PES UNIVERSITY

100 feet Ring Road, BSK 3rd Stage Bengaluru 560085

Department of Computer Science and Engineering

B. Tech. CSE - 6th Semester Jan – May 2022

UE19CS344 DATABASE TECHNOLOGIES (DBT)

Project Report

Crime data analysis

PES2UG19CS013: Achyut Jagini PES2UG19CS230: Melavoy Nithin Reddy PES2UG19CS189: Koduru Bharath Subba Reddy

Table of Contents

1. Introduction
2. Installation of Software
3. Input Data
   1. Source
   2. Description
4. Streaming Mode Experiment
   1. Description
   2. Windows
   3. Results
5. Batch mode Experiment
   1. Description
   2. Data Size
   3. Results
6. Comparison of Streaming vs Batch mode Results
7. Conclusion
8. References

Introduction

This project aims to perform analysis of crime data using spark and kafka.

Installation of software

Software installed and used are Spark

Kafka Zookeeper

Kafka Configuration Manager(CMAK) Mongodb database

Input data Source

For spark streaming ,data input from crime-train.csv and test.csv datasets.

For kafka ,data produced and consumed police-department-calls-for-service.json

Description

**Train.csv** has columns Dates ,Category, Descript, DayOfWeek,PdDistrict,Resolution,Address,X,Y.

**police-department-calls-for-service.json**

Is json data with format "crime\_id": "183653763",

"original\_crime\_type\_name": "Traffic Stop", "report\_date": "2018-12-31T00:00:00.000", "call\_date": "2018-12-31T00:00:00.000", "offense\_date": "2018-12-31T00:00:00.000",

"call\_time": "23:57",

"call\_date\_time": "2018-12-31T23:57:00.000", "disposition": "ADM",

"address": "Geary Bl/divisadero St", "city": "San Francisco",

"state": "CA",

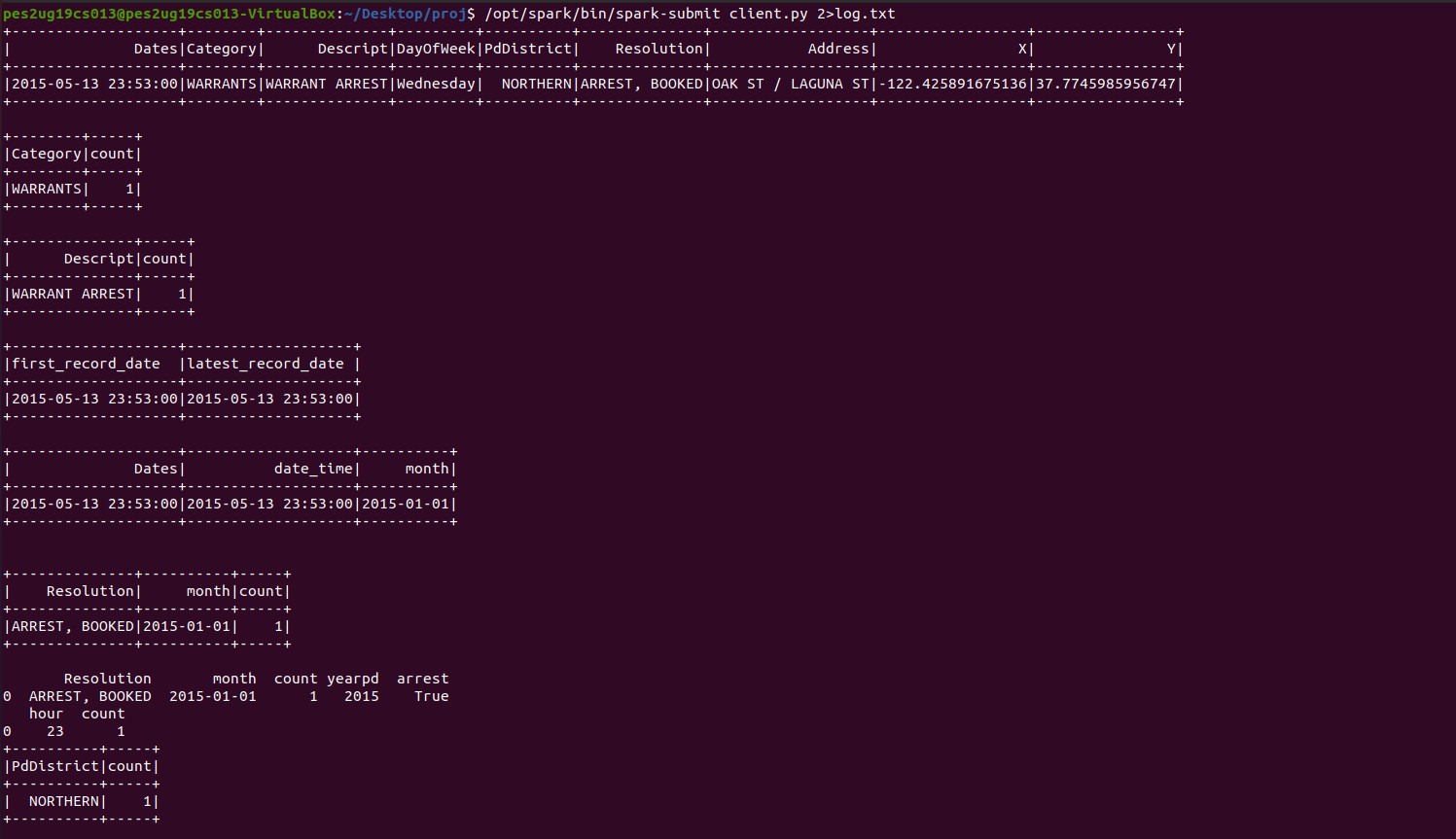
"agency\_id": "1",

"address\_type": "Intersection","common\_location": ""

