SparkSETUPTemplate

July 3, 2019

[8]:
```
import findspark
findspark.init()
findspark.find()
import pyspark
findspark.find()
from pyspark.sql import functions as F
```

[3]:
```
from pyspark import SparkContext, SparkConf
from pyspark.sql import SparkSession
conf = pyspark.SparkConf().setAppName('Ameet').setMaster('local')
sc = pyspark.SparkContext(conf=conf)
spark = SparkSession(sc)
```

[4]:
```
nums = sc.parallelize([1,2,3,4])
nums.map(lambda x: x*x).collect()
```

[4]:
```
[1, 4, 9, 16]
```

[31]:
```
ADD="C:\\Users\\ameet.
 ↪chaubal\\Documents\\projects\\united\\address\\address\\address.csv"
ADD_HIST = "C:\\Users\\ameet.
 ↪chaubal\\Documents\\projects\\united\\address\\address_hist\\address_hist.
 ↪csv"
ADD_COLS = ["customer_id", "transaction_dtmz", "effective_dtmz"]
ADDHIST_COLS = ["customer_id", "update_id", "active_dtmz", "inactive_dtmz",␣
 ↪"effective_dtmz"]
ADD_FILTER_COL = "transaction_dtmz"
ADD_FILTER_COL2="effective_dtmz"
HIST_FILTER_COL = "active_dtmz"
HIST_FILTER_COL2 = "effective_dtmz"
ADD_S1 = "2019-06-26 00:00:00"
ADD_E1 = "2019-06-27 00:00:00"
ADD_S2 = "2019-05-31 00:00:00"
ADD_E2 = "2019-06-01 00:00:00"
```

[6]:
```
addDF=spark.read.csv(ADD, header='true')
```

[7]:
```
addDF.printSchema()
```

```
root
 |-- customer_id: string (nullable = true)
 |-- channel_type_code: string (nullable = true)
 |-- country_code: string (nullable = true)
 |-- channel_type_sequence_number: string (nullable = true)
 |-- channel_code: string (nullable = true)
 |-- address_line_1: string (nullable = true)
 |-- address_line_2: string (nullable = true)
 |-- address_line_3: string (nullable = true)
 |-- city: string (nullable = true)
 |-- apartment_number: string (nullable = true)
 |-- job_title: string (nullable = true)
 |-- postal_code: string (nullable = true)
 |-- state: string (nullable = true)
 |-- wrong_mail_date: string (nullable = true)
 |-- county: string (nullable = true)
 |-- update_dtmz: string (nullable = true)
 |-- ims_country_code: string (nullable = true)
 |-- company_name: string (nullable = true)
 |-- remark: string (nullable = true)
 |-- updated_by_id: string (nullable = true)
 |-- effective_dtmz: string (nullable = true)
 |-- discontinue_dtmz: string (nullable = true)
 |-- insert_id: string (nullable = true)
 |-- insert_dtmz: string (nullable = true)
 |-- address_verified_indicator: string (nullable = true)
 |-- last_name: string (nullable = true)
 |-- first_name: string (nullable = true)
 |-- validation_type: string (nullable = true)
 |-- validation_date: string (nullable = true)
 |-- validation_error: string (nullable = true)
 |-- validation_matchkey: string (nullable = true)
 |-- transaction_dtmz: string (nullable = true)
 |-- host: string (nullable = true)
 |-- ip_address: string (nullable = true)
 |-- host_user: string (nullable = true)
 |-- q_owner: string (nullable = true)
 |-- table: string (nullable = true)
 |-- type_cdb: string (nullable = true)
 |-- foundry_insert_timestamp_millisec: string (nullable = true)
 |-- foundry_insert_timestamp_nanosec: string (nullable = true)
```

[32]:
```
add_fil_df = addDF.filter(F.col(ADD_FILTER_COL).between(ADD_S1, ADD_E1) & F.
↪col(ADD_FILTER_COL2).between(ADD_S2, ADD_E2))
```

[33]:
```
add_fil_df = add_fil_df.select(ADD_COLS)
```

[34]:
```
add_fil_df.sort(F.col("effective_dtmz").asc()).show(truncate=False)
```

```
+-----------+-----------------------+-----------------------+
|customer_id|transaction_dtmz       |effective_dtmz         |
+-----------+-----------------------+-----------------------+
|144408644  |2019-06-26T05:01:12.000Z|2019-05-31T02:36:42.000Z|
|61021510   |2019-06-26T04:02:06.000Z|2019-05-31T03:02:28.000Z|
|80353407   |2019-06-26T17:40:05.000Z|2019-05-31T11:10:22.000Z|
|154639797  |2019-06-26T19:23:00.000Z|2019-05-31T14:45:42.000Z|
|154644013  |2019-06-26T15:37:55.000Z|2019-05-31T16:32:56.000Z|
|154648286  |2019-06-26T18:49:35.000Z|2019-05-31T19:42:49.000Z|
|154648286  |2019-06-26T18:49:35.000Z|2019-05-31T19:42:49.000Z|
|154647896  |2019-06-26T02:02:08.000Z|2019-05-31T19:57:52.000Z|
+-----------+-----------------------+-----------------------+
```

[ ]:
```

```