\

|  |
| --- |
|  |
| Philip Morris International DWH |



Contents

[1. Business Description 3](#_Toc412572569)

[1.1. Business background 3](#_Toc412572570)

[1.2. Problems because of poor data management 3](#_Toc412572571)

[1.3. Benefits from implementing a Data Warehouse 3](#_Toc412572572)

[2. Dimensions of a Business 3](#_Toc412572573)

[3. Logical Scheme 3](#_Toc412572574)

[4. Data Flow 3](#_Toc412572575)

[5. Fact Table Partitioning Strategy 3](#_Toc412572576)

[6. Strategy of Parallel Load 3](#_Toc412572577)

[7. Report Layouts 3](#_Toc412572578)

# Business Description

## Business background

Philip Morris International Inc. (PMI) (NYSE: PM) is an American global cigarette and tobacco company, with products sold in over 180 countries outside the United States. The most recognized and best selling product of the company is Marlboro.

The company's operational headquarters are in Lausanne, Switzerland, although the corporate headquarters remain in New York. It does not operate in the United States, with Philip Morris brands there still owned by PMI's former owner Altria.

Because tobacco, the main constituent of cigarettes, is the single greatest cause of preventable death globally[3] and is addictive, the company's operations (and its competitors') are highly controversial and are increasingly the subject of litigation and restrictive legislation from governments concerned about the health impacts of its products.

Philip Morris International has six multi-billion US$ brands including:

1. Dji Sam Soe 234 was launched in 1913 and is a brand of Kretek cigarettes. It is the best seller of Kretek cigarettes in Indonesia.

2. L&M was launched by Liggett & Myers in 1953 with the tagline: "American cigarettes of the highest quality with the best filter." L&M variants include full flavor shorts, full flavor 100s, lights, ultra lights, menthol shorts, menthol 100s, menthol light shorts, menthol light 100s, Turkish Blend shorts, Turkish Blend 100s, and L&M Mild Kretek.

3. Longbeach include in Australia and Indonesia in 1999. Longbeach variant include: Longbeach Filter and Longbeach Mild.

4. Marlboro was launched in 1904. Marlboro is the premium brand. Marlboro variants include: Marlboro Special, Marlboro Menthol, Marlboro Lights, Marlboro Lights Menthol, Marlboro Mix-9 Filter Kretek, Marlboro Flavor Plus, Marlboro Black Menthol, and Heatsticks, a heated tobacco product.[10]

5. ST Dupont Paris is the brand cigarette designed by Simon Tissot Dupont in 1902. With the black packaging. ST Dupont Paris variants include: filter, lights, menthol, and menthol lights.

6. U Mild was launched in Indonesia in May 22, 1998 after Indonesian revolution. U Mild is a Mild Kretek cigarette sold in Indonesia.

In the end of 2016 Philip Morris presented the new product called IQOS - smokeless cigarette–one of the latest products from Philip Morris International, manufacturers of the legendary Marlboro cigarettes. The product is a hybrid between the benefits of electronic cigarettes and the design of traditional types – making it perfect for smokers looking to switch to vaping or even for occasional use. Given the known dangers of smoking traditional cigarettes, switching to vaping or just implementing a sporadic alternative is a good option.

The iQOS uses refills of real tobacco, but doesn’t burn it like the conventional cigarette which prevents the ingestion of combustive, charred materials. Rather, the iQOS ecig heats the tobacco just prior to the point of combustion, producing tobacco-flavored vapor.

## Problems because of poor data management

A brand new product (iQOS) appeared in the last year, but due to the poor management, it didn’t get much popularity worldwide. As a result new production items had produced very little profit compared to the forecasts and the amount of resources that was spent on its production. All of this caused financial problems of the company and damaged the economic policy in 2017.

## Benefits from implementing a Data Warehouse

So as to recover the company’s brand name, the board of directors decided to build a detailed plan of the economic situation on the current date and to provide detailed information about all products and its sales amounts along with new company’s production items. The main goal of the research is to get proper information about popularity and economic profitability of new products. Another task is to gather information about “old” production items. The research should be based on prebuild Data Warehouse. Data Warehouse should also be able to aggregate all information about sales for future years till 2030.

# Dimensions of a Business

Starting with identifying 4 steps of creating data warehouse

1. Select the business process

Getting detailed information about brand-new product contribution to the company’s production process

2. Declare the Grain

Total sales and quantity amount of new products compared

3. Identify the Dimensions

* Products;
* Departments;
* Consumers;
* Time;
* Promotion;
* Pay State.

