# **Data Modeling documentation - Yalo Case**

#### **Business Overview**

Local retail company focused in beverage sales, localized in lowa-US, which wants to use data analytics to understand financial performance and get insights for next actions, for both finance and sales department (main stakeholders)

#### **Tech Stack**

#### **Data Source**

Considering a paid service that is an API Endpoint with daily information about all liquor sales in lowa.

#### **Tools**

Considering a rather small company, the data stack will follow a full open source scenario

- Linux Server
- Python
- PostgreSQL
- dbt
- Airflow
- Power BI / Looker Studio

#### **Ingestion Phase**

• Python Script + Airflow

A python script that will use a GET request into the API to gather the **previous day** data, every morning at 8AM.

To avoid risks of losing data or any eventual issue in the script, the code will check the last updated date on the PostgreSQL bronze layer and will get any day that is not present. So, in case of a failed run, it may get the missing days as well Everything orchestrated by airflow.

At the end of the code, the python script will return a .parquet file for each day in a folder, that would be the raw layer

## **Data Modeling**

### **Bronze Layer**

The bronze layer will have a single table called *t\_bronze\_iowa\_sales*, that would be inserted by a Python script in PostgreSQL database db bronze iowa.

The ingestion type would be **incremental**. Only adding new days into the *t\_bronze\_iowa\_sales* .

At the beginning of dbt model run, it will perform quality tests into this table to check

- Duplicity
- NULL values

Stopping the pipeline in case of any of these issues appears, sending a mail/message to the team to verify what happened.

This table will only be accessed by Data & Analytics engineering team.

t_bronze_iowa_sales	
invoice_and_item_number	TEXT
date	DATE
store_number	TEXT
store_name	TEXT
address	TEXT
city	TEXT
zip_code	TEXT
store_location	TEXT
county_number	TEXT
county	TEXT
category	TEXT
category_name	TEXT
vendor_name	TEXT
vendor_number	TEXT
item_number	TEXT
item_description	TEXT
pack	INTEGER
bottle_volume_ml	INTEGER
state_bottle_cost	NUMERIC(10,2)
state_bottle_retail	NUMERIC(10,2)
bottles_sold	INTEGER
sale_dollars	NUMERIC(12,2)
volume_sold_liters	NUMERIC(12,2)
volume_sold_gallons	NUMERIC(12,2)

#### Silver Layer

Here we will apply the Star Schema methodology to model this data, based on the business needs:

fact Table: t\_fact\_sales

- A materialized table with informations about sales, in a daily granularity

#### dimension tables:

nm\_category\_name

TEXT

### separated by domain

dates: t\_dim\_dates

- with information about weekday, year, quarter, etc

stores: t\_dim\_stores

- detailed view of stores and its locations

products: t\_dim\_products

- details such as pack size, volume per bottle, etc

vendors: t\_dim\_vendorsvendor names

t\_dim\_dates t\_dim\_stores DATE INTEGER id\_store\_number (PK) nm\_store\_name dt\_day ds\_store\_numbe TEXT dt\_month\_name TEXT city dt ouarter INTEGER zip\_code TEXT dt\_year\_quarter county\_number TEXT TEXT county TEXT TEXT store\_location TEXT dt\_weekday t\_fact\_sales id\_invoice\_and\_item\_number (PK) TEXT dt\_date (FK) DATE id\_store\_number (FK) id\_item\_number (FK)
id\_vendor\_number (FK) TEXT nr\_bottles\_sold INTEGER nr\_state\_bottle\_cost nr\_state\_bottle\_retail NUMERIC(10,2) NUMERIC(10,2) nr\_sale\_dollars nr\_Volume\_sold\_liters NUMERIC(10,2) NUMERIC(10,2) nr\_volume\_sold\_gallons nr\_net\_raVenue\_per\_bottle NUMERIC(10,2) NUMERIC(10,2) nr\_total\_net\_revenue NUMERIC(10.2) t\_dim\_product t\_dim\_vendor id\_item\_number (PK) (PK) did\_vendor\_number id\_item\_detail nr\_bottles\_pack TEXT nm\_Vendor\_name mr\_bottle\_volume INTEGER

Everything would be orchestrated by dbt + Airflow, with a daily model run. At the end of each run, there will be tests as well.