Real time processing

Total time-1220:33:35 5:00s/it



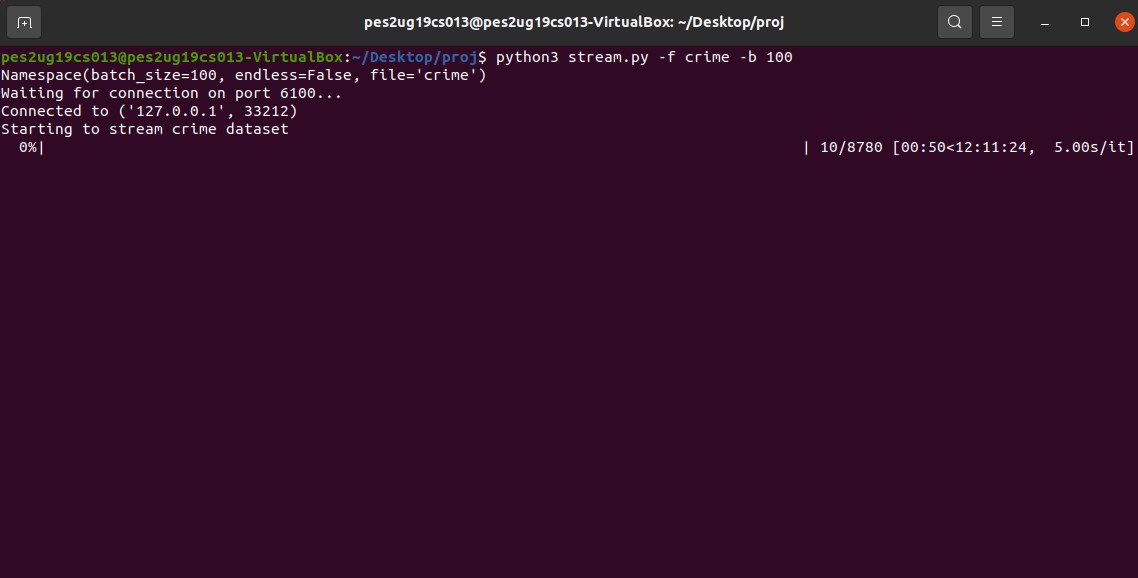


Batch Streaming experiment

Data streamed using stream.py file

python3 stream.py -f crime -b 100(can vary 100 window size)