4. Identify the Facts

“Total sales amount” as FACT\_SALES

Entities:

* **Product Category**

Categories of the products

Examples: cigarette, tobacco heating products, vapor products

* **Product**

Brands, produced by PMI company

Examples: Marlboro, Most popular product of the company, global,

* **Location**

Addresses, cities, countries and regions

Example: 10-3 Hyde Park, NY, USA, North America

* **Consumers**

Consumers of the PMI products

Examples: Japan retail agency

* **Price index**

Price index related to appropriate country. Prices of PMI products vary for different countries and each country has its index depending on which, the prices are set.

Example: Russia Price Index 1.3, USA Price Index 1.7

* **Departments**

PMI agency all over the world

Example: St. Petersburg Main Agency, 2201

* **Promotions**

Promotions, which take place in some parts of the year

Example: Promotion #276421 Brand-new product

* **Sale**

Sale of some product to the customer

Example: PRODUCT, PRICE, QUANTITY, CONSUMER, DEPARTMENT, DATE

As a result, we get such **DIMENSIONS:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Dimension** | **Field** | **Description** | **Note** |
| **PRODUCTS** | PRODUCT\_ID | Primary key | autoincrement |
| PRODUCT\_CODE | Natural Key |  |
| PRODUCT\_NAME | Product name |  |
| PRODUCT\_CATEGORY | Category of product |  |
| PRODUCT\_SUBCATEGORY | Type of product | [standard/slim/...] |
| PACKAGE\_SIZE | Size of package |  |
| PACKAGE\_COLOR | Color of package |  |
| PRODUCT\_SIZE | Product\_Item size |  |
| TOBACCO | Amount of tobacco in one product item | '99' if no tobaco |
| ITEMS\_PACKAGE | Amount of items In one package |  |
| PRODUCT\_IMAGE | Image of product |  |
| LOCALIZATION | Localization of product | [global,local] |
| START\_DT | Start date of product production |  |
| END\_DT | End date of product production |  |
| IS\_ACTIVE | Indicator of row values status at the current “storage” day | [0/1] |
| STD\_PRICE | Standard price for product item without price index |  |
| **DEPARTMENTS** | DEPARTMENT\_ID | Primary key | autoincrement |
| DEPARTMENT\_CODE | Natural Key |  |
| DESCRIPTION | Description of department |  |
| ADDRESS | Address |  |
| CITY | City |  |
| COUNTRY | Country |  |
| REGION | Region |  |
| **CONSUMERS** | CONSUMER\_ID | Primary key | autoincrement |
| CUSTOMER\_CODE | Natural Key |  |
| CONSUMER\_TITLE | Name of agency/department/employee |  |
| ADDRESS | Address |  |
| CITY | City |  |
| COUNTRY | Country |  |
| REGION | Region |  |
| POSTAL\_CODE | Postal Code |  |
| PHONE | Consumer phone number |  |
| PRICE\_INDEX | Price index for a consumer | depend on the location |
| **DATE** | FULL\_DATE | Primary Key | Natural key |
| DAY\_WEEK | Number of day in the week |  |
| DAY\_SHORT | Short name of the day | [FRI] |
| DAY\_LONG | Full name of the day | [FRIDAY] |
| DAY\_MONTH | Number of day in the month | [23] |
| DAY\_YEAR | Number of day in the year | [234] |
| MONTH\_YEAR | Number of month in the year | [4] |
| MONTH\_SHORT | Short name of the month | [JANUARY] |
| MONTH\_LONG | Full name of the month | [JAN] |
| QUARTER | - | [1,2,3,4] |
| YEAR\_HALF | - | [1,2] |
| YEAR | - |  |
| **PROMOTION** | PROMO\_ID | Primary key | autoincrement |
| PROMO\_CODE | Natural Key |  |
| PROMO\_NAME | Name of promotion |  |
| DESC | Description of promotion |  |
| START\_DT | Start date of promotion |  |
| END\_DT | End date of promotion |  |
| IS\_ACTIVE | Indicator of row values status at the current “storage” day | [0/1] |
| PROMO\_COST | Cost of promotion |  |
| **PAY\_STATE** | PAYSTAT\_ID | Primary key (Natural Key) | [0/1] |
| DESCRIPTION | [0] - payment is not full [1] - payment is full |  |

