**Assignment 1**

**Assignment Submission Date:** AUG 28th 2021

**Title**: Using Hadoop for historical sales data analysis

**Business context:**

A large multi-national retail chain has sales orders data across regions and different sales channels for a large variety of item types. The business team wants to use this data to analyze various aspects of sales - e.g. top selling items in a region, regions with maximum profit in a certain item type, if there is a significant difference in revenue in two item types across regions etc.

**Problem statement:**

As the data analytics team, use the sales transaction data set with about 100K records to answer these questions below —

1. Average unit\_price by country for a given item type in a certain year
2. Total units\_sold by year for a given country and a given item type
3. Find the max and min units\_sold in any order for each year by country for a given item type. Use a custom partitioner class instead of default hash based.
4. What are the top 10 order id for a given year by the total\_profit

You have to show the above analysis working on a Hadoop system using map reduce code, preferably in Java or Python. You can do data preparation steps as required before running a MapReduce job to answer these questions above.

**Submission materials:**

* Working documented code

Data set:

[https://drive.google.com/drive/folders/1fJyk6GE13xw0P-OKWwW8fxa-ny1G0TPp?usp=sharing](https://drive.google.com/drive/folders/1fjyk6ge13xw0p-okwww8fxa-ny1g0tpp?usp=sharing)

