University of Ottawa
School of Electrical Engineering and Computer Science
CSI4142 Project 2019

Crime Data Mart

This document contains the requirements for the data staging part of the project.

Instructions

1. Complete this project in a group of (3) students.
2. Submit your deliverable before the due date using the group locker in the Virtual Campus.
3. Demonstrate your project to the corrector, in a 15 minute time slot during the week of 24 February, in SITE 4-010. (We will create a google sheet for this purposes.)
4. Use a database management system (DBMS) such as PostgreSQL to complete this project.

Deliverables:

You are asked to submit the following details using your group locker:

- i) the scripts to create the database schema
- ii) the scripts to stage the data into the dimensional tables and the fact table
- iii) a document containing your **one-page** high level data staging plan, similar to the Electricity data mart example covered in class.

All group members should attend the demonstrations.

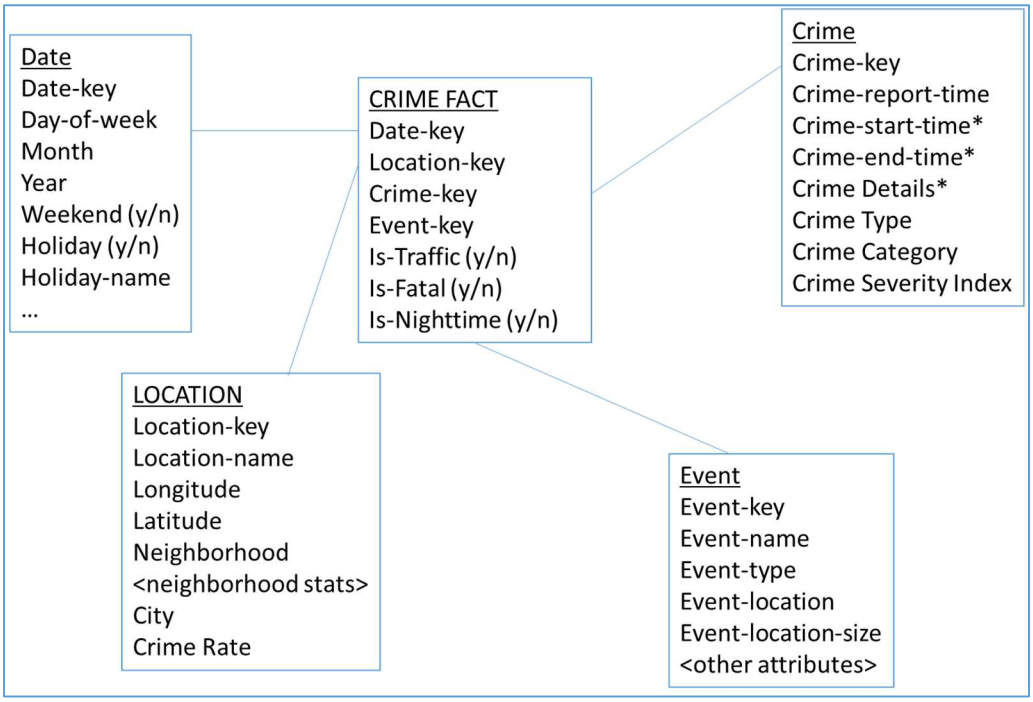
Note that you will be asked to rate one another's participation, and this rating will be reflected in your final mark. For instance, if a group obtains 100% and a team member A receives an average rating of 8/10 from the other two members, A's mark will be 80%.

Your task:

Your task is to create the crime data mart, i.e. complete the physical design and data staging. Use the data from the Denver and Vancouver datasets that cover the same time period, as contained in:
https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-crime and
https://geodash.vpd.ca/opendata/

You should augment these sources with additional information about Events and Demographic information about neighborhoods.

Here is the dimensional model of the proposed data mart. You may use this model "as is", or modify or 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. Supplement the original data with enriched data from other sources, such as e.g. population statistics, notably when considering locations.
3. Use the full Date dimension from Kimball, as discussed in class.
4. The data mart contains concept hierarchies on the Date, Crime, Event and Location dimensions.
5. Note that the Vancouver data lacks details of the crime-start-time, crime-end-time and crime details. It is important to map the two sources to ensure that the crime types and crime categories are similar.
6. The Event dimension tracks events that are categorised as types, such as music festivals, sports events and family activities.
7. We maintain three facts, namely Is-Traffic (y/n), Is-Fatal (y/n) and Is-Nighttime (y/n) where "yes" is staged as "1" and "no" is staged as "0". Implicitly, Is-Nighttime can also be used to run queries when determining whether the crime was committed during the day.
8. In this model, we add an attribute crime severity to the Crime dimension, where we rate an offense as "violent, non-violent, youth", etc. It follows that this could also have been implemented as a fact/measure, in cases where the want to count the number of crimes by severity. See

https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3510002601&pickMembers%5B0%5D=1.37 for the Vancouver index.

9. The crime rate of a city is calculated by dividing the number of reported crimes by the total population; the result is multiplied by 100,000. The information is stored in the Location dimension.

Mark Allocation (100 marks in total)

1. **(10 marks)** Submit a **one-page high level schematic** detailing the high level plan you followed.

2. **(20 marks) Physical Design:** Create the physical schema of the data mart using the DBMS of your choice.

3. **(70 marks) Data staging**: Extract and transform the data and load all rows into the data mart.
    a. **(15 marks)** Staging of Denver data, including surrogate key generation and referential integrity enforcement.
    b. **(15 marks)** Staging of Vancouver data, including surrogate key generation and referential integrity enforcement.
    c. **(20 marks)** Mapping of Denver and Vancouver data – fusion of categories, types, and so on; handling of NULL values.
    d. **(10 marks)** Staging of Event data.
    e. **(10 marks)** Staging of statistical data involving neighborhoods, crime rates and crime severity indexes.