MCS: Global Master in Big Data & Business Analytics

The Knowledge discovery process

CRISP-DM | Quick guide

Prepared by

Group A

Abdulaziz Alhagbani, Ckalib Nelson, Michael Wagdy, Sarang Zendehrooh, Nabil Massri

Professor

Antonio Pita Lozano

IE School of Human Sciences and Technology

Madrid, Spain

www.ie.edu

June, 2020

Table of Contents

Introduction

We, at Unilever as one of the largest FMCG companies in the world, and with over 400 brands across the globe, invest a big part of our revenue in market research, marketing analytics and social media listening in order to increase our brands engagements and to have certain control of the rumors about our brands and corporate image. In view of this, we must ensure that our processes always follow the highest standards in terms of quality, reliability and robustness.

This document is intended to support our analytical teams to build their data mining projects based on CRISP-DM, which is a model that has become the leading methodology in data mining. This paper is expected to serve as a quick guide for the tasks required in each phase and a refreshment for our teams since it is understood that certain knowledge on this methodology is already accomplished.

Because of its industry and tool independence, it provides guidelines for organized and transparent execution of any project. It has to be noted that the variety of data and related issues as well as business objectives may require various degrees of flexibility in applying the CRISP-DM reference model depending on each particular project. This process is divided into six consecutive phases that can be observed in the below figure. It is important to note that most of the phases have an iterative nature so that to achieve the best results.

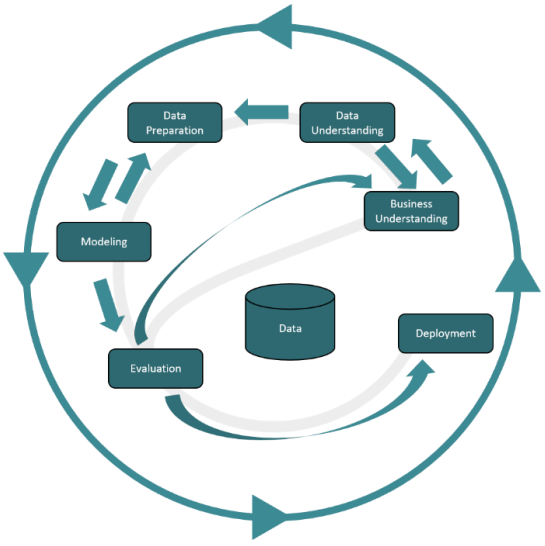

Figure - 1: Chapman's CRISP-DM diagram

This document, apart from describing the methodology for applying this model, will discuss each phase from CRISP-DM alongside a previously completed project, adding richness to the guide.

At Unilever, as a customer-centric organization, we need to deeply understand our customers' needs and fulfill them better than anyone else, and to do so in an efficient manner in this fast changing world we require data. Yet, what separates us from our competitors is our ability to transform our data into insights about consumers' motivations and to turn those

insights into strategy. The example described in this document refers to the successful project of the “insight engine” which accomplishes the aforementioned objectives. It was developed in 2016 and led by Keith Weed (CMO) and Stan Sthanunathan (Head of Insights).

Business Understanding

Determine Business Objectives

Background

This section must describe the business perspective of the data mining process, or the “why?” It should give a brief explanation about the situation of the specific department or branch which this project will mainly impact, explain what we are trying to accomplish and gather all the known information about the organization’s business situation at the given moment.

Example

Most of our stakeholders are currently receiving data from several sources in several formats, which complicates the task of getting insights and business value from the data. We will try to accomplish a more robust and reliable way of digesting our data with the aim of providing a single data source with a consistent format easing the way insights are obtained.

Business Objectives

We shall consider the business success once the useful business outcomes are achieved in line with Unilever’s higher standards and verified by the unit business managers. In order to define this business success we must follow the below criteria:

- The business outcome must be specific
- The business outcome should be measurable
- In case of deciding to set a subjective business success criteria, the following approvals must be obtained:
    - Line Manager
    - Business unit manager
    - Head of department

Example

To build a complete All-in-one analytics platform with full automation to support the different business stakeholders with all types of data work in a self service manner. This platform will Extract, Store, Analyze, Model, Visualize and publish all data types from different data sources (CRM, Social Media, Partners Data like Amazon sales logs, SKUs data files, and any other data files.).

Business Success Criteria

In this section we must describe the specific criteria which will allow us to determine the success rate of our project in an objective manner. The standard mentioned in this section should be ideally aligned with our business objectives. The more specific and measurable criteria we define, the better.

Example

The platform the we aim to build should have the following capabilities:

1. Connectors with all the available and future data types and engines
2. Social Media Listening
3. Data Transformation, Aggregation, manipulation and lookups
4. Data Visualization with a rich designer engine including coloring, charting, and pivoting the data
5. Machine Learning and AI on Business Question-based
6. Publisher to share the data, analysis, reports and dashboard with internal or external users
7. Store the data, reports and analysis models for any revisiting. improvement or enhancements or developing similar in the future
8. Accessible from everywhere
9. Data Replication to another site / data center
10. Data, Application and Network Security
11. Standard environments availability Development, Testing, UAT and Production

Assess Situation

Inventory of Resources

In this section a list of all the resources that will be required to carry out this project must be made, including but not limited to personnel, hardware and software. Then, an assessment of the currently available assets should be carried out in order to identify the specific needs.

Example

We have listed below the necessary skills, software and hardware to successfully perform this project mentioning whether it is available or not or whether it will be outsourced. Since at Unilever we used to work with business and technical consulting firms, for market research and monitoring our ads and campaigns, our team does not seem to ideally fit to implement such a project with this level of maturity, completeness and accuracy, however keeping track of all the resources required and deploying them at the right time will help achieving our business goals.

Personnel		Software	
Skill	Availability	Item	Availability
Project Management	Available	Architecture Framework	Not Available
Data & Information Architect	Not Available	Big Data Platform	Not Available
Application Integration	Not Available	Enterprise Service Bus	Not Available
Data Scientist	Not Available	Machine Learning Engine	Not Available
Business Analyst	Partially Available	Data Visualization	Not Available
Data Management	Not Available	Data Integration, Quality, Profiling, ..etc.	Not Available
Domain Expert	Available	Different Apps for Social Media Listeners, Influencer Analysis, Brand Competition	Not Available

The above table is an initial expectation for the required expertises and softwares.

Requirements, Assumptions and Constraints

In this section we must list all of our requirements, including but not limited to scheduling, quality of results, security and legal constraints (if any). Please note that the legal aspects must always be consulted with our legal department.

Furthermore, all the assumptions made must be documented, those assumptions can be related to the data quality, external factors or even to whom will this data findings be presented. And last but not least, we should spell out all the constraints associated with budgets, schedules, data sources, data accessibility, etc.

Example

The timespan to complete the project is 18 months from the final approval from our CFO. Due to the nature of this project, we require entire technical and business support as well as quick decision makers in order to overcome any hurdles that we might encounter, that's why the key stakeholders are involved.

In relation to the legal aspect, since our project will be developed through a consultancy firm, the selected company will do the due diligence to provide all legal advice and to put the necessary boundaries to implement the project in a safe manner and in line with Unilever's code of ethics and good practice.

Following are the assumptions made on our data:

1. Univariate outliers by standardizing
2. Missing data we will either drop the observation or do a mean substitution
3. When the data is not normally distributed, it will be transformed to logarithmic
4. Linearity
5. Independence

We assume that this data will be presented to the stakeholders at a regional level, CDO and CIO, hence the results presentation shall be made accordingly.

The constraints for this given project are very controlled since we will not be facing issues with the rights to data sources not accessing the data considering that those resources are internal and already run through the specific Unilever's approval flow.

Risks and Contingencies

All the potential risks that might jeopardize the quality, timespan or budget of the project must be listed here along with their contingency plan.

