

|  |
| --- |
| Business Template  **Olist Store** |
| Project – Olist – Brazil – DataNeophyte |

Contents

[1 Business Description 3](#_Toc22660618)

[1.1 Business background 3](#_Toc22660619)

[1.2 Problems because of poor data management 3](#_Toc22660620)

[1.3 Benefits from implementing a Data Warehouse 3](#_Toc22660621)

[2 Dimensions of a Business 3](#_Toc22660622)

[3 Logical Scheme 5](#_Toc22660623)

[4 Data Flow 6](#_Toc22660624)

[5 Fact Table Partitioning Strategy 8](#_Toc22660625)

[6 Strategy of Parallel Load 8](#_Toc22660626)

[7 Report Layouts 8](#_Toc22660627)

# 

# Business Description

## Business background

Very important parts of business for companies, starting from small online shops to large retailers, are followings:

• efficiency of marketing actions

• organization of logistics services

Olist Store is the largest online e-commerce store in Brazil, which provides platform for selling, advertising activities, etc.

In term of logistics services, Olist Store provides them through Olist logistics partners.

## Problems because of poor data management

The main problem of poor data management is inability to develop decision making system, it affects on loss of profits and share of the market.

## Benefits from implementing a Data Warehouse

Benefits of implementing a data warehouse are following:

• Centralized storage of major business data, which can affect on decision making

• Possibility to calculate any metrics from stored data in terms of major dimensions.

• Possibility to implement BI tools using data, which are stored into data warehouse

• Execution of ad hoc queries

• Basis for creation any analytical products to increase user experience: recommendation systems, reports of descriptive statistics

# Dimensions of a Business

2.1 Identification of the business process

Business process is selling of products through online store to customers. Generated information of selling covers the following entities: customers, suppliers, logistic partners, products, etc.

## 2.2 Identification of the grain

Grain is fact of sale, which is made by customer. This fact represents as decomposed as possible for further construction of reporting.

## 2.3 Identification of dimensions

The following dimensions are used to describe the grain:

• **DIM\_CUSTOMERS**

o Contains general information about customers

• **DIM\_GEOLOCATIONS**

o Contains detailed geolocational data

• **DIM\_PRODUCTS\_SCD**

o Contains information about selling products

• **DIM\_SUPPLIERS**

o Contains general information about suppliers

• **DIM\_DATES**

o Contains detailed data about dates

• **DIM\_PAYMENT\_TYPES**

o Contains general information about types of payment

• **DIM\_LOGISTIC\_PARTNERS**

o Contains general information about logistic partners

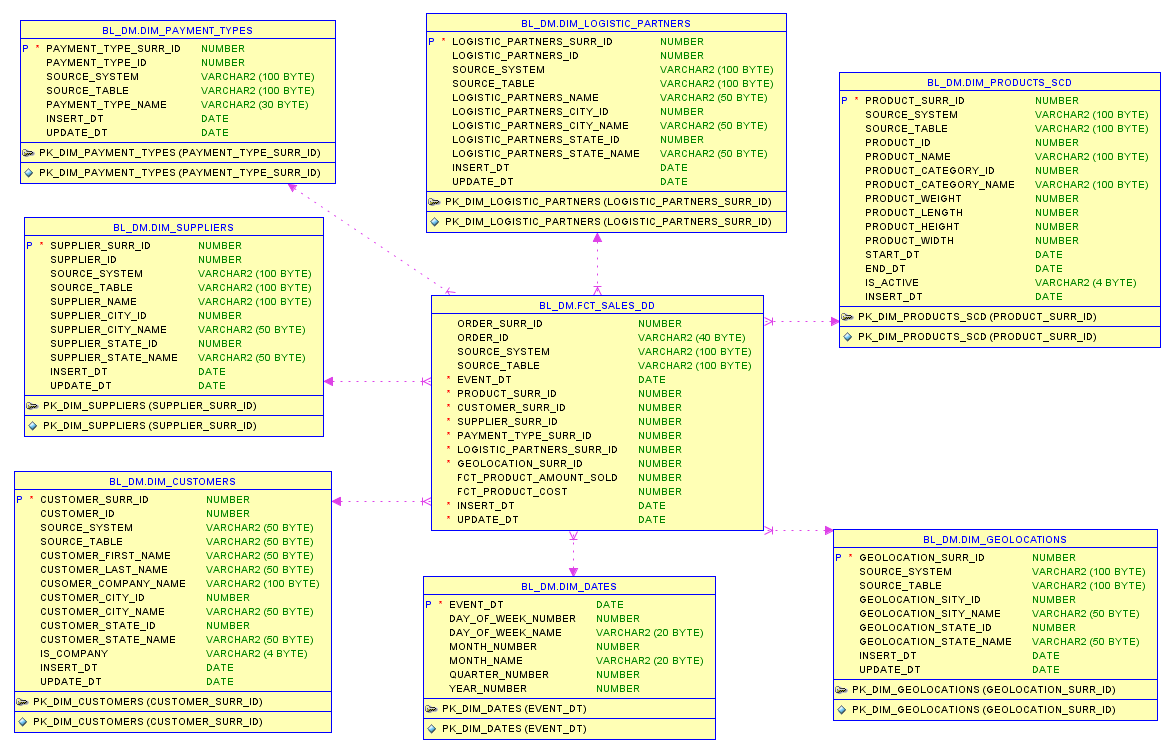
## 2.4 Identification of the fact

Every single row into the fact table is a grain, which is unique. This uniqueness is achieved through combination of dimensions values and degenerate dimension “ORDER\_ID”. Fact table contains the following measurements:

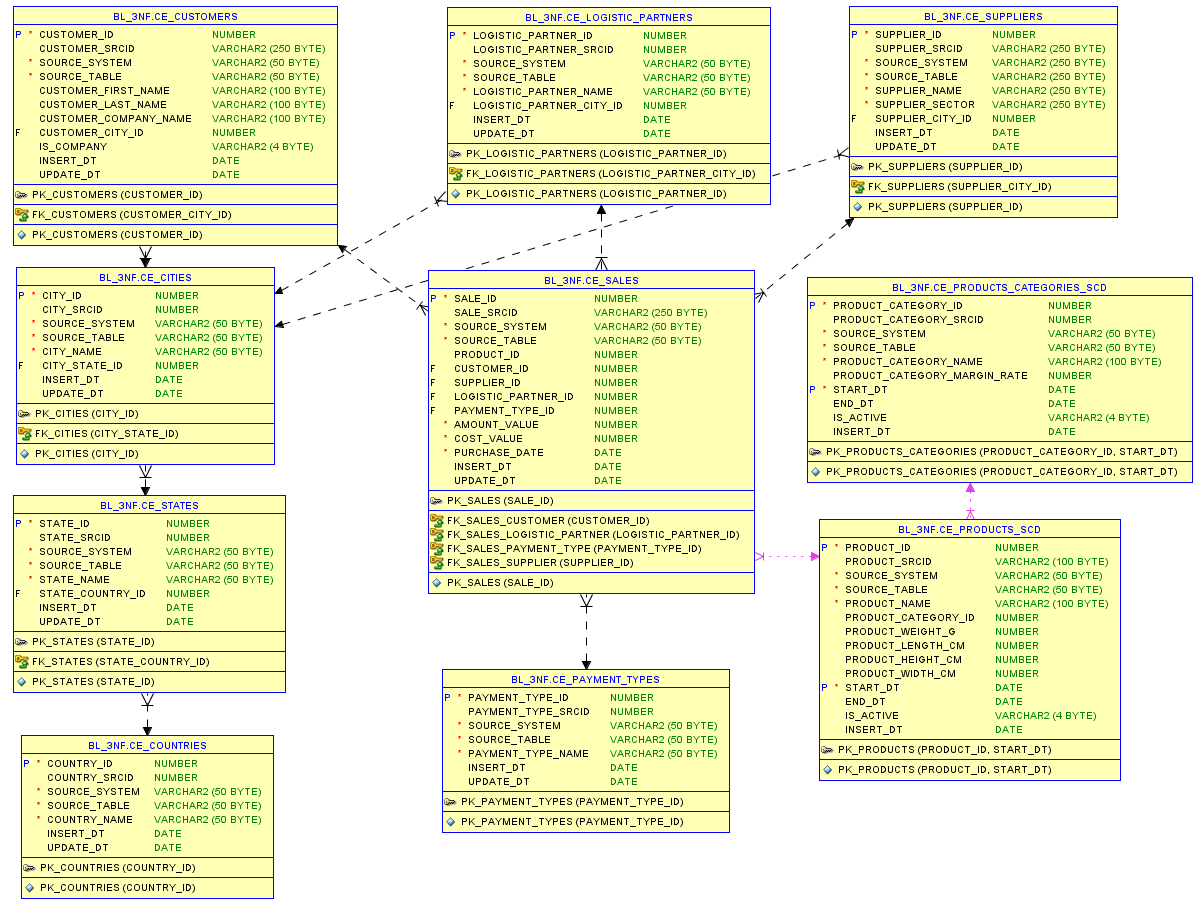
• **FCT\_ PRODUCT\_AMOUNT\_SOLD** - sale amount of product sold

