CSV to Pandas Data Frame

Let us see how we can create **Pandas Data Frames** using data from files. `read_csv` is the most popular API to create a Data Frame by reading data from files.

- Here are some of the important options.
  - sep or delimiter
  - header or names
  - index_col
  - dtype
  - and many more
- We have several other APIs which will facilitate us to create Data Frame
  - read_fwf
  - read_table
  - pandas.io.json
  - and more
- Here is how we can create a Data Frame for orders dataset.
  - Delimiter in our data is , which is default for Pandas `read_csv`.
  - There is no Header and hence we have to set keyword argument `header` to None.
  - We can pass the column names as a list using keyword argument `columns`.
  - Data types of each column are typically inferred based on the data, however we can explicitly specify Data Types using `dtype`.

> **Note**
>
> We will be running this notebook from other notebooks to create orders and order_items data frames while exploring Pandas libraries.
>
> Make sure you comment out all the informational lines, so that output is not printed when we invoke this notebook from other notebooks.

```
import pandas as pd
```

```
# pd.read_csv?
```

```
%%sh

# ls -ltr /data/retail_db/orders/part-00000
```

```
%%sh

# tail /data/retail_db/orders/part-00000
```

```
%%sh

# head /data/retail_db/orders/part-00000
```

```
orders_path = "/data/retail_db/orders/part-00000"
```

```
orders_schema = [
    "order_id",
    "order_date",
    "order_customer_id",
    "order_status"
]
```

```
orders = pd.read_csv(orders_path,
                     delimiter=',',
                     header=None,
                     names=orders_schema
                    )
```

```
# orders
```

```
# orders.head(10)
```

```
order_items_path = "/data/retail_db/order_items/part-00000"
```

```
%%sh

# ls -ltr /data/retail_db/order_items/part-00000
```

```
%%sh

# tail /data/retail_db/order_items/part-00000
```

```
%%sh

# head /data/retail_db/order_items/part-00000
```

```
order_items_schema = [
    "order_item_id",
    "order_item_order_id",
    "order_item_product_id",
    "order_item_quantity",
    "order_item_subtotal",
    "order_item_product_price"
]
```

```
order_items = pd.read_csv(order_items_path,
                    delimiter=',',
                    header=None,
                    names=order_items_schema
                   )
```

```
# order_items
```

```
# order_items.head(10)
```