Example

This project is very important for Unilever business and its digital transformation plan, thus a specific contingency committee has been created that will be gathered upon a concise risk arise, nevertheless below are some of the expected risks can have an impact this project progress are the following:

1. Part of the Unilever business units do not have a full view on what they need and what they are looking to achieve. In order to overcome this risk, the teams must be gathered to recapitulate and to come into a conclusion
2. The selected company to build this project might not have a relevant experience in the FMCG field. The importance of this risk will have to be assessed upon the receipt of the proposals and subsequent study of the same
3. Financial risks due to an unexpected economic global crisis. To minimize the damages of this risk, it will be key that the contract made with the consulting firm contain all the provisions for a sudden unexpected stoppage of the project. Legal and financial teams must be fully involved from the beginning to decrease this exposure

Terminology

Gather a glossary of terminology relevant to the project, for both aspects; business and data mining terminology. For the latter, you may refer to Unilever's data glossary which is being updated on a monthly basis by our global IT and Data teams.

Example

Thanks to Unilever's data glossary, we had only to focus on the business terminology, where we recommended Unilever business experts to prepare a terminology dictionary that has all the definitions that describe all the used terminologies and metrics they use and the same will be shared with the AP.

Costs and Benefits

A cost-benefit analysis of the project has to be performed in line with Unilever's templates and approval process. The key of this phase is to compare the costs of the project with the potential benefits of the same.

Example

The aim of this project is to save vast amounts of money to our company in market research that once took months and cost millions will be able to be performed for a fraction of that price and in mere days. The percentages of investment can be summarized as follows:

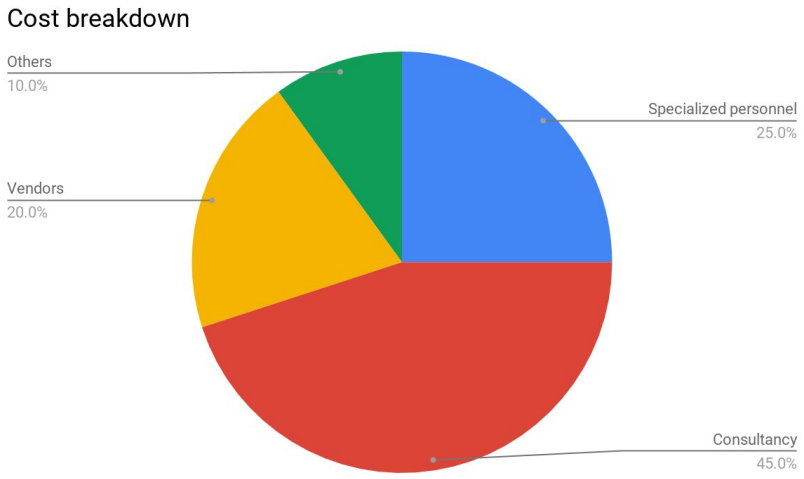

Figure - 2: Cost distribution

The benefits that will be enabled through this platform will be as follows:

Benefits	*Exhibit*
Be able to work with all the data types	*↑ Efficiency*
Uncover new business opportunities	*Beauty & personal care: 45M $* *Foods & refreshment: 35M $* *Home care: 30M $*
Upskill our teams	*↑ Productivity*
Higher understanding of the market and our competency	*Beauty & personal care: 25M $* *Foods & refreshment: 20M $* *Home care: 15M $*
Better monitoring of our brands	*↑ Auditing*
Digitalize our environment	*↑ Awareness*

Determine Data Mining Goals

This phase is related to the technical outputs expected of the project. A data mining goal should define the project objective in technical terms, enabling the technical team to better understand the goals while looking at the data and extracting insights. Two main activities shall be performed:

1. Translate the business queries into data mining goals
2. Specify the data mining problem type

Example

The amount of business queries for this project is unlimited, since the main goal is to create a unified platform that processes business queries across all the organization, therefore the data mining problem refers to the business intelligence required by each department. Many modeling techniques will be applied such as clustering, description and summarization, prediction, classification and segmentation.

Produce project plan

Project Plan

The plan required to achieve the data mining goals should be described in this specific section along with the duration of each activity, resources required, inputs, outputs, and dependencies. An estimation on the efforts and resources required in each phase should be shown marking the decision and review points across the project. A gantt diagram might be handy for a quick overview of the plan, or a more advanced approach with Oracle Primavera can be performed.

Example

Figure - 3: Project program *(internal source)*

Initial Assessment of Tools and Techniques

In this phase, the project team performs the first assessment of the tools and techniques selected in the previous steps to verify that they fit the project requirements and specifications.

Ready for the next step?

Before starting to explore your data, make sure that you have the answer to the following questions.

<u>From a business perspective:</u>

- What does your business hope to gain from this project?
- How will you define the successful completion of our efforts?
- Do you have the budget and resources needed to reach our goals?
- Do you have access to all the data needed for this project?

- Have you and your team discussed the risks and contingencies associated with this project?
- Do the results of your cost/benefit analysis make this project worthwhile?

From data mining perspective

- Do you have an idea about which data mining techniques might produce the best results?
- How will you know when your results are accurate or effective enough?
- How will the modeling results be deployed?
- Have you considered deployment in your project plan?
- Does the project plan include all phases of CRISP-DM?
- Are risks and dependencies called out in the plan?

Data understanding

Collect Initial Data

Initial Data Collection Report

Here the practical works come into the scene; we must start acquiring the data, or simply access the data that we have listed in the previous steps. This initial collection might include data loading, if required for data understanding. The selection of the data should be always made in line with our data mining goals.

Example

Our data will be collected for the purpose of creating a Data Lake environment based on a Big Data Platform. Unilever have different data sources, which can be listed as follows:

1. CRM Data: this includes the complaints and service requests information related to the consumers claims and it's stored on Oracle database. We expect that this data will be extracted as is to the distributed file systems HDFS as a part of the data lake, their tables will be prefixed with "CRM".
2. Data Files: all the data files are distributed on all the departments and sections of excel or csv formats will be initially integrated and loaded into the data lake through the Data Management tool will be acquired, these data tables will be prefixed with "DAT_DEPT". Dept: Stands for the department name and it will be secured to be only available and visible to the employees of the specific department. After going live,

there will be an option to update these files through a button called “Upload” to extract the new data files and update them on the final platform.

3. Social Media Data: this is a dynamic information and it will be customizable through the same platform by selecting the topic / hashtags and then the engine will integrate with the social media listener to execute the same and save the results directly into the data lake for different analysis methods and insight visualization.
4. Influencer Analysis Data: It will be very similar to the social media data in concept but with different execution engines, as the end user will choose an option to perform an influencer analysis and then set some parameter values as the following:
    a. Influencer relevant Category. (Personal Care, Home Care, Hair Care, Food, .. etc.)
    b. Region (Europe, Middle East, North Africa, ...etc.)
    c. Country (it is an optional parameter that will be addressed if needed)
    d. Language spoken (This will be more relevant to the targeted audience language)
    e. Platform used (Which Social Media platform used: Facebook, Twitter, Instagram, ...etc.)
5. Sales Logs from Partners: thanks to the partnership between Unilever and Amazon, this data will be helpful to analyze the online sales progress. Unilever will offer an integration service with the partner to collect the sales data on a real time basis through an API.

Describe Data

Data Description Report

Assess the properties of the acquired data in a generic manner for each datasource, describing the relevant features of our data, such as format and quantity. During this step would be sometimes interesting to perform an EDA to get a glimpse of our data.

Example

Following the same sequence as the previous step, below is an elemental description of our data. Due to the nature of our business goal, getting an in-depth description of our data does not align with our main project objective.