• **FCT\_ PRODUCT\_COST** – cost amount of product sold

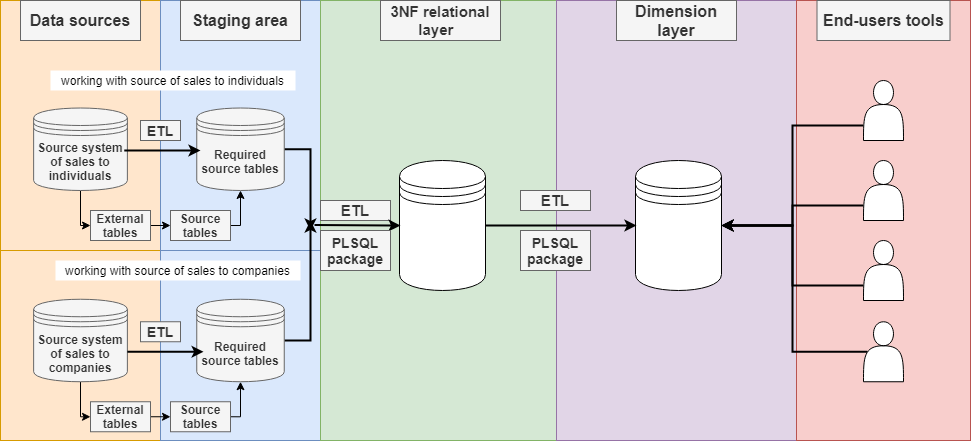
## 2.5 The Dimensional Model Schema



## 2.6 The 3nf Model Schema



# Logical Scheme

****

Logical schema is implemented in the following steps:

1. Creating external tables in separate schemas, using csv files of dataset. Separating on two schemas is required, because data come from two different source systems.

2. Then, at staging area: creating source tables from external tables into the relevant schemas.

3. At 3NF relational layer loaded data from staging area are normalized to 3NF. Rows of dimensions receive surrogate keys, it is generated by source triplet (source\_id, source\_system, source\_table).

4. At Dimension layer data from 3NF relational layer are denormalized to Dimensional model, and rows of created dimensions receive surrogate key, which based on source triplet (source\_id (surrogate key from 3NF relational layer), source\_system, source\_table).

5. Data of Dimension layer is ready for further work.

# Data Flow

Data for project come from two different source systems (retail sales and business sales) by csv files.

