Lab Task 4

Python code explanation:

- I read weather.json file in a pyspark dataframe.
- Then I run query for filtering the records where feels_like temperature for the day is below 14 degrees Celsius.
- For this query I explode the daily array in the dataframe. Here explode function will explode the daily array and give that column “col” name [1] [2] [3]. So, I used withColumn function to rename that “col” column to “daily” column [4].
- Then I used select function in which I mentioned the names of the columns that I want to get.
- Then used filter function of the dataframe to filter the data.
- At last, I saved the filtered dataframe contect to the cold_weather.json file.

You can find the weather data in weather.json file, filtered data in cold_weather.json file and python code to read json, filter and write json in main.py file.

Output screenshot:

```
>>> from pyspark.sql.functions import explode
>>> file_name = "weather.json"
>>> weather = spark.read.option("multiline", "true").json(file_name)
>>> df = weather.withColumn("daily", explode(weather.daily)).select("lat", "lon", "timezone", "timezone_offset", "daily").filter("daily.feels_like.day < 14")
>>> df.printSchema()
root
 |-- lat: double (nullable = true)
 |-- lon: double (nullable = true)
 |-- timezone: string (nullable = true)
 |-- timezone_offset: long (nullable = true)
 |-- daily: struct (nullable = true)
 |    |-- clouds: long (nullable = true)
 |    |-- dew_point: double (nullable = true)
 |    |-- dt: long (nullable = true)
 |    |-- feels_like: struct (nullable = true)
 |    |    |-- day: double (nullable = true)
 |    |    |-- eve: double (nullable = true)
 |    |    |-- morn: double (nullable = true)
 |    |    |-- night: double (nullable = true)
 |    |-- humidity: long (nullable = true)
 |    |-- moon_phase: double (nullable = true)
 |    |-- moonrise: long (nullable = true)
 |    |-- moonset: long (nullable = true)
 |    |-- pop: double (nullable = true)
 |    |-- pressure: long (nullable = true)
 |    |-- rain: double (nullable = true)
 |    |-- sunrise: long (nullable = true)
 |    |-- sunset: long (nullable = true)
 |    |-- temp: struct (nullable = true)
 |    |    |-- day: double (nullable = true)
 |    |    |-- eve: double (nullable = true)
 |    |    |-- max: double (nullable = true)
 |    |    |-- min: double (nullable = true)
 |    |    |-- morn: double (nullable = true)
 |    |    |-- night: double (nullable = true)
 |    |-- uvi: double (nullable = true)
 |    |-- weather: array (nullable = true)
 |    |    |-- element: struct (containsNull = true)
 |    |    |    |-- description: string (nullable = true)
 |    |    |    |-- icon: string (nullable = true)
 |    |    |    |-- id: long (nullable = true)
 |    |    |    |-- main: string (nullable = true)
 |    |-- wind_deg: long (nullable = true)
 |    |-- wind_gust: double (nullable = true)
 |    |-- wind_speed: double (nullable = true)

>>> df.show()
+-------+--------+---------------+---------------+--------------------+
|    lat|     lon|       timezone|timezone_offset|               daily|
+-------+--------+---------------+---------------+--------------------+
|44.6462|-63.5736|America/Halifax|         -10800|{43, 6.06, 165574...|
+-------+--------+---------------+---------------+--------------------+

>>> df.write.mode("Overwrite").json("cold_weather.json")
>>>
```

References:

[1] “cannot resolve column due to data type mismatch PySpark,” *Stack Overflow*. [Online]. Available: https://stackoverflow.com/questions/60646254/cannot-resolve-column-due-to-data-type-mismatch-pyspark. [Accessed: 22-Jun-2022].
[2] “Spark – Create a DataFrame with Array of Struct column,” *Sparkbyexamples.com*. [Online]. Available: https://sparkbyexamples.com/spark/spark-dataframe-array-of-struct/. [Accessed: 22-Jun-2022].
[3] “PySpark explode array and map columns to rows,” *Sparkbyexamples.com*. [Online]. Available: https://sparkbyexamples.com/pyspark/pyspark-explode-array-and-map-columns-to-rows/. [Accessed: 22-Jun-2022].
[4] “Spark : How do I exploded data and add column name also in pyspark or scala spark?,” *Stack Overflow*. [Online]. Available: https://stackoverflow.com/questions/48748727/spark-how-do-i-exploded-data-and-add-column-name-also-in-pyspark-or-scala-spar. [Accessed: 22-Jun-2022].