Homework Problem Set G Submission Form

# Overview

| Your Name | Nick Videtti |
| --- | --- |
| Your SU Email | nvidetti@syr.edu |

# Instructions

Put your name and SU email at the top. Answer these questions all from the lab. When asked to include screenshots, please follow the screenshot guidelines from the first homework.

Remember as you complete the homework that it is not only about getting it right/correct. We will discuss the answers in class so it’s important to articulate anything you would like to contribute to the discussion in your answer:

* If you feel the question is vague, include any assumptions you've made.
* If you feel the answer requires interpretation or justification, provide it.
* If you do not know the answer to the question, articulate what you tried and how you are stuck.
* Highlight any doubts or questions you would like me to review.

This how you receive credit for answering questions that might not be correct. In addition, you must complete the reflection portion of the homework assignment for full credit. Since most answers will be similar this is an important part of your individual submission.

Complete Part II of this document first, then go back and complete the Reflection in Part I.

# Part I: Reflection

Use this section to reflect on your learning. To achieve the highest grade on the assignment, you must be as descriptive and personal as possible with your reflection.

1. As you completed this assignment, identify what you learned.

**This assignment really helped to show the full process of implementing a simple Cassandra table and Exercise 10 really helped me understand where a materialized view would be useful, or even just a secondary index.**

1. What barriers or challenges did you encounter while completing this assignment?

**No significant barriers as this topic seemed a bit easier than some of the others that past assignments covered. It seems keyspace, table, and field names are all stored as lowercase and that the column names in PySpark are case-sensitive, which caused a brief roadblock in Exercise 6.**

1. How prepared were you to complete this assignment? What can you do to be better prepared?

**I think knowing to refer to the asynchronous work and having the sample Jupyter notebooks on the virtual machines help quite a bit in terms of preparing for these assignments as much as we can.**

1. Rate your comfort level with this week’s material. Use the rubric provided.

4 ==> I understand this material and can explain it to others.  
**3 ==> I understand this material.**  
2 ==> I somewhat understand the material but sometimes need guidance from others.  
1 ==> I understand very little of this material and need extra help.

# Part II: Questions

Your employer (weather.com) would like you store weather sensor and forecast data. Eventually you will get readings from 2,000 cities worldwide every minute. That's 2.88 million rows each day and 1 billion rows a year! Since the data does not need to be read immediately when written across all nodes, you decide Cassandra is a good choice for this project! This data will be accessible by users so they can get weather information and historical trends for they cities they live in and visit. This should help you figure out how the data will be queried.

**QUESTIONS:   
  
For each question, include a copy of the code required to complete the question along with a screenshot of the code and a screenshot of the output.**

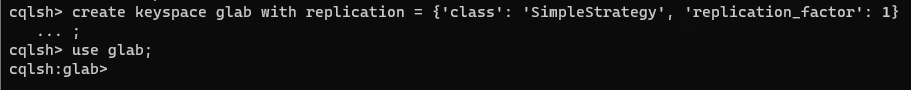
1. InCQL, create a keyspace called **glab** with a replication factor of 1 and a simple replication strategy. Use the keyspace.

