Reporting top 3 prescribers in each state with the most popularity

Name	Subhajit Bag
Course	Big Data

Objective:

1. Apply a filter to consider the prescribers only from 20 to 50 years of experience
2. Rank the prescribers based on their TRX_CNT for each state
3. Select the top 5 prescribers from each state

Github Repo: https://github.com/shuvo-iitkgp/PySpark-Top-3-prescribers

Project Overview:

Part 1: Creating the data pipeline:

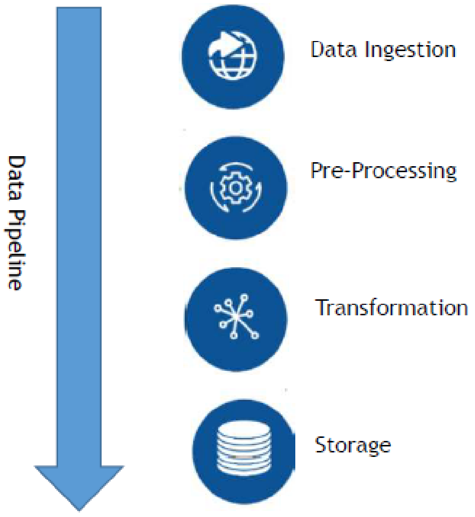
a. Copy input files to HDFS
b. Get all the variables
c. Create-objects
d. Prescriber run data ingest
e. Prescriber run data preprocessing
f. Prescriber run data transform
g. Prescriber run data extract

Part 2: Copy to local server

a. Using Microsoft Azure

Part 3: Persist

Using Hive tables and perform some operations



Integration with PySpark

```
from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \
    .master('local') \
    .appName('Testing') \
    .getOrCreate()

print("Spark Object is Created")
print(spark)
```

Creation of HDFS files

```
hdfs dfs -mkdir -p PrescPipeline/staging/dimension_city
hdfs dfs -mkdir -p PrescPipeline/staging/fact
```

Copying files to the AWS Server

```
#Step-1
### Download the Installation file using curl command.
curl "https://awscli.amazonaws.com/awscli-exe-linux-x86_64.zip" -o
"awscliv2.zip"

#Step-2
### Unzip the Installer.
unzip awscliv2.zip

#Step-3
### Run the Install Program
sudo ./aws/install

#Step-4
### Confirm the Installation
aws
aws --version
```

Persisting Data in Hive

```
### From Hive Sheel
# List Databases
show databases;

### From PySpark Shell
# List Databases
spark.sql(""" show databases""").show()
### Sample DataFrame:
sampleDF = spark.createDataFrame([('Robert',25),
                                  ('Reid',35),
                                  ('Ram',21)],
                                 ["EmpName", "EmpAge"])

### Add below property to hive-site.xml file.
/opt/hive/conf/hive-site.xml
<property>
     <name>hive.metastore.warehouse.dir</name>
     <value>/hive/user/warehouse</value>
  <description></description>
</property>

### Create Hive Table from PySpark Shell
spark.sql(""" create database prescpipeline""")
spark.sql(""" use prescpipeline""")
sampleDF.write.saveAsTable('sampleTable')
### How to create a current date column in PySpark
import datetime as date
date.datetime.now()
date.datetime.now().strftime("%Y-%m-%d")

from pyspark.sql.functions import lit
sampleDF=sampleDF.withColumn("delivery_date",
lit(date.datetime.now().strftime("%Y-%m-%d")))
sampleDF.show()

### Save the dataframe into Hive Partitioned Table
spark.sql(""" use prescpipeline""")
sampleDF.write.saveAsTable('samplePartTable',
PartitionBy='delivery_date')
```

```
### Check below at Hive prompt
use prescpipeline;
show tables;
describe formatted sampleparttable;

### Check the Hive Underlying table file at HDFS
hdfs dfs -ls
hdfs://localhost:9000/hive/user/warehouse/prescpipeline.db/samplepartta
ble
```

