databricks ETL transformations

(https://databricks.com)

Extract

```
#https://docs.databricks.com/external-data/sql-server.html
hostname = "clientserverondemand.database.windows.net"
port = 1433
database = "ClientData"
user = "sqladmin@clientserverondemand"
password = "Admin@098"
driver = "com.microsoft.sqlserver.jdbc.jdbc.SQLServerDriver"
URL = f"jdbc:sqlserver://{hostname}:{port};database = {database};user = {user}; password = {password}"


#read product table - Dimensions table
df_product = spark.read.format("jdbc").option("url",URL).option("dbtable","SalesLT.Product").load()


#read sales table - Fact table
df_sales = spark.read.format("jdbc").option("url",URL).option("dbtable","SalesLT.SalesOrderDetail").load()
```

Info about datasets

```
display(df_product)
```

Table

	ProductID	Name	ProductNumber	Color	StandardCost	ListPrice	Size
1	680	HL Road Frame - Black, 58	FR-R92B-58	Black	1059.3100	1431.5000	58
2	706	HL Road Frame - Red, 58	FR-R92R-58	Red	1059.3100	1431.5000	58
3	707	Sport-100 Helmet, Red	HL-U509-R	Red	13.0863	34.9900	null
4	708	Sport-100 Helmet, Black	HL-U509	Black	13.0863	34.9900	null
5	709	Mountain Bike Socks, M	SO-B909-M	White	3.3963	9.5000	M
6	710	Mountain Bike Socks, L	SO-B909-L	White	3.3963	9.5000	L
	711	Sport-100 Helmet, Blue	HL-U509-B	Blue	13.0863	34.9900	null

295 rows

```
display(df_sales)
```

Table

	SalesOrderID	SalesOrderDetailID	OrderQty	ProductID	UnitPrice	UnitPriceDiscount	LineTotal	row
1	71774	110562	1	836	356.8980	0.0000	356.898000	E3A
2	71774	110563	1	822	356.8980	0.0000	356.898000	5C1
3	71776	110567	1	907	63.9000	0.0000	63.900000	6DE
4	71780	110616	4	905	218.4540	0.0000	873.816000	377
5	71780	110617	2	983	461.6940	0.0000	923.388000	43A
6	71780	110618	6	988	112.9980	0.4000	406.792800	127

7	71780	110619	2	748	818.7000	0.0000	1637.400000	B12F0
8	71780	110620	1	990	323.9940	0.0000	323.994000	F117A
9	71780	110621	1	926	149.8740	0.0000	149.874000	92E50
10	71780	110622	1	743	809.7600	0.0000	809.760000	8BD33
11	71780	110623	4	782	1376.9940	0.0000	5507.976000	68699
12	71780	110624	2	918	158.4300	0.0000	316.860000	82940
13	71780	110625	4	780	1391.9940	0.0000	5567.976000	644B0
14	71780	110626	1	937	48.5940	0.0000	48.594000	7F5FE
15	71780	110627	6	867	41.9940	0.0000	251.964000	AC788
16	71780	110628	1	985	112.9980	0.4000	67.798800	2C10A
17	71780	110629	2	989	323.9940	0.0000	647.988000	654FB

542 rows

```
df_product.dtypes

Out[16]: [('ProductID', 'int'),
 ('Name', 'string'),
 ('ProductNumber', 'string'),
 ('Color', 'string'),
 ('StandardCost', 'decimal(19,4)'),
 ('ListPrice', 'decimal(19,4)'),
 ('Size', 'string'),
 ('Weight', 'decimal(8,2)'),
 ('ProductCategoryID', 'int'),
 ('ProductModelID', 'int'),
 ('SellStartDate', 'timestamp'),
 ('SellEndDate', 'timestamp'),
 ('DiscontinuedDate', 'timestamp'),
 ('ThumbNailPhoto', 'binary'),
 ('ThumbnailPhotoFileName', 'string'),
 ('rowguid', 'string'),
 ('ModifiedDate', 'timestamp')]


df_sales.dtypes

Out[17]: [('SalesOrderID', 'int'),
 ('SalesOrderDetailID', 'int'),
 ('OrderQty', 'smallint'),
 ('ProductID', 'int'),
 ('UnitPrice', 'decimal(19,4)'),
 ('UnitPriceDiscount', 'decimal(19,4)'),
 ('LineTotal', 'decimal(38,6)'),
 ('rowguid', 'string'),
 ('ModifiedDate', 'timestamp')]
```

Transform

```
#Keeping necessary columns
df_products_clean = df_product[['ProductID','Weight','Name','Size']]


#Removing nulls with some value
df_products_clean1 = df_products_clean.na.fill({"Size":50,"Weight":100})
display(df_products_clean1)
```

