Coding Challenge PySpark

Name: Divya Sree Murali **Date: 26/11/2024**

1.Explain ETL (Extract, Transform, Load) with PySpark(in your own words)

ETL (Extract, Transfer, Load) is a process used to manage data workflows. It is used to Combine data from multiple sources into a unified view and format data for better analysis. PySpark is the Python API for Apache Spark that provides a scalable and efficient way to implement ETL. By leveraging PySpark, the ETL process becomes more efficient and scalable.

1. **Extract:**

This phase involves fetching of raw data from various sources like relational databases, files, streaming data etc. PySpark supports multiple formats like CSV, JSON files. To read the data, PySpark uses SparkSession.read keyword

2. **Transform:**

The raw data extracted often requires cleaning, enrichment, or reshaping to meet business requirements. PySpark provides tools to perform these transformations at scale:

- Use Data Frame API for operations like filtering, joining, grouping, or aggregating.
- Use PySpark SQL for querying data with SQL-like syntax.
- Apply functions using with Column or select to modify or add new columns.

3. **Load**

The processed data is then written to a target location, such as a data warehouse, database, or storage system. PySpark supports writing data in various formats and to different storage backends and Formats supported include CSV, Parquet, ORC, JSON, etc.

The Advantages of using PySpark for ETL is that:

PySpark is highly scalable such that it can handle large datasets distributed across clusters. It is flexible and hence supports multiple data sources and formats. It is optimized speed for parallel processing. It is versatile and can easily integrate with numerous data sources and works well with other Big Data Tools

2..Using spark sql Transformations such as Filter, Join, Simple Aggregations, GroupBy on the case study dataset

Using spark sql

In 2 minutes (2s) 2

```python
from pyspark.sql import SparkSession

# Initialize Spark Session
spark = SparkSession.builder.appName("CreditCardAnalysis").getOrCreate()

# Load dataset
credit_card_df = spark.read.csv("/FileStore/tables/credit_card-1.csv", header=True, inferSchema=True)
new_credit=spark.read.csv("/FileStore/tables/book.csv", header=True, inferSchema=True)

# Create a temporary SQL view
credit_card_df.createOrReplaceTempView("credit_card")
new_credit.createOrReplaceTempView("new")
```

(5) Spark Jobs

credit_card_df: pyspark.sql.dataframe.DataFrame = [RowNumber: integer, CustomerId: integer ... 11 more fields]

new_credit: pyspark.sql.dataframe.DataFrame = [RowNumber: integer, CustomerId: integer ... 11 more fields]

1.Filter

List all customers having credit score greater tan 830

List all female customers

05:23 PM (1s) 3 Python

```python
# 1. Filter transactions where credit score is above 830
spark.sql("SELECT * FROM credit_card WHERE CreditScore > 830").show()
```

(2) Spark Jobs

RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	IsActiveMember	EstimatedSalary	Exited
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	79084.1	0
25	15625047	Yen	846	France	Female	38	5	0.0	1	1	187616.16	0
39	15717426	Armstrong	850	France	Male	36	7	0.0	1	1	40812.9	0
44	15755196	Lavine	834	France	Female	49	2	131394.56	1	0	194365.76	1
181	15716334	Rozier	850	Spain	Female	45	2	122311.21	1	1	19482.5	0
201	15604482	Chiemezie	850	Spain	Male	30	2	141040.01	1	1	5978.2	0
224	15733247	Stevenson	850	France	Male	33	10	0.0	1	0	4861.72	1
260	15607178	Welch	850	Germany	Male	38	3	54901.01	1	1	140075.55	0
346	15763859	Brown	840	France	Female	43	7	0.0	2	0	90908.95	0
357	15611759	Simmons	850	Spain	Female	57	8	126776.3	2	1	132298.49	0
412	15760431	Pino	850	France	Male	38	1	0.0	2	1	80006.65	0
452	15785798	Uchechukwu	850	France	Male	40	9	0.0	2	1	119232.33	0
467	15663252	Olisanugo	850	Spain	Female	32	9	0.0	2	1	18924.92	0
521	15671256	Macartney	850	France	Female	35	1	211774.31	1	0	188574.12	1
522	15653547	Madukwe	850	France	Male	56	7	131317.48	1	1	119175.45	0
541	15667896	De Luca	833	France	Male	37	8	151226.18	2	1	136129.49	0
546	15615457	Burns	842	Spain	Female	44	2	112652.08	2	0	126644.98	0

