VNGRS

DWH Specialist - Challenge

BACKGROUND

The marketing department wants to publish a new campaign to increase the company's revenue. Before that, they wanted to know more about their customers and the current situation. The marketing department and some other departments want a data model that can easily be understood by them and can effectively generate the analyses/reports they aim to see. The metadata of the datasets you have is as follows:

Datasets:

1. Customers Data
   - *customer_id:* the key to the orders dataset. Each order has a unique customer_id
   - *customer_unique_id:* unique identifier of a customer
   - *customer_zip_code:* first five digits of customer zip code
   - *customer_city:* customer city name
   - *customer_state:* customer state
2. Orders Data
   - *order_id:* the unique identifier of the order
   - *customer_id:* the key to the customer dataset. Each order has a unique customer_id
   - *order_status:* reference to the order status (delivered, shipped, etc)
   - *order_purchase_timestamp:* shows the purchase timestamp
   - *order_approved_at:* shows the payment approval timestamp
   - *order_delivered_carrier_date:* shows the order posting timestamp
   - *order_delivered_customer_date:* shows the actual order delivery date to the customer.
   - *order_estimated_delivery_date:* shows the estimated delivery date that was informed to the customer at the purchase moment
3. Payments Data
   - *order_id:* the unique identifier of an order
   - *payment_sequential:* a customer may pay an order with more than one payment method. If he does so, a sequence will be created to
   - *payment_type:* method of payment chosen by the customer
   - *payment_installments:* number of installments chosen by the customer

   - *payment_value:* transaction value
4. Order Items Data
   - *order_id:* order unique identifier
   - *order_item_id:* sequential number identifying the number of items included in the same order
   - *product_id:* product unique identifier
   - *seller_id:* seller unique identifier
   - *shipping_limit_date:* shows the seller shipping limit date for handling the order over to the logistic partner
   - *price:* the item price
   - *freight_value:* item freight value item (if an order has more than one item the freight value is split between items)
5. Products Data
   - *product_id:* unique product identifier
   - *product_category_name:* root category of product, in Portuguese
   - *product_name_lenght:* number of characters extracted from the product name
   - *product_description_lenght:* number of characters extracted from the product description
   - *product_photos_qty:* number of product published photos
   - *product_weight_g:* product weight measured in grams
   - *product_length_cm:* product length measured in centimeters
   - *product_height_cm:* product height measured in centimeters
   - *product_width_cm:* product width measured in centimeters
6. Product Category Name Data
   - *product_category_name:* category name in Portuguese
   - *product_category_name_english:* category name in English

OBJECTIVES

Task

- Draw a Suitable Data Model
- Develop the Data Model (SQL, ETL tool, or a Programming Language)
- Calculate Reporting Metrics (SQL)

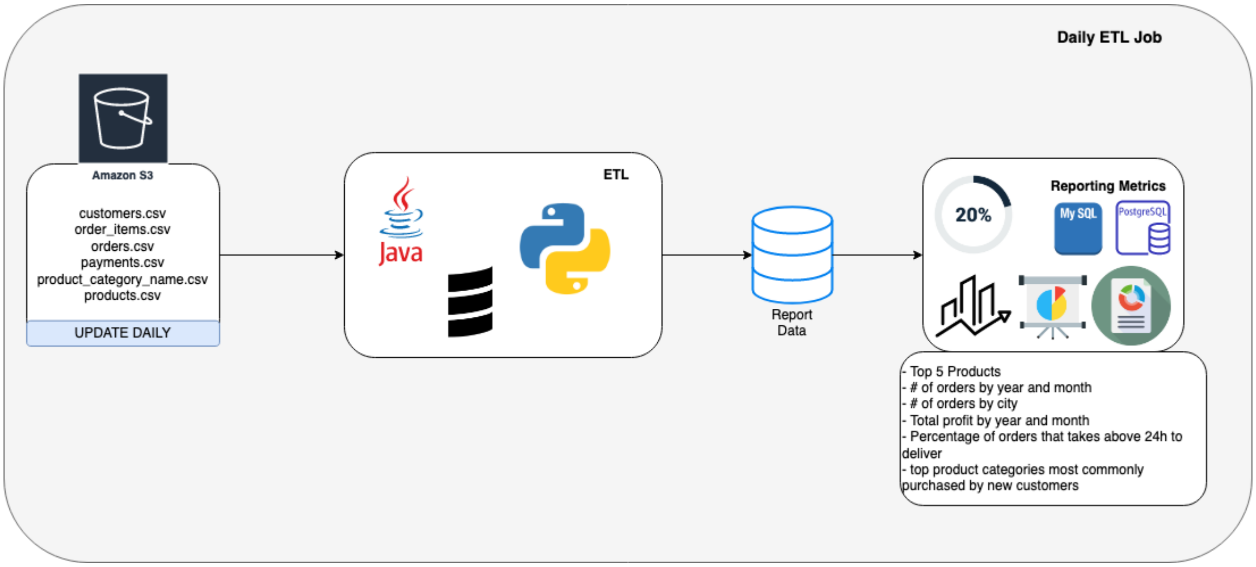
You have been tasked with creating an appropriate data model that represents the data in a simple way and can produce the necessary metrics easily. And after that the second task is creating the reports for expected metrics.

