TASK 4

Script to calculate the moving average and standard deviation of the last 10 transactions for each card_id for the data present in Hadoop and NoSQL database. If the total number of transactions for a particular card_id is less than 10, then calculate the parameters based on the total number of records available for that card_id. The script should be able to extract and feed the other relevant data ('postcode', 'transaction_dt', 'score', etc.) for the look-up table along with card_id and UCL. **(PreAnalysis.pdf)**

1: Loading data into dataframes and performing necessary transformation

Dealing with CARD_TRANSACTIONS.csv

In [8]:
```
spark.conf.set("spark.sql.legacy.timeParserPolicy", "LEGACY")
```

In [9]:
```
card_transaction_schema = StructType([
    StructField("card_id", StringType(), True),
    StructField("member_id", StringType(), True),
    StructField("amount", IntegerType(), True),
    StructField("postcode", StringType(), True),
    StructField("pos_id", StringType(), True),
    StructField("transaction_date", StringType(), True),
    StructField("status", StringType(), True)
])
```

In [11]:
```
card_transactions= spark.read \
            .format("csv") \
            .option("header","true") \
            .schema(card_transaction_schema) \
            .load("s3://aws-logs-891377364898-us-east-1/capstone/card_transactions.csv")
```

In [12]:
```
card_transactions.show(5)
```

▸ Spark Job Progress

```
+---------------+---------------+-------+--------+---------------+-------------------+-------+
|        card_id|      member_id| amount|postcode|         pos_id|   transaction_date| status|
+---------------+---------------+-------+--------+---------------+-------------------+-------+
|348702330256514|000037495066290|9084849|   33946|614677375609919|11-02-2018 00:00:00|GENUINE|
|348702330256514|000037495066290| 330148|   33946|614677375609919|11-02-2018 00:00:00|GENUINE|
|348702330256514|000037495066290| 136052|   33946|614677375609919|11-02-2018 00:00:00|GENUINE|
|348702330256514|000037495066290|4310362|   33946|614677375609919|11-02-2018 00:00:00|GENUINE|
|348702330256514|000037495066290|9097094|   33946|614677375609919|11-02-2018 00:00:00|GENUINE|
+---------------+---------------+-------+--------+---------------+-------------------+-------+
```

In [13]:
```
from pyspark.sql.functions import to_timestamp

card_transactions = card_transactions.withColumn(
    "transaction_date",
    to_timestamp("transaction_date", "MM-dd-yyyy HH:mm:ss")
)
```

In [14]:
```
df_transactions=card_transactions.filter(card_transactions.status !="FRAUD")
```

In [15]:
```
df_transactions.show(5)
```

▸ Spark Job Progress

```
+---------------+---------------+-------+--------+---------------+-------------------+-------+
|        card_id|      member_id| amount|postcode|         pos_id|   transaction_date| status|
+---------------+---------------+-------+--------+---------------+-------------------+-------+
|348702330256514|000037495066290|9084849|   33946|614677375609919|2018-11-02 00:00:00|GENUINE|
|348702330256514|000037495066290| 330148|   33946|614677375609919|2018-11-02 00:00:00|GENUINE|
|348702330256514|000037495066290| 136052|   33946|614677375609919|2018-11-02 00:00:00|GENUINE|
|348702330256514|000037495066290|4310362|   33946|614677375609919|2018-11-02 00:00:00|GENUINE|
|348702330256514|000037495066290|9097094|   33946|614677375609919|2018-11-02 00:00:00|GENUINE|
+---------------+---------------+-------+--------+---------------+-------------------+-------+
only showing top 5 rows
```

2:Creating UCL

Calculating UCL

In [18]:
```
from pyspark.sql import functions as f
```

In [19]:
```
window=Window.partitionBy(df_transactions.card_id).orderBy(df_transactions.transaction_date.desc())
```

In [20]:
```
ucl_df=df_transactions.select("*",f.rank().over(window).alias("rank")).filter(f.col('rank')<=10)
```