05:24 PM (<1s) 4 Python

```python
# 2. FIlter for female customer

spark.sql("SELECT * FROM credit_card WHERE Gender = 'Female'").show()
```

(1) Spark Jobs

RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	IsActiveMember	EstimatedSalary	Exited
1	15634602	Hargrave	619	France	Female	42	2	0.0	1	1	101348.88	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	1	112542.58	0
3	15619304	Onio	502	France	Female	42	8	159660.8	3	0	113931.57	1
4	15701354	Boni	699	France	Female	39	1	0.0	2	0	93826.63	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	79084.1	0
8	15656148	Obinna	376	Germany	Female	29	4	115046.74	4	0	119346.88	1
13	15632264	Kay	476	France	Female	34	10	0.0	2	0	26260.98	0
14	15691483	Chin	549	France	Female	25	5	0.0	2	0	190857.79	0
15	15600882	Scott	635	Spain	Female	35	7	0.0	2	1	65951.65	0
18	15788218	Henderson	549	Spain	Female	24	9	0.0	2	1	14406.41	0
20	15568982	Hao	726	France	Female	24	6	0.0	2	1	54724.03	0
22	15597945	Dellucci	636	Spain	Female	32	8	0.0	2	0	138555.46	0
23	15699309	Gerasimov	510	Spain	Female	38	4	0.0	1	0	118913.53	1
25	15625047	Yen	846	France	Female	38	5	0.0	1	1	187616.16	0
29	15728693	McWilliams	574	Germany	Female	43	3	141349.43	1	1	100187.43	0
31	15589475	Azikiwe	591	Spain	Female	39	3	0.0	3	0	140469.38	1

2.Joins

Inner Join

```python
# 2.Using joins
# 1. Inner join on customer_id

spark.sql("""
    SELECT cc.*
    FROM credit_card cc
    INNER JOIN new n
    ON n.CustomerId = cc.CustomerId
""").show()
```

RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	IsActiveMember	EstimatedSalary	Exited
1	15634602	Hargrave	619	France	Female	42	2	0.0	1	1	101348.88	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	1	112542.58	0
3	15619304	Onio	502	France	Female	42	8	159660.8	3	0	113931.57	1
4	15701354	Boni	699	France	Female	39	1	0.0	2	0	93826.63	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	79084.1	0

Full outer join

```python
# 2. Outer join on customer_id
spark.sql("""
    SELECT cc.*
    FROM credit_card cc
    full OUTER JOIN new n
    ON n.CustomerId = cc.CustomerId
""").show(5)
```

RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	IsActiveMember	EstimatedSalary	Exited
1288	15565701	Ferri	698	Spain	Female	39	9	161993.89	1	0	90212.38	0
4199	15565706	Akobundu	612	Spain	Male	35	1	0.0	1	1	83256.26	1
7091	15565714	Cattaneo	601	France	Male	47	1	64430.06	2	1	96517.97	0
2021	15565779	Kent	627	Germany	Female	30	6	57809.32	1	0	188258.49	0
3698	15565796	Docherty	745	Germany	Male	48	10	96048.55	1	0	74510.65	0

Left join

```
# 3.Left join
spark.sql("""
    SELECT cc.*
    FROM credit_card cc
    left JOIN new n
    ON n.CustomerId = cc.CustomerId
""").show(5)
```

```
+---------+----------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|RowNumber|CustomerId| Surname|CreditScore|Geography|Gender|Age|Tenure|  Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|
+---------+----------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|        1|  15634602|Hargrave|        619|   France|Female| 42|     2|      0.0|            1|             1|      101348.88|     1|
|        2|  15647311|    Hill|        608|    Spain|Female| 41|     1| 83807.86|            1|             1|      112542.58|     0|
|        3|  15619304|    Onio|        502|   France|Female| 42|     8| 159660.8|            3|             0|      113931.57|     1|
|        4|  15701354|    Boni|        699|   France|Female| 39|     1|      0.0|            2|             0|       93826.63|     0|
|        5|  15737888|Mitchell|        850|    Spain|Female| 43|     2|125510.82|            1|             1|        79084.1|     0|
+---------+----------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
only showing top 5 rows
```

