ETL Project Report

Jason Cook

**Introduction**

This assignment required extracting data from an external source, transforming the data to meet specific requirements, and then loading the transformed data into a database for usage by others. The raw data came from a dataset published by the Centers for Dieses Control (CDC) containing information on deaths in the United States from 2005 – 2015.

The dataset contains two files for each year:

* The first file is a CSV with roughly 2.5 million rows as well as 77 columns with details about each death including the time, place, and manner of death as well as information about the decedent such as resident status, race, and education. The CSV file does not contain written descriptions of the information in each column; instead there are key values of different data types (integer, float, and object). I assume this is due to the robust nature of the data to minimize processing. These key values relate to descriptions that are found in the second file for each year.
* The second file is a json collection of dictionaries where each object in the collection is used to decode the values from the related column in the CSV file. Some of the keys are integers while others are objects.

The main purpose of the project is to select columns from the CSV file and then reference the corresponding json object that relates to the columns chosen so the actual descriptions are shown versus the key value placeholders. To do this we needed to read the data into a pandas dataframe and then export it to a mongo database so SQL queries can be run on the synthesized tables.

**ETL**