In [21]:
```
ucl_df.show()
```

▸ Spark Job Progress

```
+---------------+---------------+-------+--------+---------------+-------------------+-------+----+
|        card_id|      member_id| amount|postcode|         pos_id|   transaction_date| status|rank|
+---------------+---------------+-------+--------+---------------+-------------------+-------+----+
|340028465709212|009250698176266|8696557|   24658|246987608008994|2018-02-01 03:25:35|GENUINE|   1|
|340028465709212|009250698176266|9291309|   31322|057678222018909|2017-12-08 08:29:54|GENUINE|   2|
|340028465709212|009250698176266|8370505|   84056|799222285691793|2017-12-07 02:51:29|GENUINE|   3|
|340028465709212|009250698176266|6503191|   84776|072517912051441|2017-09-11 07:18:21|GENUINE|   4|
|340028465709212|009250698176266|1325446|   82630|084349436605018|2017-09-03 07:48:26|GENUINE|   5|
|340028465709212|009250698176266|8884049|   25537|553994467681624|2017-07-10 09:17:12|GENUINE|   6|
|340028465709212|009250698176266|9687739|   51542|498255481028848|2017-05-07 11:05:55|GENUINE|   7|
|340028465709212|009250698176266|6019037|   42718|725531606326617|2017-04-01 05:54:11|GENUINE|   8|
|340028465709212|009250698176266|2588786|   35201|740589158553178|2017-03-03 06:05:33|GENUINE|   9|
|340028465709212|009250698176266|7381252|   63951|305578834263571|2017-01-06 19:31:24|GENUINE|  10|
|340054675199675|835873341185231|9728785|   77373|901111066157760|2018-10-01 02:47:11|GENUINE|   1|
|340054675199675|835873341185231|2223104|   35973|512092007363081|2018-09-01 10:59:10|GENUINE|   2|
|340054675199675|835873341185231|7914699|   41844|268399801399464|2017-12-12 07:04:51|GENUINE|   3|
|340054675199675|835873341185231|2801001|   27972|770952296643375|2017-12-05 02:41:20|GENUINE|   4|
|340054675199675|835873341185231|5639516|   15736|824389119218918|2017-11-03 00:00:00|GENUINE|   5|
|340054675199675|835873341185231|2515202|   93648|929096651859256|2017-10-03 22:25:49|GENUINE|   6|
|340054675199675|835873341185231|2790716|   97057|503357356462201|2017-08-09 06:33:31|GENUINE|   7|
|340054675199675|835873341185231|9574653|   37885|498474057594510|2017-07-10 18:03:53|GENUINE|   8|
|340054675199675|835873341185231|4803251|   56164|550820714242247|2017-07-08 02:05:43|GENUINE|   9|
|340054675199675|835873341185231|7573707|   12024|115234873204918|2017-06-12 08:52:38|GENUINE|  10|
+---------------+---------------+-------+--------+---------------+-------------------+-------+----+
only showing top 20 rows
```

```
In [22]: ucl_df_moving_average=ucl_df.groupBy("card_id").agg(f.round(f.avg("amount"),2).alias("moving_avg"),
                                                            f.round(f.stddev("amount"),2).alias("Std_dev"))

In [23]: ucl_df_moving_average.show(5)
```

▸ Spark Job Progress

```
+---------------+----------+----------+
|        card_id|moving_avg|   Std_dev|
+---------------+----------+----------+
|340028465709212| 6874787.1|2860406.31|
|340054675199675| 5556463.4|2969500.07|
|340082915339645| 6015770.9|3667804.91|
|340134186926007| 4463173.5| 3144603.5|
|340265728490548| 5186170.6|3578397.38|
+---------------+----------+----------+
only showing top 5 rows
```

```
In [24]: ucl_df_final= ucl_df_moving_average.withColumn("UCL",col("moving_avg") + 3*col("Std_dev"))

In [26]: ucl_df_final=ucl_df_final.select("card_id","UCL")

In [27]: ucl_df_final.show(5)
```