Right join

```
# Right join
spark.sql("""
    SELECT cc.*
    FROM credit_card cc
    Right JOIN new n
    ON n.CustomerId = cc.CustomerId
""").show(5)
```

```
+---------+----------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|RowNumber|CustomerId| Surname|CreditScore|Geography|Gender|Age|Tenure|  Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|
+---------+----------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|        1|  15634602|Hargrave|        619|   France|Female| 42|     2|      0.0|            1|             1|      101348.88|     1|
|        2|  15647311|    Hill|        608|    Spain|Female| 41|     1| 83807.86|            1|             1|      112542.58|     0|
|        3|  15619304|    Onio|        502|   France|Female| 42|     8| 159660.8|            3|             0|      113931.57|     1|
|        4|  15701354|    Boni|        699|   France|Female| 39|     1|      0.0|            2|             0|       93826.63|     0|
|        5|  15737888|Mitchell|        850|    Spain|Female| 43|     2|125510.82|            1|             1|        79084.1|     0|
+---------+----------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
only showing top 5 rows
```

Left Anti join

```
#Left Anti join
spark.sql("""
    SELECT cc.*
    FROM credit_card cc
    LEFT ANTI JOIN new n
    ON n.CustomerId = cc.CustomerId
""").show(5)
```

```
+---------+----------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|RowNumber|CustomerId| Surname|CreditScore|Geography|Gender|Age|Tenure|  Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|
+---------+----------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|        1|  15634602|Hargrave|        619|   France|Female| 42|     2|      0.0|            1|             1|      101348.88|     1|
|        2|  15647311|    Hill|        608|    Spain|Female| 41|     1| 83807.86|            1|             1|      112542.58|     0|
|        3|  15619304|    Onio|        502|   France|Female| 42|     8| 159660.8|            3|             0|      113931.57|     1|
|        4|  15701354|    Boni|        699|   France|Female| 39|     1|      0.0|            2|             0|       93826.63|     0|
|        5|  15737888|Mitchell|        850|    Spain|Female| 43|     2|125510.82|            1|             1|        79084.1|     0|
+---------+----------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
```

Left semi join

```
#left semi join
spark.sql("""
    SELECT cc.*
    FROM credit_card cc
    LEFT SEMI JOIN new n
    ON n.CustomerId = cc.CustomerId
""").show(5)
```

RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	IsActiveMember	EstimatedSalary	Exited
6	15574012	Chu	645	Spain	Male	44	8	113755.78	2	0	149756.71	1
7	15592531	Bartlett	822	France	Male	50	7	0.0	2	1	10062.8	0
8	15656148	Obinna	376	Germany	Female	29	4	115046.74	4	0	119346.88	1
9	15792365	He	501	France	Male	44	4	142051.07	2	1	74940.5	0
10	15592389	H?	684	France	Male	27	2	134603.88	1	1	71725.73	0

3.Simple Aggregations

Calculate total Balance Amount of all customers

Calculate revenue of Balance Amount(Calculate average)

In 2 minutes (1s) 6

```python
# Using Simple Aggregations
# 1. Total Balance Amount
spark.sql("SELECT SUM(Balance) AS total_amount FROM credit_card").show()

#2.Average of Balance amount
spark.sql("SELECT avg(Balance) AS Average FROM credit_card").show()
```

(4) Spark Jobs

```
+------------------+
|      total_amount|
+------------------+
|7.648588928799961E8|
+------------------+

+-----------------+
|          Average|
+-----------------+
|76485.88928799961|
+-----------------+
```

4.Group By

Calculate revenue genereated in each geographical region

05:28 PM (2s) 7

```python
# Using group by
# 2. Average transaction amount by region
spark.sql("""
    SELECT Geography, AVG(Balance) AS avg_spent
    FROM credit_card
    GROUP BY Geography
""").show()
```

