## Problem Statement:

Given 3 datasets related to online products sales in the year 2019 for months Jan, Feb and March for a given store.

Need to do data engineering using PySpark or Spark(Scala/Java) on these datasets to obtain the following objectives

1. Cleanse the data removing blank rows
2. Get the date on which max sales was done by product in these 3 months
3. Get the date on which max sales was done for all products in these 3 months
4. Get the average sales value for each product in these 3 months
5. Create a combined dataset merging all these 3 datasets with order by date in desc order and add a new column which is “salesdiff” where this column will contain the difference of the sales in the current row (current date of that row) and the next row (previous date of that row, as the date columns are sorted by desc) grouped on the product

For the last row, next row will be blank so consider the sales as 0

1. Get the orderId and purchase address details who made max sales in all the 3 months
2. Extract city from the purchase address column which is 2nd element in , delimited separated string and determine the city from where more orders came in all these 3 months
3. Get the total order count details for each city in all the 3 months
4. Create partition by “bi-weekly” order dates and save the data in physical storage in DBFS or ADLS or local storage or delta table or S3 or HDFS.

Note:

* Sales value calculated by qty \* price
* orders count can be determined based on orderId(one orderId means 1 order)