Total time-12:11:24 5.00 s/iter window size 100

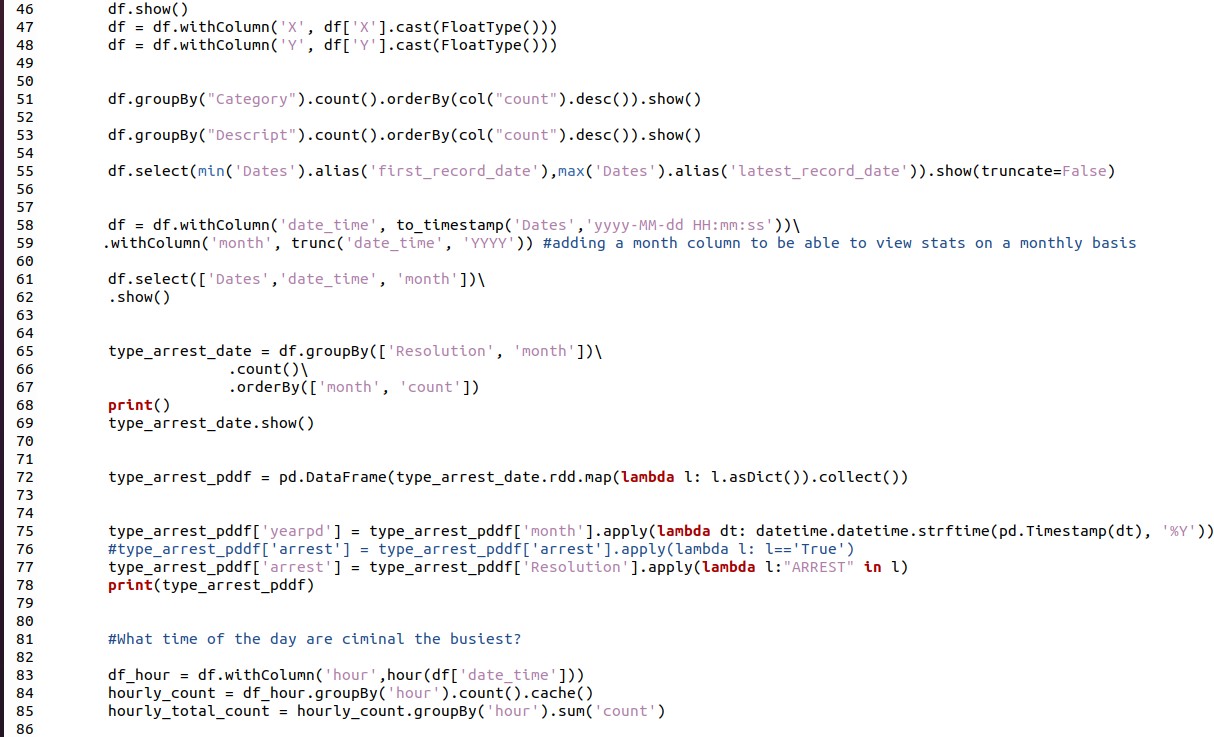


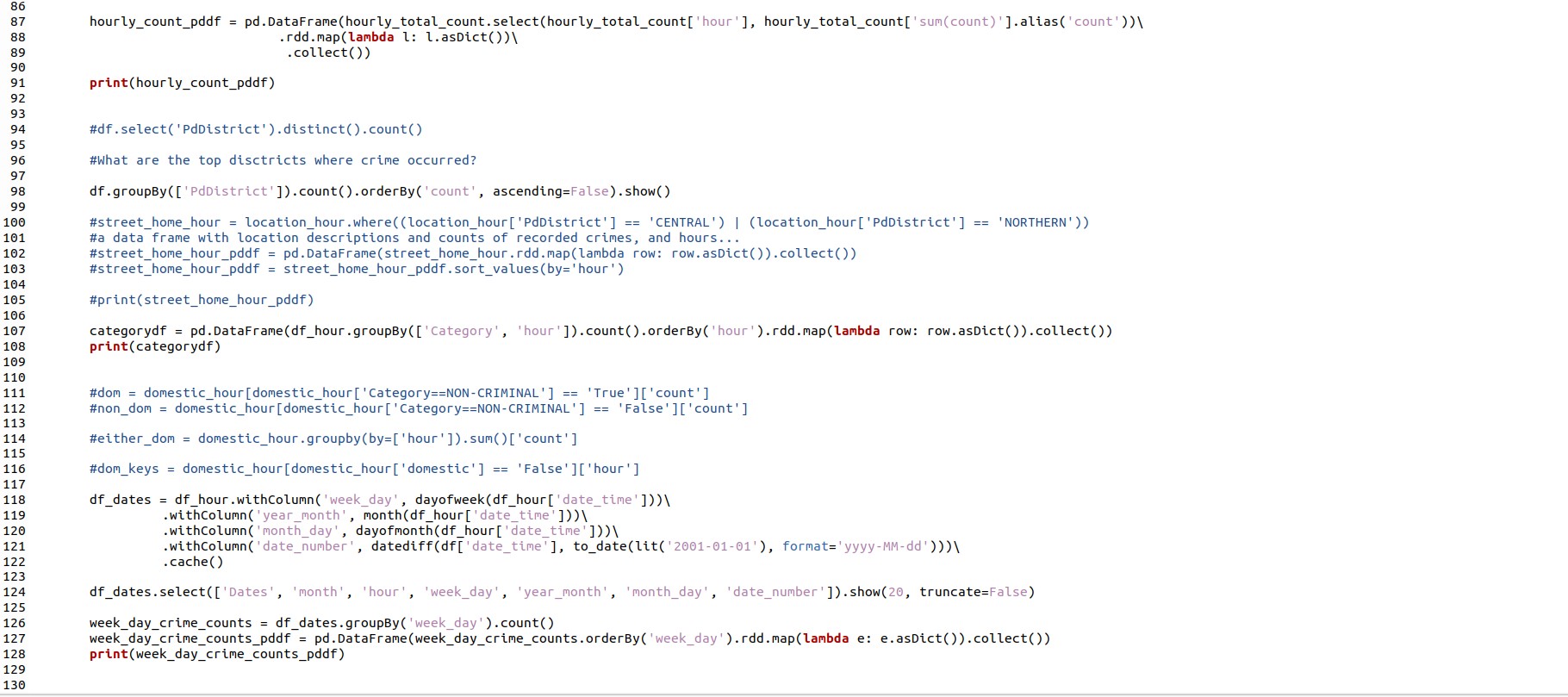
