name,
8      total as expected_result,
9      ids as result,
10     total = ids AS test_pass
11  FROM prep
12
13
14
```

Figure 23: adgroup_utm_map test script

This script above counts the total number of rows and the number of join_id (account_id||campaign_id||ad_group_id) which should be the same. Each row represents a specific marketing channel with its relevant utm.

➤ Test_script_4:



```

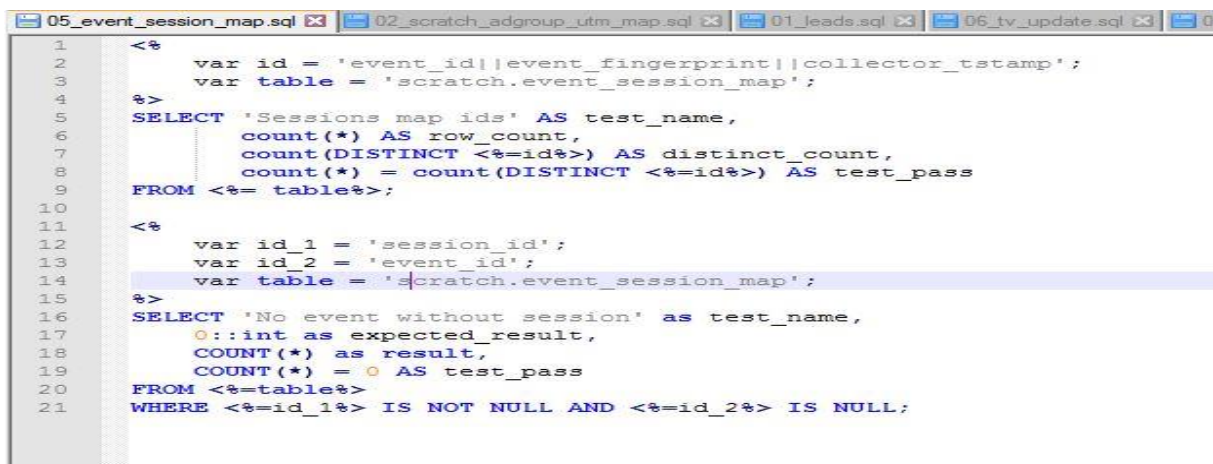
1  <%
2      var id = 'utm_id';
3      var table = 'dims.channels';
4  %>
5  SELECT 'Dups: <%=table%>' as test_name,
6      COUNT(*) AS row_count,
7      COUNT(DISTINCT <%=id%>) AS distinct_count,
8      COUNT(*) = COUNT(DISTINCT <%=id%>) AS test_pass
9  FROM <%= table%>;

```

Figure 24: Channel dimension test script

This script above, as the previous script, counts the total number of rows and the number of utm_id (which is the id of each specific channel) in the dims.channels table and returns false if we have missing utm_ids.

➤ Test_script_5:



```

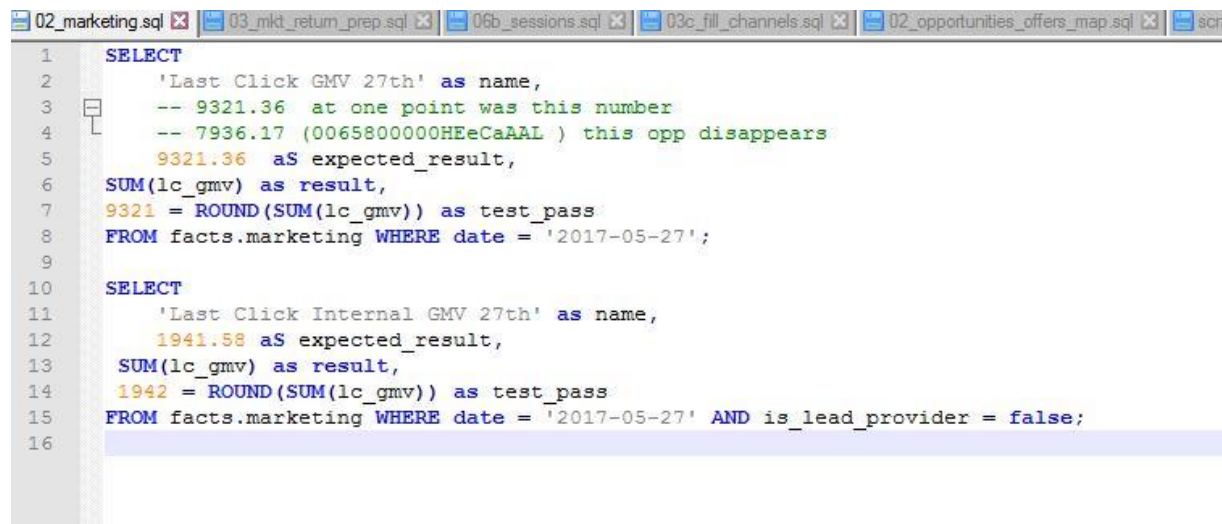
1  <%
2      var id = 'event_id||event_fingerprint||collector_tstamp';
3      var table = 'scratch.event_session_map';
4  %>
5  SELECT 'Sessions map ids' AS test_name,
6      count(*) AS row_count,
7      count(DISTINCT <%=id%>) AS distinct_count,
8      count(*) = count(DISTINCT <%=id%>) AS test_pass
9  FROM <%= table%>;
10
11  <%
12      var id_1 = 'session_id';
13      var id_2 = 'event_id';
14      var table = 'scratch.event_session_map';
15  %>
16  SELECT 'No event without session' as test_name,
17      0::int as expected_result,
18      COUNT(*) as result,
19      COUNT(*) = 0 AS test_pass
20  FROM <%=table%>
21  WHERE <%=id_1%> IS NOT NULL AND <%=id_2%> IS NULL;

```

Figure 25: event_session_map test script

This script above checks the number of ids and the number of rows in the scratch.event_session_map table. It also checks that there is no event without its corresponding session_id.

➤ Test_script_6:



```
1  SELECT
2      'Last Click GMV 27th' as name,
3      -- 9321.36 at one point was this number
4      -- 7936.17 (0065800000HECaAAL ) this opp disappears
5      9321.36 as expected_result,
6      SUM(lc_gmv) as result,
7      9321 = ROUND(SUM(lc_gmv)) as test_pass
8  FROM facts.marketing WHERE date = '2017-05-27';
9
10 SELECT
11     'Last Click Internal GMV 27th' as name,
12     1941.58 as expected_result,
13     SUM(lc_gmv) as result,
14     1942 = ROUND(SUM(lc_gmv)) as test_pass
15 FROM facts.marketing WHERE date = '2017-05-27' AND is_lead_provider = false;
16
```

Figure 26: facts.marketing test script

This script is a test that we use run manually in our front end up for a specific date to check the sum of our metrics (nmv, gmv..). The sum of nmv for example for a specific date should be the same in the facts.marketing and facts.sales_revenue.

5. Running the ETL jobs:

The different SQL jobs mentioned above are the ETL jobs we use to create the marketing datamart. The way they are called in the ETL application is through a JSON manifest file. The JSON manifest file define the order of the running the SQL scripts as we discussed earlier.

