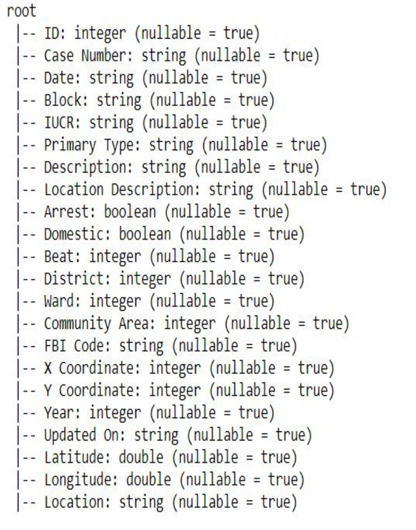
CRIME RATE ANALYSIS IN CHICAGO

Group 12: Ankur Shukla, Santhosh Kumar Nagarajan, Lukas Enander, Arianna Delsante

May 23, 2018

Uppsala University, 2018

Data and analysis



```
root
 |-- ID: integer (nullable = true)
 |-- Case Number: string (nullable = true)
 |-- Date: string (nullable = true)
 |-- Block: string (nullable = true)
 |-- IUCR: string (nullable = true)
 |-- Primary Type: string (nullable = true)
 |-- Description: string (nullable = true)
 |-- Location Description: string (nullable = true)
 |-- Arrest: boolean (nullable = true)
 |-- Domestic: boolean (nullable = true)
 |-- Beat: integer (nullable = true)
 |-- District: integer (nullable = true)
 |-- Ward: integer (nullable = true)
 |-- Community Area: integer (nullable = true)
 |-- FBI Code: string (nullable = true)
 |-- X Coordinate: integer (nullable = true)
 |-- Y Coordinate: integer (nullable = true)
 |-- Year: integer (nullable = true)
 |-- Updated On: string (nullable = true)
 |-- Latitude: double (nullable = true)
 |-- Longitude: double (nullable = true)
 |-- Location: string (nullable = true)
```

The chosen dataset reports crimes in Chicago from year 2001 up to now

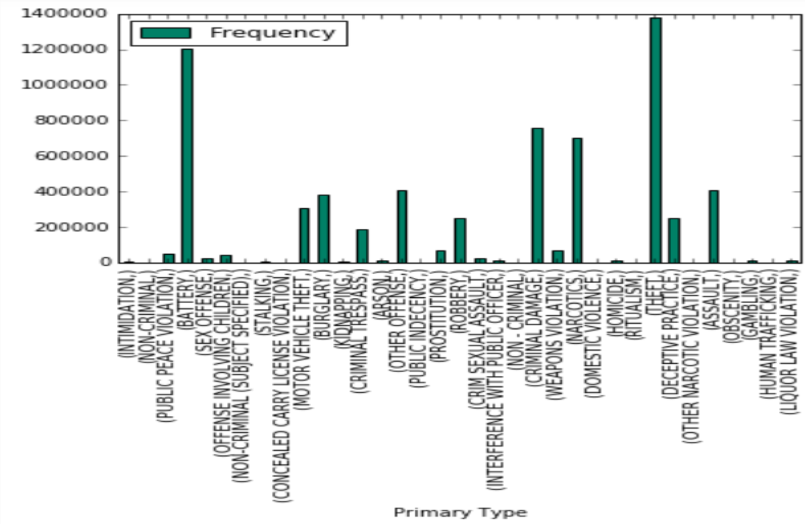
- 6,6 million reported crimes (rows)
- 22 columns (Type of crime, location, arrested, etc)

Data and analysis

Possible analysis of the data:

- Which type of crime is the most frequent
- Number of crimes per year
- Where most of the crimes take place

Preliminary results



Technology and methods

We choosed to work with Apache Spark.

Pypark is used for analyzing the dataset due to its **in-memory** RDD computation technique, **scalability** and **fault tolerance**.

- Unused columns were dropped to reduce memory consuption.
- First worked on a smaller version of the dataset for testing purposes.
- Lastly we used PySpark's functions to calculate various statistics e.g. the most common primary crime type, where do most crimes take place.

Scalability studies

Proposed scalability experiments:

- Check the run time for one dataset over different clusters holding up to 3 nodes.
- Increase the size of the dataset and observe the what effect does it have on different cluster configuration?

Another Dataset

Red Light Camera Violations dataset reflects the daily volume of violations created by the City of Chicago Red Light Program for each camera from 1st July 2014 to present.

- 3,74,756 rows with around 2,081,230 violations
- 10 columns (Camera ID, violation count, address, violation date, etc)

Possible analysis of the data:

- Compare the number of violations over Camera ID or locations.
- Compare the total violation numbers over each year (2014 - Present)