1. CRM Data: it’s a relational database based on an Oracle instance. It has about 25 tables. Most of the column data types are Numeric, String and Dates. It simply describes the service requests raised from consumers, distributors and sellers. The data will be available only for self-service analytics
2. Data Files: are based on excel and csv formats and they have basic different data types of Numeric, String, and dates. These files have different information related to each department, so the data is needed only for self-service analytics as well
3. Social Media Data: This data will be on-demand collected through the different social listener platforms, and it will include the following information in csv file format:
    a. Mention Date
    b. Author Name

    c. Author Gender
    d. Author Email Address
    e. Mention Content
    f. Author Reach
    g. Author Impression score
    h. Mention location
    i. Mention url
    j. Mention Language
    k. Continent
    l. Account Type
    m. Channel Name
    n. Hashtag
    o. Kred Influence
    p. Kred Outreach
    q. mozRank Score
4. Influencer Analysis Data: This data will be on-demand collected through the influencer analysis platform, and it will include the following information in csv file format:
    a. Influence Score
    b. Influencer bio
    c. Influence Channel
    d. Average Age Followers
    e. Dominant Gender
    f. Top 5 Countries
    g. Top Cities
    h. Engagement Level
    i. True Reach
    j. Influential topics
    k. Top Location
    l. Number of fans
    m. Categories scores
    n. Hashtags mentioned
5. Amazon Sales: this will be a .json file that will be including the sales information for each Unilever's product related transaction, and will include the following fields:
    a. Transaction Date
    b. Buyer name
    c. Buyer email address
    d. Quantity
    e. Product main category
    f. Brand
    g. Item SKU
    h. Price
    i. Paid Amount
    j. Unilever profit
    k. Amazon commission

Explore Data

In this task we must focus on dealing with the data mining questions that can be tackled using querying, visualization, and reporting techniques. Our goal is not to answer directly our data mining goals, but to polish our data description, and feed into the transformation and other data preparation steps required before we enter into further data analysis.

Example

In our project the data description performed in the previous step would be enough to accomplish the whole project's objective, however below we further describe the approaches followed for each data source:

1. CRM Data: Unilever expects to have a proper DWH model for the CRM application that applies the multidimensional approach to facilitate self-service analytics for the business stakeholders. Here is an example:

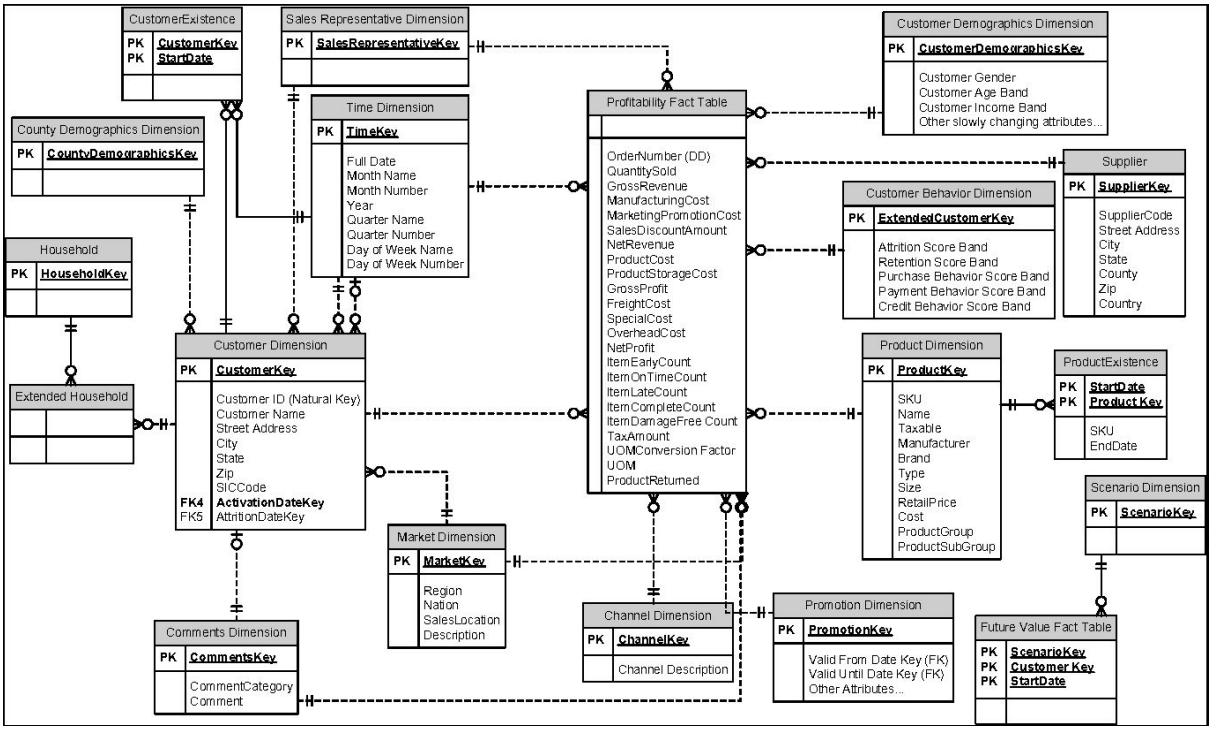

Figure - 4: Proposed CRM data warehouse model

2. Data Files: These files will not be analyzed, The AP needs to extract these files into the data lake without any needed analysis but only having the capability of visualization, standard charting and reporting features.

3. Social Media Data: The standard data file will be in csv format and will have 112 columns and will be including all the collected information from each mention / post extracted through each query. This file format will be shared once Unilever and AP agree on the platform. Here is an example

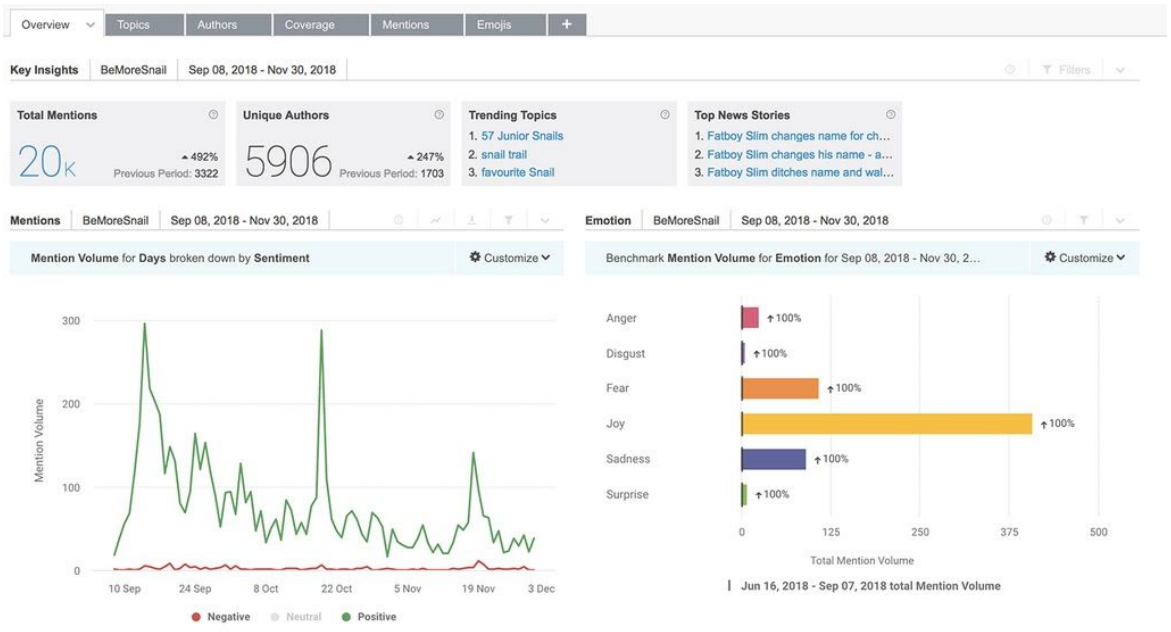

Figure - 5: Social medial data format

4. Influencer Analysis Data: The data will be extracted through the influencer analysis platform and its template will be fixed once Unilever and AP agree on the platform. Here is an example:

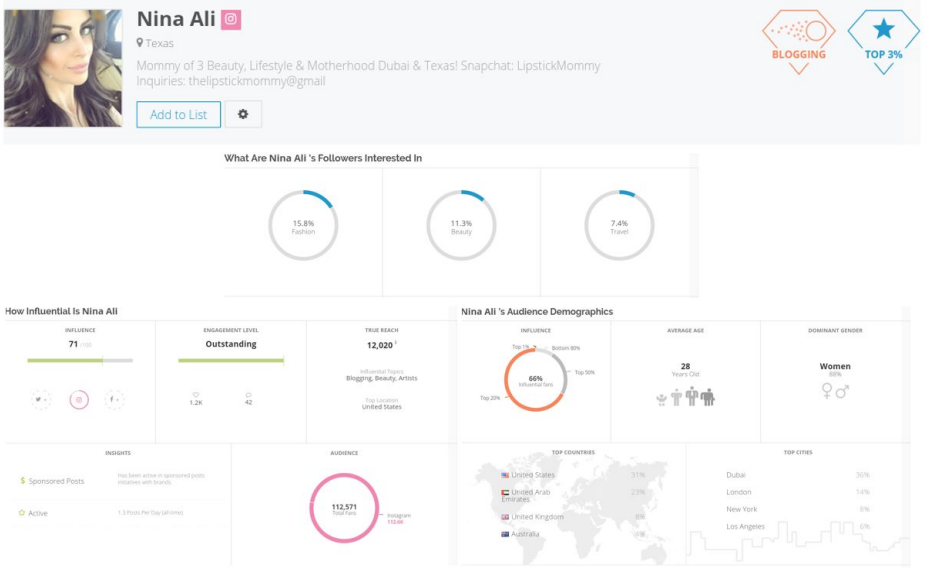

Figure - 6: Influences data format

5. Amazon Sales Data: The data will be extracted as a json file format that has the needed information, this file will be integrated through an API within a specific time basis. Here is an example:

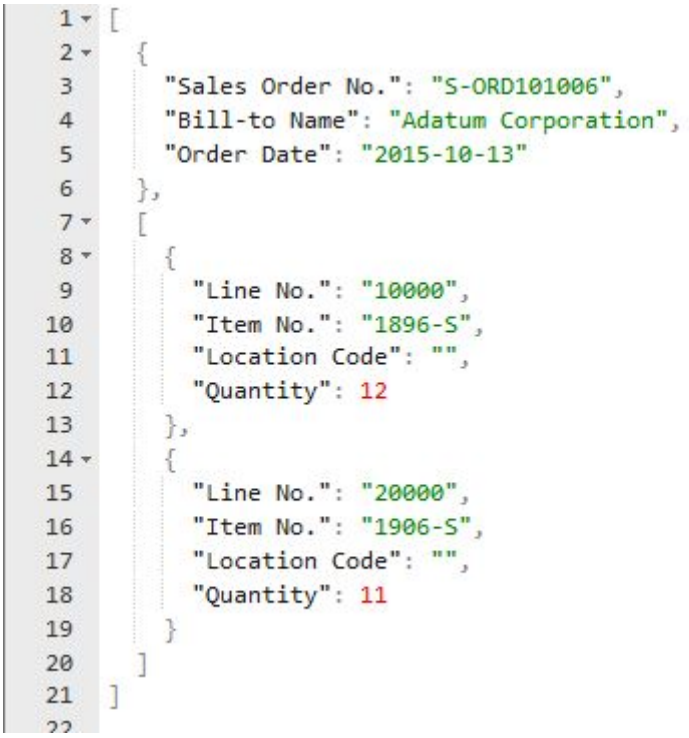

Figure -7: Amazon sales data format

Verify Data Quality

Here is where we need to verify the quality of the data in order to check how complete it is, whether there is a certain pattern of error that keeps on repeating itself and decide what to do with the missing values. We must have sufficient business knowledge in order to make the right decisions with the data and avoid any manipulation that might affect the final output.

Example

The appropriate methodology to verify the data quality will be applied for each datasource in order to maximize the accuracy of our results:

The CRM Data warehouse and OLAP will be tested and evaluated against the OLTP through designed projects for each section like Customer Orders, Complaints, Contacts, Price Lists, and Service Requests.

The Data files will be revised with the concerned stakeholders who work and handle these files during their daily operations. Their broad business and specific technical knowledge will serve our best benefit for this task.

Social Media & Influencer Analysis parameters will be revised and tested with various cases to ensure that the results match the business team expectations. Extracted Mentions will be tested and reviewed to confirm that it doesn't have spasms, irrelevant

posts/comments, and ensure that the conditions are properly applied. The business teams will share their experience with the AP business and data analysts about the common mistakes and irrelevant mentions and posts that always appear with any related social query.

Amazon Sales data will be tested through different reports for specific previous dates to ensure that the numbers are matched.

As example, Unilever will use the following reports for testing:

1. Total sold quantity and amount per Day / Month / Quarter and Year
2. Total sold quantity and amount per Product
3. Total sold quantity and amount per Category
4. Total sold quantity and amount per SKU

Ready for the next step?

Before preparing the data for modeling, let us consider the following points:

- How well do we understand the data?
- Are all data sources clearly identified and accessed? Are we aware of any problems or restrictions?
- Have we identified key attributes from the available data?
- Did these attributes help us to formulate our hypotheses?
- Have we noted the size of all data sources
- Are we able to use a subset of data where appropriate?
- Have we computed basic statistics for each attribute of interest?
- Did we use exploratory graphics to gain further insight into key attributes? Did this insight reshape any of our hypotheses?
- What are the data quality issues for this project? Do we have a plan to address these issues?
- Are the data preparation steps clear? For instance, do we know which data sources to merge and which attributes to filter or select?

Data Preparation

Select data

In this section you must decide on the data that will be used on the analysis based on the business goals that you intend to achieve. The selection criteria must be based on the following:

- Mining goals
- Technical constraints
- Quality

The key point that should not be missed here is to mention the underlying reason for selecting (or not selecting) certain data.

Example

Our data comes from several sources as described in the data understanding section, and due to the nature and objective of our project the data selection would be mostly performed by the specific business analytics unit, therefore our aim would minimize our interference in the potential business queries by cutting off data that might be relevant to certain units, therefore the data selection will vary from dataset to another:

CRM Data warehouse will cover all the application data tables and attributes to offer the best insightful information for business stakeholders and management teams.

Data Files will be fully considered in order to keep or increase the different team productivity and any petition to add or remove any attribute or file, will be approved based on the business teams request.

For Social Media Mentions and Influencer Analysis, we will be utilizing the following attributes:

1. Social Media
    a. Mention Date
    b. Author Name
    c. Author Gender
    d. Author Email Address
    e. Mention Content
    f. Author Reach
    g. Author Impression score
    h. Mention location
    i. Mention url
    j. Mention Language
    k. Continent
    l. Account Type
    m. Channel Name
    n. Hashtag
    o. Kred Influence
    p. Kred Outreach
    q. mozRank Score
2. Influencer Analysis

a. Influence Score
b. Influencer bio
c. Influence Channel
d. Average Age Followers
e. Dominant Gender
f. Top 5 Countries
g. Top Cities
h. Engagement Level
i. True Reach
j. Influential topics
k. Top Location
l. Number of fans
m. Categories scores
n. Hashtags mentioned

Amazon Sales Data will be integrated through an API as per an agreed template format to include specific attributes, therefore Unilever will consider all these attributes to cover an insightful reports, for example:

1. Total Sales Amount per time basis
2. Product Sales competition within the same category per time basis
3. Category Sales competition
4. Revenue Sales Forecasting per Category
5. Correlation score on the category / product level.

Clean Data

The selected data must match the quality and cleanliness required by the analysis techniques selected. This includes the estimation or dismissal of missing values and the insertion of suitable defaults that supports the modelling techniques. Is important to get rid of the noise in the fields that lack significance. These modifications must always be documented and in line with the final business goals.