**Fact Table:**

|  |  |  |  |
| --- | --- | --- | --- |
| **FACT** | **Field** | **Description** | **Note** |
| **FACT\_SALES** | PRODUCT\_ID | Foreign Key | Composite PK |
| DEPARTMENT\_ID | Foreign Key | Composite PK |
| CONSUMER\_ID | Foreign Key | Composite PK |
| DATE\_ID | Foreign Key | Composite PK |
| PROMO\_ID | Foreign Key | Composite PK |
| PAYSTAT\_ID | Foreign Key | [1,2,3] |
| COST | Amount of sale | [230000] |
| QUANTITY | Quantity of product per sale | [3500] |

# Logical schema

Provided dimensions and fact table build the Star schema of data warehouse. Star schema was chosen for the following reasons:

* Maximum performance and ease of access;

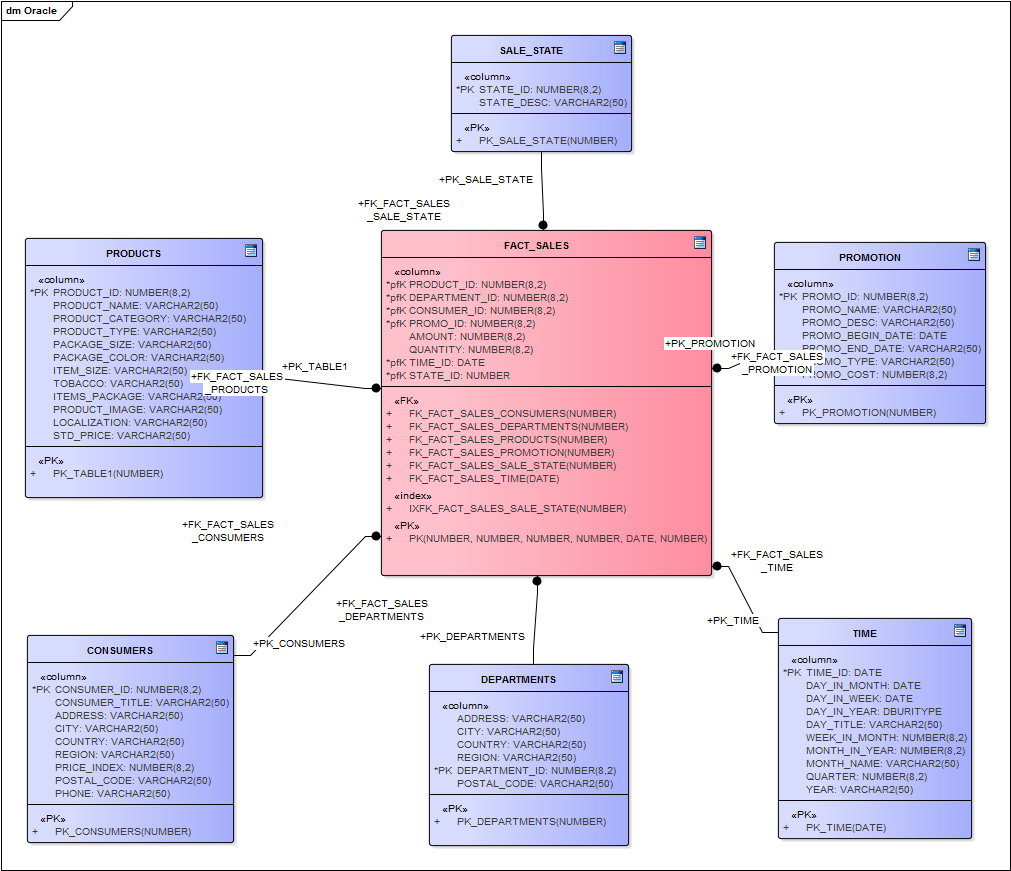
PMI Corporation is ready to invest much in local storage to get rid of redundant connections so as to make access to all data without extra joins and do not pay much attention about redundancy in data.

