CISC 499 Project

***Supervisor*: Patrick Martin**

Data Analysis using the Hadoop Ecosystem

Data analysis is the process of obtaining an optimal and realistic decision or conclusion based on existing data. In the current era of *big data*, which involves large amounts of data of varying types and qualities, a wealth of new tools have been developed to perform data analysis. The *Apache Hadoop Ecosystem* is a set of open-source tools for big data analysis based on Hadoop, an open-source version of the MapReduce parallel programming framework, and HDFS (Hadoop Distributed File System).

The aim of the project is to first produce a report outlining the capabilities offered by the Hadoop Ecosystem and then to use tools of the Ecosystem to perform data analysis on an existing database. The database is a set of relational tables provided by the Canadian Primary Care Sentinel Surveillance Network. Using the Hadoop Ecosystem, the data will be loaded into HDFS and data analysis will be performed to answer a set of research questions. The work will be done on an OpenStack cloud running in the School of Computing.

Background: Familiarity with relational databases, data mining, Java and working with large software packages are required. Familiarity with cloud environments, preferably OpenStack, is recommended.