**University of Bridgeport**

**Fall 2015**



**Project 1: Phase 3 (Revised)**

**Dataset: Global Database of Events, Language, and Tone (GDELT)**

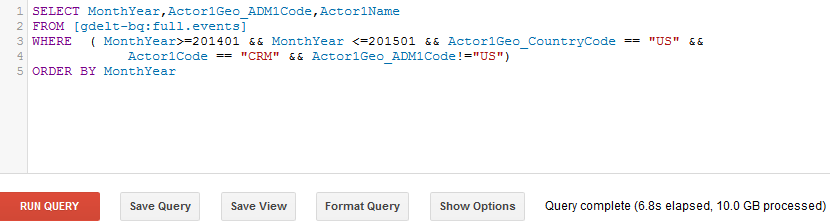
**Prof. Jeongkyu Lee**

**By Dhruv Thakkar**

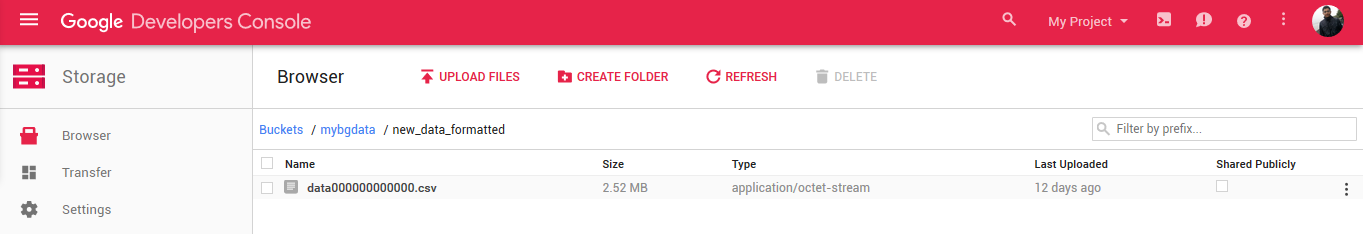
**- 0985910**

**The Dataset:**

GDELT, as previously mentioned is an open source dataset. The following was the final query that was ran to get the dataset:



The output was a BigQuery table of the size 102044 entries. Since the data was so huge it needed to be stored and then downloaded in the required format. Hence, the result was then exported to Google’s Bucket from where the data was downloaded in csv format.



**HDFS Data Upload:**

The csv file was then uploaded to the Virtual Machine (VM) and then into HDFS using the following command:

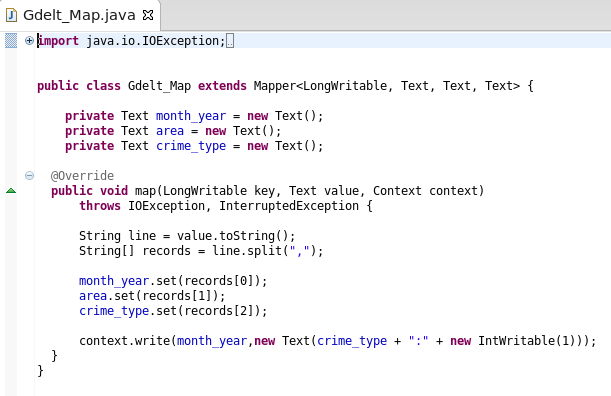


The uploaded file later will be used to run MapReduce programs.