* Simplification of preparing data for reporting

Using Star makes does not require special IT staff for ETL and building data marts for appropriate report because almost everything will be aggregated in dimensions and the only thing to do will be some simple grouping operations,

As far as dimensions were defined in the previous chapter, PMI data warehouse gets next model:

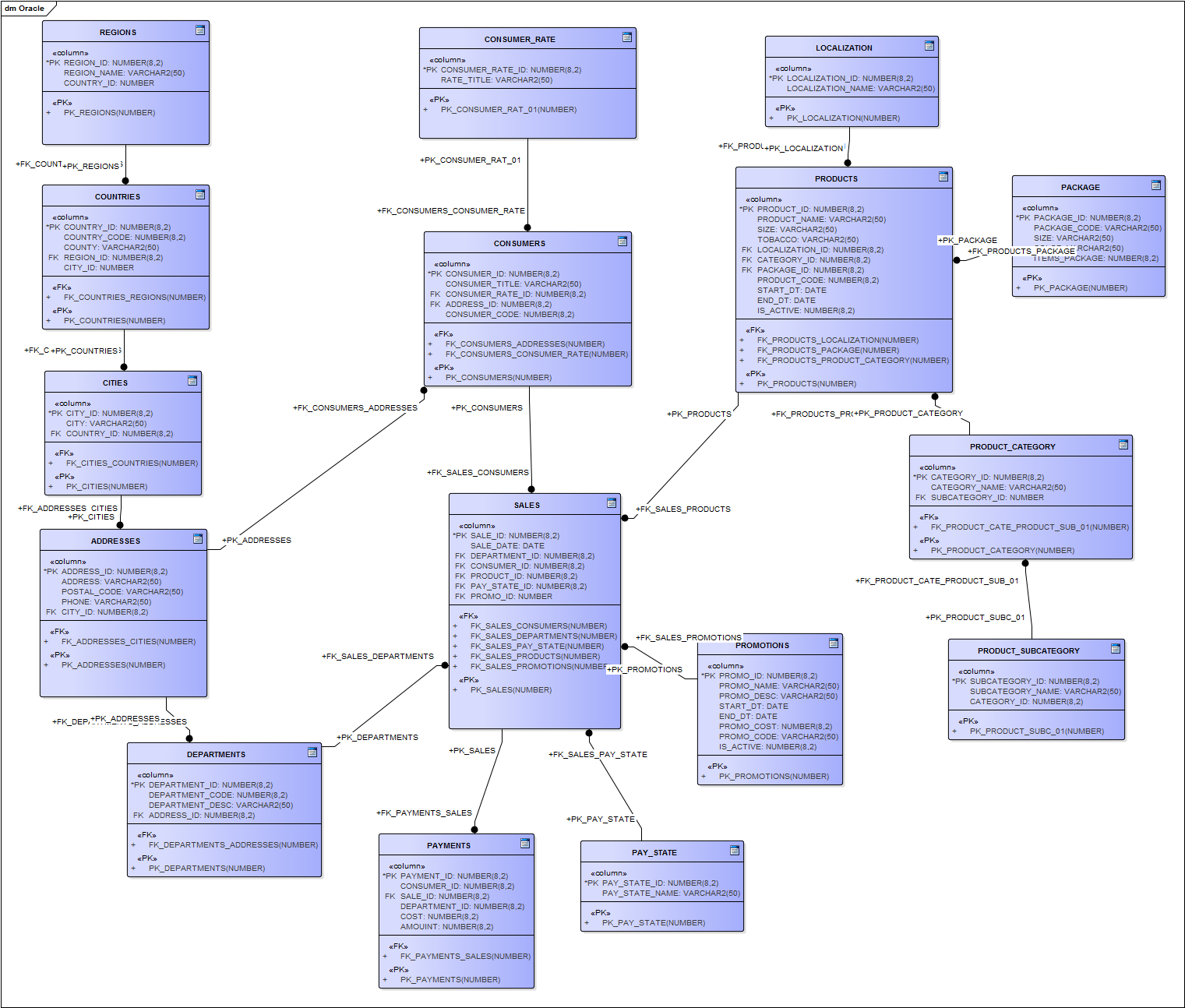
#### Star schema:

****

**Picture – PMI Star schema model**

As far as star schema is built, the model of PMI database in 3rd normal form can be seen below.