```

"version": "1.0.0",
"title": "Event Loader",
"sequence": [
  {
    "t": "Opp Ranking",
    "path": "00_prep/01_offer_version_ranking.sql",
    [
      {
        "t": "Opp map",
        "path": "00_prep/02_opportunities_offers_map.sql",
        "test": true,
      },
      {
        "t": "inbound calls",
        "path": "facts/08_inbound_calls_ld.sql",
      },
      {
        "t": "outbound calls",
        "path": "facts/09_outbound_calls_ld.sql",
      }
    ],
    {
      "t": "Leads Conversions",
      "path": "dime/01_leads.sql",
      "test": true,
    },
    {
      "t": "r",
      "path": "dime_scratch/scratch.pipeline_events.sql",
    },
    {
      "t": "r",
      "path": "dime_scratch/scratch.img_events.sql",
    },
    {
      "t": "START",
      "path": "01_events/00_start_load.sql",
    },
    {
      "t": "Duplicates",
      "path": "01_events/01_duplicates.sql",
    },
    [
      {
        "t": "Adgroup_UTMS_MAP",
        "path": "01_events/02_scratch_adgroup_utm_map.sql",
        "test": true,
      },
      {
        "t": "User Mapping",
        "path": "01_events/03a_user_map.sql",
      }
    ],
    {
      "t": "Prep Channels",
      "path": "01_events/03b_channels_prep.sql",
      "params": {
        "recreate": true,
      },
    },
    {
      "t": "Channels - Events",
      "path": "01_events/03c_fill_channels.sql",
      "params": {
        "recreate": true,
        "sourceTable": "scratch.channel_prep"
      },
      "test": true
    },
  ],
  {
    "title": "Events Sessions Map",
    "path": "01_events/05_event_session_map.sql",
    "params": {
      "recreate": true,
    },
    "test": true
  },
  {
    "title": "Sessions Prep",
    "path": "01_events/06a_session_prep.sql",
    "params": {
      "recreate": true,
    },
  },
  {
    "title": "Sessions",
    "path": "01_events/06b_sessions.sql",
    "params": {
      "recreate": true,
    },
    "test": true,
  },
  {
    "title": "Onsite Conversions",
    "path": "02_marketing/01_onsite_conversions.sql",
  },

```

Figure 27: JSON manifeste script

Key	Value
"T"	The title of the iteration
"Path"	The path of the SQL file
"test"	Takes true or false. It define if the SQL file has a corresponding test file.
"recreate"	Takes true or false. It define that the table in the SQL file needs to be recreated or not.
"sourceTable"	Some of the SQL scripts uses different table in each run, such as the fill_chnnels script.

Table 3: JSON manifeste script description

6. Conclusion

In this chapter we have presented the ETL phase of our project. The in-house ETL application enables to execute SQL scripts into our Redshift database. The application also enables us to schedule this ETL jobs to run automatically in an interval of time, giving us near real time ETL.

The different report and analysis created based on our datamart are discussed in the next chapter.

Chapitre V :

Data

exploration

1. Introduction

In our modern world, the visualization of information has evolved and is used to communicate information of all kinds: economic, social, military, health, etc. Graphical representations are the best way to make decisions by facilitating to grasp difficult concepts or to identify new patterns. They help to better process information by reducing the requirements on attention, working memory and long-term memory. In a business environment, visualizations can have two broad goals, which sometimes overlap: Explanatory and Exploratory. Data explanation allows solving specific problems and the exploration of large data sets allows understanding it better. Generally speaking, graphical representations are used to gain better insight of the problem we are studying.

2. Power BI reporting

Power BI is a Business Intelligence solution and an analytics tool developed by Microsoft to enable organizations to aggregate, analyze, and visualize data from multiple sources as well as securely share information (insights) gained through analytics in the form of reports and dashboards. This solution focuses on self-service to enable all employees to understand the data and exploit it. It also unifies all data sources. This facilitates the creation of a data model.

Power BI is actually composed of two platforms:

- ❖ Power BI desktop: enables you to connect to, transform, and visualize your data. With **Power BI Desktop**, you can connect to multiple different sources of data, and combine them (often called modeling) into a data model that lets you build visuals.
- ❖ Power BI Service (cloud): Most users who work on Business Intelligence projects use **Power BI Desktop** to create reports, and then use the **Power BI service** to share their reports with others. In addition to that, the web service provides the ability to create dashboards. [13]

2.1 Existing reporting system:

The existing reporting system consist of creating using Power BI desktop to connect to our Redshift database and create a data model. After creating the data model in power BI desktop, we can use the publish feature to publish this data model to Power BI service which creates a dataset that we can use to create reports and dashboards. See figures below.

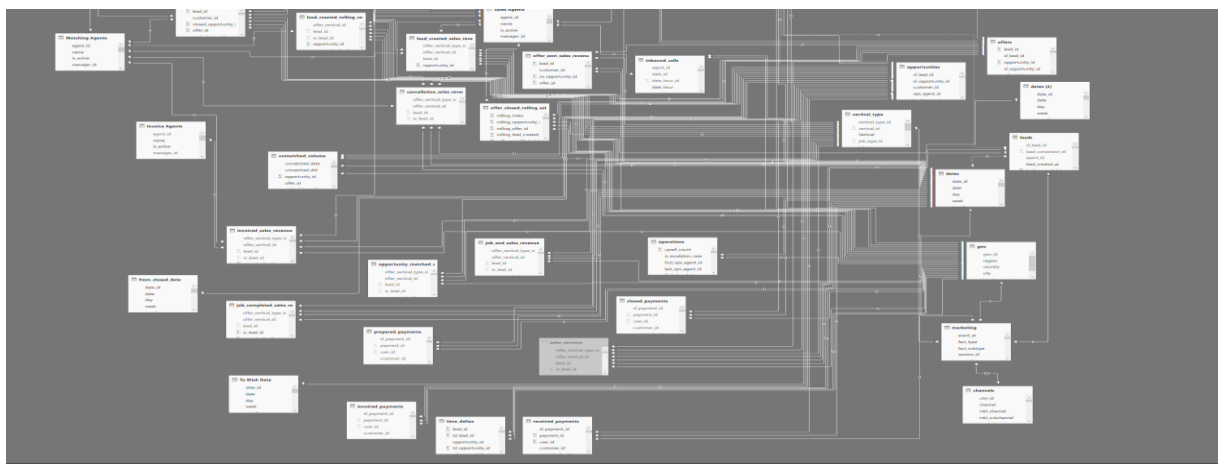


Figure 28: power BI data model

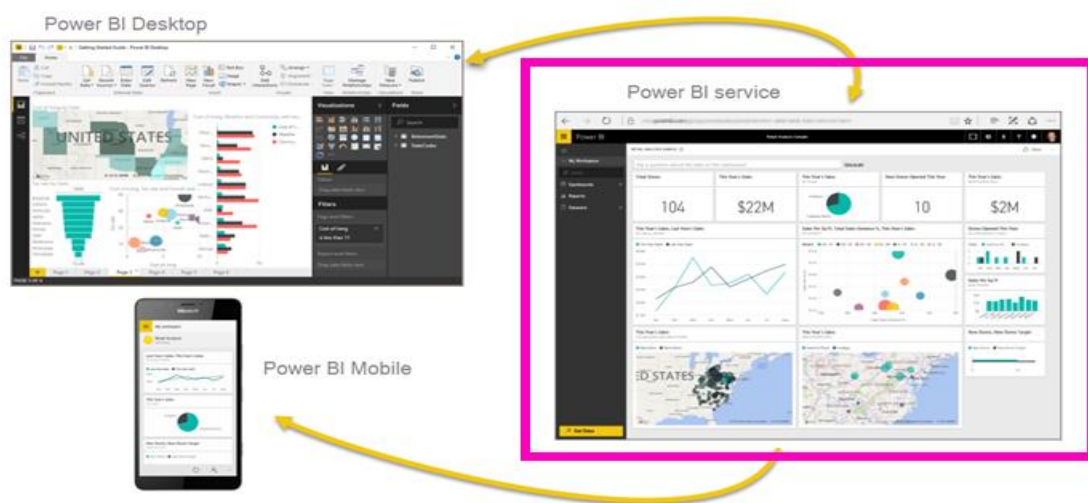


Figure 29: publish data model from desktop to service

This reporting system actually presents two issues:

- ❖ The first issue is that, Homebell is a growing startup, which means business processes often change, and with that change the data warehouse structure also changes, which means you have to propagate these changes to the data model in Power BI desktop and all the reports based on that data model (dataset).
- ❖ The second issue is that for some analysis, we need an advanced changing of data that the Power BI features does not give and we need to actually build our dataset using SQL, not using the data model itself.

2.2 New reporting system (streaming dataset)

The solution to the problems mentioned above is to use the streaming datasets and Power BI REST APIs in Power BI service. The streaming datasets in Power BI will replace the dataset created from the data model in Power BI desktop.

In practice, streaming datasets and their accompanying streaming visuals are best used in situations when it is critical to minimize the latency between when data is pushed and when it is visualized. In addition, it is best practice to have the data pushed in a format that can be visualized as-is, without any additional aggregations.

In order to push data to this streaming dataset, we will use the Power BI REST APIs which can be used to create and send data to a streaming dataset.

Using any programming language that supports REST calls, we can create apps, our Node.js application in our case that integrates with a Power BI in real time. [14]

The steps to push data into a streaming dataset using the Power BI REST APIs are the following:

- Create a streaming dataset using Power BI interface. We should also make sure to enable Historic data if we want Power BI to store the data that's sent through this data stream, because otherwise, Power BI will only store the data in the cache memory for only one hour.

Figure 30: Streaming dataset creation

- Once we successfully create our streaming dataset, we are provided with a REST API URL endpoint, which our Node JS application can call using POST requests to push the data to Power BI streaming data dataset we created.