▸ Spark Job Progress

```
+---------------+--------------------+
|        card_id|                 UCL|
+---------------+--------------------+
|340028465709212|       1.545600603E7|
|340054675199675|       1.446496361E7|
|340082915339645|1.7019185630000003E7|
|340134186926007|         1.3896984E7|
|340265728490548|       1.592136274E7|
+---------------+--------------------+
only showing top 5 rows
```

3: Getting Latest Transaction Records

From UCL_DF we have latest transactions with rank 1

In [28]:
```
ucl_df.show()
```

▸ Spark Job Progress

```
+---------------+---------------+-------+--------+---------------+-------------------+-------+----+
|        card_id|      member_id| amount|postcode|         pos_id|   transaction_date| status|rank|
+---------------+---------------+-------+--------+---------------+-------------------+-------+----+
|340028465709212|009250698176266|8696557|   24658|246987608008994|2018-02-01 03:25:35|GENUINE|   1|
|340028465709212|009250698176266|9291309|   31322|057678222018909|2017-12-08 08:29:54|GENUINE|   2|
|340028465709212|009250698176266|8370505|   84056|799222285691793|2017-12-07 02:51:29|GENUINE|   3|
|340028465709212|009250698176266|6503191|   84776|072517912051441|2017-09-11 07:18:21|GENUINE|   4|
|340028465709212|009250698176266|1325446|   82630|084349436605018|2017-09-03 07:48:26|GENUINE|   5|
|340028465709212|009250698176266|8884049|   25537|553994467681624|2017-07-10 09:17:12|GENUINE|   6|
|340028465709212|009250698176266|9687739|   51542|498255481028848|2017-05-07 11:05:55|GENUINE|   7|
|340028465709212|009250698176266|6019037|   42718|725531606326617|2017-04-01 05:54:11|GENUINE|   8|
|340028465709212|009250698176266|2588786|   35201|740589158553178|2017-03-03 06:05:33|GENUINE|   9|
|340028465709212|009250698176266|7381252|   63951|305578834263571|2017-01-06 19:31:24|GENUINE|  10|
|340054675199675|835873341185231|9728785|   77373|901111066157760|2018-10-01 02:47:11|GENUINE|   1|
|340054675199675|835873341185231|2223104|   35973|512092007363081|2018-09-01 10:59:10|GENUINE|   2|
|340054675199675|835873341185231|7914699|   41844|268399801399464|2017-12-12 07:04:51|GENUINE|   3|
|340054675199675|835873341185231|2801001|   27972|770952296643375|2017-12-05 02:41:20|GENUINE|   4|
|340054675199675|835873341185231|5639516|   15736|824389119218918|2017-11-03 00:00:00|GENUINE|   5|
|340054675199675|835873341185231|2515202|   93648|929096651859256|2017-10-03 22:25:49|GENUINE|   6|
|340054675199675|835873341185231|2790716|   97057|503357356462201|2017-08-09 06:33:31|GENUINE|   7|
|340054675199675|835873341185231|9574653|   37885|498474057594510|2017-07-10 18:03:53|GENUINE|   8|
|340054675199675|835873341185231|4803251|   56164|550820714242247|2017-07-08 02:05:43|GENUINE|   9|
|340054675199675|835873341185231|7573707|   12024|115234873204918|2017-06-12 08:52:38|GENUINE|  10|
+---------------+---------------+-------+--------+---------------+-------------------+-------+----+
only showing top 20 rows
```

