**Assessment**

**1.**An Indian MNC has taken over a small software company with around 15 employees. The software company has provided the csv files of Employees details and their organisation structure.

Task is to upload the data into Bigquery data warehouse and transform the data according to the requirements of the data analysts

**Departments**

department\_id

department\_name

manager\_id

location\_id

**Employees**

Employee\_id

first\_name

last\_name

hire\_date

salary

manager\_id

department\_id

**Job\_history**

Employee\_id

Start\_date

end\_date

department\_id

**Locations**

Location\_id

Postal\_code

City

country\_id

**Task:**

**Create a project in GCP.**

**Create dataset in the bigquery**

**1) employees\_source(datasource)**

Create the below tables in employees\_data\_source using the csv files.

Ex: bigquery\_project\_id  
 🡪 employees\_datasource

→ employees

→ departments

→ locations

→ job\_history

→ Create required stg\_models in the dbt\_transformation schema.

* Final dbt model should be a table with following columns

**Employee\_id, Employee\_name(**full\_name)**, salary\_in\_rupees, department\_name, number\_of\_years\_worked, city**

* Create a view to get the manager\_name and employees in his department.

**Manager\_name, department\_name, employee\_name**

* Find the employees with 2nd Highest salary in each department.

**Employee\_id, Employee\_name(**full\_name)**, salary\_in\_rupees, department\_name**

* Data analysts require the data of employees who have resigned from various cities with their last withdrawn salary. Create a view with following columns.

**Employee\_id, Employee\_name(**full\_name)**, salary\_in\_rupees, department\_name,last\_working\_date, city**

* For the above models write the unit tests/singular /generic tests that are required.
* prepare documentation with descriptions for all the columns and generate the lineage graph from the sources
* Follow dbt best practices in creating the folder structure and indentation in the sql code.
* Create a deploy job and create a new dataset/schema in the Bigquery for production.

**Live\_data\_(any name u want).**

Deploy all the models created to this particular schema.

------------------------------------------------------------------------------------------------------------------

**2. Context:**

Target is a globally renowned brand and a prominent retailer in the United States. Target makes itself a preferred shopping destination by offering outstanding value, inspiration, innovation and an exceptional guest experience that no other retailer can deliver.

This particular business case focuses on the operations of Target in Brazil and provides insightful information about 100,000 orders placed between 2016 and 2018. The dataset offers a comprehensive view of various dimensions including the order status, price, payment and freight performance, customer location, product attributes, and customer reviews.

By analyzing this extensive dataset, it becomes possible to gain valuable insights into Target's operations in Brazil. The information can shed light on various aspects of the business, such as order processing, pricing strategies, payment and shipping efficiency, customer demographics, product characteristics, and customer satisfaction levels.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

The data is available in 8 csv files:

1. customers.csv
2. sellers.csv
3. order\_items.csv
4. geolocation.csv
5. payments.csv
6. reviews.csv
7. orders.csv
8. products.csv

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The column description for these csv files is given below.

The **customers.csv** contain following features:

|  |  |
| --- | --- |
| **Features** | **Description** |
| customer\_id | ID of the consumer who made the purchase |
| customer\_unique\_id | Unique ID of the consumer |
| customer\_zip\_code\_prefix | Zip Code of consumer’s location |
| customer\_city | Name of the City from where order is made |
| customer\_state | State Code from where order is made (Eg. são paulo - SP) |

The **sellers.csv** contains following features:

|  |  |
| --- | --- |
| **Features** | **Description** |
| seller\_id | Unique ID of the seller registered |
| seller\_zip\_code\_prefix | Zip Code of the seller’s location |
| seller\_city | Name of the City of the seller |
| seller\_state | State Code (Eg. são paulo - SP) |

The **order\_items.csv** contain following features:

|  |  |
| --- | --- |
| **Features** | **Description** |
| order\_id | A Unique ID of order made by the consumers |
| order\_item\_id | A Unique ID given to each item ordered in the order |
| product\_id | A Unique ID given to each product available on the site |
| seller\_id | Unique ID of the seller registered in Target |
| shipping\_limit\_date | The date before which the ordered product must be shipped |
| price | Actual price of the products ordered |
| freight\_value | Price rate at which a product is delivered from one point to another |

The **geolocations.csv** contain following features:

|  |  |
| --- | --- |
| **Features** | **Description** |
| geolocation\_zip\_code\_prefix | First 5 digits of Zip Code |
| geolocation\_lat | Latitude |
| geolocation\_lng | Longitude |
| geolocation\_city | City |
| geolocation\_state | State |

