**Big Data Spring 2018**

**Project 2**

Nicola Eldering

Yifan Chen

**Member 1: Nicolas Eldering**  
My teammate and I agree that I handled 50% of the overall project. My specific tasks included:

* Task 1: I wrote half of the Pig code for question 1.
* Task 2: I wrote the Pig code for question 3.

**Member 2: Yifan Chen**  
My teammate and I agree that I handled 50% of the overall project. My specific tasks included:

* Task 1: I wrote half of the Pig code for question 1.
* Task 2: I wrote the Pig code for question 2

Question 1:

On the average, which state has the highest amount of precipitation in September 2013?

**rainQ1.pig**

-------------------------------------------Start of rainQ1 ----------------------------------

line = LOAD'hdfs://sophia:9000/user/hduser/input/3240sep2013.dat' USING PigStorage() AS (

total:chararray

); -- bring in data from the hdfs

stateID\_precipitation = FOREACH line GENERATE SUBSTRING(total, 3, 5) as stateID, flatten(STRSPLIT(total, ' 2500 ', 2)) as (noMatter, precipitation); -- grab the state id and total rain

stateID\_precipitation\_rtrim = FOREACH stateID\_precipitation GENERATE stateID, RTRIM(precipitation) as precipitation; -- clean our precipitation data

stateID\_precipitation\_withoutI = FILTER stateID\_precipitation\_rtrim by not(precipitation matches '.\*I\\b'); -- remove unneeded I records

stateID\_rainfall = FOREACH stateID\_precipitation\_withoutI GENERATE $0, (int)(SUBSTRING($1, 0, 5)) as rainfall; -- allow us to group by state

septState = GROUP stateID\_rainfall by stateID; -- group our data by state

septClean = FOREACH septState GENERATE group as stateID, SUM(stateID\_rainfall.rainfall)/30 as rainfall;

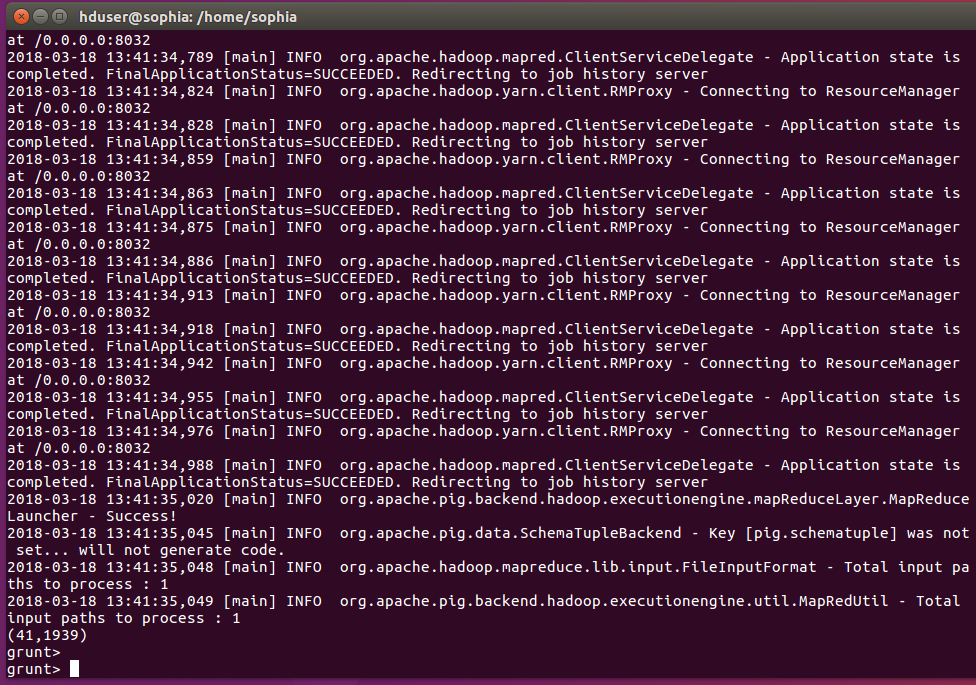
septSort = order septClean by rainfall desc; -- sort so we can grab the max

max = LIMIT septSort 1; -- grab our max result

dump max; -- show our results

---------------------------------------End of Script --------------------------

First step in our pig program is to bring in September’s rainfall data and grab the useful section. We then break it up by state and group it by stateID and calculate the average rainfall (rainfall/ days of month). We finally sort it and limit it by one to grab the state with the highest average rainfall in September.



Question 2:

Which state has the lowest precipitation in 2013?