```
function addRows(params) {
  let { group, dataset, table, rows } = params;
  group = group || HOMEBELL_GROUP;
  table = table || 'RealTimeData' // this is the default by POWER BI
  return authReq({
    path: '/groups/${group}/datasets/${dataset}/tables/${table}/rows',
    TYPE: 'post',
    body: {
      rows
    }
  })
}
```

Figure 31: Power BI script

2.3 Reports & dashboards

In this part, we are going to present the different reports and dashboards created using the Power BI streaming datasets and Power BI REST APIs. We will start with the marketing related reports, and then we are going to present other reports created for other departments that wanted to have real time reports as well.

Since Homebell is just introducing their own marketing campaigns to get leads internally, One of the biggest questions that's need to be answered for marketing is how well the marketing metrics are doing compared to the previous month and compared the non-marketing leads, leads that are coming from lead providers.

For this, we created a cumulative report that measures the number of leads, valid leads, sessions and marketing cost for a given month comparing to the previous month in a cumulative way.



Figure 32: Month over month marketing metrics

We can see in this report above the evolution of number of sessions, leads and marketing cost for the month of June comparing to July in the top part of the report. In the bottom part we see the same comparison for the same metrics but in a non-cumulative way.

This are the overall numbers without selecting any marketing channel. And of course we have a filter so we can select a specific marketing channel (Bing, Google, Facebook, lead providers) and see the performance of this channel.



Figure 33: Number of leads per marketing sub-channel per day

As we can see in the report above, we can follow the number of leads per marketing sub-channel for the current month. We took the case of Yahoo in this chart.



Figure 34: Distribution of number of leads per marketing sub-channel per day

We can also select all marketing sub-channel in the same report and this way we can see the distribution of number of leads between all the lead sources per day. We can see that the lead coming from lead providers is still the number one source of leads for Homebell, followed by Google paid search.



Figure 35: Distribution of number of leads per marketing sub-channel per day

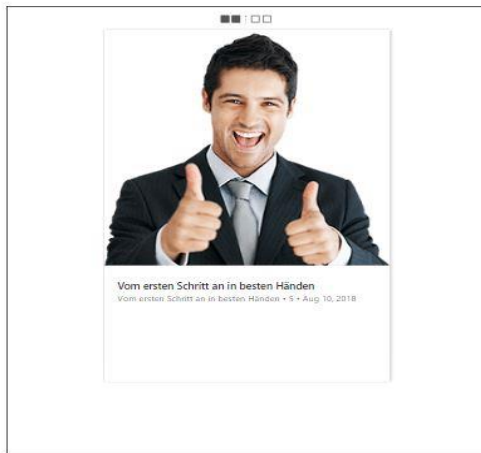
The report above represents two metrics: the lead to qualified (L2Q) which is the rate of potential leads becoming qualified and the offer won conversion rate (WCR) which is the rate qualified leads leading to a won offer per leads source. We took the example of Skydreams here which is our biggest lead provider.



Figure 36: Sales dashboard

The dashboard above is for the sales department. It represents different metrics per different axe of analysis:

- The lead count per job verticals
- The amount of the NMV and the number of Won opportunities by month and region.
- The booked margin, the booked NMV and the margin percentage per month.
- The invoiced NMV by month and by paid status (cancelled, not paid, paid, paid partially)



*Ich war vom ersten Schritt an in guten
Händen.
Alles wurde ausführlich per Telefon erklärt.
Ich war zu keinem Zeitpunkt auf mich
alleine gestellt.
Die Handwerker haben ihre Arbeit
pünktlich und sehr gut erledigt.
Ich hatte überhaupt keine Arbeit :)*
Au

Figure 37: Customers reviews report

The above report is a report that shows the latest customers reviews in the screen in front of the sales agent. This is supposed to boost their productivity and get an idea about the customer's satisfaction.

3. Google Sheets reports

When it comes to reports, Homebell rely heavily on Google Sheets on a daily basis in all the departments. This because Google Sheet can be very handy when it comes to seeing the raw data and the exact value of the different KPIs. It is also because the fact that Google Sheet is hosted in the cloud so it is easily accessible by anyone, and can be designed to solve almost any problem. From real-time editing over the internet to scheduling to data visualization, and, most recently, offline access to files, spreadsheets is essential to Homebell.

3.1 Existing reporting system

The Existing reporting system in Google Sheets is also one of the reason why Homebell decision makers rely on it as a reporting tool on a daily basis, and this is because anyone can update the Google Sheets report at any moment without any

coding thanks to our BI-Grabber add-on. The BI- Grabber is a Google app script which is a scripting language for application development in Google Suite Platform based on JavaScript. It works as the following:

- When creating a report in Google Sheet, we start by creating three different tabs (sheets) in the report, one for the final report itself used by decision makers, one for the raw data upon which is built the report, and one named “queries”.
- Creating the query that will return the data we need to build the report and adding it to the “queries” tab in the second column (B) with a specific title in the first column (A). The title of the query should be the same name as the title of the raw data sheet. This is because each query in the “queries” tab should return data to its specific raw data sheet.
- After that, going to add-on and running the BI-Grabber.
- The script of the BI-Grabber will read the content of the “queries” sheet and send the queries to an API URL created in our application and published in the Homebell server.
- The API methods of our Node JS application will send the result set of the query in a JSON format to the BI-Grabber script, which then will update the data in the correspond raw data sheet with the new result set sent by the Homebell server.
- Now we can create a report in the “report” tab based on the raw data tab and aggregate data using Google Sheets formula.
- To update this report, the report user will just run the BI-Grabber which will update the raw data which then will update the report based upon on.

This system allows Google Sheets users to update any report they are using and get updated data from the data warehouse.

3.2 Reports

vertical	offer	closed	booked	potential nm	nmv	booked nmv	offer to close % (#)	close to booked % (#)	offer to close % (NMV)	close to booked % (NMV)
constructional_plastering	193	15	12	823974	38144	32874	7.8%	80.0%	6.1%	86.2%
custom	356	42	34	1393768	284241	256727	11.8%	81.0%	20.4%	90.0%
floor_tiling	96	9	8	455278	17408	16027	9.4%	88.9%	3.8%	92.1%
lacquering	727	66	58	1416932	114471	103511	9.1%	87.9%	8.1%	90.4%
lamine	458	44	38	704942	85449	58943	9.8%	88.4%	9.3%	90.1%
paint_outside_items	114	12	10	243901	21312	18272	10.5%	83.3%	8.7%	88.7%
painting	2377	204	179	4808451	400455	347871	8.8%	87.7%	8.7%	88.8%
parquet	734	44	37	1849996	114930	92234	6.0%	84.1%	6.2%	80.3%
plastering	65	4	1	233171	1498	-4117	6.2%	25.0%	0.6%	-274.8%
pvc	159	6	5	292331	7990	7224	3.8%	83.3%	2.7%	90.4%
scaffolding	56	6	5	183876	12759	11455	10.7%	83.3%	6.9%	88.8%
wall_tiling	41	2	1	140384	5442	2426	4.9%	50.0%	3.0%	44.6%
wallpapering	582	52	41	1287876	116474	97887	8.9%	73.8%	9.0%	83.9%
carpet	129	14	13	258550	35852	35344	10.9%	82.9%	13.9%	99.1%
dry_walling	68	7	6	203300	22250	13180	10.3%	85.7%	10.9%	59.2%
exterior_painting	83	6	6	358943	20350	16281	7.2%	100.0%	5.7%	80.0%
exterior_plastering	34	6	3	248948	26995	11999	17.6%	50.0%	12.1%	40.0%
vinyl	182	17	16	473432	32951	24716	9.3%	84.1%	7.0%	75.0%

Figure 38: Conversion rates per verticals report

This Google Sheet report above is an aggregation of the different metrics (number of offers, leads, booked offers, potential NMV, and actual NMV) as well as the conversion rates (lead to qualified, offer to close, offer to booked) per vertical. The user can also select the country (Germany or Netherlands) and the source of leads (Internal or external).

By #	Maurice Ebbinge	Simon Reslan	Dif
# Total Leads	1,354	1,225	11%
# Total Leads (Last Agent)	525	552	-10%
% Leads to Valid	83.38%	79.59%	5%
# Valid Leads	1,129	975	16%
# Valid Leads (Last Agent)	525	552	-10%
Reached Leads	339	426	-20%
# Opportunities	295	445	-34%
% Lead to Offer	17.73%	29.31%	-40%
% Valid Lead to Offer	21.26%	36.82%	-42%
# Offers	240	359	-33%
% Offer to Close	6.67%	13.09%	-49%
% Valid Lead to Close	1.42%	4.82%	-71%
	3.05%	7.21%	-58%
# Close - Pre Cancel	18	47	-66%
# Cancels	1	14	-93%
# Close - Post Cancel	15	33	-55%
Avg time lead to qualified (h)	30	19	61%
Avg time qualified to offer (h)	58	119	-53%
Avg time offer to close (h)	265	265	0%
Avg call attempts	1.00	1.00	0%

Figure 39: Sales overview report

The report above is highly used by the company CEO to monitor the productivity of the sales agents in terms of leads, offers created, cancelled opportunity and closed offers etc. It also offers the ability to compare two sales agents in specific timeframe that he select.

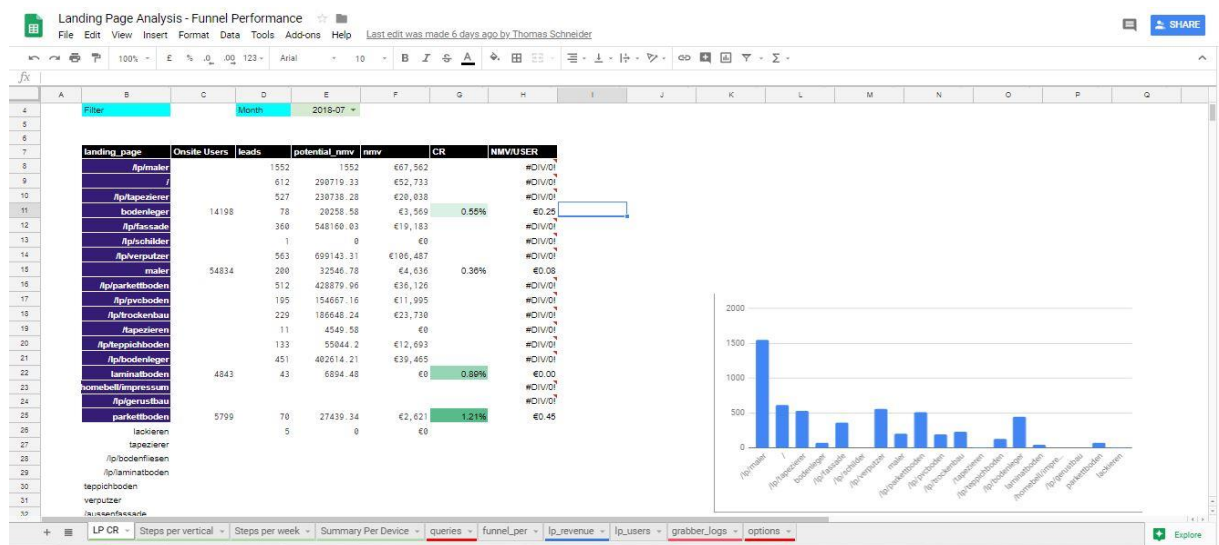


Figure 40: Landing page analysis report

The report above is both for the marketing and the product teams to analyse the activity per landing pages. This enable the marketing team to track which landing pages are doing well and which not, which can give an idea about how well is doing the ad that direct to this landing page. This report is very useful to the product team as well to track the different users per landing page to improve the user experience in the Homebell website based on this results.

Top View Marketing Report

File Edit View Insert Format Data Tools Add-ons Help Scripts Last edit was made 3 days ago by Business Intelligence System Account

100% 123 Arial 10

Filter

By Date: 2018-12-01 to 2018-12-31

By Region: both

By Vertical: advertiser

Reporting by Channel

Channel*	Sessions	Leads	Valid Leads	qualified leads	Offers sent	Offers above 500	Offers above 1000	Offers above 1500	Closed	Booked	Completed	Costs	Won NMV	Booked NMV	Completed NMV	Booked Margin	Completed Margin	Lead
brand	2,975	0	0	0	0	0	0	0	0	0	0	€0	€0	€0	€0	€0	€0	#C
social paid	13	0	0	0	0	0	0	0	0	0	0	€34	€0	€0	€0	€0	€0	#C
paid search	24,136	643	555	180	115	111	87	65	8	5	0	€0	€7,327	€3,490	€026	€0	€042	#C
referrer	5,777	0	0	0	0	0	0	0	0	0	0	€0	€0	€0	€0	€0	€0	#C
crm	4,071	0	0	0	0	0	0	0	0	0	0	€0	€0	€0	€0	€0	€0	#C
organic search	3,302	0	0	0	0	0	0	0	0	0	0	€0	€0	€0	€0	€0	€0	#C
display remarketing	1,903	2	2	1	1	1	1	1	0	0	0	€0	€0	€0	€0	€0	€0	#C
lead provider	0	771	672	266	177	175	127	98	13	5	0	€4,080	€22,579	€5,893	€2,534	€0	€771	#C
unknown	28	1,198	935	330	208	199	137	94	15	6	0	€1,399	€23,716	€4,592	€1,379	€0	€1,195	#C
social organic	39	0	0	0	0	0	0	0	0	0	0	€0	€0	€0	€0	€0	€0	#C
native	13,812	199	173	38	13	13	10	8	2	1	0	€0	€3,470	€1,359	€159	€0	€199	#C
display acquisition	381	5	5	0	0	0	0	0	0	0	0	€0	€0	€0	€0	€0	€0	#C
inbound call	0	13	12	12	4	4	3	2	4	4	0	€0	€39,479	€39,479	€11,537	€0	€13	#C
Total	56,497	2,809	2,364	827	516	503	365	266	42	21	0	€5,513	€96,571	€54,783	€16,535	€0	€2,807	#C
Total (- lead providers)	56,497	2,038	1,692	561	339	328	238	168	29	16	0	1,433	73,692	48,890	14,001	€0	€2,807	#C
CR%					5.0%													
% of Internal Lead					72.8%													
opportunity to see repeat rate					0.3656180745													

By Channel By Subchannel By Vertical queries raw_data sources UTM attribution to MKT channel invoices KPIS grabber_logs options

Figure 41: Top view marketing report

The report above is the Top view marketing reports that is used on a daily basis by the marketing team. This report aggregates the top marketing metrics per vertical, marketing channels and marketing sub-channels.

LEAD AGE @ FIRST CALL / REACH

File Edit View Insert Format Data Tools Add-ons Help Last edit was on September 17

100% 123 Arial 10

REGION DE

CREATED 3,000 [called & reached sections in development]

LEADS CREATED

DAY	FROM	TO
01/07/2018	09:00:00	18:30:00

Fields colored like this are editable

Called

		.25 quantile age @ first call (hours)	.75 quantile age @ first call (hours)
called leads	3,000	0.1	0.7
average of lowest	90%	0.1	0.7
average age @ first call (hours)	5.4	0.1	14.5

Reached

		.25 quantile age @ first reach (hours)	.75 quantile age @ first reach (hours)
reached leads	3,544	0.1	1.1
average of lowest	90%	0.1	1.1
average age @ first reach (hours)	7.8	0.1	15.6

report queries test creation_contact_reach grabber_logs options

Figure 42: Lead age in the funnel report

This report above concerns the lead age in our funnel. The lead age is how much time does the leads stays in our calling funnel after it is created before a sales agent calls him. This report is used to track how well the sales agents are doing in terms of calling the leads created.

4. Conclusion

Power BI and Google Sheets are both complementary reporting tools in Homebell. Power BI is good for seeing the big picture and monitoring the KPIs of the company and monitoring how different areas of a business are performing while Google Sheets is there to explore data and compare exact numbers and aggregating the raw data which can be used to better understand and improve business performance. In the next chapter we will use data analytics to go further into details and take a deep dive into some of the business questions that we cannot analyze using Power BI or Google Sheets.

Chapter VI : Data Analysis

1. Introduction

In this chapter, we are going to take data analysis in Homebell to the next step, a new step. Since Homebell is a newly created start-up, there was not any attempt before to do some deep dive analysis. For that we decided to work the business development team, a department that has a lot of business questions that could be answered with data analysis.

One of the biggest current question for the business development team is to lower the cancellation rate for opportunities.

2. Cancellation reasons analysis :

Each opportunity can be cancelled whether before it is booked (affected to a partner) or after it is booked. Many reason could cause the cancellation of an opportunity: the lead cancels after Homebell spending too much time finding a partner, the partner cancels after discovering that the amount of work or the working condition are different from what was agreed on..

Understanding the different reason for the cancellation rate, as well as the different features and variable that affect this rate could be very beneficial to the business development team in order to lower this rate and to go deep into the drivers of this metric.

2.1 Preface:

The report is roughly divided into two parts: The first part provides information about cancellations on the level of partners, agents, verticals & vertical combinations, cities and months. It helps identifying abnormal cancellation behavior of individual subjects. The second part approximates the importance of the different cancellation drivers through a logistic regression and a decision tree model.

The report can be rerun at any time to reproduce this html output with an updated dataset. The underlying dataset of this analysis contains all valid opportunities that were closed after 01.01.2017 and have been either cancelled or completed. At the creation of this document the dataset consists of 5900 opportunities.

2.2 Descriptive analysis:

In this part the dataset is being filtered and aggregated in different ways according to the subject. The resulting browsable table serves for identifying subjects with high cancellation rates or other remarkable behavior.

The information shown for each subject include:

- `n_jobs` = number of opportunities
- `c_rate_prebook_num` = prebook cancellation rate by number of opportunities
- `c_rate_prebook_nmv` = prebook cancellation rate by nmv
- `c_rate_postbook_num` = prebook cancellation rate by number of opportunities
- `c_rate_postbook_nmv` = postbook cancellation rate by nmv
- `c_rate_total_num` = overall cancellation rate by number of opportunities
- `c_rate_total_nmv` = overall cancellation rate by nmv
- `cr_no_partner_found_num` = percent of cancelled opps where cancellation reason is “no partner found” by number of opportunities
- `cr_partner_problem_num` = percent of cancelled opps where cancellation reason has been a problem with the partner by number of opportunities
- `med_invoiced_margin_nmv` = median margin by booked nmv
- `med_invoiced_margin` = median margin by invoice net amount
- `med_invoiced_nmv` = median opportunity size by nmv
- `med_invoiced_nmv` = median opportunity size by net invoiced
- `med_close_to_cancel` = median days from last close to cancellation
- `d14_cancels` = percent of cancelled opps with cancellation after < 14 days from last close.
- `med_book_to_cancel` = median days from booking to cancellation

The table is then followed by summary statistics, stating the quintiles, median, min/max and mean for comparability of the individual values in the table.

2.2.1: Cancellations by partners

For this analysis, we selected all partners (except black sheep, dropout and disabled) with at least 5 opportunities and one cancellation.

This descriptive analysis can be used to identify partners with high cancellation rates, as well as partners who might be stealing jobs (short timeframe between booking date and cancellation date as reflected by med_book_to_cancel in combination with a high cancellation rate).

Show entries

Search:

	partner_link	partner_status	n_jobs	c_rate_total_num	c_rate_total_nmv	cr_no_partner_found_num	cr_partner_problem_num	med_invoiced_margin_nmv
1	https://de.homebell.com/partner/admin/users/100	plus_partner	10	0.5	0.38	0.2	0	0.25
2	https://de.homebell.com/partner/admin/users/1005	plus_partner	27	0.19	0.09	0	0	0.2
3	https://de.homebell.com/partner/admin/users/1008	plus_partner	10	0.1	0.04	0	1	0.32
4	https://de.homebell.com/partner/admin/users/111	partner	23	0.35	0.39	0.12	0.25	0.3
5	https://de.homebell.com/partner/admin/users/1121	partner	7	0.29	0.17	0	0	0.2
6	https://de.homebell.com/partner/admin/users/114	plus_partner	14	0.21	0.1	0	0	0.29
7	https://de.homebell.com/partner/admin/users/1156	plus_partner	7	0.29	0.23	0	0	0.15
8	https://de.homebell.com/partner/admin/users/1160	partner	9	0.22	0.28	0.5	0	0.2
9	https://de.homebell.com/partner/admin/users/1172	plus_partner	8	0.25	0.4	0	0	0.12
10	https://de.homebell.com/partner/admin/users/1203	partner	6	0.17	0.14	0	0	0.13

Showing 1 to 10 of 91 entries

Previous 2 3 4 5 ... 10 Next

Figure 43: Cancellation by partners results

Summary Statistics:

