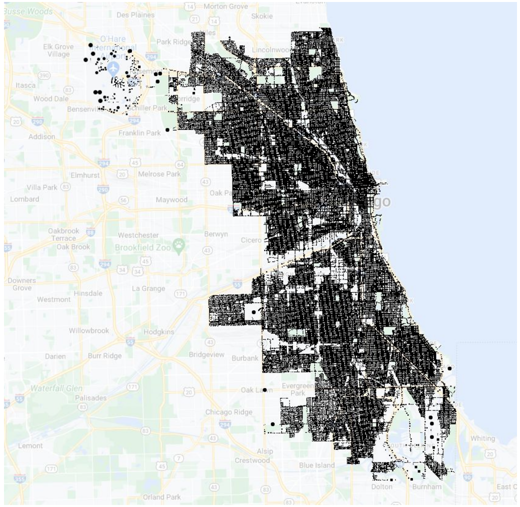
Chicago Crimes Project

Kim Morris, Neha Mathews, and Cynthia Wen
Group 11

Project Introduction



Tasks

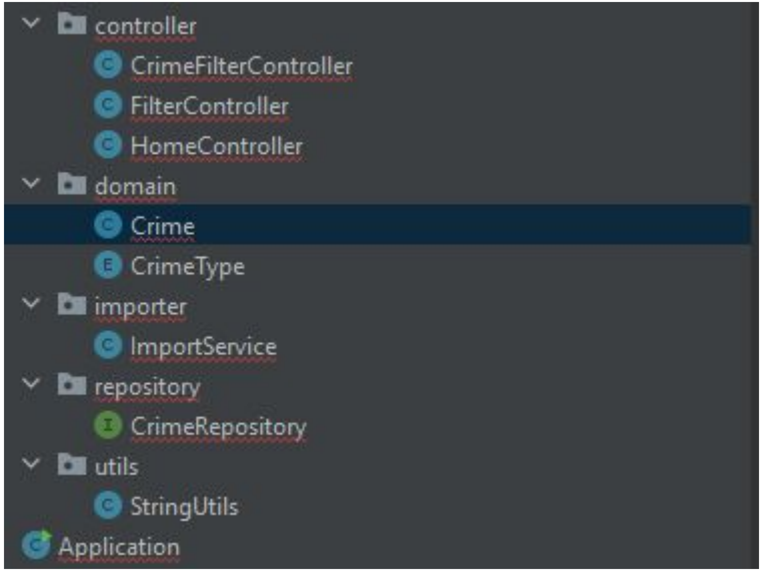
- Analyze the Chicago Crimes dataset
- Clean the data and convert to parquet file
- Create a choropleth map of the number of crimes per ZIP Code
- Create a bar chart of the number of each type of crime that occured within a specific date range

Big Data System

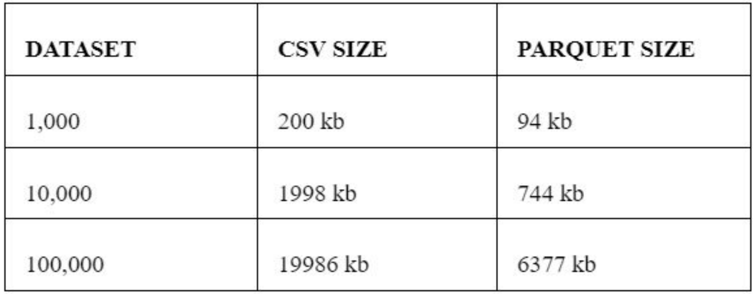
- SparkSQL
  - Easier to write analytical queries
  - Parquet files are column-formatted and recommended for analytical queries

Task 1

- Clean up the dataset- Controller
  - Removing empty entry or duplicates
  - Remove unnecessary columns
  - Remove null value
- Load the Zip code data
- Create parquet files for 1K, 10k, 100k data
  - Size decrease drastically



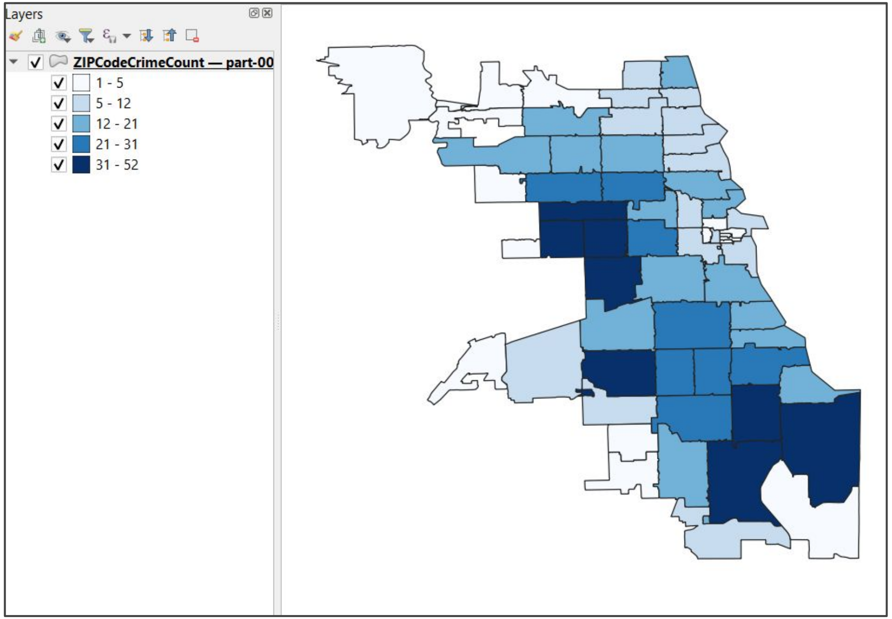
```
controller
    CrimeFilterController
    FilterController
    HomeController
domain
    Crime
    CrimeType
importer
    ImportService
repository
    CrimeRepository
utils
    StringUtils
Application
```