#### 3NF model:



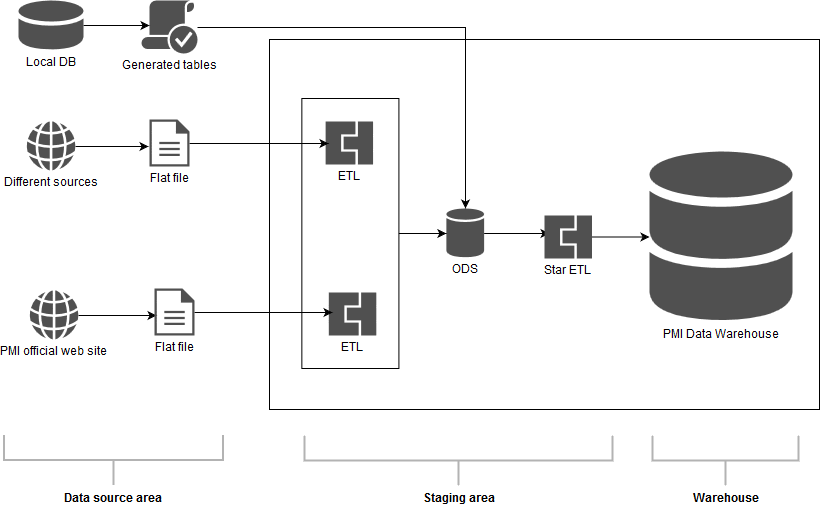
**Picture – PMI 3rd normal form model**

Description of PMI tables can be seen below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Table** | **Field** | **Description** | **Note** |
| **PRODUCT** | PRODUCT\_ID | Primary Key | autoincrement |
| PRODUCT\_CODE | Natural Key |  |
| PRODUCT\_NAME | Name of the Brand item |  |
| SIZE | Size of the item |  |
| START\_DT | Start date of product production |  |
| END\_DT | End date of product production |  |
| IS\_ACTIVE | Indicator of row values status at the current “storage” day |  |
| TOBACCO | Grams of tobacco per one item |  |
| PACKAGE\_ID | Foreign Key |  |
| LOCALIZATION\_ID | Localization of product |  |
| CATEGORY\_ID | Foreign Key |  |
| STD\_PRICE | Standard price for product item without price index |  |
| **PRODUCT CATEGORY** | CATEGORY\_ID | Primary Key | autoincrement |
| CATEGORY\_NAME | Name of the category |  |
| **PRODUCT SUBCATEGORY** | SUBCATEGORY\_ID | Primary Key | autoincrement |
| CATEGORY\_ID | Foreign Key |  |
| SUBCATEGORY\_NAME | Name of the subcategory |  |
| **PACKAGE** | PACKAGE\_ID | Primary Key | autoincrement |
| PACKAGE\_CODE | Natural Key/Code of the package |  |
| SIZE | Size of package |  |
| COLOR | Color of package |  |
| ITEMS\_PACKAGE | Number of items in one package |  |
| **LOCALIZATION** | LOCALIZATION\_ID | Primary Key |  |
| LOCALIZATION\_NAME | Name of localization | [local,global] |
| **DEPARTMENTS** | DEPARTMENT\_ID | Primary Key | autoincrement |
| DEPARTMENT\_CODE | Natural Key |  |
| ADDRESS\_ID | Address of the department |  |
| DEPARTMENT\_DESC | Description of the department |  |
| **CONSUMERS** | CONSUMER\_ID | Primary key | autoincrement |
| CONSUMER\_CODE | Natural Key |  |
| CONSUMER\_TITLE | Name of agency/department/employee |  |
| CONSUMER\_RATE\_ID | Foreign Key |  |
| ADDRESS\_ID | Foreign Key |  |
| **CONSUMERS\_RATE** | CONSUMER\_RATE\_ID | Primary key (Natural key) | autoincrement |
| RATE\_TITLE | Name of the rate |  |
| **ADDRESSES** | ADDRESS\_ID | Primary key | autoincrement |
| ADDRESS | Address |  |
| POSTAL\_CODE | - |  |
| PHONE | - |  |
| CITY\_ID | Foreign Key |  |
| **CITY** | CITY\_ID | Primary Key | autoincrement |
| CITY | Name of the City |  |
| COUNTRY\_ID | Foreign Key |  |
| **COUNTRY** | COUNTRY\_ID | Primary Key | autoincrement |
| COUNTRY\_CODE | Natural Key |  |
| COUNTRY | Name of the country |  |
| REION\_ID | Foreign Key |  |
| **REGION** | REGION\_ID | Primary Key | autoincrement |
| REGION | Name of the region |  |
| **PAYMENTS** | PAYMENT\_ID | Primary Key | autoincrement |
| CONSUMER\_ID | Foreign Key |  |
| SALE\_ID | Foreign Key |  |
| DEPARTMENT\_ID | Foreign Key |  |
| COST | Cost of products in payment |  |
| AMOUNT | Amount of products in payment |  |
| **SALES** | SALE\_ID | Primary Key | autoincrement |
| SALE\_DATE | Foreign Key |  |
| DEPARTMENT\_ID | Foreign Key |  |
| CONSUMER\_ID | Foreign Key |  |
| PRODUCT\_ID | Foreign Key |  |
| PAY\_STATE\_ID | Foreign Key |  |
| **PAY\_STATE** | PAY\_STATE\_ID | Primary Key | autoincrement |
| DESCRIPTION | Name of the payment status |  |
| **PROMOTIONS** | PROMO\_ID | Primary Key | autoincrement |
| PROMO\_CODE | Natural Key |  |
| PROMO\_NAME | Name of the promotion |  |
| PROMO\_DESC | Description of the promotion |  |
| START\_DT | Start date of promotion |  |
| END\_DT | End date of promotion |  |
| IS\_ACTIVE | Indicator of row values status at the current “storage” day |  |
| PROMO\_COST | Cost of the promotion |  |