```
##      n_jobs      c_rate_total_num c_rate_total_nmv
## Min.   : 6.00   Min.   :0.0500   Min.   :0.0100
## 1st Qu.: 8.00   1st Qu.:0.1400   1st Qu.:0.1150
## Median :14.00   Median :0.2100   Median :0.2400
## Mean   :21.14   Mean   :0.2438   Mean   :0.2715
## 3rd Qu.:22.00   3rd Qu.:0.3150   3rd Qu.:0.3700
## Max.   :115.00   Max.   :0.7000   Max.   :0.8300
##
## cr_no_partner_found_num cr_partner_problem_num med_invoiced_margin_nmv
## Min.   :0.00000   Min.   :0.000   Min.   :0.0500
## 1st Qu.:0.00000   1st Qu.:0.000   1st Qu.:0.2000
## Median :0.00000   Median :0.000   Median :0.2350
## Mean   :0.06253   Mean   :0.156   Mean   :0.2303
## 3rd Qu.:0.00000   3rd Qu.:0.235   3rd Qu.:0.2800
## Max.   :1.00000   Max.   :1.000   Max.   :0.3700
## NA's   :1
## med_invoiced_margin med_invoiced_nmv med_invoiced_net_amount
## Min.   :0.0500   Min.   : 425.5   Min.   : 421.6
## 1st Qu.:0.2000   1st Qu.: 893.7   1st Qu.: 893.7
## Median :0.2350   Median :1109.9   Median :1082.3
## Mean   :0.2303   Mean   :1199.7   Mean   :1177.9
## 3rd Qu.:0.2800   3rd Qu.:1433.8   3rd Qu.:1433.8
## Max.   :0.3700   Max.   :2947.4   Max.   :2926.4
## NA's   :1
## med_close_to_cancel d14_cancels med_book_to_cancel
## Min.   : 3.00   Min.   :0.0000   Min.   : 1.00
## 1st Qu.:13.50   1st Qu.:0.0000   1st Qu.: 9.00
## Median :26.00   Median :0.3300   Median :22.50
## Mean   :43.61   Mean   :0.3424   Mean   :38.19
## 3rd Qu.:58.75   3rd Qu.:0.5000   3rd Qu.:46.50
## Max.   :323.00   Max.   :1.0000   Max.   :319.00
##
```

Figure 44: Cancellation by partners results summary

2.2.2: Cancellations by agents

For this analysis, we selected all active agents with at least 10 opportunities and 1 cancellation.

This descriptive analysis can be used to identify agents with high cancellation rates, as well as agents who might be pitching with the cancellation possibility (high percentage of cancellations within 14 days as reflected by d14_cancels combined with a high overall cancellation rate).

Show 10 entries

Search:

	first_sales_agent	n_jobs	c_rate_prebook_num	c_rate_prebook_nmv	c_rate_postbook_num	c_rate_postbook_nmv	c_rate_total_num	c_rate_total_nmv	cr_no_partner_found_num
1	Benjamin Relling	60	0.32	0.33	0.22	0.27	0.53	0.61	0.34
2	Christian Beier	106	0.35	0.28	0.23	0.35	0.58	0.64	0.25
3	Christian Berger	29	0.38	0.45	0.24	0.2	0.62	0.64	0.44
4	Claus Philip Lehmann	23	0.35	0.47	0.09	0.18	0.43	0.65	0.4
5	David de Gier	43	0.28	0.17	0.12	0.15	0.4	0.32	0.24
6	Eline Oostra	34	0.38	0.43	0.21	0.19	0.59	0.63	0.15
7	Emil Devrim	11	0.27	0.07	0.36	0.56	0.64	0.63	0.29
8	Frank Lichtenberg	141	0.23	0.15	0.21	0.25	0.43	0.4	0.18
9	Hein Wouters	124	0.17	0.1	0.15	0.12	0.31	0.22	0.23
10	Johannes Schnabel	368	0.27	0.24	0.18	0.24	0.45	0.47	0.28

Showing 1 to 10 of 25 entries

Previous 1 2 3 Next

Figure 45: Cancellation by agents results

Summary Statistics:

