# Spark - Assignment

## Problem-1

The dataset consist of details of all the laptops models(SKUs) listed on the e-commerce website. Details like Name, brand, selling price, MRP, discount, ratings, rating count, details, etc. The  data is retrieved from e-commerce website by some process.
There are 8 columns and more than 200+ rows in the dataset.
The columns are as follows:
- Name ( name of the product)
- Brand ( Brand)
- Selling Price
- MRP
- Discount
- ratings
- no_of_ratings
- Details

1. Read the given csv file into Spark by defining the Schema (Assign datatypes as per your understanding
2. Print the Schema constructed
3. Identify the unique number of rows in the dataset.
4. If there are any null values drop the null values.
5. Find the most expensive laptop from the given dataset
6. Find the cheapest laptop from the given dataset
7. Find the 10 most expensive and 10 most cheapest laptop from the dataset
8. Identify the laptop with the most and the least discount respectively
9. Find the laptop with highest ratings
10. Find the most expensive and least expensive 'ASUS' laptop.

## Problem-2

Each file contains anywhere from around 9000 to 26000 rows and 6 columns.
The columns are as follows:
Order ID, Product, Quantity Ordered, Price Each, Order Date, Purchase Address

## Tasks

# **Spark - Assignment**

- Merge 4 months of sales data into a single CSV file
- Create new 'Month' column from 'Order Date' column
- Add a Sales column
- Add a City column

## Questions

- What was the best month for sales?
- How much was earned the month that had highest sales?
- What city has the highest sales?
- What time should we display advertisements to maximize likehood of customers buying products?
- What products are most often sold together?
- What product sold the most?