(2) Spark Jobs

```
+---------+-----------------+
|Geography|        avg_spent|
+---------+-----------------+
|  Germany|119730.11613391782|
|   France|  62092.6365157559|
|    Spain| 61818.14776342349|
+---------+-----------------+
```

Using PySpark

Initializing Pyspark session

```
In 2 minutes (2s)                9

from pyspark.sql import SparkSession

# Initialize Spark Session
spark = SparkSession.builder \
    .appName("Credit Card Analysis") \
    .getOrCreate()

# Load the dataset
file_path = '/FileStore/tables/credit_card-1.csv'
credit_df = spark.read.csv(file_path, header=True, inferSchema=True)

filenew = '/FileStore/tables/book.csv'
new_df=spark.read.csv(filenew, header=True, inferSchema=True)

(4) Spark Jobs
credit_df: pyspark.sql.dataframe.DataFrame = [RowNumber: integer, CustomerId: integer ... 11 more fields]
new_df: pyspark.sql.dataframe.DataFrame = [RowNumber: integer, CustomerId: integer ... 11 more fields]
```

1.Filter

List all customers having credit score greater tan 830

```
In 2 minutes (<1s)                10                Python

# Filter transactions where CreditScore > 830 using DataFrame API
filtered_df = credit_df.filter(credit_df["CreditScore"] > 830)

# Show the results
filtered_df.show()

(1) Spark Jobs
filtered_df: pyspark.sql.dataframe.DataFrame = [RowNumber: integer, CustomerId: integer ... 11 more fields]

+---------+----------+---------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|RowNumber|CustomerId|  Surname|CreditScore|Geography|Gender|Age|Tenure|  Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|
+---------+----------+---------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|        5|  15737888| Mitchell|        850|    Spain|Female| 43|     2|125510.82|            1|             1|        79084.1|     0|
|       25|  15625047|      Yen|        846|   France|Female| 38|     5|      0.0|            1|             1|      187616.16|     0|
|       39|  15717426|Armstrong|        850|   France|  Male| 36|     7|      0.0|            1|             1|        40812.9|     0|
|       44|  15755196|   Lavine|        834|   France|Female| 49|     2|131394.56|            1|             0|      194365.76|     1|
|      181|  15716334|   Rozier|        850|    Spain|Female| 45|     2|122311.21|            1|             1|        19482.5|     0|
|      201|  15604482|Chiemezie|        850|    Spain|  Male| 30|     2|141040.01|            1|             1|         5978.2|     0|
|      224|  15733247|Stevenson|        850|   France|  Male| 33|    10|      0.0|            1|             0|        4861.72|     1|
|      260|  15607178|    Welch|        850|  Germany|  Male| 38|     3| 54901.01|            1|             1|      140075.55|     0|
|      346|  15763859|    Brown|        840|   France|Female| 43|     7|      0.0|            2|             0|       90908.95|     0|
|      357|  15611759|  Simmons|        850|    Spain|Female| 57|     8| 126776.3|            2|             1|      132298.49|     0|
```

List all female customers

```
# Filter for female customers using DataFrame API
female_customers_df = credit_df.filter(credit_df["Gender"] == "Female")

# Show the results
female_customers_df.show()
```