DATASET	CSV SIZE	PARQUET SIZE
1,000	200 kb	94 kb
10,000	1998 kb	744 kb
100,000	19986 kb	6377 kb

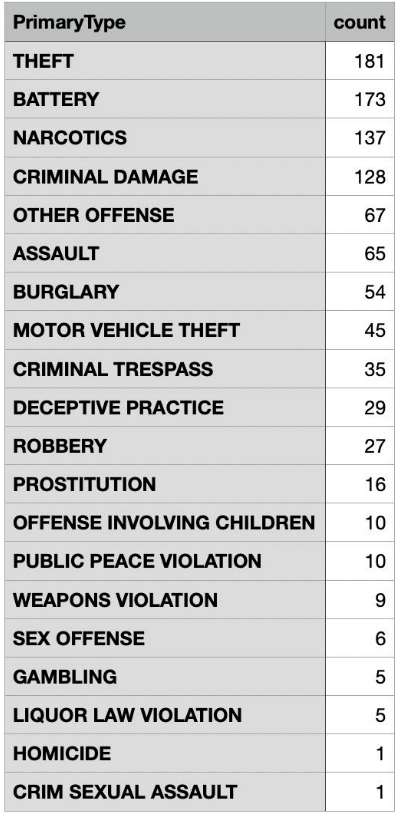
Task 2

- Load the parquet file into a BeastScala project
- Create a view using an SQL query that selected the ZIPCode and count of all crimes grouped by ZIPCode
- Use Beast to load the ZIP Code dataset and convert it to a dataframe
- Join the two views using an equi-join query and save the output as a Shapefile
- Import the file into QGIS to produce a choropleth map

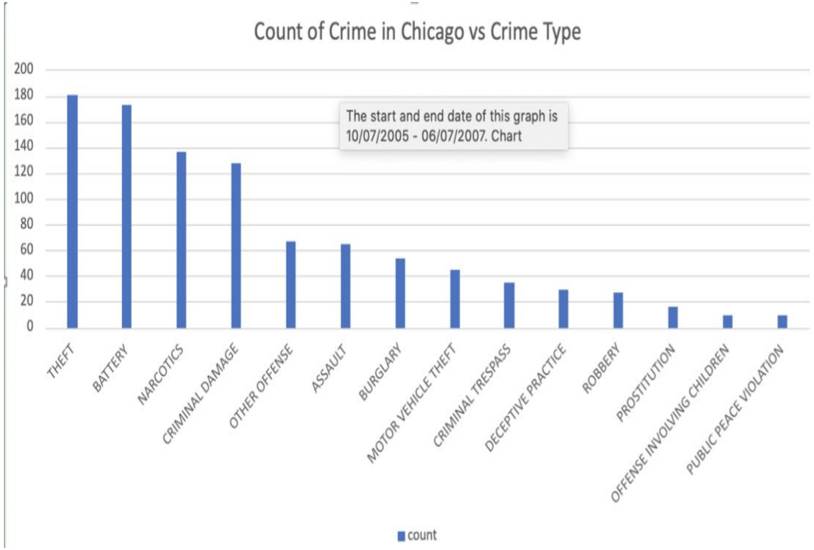


Task 3

- Read the Chicago Crimes 10K Parquet File from task 1
- Write an SQL Query to filter data of crimes committed between start/end date which are command-line arguments
- GroupBy() and Aggregation
- Output is a csv file with the crime and the count of each crime
  - Use data to create a bar chart in excel



PrimaryType	count
THEFT	181
BATTERY	173
NARCOTICS	137
CRIMINAL DAMAGE	128
OTHER OFFENSE	67
ASSAULT	65
BURGLARY	54
MOTOR VEHICLE THEFT	45
CRIMINAL TRESPASS	35
DECEPTIVE PRACTICE	29
ROBBERY	27
PROSTITUTION	16
OFFENSE INVOLVING CHILDREN	10
PUBLIC PEACE VIOLATION	10
WEAPONS VIOLATION	9
SEX OFFENSE	6
GAMBLING	5
LIQUOR LAW VIOLATION	5
HOMICIDE	1
CRIM SEXUAL ASSAULT	1



The start and end date of this graph is 10/07/2005 - 06/07/2007. Chart

Thank you!!!

Multiple Choice Question

Why did our team choose to use SparkSQL as our Big Data System for all three tasks in this project?

A: Using SparkSQL made running analytic queries on the Parquet files easier

B: Using SparkSQL made running transactional queries easier

C: Using SparkSQL made running the Parquet files easier to use since it is row formatted data