Homework IST769 Unit E - Schemaless SQL

## Agenda

1. Your Questions
2. Go over Problem Set
3. Unit Coursework Activities

## 1. Your Questions

Ask any questions you have here!

* How to Interactive Plotting
* Problems in setting up Drill
* Question 8

## 2. Go Over Problem Set E

1. Configure a Drill storage plugin for the Minio **labe** bucket. Then write a drill query for **syracuse-ny.csv** to demonstrate you can read the file with headers.

"extractHeader": true

Select \*

From labe.`syracuse-ny.csv`

Limit 100;

2. Write a Drill SQL Query to get the overall average min and max temperatures by year and month. Use drill’s SPLIT() function to separate Year, Month. You might need to use cast() to ensure the min and max temperatures are numeric types. You output should include 4 columns: Year, Month, the average minimum temperature for that month, and the average maximum temperature for that month.

Data:

<https://raw.githubusercontent.com/mafudge/datasets/master/weather/syracuse-ny.csv>

WITH temp\_table AS (

SELECT CAST(SPLIT(EST, '-')[0] AS INT) AS YEAR,

CAST(SPLIT(EST, '-')[1] AS INT) AS MONTH,

CAST(`Max\_TemperatureF` AS INT) AS AVG\_MAX\_TEMP,

CAST(`Min\_TemperatureF` AS INT) AS AVG\_MIN\_TEMP

FROM `labe`.`syracuse-ny.csv`

)

SELECT YEAR,

MONTH,

AVG(AVG\_MAX\_TEMP) AS AVG\_MAX\_TEMP,

AVG(AVG\_MIN\_TEMP) AS AVG\_MIN\_TEMP

FROM temp\_table

GROUP BY YEAR,MONTH

ORDER BY YEAR,MONTH

>>

WITH VT AS (

SELECT

SPLIT(EST, '-')[0] AS Year

, SPLIT(EST, '-')[1] AS Month

, CAST(Min\_TemperatureF AS DOUBLE) AS Min\_TemperatureF

, CAST(Max\_TemperatureF AS DOUBLE) AS Max\_TemperatureF

FROM labe.`syracuse-ny.csv`

)

SELECT

Year

, Month

, AVG(Min\_TemperatureF) AS Avg\_Min\_TemperatureF

, AVG(Max\_TemperatureF) AS Avg\_Max\_TemperatureF

FROM VT

GROUP BY Year, Month;

select Split(`EST`, '-')[0] as Year, Split(`EST`, '-')[1] as Month , avg(CAST(`Max\_TemperatureF` AS INT)) as The\_Average\_Maximum\_Temperature\_for\_the\_Month, avg(CAST(`Min\_TemperatureF` AS INT)) as THE\_Average\_Minimum\_Temperature\_for\_the\_Month

from minio.`syracuse-ny.csv`

group by Split(`EST`, '-')[0],Split(`EST`, '-')[1]

3. Create a view called **monthly\_syracuse\_weather\_averages** from the query you wrote in question 2 and store it back on the **labe** bucket. (If you cannot get question 2 working, use a similar query). Provide your drill SQL code and a screenshot showing the view file is on the Minio bucket.

NOTE: If you get an error about an immutable object, you need to change your storage config so you can write to the storage location.

Create view monthly\_syracuse\_weather\_averages as

with monthly\_syracuse\_weather\_averages as (

select SPLIT(`EST`,'-')[0] as Year, SPLIT(`EST`,'-')[1] as Month, AVG(CAST(`Max\_TemperatureF` as INT)) as The\_Average\_Maximum\_Temperature\_for\_the\_Month, AVG(CAST(`Min\_TemperatureF` as INT)) as THE\_Average\_Minimum\_Temperature\_for\_the\_Month

from minio.`syracuse-ny.csv`

group by SPLIT(`EST`,'-')[0],SPLIT(`EST`,'-')[1])

select \* from monthly\_syracuse\_weather\_averages

>>

​​CREATE VIEW labe.`monthly\_syracuse\_weather\_averages` AS

WITH VT AS (

SELECT

CAST(SPLIT(EST, '-')[0] AS INT) AS Year

, CAST(SPLIT(EST, '-')[1] AS INT) AS Month

, CAST(Min\_TemperatureF AS DOUBLE) AS Min\_TemperatureF

, CAST(Max\_TemperatureF AS DOUBLE) AS Max\_TemperatureF

FROM labe.`syracuse-ny.csv`

)

SELECT

Year

, Month

, AVG(Min\_TemperatureF) AS Avg\_Min\_TemperatureF

, AVG(Max\_TemperatureF) AS Avg\_Max\_TemperatureF

FROM VT

GROUP BY Year, Month;

4. Use the view you created in question 3 to show the weather data only the month of July.

>>

SELECT \*

FROM labe.`monthly\_syracuse\_weather\_averages`

WHERE Month = 7;

5. Configure spark to read from Minio **labe** bucket, then load **syracuse-ny.csv** into a DataFrame as register it as the table **weather**

6. Rewrite question 2 using pure Spark SQL and the **weather** temp view. NOTE: There will be some subtle differences with how you must write the code, so be sure to **printSchema()** so you can see what the columns are.

7. Save the output from the DataFrame in question 6 to the temp view **monthly\_syracuse\_weather\_averages**. Prove the view is there by querying it.

8. CHALLENGE YOURSELF! At the bottom of the **work/content/E-Drill-Spark.ipynb**  file there is a section Called “Big Data to Small Data”. Try to write a complete program that:

a. Inputs a month 1 – 12 at run-time.

b. Displays a scatter plot of min/max average monthly temperatures, where year is on the X-Axis.