**cqlsh> create keyspace glab with replication = {'class': 'SimpleStrategy', 'replication\_factor': 1}**

**... ;**

**cqlsh> use glab;**

**cqlsh:glab>**



1. In Spark, set up a Spark session that is ready to talk with Cassandra.

**import pyspark**

**spark = pyspark.sql.SparkSession \**

**.builder \**

**.master("local") \**

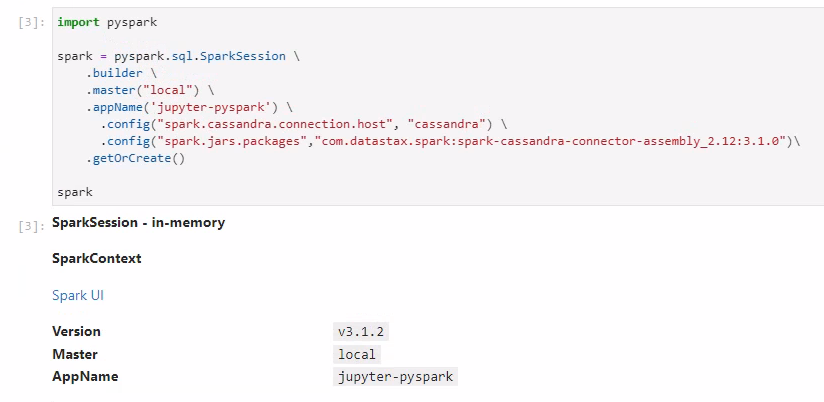
**.appName('jupyter-pyspark') \**

**.config("spark.cassandra.connection.host", "cassandra") \**

**.config("spark.jars.packages","com.datastax.spark:spark-cassandra-connector-assembly\_2.12:3.1.0")\**

**.getOrCreate()**

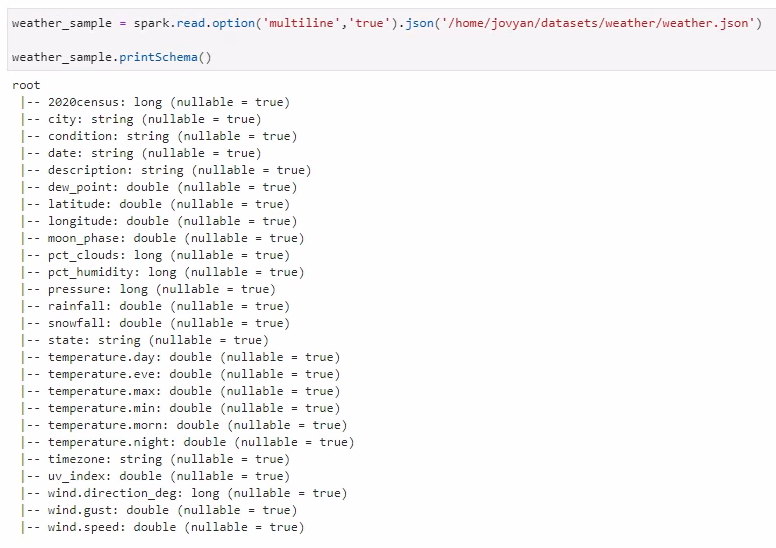
**spark**



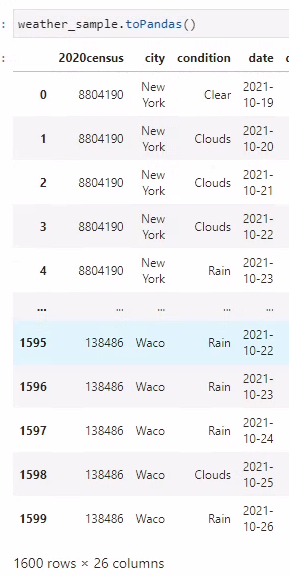
1. To deal with the amount of data associated with the weather.com data set, you decide to start with a smaller sample data set. The data set contains 7 days of weather information for major U.S. cities, with one row being weather information for a single city on a single day. Load the data set located at **/home/jovyan/datasets/weather/weather.json** and use printSchema() to inspect the schema.

**weather\_sample = spark.read.option('multiline','true').json('/home/jovyan/datasets/weather/weather.json')**

**weather\_sample.printSchema()**



1. Look at rows of data in the sample data set. Profile the data to determine what should be used as the partition and cluster key:
   1. First: Find the minimal candidate key. Which columns serve as a key for each row?   
      NOTE: You canNOT use 2020census as that is a population figure and coincidentally unique.
   2. Next: Prove your key works. In Spark:
      1. Get a count of rows in the entire DataFrame.
      2. Get a count rows when you select your key columns and use distinct() to remove duplicates.
      3. If the row counts are the name, that’s a candidate key. Include the code and output in the screenshot.
   3. A Cassandra row key consists of a partition and cluster key.   
      For this example, use the column that will guarantee to be storing data in increasing order over time (append only) as your cluster key. The other column (or columns) should be the partition key.



