CS4830 – Big Data Laboratory

Assignment Lab 5

Ishan Chokshi – BE19B018

January-May 2023

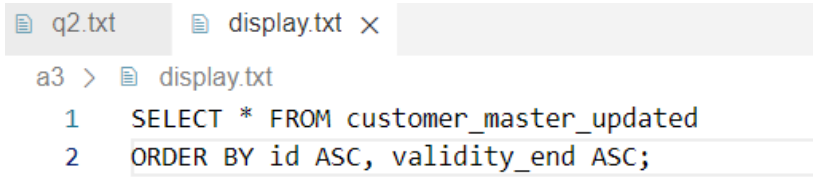
Indian Institute of Technology Madras

Ishan Chokshi
BE19B018

1. **Write PySpark code to implement SCD Type II on the givensample customer master dataframe.**
   Relevant code was written and comments were added. I have tried to keep the code structure similar to the sample python code sent. Comments are slightly modified. Functions were created to carry out the various updates. The code was run on google cloud shell. Snapshot of output is attached below:

```
be19b018@cloudshell:~/a3 (be19b018)$ python3 q1.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/03/12 14:47:10 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
+---+------+----------+-------------+-----------+
| id|  name|       dob|validity_start|validity_end|
+---+------+----------+-------------+-----------+
|  1|Harsha|20-08-1990|   01-01-1970|  12-03-2023|
|  2|Goldie|11-02-1990|   01-01-1970|  12-12-9999|
|  3| Divya|25-12-1990|   01-01-1970|  12-12-9999|
|  1|Harsha|05-09-1990|   12-03-2023|  12-12-9999|
+---+------+----------+-------------+-----------+
```

2. **Write SparkSQL code to implement SCD Type II on the given sample customer master dataframe**.
   The SQL query was written in a text (q2.txt) file and run on google cloud shell using the spark-sql command. To display the updated table, I created a separate SQL query in another txt file. The screenshot of the text file is show below. I added a few lines to ensure the id values are shown in ascending order. Snapshots are attached below:




```
be19b018@cloudshell:~/a3 (be19b018)$ spark-sql -f display.txt
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
23/03/12 17:07:15 WARN NativeCodeLoader: Unable to load native-hadoop
23/03/12 17:07:18 WARN HiveConf: HiveConf of name hive.stats.jdbc.time
23/03/12 17:07:18 WARN HiveConf: HiveConf of name hive.stats.retries.
23/03/12 17:07:21 WARN ObjectStore: Version information not found in
23/03/12 17:07:21 WARN ObjectStore: setMetaStoreSchemaVersion called
Spark master: local[*], Application Id: local-1678640837459
23/03/12 17:07:24 WARN SessionState: METASTORE_FILTER_HOOK will be ig
1       Harsha  20-08-1990      01-01-1970      12-03-2023
1       Harsha  05-09-1990      12-03-2023      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
Time taken: 4.785 seconds, Fetched 4 row(s)
be19b018@cloudshell:~/a3 (be19b018)$
```