Example

We will be using the data cleaning process to address the problems noted in the data quality report. Due to the nature of our project, the below data issues will be addressed with each business unit, in order to ensure high data reliability a unnecessary noise

- Missing data
- Data error
- Measurement errors

Construct Data

This phase consists of trying to create new attributes from the available data through the combination of the same with the objective of enabling a further business understanding and assisting during the modelling stage. It is important to create new attributes that are business related and aligned with the project's objectives so to avoid unnecessary noise.

Example

Designing a data warehouse or data mart will need to have aggregated attributes as measures to show insightful information. For instance, in the CRM case we will be requiring the following measures to be available:

Sales Module:

a. Lead conversion rate.
b. Number of opportunities generated from a lead.
c. Number of leads generated.
d. Number of customers generated from a sales lead.
e. Number of orders generated from a lead,
f. Average number of quotes from a sales lead.
g. Revenue forecasted for an opportunity.
h. Order amount for an opportunity.
i. Number of days estimated for an opportunity to generate an order.
j. Total budget amount for all of the opportunities generated in a specific quarter.
k. Revenue allocation percentage for a sales person.
l. Qualification goal of an opportunity based on its selected qualification.
m. Qualification score of an opportunity compared to the opportunity qualification goal, answering whether the opportunity is qualified.
n. Difference between actual opportunity sales days and estimated opportunity sales days.
o. Number of orders generated from a lead.
p. Number of orders generated from an opportunity.
q. Number of transactions associated with an order.
r. Number of service orders associated with a particular case.
s. Number of RMAs associated with a particular case.
t. Number of orders and corresponding revenue averaged per month, for a particular sales rep.
u. List price for a product.
v. Average quantity ordered for a product.
w. Price of the periodic recurring charge for a product.
x. Frequency of the recurring charge for that instance of the product sale.
y. Actual price used for an order.
z. Average discount and selling price for a product.
aa. Average quote price for a product.
bb. Average discount for a product.
cc. List price for a product.
dd. Quantity of product used for a set of quotes for that product.
ee. Average quota amount for a sales person,
ff. Total forecasted revenue amount in a sales timeframe.

gg. Actual revenue for a sales person in a sales timeframe.
hh. Average discount and selling price for a sales representative.
ii. Discount amount offered to a customer by a sales representative.
jj. Number of quotes for a particular product opportunity.

Figure - 8: Example of CRM data

Service Module:

a. How long a case has been open.
b. Number of cases with an open status.
c. Number of opened cases.
d. Number of cases with a closed status.
e. Number of reopened cases.
f. Number of cases closed on the same day.
g. Number of times a case is associated with other business processes, such as a case generating a sales lead or an order.
h. Number of cases generated form a sales lead.
i. Number of cases generated for an order.
j. Number of cases generated with an RMA.
k. Number of cases followed with a customer survey.
l. Number of cases related to a product defect.
m. Length of time to solve cases for a specific product.
n. Number of cases resolved using solutions from the solution library.
o. Average number of resolutions for a support case.
p. Number of cases closed without resolution.
q. Number of customer surveys.
r. Weight value for a survey script.
s. Score for a survey script.
t. Customer satisfaction with the results of their cases.
u. Number of open, closed, or pending cases with respect to the snapshot date.
v. Age of a case with respect to the snapshot date.

w. Average number of days taken to close support cases.
x. Rates of first call resolution.
y. Average wait time for customer calls.
z. Average call volume for a specific month.
aa. Average duration for customer calls.
bb. Number of outbound or inbound calls for a specific day.
cc. Average number of agents handling a customer call.
dd. Total call time in seconds.
ee. Total call conference time in seconds.
ff. Number of chats for a specific day.
gg. Number of chats related to cases.
hh. Number of chats related to an order.
ii. Chat length.
jj. Total number of inbound emails to support customers.
kk. Number of SPAM emails to support customers.
ll. Number of outbound emails.
mm. Number of closed emails when supporting customers.
nn. Number of times an email was assigned.
oo. Average agent assignment for handling emails from customers who requested support.
pp. Average time duration before an email is handled by an agent.
qq. Number of entitlements for a specific customer and product.
rr. Which agreements and entitlements exist for my customers.
ss. Number of days does a particular service agreement expire.
tt. Number of times a particular service agreement renewed.
uu. Total number of customer interactions associated with a sales lead.
vv. Total number of customer interactions associated with a support case.
ww. Total number of customer interactions associated with a service order.
xx. Total number of customer interactions associated with an order capture.
yy. Number of calls/emails/chats received for a particular day.
zz. Number of times a particular resolution was used to solve a problem.
aaa. Number of cases were solved with a particular solution.
bbb. Number of times the solution actually solves the problem.

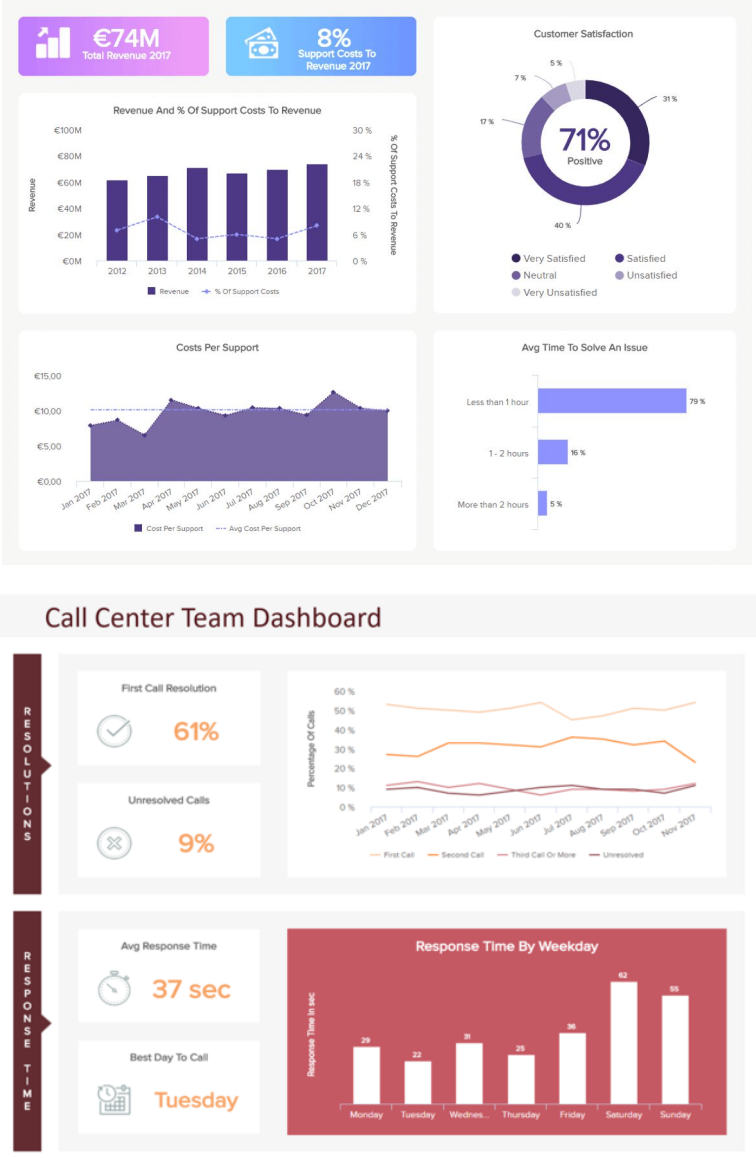

Figure - 9 & 10: Example data from service module

Integrate Data

This step covers the integration of various tables, if required, to assist on achieving the latest mining goals. The tables joined must have different information about the same objects. It might be also useful to generate aggregate values which will enable us to summarize information from multiple tables.

Example

Unilever needs to ensure that the data warehouse / marts integrity will be fully implemented and validated by the AP according to the standards and best practices.

