**Apache Spark—Real Time Project—Crime Categorization**

**Analyze and categorize data into different kinds of crimes**

SF Police Department wants to categorize crimes that were registered in different police departments. The data given by SF Police department has the following columns:

Data Fields

Dates - timestamp of the crime incident

Category - category of the crime incident (only in train.csv). This is the target variable you are going to predict.

Descript - detailed description of the crime incident (only in train.csv)

DayOfWeek - the day of the week

PdDistrict - name of the Police Department District

Resolution - how the crime incident was resolved (only in train.csv)

Address - the approximate street address of the crime incident

X - Longitude

Y - Latitude

Your company is contacted by SF Police Department to help them categorize various crimes registered. You have been given the following tasks.

1. Create a Module to initialize spark context and sqlcontext for spark.

2. Create a Module to load test dataset and training dataset and create two data frames as train and test.

3. Encode categorical data in training and test data into a format which can be accepted by mllib algorithms (Create label points).

4. Train a model on training data.

5. Categorize the crimes mentioned in test data and store results in a text file.

Data Ref: <https://www.kaggle.com/c/sf-crime>

The total time required to complete this task is 8 hours.

Note: The dataset required for this project can be accessed either from the above link or downloaded from the “Download Center”.