# JOB-A-THON - June 2021

## Marketplace Feature Table

Your Client ComZ is an ecommerce company. The company wants to focus on targeting the right customers  with the right products to increase overall revenue and conversion rate.

To target the right customers with the right products, they need to build an ML model for marketing based on user interaction with products in the past like number of views,  most viewed product, number of activities of user, vintage of user and others.

ComZ has contacted the Data Science and Engineering team to use this information to fuel the personalized advertisements, email marketing campaigns, or special offers on the landing and category pages of the company's website.

You, being a part of the data engineering team, are expected to “Develop input features”  for the efficient marketing model given the Visitor log data and User Data.

**Approach Steps:**

1. Dropped the rows which doesn’t have UserID in the **VisitorLogData** for non-registered users and also dropped rows which doesn’t have **VisitDateTime.**

2. Replaced the empty string to null to maintain consistency across the data set

3. Converted the string format into date column

* datetime format “2018-05-07 04:28:45.970” has been converted using **to\_date**
* unix datetime format “1527051855673000000”has been divided by 1000000000 to handle “ns” and later converted by using **to\_timestamp** function

4. Filtered the **VisitorLogData** for last 21 days

5. Normalization of data has been done by converting data into lowercase for the columns ["ProductID","Activity", "Browser", "OS", "City", "Country"]

6. Processed data is passed into feature generation funtion to create the required features to the data

**Feature generation function:**

Function build ETL Pipeline such that passing the information of user data and log data,It can generate the input feature table automatically

1. User vs Timeline (User \* 21 days date) dataframe is created, which is further to draw insights on user level

2. Data is merged with users Data, Vistor Log data to created a final merged file and further analysis is done on this dataframe

3. Different approaches are followed to generate the further which are described below:

**User Vintage:**

* Total age of the customer is calculated using difference of reference date (28-May-2018) and Signup Date

**Most\_Active\_OS:**

* Grouped the data by UserID and OS and removed null values from the group data
* Most Active OS is taken by looking the Max values of OS after removing the null values (Similar to filling the missing values by mode on group and identifying the Active OS)

**Recently\_Viewed\_Product**

* Most recently viewed (page loads) product by the user.
* Filter the data where activity is pageload and removing null values associated to Product ID
* Data is order the data by UserID and descending order of VisitDateTime

and removed demove duplicates by UserID to capture the Recently viewed product

**No\_of\_days\_Visited\_7\_Days:**

* Filtered the data for last 7 days and webClientID to identify users visits, this is stored into different dataframe
* From the new dataframe, we can find the distinct VisitDateTime to capture the No of days a user visited in 7 days

**Pageloads\_last\_7\_days & Clicks\_last\_7\_days:**

* From the dataframe in previous step pivot the dataframe with UserID and Activity and further drop down the “null” column from the data

**Most\_Viewed\_product\_15\_Days:**

* Filtered the data for last 15 days with ProductID are not null to create a new dataframe
* Using the dataframe created calculated the values count by using groupby UserID, ProductID and rearrange the data by UserID and descending order of Count of values from groupby function
* Capture First row for each user to identify the most viewed product in last 15 days

Once the feature are created then merged all the features into one dataframe and fill the missing values in the column **Most\_Viewed\_product\_15\_Days** and **Recently\_Viewed\_Product** with **Product101.** Fill the missing of the other columns with 0. Once data is prepared order it using **UserID**

**Tools Used:**

- PySpark

- Pandas