The metrics that will be calculated are:

- Top 5 products according to number of sales
- Number of orders by year and month
- Number of orders by city
- Total profit by year and month
- Percentage of orders that takes above 24h to deliver
- Top 5 product categories most commonly purchased by new customers*

  ** new customer: a customer that gave an order for the first time*

Please create reporting data(modeled data) and calculate necessary metrics.


Daily ETL Job
Amazon S3
customers.csv
order_items.csv
orders.csv
payments.csv
product_category_name.csv
products.csv
UPDATE DAILY
ETL
Java
Report Data
Reporting Metrics
20%
My SQL
PostgreSQL
- Top 5 Products
- # of orders by year and month
- # of orders by city
- Total profit by year and month
- Percentage of orders that takes above 24h to deliver
- top product categories most commonly purchased by new customers

Tools/Language and Key Points

You can use SQL, any programming language (e.g Python) or an ETL tool of your choice to solve ETL tasks.
Reporting Metrics must be done with **SQL.**

The key point is that all code you submit must be **executable, readable and scalable**. The datasets sent to you are sample datasets, and it should not be forgotten that much larger data can be worked with in real life.

EVALUATION

Time management is left to the candidates, as they set their own deadline. Note that submitting work on the deadline agreed upon is still important. They should submit the whole project, which in turn will be compiled and executed.

All steps required to compile and execute the code should be documented, in whatever reasonable way the candidates see fit. In the absence of such a guide and / or if the code cannot be compiled and run trivially, it may be evaluated just for code quality.

SUBMISSION

Please submit these files below

- ETL Script (if it is written by a language) or ETL tool package along with the code
- Reporting Scripts
- Reporting files:
  - top_5_products.csv
  - number_of_orders_year_month.csv
  - number_of_orders_city.csv
  - total_profit.csv
  - deliver.csv
  - New_customers_product_categories.csv
- Modeled data can be written under the same bucket that we provide datasets or a database that you've chosen. If you choose to dump the data model's data to S3, you can provide the tables as files in CSV format. Reporting metrics should be dumped to S3 in any case.
  For modeled data;
  "{S3_BUCKET}/report_data/YYYY-MM-DD/file_name.csv"
  For reporting metrics;
  "{S3_BUCKET}/metrics/YYYY-MM-DD/{METRIC_NAME}.csv"
  Please note that we won't provide you any access keys to write results to the bucket, you can test your code on any other bucket/account that you already have permission to write.

If you have additional files you want to share you can send them too.

PROVIDED MATERIALS

- Data Path: s3://vngrs-recruitment/bi-specialist/case/etl-case/data/
- Alternative Data Path: https://drive.google.com/drive/folders/1nU3WCIkPnJIT70v5FecBRZ5OYetMvqpU?usp=sharing

Please, do not hesitate to contact us if you require further information.