Step 1: Connect to BDCE

**The data being used is a 11.43 GB sample file from a larger file.** The first step is to connect to Oracle BDCE.

1. ssh amarro15@129.150.69.91

Step 2: Get Dataset

**The data being used is already uploaded to team mate directory.** The next step is to copy the file to personal user directory. Additional steps are given to check if code was successful.

1. hdfs dfs -mkdir /practice
2. hdfs dfs -cp /user/fmamagh2/Group5/arcos\_all\_washpost1.tsv /user/amarro15/practice/
3. hdfs dfs -ls /user/amarro15/practice
4. hdfs dfs -cat /user/amarro15/practice/arcos\_all\_washpost1.tsv | head -n 2

Step 3: Loading Data With PIG

**Pig will be used to make our large file into smaller files to make state-specific visualizations.** This step is to enter the Pig grunt shell environment and create a new relation and schema.

$ pig

data = LOAD '/user/amarro15/practice/arcos\_all\_washpost1.tsv' AS (reporter\_dea\_no:chararray, reporter\_bus\_act:chararray, reporter\_name:chararray, reporter\_addl\_co\_info:chararray, reporter\_address1:chararray, reporter\_address2:chararray, reporter\_city:chararray, reporter\_state:chararray, reporter\_zip:chararray, reporter\_county:chararray, buyer\_dea\_no:chararray, buyer\_bus\_act:chararray, buyer\_name:chararray, buyer\_addl\_co\_info:chararray, buyer\_address1:chararray, buyer\_address2:chararray, buyer\_city:chararray, buyer\_state:chararray, buyer\_zip:chararray, buyer\_county:chararray, transaction\_code:chararray, drug\_code:chararray, ndc\_no:chararray, drug\_name:chararray, quantity:int, unit:int, action\_indicator:chararray, order\_form\_no:chararray, correction\_no:chararray, strength:int, transaction\_date:chararray, calc\_base\_wt\_in\_gm:double, dosage\_unit:int, transaction\_id:chararray, product\_name:chararray, ingredient\_name:chararray, measure:chararray, mme\_conversion\_factor:int, combined\_labeler\_name:chararray, revised\_company\_name:chararray, reporter\_family:chararray, dos\_str:float);

Step 4: Filtering Data With PIG

**The file is now ready to be filtered to specific data.** The following code will create top distributors by city/state. There is the option to choose which state or drug name that the table should contain.

drug\_data = FILTER data BY drug\_name == 'OXYCODONE';

california\_subset = FILTER drug\_data BY buyer\_state == 'CA' AND reporter\_bus\_act == 'DISTRIBUTOR';

grouped = GROUP California\_subset BY reporter\_city;

totals = FOREACH grouped GENERATE group, SUM(California\_subset.quantity) AS city\_count;

sorted = ORDER totals BY city\_count DESC;

top\_ten = LIMIT sorted 100;

Step 5: Storing Data With PIG

**Last step in Pig is to store the output.** DUMP the file to check if it is working, and store the output in a .csv file. Quit pig.

DUMP top\_ten;

STORE top\_ten INTO 'output/top\_ten' USING PigStorage(',');

Quit

Step 6: Downloading Output File

**The pig output will now be moved into local filesystem and downloaded into personal laptop/desktop.** Once exiting pig, move and confirm the file is in correct order. Then download file.

hdfs dfs -get output/top\_ten/part-r-00000 top\_ten.csv

cat top\_ten.csv | tail -n 2

scp amarro15@129.150.69.91:/home/amarro15/top\_ten.csv .

Step 7: Visualization

**The last step is to upload into visualization software.** The first visualization was loaded into SAP Predictive Analytics. The second graph was created by Microsoft Excel Power Maps.

Chart

Description automatically generated

