University of Ottawa
School of Electrical Engineering and Computer Science
CSI4142 Project 2017

Canadian Disaster Database

This document contains the requirements as discussed in class.

Instructions

1. Complete this project in a group of two (2) to three (3) students.
2. Demonstrate the project on **April 2nd, 2018** in a 15 minute timeslot, as allocated by the TA.
3. Use a database management system (DBMS) such as PostgreSQL to complete this project. As explained in class, you may use a Dashboard Template or you could also create your own frontend for the OLAP analytics. Similarly, the data mining component may be implemented using the R or Python languages, or by using a system such RapidMiner or WEKA.

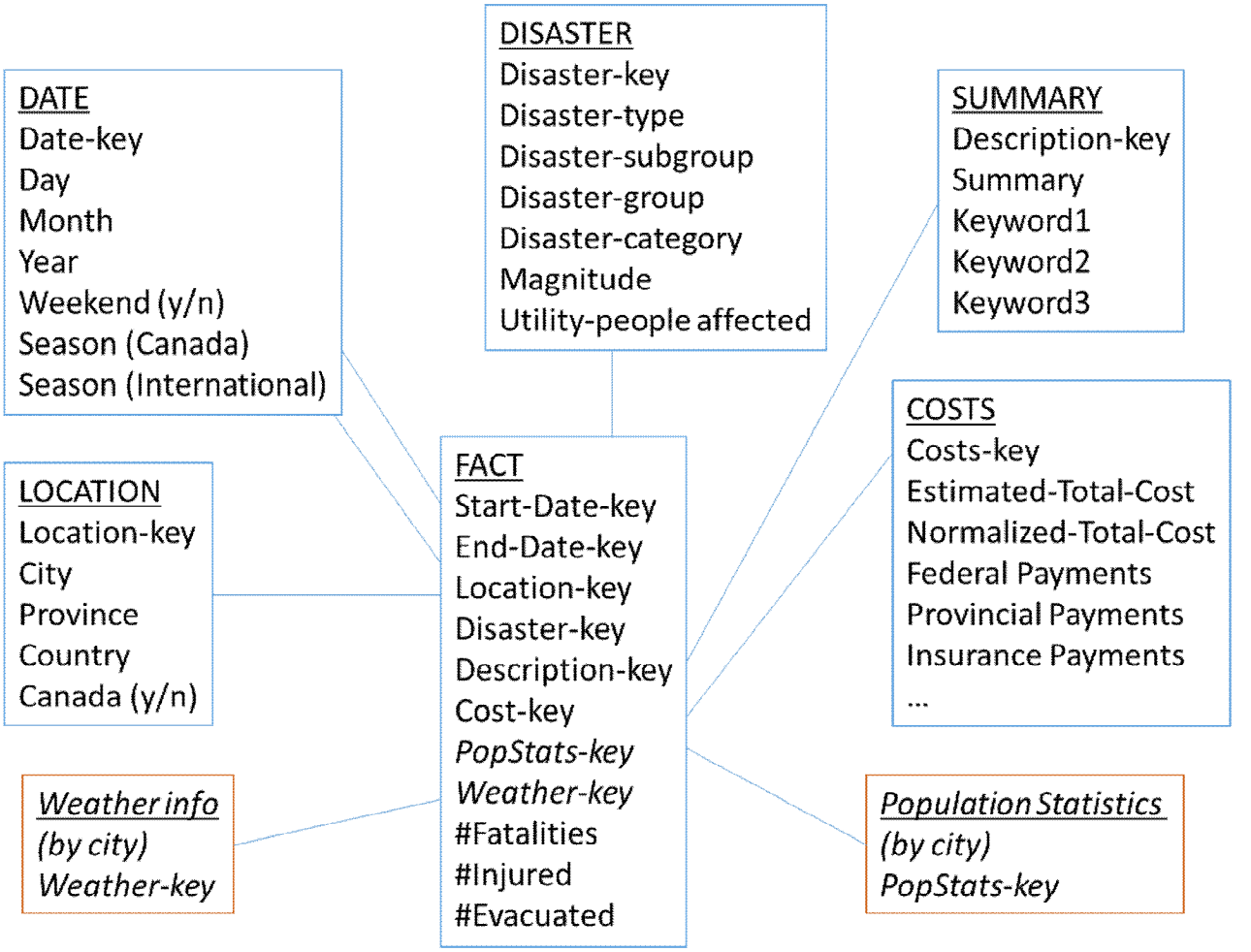
Deliverables:

Submit all your source code, together with a one-page high level data staging plan through the Virtual Campus**, before April 2nd, 2018 at 08h00**. (That is, before the demonstrations will start.)
Note that all group members should submit the source code, not just one per group.
All group members should attend the project demonstrations.