Unilever has a conceptual Data Lake architecture as following:

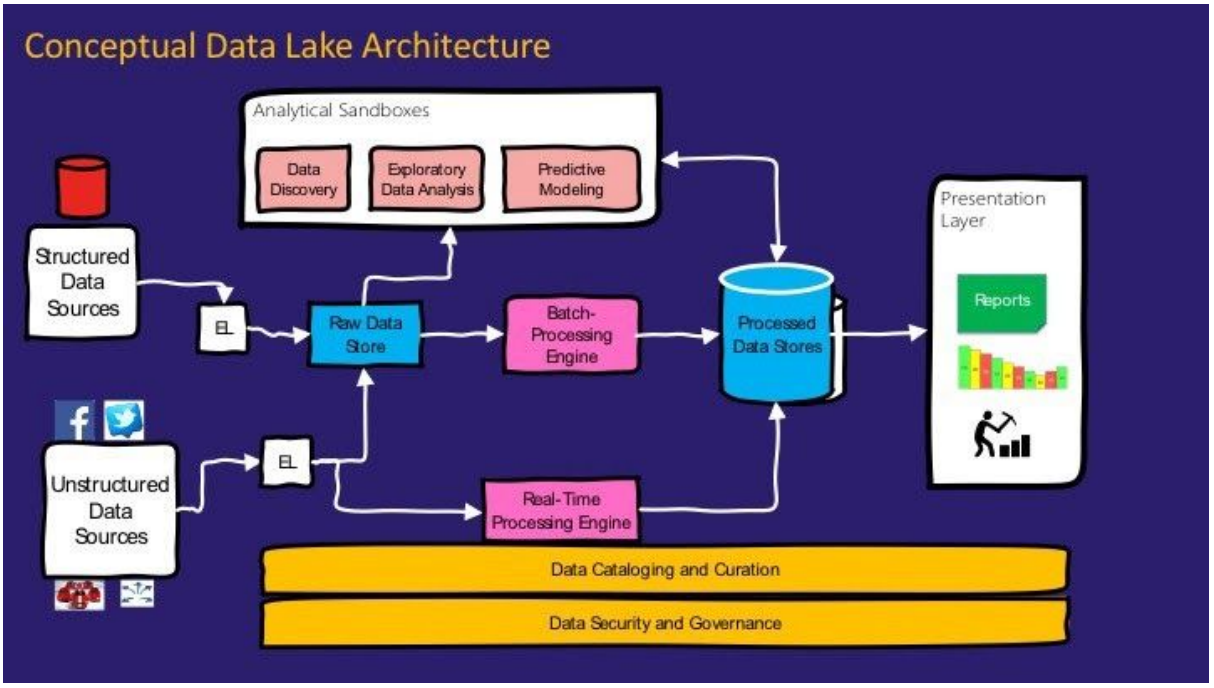

Figure - 11: Data lake architecture

This should be communicated with the Unilever Technical Team to review and confirm the final data architecture that matches the requirements and ensure that all the data points are integrated and accurate.

Format Data

In this step we must focus on adapting the data to the model deployed in the next phase. It is crucial to not interfere or alter the business meaning with these changes in the format. Having a deep knowledge of the tools that will be used in the next step is a must so to elude possible pitfalls.

Example

The data source, especially the structured data like CRM and Amazon Sales, will keep working as it is and no difference will happen on them unless there's any new functionality of process change will be applied.

In the case of applying any new functionality, this will be communicated with the Analytics team to reflect these changes to the analytical side as well.

The great main idea is to have a raw data zone automatically refreshed on a timestamp-basis, from this step the analytics team will go forward in their regular process to analyze, model , predict and visualize the insights.

This will help a lot to not impact the source system(s) for any data analytical purpose.

Ready for the next step?

Before starting building our models, let us make sure that we have answered the following questions.

- Is all our data accessible and ready for our models?
- Based upon our initial exploration and understanding, were we able to select relevant subsets of data?
- Have we cleaned the data effectively or removed unnecessary noise?
- Are multiple data sets integrated properly? Were there any merging problems that should be documented?
- Have we researched the requirements of the modeling tools that we plan to use?
- Are there any formatting issues we can address before modeling?

Modeling

Select Modeling Technique

In this phase we must select the technique that best satisfies our final business goals. It is always advisable to start filtering only for those that will be appropriate for the problem and always be aware of the political/business constraints. We also need to consider other restrictions specific to the model that we chose, i.e; the time it would take to run the model vs deadlines, how deep our knowledge is about the specific model, etc.

Example

Unilever site will be private cloud-based that will have a centralized data management platform including more than a model to achieve the business goals.

Because of the lack of information and decision making support tools, Unilever will propose separate solutions for each data source as follows:

1- CRM Data Source: The main objective for this information is to have a complete business intelligence platform that shows the insights, opportunities and trends related to the sales and service modules.

This business intelligence system will be based on a Data warehouse model including all the needed considerations for OLAP cubes, aggregated tables, product hierarchies, lead segments, customer layers and pre-defined measures.

- The DWH model will be auto refreshed on a daily basis within non work hours / the most relaxed time / non rush hours for the centralized CRM database. This will be revised by Unilever's infrastructure team to decide the proper time.
- Unilever will use ELT methodology to replicate the new updates into the Data Lake.
- Data Lake has 2 main areas: Raw Data Store and Processed Data Store.
- The refreshment approach will use EL method to extract the raw data from the main data sources and load it into the Raw Data Store then Transform this data into the needed format / aggregation in the Processed Data Store.
- Cloudera Data Flow CDF technology will implement the above approach.
- SCD type will be agreed with the AP but as for now Unilever expects to implement type 2 to maintain the historical data for reporting and analysis purposes.
- The next step of analyzing and visualizing the data will be handled by Dataiku.
- Unilever expects different predictive analytics algorithms to be implemented above this information to have a vision for the revenue forecasting, Customer Segmentations and to identify more opportunities through the revenue factor's correlations.

2- Social Media Listening & Influencer Analysis:
The data will be extracted from social media channels by 3rd Party Applications

- The 3rd Party applications will extract the data as per the queries parameter inputs from the business stakeholders.
- The extracted data will be replicated into the Raw Data Store on Data Lake using CDF data streaming.
- Once the data is replicated into Unilever's Data Lake, it will appear on Dataiku as a Dataset / data source that has the same query name to be analyzed and visualized.
- This will happen every pre-defined time stamp to ensure that the new updates are extracted and considered.
- There are 2 main reasons to extract this data from Application Cloud into Unilever's Data Lake:
    a. These cloud-based applications do the cost estimation based on the used spaces and the number of extracted records (as a total).
    b. The data will be stored in Unilever's site for any analysis purpose.
- To maintain and save Unilever's site of having obsolete data files or very high storage usage, The new updated data files will replace the old ones as the new files will include old and new data in the same new file.
- To save the cost and resources from the cloud-based 3rd parties, it will be configured to purge the data files older than a month.

<u>3- Amazon Data Source:</u> this 3rd party information is integrated through an API.

- Unilever will use CDF stream processing to replicate this data into Data Lakes' Raw Data Store.

- The analysis models process the Raw Data and store its results into the Processed Data Store.
- Dataiku will take care of refreshing the analysis models and visualizing them.
- The analysis models and their reports / dashboard will be refreshed twice a day.

4- Data Files: For the departments' and stakeholder's data files, they use them on a daily basis.

- The AP will be responsible to upload all the files into Raw Data Store for the first time then it will be the stakeholders and departments' focal points responsible to upload the new updated files / add new files.
- AP will set up the needed analysis and visualization for the frequent dashboards and reports for the stakeholders and departments management.
- Any new data files uploaded and analyzed after the support period, will be the stakeholder's self-service responsibilities with Unilever's team internal support.
- Each Department has a separate folder to upload their datasets and structure it as per their needs, this department employees only have access to this folder and its content.
- The Department Director has the full rights to manage the department folder, give and revoke access from each user in line with the security and confidentiality levels
- There's an automated report to be sent on monthly basis to each department's focal point including the below information to clean the not needed files and save the used storage:
    a. File name
    b. File Owner
    c. Creation Date
    d. Last Updated Date
    e. File Size

Generate Test Design

This task entails moving forward with the most relevant modeling technique and ensuring each phase of the model is executed properly. We must start defining the quality measures that will be used and depict the intended plan for training, testing, and evaluating the models. You can think of this as the blueprint of our model.