Integration test

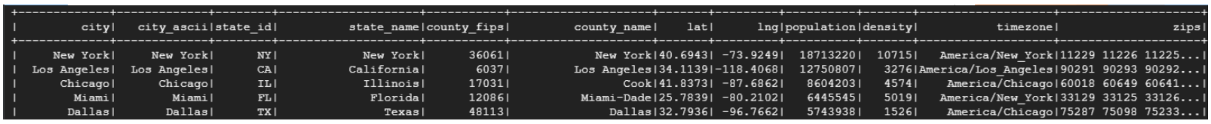
```
import unittest
first=1.345
second=1.346
decimal=2
message="First and Second inputs are not equal."
delta=0.01

class DemoTest(unittest.TestCase):
    def test_almost_equal1(self):
        self.assertAlmostEqual(first, second,None,message,delta)

if __name__ == '__main__':
    unittest.main()
```

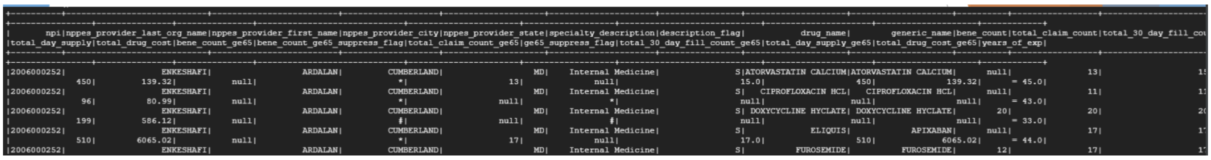
Output

Input City Dimension Layout



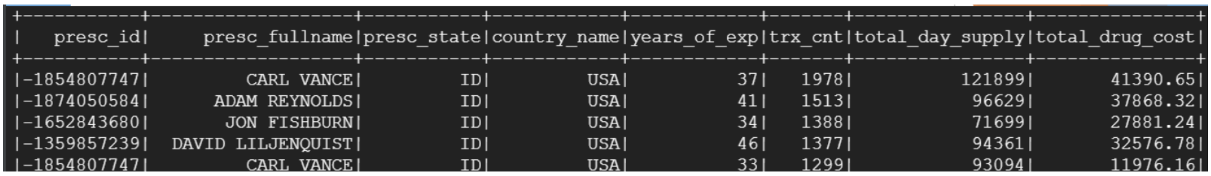
```
+------------+------------+--------+--------------------+-----------+----------------+-------+---------+----------+-------+-------------------+--------------------+
|        city|  city_ascii|state_id|          state_name|county_fips|     county_name|    lat|      lng|population|density|           timezone|                zips|
+------------+------------+--------+--------------------+-----------+----------------+-------+---------+----------+-------+-------------------+--------------------+
|    New York|    New York|      NY|            New York|      36061|        New York|40.6943| -73.9249|  18713220|  10715|   America/New_York|11229 11226 11225...|
| Los Angeles| Los Angeles|      CA|          California|       6037|     Los Angeles|34.1139|-118.4068|  12750807|   3276|America/Los_Angeles|90291 90293 90292...|
|     Chicago|     Chicago|      IL|            Illinois|      17031|            Cook|41.8373| -87.6862|   8604203|   4574|    America/Chicago|60018 60649 60641...|
|       Miami|       Miami|      FL|             Florida|      12086|      Miami-Dade|25.7839| -80.2102|   6445545|   5019|   America/New_York|33129 33125 33126...|
|      Dallas|      Dallas|      TX|               Texas|      48113|          Dallas|32.7936| -96.7662|   5743938|   1526|    America/Chicago|75287 75098 75233...|
```

