DALHOUSIE UNIVERSITY

Faculty of Computer Science

Assignment - 4

In

The Class of

CSCI5408: Database Management, Warehousing, Analytics

By

Deep Patel
B00865413
dp889845@dal.ca

Submitted to

Dr. Saurabh Dey

Department of Computer Science
Dalhousie university.

Date: 15th April 2021

Selection of Measurable Fields:

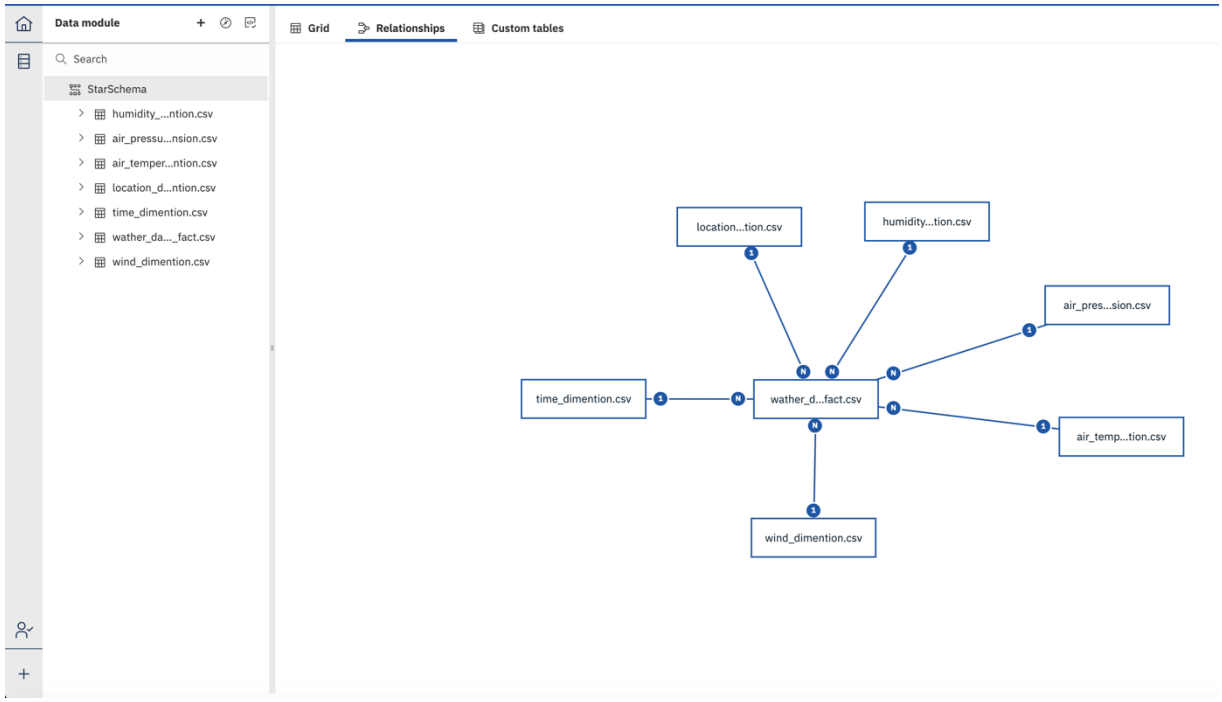
1. **Identification of dimension tables:**
   A dimension table is the structure that helps in creating a fact table. In the given example of the weather dataset the dimension tables could be as per given below;
   1. air_pressure – This could be a dimension table to calculate fact table as air pressure is imported to identify rain.
   2. time – Time could be one of the factors to took into consideration while calculating rain and temperature during particular time.
   3. Humidity – Humidity of the air could also be the reason behind the raining.
   4. Location – Location is the important factor for which we can calculate the amount of raining and temperature on particular location at particular time.
   5. Wind – The amount of wind helps in determining the temperature of the place on the given particular time therefore this could be a dimension table for a fact table of temperature.

2. **Identification of fact tables:**
   Fact table is the table containing all the derived attributes from all other dimension tables to perform particular calculation on that data to get needed outcome.
   1. Weather_dataset – In the fact table of calculating rain perception can be done using factors like humidity, air pressure location and time.
   2. air_temperature – Temperature is the fact table which can be identified using location, time, wind and pressure, and solar radiation.