▸ (1) Spark Jobs

▸ female_customers_df: pyspark.sql.dataframe.DataFrame = [RowNumber: integer, CustomerId: integer ... 11 more fields]

```
+---------+----------+----------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|RowNumber|CustomerId|   Surname|CreditScore|Geography|Gender|Age|Tenure|  Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|
+---------+----------+----------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|        1|  15634602|  Hargrave|        619|   France|Female| 42|     2|      0.0|            1|             1|      101348.88|     1|
|        2|  15647311|      Hill|        608|    Spain|Female| 41|     1| 83807.86|            1|             1|      112542.58|     0|
|        3|  15619304|      Onio|        502|   France|Female| 42|     8| 159660.8|            3|             0|      113931.57|     1|
|        4|  15701354|      Boni|        699|   France|Female| 39|     1|      0.0|            2|             0|       93826.63|     0|
|        5|  15737888|  Mitchell|        850|    Spain|Female| 43|     2|125510.82|            1|             1|        79084.1|     0|
|        8|  15656148|    Obinna|        376|  Germany|Female| 29|     4|115046.74|            4|             0|      119346.88|     1|
|       13|  15632264|       Kay|        476|   France|Female| 34|    10|      0.0|            2|             0|       26260.98|     0|
|       14|  15691483|      Chin|        549|   France|Female| 25|     5|      0.0|            2|             0|      190857.79|     0|
|       15|  15600882|     Scott|        635|    Spain|Female| 35|     7|      0.0|            2|             1|       65951.65|     0|
|       18|  15788218| Henderson|        549|    Spain|Female| 24|     9|      0.0|            2|             1|       14406.41|     0|
|       20|  15568982|       Hao|        726|   France|Female| 24|     6|      0.0|            2|             1|       54724.03|     0|
|       22|  15597945|  Dellucci|        636|    Spain|Female| 32|     8|      0.0|            2|             0|      138555.46|     0|
|       23|  15699309| Gerasimov|        510|    Spain|Female| 38|     4|      0.0|            1|             0|      118913.53|     1|
|       25|  15625047|       Yen|        846|   France|Female| 38|     5|      0.0|            1|             1|      187616.16|     0|
|       29|  15728693|McWilliams|        574|  Germany|Female| 43|     3|141349.43|            1|             1|      100187.43|     0|
```

2.Joins

▶ ✓ In 2 minutes (4s) 14

```
#Joins
joined_df = credit_df.join(new_df, credit_df["CustomerId"] == new_df["CustomerId"], "inner").show()
joined_df = credit_df.join(new_df, credit_df["CustomerId"] == new_df["CustomerId"], "outer").show()
joined_df = credit_df.join(new_df, credit_df["CustomerId"] == new_df["CustomerId"], "left").show()
joined_df = credit_df.join(new_df, credit_df["CustomerId"] == new_df["CustomerId"], "right").show()
joined_df = credit_df.join(new_df, credit_df["CustomerId"] == new_df["CustomerId"], "leftanti").show()
joined_df = credit_df.join(new_df, credit_df["CustomerId"] == new_df["CustomerId"], "leftsemi").show()
```

Innerjoin output

```
+---------+----------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+---------+-
---------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|RowNumber|CustomerId| Surname|CreditScore|Geography|Gender|Age|Tenure|  Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|RowNumber|C
ustomerId| Surname|CreditScore|Geography|Gender|Age|Tenure|  Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|
+---------+----------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+---------+-
---------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|        1|  15634602|Hargrave|        619|   France|Female| 42|     2|      0.0|            1|             1|      101348.88|     1|        1|
15634602|Hargrave|        619|   France|Female| 42|     2|      0.0|            1|             1|      101348.88|     1|
|        2|  15647311|    Hill|        608|    Spain|Female| 41|     1| 83807.86|            1|             1|      112542.58|     0|        2|
15647311|    Hill|        608|    Spain|Female| 41|     1| 83807.86|            1|             1|      112542.58|     0|
|        3|  15619304|    Onio|        502|   France|Female| 42|     8| 159660.8|            3|             0|      113931.57|     1|        3|
15619304|    Onio|        502|   France|Female| 42|     8| 159660.8|            3|             0|      113931.57|     1|
|        4|  15701354|    Boni|        699|   France|Female| 39|     1|      0.0|            2|             0|       93826.63|     0|        4|
15701354|    Boni|        699|   France|Female| 39|     1|      0.0|            2|             0|       93826.63|     0|
|        5|  15737888|Mitchell|        850|    Spain|Female| 43|     2|125510.82|            1|             1|        79084.1|     0|        5|
15737888|Mitchell|        850|    Spain|Female| 43|     2|125510.82|            1|             1|        79084.1|     0|
+---------+----------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+---------+-
```