The **payments.csv** contain following features:

|  |  |
| --- | --- |
| **Features** | **Description** |
| order\_id | A Unique ID of order made by the consumers |
| payment\_sequential | Sequences of the payments made in case of EMI |
| payment\_type | Mode of payment used (Eg. Credit Card) |
| payment\_installments | Number of installments in case of EMI purchase |
| payment\_value | Total amount paid for the purchase order |

The **orders.csv** contain following features:

|  |  |
| --- | --- |
| **Features** | **Description** |
| order\_id | A Unique ID of order made by the consumers |
| customer\_id | ID of the consumer who made the purchase |
| order\_status | Status of the order made i.e. delivered, shipped, etc. |
| order\_purchase\_timestamp | Timestamp of the purchase |
| order\_delivered\_carrier\_date | Delivery date at which carrier made the delivery |
| order\_delivered\_customer\_date | Date at which customer got the product |
| order\_estimated\_delivery\_date | Estimated delivery date of the products |

The **reviews.csv** contain following features:

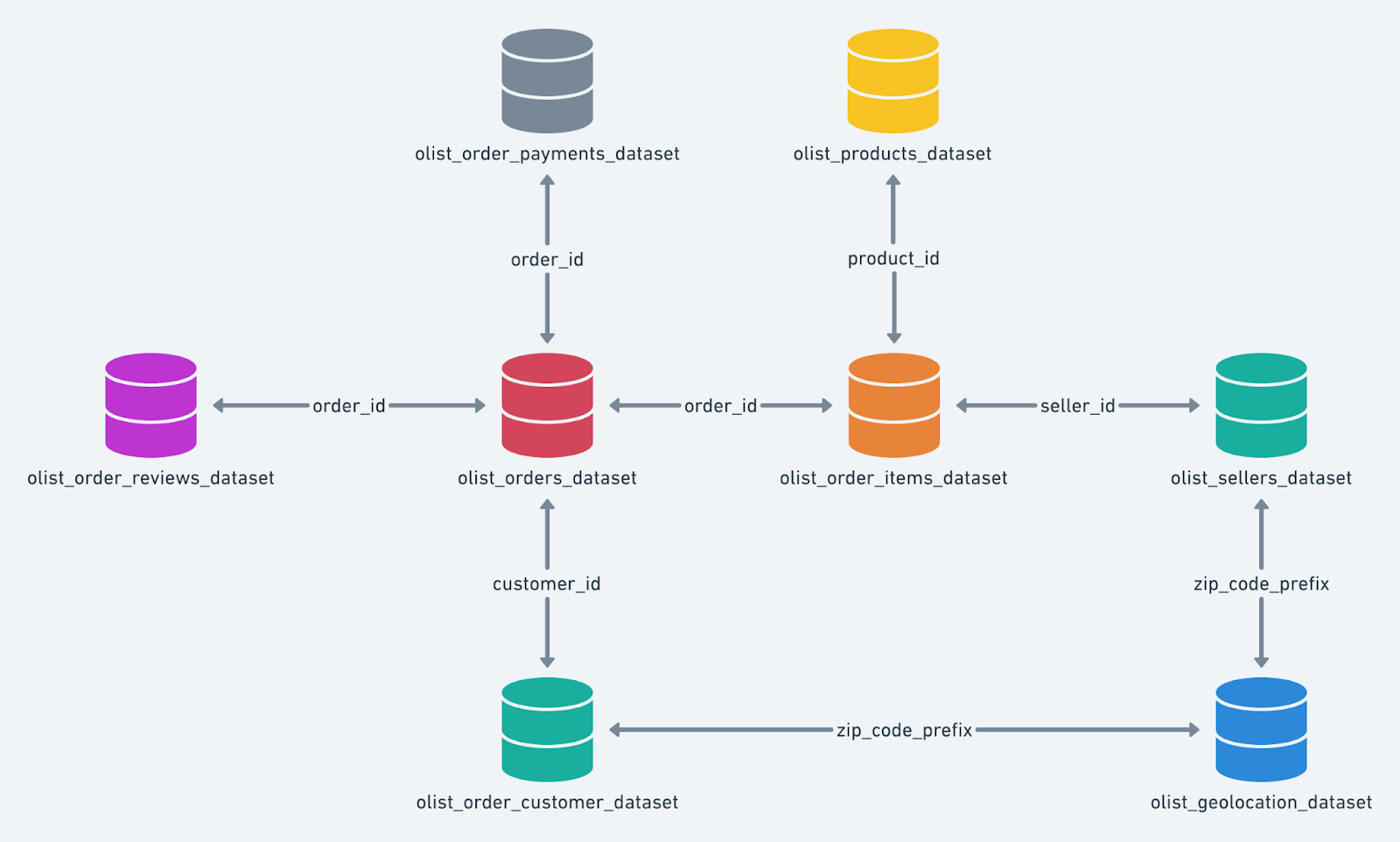
|  |  |
| --- | --- |
| **Features** | **Description** |
| review\_id | ID of the review given on the product ordered by the order id |
| order\_id | A Unique ID of order made by the consumers |
| review\_score | Review score given by the customer for each order on a scale of 1-5 |
| review\_comment\_title | Title of the review |
| review\_comment\_message | Review comments posted by the consumer for each order |
| review\_creation\_date | Timestamp of the review when it is created |
| review\_answer\_timestamp | Timestamp of the review answered |