Example

Unilever needs to have a data quality report for each data source in both areas Raw Data and Processed Data to be always matched and reconciled with the transactional source.

- Raw Data Area Quality Assurance will assure and show the result of the full data matching between the Data Lake's Raw Data Area and the data sources for specific cut-off date.

- This scenario will be tested at least for 3 times to ensure the data accuracy for all the data types, number of records, total amounts, .. etc. are the same and 100% matching the transactional source.
- The business modelling layer / processed data area will be revised for the same cutoff dates to ensure the exact same aggregations reflect what have been tested in the transactional source and raw data area.
- The predictive analytics models results will be reviewed and compared with other research and studies that have been done by Unilever before to confirm the trends, opportunities identified, and proper customer segmentations.
- The predictive analytics models measures will be monitored by Unilever's analytics team before the AP delivery to ensure its future support through reviewing the accuracy, mean squared errors, R square, … etc.
- Application Integration Interface API will be monitored and notified immediately to Unilever's Technology team in case there's any error.
- As CDF technology has great capabilities to identify the bottlenecks through the data extraction, loading or transformation, it will be reviewed by Unilever's data engineering team to address any improvement needed in any data workflow.
- Data Workflows will update a status table with the workflow status after its completion including the workflow name, starting date time, ending date time, duration, the completion status (Successful, Aborted, Cancelled, or Failed) and error message (if applicable)
- The log table will be historical based information for analysis purposes for Unilever's technical team.
- Email notification will be sent to the analytics team, in case there's any workflow error.
- Data Mapping will be reviewed to ensure the data completeness.
- Plan the required training for each business unit aligned with the project plan.
- Plan the unit, integration, smoke, system, regression and user acceptance test for each phase.

Build Model

Run the model in the predefined tool and always record the parameters selected and their respective values. Providing a brief explanation on why certain values or parameters were chosen would be of a big help for future projects or if we want to tweak our model to assess different business needs.

Example

Our tems expects the overall model's architecture to be as the follows:

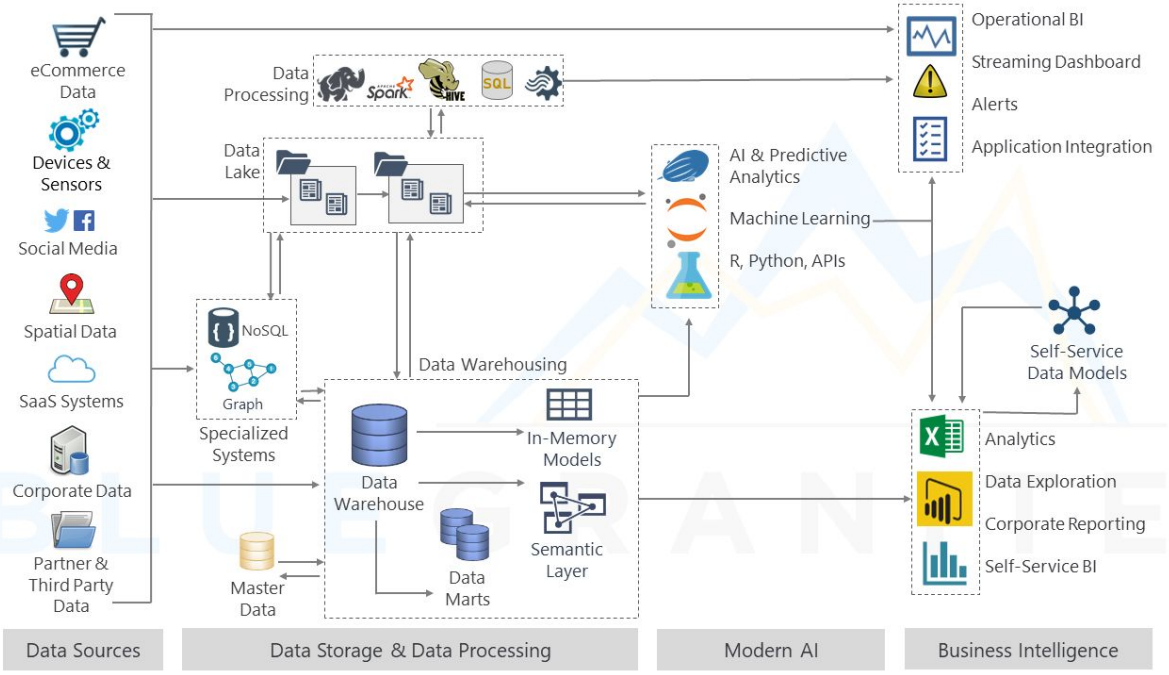

Figure - 12: Insights engine architecture

- All the data sources to be loaded into Data Lake's Raw Data area then to be processed and stored into the Processed data area.
- The main data processing engine is the big data platform.
- There the processed data will be integrated between the data types that come from structured data warehouse / data mart with other data types that are already processed using other methods.
- All the data types will be available for applying any type of predictive analytics using machine learning and AI algorithms.
- The Raw data will be always accessible to the modern AI area to have the benefits of applying any pure calculation on this data directly.
- The output from the processed data are as well as the modern AI area will be available for the visualization capabilities.
- Currently, for this project and as an essential digital transformation step, Unilever will use Dataiku as an essential part for the modern AI.
- Unilever will use the visualization capability through 2 tools: Dataiku and Tableau.

Asses Model

Here we must interpret and assess the model from a technical point of view to see how fit it is compared to the initial business goals and criteria. Fine-tuning the parameters based on this evaluation to get an improved model would be the final goal.

Example

Unilever will asses the whole proposed solution by checking the following criteria:

- Scalability:
    a. How much scalable the infrastructure resources?
    b. How many data sources can be added into the ingestion phase?
    c. How easy can Unilever add any more resources for CPU, Memory or storage?
    d. What if there's any new source engine to be connected ?
    e. What's the modern AI scalability to add any new tool like Rapid Miner, Alteryx, PyTorch?
    f. Can Unilever use any other visualization tool as a replacement or in parallel with Tableau?

    - Performance:
        a. How long does CRM data refresh take on a daily basis?
        b. How fast is API data integration?
        c. How fast is the data streaming ?
        d. How many Developers / Architects / Testers will be needed to handle the whole solution?
        e. How long does it take to generate a report through tableau / dataiku?

    - Automation:
        a. How much the solution is human based?
        b. Does the data refreshment need any human interaction?
        c. Does data quality assurance need human check?
        d. Should the data analytics team investigate the non matched / non reconciled information to be identified?
        e. Can the data architecture identify the data flow bottlenecks and alert the analytics team?
        f. Can the cloud system alert the technology team once there's a specific shortage of CPU, Memory or Storage?
    - Re-Usability:
        a. Can the business user reuse their visualization template with another data source / report?
        b. Can the data analytics team reuse a model for another purpose?
        c. Can the business users reuse a social media query?
    - Usability:
        a. How easy is the whole platform to the business user?
        b. How easy is uploading a new dataset?
        c. How easy is filling social media query parameters?
        d. How easy is creating a self-service reporting?
    - Documentation:
        a. Did the AP deliver technical documentation?
        b. Did the AP deliver a user guide?
        c. Did the AP deliver the data mapping sheets?

d. Did the AP deliver a compliance sheet against Unilever requirements?

Ready for the next step?