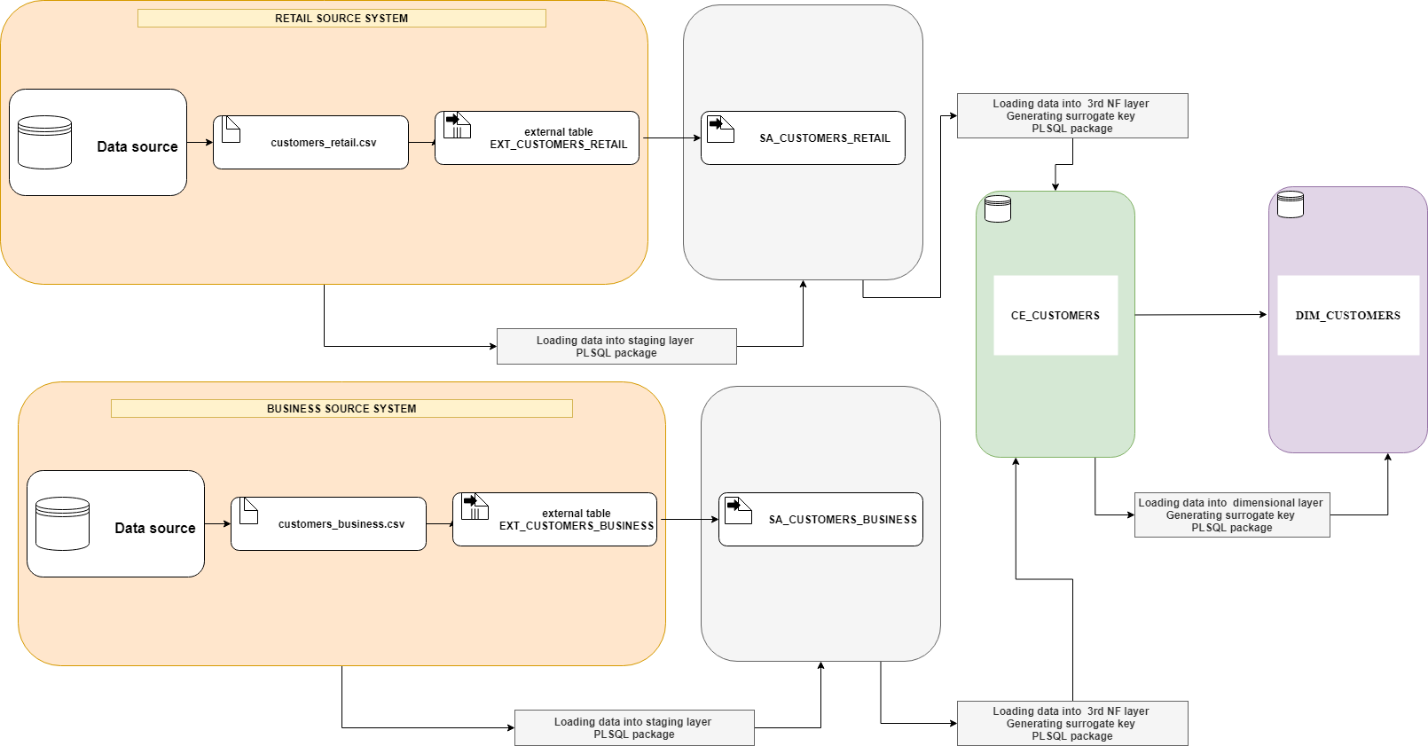
Both of source system contain the following identical entities:

* SALES
* CUSTOMERS

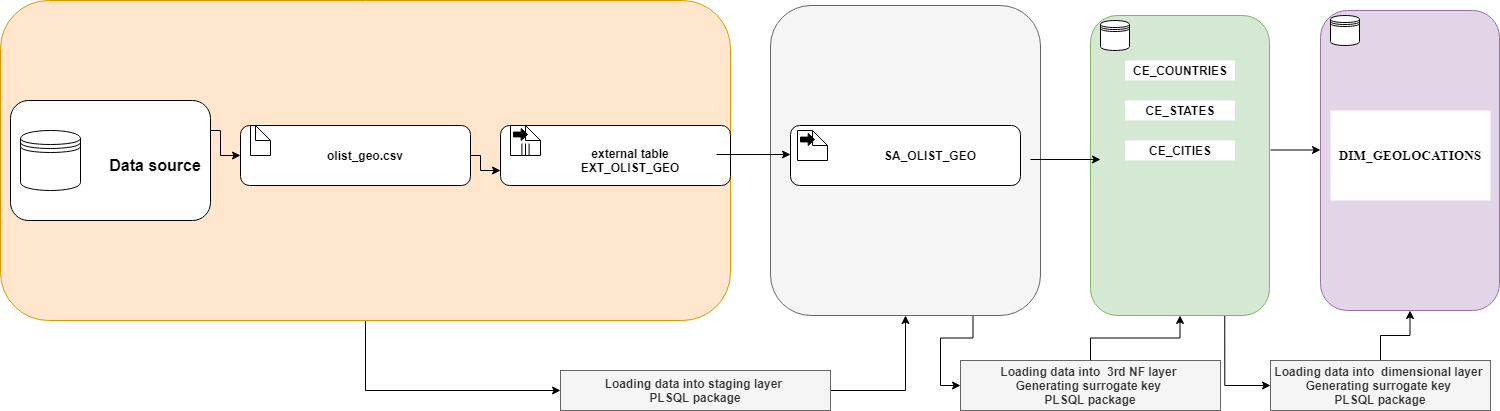
Other’s entities are included only into retail source system.

