**BigQuery Practice Project: thelook\_ecommerce**

**Datasets:**

* **Orders** (order\_id, user\_id, status, gender, created\_at, num\_of\_item)
* **Order items** (order\_id, user\_id, product\_id, inventory\_item\_id)
* **Events** – browser details (id, user\_id, session\_id, browser, ip\_address, postal\_code, traffic\_source, uri (unique , event\_type)
* **Products** (id, distribution\_center\_id)
* **Users** (id, lat, lon, traffic\_source)

**Questions: Products**

What were

**Questions: Order related**

What were the most ordered items?

What were the most profitable items?

Who were the most loyal customers?

What was the most profitable category?

What was the average number of orders?

How many items per order on average?

**Questions: Events**

What was the average duration of a session?

What is the relationship between session and order number? How about for number of items?

What was the most utilized browser? Is there a statistical significance between the kind of browsers that users used and whether they placed an order?

What was the most repeated ip\_address?

Where are most of the users who placed orders from (lat/lon on postal\_code)?

What was the most common traffic source?

Something with URI?

Where are most users from?

**Question: Statistical Analysis and Predictive modeling**

What are the correlations between the different variables?

Are there any linear relationships between the variables?

Can we predict the likelihood that a customer will order based on X features?

Can we predict the amount spent?

**Getting the Data:**

Query the tables to get the data frame I want from BigQuery

Users – (id) – Products – (id) – Events

Events – (user id) Order items

Order – (order\_id) – Order items

**Orders** (order\_id, user\_id, status, gender, created\_at, num\_of\_item)

**Order items** (order\_id, user\_id, product\_id, inventory\_item\_id)

**Products** (id, distribution\_center\_id)

**Events** – browser details (id, user\_id, session\_id, browser, ip\_address, postal\_code, traffic\_source, uri (unique , event\_type)

**Users** (id, lat, lon, traffic\_source)

Query pseudo code

SELECT \* from Users

LEFT JOIN Events

ON Users.id = Events.user\_id

LEFT JOIN Order\_items

ON Users.id = Order\_items.user\_id

LEFT JOIN Orders

ON Order\_items.order\_id = Orders.order\_id

**Models: Brainstorm**

Customer segmentation using Users table

Hypothesis testing using Events table – is there a statistical significance between browsers and session time,

Product recommendation using Product and Orders tables

Network Analysis with Users to identify most influential user collaborations