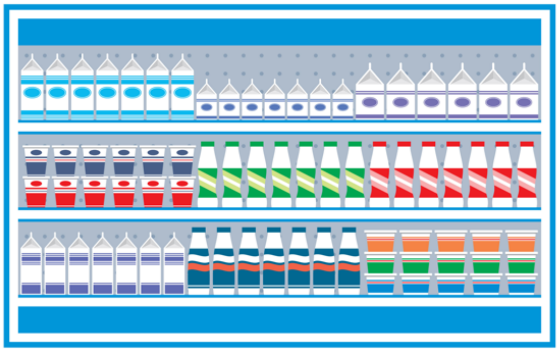
Shoppers dataset description

This dataset is part of the Acquire Valued Shoppers Challenge on Kaggle. The aim of the Kaggle competition is to predict which shoppers, when presented a promotionnal offer, are most likely to repeat purchase. However, we will only explore a part of the original dataset to learn RDD operations and queries in Spark. The original Kaggle dataset provided the pre-offer shopping history, at the basket-level, for a large set of shoppers who were targeted for an acquisition campaign. The incentive offered to that shopper and their post-incentive behavior is also provided.



This challenge provided almost **350 million rows** of completely anonymised transactional data from over **300,000 shoppers**. Since the dataset requires about 22GB of space, we will only use a portion of the transactions data for our exercise. Precisely, topTransactions.csv contains the history of five customers that have the most repeated purchases after redeeming an offer.

This data captures the process of offering incentives (a.k.a. coupons) to a large number of customers and forecasting those who will become loyal to the product. Let's say 100 customers are offered a discount to purchase two bottles of water. Of the 100 customers, 60 choose to redeem the offer. These 60 customers are the focus of this competition. You are asked to predict which of the 60 will return (during or after the promotional period) to purchase the same item again.

For each customer, you are given a minimum of a year of shopping history prior to each customer's incentive. The transaction history contains all items purchased, not just items related to the offer. Only one offer per customer is included in the data.

Files

You are provided four relational files:

- **topTransactions.csv** - contains transaction history for the top 5 customers (in terms of repeat purchases after redeeming an offer) for a period of at least 1 year prior to their offered incentive
- **history.csv** - contains the incentive offered to each customer and information about the behavioral response to the offer
- **offers.csv** - contains information about the offers

Fields

All of the fields are anonymized and categorized to protect customer and sales information. The specific meanings of the fields will not be provided (so don't bother asking). Part of the challenge of this competition is learning the taxonomy of items in a data-driven way.

history
id - A unique id representing a customer
chain - An integer representing a store chain
offer - An id representing a certain offer
market - An id representing a geographical region
repeattrips - The number of times the customer made a repeat purchase
repeater - A boolean, equal to repeattrips > 0
offerdate - The date a customer received the offer

transactions
id - see above
chain - see above
dept - An aggregate grouping of the Category (e.g. water)
category - The product category (e.g. sparkling water)
company - An id of the company that sells the item
brand - An id of the brand to which the item belongs
date - The date of purchase
productsize - The amount of the product purchase (e.g. 16 oz of water)
productmeasure - The units of the product purchase (e.g. ounces)
purchasequantity - The number of units purchased
purchaseamount - The dollar amount of the purchase

offers
offer - see above
category - see above
quantity - The number of units one must purchase to get the discount
company - see above
offervalue - The dollar value of the offer
brand - see above

The transactions file can be joined to the history file by (id,chain). The history file can be joined to the offers file by (offer). The transactions file can be joined to the offers file by (category, brand, company). A negative value in productquantity and purchaseamount indicates a return.