client.py

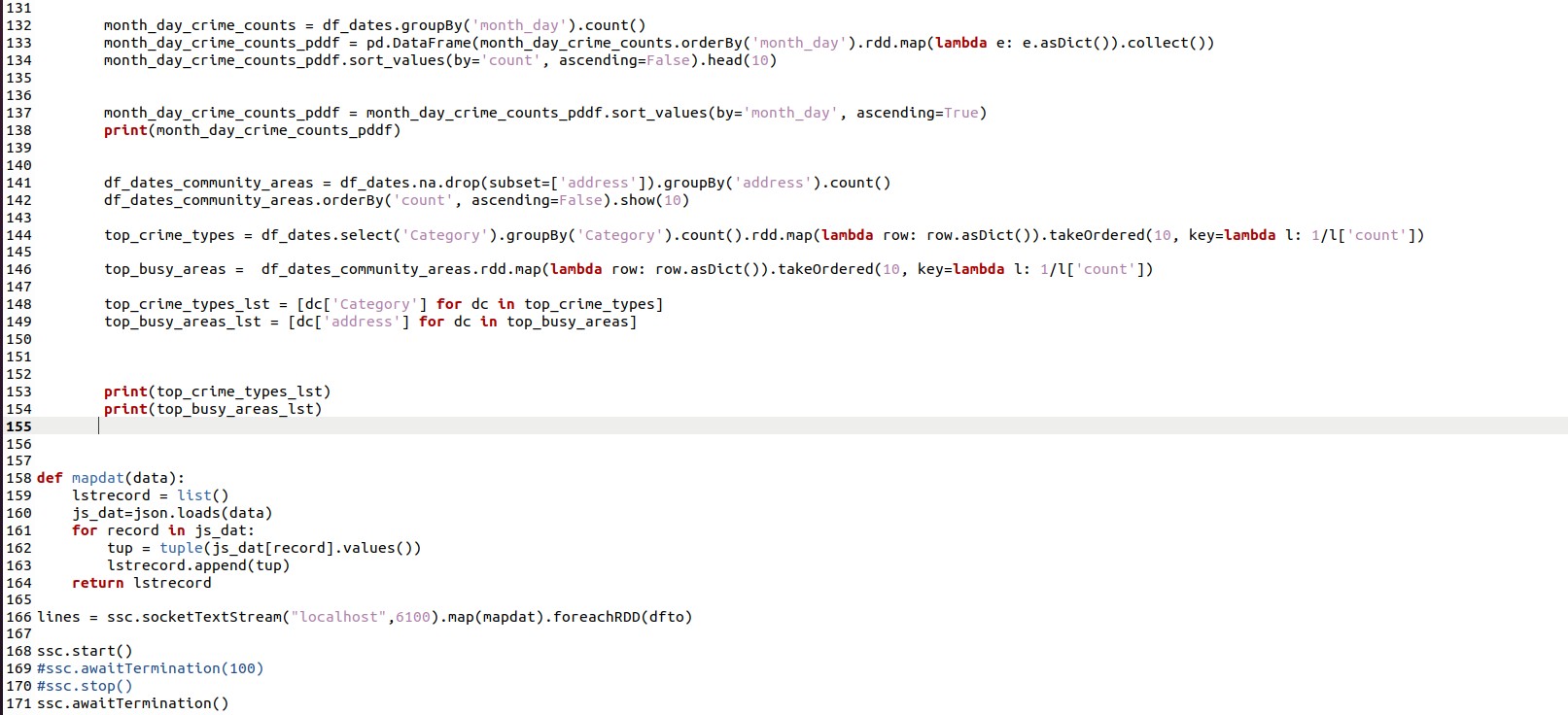
/opt/spark/bin/spark-submit client.py 2>log.txt

