**ETL Project Technical Report**

Provided by: Group #2

*Katherine Morriss*

*Cindy Pappas*

*Ergin Bostanci*

*Matt Mullinazi*

**PROJECT SUMMARY**

This report outlines the steps taken to compile and migrate 2015 crime data for Los Angeles, California and Kansas City, Missouri into a production database. The report will cover the data selection considerations, data cleanup process, data formatting and final production database layout. The team ran into one issue when loading the data into the MySQL database. The error identification troubleshooting process and solution are outlined at the end of the report.

**ROLES AND RESPONSIBILITIES**

Data selection: *Team effort*

Data cleanup and transformation: *Cindy Pappas*

MySQL database creation: *Ergin Bostanci*

Troubleshooting: *Team effort*

Solution: *Katherine Morriss*

Technical Report: *Team effort, primarily Katherine and Cindy*

**PROCESS**

**Data selection**: the team researched available datasets to find two cities of interest with similar data and decided upon Kansas City, Missouri and Los Angeles, California. The data sources and formats are as follows:

* Kansas City - Open Data KC (CSV file)
* Los Angeles - Kaggle (CSV file)

Both datasets had multiple years to choose from so we picked 2015 to focus on as it is both recent and most complete in both datasets.

**Data cleanup process**: Using a Jupyter notebook, the datasets were read into Pandas as dataframes. Each dataset included a unique identifier (incident\_id) for each crime reported, however, there were duplicates in both, so we removed them from each dataframe.

**Pandas dataframe format**: Once the data was transformed, we created a filtered dataframe for specific columns. Even though the data was similar, the column headings need to be changed for consistency. The dataframe columns are as follows:

* date\_reported
* crime\_code
* crime\_description
* city

**Final production MySQL database format**: Using a local instance of a MySQL database we created a new schema. In this schema we created two tables and identified the columns and datatypes for each (code is provided below for your reference).

CREATE DATABASE crimes\_db;

USE crimes\_db;

CREATE TABLE la\_crimes (

incident\_id BIGINT PRIMARY KEY,

date\_reported DATE,

crime\_code INT,

crime\_description TEXT,

city TEXT

);

CREATE TABLE kc\_crimes (

incident\_id BIGINT PRIMARY KEY,

date\_reported DATE,

crime\_code INT,

crime\_description TEXT,

city TEXT

);

Once the database was created in MySQL, we created the connection in the Pandas Jupyter notebook using the local host information and ‘create\_engine’ function. We confirmed the successful connection by using the following function:

* engine.table\_names()

The data was then imported into the MySQL database tables for each city using the following function:

* .to\_sql

**Error identification troubleshooting and solution process**: When trying to import the data to MySQL the kernel would freeze up and eventually error out. The error details were stated as:

* (pymysql.err.OperationalError) (2006, "MySQL server has gone away”)

After researching the error via Google and StackOverflow, Katherine was able to identify the cause of the error to be the size of the CSV file causing Pandas to freeze due to limited memory. Katherine was then able to find a solution by using the following function to breakdown the dataset into manageable chunks for processing:

* chunksize=1000

*#We then awarded the ‘Jupyter Bad-Ass’ Achievement to Katherine for her game winning kill (finding the solution to our problem) and bringing peace to all the Land*

**CONCLUSIONS**

The 2015 crime data for Kansas City, Missouri and Los Angeles, California has been successfully loaded into a production MySQL database. It is important to note when using large datasets in future projects we will want to be sure to break down the data into manageable chunks to avoid processing errors.

Suggested joins on the ‘date\_reported’ and/or ‘crime\_descriptions’ columns might identify if there are trends in these scenarios:

* Criminal activity at certain times in the year
* Similar types of crimes committed at certain times of the year