In [29]:
```
latest_transaction= ucl_df.filter(ucl_df.rank==1)
```

In [31]:
```
latest_transaction.show()
```

▸ Spark Job Progress

```
+---------------+---------------+-------+--------+---------------+-------------------+-------+----+
|        card_id|      member_id| amount|postcode|         pos_id|   transaction_date| status|rank|
+---------------+---------------+-------+--------+---------------+-------------------+-------+----+
|340028465709212|009250698176266|8696557|   24658|246987608008994|2018-02-01 03:25:35|GENUINE|   1|
|340054675199675|835873341185231|9728785|   77373|901111066157760|2018-10-01 02:47:11|GENUINE|   1|
|340082915339645|512969555857346|5427525|   76679|578416739727226|2017-11-05 00:21:19|GENUINE|   1|
|340134186926007|887711945571282|1119931|   99769|767551517007258|2017-12-05 07:57:21|GENUINE|   1|
|340265728490548|680324265406190| 132839|   65501|716737630651927|2017-12-10 17:26:43|GENUINE|   1|
|340268219434811|929799084911715|6617581|   57030|913649766678378|2018-12-01 12:43:17|GENUINE|   1|
|340379737226464|089615510858348|4242710|   96105|285501971776349|2018-11-01 19:09:55|GENUINE|   1|
|340383645652108|181180599313885|8910036|   30634|167239770560183|2017-12-11 17:49:09|GENUINE|   1|
|340803866934451|417664728506297|6673944|   38581|686884082342988|2018-10-01 20:20:34|GENUINE|   1|
|340889618969736|459292914761635|6189940|   26202|569969474793213|2017-10-10 19:48:00|GENUINE|   1|
|340924125838453|188119365574843|8127132|   46392|234052426105135|2017-12-10 15:41:23|GENUINE|   1|
|341005627432127|872138964937565| 513792|   85352|114565288201779|2018-11-01 19:09:55|GENUINE|   1|
|341029651579925|974087224071871|6332570|   20137|785531801444257|2018-05-01 17:27:03|GENUINE|   1|
|341311317050937|561687420200207|5716682|   72166|747010723019526|2017-12-11 12:33:09|GENUINE|   1|
|341344252914274|695906467918552|7053189|   28325|202914405057619|2018-12-01 02:11:16|GENUINE|   1|
|341363858179050|009190444424572|7663138|   17082|662551622829419|2017-11-12 04:44:49|GENUINE|   1|
|341519629171378|533670008048847|9936487|   43156|295724174902824|2018-09-01 00:27:09|GENUINE|   1|
|341641153427489|230523184584316|7645102|   71254|338665811158737|2017-09-04 02:54:52|GENUINE|   1|
|341719092861087|304847505155781|1923178|   27942|704188803772359|2018-12-01 00:25:55|GENUINE|   1|
|341719092861087|304847505155781|2754624|   27942|704188803772359|2018-12-01 00:25:55|GENUINE|   1|
+---------------+---------------+-------+--------+---------------+-------------------+-------+----+
only showing top 20 rows
```

4: Joining with other tables to get the final lookuptable

```
In [34]: look_up_table=look_up_table.join(ucl_df_final,on="card_id",how="left")

In [35]: look_up_table.show(5)
```

▸ Spark Job Progress

```
+---------------+---------------+---------------+-------------------+--------+--------------------+
|        card_id|      member_id|         pos_id|   transaction_date|postcode|                 UCL|
+---------------+---------------+---------------+-------------------+--------+--------------------+
|340028465709212|009250698176266|246987608008994|2018-02-01 03:25:35|   24658|       1.54560060 3E7|
|340054675199675|835873341185231|901111066157760|2018-10-01 02:47:11|   77373|       1.44649636 1E7|
|340082915339645|512969555857346|578416739727226|2017-11-05 00:21:19|   76679|1.7019185630000003E7|
|340134186926007|887711945571282|767551517007258|2017-12-05 07:57:21|   99769|         1.3896984E7|
|340265728490548|680324265406190|716737630651927|2017-12-10 17:26:43|   65501|       1.59213627 4E7|
+---------------+---------------+---------------+-------------------+--------+--------------------+
only showing top 5 rows
```

```
In [36]: look_up_table=look_up_table.join(member_score, on="member_id",how="left")

In [37]: look_up_table.show()
```