Full outer join output

```
+---------+----------+----------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|RowNumber|CustomerId|   Surname|CreditScore|Geography|Gender|Age|Tenure|  Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|RowNumber
|CustomerId|Surname|CreditScore|Geography|Gender| Age|Tenure|Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|
+---------+----------+----------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+---------
+---------+----------+----------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|     1288|  15565701|     Ferri|        698|    Spain|Female| 39|     9|161993.89|            1|             0|       90212.38|     0|     null
|     null|     null|      null|       null|  null|null|  null|    null|         null|          null|           null|  null|
|     4199|  15565706|  Akobundu|        612|    Spain|  Male| 35|     1|      0.0|            1|             1|       83256.26|     1|     null
|     null|     null|      null|       null|  null|null|  null|    null|         null|          null|           null|  null|
|     7091|  15565714|  Cattaneo|        601|   France|  Male| 47|     1| 64430.06|            2|             1|       96517.97|     0|     null
|     null|     null|      null|       null|  null|null|  null|    null|         null|          null|           null|  null|
|     2021|  15565779|      Kent|        627|  Germany|Female| 30|     6| 57809.32|            1|             0|      188258.49|     0|     null
|     null|     null|      null|       null|  null|null|  null|    null|         null|          null|           null|  null|
|     3698|  15565796|  Docherty|        745|  Germany|  Male| 48|    10| 96048.55|            1|             0|       74510.65|     0|     null
|     null|     null|      null|       null|  null|null|  null|    null|         null|          null|           null|  null|
|     3417|  15565806|    Toosey|        532|   France|  Male| 38|     9|      0.0|            2|             0|       30583.95|     0|     null
|     null|     null|      null|       null|  null|null|  null|    null|         null|          null|           null|  null|
|     6882|  15565878|     Bates|        631|    Spain|  Male| 29|     3|      0.0|            2|             1|      197963.46|     0|     null
|     null|     null|      null|       null|  null|null|  null|    null|         null|          null|           null|  null|
```

Left join output

```
+---------+----------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+---------+-
---------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|RowNumber|CustomerId| Surname|CreditScore|Geography|Gender|Age|Tenure|  Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|RowNumber|C
ustomerId| Surname|CreditScore|Geography|Gender|Age|Tenure|  Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|
+---------+----------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+---------+-
---------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|        1|  15634602|Hargrave|        619|   France|Female| 42|     2|      0.0|            1|             1|      101348.88|     1|        1|
15634602|Hargrave|        619|   France|Female| 42|     2|      0.0|            1|             1|      101348.88|     1|
|        2|  15647311|    Hill|        608|    Spain|Female| 41|     1| 83807.86|            1|             1|      112542.58|     0|        2|
15647311|    Hill|        608|    Spain|Female| 41|     1| 83807.86|            1|             1|      112542.58|     0|
|        3|  15619304|    Onio|        502|   France|Female| 42|     8| 159660.8|            3|             0|      113931.57|     1|        3|
15619304|    Onio|        502|   France|Female| 42|     8| 159660.8|            3|             0|      113931.57|     1|
|        4|  15701354|    Boni|        699|   France|Female| 39|     1|      0.0|            2|             0|       93826.63|     0|        4|
15701354|    Boni|        699|   France|Female| 39|     1|      0.0|            2|             0|       93826.63|     0|
|        5|  15737888|Mitchell|        850|    Spain|Female| 43|     2|125510.82|            1|             1|        79084.1|     0|        5|
15737888|Mitchell|        850|    Spain|Female| 43|     2|125510.82|            1|             1|        79084.1|     0|
+---------+----------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+---------+-
```