**The minimal candidate key appears to be a combination of city and date.**

**weather\_sample\_view = weather\_sample.createOrReplaceTempView("weather\_sample\_view")**

**spark.sql('''**

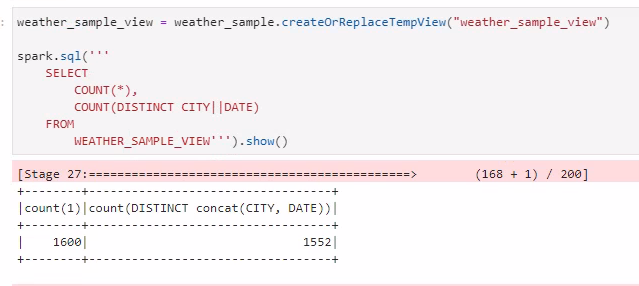
**SELECT**

**COUNT(\*),**

**COUNT(DISTINCT CITY||DATE)**

**FROM**

**WEATHER\_SAMPLE\_VIEW''').show()**

****

**Perhaps not… let’s investigate…**

**spark.sql('''**

**SELECT**

**CITY,**

**DATE,**

**COUNT(\*)**

**FROM**

**WEATHER\_SAMPLE\_VIEW**

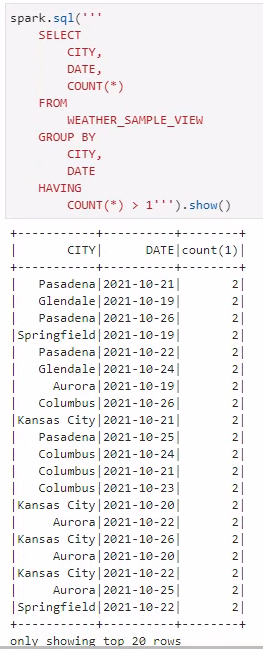
**GROUP BY**

**CITY,**

**DATE**

**HAVING**

**COUNT(\*) > 1''').show()**

****

**spark.sql('''**

**SELECT**

**\***

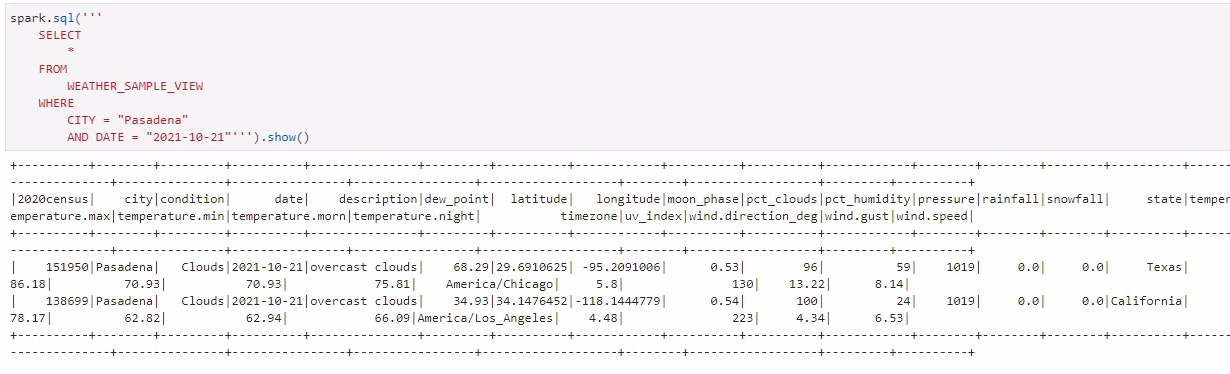
**FROM**

**WEATHER\_SAMPLE\_VIEW**

**WHERE**

**CITY = "Pasadena"**

**AND DATE = "2021-10-21"''').show()**

****

**It looks like the states are different in these two records! Maybe a good key would instead be a combination of City, State, and Date.**

**weather\_sample\_view = weather\_sample.createOrReplaceTempView("weather\_sample\_view")**

**spark.sql('''**

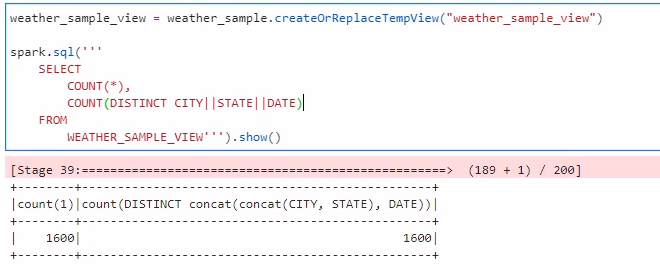
**SELECT**

**COUNT(\*),**

**COUNT(DISTINCT CITY||STATE||DATE)**

**FROM**

**WEATHER\_SAMPLE\_VIEW''').show()**

****

**Eureka!!**

**For the cluster key to sort data by date and the remaining columns to serve as the partition key, we should have the following.**