▸ Spark Job Progress

```
+---------------+---------------+---------------+-------------------+--------+--------------------+-----+
|      member_id|        card_id|         pos_id|   transaction_date|postcode|                 UCL|score|
+---------------+---------------+---------------+-------------------+--------+--------------------+-----+
|009250698176266|340028465709212|246987608008994|2018-02-01 03:25:35|   24658|       1.54560060 3E7|  233|
|835873341185231|340054675199675|901111066157760|2018-10-01 02:47:11|   77373|       1.44649636 1E7|  631|
|512969555857346|340082915339645|578416739727226|2017-11-05 00:21:19|   76679|1.7019185630000003E7|  407|
|887711945571282|340134186926007|767551517007258|2017-12-05 07:57:21|   99769|         1.3896984E7|  614|
|680324265406190|340265728490548|716737630651927|2017-12-10 17:26:43|   65501|       1.59213627 4E7|  202|
|929799084911715|340268219434811|913649766678378|2018-12-01 12:43:17|   57030|1.5997032759999998E7|  415|
|089615510858348|340379737226464|285501971776349|2018-11-01 19:09:55|   96105|       1.43342585 9E7|  229|
|181180599313885|340383645652108|167239770560183|2017-12-11 17:49:09|   30634|       1.55777909 3E7|  645|
|417664728506297|340803866934451|686884082342988|2018-10-01 20:20:34|   38581|       1.24546391 5E7|  502|
|459292914761635|340889618969736|569969474793213|2017-10-10 19:48:00|   26202|       1.28466795 8E7|  330|
|188119365574843|340924125838453|234052426105135|2017-12-10 15:41:23|   46392|       1.30961878 4E7|  644|
|872138964937565|341005627432127|114565288201779|2018-11-01 19:09:55|   85352|       1.16091954 1E7|  629|
```

5: Final Look_up_table

In [42]:
```
look_up_table.show()
```

▸ Spark Job Progress

```
+---------------+-------------------+-----+--------+--------------------+
|        card_id|   transaction_date|score|postcode|                 UCL|
+---------------+-------------------+-----+--------+--------------------+
|340028465709212|2018-02-01 03:25:35|  233|   24658|       1.545600603E7|
|340054675199675|2018-10-01 02:47:11|  631|   77373|       1.446496361E7|
|340082915339645|2017-11-05 00:21:19|  407|   76679|1.7019185630000003E7|
|340134186926007|2017-12-05 07:57:21|  614|   99769|         1.3896984E7|
|340265728490548|2017-12-10 17:26:43|  202|   65501|       1.592136274E7|
|340268219434811|2018-12-01 12:43:17|  415|   57030|1.5997032759999998E7|
|340379737226464|2018-11-01 19:09:55|  229|   96105|       1.433425859E7|
|340383645652108|2017-12-11 17:49:09|  645|   30634|       1.557779093E7|
|340803866934451|2018-10-01 20:20:34|  502|   38581|       1.245463915E7|
|340889618969736|2017-10-10 19:48:00|  330|   26202|       1.284667958E7|
|340924125838453|2017-12-10 15:41:23|  644|   46392|       1.309618784E7|
|341005627432127|2018-11-01 19:09:55|  629|   85352|       1.160919541E7|
|341029651579925|2018-05-01 17:27:03|  682|   20137|       1.406252373E7|
|341311317050937|2017-12-11 12:33:09|  440|   72166|1.4858326579999998E7|
|341344252914274|2018-12-01 02:11:16|  643|   28325|        1.15808818E7|
|341363858179050|2017-11-12 04:44:49|  559|   17082|       1.411977277E7|
|341519629171378|2018-09-01 00:27:09|  408|   43156|         1.3783303E7|
|341641153427489|2017-09-04 02:54:52|  475|   71254|       1.3282300016E7|
|341719092861087|2018-12-01 00:25:55|  442|   27942|       1.348377687E7|
|341722035429601|2018-08-01 12:09:49|  552|   94501|1.5875301370000001E7|
+---------------+-------------------+-----+--------+--------------------+
only showing top 20 rows
```