

**CSCI 5408**

**Data Management, Warehousing and Analytics**

**Assignment 4**

**Name : Sagarkumar Pankajbhai Vaghasia**

**CSID : vaghasia**

**Banner ID : #B00878629**

GitLab Link : <https://git.cs.dal.ca/vaghasia/csci5408_s22_sagarkumar_vaghasia_b00878629/-/tree/main/A4>

**Problem #1**

**Business Intelligence Reporting using Cognos.**

**Step 1:** Firstly, I have downloaded the weather dataset from Kaggle website[1]. Weather stations are located in the southeast region of Brazil.

**Step 2:** From the first look of the downloaded dataset, it can be seen that all the tables (central\_west, north, northeast, south, southeast) have same columns except two tables (columns\_description and stations).

**Step 3: Measurable fields and dimensions**

Measurable fields and dimensions selected for the given dataset with the reasons are described below:

Field -1: max temperature in previous hour OR min temperature in previous hour. Dimensions: “region”, “state”, “station”, “date”, “hour”.

This illustrates dicing when we want the max temperature on hourly basis of specific region for specific date. Also, slicing can be applied when we want it for any one dimension only.

Field -2: atmospheric pressure at station height.

Dimensions: “station”, “height”.

This illustrates dicing where we can measure the pressure by station and height.

Field -3: wind speed.

Dimensions: “wind direction”.

This illustrates slicing where we can measure average wind speed for specific direction.

Field -4: radiation.

Dimensions: “region”, “state”, “station”, “date”, “hour”.

We can measure radiation on hourly basis OR for specific date OR region wise OR state wise OR station wise. The slicing can be applied if we measure radiation based on hours only and dicing is applied if we measure radiation based on multiple dimensions like hour, date and state.

Field -5: total precipitation.

Dimensions: “region”, “state”, “station”, “date”, “hour”.

The reason for selecting total precipitation is same as of radiation.

Filed -6: air temperature.

Dimensions: “region”, “state”, “station”, “date”, “hour”.

**Step 4:** The column names in all the csv files are in different languages. So, I translated it to English by using google translate and columns\_description table.

There is one cell in a column named as “abbreviation” which is blank for “region” entry which I assigned as reg.

**Step 5:** I created IBM Cognos account by following instructions of laboratory tutorial[2][5].

Graphical user interface, website

Description automatically generated

Figure 1: IBM Cognos

I decided to implement star schema using the following four dimensions :

* Date
* Region
* State
* Station

Cognos do not accept files larger than 100 MB . Thus, I manually picked 4000 rows from given CSV files.

Then, I uploaded those files to the Cognos.

I have developed those dimension tables using Microsoft Excel.

**Step 6:**

* I choose to go with star schema.
* The first dimension I took was date. Based on date we can get answers for the questions like “Total precipitation for specific date” and also we can visualize and analyze it.
* I imported weather table as fact table which I generated from central\_west by picking random 4000 rows. I have also imported state, region, station, date tables as dimension tables.

After that, I developed star schema as shown in the below image.

Graphical user interface, diagram

Description automatically generated

Figure 2: Star Schema

* It can be seen that there are one to many relationships between dimension table and fact table as one value of dimension table is connected with many value of fact table.

**Step 7:** I have performed visual analysis using IBM Cognos exploration.

Visual Analysis 1

Graphical user interface, chart

Description automatically generated

Figure 3: Visual Analysis 1

Visual Analysis 2

Graphical user interface

Description automatically generated

Figure 4: Visual Analysis 2 [Bar-chart]

Visual Analysis 3

A screenshot of a computer

Description automatically generated with medium confidence

Figure 5: Visual Analysis 3

Visual Analysis 4

Graphical user interface, application

Description automatically generated

Figure 6: Visual Analysis 4 [Bar-chart]

Visual Analysis 5

Graphical user interface

Description automatically generated

Figure 7: Visual Analysis 5 [Pie-chart]

**Problem #2**

**Sentiment Analysis**

For sentiment analysis, I have developed a script without using any additional java external library.

My code will generate txt file as an output.

**Steps:**

* I have taken the reference of the parser which I used in the last assignment. Now using that parser, the code will parse news articles into array list.
* Now, I will generate list of bags of words where each bag of words contains news description.
* After that, I developed a Hashset for positive and negative words[3][4].
* I have also developed one model class having following variables: id, content, contentFrequencyCount, matchedWords, polarity, positiveScore, negativeScore.
* In the logic of Sentiment Analysis, code will go through list of bag words and it will compare with the previously developed hashset of positive and negative words which will ultimately store the result.
* At the end, I simply generated a txt file for the output.
* The image of the output txt file is attached below.

**A picture containing application

Description automatically generated**

**Problem #2**

**Semantic Analysis**

For sentiment analysis I have developed a script without using any additional java external library.

My code will generate txt file as an output.

I have developed TF-IDF and TermFrequency table in the form of txt file using core JAVA.

**Steps:**

* I have taken the reference of the parser which I used in the last assignment. Now using that parser, the code will parse news articles into array list.
* Then, I developed two classes: SemanticAnalysis and SemanticAnalysisModel.
* We will travel through all the documents and calculate how many documents contains the search word.
* After that, based on the list I deveoped a tabluar data in txt file.
* You can better understand the working of the code by looking at it.
* At the end, I simply generated a txt file for the output.
* The image of the output txt file is attached below.

**A picture containing diagram

Description automatically generated**

**References :**

[1] P. em I.-D. Mestrado, “Climate weather surface of Brazil – hourly,” *Kaggle.* [Online]. Available: https://www.kaggle.com/datasets/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region?resource=download [Accessed: 20-Oct-2021]

[2] “Lab9\_IBM\_Cognos\_data\_module,” *Brightspace.com*. [Online]. Available: https://dal.brightspace.com/d2l/le/content/221749/viewContent/3063667/View. [Accessed: 20-Jul-2022].

[3] Marcin, *Negative-words.Txt*. .

[4] Marcin, *Positive-words.Txt*. .

[5] “IBM Cognos Analytics,” *IBM Cognos Analytics*. [Online]. Available: https://www.ibm.com/products/cognos-analytics. [Accessed: 23-Jul-2022].