**Partition Key: (state, city)**

**Cluster Key: date**

1. With your keys figured out, it’s time to create your table. Using the CQL Shell, write an CQL query to create a table called **daily\_city\_weather**. Include all columns in the source data set, and make sure to set your partition and cluster keys, as designed. Show the CQL query and the output in the screenshot. Include an additional screenshot of the describe command on this table.  
   ADVICE: Write your create table in a text editor then paste it into CQL, as the command line can be a tad unforgiving.

**CREATE TABLE**

**DAILY\_CITY\_WEATHER**

**(STATE VARCHAR,**

**CITY VARCHAR,**

**DATE DATE,**

**CENSUS\_2020 INT,**

**CONDITION VARCHAR,**

**DESCRIPTION TEXT,**

**DEW\_POINT DOUBLE,**

**LATITUDE DOUBLE,**

**LONGITUDE DOUBLE,**

**MOON\_PHASE DOUBLE,**

**PCT\_CLOUDS DOUBLE,**

**PCT\_HUMIDITY DOUBLE,**

**PRESSURE DOUBLE,**

**RAINFALL DOUBLE,**

**SNOWFALL DOUBLE,**

**TEMPERATURE\_DAY DOUBLE,**

**TEMPERATURE\_EVE DOUBLE,**

**TEMPERATURE\_MAX DOUBLE,**

**TEMPERATURE\_MIN DOUBLE,**

**TEMPERATURE\_MORN DOUBLE,**

**TEMPERATURE\_NIGHT DOUBLE,**

**TIMEZONE VARCHAR,**

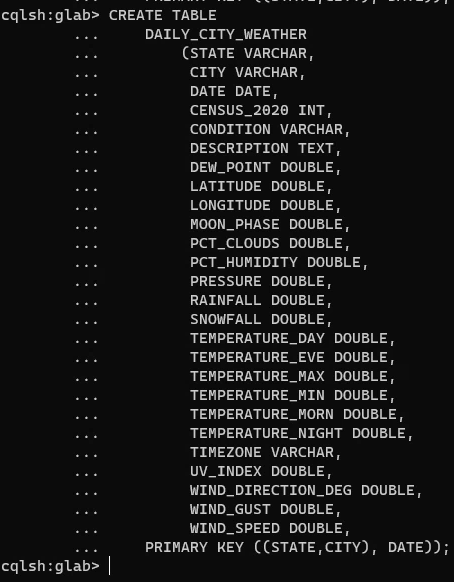
**UV\_INDEX DOUBLE,**

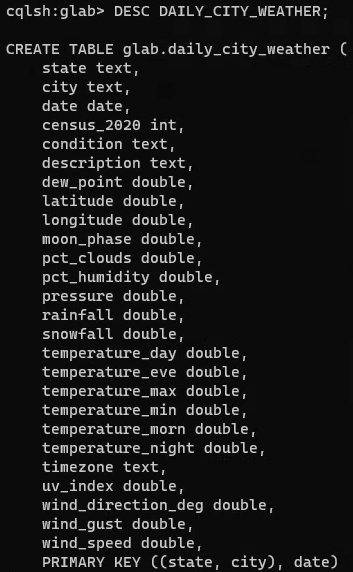
**WIND\_DIRECTION\_DEG DOUBLE,**

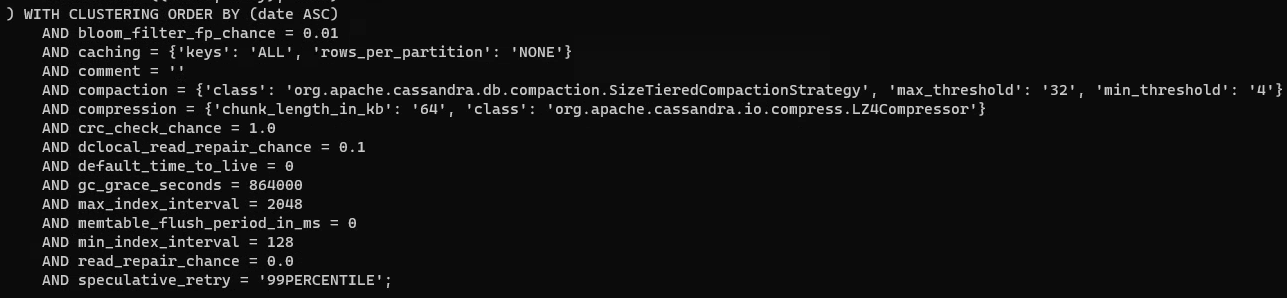
**WIND\_GUST DOUBLE,**

**WIND\_SPEED DOUBLE,**

**PRIMARY KEY ((STATE,CITY), DATE));**







1. Write Spark code to save the JSON DataFrame into your Cassandra table. Make sure the column names are the same. Read the data back out and make sure you have the same number of rows in the DataFrame and in the Cassandra table. This will be further proof that your Cassandra row key is set up correctly. Provide Spark code to save the data to Cassandra and then a screenshot of a select statement and output in the CQL shell.