```
##      n_jobs      c_rate_prebook_num      c_rate_prebook_nmvm      c_rate_postbook_num
## Min. : 11.0      Min. :0.0600      Min. :0.0200      Min. :0.0900
## 1st Qu.: 34.0      1st Qu.:0.1800      1st Qu.:0.1500      1st Qu.:0.1600
## Median : 69.0      Median :0.2500      Median :0.1900      Median :0.2000
## Mean : 99.2      Mean :0.2452      Mean :0.2208      Mean :0.1964
## 3rd Qu.:141.0      3rd Qu.:0.3200      3rd Qu.:0.2700      3rd Qu.:0.2200
## Max. :368.0      Max. :0.4700      Max. :0.4900      Max. :0.3600
##      c_rate_postbook_nmvm      c_rate_total_num      c_rate_total_nmvm
## Min. :0.0700      Min. :0.2000      Min. :0.1500
## 1st Qu.:0.1500      1st Qu.:0.3200      1st Qu.:0.3200
## Median :0.2100      Median :0.4300      Median :0.4500
## Mean :0.2352      Mean :0.4404      Mean :0.4572
## 3rd Qu.:0.2900      3rd Qu.:0.5300      3rd Qu.:0.6300
## Max. :0.5600      Max. :0.8100      Max. :0.8100
##      cr_no_partner_found_num      cr_partner_problem_num      med_invoiced_margin_nmvm
## Min. :0.0000      Min. :0.0000      Min. :0.1600
## 1st Qu.:0.0900      1st Qu.:0.0100      1st Qu.:0.2100
## Median :0.2400      Median :0.0600      Median :0.2500
## Mean :0.2252      Mean :0.0872      Mean :0.2424
## 3rd Qu.:0.3000      3rd Qu.:0.1300      3rd Qu.:0.2800
## Max. :0.4800      Max. :0.2700      Max. :0.3000
##      med_invoiced_margin      med_invoiced_nmvm      med_invoiced_net_amount
## Min. :0.1600      Min. : 852.2      Min. : 852.2
## 1st Qu.:0.2100      1st Qu.: 997.0      1st Qu.: 945.7
## Median :0.2500      Median :1069.6      Median :1059.6
## Mean :0.2424      Mean :1314.6      Mean :1296.3
## 3rd Qu.:0.2800      3rd Qu.:1630.5      3rd Qu.:1620.4
## Max. :0.3000      Max. :2219.2      Max. :2219.2
##      med_close_to_cancel      d14_cancels
## Min. : 5.00      Min. :0.1200
## 1st Qu.:10.00      1st Qu.:0.4400
## Median :14.50      Median :0.5100
## Mean :14.66      Mean :0.5100
## 3rd Qu.:17.50      3rd Qu.:0.6000
## Max. :31.00      Max. :0.7600
```

Figure 46: Cancellation by agents results summary

2.2.3: Cancellations by city

For this analysis, we selected cities with at least 10 jobs and wish-start time during the past 365 days.

This descriptive analysis can be used to identify an insufficient partner network in specific cities (high pre-book cancellation rate & high rate of “no partner found” cancellation reason as reflected by cr_no_partner_found_num).

Show 10 entries

Search:

	city_bucket	n_jobs	c_rate_prebook_num	c_rate_prebook_nmvm	c_rate_postbook_num	c_rate_postbook_nmvm	c_rate_total_num	c_rate_total_nmvm	cr_no_partner_found_num	c
1	Aachen	11	0.09	0.25	0.09	0.07	0.18	0.32	0	
2	Amsterdam Binnenstad en Oostelijk Havengebied	14	0.14	0.08	0.14	0.31	0.29	0.39	0.25	
3	Augsburg	20	0.05	0.02	0.5	0.78	0.55	0.8	0	
4	Berlin	455	0.18	0.18	0.16	0.15	0.35	0.33	0.18	
5	Bielefeld	38	0.34	0.09	0.08	0.04	0.42	0.13	0.38	
6	Bochum	19	0.05	0.03	0.21	0.15	0.26	0.17	0	
7	Bonn	23	0.26	0.13	0.26	0.35	0.52	0.48	0.08	
8	Braunschweig	21	0.43	0.43	0.1	0.11	0.52	0.54	0.64	
9	Bremen	50	0.2	0.08	0.16	0.21	0.36	0.29	0.33	
10	Chemnitz	20	0.25	0.32	0.2	0.24	0.45	0.57	0.33	

Figure 47: Cancellation by city results

Summary Statistics:

```
##      n_jobs      c_rate_prebook_num c_rate_prebook_nmv
##  Min.   : 11.00      Min.   :0.0000      Min.   :0.0000
##  1st Qu.: 15.50      1st Qu.:0.1550      1st Qu.:0.0950
##  Median : 25.50      Median :0.2050      Median :0.1750
##  Mean   : 46.22      Mean   :0.2282      Mean   :0.1802
##  3rd Qu.: 47.00      3rd Qu.:0.3175      3rd Qu.:0.2375
##  Max.   :455.00      Max.   :0.5000      Max.   :0.6400
##
##      c_rate_postbook_num c_rate_postbook_nmv c_rate_total_num c_rate_total_nmv
##  Min.   :0.0000      Min.   :0.0000      Min.   :0.1700      Min.   :0.0900
##  1st Qu.:0.1425      1st Qu.:0.1300      1st Qu.:0.3600      1st Qu.:0.3300
##  Median :0.1950      Median :0.2250      Median :0.4150      Median :0.3950
##  Mean   :0.1974      Mean   :0.2468      Mean   :0.4258      Mean   :0.4280
##  3rd Qu.:0.2300      3rd Qu.:0.3300      3rd Qu.:0.5100      3rd Qu.:0.5475
##  Max.   :0.5000      Max.   :0.7800      Max.   :0.7900      Max.   :0.8000
##
##      cr_no_partner_found_num cr_partner_problem_num med_invoiced_margin_nmv
##  Min.   :0.0000      Min.   :0.0000      Min.   :0.1000
##  1st Qu.:0.1025      1st Qu.:0.0000      1st Qu.:0.1800
##  Median :0.2300      Median :0.0750      Median :0.2300
##  Mean   :0.2172      Mean   :0.0942      Mean   :0.2218
##  3rd Qu.:0.3300      3rd Qu.:0.1475      3rd Qu.:0.2700
##  Max.   :0.6400      Max.   :0.6700      Max.   :0.3200
##
##      NA's :1
##
##      med_invoiced_margin med_invoiced_nmv med_invoiced_net_amount
##  Min.   :0.1000      Min.   : 596.9      Min.   : 596.9
##  1st Qu.:0.1800      1st Qu.: 967.6      1st Qu.: 930.5
##  Median :0.2300      Median :1137.8      Median :1093.0
##  Mean   :0.2218      Mean   :1214.2      Mean   :1176.9
##  3rd Qu.:0.2700      3rd Qu.:1401.9      3rd Qu.:1324.6
##  Max.   :0.3200      Max.   :2138.1      Max.   :2138.1
##
##      NA's :1
##
##      med_close_to_cancel d14_cancels      med_close_to_book mean_close_to_book
##  Min.   : 2.00      Min.   :0.0000      Min.   : 0.00      Min.   : 1.100
##  1st Qu.:13.12      1st Qu.:0.3925      1st Qu.: 1.00      1st Qu.: 3.561
##  Median :16.00      Median :0.5000      Median : 1.00      Median : 4.679
##  Mean   :20.01      Mean   :0.4546      Mean   : 1.91      Mean   : 5.442
##  3rd Qu.:23.38      3rd Qu.:0.5600      3rd Qu.: 2.75      3rd Qu.: 6.981
##  Max.   :73.00      Max.   :0.7500      Max.   :10.00      Max.   :13.900
##
```

Figure 48: Cancellation by city results summary

2.2.4: Cancellations by city and vertical

For this analysis, we selected cities with at least 10 jobs and wish-start time during the past 365 days) split by vertical.

This descriptive analysis can be used to identify an insufficient partner network in a city for a specific vertical. Can also be used for deciding on discontinuing specific verticals in a city. Each vertical is represented in an own table, containing all cities where it has been performed. In this report we are going to show only the Painting vertical but the analysis was done for all the verticals.

Show 10 entries

Search:

	city_bucket	n_jobs	c_rate_prebook_num	c_rate_prebook_nmv	c_rate_postbook_num	c_rate_postbook_nmv	c_rate_total_num	c_rate_total_nmv	cr_no_partner_found_num
1	Augsburg	13	0.08	0.04	0.46	0.77	0.54	0.8	0
2	Berlin	257	0.14	0.13	0.19	0.17	0.32	0.29	0.11
3	Bielefeld	20	0.25	0.07	0.05	0.03	0.3	0.1	0.17
4	Bonn	11	0	0	0.55	0.79	0.55	0.79	0
5	Braunschweig	15	0.47	0.44	0.13	0.18	0.6	0.62	0.56
6	Bremen	26	0.19	0.06	0.08	0.16	0.27	0.21	0.43
7	Dortmund	30	0.1	0.08	0.23	0.37	0.33	0.45	0.2
8	Dresden	28	0.11	0.21	0.32	0.17	0.43	0.38	0.17
9	Duisburg	14	0	0	0.36	0.32	0.36	0.32	0
10	Düsseldorf	27	0.19	0.09	0.26	0.51	0.44	0.61	0.25

Showing 1 to 10 of 35 entries

Previous 1 2 3 4 Next

Figure 49: Cancellation by city & verticals results