The **products.csv** contain following features:

|  |  |
| --- | --- |
| **Features** | **Description** |
| product\_id | A Unique identifier for the proposed project. |
| product\_category\_name | Name of the product category |
| product\_name\_lenght | Length of the string which specifies the name given to the products ordered |
| product\_description\_lenght | Length of the description written for each product ordered on the site |
| product\_photos\_qty | Number of photos of each product ordered available on the shopping portal |
| product\_weight\_g | Weight of the products ordered in grams |
| product\_length\_cm | Length of the products ordered in centimeters |
| product\_height\_cm | Height of the products ordered in centimeters |
| product\_width\_cm | Width of the product ordered in centimeters |

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Dataset schema:**



**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Problem Statement:**

Data analysts in your team wants the following transformed models in the Datawarehouse so that they can directly make them into charts and display them to stakeholders for business analysis. You have been assigned the task of creating new tables/views in the database **Live\_data\_**(Data analyst/ scientists only has access to this database) for the given dataset.

Build the following models after analyzing the dataset.

1. **Import the dataset to bigquery. Create a source schema in your project-dataset(bigquery) and do usual exploratory analysis steps like checking the structure & characteristics of the dataset.**
2. **In-depth Exploration:**
   1. Get the time range between which the orders were placed and calculate range in number of months.
   2. Count the number of Cities and States in our dataset.
   3. Is there a growing trend in the no. of orders placed over the past years?
   4. During what time of the day, do the Brazilian customers mostly place their orders? (Dawn, Morning, Afternoon or Night)
      * 0-6 hrs : Dawn
      * 7-12 hrs : Mornings
      * 13-18 hrs : Afternoon
      * 19-23 hrs : Night
3. **Evolution of E-commerce orders in the Brazil region:**
   1. Get the month on month no. of orders placed in each state.
   2. How are the customers distributed across all the states?
4. **Impact on Economy: Analyze the money movement by e-commerce by looking at order prices, freight and others.**
   1. Get the % increase in the cost of orders from year 2017 to 2018 (include months between Jan to Aug only).  
      You can use the "payment\_value" column in the payments table to get the cost of orders.
   2. Calculate the Total & Average value of order price and order freight for each state.
5. **Analysis based on sales, freight and delivery time.**
   1. Find the no. of days taken to deliver each order from the order’s purchase date as delivery time.  
      Also, calculate the difference (in days) between the estimated & actual delivery date of an order.  
      Do this in a single query.  
        
      You can calculate the delivery time and the difference between the estimated & actual delivery date using the given formula:
      * **time\_to\_deliver** = order\_delivered\_customer\_date - order\_purchase\_timestamp
      * **diff\_estimated\_delivery** = order\_estimated\_delivery\_date - order\_delivered\_customer\_date
   2. Find out the top 5 states with the highest & lowest average freight value and average delivery time.  
      **Note: Use dbt seed – create a csv with full names of states of brazil as we only have abbrevations in dataset.**
   3. Find out the top 5 states where the order delivery is really fast as compared to the estimated date of delivery.  
      You can use the difference between the averages of actual & estimated delivery date to figure out how fast the delivery was for each state.
6. **Analysis based on the payments:**
   1. Find the month on month no. of orders placed using different payment types.
   2. Find the no. of orders placed on the basis of the payment installments that have been paid.