Right join output

```
+---------+----------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+---------+-
---------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|RowNumber|CustomerId| Surname|CreditScore|Geography|Gender|Age|Tenure|  Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|RowNumber|C
ustomerId| Surname|CreditScore|Geography|Gender|Age|Tenure|  Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|
+---------+----------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+---------+-
---------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|        1|  15634602|Hargrave|        619|   France|Female| 42|     2|      0.0|            1|             1|      101348.88|     1|        1|
15634602|Hargrave|        619|   France|Female| 42|     2|      0.0|            1|             1|      101348.88|     1|
|        2|  15647311|    Hill|        608|    Spain|Female| 41|     1| 83807.86|            1|             1|      112542.58|     0|        2|
15647311|    Hill|        608|    Spain|Female| 41|     1| 83807.86|            1|             1|      112542.58|     0|
|        3|  15619304|    Onio|        502|   France|Female| 42|     8| 159660.8|            3|             0|      113931.57|     1|        3|
15619304|    Onio|        502|   France|Female| 42|     8| 159660.8|            3|             0|      113931.57|     1|
|        4|  15701354|    Boni|        699|   France|Female| 39|     1|      0.0|            2|             0|       93826.63|     0|        4|
15701354|    Boni|        699|   France|Female| 39|     1|      0.0|            2|             0|       93826.63|     0|
|        5|  15737888|Mitchell|        850|    Spain|Female| 43|     2|125510.82|            1|             1|        79084.1|     0|        5|
15737888|Mitchell|        850|    Spain|Female| 43|     2|125510.82|            1|             1|        79084.1|     0|
+---------+----------+--------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+---------+-
```

Left anti join output

RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	IsActiveMember	EstimatedSalary	Exited
6	15574012	Chu	645	Spain	Male	44	8	113755.78	2	0	149756.71	1
7	15592531	Bartlett	822	France	Male	50	7	0.0	2	1	10062.8	0
8	15656148	Obinna	376	Germany	Female	29	4	115046.74	4	0	119346.88	1
9	15792365	He	501	France	Male	44	4	142051.07	2	1	74940.5	0
10	15592389	H?	684	France	Male	27	2	134603.88	1	1	71725.73	0
11	15767821	Bearce	528	France	Male	31	6	102016.72	2	0	80181.12	0
12	15737173	Andrews	497	Spain	Male	24	3	0.0	2	0	76390.01	0
13	15632264	Kay	476	France	Female	34	10	0.0	2	0	26260.98	0
14	15691483	Chin	549	France	Female	25	5	0.0	2	0	190857.79	0
15	15600882	Scott	635	Spain	Female	35	7	0.0	2	1	65951.65	0
16	15643966	Goforth	616	Germany	Male	45	3	143129.41	2	1	64327.26	0
17	15737452	Romeo	653	Germany	Male	58	1	132602.88	1	0	5097.67	1
18	15788218	Henderson	549	Spain	Female	24	9	0.0	2	1	14406.41	0

Left semi join output

RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	IsActiveMember	EstimatedSalary	Exited
1	15634602	Hargrave	619	France	Female	42	2	0.0	1	1	101348.88	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	1	112542.58	0
3	15619304	Onio	502	France	Female	42	8	159660.8	3	0	113931.57	1
4	15701354	Boni	699	France	Female	39	1	0.0	2	0	93826.63	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	79084.1	0

3.Simple Aggregations

Calucate sum and avg of balance amount of all customers

In 2 minutes (1s) 12

```python
from pyspark.sql.functions import sum, avg
# 1. Total Balance Amount
total_balance = credit_df.agg(sum("Balance").alias("TotalBalance"))
total_balance.show()

# 2. Average of Balance Amount
average_balance = credit_df.agg(avg("Balance").alias("AverageBalance"))
average_balance.show()
```

(4) Spark Jobs

total_balance: pyspark.sql.dataframe.DataFrame = [TotalBalance: double]

average_balance: pyspark.sql.dataframe.DataFrame = [AverageBalance: double]

```
+------------------+
|      TotalBalance|
+------------------+
|7.648588928799961E8|
+------------------+

+-----------------+
|   AverageBalance|
+-----------------+
|76485.88928799961|
+-----------------+
```

4.Group by

Calculate average transaction amount of each location

In 2 minutes (1s) 12

```python
# Group by Geography and calculate the average transaction amount using DataFrame API
avg_transaction_by_region = credit_df.groupBy("Geography").avg("Balance").alias("avg_spent")

# Show the results
avg_transaction_by_region.show()
```

(2) Spark Jobs

avg_transaction_by_region: pyspark.sql.dataframe.DataFrame = [Geography: string, avg(Balance): double]

```
+---------+-----------------+
|Geography|     avg(Balance)|
+---------+-----------------+
|  Germany|119730.11613391782|
|   France|  62092.6365157559|
|    Spain| 61818.14776342349|
+---------+-----------------+
```