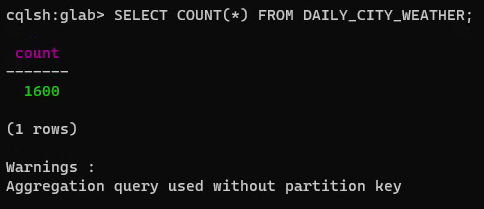
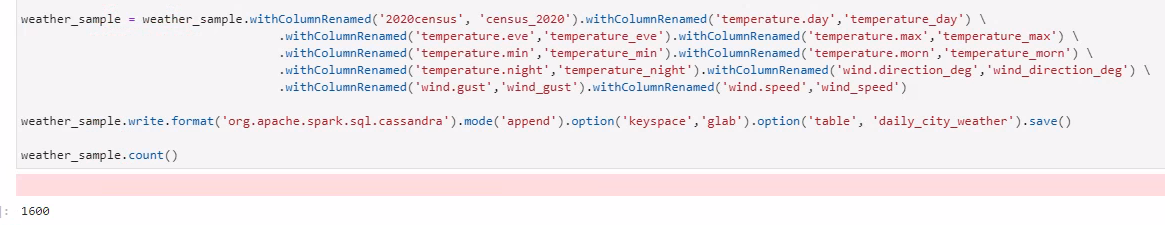
**weather\_sample = weather\_sample.withColumnRenamed('2020census', 'census\_2020').withColumnRenamed('temperature.day','temperature\_day') \ .withColumnRenamed('temperature.eve','temperature\_eve').withColumnRenamed('temperature.max','temperature\_max') \ .withColumnRenamed('temperature.min','temperature\_min').withColumnRenamed('temperature.morn','temperature\_morn') \ .withColumnRenamed('temperature.night','temperature\_night').withColumnRenamed('wind.direction\_deg','wind\_direction\_deg') \**

**.withColumnRenamed('wind.gust','wind\_gust').withColumnRenamed('wind.speed','wind\_speed')**

**weather\_sample.write.format('org.apache.spark.sql.cassandra').mode('append').option('keyspace','glab').option('table', 'daily\_city\_weather').save()**

**weather\_sample.count()**

**1600**



1. Write a CQL shell query to get the condition, description, and daytime temperatures for "Syracuse, NY"; include all dates.

**cqlsh:glab> SELECT**

**... DATE,**

**... CONDITION,**

**... DESCRIPTION,**

**... TEMPERATURE\_DAY**

**... FROM**

**... DAILY\_CITY\_WEATHER**

**... WHERE**

**... STATE = 'New York'**

**... AND CITY = 'Syracuse';**



1. Write the same query as Question 7, but using Spark SQL. Register the data from Cassandra as the Temp View **daily\_city\_weather**, then use Spark SQL to filter on “Syracuse, NY.” Instead of showing the output, **explain()** the Spark query to prove the filter is being passed-through to Cassandra. (The filter should NOT be happening in Spark—Welcome to big data country!)