Description of data flow to BL\_DM schema:

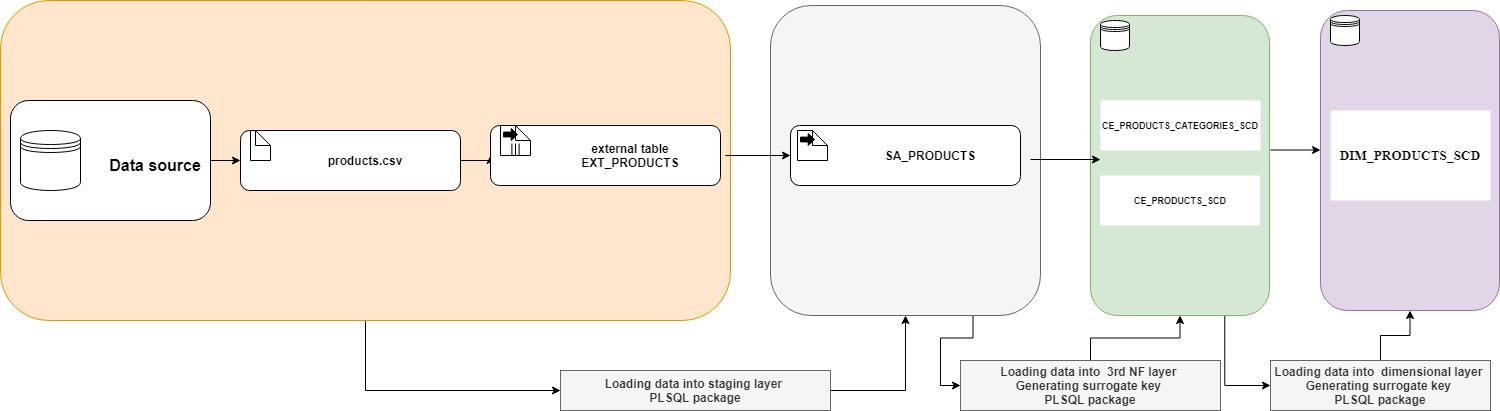
**DIM\_CUSTOMERS:**

****

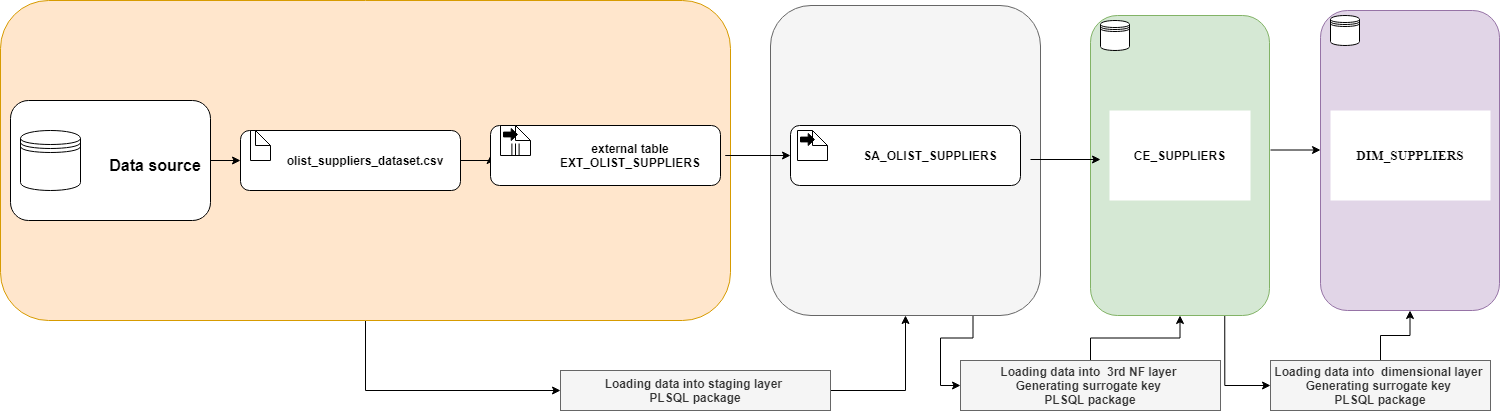
**DIM\_GEOLOCATIONS:**

****

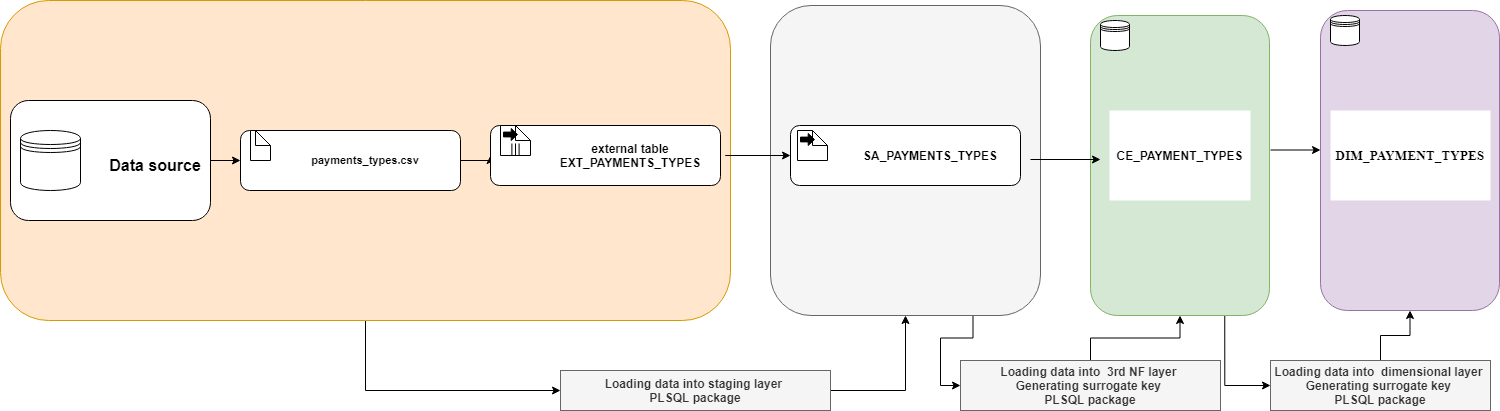
**DIM\_PRODUCTS\_SCD**

****

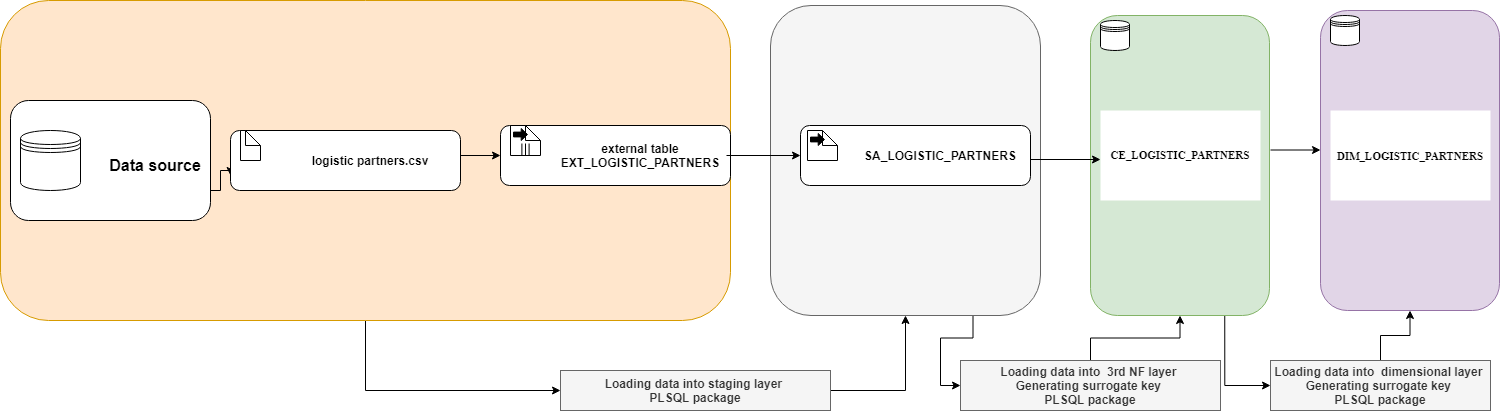
**DIM\_SUPPLIERS**

****

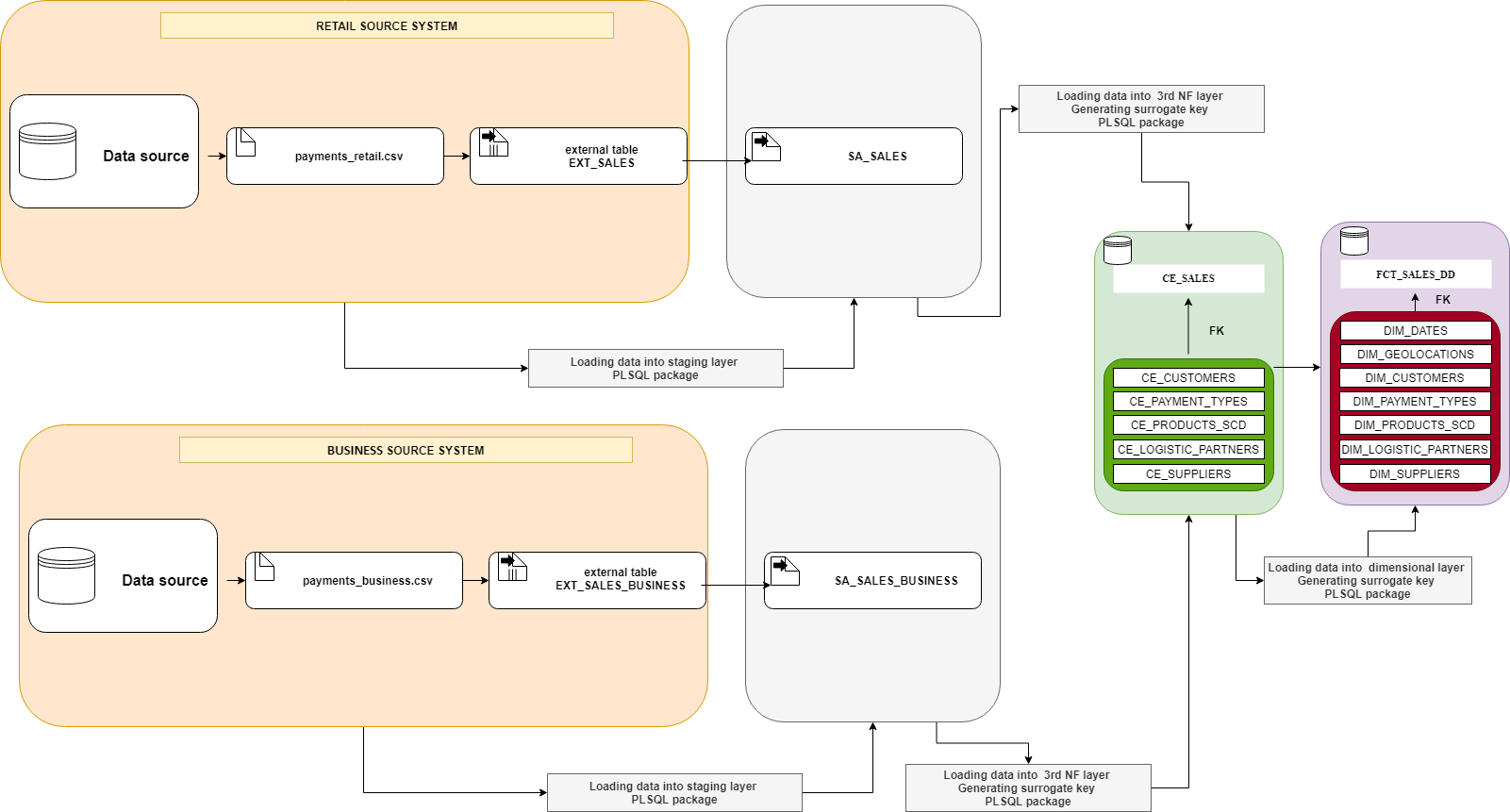
**DIM\_PAYMENT\_TYPES**

****

**DIM\_LOGISTIC\_PARTNERS**

****

**FCT\_SALES\_DAILY**

****

# Fact Table Partitioning Strategy

Using partitions for the fact tables makes it possible to use partition pruning. It means that, if WHERE condition will have some attribute, which is used as key for partition, Oracle will scan only specific partitions, instead of scanning full table.

As key for partitions to the fact tables at BL\_3NF and BL\_DM layers I chose attributes, which contain dates of transactions, because they are high rated for SELECT statement. Column PURCHASE\_DATE has been selected as key of partition for the table CE\_SALES, also INSERT\_DT has been selected as key of partition for the table FCT\_SALES\_DD.

# Strategy of Parallel Load

# Report Layouts