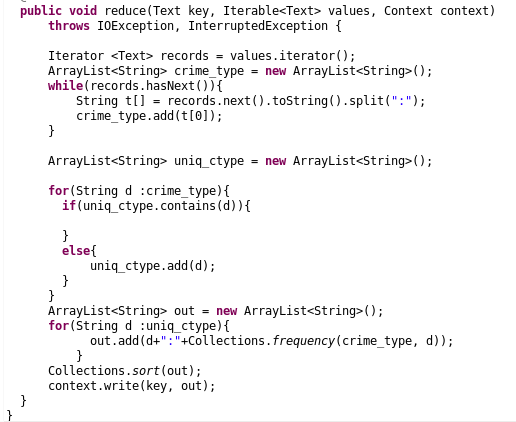
**MapReduce I:**

**Problem Set:** To get the month-wise crime name and count in alphabetical order.

**Mapper:**



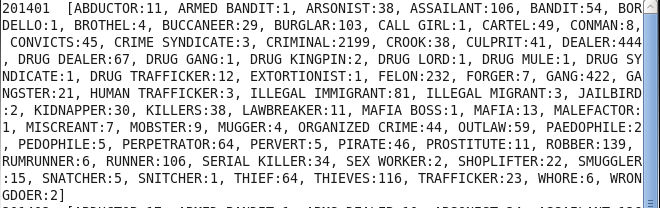
**Reducer:**



The following was the command passed to run the MapReduce:



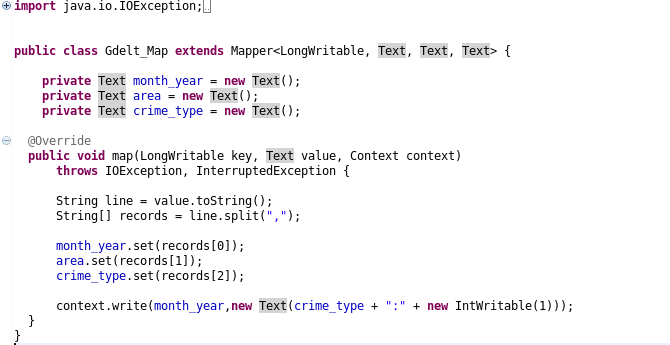
**Output:**



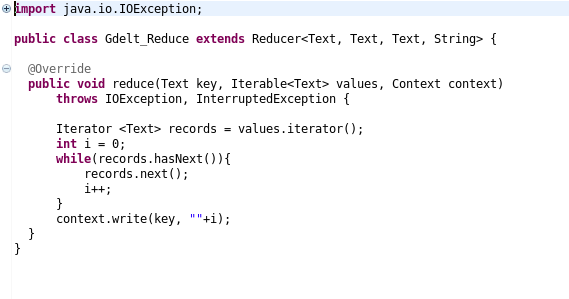
**MapReduce II:**

**Problem Set:** Get the month-wise crime count (total).

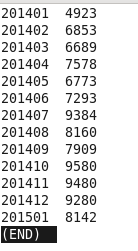
**Mapper:**



**Reducer:**



**Output:**



The value “201401” be read as “2014 January”. Hence, the total amount of crime committed in the month of January 2014 was 4923.

**Hive:**

In order to run hive queries, first the data needs to be imported into Hive. The following queries were used in order to populate data into HIVE:

Database creation:

>CREATE DATABASE IF NOT EXISTS db1;

>CREATE TABLE IF NOT EXISTS tb1 (monthyear String, area\_code String, crime\_type String)

>COMMENT “GDELT DB”

>ROW FORMAT DELIMITED

>FIELDS TERMINATED BY ‘,’

>LINES TERMINATED BY ‘\n’

>STORED AS TEXTFILE;

Inserting data into database:

>LOAD DATA LOCAL INPATH ‘/home/training/Desktop/data.csv’

>OVERWRITE INTO TABLE tb1;

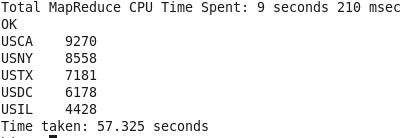
**Hive I:**

**Problem Set:** Get the top 5 states with the highest crime.

**Query:**



**Result:**



As per the GDELT data, California (CA) has the highest crime rate of 9270, followed by New York (NY) at 8558, Texas (TX) at 7181, Washington (DC) at 6178and Illinois (IL) at 4428.