Code

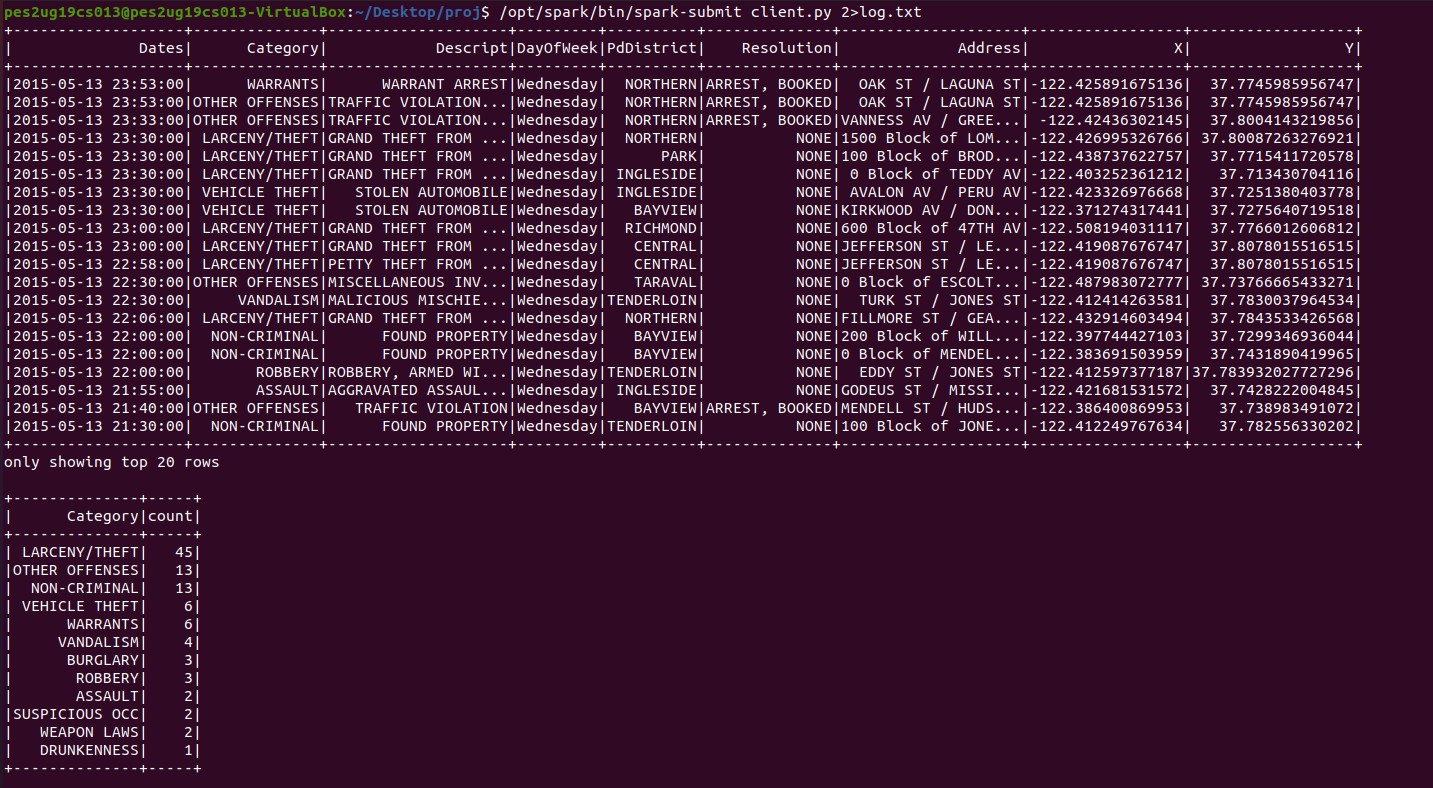


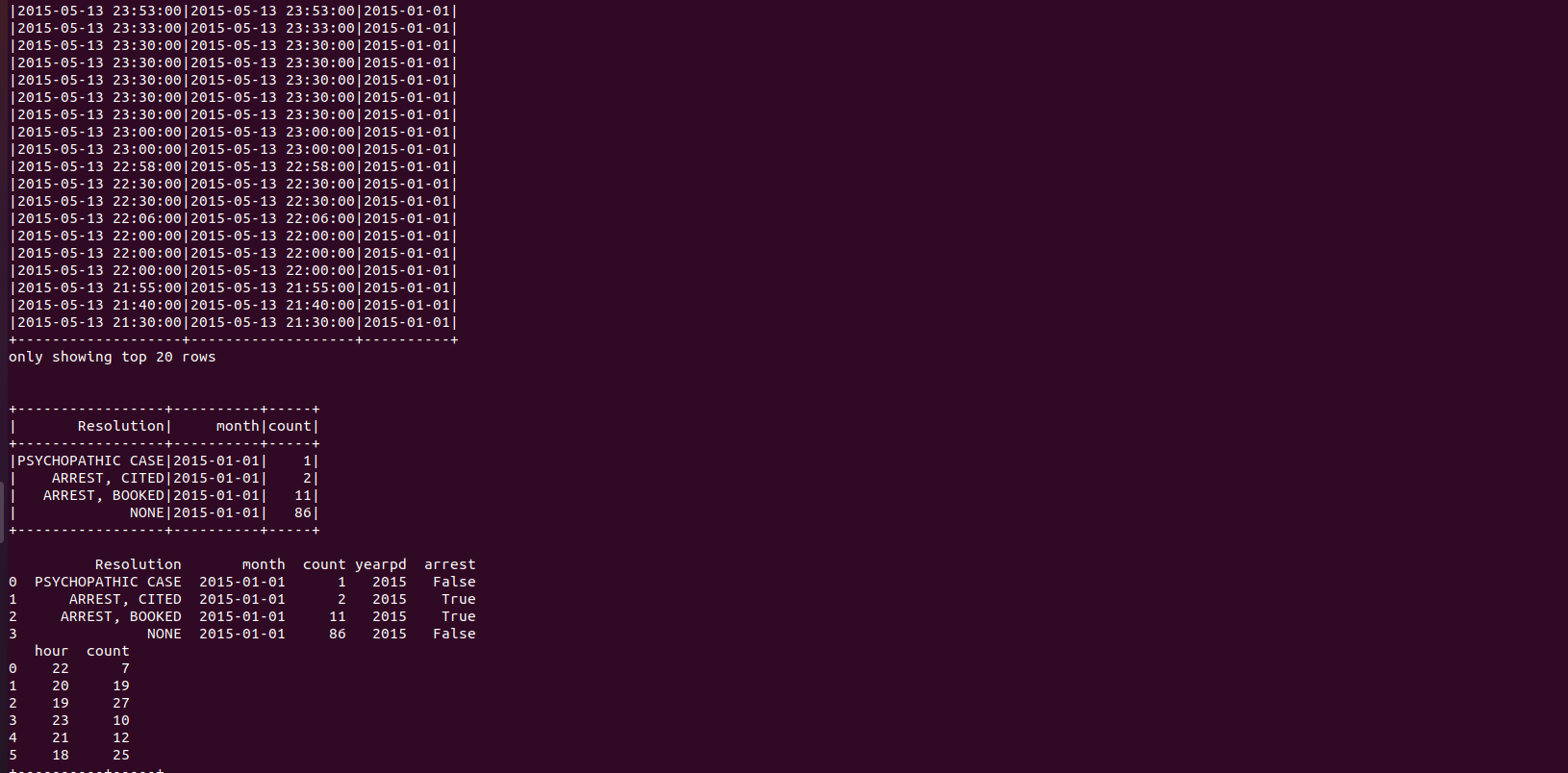