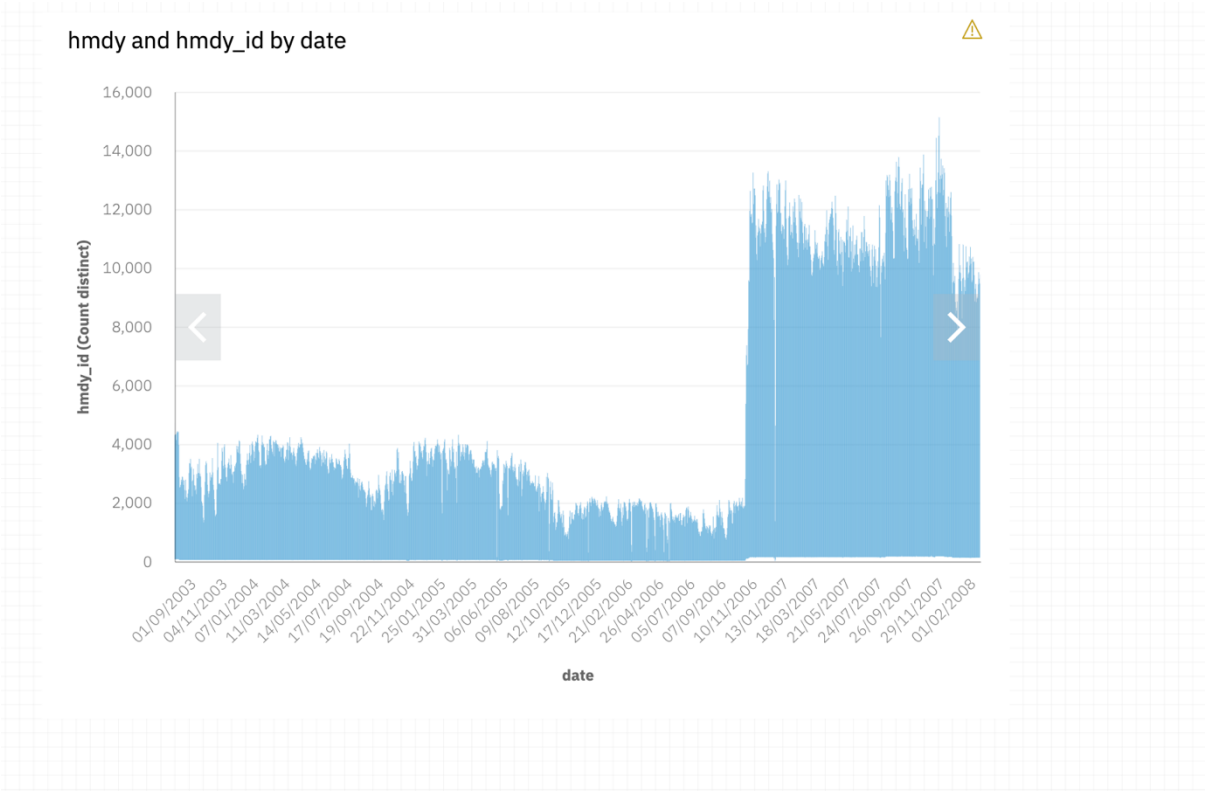
Cleaning of Dataset:

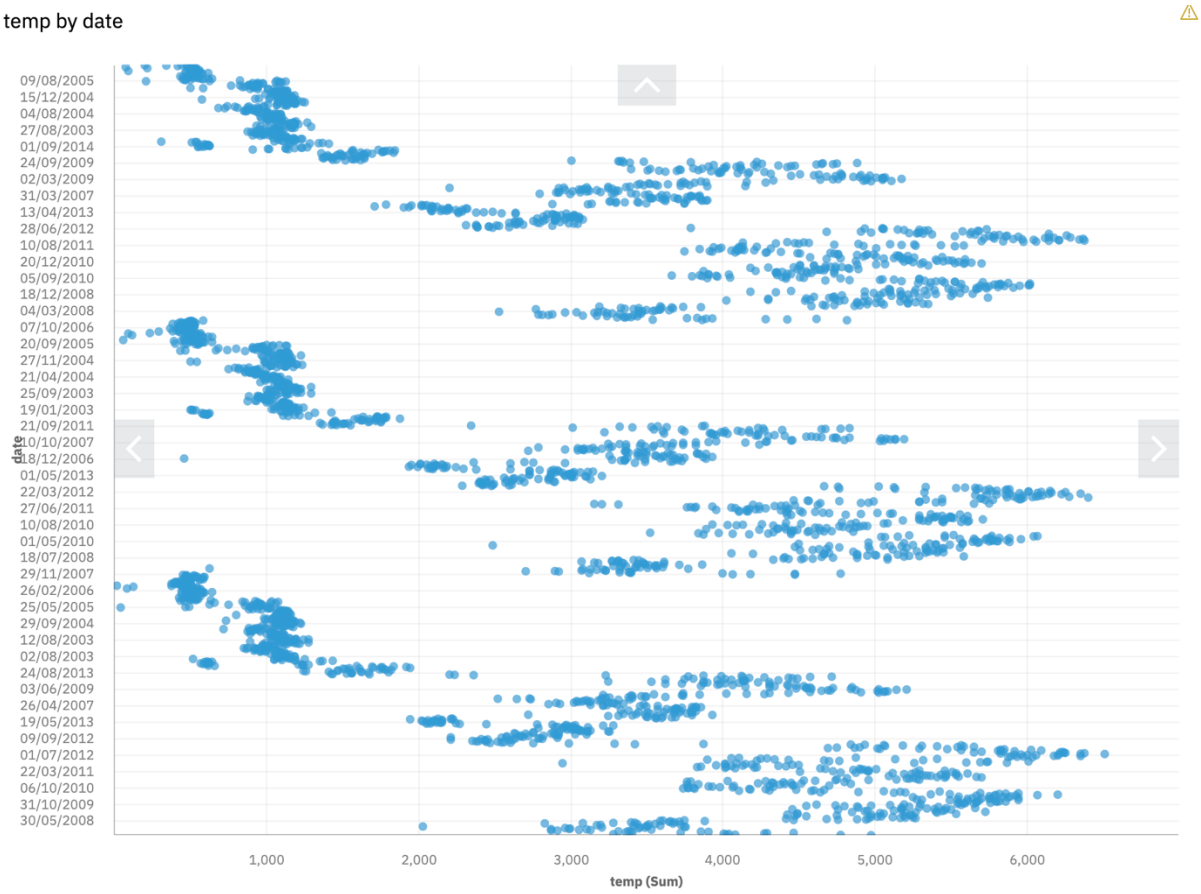
1. For cleaning first thing to do was cleaning the name of the city to which can be done using Numbers in MAC by opening it as it is in particular format.
2. From the million records I selected only 500000 records with equal amount of data with each having same values.
3. Replaced all the null values in the integer field with value zero(0).
4. Removed all the data rows where multiple columns data is null or zero (0).
5. Changed the format of date and time.
6. Added id field in each dimension table.

Star Schema:



Visualizations:





Problem 2#

Sentiment analysis:

File: assignment_4_sentiment_analysis_deep.py

Output:

{'RT': 1, ':': 1, 'Science': 1, 'denying': 1, 'nationalist': 1, 'who': 1, 'compared': 1, 'the': 2, 'pandemic': 1, 'to': 2, 'a': 1, 'common': 1, 'flu': 1, 'wants': 1, 'you': 1, 'bel
Polarity for this tweet is: NEGATIVE
{'RT': 1, ':': 1, 'Football': 1, 'fans': 1, 'wearing': 1, 'masks': 1, 'during': 2, 'a': 1, 'football': 1, 'match': 1, 'in': 1, 'Atlanta': 1, 'the': 1, '1918': 1, 'flu': 1, 'pandemi
Polarity for this tweet is: POSITIVE
{'So': 1, 'how': 1, 'to': 1, 'we': 2, 'make': 1, 'sure': 1, 'this': 1, "doesn't": 1, 'happen': 1, "again.Let's": 1, 'face': 1, 'it': 1, 'as': 1, 'all': 1, 'know': 1, 'winter': 1, '
Polarity for this tweet is: NONE
{'History': 1, 'repeats': 1, 'itself': 1, 'strangely': 1, 'before': 1, 'Covid19': 1, 'there': 1, 'was': 1, 'the': 1, 'Spanish': 1, 'flu': 1, 'so': 1, 'that': 1, 'is': 1, 'why': 1,
Polarity for this tweet is: NEGATIVE
{'Not': 1, 'sure': 1, 'how': 1, 'old': 1, 'he': 2, 'is': 1, 'but': 1, 'two': 1, 'facts': 1, 'can': 1, 'tell:': 1, '94+%': 1, 'positive': 1, 'PCR': 1, 'tests': 1, 'are': 1, 'FALSE':
Polarity for this tweet is: POSITIVE
{'RT': 1, ':': 1, 'Available': 1, 'from': 1, 'Helios': 1}
Polarity for this tweet is: NONE
{'Summer.': 1, 'Watch': 1, 'flu': 2, '(Covid)': 1, 'cases': 1, 'spike': 1, 'again': 1, 'this': 1, 'season.': 1, 'The': 1, 'vaccine': 1, 'rollout': 1, 'is': 1, 'slow': 1, 'for': 1,
Polarity for this tweet is: NEGATIVE
{'RT': 1, ':': 1, 'Professor': 1, 'elains': 1, 'the': 1, 'normal': 1, 'flu': 1, 'vs': 1, 'coronavirus': 1, 'in': 1, 'one': 1, 'minute...': 1}
Polarity for this tweet is: NONE
{'RT': 1, ':': 1, '"In': 1, 'summary:': 1, 'The': 2, 'corona': 1, 'crime': 2, 'is': 2, 'swine': 1, 'flu': 1, '2.0': 1, 'and': 1, '"Corona"': 1, 'mostly': 1, 'Influenza': 1, '2.0:':
Polarity for this tweet is: NEGATIVE
{"Isn't": 1, 'it': 1, 'amazing': 1, 'that': 1, 'No': 1, 'cases': 1, 'of': 1, 'Influenza': 1, 'were': 1, 'reported': 1, 'for': 1, '2020-2021': 1, 'fl…': 1}
Polarity for this tweet is: POSITIVE
{'RT': 1, ':': 1, '"There': 1, 'is': 1, 'no': 2, 'flu.': 1, 'There': 1, 'has': 1, 'been': 1, 'human': 1, 'influenza': 1, 'in': 1, 'Alberta': 1, 'this': 1, 'season,"': 1, 'said': 1,
Polarity for this tweet is: NONE
{'RT': 1, ':': 1, 'No': 1, 'he': 2, "didn't": 1, 'hear': 1, 'about': 1, 'the': 1, 'pandemic': 1, 'thought': 1, 'it': 1, 'was': 1, 'flu': 1}
Polarity for this tweet is: NONE
{'RT': 1, ':': 1, 'The': 1, 'same': 1, 'people': 1, 'who': 1, 'said...': 1, 'COVID': 1, 'is': 1, 'like': 1, 'flu': 1, 'Masks': 1, 'don’t': 1, 'work': 1, 'BCG': 1, 'will': 1, 'prote
Polarity for this tweet is: POSITIVE
{"There's": 1, 'no': 1, 'basis': 1, 'to': 2, 'lockdown': 1, 'for': 1, 'a': 1, 'bad': 1, 'flu': 1, 'season.': 1, 'Worldwide': 1, 'other': 1, 'countries': 1, "wouldn't": 1, 'gain': 1
Polarity for this tweet is: NONE
{'RT': 1, ':': 1, 'The': 1, 'oldest': 1, 'person': 1, 'in': 3, 'Costa': 1, 'Rica.': 1, '121': 1, 'years': 1, 'old': 1, 'today.': 1, 'Got': 1, 'the': 1, 'Spanish': 1, 'flu': 1, '191
Polarity for this tweet is: NONE
{'RT': 1, ':': 1, 'This': 1, 'is': 1, 'the': 2, 'same': 1, 'anti-science': 1, 'person': 1, 'who': 1, 'said': 1, 'Covid': 1, 'was': 1, 'nothing': 1, 'to': 1, 'be': 1, 'afraid': 1, '
Polarity for this tweet is: NONE
{'RT': 1, ':': 1, 'So': 1, 'what': 1, 'have': 1, 'we': 1, 'learnt': 1, 'this': 1, 'morning.': 1, 'Just': 1, 'facts...': 1, 'Ireland': 1, 'is': 1, 'looking': 1, 'at': 1, 'another':
Polarity for this tweet is: NONE
{'No': 1, 'he': 2, "didn't": 1, 'hear': 1, 'about': 1, 'the': 1, 'pandemic': 1, 'thought': 1, 'it': 1, 'was': 1, 'flu': 1}

Semantic analysis:

File: assignment_4_semantic_analysis_deep.py

Output:

```
Query word : Document containing term(df) : Total Documents(N)/ number of documents term appeared(df) : Log10(N/df)
flu: 244 : 2.0491803278688523 : 1.1305108091522382
snow: 88 : 5.681818181818182 : 1.3880144282034426
cold: 2 : 250.0 : 8.965784284662087

documentNo : Total Words(m): frequency(f)
0 : 22 : 0
1 : 19 : 0
2 : 6 : 0
3 : 15 : 1
4 : 17 : 2
5 : 12 : 3
6 : 25 : 4
7 : 25 : 5
8 : 25 : 6
9 : 23 : 6
10 : 28 : 6
11 : 17 : 7
12 : 21 : 8
13 : 24 : 8
14 : 17 : 8
15 : 24 : 8
16 : 7 : 9
17 : 25 : 10
18 : 17 : 11
19 : 25 : 12
20 : 25 : 13
21 : 23 : 13
22 : 12 : 13
23 : 25 : 14
```

The highest frequency/total words is 14/25 = 0.56