Preparing Data Sets

We will be primarily using orders and order_items data set to understand about manipulating collections.

- orders is available at path **/data/retail_db/orders/part-00000**
- order_items is available at path **/data/retail_db/order_items/part-00000**
- orders - columns
  - order_id - it is of type integer and unique
  - order_date - it can be considered as string
  - order_customer_id - it is of type integer
  - order_status - it is of type string
- order_items - columns
  - order_item_id - it is of type integer and unique
  - order_item_order_id - it is of type integer and refers to orders.order_id
  - order_item_product_id - it is of type integer and refers to products.product_id
  - order_item_quantity - it is of type integer and represents number of products as an order item with in an order.
  - order_item_subtotal - it is item level revenue (product of order_item_quantity and order_item_product_price)
  - order_item_product_price - it is product price for each item with in an order.
- orders is parent data set to order_items and will contain one record per order. Each order can contain multiple items.
- order_items is child data set to orders and can contain multiple entries for a given order_item_order_id.

Print to PDF

Task 1 - Read orders into collection

Let us read orders data set into the collection called as **orders**. This will be used later.

```
orders_path = '/data/retail_db/orders/part-00000'
# C:\\users\\itversity\\Research\\data\\retail_db\\orders\\part-00000
orders_file = open(orders_path)
```

```
orders_raw = orders_file.read()
```

```
orders = orders_raw.splitlines()
```

```
orders[:10]
```

```
['1,2013-07-25 00:00:00.0,11599,CLOSED',
 '2,2013-07-25 00:00:00.0,256,PENDING_PAYMENT',
 '3,2013-07-25 00:00:00.0,12111,COMPLETE',
 '4,2013-07-25 00:00:00.0,8827,CLOSED',
 '5,2013-07-25 00:00:00.0,11318,COMPLETE',
 '6,2013-07-25 00:00:00.0,7130,COMPLETE',
 '7,2013-07-25 00:00:00.0,4530,COMPLETE',
 '8,2013-07-25 00:00:00.0,2911,PROCESSING',
 '9,2013-07-25 00:00:00.0,5657,PENDING_PAYMENT',
 '10,2013-07-25 00:00:00.0,5648,PENDING_PAYMENT']
```

```
len(orders) # same as number of records in the file
```

```
68883
```

Task 2 - Read order_items into collection

Let us read order_items data set into the collection called as **order_items**. This will be used later.

```
order_items_path = '/data/retail_db/order_items/part-00000'
# C:\\users\\itversity\\Research\\data\\retail_db\\order_items\\part-00000
order_items_file = open(order_items_path)
```

```
order_items_raw = order_items_file.read()
```

```
order_items = order_items_raw.splitlines()
```

```
order_items[:10]
```

```
['1,1,957,1,299.98,299.98',
 '2,2,1073,1,199.99,199.99',
 '3,2,502,5,250.0,50.0',
 '4,2,403,1,129.99,129.99',
 '5,4,897,2,49.98,24.99',
 '6,4,365,5,299.95,59.99',
 '7,4,502,3,150.0,50.0',
 '8,4,1014,4,199.92,49.98',
 '9,5,957,1,299.98,299.98',
 '10,5,365,5,299.95,59.99']
```

```
len(order_items) # same as number of records in the file
```

```
172198
```

By Durga Gadiraju