```
##      n_jobs      c_rate_prebook_num c_rate_prebook_nmv c_rate_postbook_num
##  Min. : 11.0      Min. :0.0000      Min. :0.0000      Min. :0.0500
## 1st Qu.: 14.0      1st Qu.:0.1250      1st Qu.:0.0550      1st Qu.:0.1350
## Median : 20.0      Median :0.1900      Median :0.1300      Median :0.2100
## Mean   : 34.0      Mean   :0.1906      Mean   :0.1534      Mean   :0.2223
## 3rd Qu.: 31.5      3rd Qu.:0.2350      3rd Qu.:0.1600      3rd Qu.:0.2650
## Max.   :257.0      Max.   :0.4700      Max.   :0.6600      Max.   :0.5500
## c_rate_postbook_nmv c_rate_total_num c_rate_total_nmv
##  Min. :0.0100      Min. :0.1300      Min. :0.0500
## 1st Qu.:0.0850      1st Qu.:0.3400      1st Qu.:0.2500
## Median :0.2000      Median :0.3900      Median :0.3800
## Mean   :0.2494      Mean   :0.4137      Mean   :0.4011
## 3rd Qu.:0.3350      3rd Qu.:0.5150      3rd Qu.:0.5300
## Max.   :0.7900      Max.   :0.6200      Max.   :0.8000
## cr_no_partner_found_num cr_partner_problem_num med_invoiced_margin_nmv
##  Min. :0.0000      Min. :0.00000      Min. :0.1100
## 1st Qu.:0.1100      1st Qu.:0.00000      1st Qu.:0.2000
## Median :0.1700      Median :0.04000      Median :0.2500
## Mean   :0.1974      Mean   :0.07914      Mean   :0.2386
## 3rd Qu.:0.3200      3rd Qu.:0.14000      3rd Qu.:0.2800
## Max.   :0.5600      Max.   :0.50000      Max.   :0.3000
## med_invoiced_margin med_invoiced_nmv med_invoiced_net_amount
##  Min. :0.1100      Min. : 337.2      Min. : 337.2
## 1st Qu.:0.2000      1st Qu.: 873.4      1st Qu.: 847.9
## Median :0.2500      Median :1019.6      Median :1016.3
## Mean   :0.2386      Mean :1129.4      Mean :1089.0
## 3rd Qu.:0.2800      3rd Qu.:1345.7      3rd Qu.:1303.7
## Max.   :0.3000      Max. :2462.2      Max. :2043.5
## med_close_to_cancel d14_cancels med_close_to_book mean_close_to_book
##  Min. : 2.5      Min. :0.1500      Min. :0.000      Min. : 0.000
## 1st Qu.:11.0      1st Qu.:0.3450      1st Qu.:1.000      1st Qu.: 2.753
## Median :15.0      Median :0.5000      Median :1.000      Median : 3.567
## Mean   :16.6      Mean :0.5123      Mean :1.643      Mean : 4.334
## 3rd Qu.:21.0      3rd Qu.:0.6350      3rd Qu.:2.000      3rd Qu.: 5.114
## Max.   :36.5      Max. :1.0000      Max. :6.000      Max. :13.923
```

Figure 50: Cancellation by city & verticals results summary

2.3 Feature Importance Approximation:

This part tries to analyze the impact that each cancellation driver has on the cancellation probabilities of an opportunity. We analyze pre-book cancellations and post-book cancellations each on their own. As for now we use the following information as predictor variables:

- vorlauf: time from date of (last) close to projects wish start.
- close2book: = time to match. The time that it takes us to find a partner (only postbook cancellation)

- bin_rematched: did we have to change the partner? (only postbook cancellation)
- nmv: as proxy for the size of the project
- margin: as proxy for how low our partner-price is and how high the price that the consumer gets (only postbook cancellation)
- vertical_combo_c_rate: the cancellation rate of the opportunity's underlying vertical-combination
- partner_c_rate: the cancellation rate of the partner (only postbook cancellation)
- agent_c_rate: the cancellation rate of the agent
- is_image_uploaded: did the customer upload any images
- epd: did we grant the customer an early payment discount and how high was it

2.3.1 Pre-book cancellation dataset:

❖ F-Selector Feature importance on pre-book cancellation

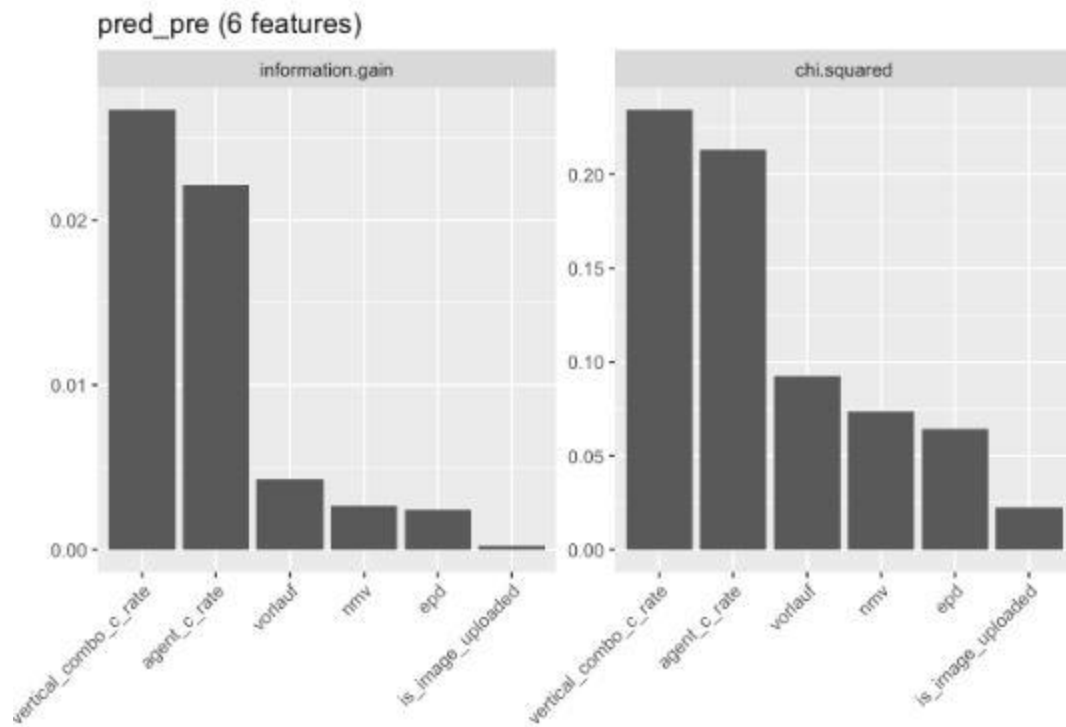


Figure 51: Features importance on the pre-book dataset

❖ Logistic regression on pre-book cancellation:

