# Assignment 5.1

Based on the data contained in tasks/4\_data\_pipelines/day\_1\_introduction/daily\_assignment/data directory, use PySpark to read, filter and join the data from CSV files and answer the following questions:

What are the daily total sales for the store with id 1?

What are the mean sales for the store with id 2?

What is the email of the client who spent the most when summing up purchases from all of the stores?

Which 5 products are most frequently bought across all stores?

# Problem # 1

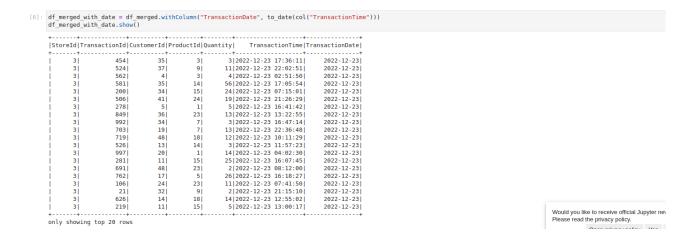
What are the daily total sales for the store with id 1?

#### Type casting Quantity into Integer

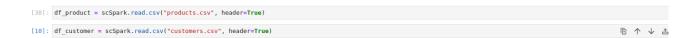
```
[5]: df_merged = df_merged.withColumn(
    "Quantity", df_merged["Quantity"].cast(IntegerType())
)
df_merged
```

[5]: DataFrame[StoreId: string, TransactionId: string, CustomerId: string, ProductId: string, Quantity: int, TransactionTime: string]

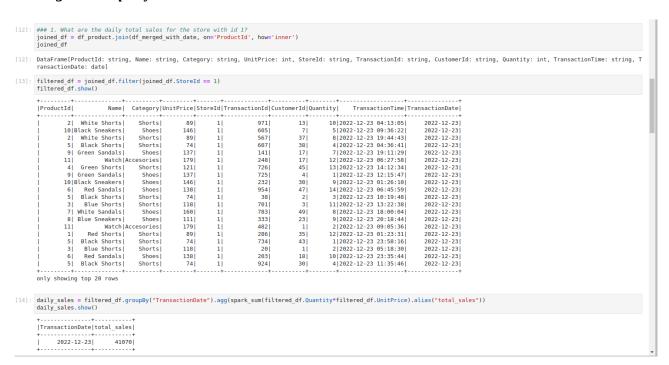
### Splitting date from transaction time give alias transaction data



#### Storing customer and product into variable



#### Solving the 1<sup>st</sup> query



#### Solving the 2nd query

#### Solving the 3rd query

## Solving the 4rt query