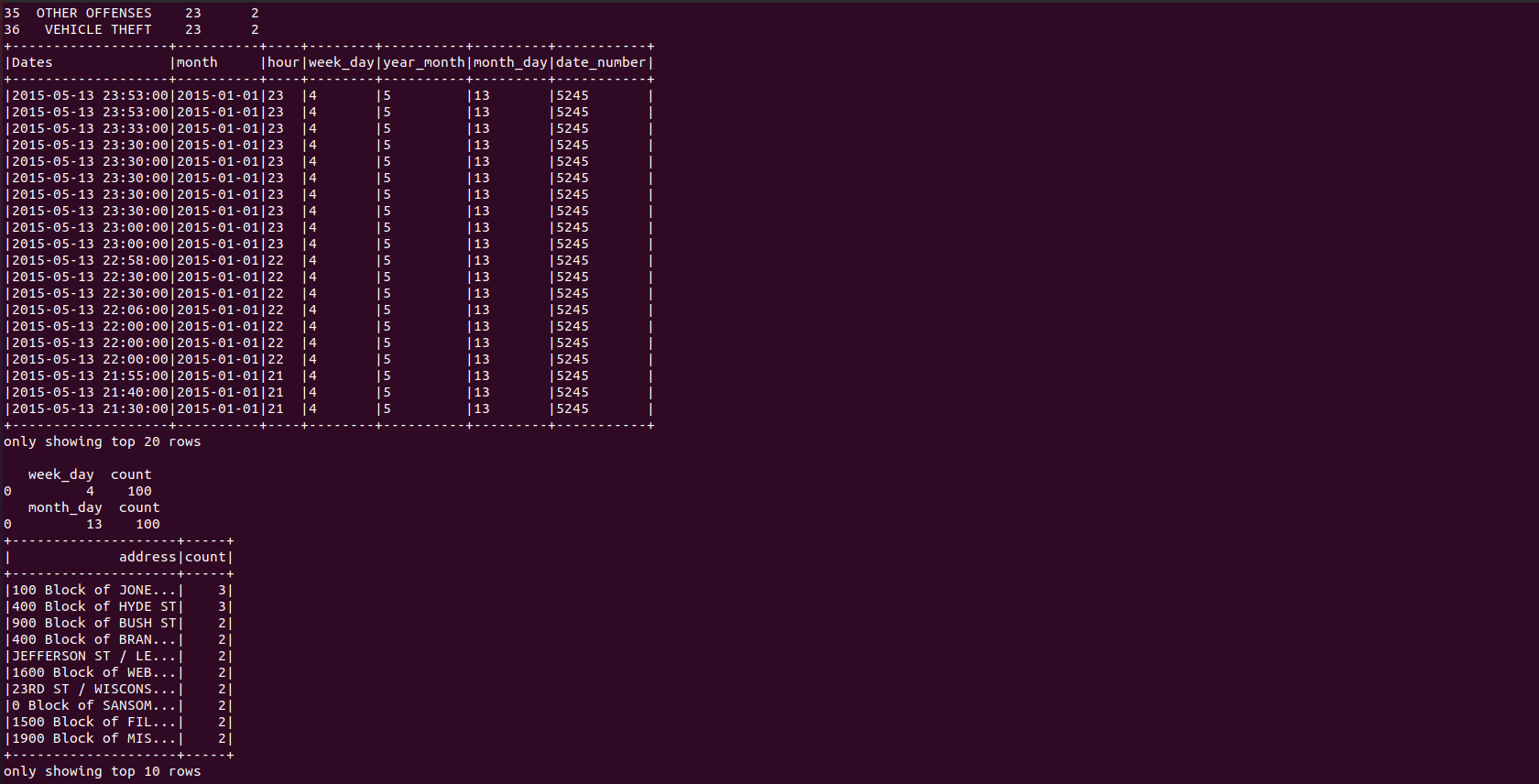
Output