```
## auc.test.mean acc.test.mean
##      0.7210851      0.7938628
##
## Call: stats::glm(formula = f, family = "binomial", data = getTaskData(.task,
##      .subset), weights = .weights, model = FALSE)
##
## Coefficients:
##      (Intercept)          vorlauf              nmv
##          -1.4517          -0.0873          -0.2439
## vertical_combo_c_rate    agent_c_rate    is_image_uploaded1
##          0.5905          0.4860          -0.1260
##          epd
##          -0.1558
##
## Degrees of Freedom: 5898 Total (i.e. Null); 5892 Residual
## Null Deviance:      6117
## Residual Deviance: 5483  AIC: 5497

## Model for learner.id=classif.logreg; learner.class=classif.logreg
## Trained on: task.id = pred_pre; obs = 5899; features = 6
## Hyperparameters: model=FALSE
```

Figure 52: Logistic regression on the pre-book dataset

❖ Decision Tree's AUC & Accuracy on pre-book dataset

```
## auc.test.mean acc.test.mean
##      0.6898613      0.7760675
```

Figure 53: Decision Tree's accuracy on the pre-book dataset

❖ Interpretation:

We can see that the most important features with the most impact on the pre-book cancellation probabilities are the vertical_combo cancellation rate and the agent cancellation rate. We can also see that the accuracy of the logistic regression model is about 80% comparing to the accuracy of the decision tree model which is 78%.

2.3.1 Post-book cancellation dataset:

❖ F-Selector Feature importance on post-book cancellation

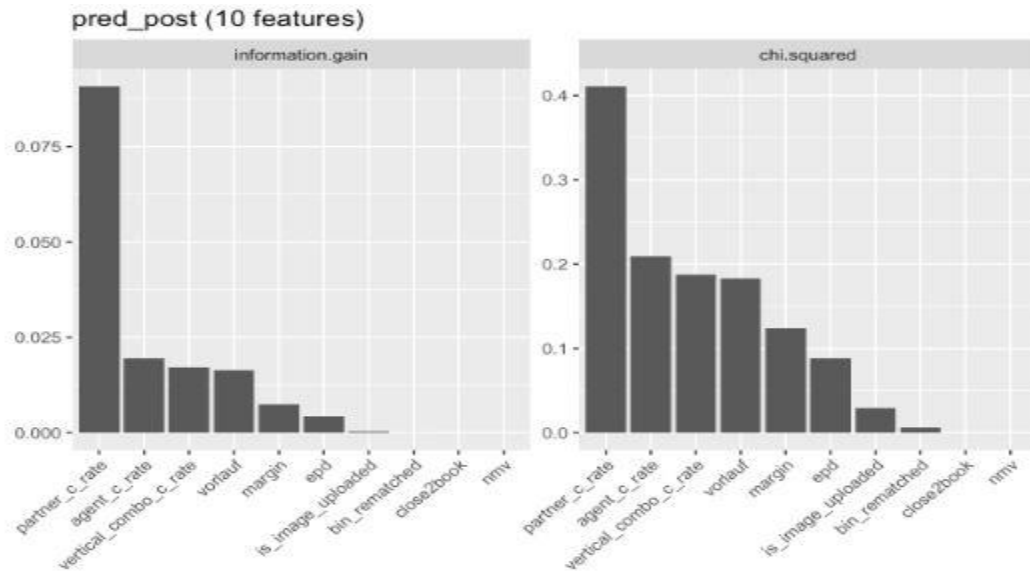


Figure 54: Features importance on the post-book dataset

❖ Logistic regression on post-book cancellation:

```
## auc.test.mean acc.test.mean
##      0.8260642      0.8323276
##
## Call: stats::glm(formula = f, family = "binomial", data = getTaskData(.task,
##      .subset), weights = .weights, model = FALSE)
##
## Coefficients:
##      (Intercept)          vorlauf          close2book
##          -1.68510           0.45218           0.06807
##      bin_rematched1          nmw          margin
##          -0.17761           0.15332           0.32051
## vertical_combo_c_rate    partner_c_rate    agent_c_rate
##          0.44492           1.13309           0.51409
## is_image_uploaded1          epd
##          0.03093          -0.20004
##
## Degrees of Freedom: 4639 Total (i.e. Null); 4629 Residual
## Null Deviance:      4875
## Residual Deviance: 3595  AIC: 3617

## Model for learner.id=classif.logreg; learner.class=classif.logreg
## Trained on: task.id = pred_post; obs = 4640; features = 10
## Hyperparameters: model=FALSE
```

Figure 55: Logistic regression on the post-book dataset

❖ Decision Tree's AUC & Accuracy on post-book dataset

```
## auc.test.mean acc.test.mean  
##      0.7646007      0.8099132
```

Figure 56: Decision tree accuracy on the post-book dataset

❖ Interpretation:

We can see that the most important features with the most impact on the post-book cancellation probabilities are the partner cancellation rate and the agent cancellation rate, which makes sense since the opportunity now is affected to a partner (booked) therefore there is higher chances of being cancelled by the partner selected. We can also see that the accuracy of the logistic regression model is about 83% comparing to the accuracy of the decision tree model which is about 80%. The accuracy of the logistic regression model is higher on post-book dataset than on the pre-book dataset since it has a lower AIC value. This means that we have more chance on predicting cancellation probability of an opportunity after it is booked (affected to a partner).

3. Conclusion

We tried in this chapter to build some statistic models using logistic regression and decision tree to predict the cancellation rate of an opportunity. It is necessary to say that this work is not complete as there are other prediction models that could be used which may have better accuracy. This work should be the basis to develop other models and to answer other questions a part of the cancellation rate.

General Conclusion

Data is at the heart of strategic decision making in business whether you run a small growing start-up or a huge multi-national corporation. When you ensure high data quality, you can provide insight that helps you to answer key business questions. Data leads to insight; business owners and managers can turn that insight into decisions and actions that improve the business.

Working with huge quantities of data and messy data is always hard, but knowing how to deal with it, and automating all the regular processes will provide more time to get your data fresh and ready for analytics.

It is in this exciting and intellectually challenging and motivational setting that I completed my internship graduation. The past months have been very instructive for me. Homebell has offered me opportunities to learn and develop myself in many areas. I gained a great deal of experience, especially in the Data Engineering and Data Analysis field as well as the project planning and monitoring areas.

I also developed my skills in the design and programming of significant software solutions. Many of the tasks and activities that I have worked on during my internship are similar to what I'm studying at the moment. I worked in many areas where I performed different tasks. I also learned what working in a team really means. The meaning of the term "deliver" has another dimension in my mind. I acquired the meaning of "rigorous work", and learned things that could not be taught in college. Working at this company allowed me to understand and grow my skill set in a quantum manner.

Webography

- [1] <https://doc.rero.ch/record/6603/files/BusinessIntelligence.pdf>
- [2] <https://de.homebell.com/>
- [3] <https://snowplowanalytics.com/blog/2016/03/16/introduction-to-event-data-modeling/>
- [4] <https://www.confluent.io/blog/building-real-time-streaming-etl-pipeline-20-minutes/>
- [5] <https://www.sisense.com/blog/top-benefits-business-intelligence-marketing/>
- [6] <https://blog.anant.us/gain-upper-hand-etl-node-js/>
- [7] <https://docs.aws.amazon.com/AmazonS3/latest/dev/Welcome.html>
- [8] <https://aws.amazon.com/fr/redshift/>
- [9] <https://www.supinfo.com/articles/single/4108-qu-est-ce-que-power-bi>
- [10] <https://rviews.rstudio.com/2017/06/08/what-is-the-tidyverse/>
- [11] [https://en.wikipedia.org/wiki/Jira_\(software\)](https://en.wikipedia.org/wiki/Jira_(software))
- [12] [https://en.wikipedia.org/wiki/Confluence_\(software\)](https://en.wikipedia.org/wiki/Confluence_(software))
- [13] <https://powerbi.microsoft.com/en-us/what-is-power-bi/>
- [14] <https://docs.microsoft.com/en-us/power-bi/service-real-time-streaming>



ESPRIT SCHOOL OF ENGINEERING

www.esprit.tn - E-mail : contact@esprit.tn

Siège Social : 18 rue de l'Usine - Charguia II - 2035 - Tél. : +216 71 941 541 - Fax. : +216 71 941 889

Annexe : Z.I. Chotrana II - B.P. 160 - 2083 - Pôle Technologique - El Ghazala - Tél. : +216 70 685 685 - Fax. : +216 70 685 454