**question2.pig**

-------------------------------------------Start of question2.pig----------------------------------

line = LOAD'hdfs://sophia:9000/user/precipitation/3240{jan,feb,mar,apr,may,jun,jul,aug,sep,oct,nov,dec}2013.dat' USING PigStorage() AS (

total:chararray

);

stateID\_precipitation = FOREACH line GENERATE SUBSTRING(total, 3, 5) as stateID, flatten(STRSPLIT(total, ' 2500 ', 2)) as (nomatter, precipitation);

stateID\_precipitation\_rtrim = FOREACH stateID\_precipitation GENERATE stateID, RTRIM(precipitation) as precipitation;

stateID\_precipitation\_withoutI = FILTER stateID\_precipitation\_rtrim by not(precipitation matches '.\*I\\b');

stateID\_rainfall = FOREACH stateID\_precipitation\_withoutI GENERATE $0, (int)(SUBSTRING($1, 0, 5)) as rainfall;

stateID\_rainfall\_group = GROUP stateID\_rainfall by stateID;

total\_precipitation = FOREACH stateID\_rainfall\_group GENERATE group as stateID, SUM(stateID\_rainfall.rainfall) as totalRainfall;

total\_precipitation\_ordered = ORDER total\_precipitation BY totalRainfall;

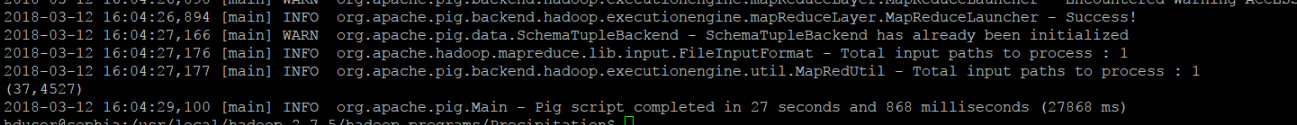
total\_precipitation\_limited = LIMIT total\_precipitation\_ordered 1;

DUMP total\_precipitation\_limited;

---------------------------------------End of Script --------------------------

First step in our pig program is to bring in the whole year’s rainfall data and grab the useful section. We then break it up by state and group it by stateID and calculate the total rainfall (SUM(stateID\_rainfall.rainfall)). We finally sort it by total rainfall in ASC and limit it by one to grab the state with the lowest precipitation in 2013.

(37, 4527)

****

Question 3:

Which states had more than an inch of precipitation on July 4, 2013?

**rainQ3.pig**

---------------------------------------Start of rainQ3.pig -----------------------

july = LOAD 'hdfs://sophia:9000/user/hduser/input/3240jul2013.csv' USING PigStorage(',') as (dtype:chararray,state:int,id:int,misc:chararray,qualifier:chararray,year:int,month:int,day:int,rainfall:int); -- Bring in the loaded scrubbed data from the hdfs

julyDay = FILTER july by day == 4; -- grab only the data for July 4th

julyState = GROUP julyDay by state; -- group our July 4th data by state

julyClean = FOREACH julyState GENERATE (int)AVG(julyDay.state), SUM(julyDay.rainfall); -- sum each july 4th for each state

results = FILTER julyClean by $1 > 100; -- find the states that have more than an inch of rain

dump results; -- show the results

--------------------------------------End of Script ----------------------------

First step in our pig program is to bring in July’s data and grab only the data for July 4th. Then we group our data by state and sum each states rainfall for July 4th. Next we filter and grab only the states who’s totals are greater than 100 hundredths of an inch. Finally we dump the results to the display.

Results:

(1, 5382) (38, 2509)

(8, 3779) (40, 3549)

(9, 2095) (41, 207)

(15, 2031) (42, 593)

(22, 244) (43, 428)

(24, 123) (44, 641)

(26, 207) (46, 408)

(29, 250) (48, 471)

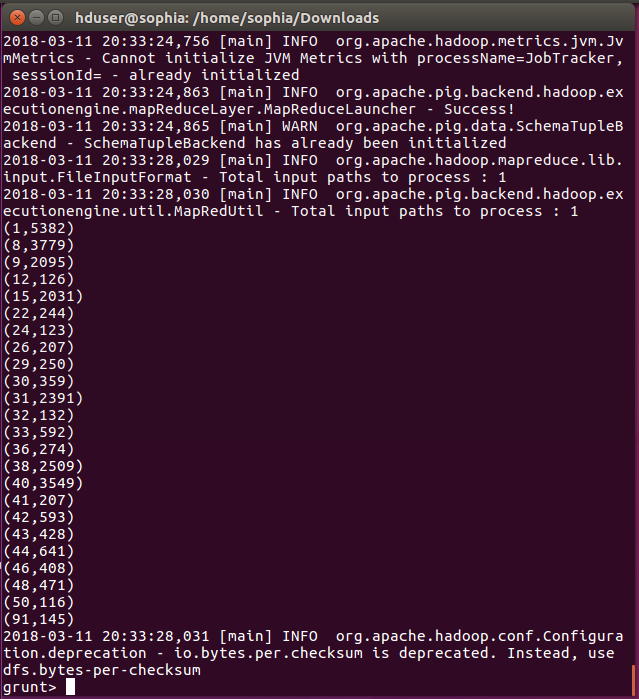
(30, 359) (50, 116)

(31, 2391) (91, 145)

(32, 132)

(33, 592)

(36, 274)

****