me19b190-assignment-1-q1

March 12, 2023

[1]:
```
pip install pyspark
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Collecting pyspark
  Downloading pyspark-3.3.2.tar.gz (281.4 MB)
                             281.4/281.4
MB 4.8 MB/s eta 0:00:00
  Preparing metadata (setup.py) … done
Collecting py4j==0.10.9.5
  Downloading py4j-0.10.9.5-py2.py3-none-any.whl (199 kB)
                            199.7/199.7 KB
20.0 MB/s eta 0:00:00
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) … done
  Created wheel for pyspark: filename=pyspark-3.3.2-py2.py3-none-any.whl
size=281824025
sha256=19643b0440f5a2a923fea47d299d590802f5d6368a637c35d213d3c3c1090877
  Stored in directory: /root/.cache/pip/wheels/6c/e3/9b/0525ce8a69478916513509d4
3693511463c6468db0de237c86
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.5 pyspark-3.3.2
```

[54]:
```
from pyspark.sql.functions import *
from pyspark.sql import SparkSession
import sys
import numpy as np

spark = SparkSession.builder.appName("SCD").getOrCreate()

customer_data=[(1,'Harsha','20-08-1990','01-01-1970','12-12-9999'),(2,'Goldie','11-02-1990','0
cols = ['id','name','dob','validity_start','validity_end']
customer_data_df = spark.createDataFrame(data = customer_data, schema =  cols)
curr_date = '12-03-2023'

updates=[['Harsha','05-09-1990']]
```

```
updates_df = spark.createDataFrame(data = updates,schema = ['name',␣
 ↪'updated_dob'])

new_record_toappend = updates_df.join(customer_data_df, on = 'name', how =␣
 ↪'inner')
#drop the dob column as the value in the updated_dob column(which is from␣
 ↪source data) needs to be appended to the existing dob
#rename the column updated_dob to dob
#change the value of the validity start date to 12-03-2023(current date)
new_record_toappend = new_record_toappend.withColumn('validity_start',␣
 ↪lit(curr_date))
new_record_toappend = new_record_toappend.drop('dob').
 ↪withColumnRenamed('updated_dob', 'dob')
new_record_toappend.show()
```

```
+------+----------+---+--------------+------------+
|  name|       dob| id|validity_start|validity_end|
+------+----------+---+--------------+------------+
|Harsha|05-09-1990|  1|    12-03-2023|  12-12-9999|
+------+----------+---+--------------+------------+
```

[55]:
```
new_record_toappend = new_record_toappend.select("name", "id", "dob",␣
 ↪"validity_start", "validity_end")
modifed_customerdata = customer_data_df.join(updates_df, 'name', 'left_outer')


#Change the validity end date to 12-03-2023 for the previous record
#so when the left outer join has null values in the updated dob it means that␣
 ↪the records should not to be modified.
#Only the case where we have updated_dob values which are not null, we have to␣
 ↪modify the validity end to curr_date
modifed_customerdata = modifed_customerdata.withColumn('validity_end',␣
 ↪when(modifed_customerdata['updated_dob'].isNotNull(),curr_date).
 ↪otherwise(modifed_customerdata['validity_end']))
#drop the updated dob column as it is redundant and not necessary for final␣
 ↪table creation
modifed_customerdata = modifed_customerdata.drop('updated_dob')

print("Modification to existing customer data:")
modifed_customerdata.show()
```

```
Modification to existing customer data:
+------+---+----------+--------------+------------+
|  name| id|       dob|validity_start|validity_end|
+------+---+----------+--------------+------------+
|Harsha|  1|20-08-1990|    01-01-1970|  12-03-2023|
```

```
|Goldie|  2|11-02-1990|    01-01-1970|  12-12-9999|
| Divya|  3|25-12-1990|    01-01-1970|  12-12-9999|
+------+---+----------+--------------+------------+
```

[56]:
```
#append the new record to existing customer data
print("Initial customer data:")
customer_data_df.show()
print("Final SCD type 2 table: ")
finaltable = modifed_customerdata.union(new_record_toappend)
finaltable.show()
```

```
Initial customer data:
+---+------+----------+--------------+------------+
| id|  name|       dob|validity_start|validity_end|
+---+------+----------+--------------+------------+
|  1|Harsha|20-08-1990|    01-01-1970|  12-12-9999|
|  2|Goldie|11-02-1990|    01-01-1970|  12-12-9999|
|  3| Divya|25-12-1990|    01-01-1970|  12-12-9999|
+---+------+----------+--------------+------------+

Final SCD type 2 table:
+------+---+----------+--------------+------------+
|  name| id|       dob|validity_start|validity_end|
+------+---+----------+--------------+------------+
|Harsha|  1|20-08-1990|    01-01-1970|  12-03-2023|
|Goldie|  2|11-02-1990|    01-01-1970|  12-12-9999|
| Divya|  3|25-12-1990|    01-01-1970|  12-12-9999|
|Harsha|  1|05-09-1990|    12-03-2023|  12-12-9999|
+------+---+----------+--------------+------------+
```