Before moving on to our final evaluation of the models, we should consider whether our initial assessment was thorough enough or not

- Are we able to understand the results of the models?
- Do the model results make sense to us from a business perspective? Are there apparent inconsistencies that need further exploration?
- From our initial glance, do the results seem to address Unilever's question?
- Have we used analysis nodes and lift or gains charts to compare and evaluate model accuracy?
- Have we explored more than one type of model and compared the results?
- Are the results of our model deployable?

Evaluation

Evaluate result

This phase will include the practical involvement of the different business units to test the processes and the platform usability and flexibility themselves, and they will have all the rights to ask for any enhancements or improvements needed in the platform.

In this phase we must assess how well the generated model from the previous steps meets the business objectives in terms of an evaluated versus the defined success criteria. The evaluation of the results should involve key personnel from the business and technical staff as well.

Example

The overall results of the experience with the insights engine are fairly easy to communicate from a business perspective: the study produced what are hoped to be great at being the sole engine for the organization to gain insights. The UX was enhanced after some feedback of the business units. To produce the final report, the analysts will try to identify some general trends in the rulesets that can be more easily explained. Most of the business stakeholders were involved in this phase. On the other hand, the business users will ensure

that the business principles are achieved and considered as per their initial requirement documents and agreement with the AP.

Review process

Assuming that our model provides satisfactory results against the defined business objectives, this step involves a review of all previous steps of the process in order to evaluate if any critical steps were missed and if the business-related facts are complete. This task can be seen as a quality control review for the overall process to check for inconsistencies or enhancements. Any missing parameter, step during the process, or technical requirements to be added for more control should be mentioned in this phase.

Example

The evaluation criteria was successfully applied and the business units were happy with the achievement. The final step is for the Unilever technical team to review the whole process to ensure that everything is in line with the expectations and no technical issues will arise in the future. This is an internal audit which must be carried out in compliance with the ongoing internal quality policies.

Determine Next Steps

This step involves the project team considering potential next steps relating to improvements of the process and model, as well as the rationale supporting or against each step. Based on the results of the previous process, this step should be decided on whether we are ready to proceed to the next phase or we need some additional iterations.

Example

Unilever is fairly confident of both the accuracy and relevancy of the project results and so is continuing to the deployment phase. At the same time, the project team is also ready to go back and include any business units that might request to be incorporated into this insight engine. At this point, we have only to wait for the green light from the decision makers and from the internal quality team to proceed with the deployment.

Deployment

Plan Deployment

Ideally we want to divide this phase into two:

- Technical deployment: with the aim of deploying in the company's IT infrastructure the model built in the analytical environment.
- Business deployment: where Unilever must integrate the model or tool built into the company's operations, selecting the appropriate threshold (if any) and designing a periodic monitoring of the tool created. This part should always be consensuated with the stakeholders and the business units involved.

This plan should summarize the deployment strategy and the necessary steps, for both, the technical deployment and business deployment.

Example

Unilever will prepare their on-premise environment and the cloud subscription with the recommended sizing from the AP. On the other hand, the business units are preparing their manual work closure plan to be ready for the automated system. Keeping in mind, that any current process will keep running normally in order to not impact the business till it ends, and then share its results through the new platform for tracking and analysis purposes.

In order to have a smooth deployment from the business perspective, training will be provided to all the business units involved with the aim of letting them take the most advantage of this platform and run it in a smooth manner without compromising any of their current activities.

In order to ensure a successful development, the main players must be involved, so the right information reaches the right people. Above all, the project team needs to keep in touch with each of these players to coordinate the deployment of results and planning for future projects.

Plan Monitoring and Maintenance

Once the final data mining tool or model becomes part of the every day business operations and its environment, we should design a maintenance strategy in order to keep this tool in optimal conditions throughout its usage and avoid any downtime which might jeopardize the business operations.

Example

The critical task for monitoring is to determine whether the insights engine created gathers and processes the right information as required by each business unit. This monitoring and maintenance can be divided as follows:

1- Technical Monitoring

Unilever IT Infrastructure team to share the maintenance strategy including the following:

a. Service level agreements SLAs including the function name, severity, priority, categorization, the expected level of support, expected time to support and how to calculate its KPI.
b. Expected periodic maintenance for platform operation enhancement,
   As example: Every 1st Saturday of every month to clear the cache, and archive the logs.
c. The monitor dashboard specification to monitor the platform performance and resources.
d. Time-basis report to be shared for the cost effective / reduction opportunities by increasing / decreasing the infrastructure components.
   As Example: the infrastructure resources being utilized 97% of each last 2 days of every month but utilized between 30 to 45% during the whole month.

2- Business Monitoring

Different business Units should always share their feedback and suggestions regarding the following:

a. Which capabilities should be added to the implemented solution to add more values.
b. Which capabilities need improvement / enhancement.
c. What data type / source can be added to make it more meaningful.
d. Which feature can be revised to give more flexibility.

Produce the Final Report

This final report is considered as summary of all previous deliverables. It should be partially technically and business oriented; it must verify how well the goals were achieved compared to the initial business and data mining goals set, as well as any deviations from the initial plans, the costs incurred and last but not least any recommendations for future works. It is very important to identify the audience for which the report is intended to.

The report should include the following points:

a. A thorough description of the original business problem
b. The process used to conduct data mining
c. Costs of the project

d. Notes on any deviations from the original project plan
e. A summary of data mining results, both models and findings
f. An overview of the proposed plan for deployment
g. Recommendations for further data mining work, including interesting leads discovered during exploration and modeling

Final Presentation

This step is where we present our results to the management team, sponsors or stakeholders. We must always maintain the level of technicalities to a level that suits our audience and must focus on the business advantages achieved with this project. We can take advantage of the report done on the previous step to extract some information and structure it in an appropriate manner.

Review Project

We should consider this step as an auto-evaluation or assessment of the whole process, where we must identify what went wrong and what was done to overcome the hardships encountered throughout the project. Keeping in mind that this document can be of crucial importance for the next team that wants to implement a data mining project.

All experiences that can be of certain use should be listed along with the persons involved in all stages of the project. Specific references to people within the organization that helped getting around the adversities encountered must be mentioned. And finally, the feedback obtained from the sponsors and stakeholders after the completion of the project is worth to be mentioned.

The review document should have answers for the following questions:

a. What are your overall impressions of the project?
b. What did you learn during the process
c. Which parts of the project went well?
d. Where did difficulties arise?
e. Was there information that might have helped ease the confusion?

References

Resources shared in IE Campus - *The knowledge discovery process subject*

https://hbr.org/2016/09/building-an-insights-engine

https://www.researchworld.com/how-is-technology-disrupting-marketing-research/

https://www.semanticscholar.org/paper/Data-warehouse-design-to-support-customer-analyses-Cunningham-Song/535c6de27c79cc7ee3d41d89df8fc07ff7d0f34d

ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/17.0/en/ModelerCRISPDM.pdf

https://docs.oracle.com/cd/E41507_01/epm91pbr3/eng/epm/pcrw/concept_UnderstandingtheCRMWarehouseStructure-399bf2.html

Figures

Figure 1
https://epicinnolabs.com/2019/11/15/data-science-magic-for-industry-of-the-21st-century/

Figure 2
Internal source

Figure 3
Internal source

Figure 4
https://www.semanticscholar.org/paper/Data-warehouse-design-to-support-customer-analyses-Cunningham-Song/535c6de27c79cc7ee3d41d89df8fc07ff7d0f34d/figure/1

Figure 5
Internal source

Figure 6
https://klear.com

Figure 7
Internal source

Figure 8
https://www.datapine.com/blog/sales-report-kpi-examples-for-daily-reports/

Figure 9 & 10
https://www.datapine.com/blog/call-center-dashboard-reports-and-data-analytics/

Figure 11
https://www.datasciencecentral.com/profiles/blogs/demystifying-data-lake-architecture

Figure 12
https://www.pinterest.com/pin/574701602449146066/?nic_v1=1aImUYUdS64ptdQfZpiYNikiwx9USW7V1xFKY93MNpMAdI08JAOYp%2BKQL8lDocTM6z