#### **Gold Layer**

Here the modeling is focused in 2 approaches

#### ABT (Analytical Base Table)

A single source of truth for BI teams and Data Scientists, with tested and clean data to link to a BI tool (Looker Studio or Power BI). And for Data Scientists, allowing them to prepare Machine Learning Models (such as Time Series forecasting). In case of a MLOps team, this would be the beginning of their workflow.

#### Performance Tables (Materialized Views)

Aggregated materialized views, which could provide status of sales per vendor, product, store and county, allowing business teams to have fast answers and also use it on BI tools, as shown in the chart below

#### Analytical Base Table ABT

t_sales_transactions	
id_i_voice_and_iten_number	TEXT
#t_date	DATE
#t_#ay	TEXT
dt_nonth	TEXT
dt_quarter	TEXT
dt_year	DATE
dt_year_quarter	TEXT
dt_year_nonth	TEXT
dt_weekday	INTEGER
mB_87018_mm	TEXT
nm_store_name	TEXT
mm_mddress	TEXT
ds_city	TEXT
ds_state	TEXT
ds_regior	TEXT
nm_zip_enda	TEXT
nn_county	TEXT
nn_item_rumber	TEXT
ds_item_description	TEXT
mr_bottles_pack	INTEGER
nr_rn_bottle_volume	INTEGER
nn_category	TEXT
nn_category_name	TEXT
nn_vendor_nunber	TEXT
nn_vendor_name	TEXT
mr_bottles_sold	INTEGER
mr_state_bottle_cost	NUMERIC(18,2)
mr_state_bottle_retail	NUMERIC(18,2)
m:_Sale_dollarS	NUMERIC(10,2)
mr_ret_revanue_per_Bottle	NUHERIC(10,2)
mr_total_net_revenue	NUMERIC(10,2)
mr_wolume_sold_liters	NUMERIC(10,2)
mr_wolume_sold_galloms	NUMERIC(10,2)