Your task:

Your task is to design and implement a Canadian Disaster Data Mart using the data, as uploaded on the Virtual Campus. More details about the data may be found at https://www.publicsafety.gc.ca/cnt/rsrcs/cndn-dsstr-dtbs/index-en.aspx
Your project will consist of i) data staging, ii) the building of OLAP queries and the design of an OLAP end user interface, and iii) some machine learning.

Here is the dimensional model of the proposed data mart. You may use this model "as is", or modify and extend it as you see fit.



1. Remember to create your own surrogate keys. Refer to the slides and/or the book by Kimball et. al. that explain how to stage the data for surrogate key lookup.
2. Recall from the JAD session that you are encouraged to supplement the original data with enriched data from other sources, such as population statistics and/or weather data.
3. The data mart contains concept hierarchies on the Date, Disaster, and Location dimensions.
4. Date is a role playing dimension.

5. You will notice that there are missing values, notably for the Costs dimension. Here, you need to create an "Empty" Costs row, where the values are all set to unknown, to be linked the Fact table.
6. Some of the data are on provincial or regional level, while others are on city level. For instance "the Prairies" refer to three provinces: Alberta, Saskatchewan, and Manitoba. In this case, you should convert the data to include the provinces and their main cities. This will lead to more rows in the Location dimension.
7. Refer to the "Typical Analytic Cycle" as described in class, for a list of typical analytic questions which should be answered when accessing your data mart.

Requirements and Mark Allocation

Note that the project is out of **100**. You obtain a mark of 115/100 if you complete all the questions. Additional work, outside the scope of this project, could earn you up to 10 additional marks.

1. **(5 marks) Physical Design:** Create the physical schema of the data mart using the DBMS of your choice.

2. **(25 marks) Data staging**: Extract and transform the data and load all rows into the data mart. Be sure to record all the steps that you followed and submit a one-page high level schematic with your source code.

3. **(25 marks) OLAP queries**: Implement the following five types of queries by traversing the concept hierarchies. (Refer to Example 4.4 on page. 146 of the Data Mining textbook (3rd edition) by Han et. al. for a description of these operations.)
   a. **Drill down.**
   b. **Roll up.**
   c. **Slice.**
   d. **Dice.**
   e. Top N or Bottom N (Iceberg queries).

   These operations enable us to answer questions such as:

   - Determine X in {*fatalities, injuries, evacuations*} per Y in {*disaster type, disaster subgroup, disaster group, disaster category, year, month, date, location (city), province, country (Canada or International)*}

     For instance, determine the total number of fatalities in Ontario during 1999.

     For instance, determine the total number of fatalities due to natural disasters in Ontario during 1999, and so on.
   - Contrast X in {*fatalities, injuries, evacuations*} per Y in {*event, event type, event category, event subgroup, year, month, date, location (city), province, country (Canada or International)*} when compared to Z in {*event, event type, event category, event subgroup, year, month, date, location (city), province, country (Canada or International)*}

For instance, contrast the number of fatalities in Ontario due to wildfires, during 1999, with the number of fatalities in Ontario due to flooding, during 1999.
For instance, contrast the number of fatalities in Ontario due to wildfires, during 1999, with the number of fatalities in Quebec due to wildfires, during 1999.
For instance, contrast the number of fatalities in Ontario due to wildfires, during December 1999, with the number of fatalities in Ontario due to flooding, in December 1999, and so on.

- Locate "hot spots" for certain types of events.
  For instance, determine the 5 cities in Canada with the most riots.
  For instance, determine the province with the most space debris.
- Calculate the trends in costs, fatalities, injuries or evacuations per event type, over the years.
  For instance, determine the increase in normalized costs of wildfires over the last 50 years.
  For instance, determine the trends in fatalities due to riots over last 100 years.

4. **(15 marks) Business Intelligence Dashboard.** Create an interface which will give a knowledge worker the ability to easily explore the data mart. (This should include graphs.)

Complete at least two of the following four questions.

5. **(15 marks) Classification.** Employ a classification algorithm (e.g. decision tree, support vector machine, lazy learner, ensemble, etc.) to explore the data. For example, you may want to construct a model that contrast the disasters in two or more provinces.
   (and/or)
6. **(15 marks) Cluster analysis.** Apply a cluster analysis algorithm to find potential groupings within the data.
   (and/or)
7. **(15 marks) Anomaly detection.** Explore the data in order to determine whether it contains global, contextual and/or collective outliers. In this question, one may employ statistical tests, cluster analysis and/or one-class learners.