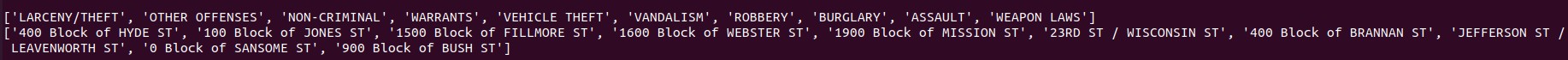






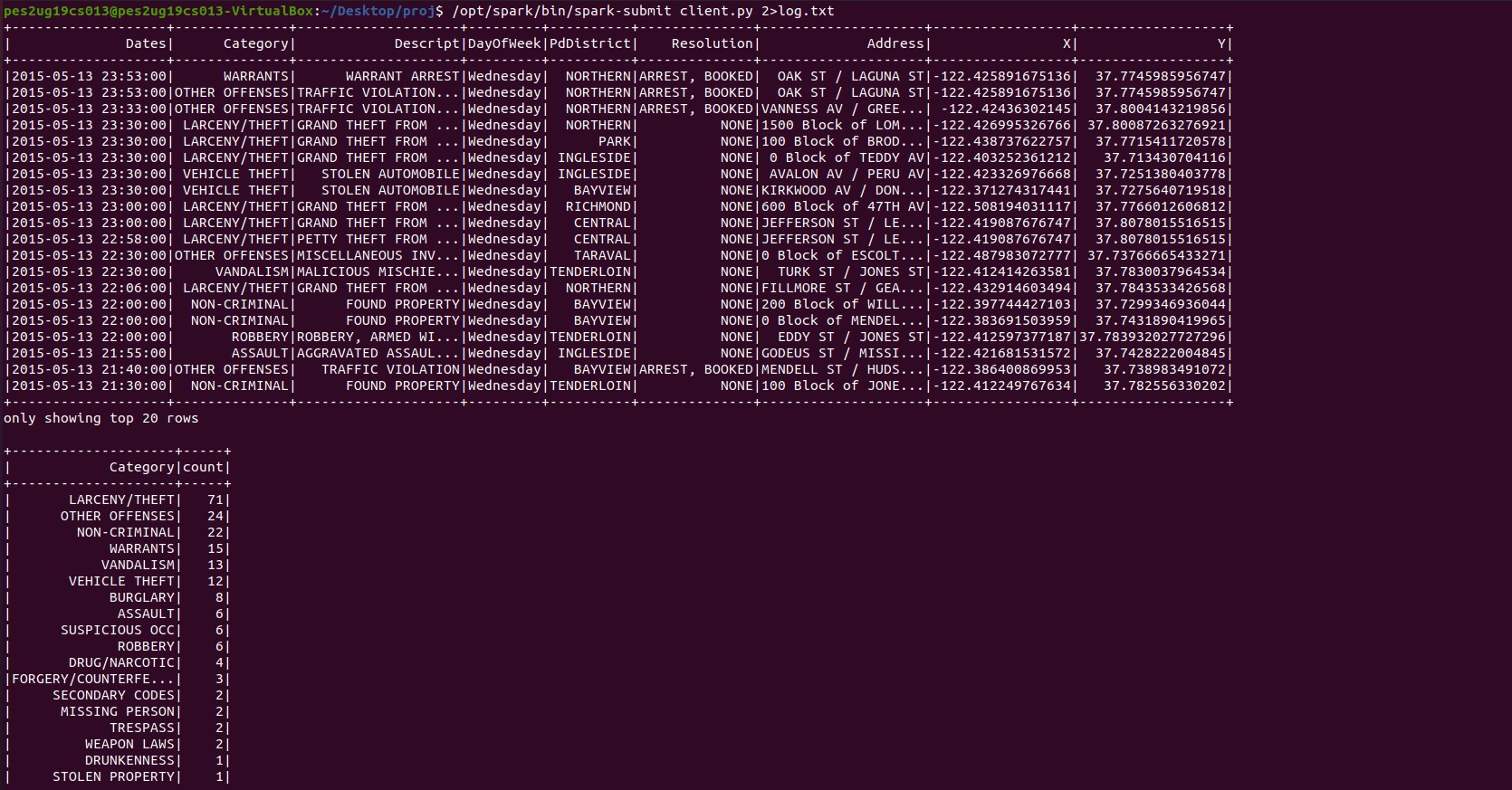