Table

	ProductID	Weight	Name	Size
1	680	1016.04	HL Road Frame - Black, 58	58
2	706	1016.04	HL Road Frame - Red, 58	58
3	707	100.00	Sport-100 Helmet, Red	50
4	708	100.00	Sport-100 Helmet, Black	50
5	709	100.00	Mountain Bike Socks, M	M
6	710	100.00	Mountain Bike Socks, L	L
7	711	100.00	Sport-100 Helmet, Blue	50

295 rows

```
#Removing duplicates from sales
df_sales_cleaned = df_sales.dropDuplicates()
display(df_sales_cleaned)
```

Table

	SalesOrderID	SalesOrderDetailID	OrderQty	ProductID	UnitPrice	UnitPriceDiscount	LineTotal	row
1	71796	111034	1	900	200.0520	0.0000	200.052000	C54

2	71856	112332	1	945	54.8940	0.0000	54.894000	47339
3	71783	110738	13	974	986.5742	0.0200	12568.955308	AC5C5
4	71784	110793	8	873	1.3740	0.0000	10.992000	2B64C
5	71935	113219	3	874	5.3940	0.0000	16.182000	040CB
6	71783	110711	6	939	37.2540	0.0000	223.524000	49FF2
7	71797	111048	4	715	29.9940	0.0000	119.976000	14C47
8	71815	111452	2	835	356.8980	0.0000	713.796000	DAEB0
9	71784	110783	3	896	200.0520	0.0000	600.156000	EC43C
10	71797	111050	15	884	29.6945	0.0500	423.146625	59634
11	71832	111062	13	867	40.5942	0.0200	517.170108	105F8
12	71902	112957	7	783	1376.9940	0.0000	9638.958000	BFDE6
13	71902	112978	4	876	72.0000	0.0000	288.000000	7C177
14	71783	110747	10	865	38.1000	0.0000	381.000000	9D1DF
15	71784	110708	8	880	32.9940	0.0000	296.946000	75F21
16	71797	111066	2	795	1466.0100	0.0000	2932.020000	45F21
17	71897	112904	1	864	38.1000	0.0000	38.100000	ECE19

542 rows

```
#Keeping necessary columns
df_sales_cleaned2 = df_sales_cleaned[['ProductID','OrderQty','UnitPrice','UnitPriceDiscount']]
df_sales_cleaned3 = df_sales_cleaned2.withColumnRenamed('ProductID','ProductID_sales')
display(df_sales_cleaned3)
```

Table

	ProductID_sales	OrderQty	UnitPrice	UnitPriceDiscount
1	900	1	2.2940	0.0000
2	948	1	54.8940	0.0000
3	974	13	986.5742	0.0200
4	873	8	1.3740	0.0000
5	874	3	5.3940	0.0000
6	939	6	37.2540	0.0000
7	715	4	29.9940	0.0000

542 rows

```
#Joining both tables
df_join = df_sales_cleaned3.join(df_products_clean1,df_sales_cleaned3.ProductID_sales == df_products_clean1.ProductID)
df_join
display(df_join)
```

Table

	ProductID_sales	OrderQty	UnitPrice	UnitPriceDiscount	ProductID	Weight	Name
1	858	6	14.6940	0.0000	858	100.00	Half-Finger Gloves, S
2	858	2	14.6940	0.0000	858	100.00	Half-Finger Gloves, S
3	858	4	14.6940	0.0000	858	100.00	Half-Finger Gloves, S
4	808	1	26.7240	0.0000	808	100.00	LL Mountain Handlebars
5	808	4	26.7240	0.0000	808	100.00	LL Mountain Handlebars
6	808	1	26.7240	0.0000	808	100.00	LL Mountain Handlebars
7	808	1	26.7240	0.0000	808	100.00	LL Mountain Handlebars

542 rows

```
df_agg = df_join.groupBy(['ProductID']).sum("Weight")
display(df_agg)
```

Table

	ProductID	sum(Weight)
1	858	300.00
2	808	400.00
3	883	800.00
4	799	16447.18
5	970	49223.60
6	918	5370.52
7	961	78298.68

142 rows

Load

```
#Mount location in ADLS
dbutils.fs.mount(
  source = "wasbs://finalstoragecontainer@finalstorageaccounts.blob.core.windows.net",
  mount_point = "/mnt/blob-storage",
  extra_configs = {"fs.azure.account.key.finalstorageaccounts.blob.core.windows.net":
                   "ozrTTmdkGmBut1qX597ZZMASvxAORaV/EmtqHgSZPC2CaenFyZN86hya6ieNlutJqRMbl9+8oBGg+AStCTUAEw=="
                  })
```

```
Out[48]: True
```

```
dbutils.fs.ls("/mnt/blob-storage")
```

```
Out[49]: []
```