Input Prescriber Fact Layout



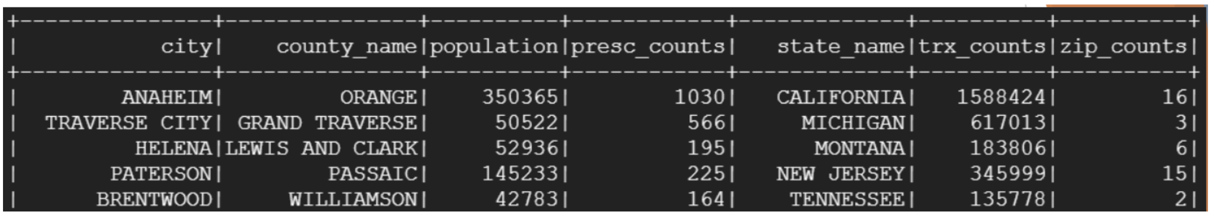
```
+----------+---------------------------+-------------------------+-------------------+--------------------+---------------------+----------------+--------------------+--------------------+----------+-----------------+--------------------
+----------------+---------------+---------------+-----------------------------+----------------------+------------------+----------------------------+---------------------+--------------------+------------+
|       npi|nppes_provider_last_org_name|nppes_provider_first_name|nppes_provider_city|nppes_provider_state|specialty_description|description_flag|           drug_name|        generic_name|bene_count|total_claim_count|total_30_day_fill_cou
|total_day_supply|total_drug_cost|bene_count_ge65|bene_count_ge65_suppress_flag|total_claim_count_ge65|ge65_suppress_flag|total_30_day_fill_count_ge65|total_day_supply_ge65|total_drug_cost_ge65|years_of_exp|
+----------+---------------------------+-------------------------+-------------------+--------------------+---------------------+----------------+--------------------+--------------------+----------+-----------------+--------------------
+----------------+---------------+---------------+-----------------------------+----------------------+------------------+----------------------------+---------------------+--------------------+------------+
|2006000252|                  ENKESHAFI|                  ARDALAN|         CUMBERLAND|                  MD|    Internal Medicine|               S|ATORVASTATIN CALCIUM|ATORVASTATIN CALCIUM|      null|               13|                  15
|             450|         139.32|           null|                            *|                    13|              null|                        15.0|                  450|              139.32|      = 45.0|
|2006000252|                  ENKESHAFI|                  ARDALAN|         CUMBERLAND|                  MD|    Internal Medicine|               S|   CIPROFLOXACIN HCL|   CIPROFLOXACIN HCL|      null|               11|                  11
|              96|          80.99|           null|                            *|                  null|                 *|                        null|                 null|                null|      = 43.0|
|2006000252|                  ENKESHAFI|                  ARDALAN|         CUMBERLAND|                  MD|    Internal Medicine|               S| DOXYCYCLINE HYCLATE| DOXYCYCLINE HYCLATE|        20|               20|                  20
|             199|         586.12|           null|                            #|                  null|                 #|                        null|                 null|                null|      = 33.0|
|2006000252|                  ENKESHAFI|                  ARDALAN|         CUMBERLAND|                  MD|    Internal Medicine|               S|             ELIQUIS|            APIXABAN|      null|               17|                  17
|             510|        6065.02|           null|                            *|                    17|              null|                        17.0|                  510|             6065.02|      = 44.0|
|2006000252|                  ENKESHAFI|                  ARDALAN|         CUMBERLAND|                  MD|    Internal Medicine|               S|          FUROSEMIDE|          FUROSEMIDE|        12|               17|                  17
```

Prescriber Report Layout



```
+-----------+-------------------+-----------+------------+------------+-------+----------------+---------------+
|   presc_id|     presc_fullname|presc_state|country_name|years_of_exp|trx_cnt|total_day_supply|total_drug_cost|
+-----------+-------------------+-----------+------------+------------+-------+----------------+---------------+
|-1854807747|         CARL VANCE|         ID|         USA|          37|   1978|          121899|       41390.65|
|-1874050584|      ADAM REYNOLDS|         ID|         USA|          41|   1513|           96629|       37868.32|
|-1652843680|       JON FISHBURN|         ID|         USA|          34|   1388|           71699|       27881.24|
|-1359857239|  DAVID LILJENQUIST|         ID|         USA|          46|   1377|           94361|       32576.78|
|-1854807747|         CARL VANCE|         ID|         USA|          33|   1299|           93094|       11976.16|
```

City Report Layout



```
+---------------+---------------+----------+------------+-------------+----------+----------+
|           city|    county_name|population|presc_counts|   state_name|trx_counts|zip_counts|
+---------------+---------------+----------+------------+-------------+----------+----------+
|        ANAHEIM|         ORANGE|    350365|        1030|   CALIFORNIA|   1588424|        16|
|  TRAVERSE CITY| GRAND TRAVERSE|     50522|         566|     MICHIGAN|    617013|         3|
|         HELENA|LEWIS AND CLARK|     52936|         195|      MONTANA|    183806|         6|
|       PATERSON|        PASSAIC|    145233|         225|   NEW JERSEY|    345999|        15|
|      BRENTWOOD|     WILLIAMSON|     42783|         164|    TENNESSEE|    135778|         2|
```