**spark.read.format("org.apache.spark.sql.cassandra") \**

**.option('keyspace','glab').option('table', 'daily\_city\_weather').load() \**

**.createOrReplaceTempView('daily\_city\_weather')**

**spark.sql('''**

**SELECT**

**DATE,**

**CONDITION,**

**DESCRIPTION,**

**TEMPERATURE\_DAY**

**FROM**

**DAILY\_CITY\_WEATHER**

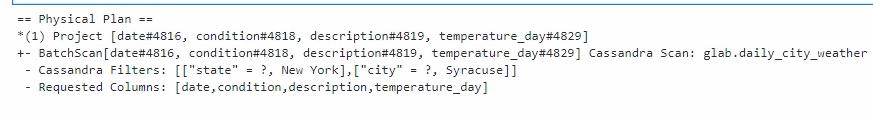
**WHERE**

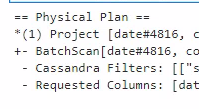
**STATE = 'New York'**

**AND CITY = 'Syracuse'**

**''').explain()**

****





1. Your company would like to now allow users to find cities where it is raining on a specific date. Specifically, they would like a query to show the city and state name, date, condition, and description for only those cities where it is raining on the given date. Write this query in Spark or Spark SQL. Which Cassandra filters are used? Show with explain and highlight in your screenshot.

**spark.sql('''**

**SELECT**

**CITY,**

**STATE,**

**DATE,**

**CONDITION,**

**DESCRIPTION**

**FROM**

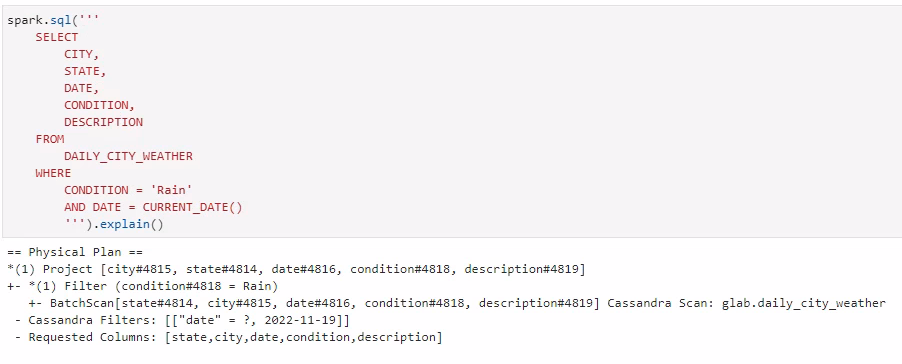
**DAILY\_CITY\_WEATHER**

**WHERE**

**CONDITION = 'Rain'**

**AND DATE = CURRENT\_DATE()**

**''').explain()**





1. Run the same query in Question 9 from the CQL command line; obviously it requires ALLOW FILTERING. Figure out how you can do an index or materialized view to avoid a costly ALLOW FILTERING operation. Include your CQL to create the index or materialized view and then include a query demonstrating it works in CQL.   
   NOTE: Our version of Cassandra and the Spark Connector does not support materialized views.

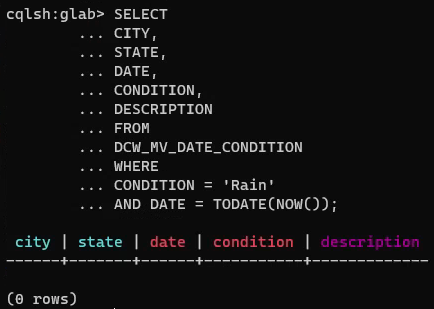
**cqlsh:glab> CREATE MATERIALIZED VIEW DCW\_MV\_DATE\_CONDITION AS**

**... SELECT \* FROM DAILY\_CITY\_WEATHER**

**... WHERE DATE IS NOT NULL AND CONDITION IS NOT NULL**

**... AND STATE IS NOT NULL AND CITY IS NOT NULL**

**... PRIMARY KEY ((DATE, CONDITION),STATE,CITY);**

****