# Data flow

Here you can see the example of data flow that will be built during the process of building DWH:



**Picture – PMI data flow model**

Data flow has 3 areas:

1. Data source area;
2. Staging area;
3. Warehouse area.

**Data source area**

On the first step main task is to choose appropriate sources with PMI company data and combine it in flat files so as to prepare data for ETL process. **Official website of PMI** Company is used as the main data source for product and department description. Some information about outdate products will be gathered from other **different sources** likeWikipedia, Numbeo.and some others. Some information like sales and date dimension will be generated manually in the **local database** and provided as .sql file that would be used in operational data store (ODS).

**Staging area**

Second step is to make **ETL** processes that will combine operations like:

* Extract;
* Clean;
* Conform;
* Transform
* Load.

External tables will be actively used on this stage. As the result of these operations, tables with cleaned data will be gained. Cleaned data will be sent to the ODS, where database in the 3rd normal form will be created.

Another step will be to transform 3rd normal form database into Star schema through some additional ETL processes.

**Warehouse area**

The result of these processes will be ready-made Data Warehouse for PMI Corporation.

**Methods, rules and the process of generation source data are the following:**

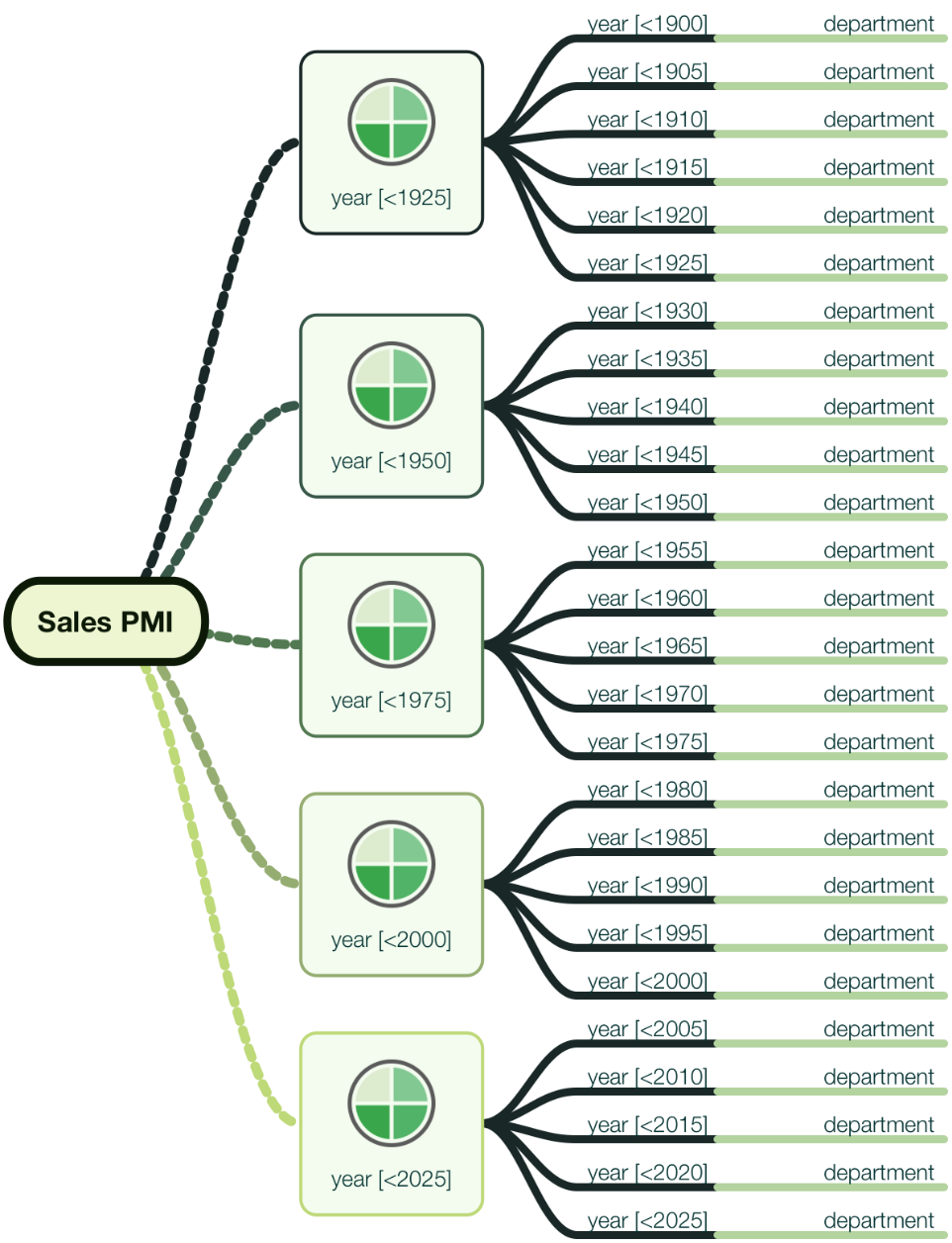
* Product description and whole information will be taken from the PMI website and Wikipedia;
* Information about departments will be taken from additional sources, which provide appropriate information about PMI deps;
* Prices for products will be taken from the source <https://www.numbeo.com/cost-of-living/prices_by_country.jsp?displayCurrency=USD&itemId=17> where there is an information about pricing in different countries. Other product pricing will be generated on the basis of this source so as to provide approximately true information about different brands of PMI;
* Pricing Index will be set up for location on the basis of source above;
* Data on different years will be generated manually;
* Data on amount of sales will be generated manually (within the framework of global brand share);
* Data on quantity of products sold will be generated manually (within the framework of global brand share).

