**Query Engine Design**

**Team Brockville**

Ahmed Mahmoud - 0966966

Joshua Prpic - 0971107

Alex Montague - 0959687

1. The current program design will require all of the data sets available to us except for the vector flag in the crime statistics. We will omit the vector flag since we will use the coordinate encoding allowing us to transverse through the data efficiently without reading in an extra field. Every other field will be used in the query engine in some capacity. The “Geo”, “Violation” and “STA” fields will all be converted to tables in order to be used alongside the coordinate encoding described in the previous design document. Thus when we read in the crime data we will avoid reading in these fields as it we would not need it and it would just take up space. The coordinate and value fields will be read in as is and stored for later use.

Any imported data we use (i.e. the census data for 2001, 2006, and 2011) will be organized in a very similar manner. All the repeated fields (i.e. Geo) will be converted into a table and then the data will be transversed using another coordinate system for the census data.

1. Our system will cover the years that are given in the crime stats data (1998 - 2015) as it is the range of crime data. However, if the data for a specific violation and/or statistic at a given year does not exist the user will be notified.

Moreover, when cross-referencing with the census data we would be only using the 2001, 2006 and 2011 data points since these are the most relevant dates when comparing to the census data.

1. The data will be organized by an automated script that the entire group will work on for the respective large data sets. As mentioned this script will break down the crime data into sub-files organized by year (the census data is already organized based on year - 2001, 2006, and 2011).

Since we will be taking in questions that relate solely on the crime data and also a cross reference of the census data with the crime data the main query engine will be most likely the most complex section. We decided that we will be using programming in “pairs” for the engine query in order to make sure it does not have any issues that any one of us may miss. By having multiple people looking at the code we can minimize these errors.

As for the less vital scripts we will split them up over time depending on what parts of the query engine is complete and who has extra time in which they can work on some of the other scripts (i.e. the graphing script).

1. Initially we will begin by creating a question format that pertains solely on the crime data to simplify the foundation of the query engine. Additionally, the user’s input will be a format specified by the script to allow the query engine to “understand” the question. The questions based solely on the crime data will take the following format:

|  |  |  |  |
| --- | --- | --- | --- |
| **Region** | **Violation** | **Characteristic** | **Timeframe** |
| * Canada * Province * City (Major) | * Murders * Robberies * Assault * etc… | * Actual Incidents * Total Charges * Rate per 100,000 * etc... | * 1998 to 2015 (for crime data) |

Based on this query format some of the possible questions include:

* In Canada, what are the total charges of first degree murder (in the entire data set)?
* In Toronto, what is the rate per 100,00 people for robberies in 2010?
* What province has the greatest actual incidents of violent crime?  
  Etc…

The goal is to hard code none of the questions and instead use the above format to accurately search the data and come up with an answer in any specified format. However, in order to have the questions interpreted correctly and to give forgiveness to the user each section of the format will be checked for correct input.

This question format will then be incremented to also allow the user to input questions that will cross reference the crime statistics with the census.

* 1. For reading/pre-processing the crime stats data we will need a single script. For the query engine we will need another script.

The reading/pre-processing script will split up the crime stats based on year into different files for more efficient traversing. The query engine will encompass all the questions and output a universal format that will be read by the output scripts to format the result into its final form.

* 1. For the outputting we will need two scripts, one for making the graphs, and the other for making table outputs.

The graph script will take in the output from the query engine and form it into a graph using R. Depending on the project timing there will be multiple types of graphs to choose from. The table output script will simply format the query engine output into a table for more clear reading.

In total we estimate that we will need 4 scripts.