# Aggregate Tables -

ny_store_smles_performance		
dt_date	DATE	
nm_store_number	TEXT	
mr_total_bottles_sold	INTEGER	
nr_total_sale_dollars	NUMERIC(18,2)	
er_totsl_net_revenue	NUMERIC(18,2)	
mr_total_yolwne_sold_liters	NUMERIC(18,2)	
mr_total_volume_sold_gallons	NUMERIC(18,2)	
nv_product_sales_performance		
ft_fate	DATE	
nm_item_rumber	TEXT	
nm_item_fescription	TEXT	
nn_category	TEXT	
nn_cotegory_name	TEXT	
mr_total_Bottles_sold	INTEGER	
nr_total_sale_dollars	NUMERIC(18,2)	
mr_total_nst_revenue	NUMERIC(18,2)	
nr_total_volume_sold_liters	NUMERIC(18,2) NUMERIC(18,2)	
nr_total_volume_sold_liters nr_total_volume_sold_pallons	NUMERIC(18,2) NUMERIC(18,2) NUMERIC(18,2)	
nr_total_volume_sold_liters nr_total_volume_sold_pallons	NUMERIC(18,2) NUMERIC(18,2)	
nr_total_volume_sold_liters nr_total_volume_sold_pallons nr_uvwretail_price nr_uvwretail_price nr_uvw_retail_price nr_vendor_soles_performance nt_nate nn_vendor_number nn_vendor_name	NUMERIC(18, 2) NUMERIC(18, 2) NUMERIC(18, 2) NUMERIC(18, 2) NUMERIC(18, 2)	
nr_total_volume_sold_liters nr_total_volume_sold_gallons nr_avwretadl_price  nw_wendor_soles_performance nt_nate nm_vendor_number nm_vendor_number nm_vendor_number nr_total_sole_dallers	NUMERIC(18, 2) NUMERIC(18, 2) NUMERIC(18, 2) NUMERIC(18, 2) NUMERIC(18, 2) NUMERIC(18, 2)	
nr_total_volume_sold_liters nr_total_volume_sold_pallons nr_uvwretail_price nr_uvwretail_price nr_uvw_retail_price nr_vendor_soles_performance nt_nate nn_vendor_number nn_vendor_name	NUMERIC(18, 2) NUMERIC(18, 2) NUMERIC(18, 2) NUMERIC(18, 2) NUMERIC(18, 2)	
nr_total_volume_sold_liters nr_total_volume_sold_gallons nr_avwww_retadl_price  nw_vendor_soles_performance At_Aste nw_vendor_number nw_vendor_number nr_total_sole_dollars nr_total_nat_revenue nr_total_nat_revenue nr_total_bottles_sols	NUMERIC(18, 2) NUMERIC(18, 2) NUMERIC(18, 2) NUMERIC(18, 2)  DATE TEXT TEXT TEXT TEXT TEXT TEXT NUMERIC(18, 2) NUMERIC(18, 2)	
nr_total_volane_sold_liters nr_total_volane_sold_gallons nr_avg_retail_price  nv_vendor_soles_performance nt_nate nn_vendor_number nn_vendor_number nn_vendor_number nr_total_sold_dallars nr_total_nt_revenue nr_total_bottles_sold  nv_county_sales_performance	NUMERIC(18, 2) NUMERIC(18, 2) NUMERIC(18, 2) NUMERIC(18, 2)  DATE TEXT TEXT TEXT NUMERIC(18, 2) NUMERIC(18, 2) NUMERIC(18, 2) NUMERIC(18, 2)	
nr_total_volume_sold_liters nr_total_volume_sold_gallons nr_wvw_retail_price  nv_vendor_soles_performance st_nate nn_vendor_sumbor nn_vendor_sumbor nn_vendor_sumbor nn_vendor_sumbor nn_total_sale_dallars nr_total_sale_dallars nr_total_sale_dallars nr_total_sale_dallars nr_total_sale_dallars nr_total_sale_dallars nr_total_sale_dallars nr_total_sale_dallars	MUMERIC(18, 2) NUMERIC(18, 2)	
nr_total_volume_sold_liters nr_total_volume_sold_gallons nr_avw_retail_price  nv_vendor_soles_performance at_nata nn_vendor_number nn_vendor_number nr_total_sale_dollars nr_total_sale_sold	MUMERIC(18, 2) NUMERIC(18, 2)	
nr_total_volume_sold_liters  nr_total_volume_sold_gallons  nr_avwretail_price  nv_vendor_soles_performance  nt_nate  nn_vendor_number  nn_vendor_number  nn_vendor_number  nn_vendor_number  nn_vendor_number  nn_total_sold_dallars  nr_total_nat_revenue  nr_total_bottles_sold  nv_county_sales_performance  nt_cote  nn_county_number  nn_county_number	NUMERIC(18, 2) INTEGER	
nr_total_volume_sold_liters  nr_total_volume_sold_pallons  nr_uva_retail_price  nw_vendor_soles_performance  ft_fate nn_vendor_name nn_vendor_name nr_total_sald_dollars  nr_total_sald_dollars  nr_total_sald_dollars  nr_total_bottles_sold  nw_county_sales_performance  ft_cote nn_county_number  nr_total_bottles_sold	MUMERIC(18, 2) NUMERIC(18, 2) INTEGER	
nr_total_volume_sold_liters nr_total_volume_sold_gallons nr_ava_retail_price  nv_wendsr_soles_performance nt_nate nn_vendor_number nn_vendor_number nr_total_sale_dollars nr_total_sale_dollars nr_total_bottles_sold  nr_total_bottles_sold nr_total_bottles_sold nr_total_bottles_sold nr_total_bottles_sold	MUMERIC(18, 2) NUMERIC(18, 2) INTEGER	
nr_total_volume_sold_liters  nr_total_volume_sold_pallons  nr_uva_retail_price  nw_vendor_soles_performance  ft_fate nn_vendor_name nn_vendor_name nr_total_sald_dollars  nr_total_sald_dollars  nr_total_sald_dollars  nr_total_bottles_sold  nw_county_sales_performance  ft_cote nn_county_number  nr_total_bottles_sold	MUMERIC(18, 2) NUMERIC(18, 2) INTEGER	

At the end of silver layer processing, there will be several tests before creating those tables and materialized views, with actions in case of trigger

- Not null Stop flow
- Duplicates Stop flow
- Sales standard deviation Warning
- Outliers Warning

## **Overall Architecture view:**