# Fact Table Partitioning Strategy

Data Warehouse is going to have a great amount of records when it starts a production usage. And, moreover, the amount of records will be rising dramatically. It means that performance will be fallen every time when new sales appear and it will have been continuing till the end of 2030 year at least. So PMI Data Warehouse is needed to be correctly partitioned.

Approximately in every 25 years company make some big step in smoking industry and provides market with brand new products. And every 5 year company have a new economic plan that regulates it sales policy in the future years. So, it was decided to build partition on mentioned date (partition on every 25 years and subpartitions on every 5 years) periods and it will be manually added range partitions. Despite it will make a great performance gain it can be not enough to manipulate with such data warehouse in a comfort way. So, additional department has partition can be added so as to divided every 5 years by department sales in approximately equal proportions due to the techniques of hash partition.

As a result PMI data warehouse will have partition structure like it shown on the picture below:



# Strategy of parallel load

# Report layouts

One of the concepts of final report is shown on the screenshot below:



**Picture – concept of final report**

This report shows some main goals that were set on the first step of building DWH. One of the task was to find the difference between sales start of new product compared to the already known production. This question appeared because of the small profit after new production line had started.

This concept report was built in TIBCO Spotfire on the example-basis values and real production items. It includes:

* 2 KPI charts (on the left of the screen) where
  + 1st chart shows total Quantity sold per category on its the first year on market;
  + 2nd chart shows total profit per category on its first year on market;

\* total profit is a calculated column where [quantity] was multiplied on [cost per item]

* Bar Chart (on the right of the screen) where there is a quantity and profit on value axis per products (grouped by categories). There is also a line that shows average quantity sold in the first year of the product on the market.

Such charts provides easy-to-understand access to the success of the products in their market sales starts.