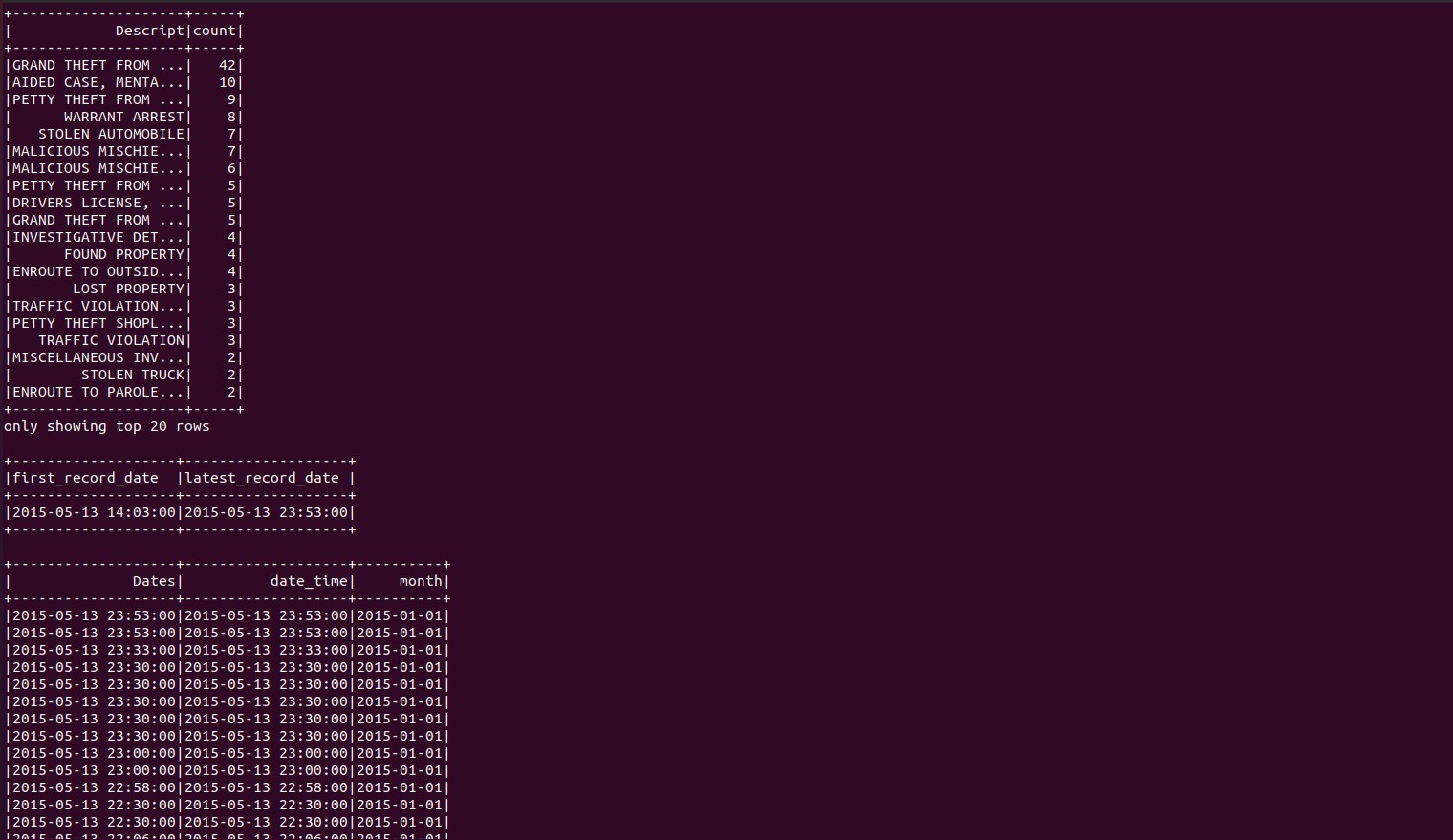


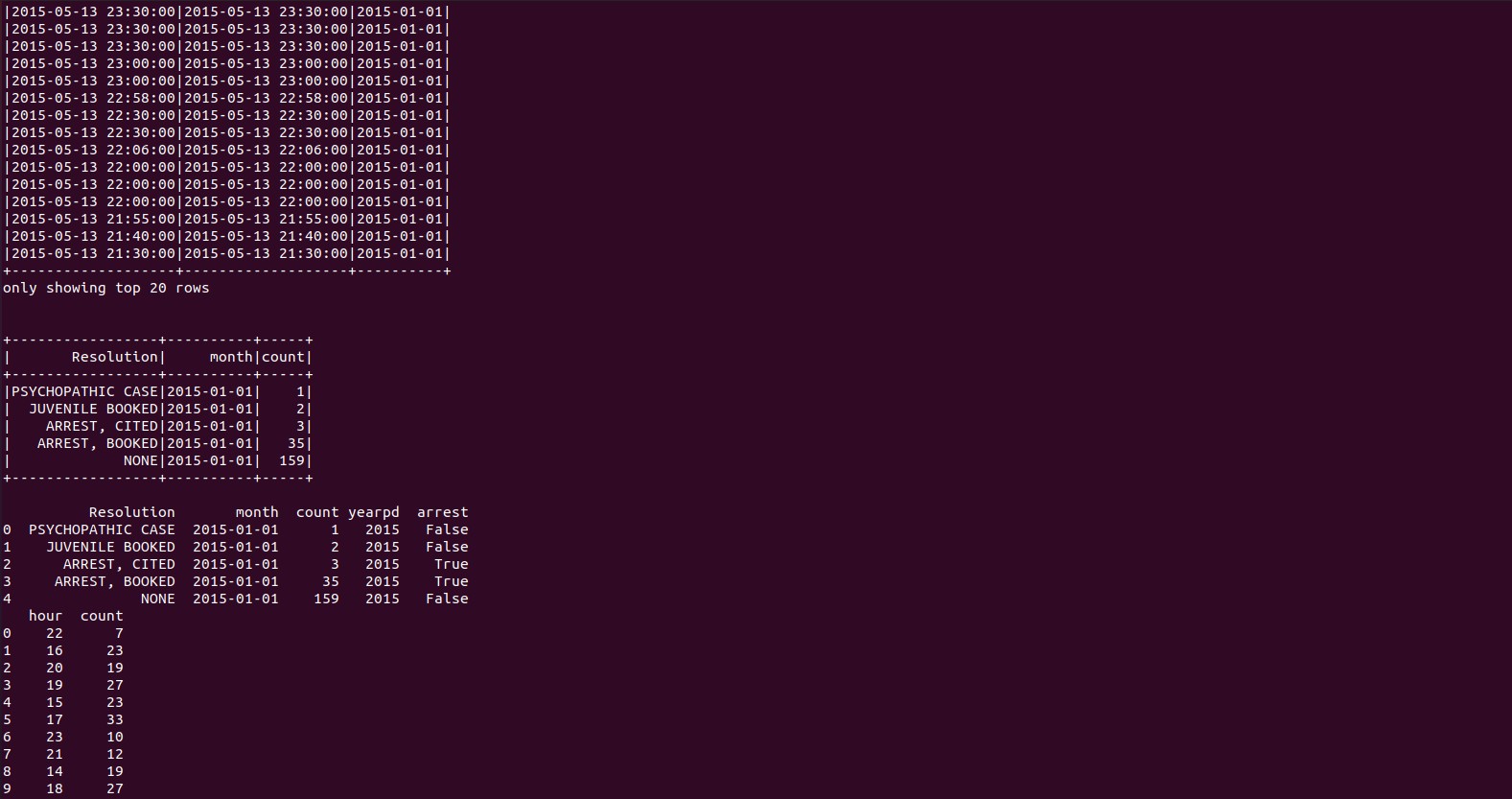


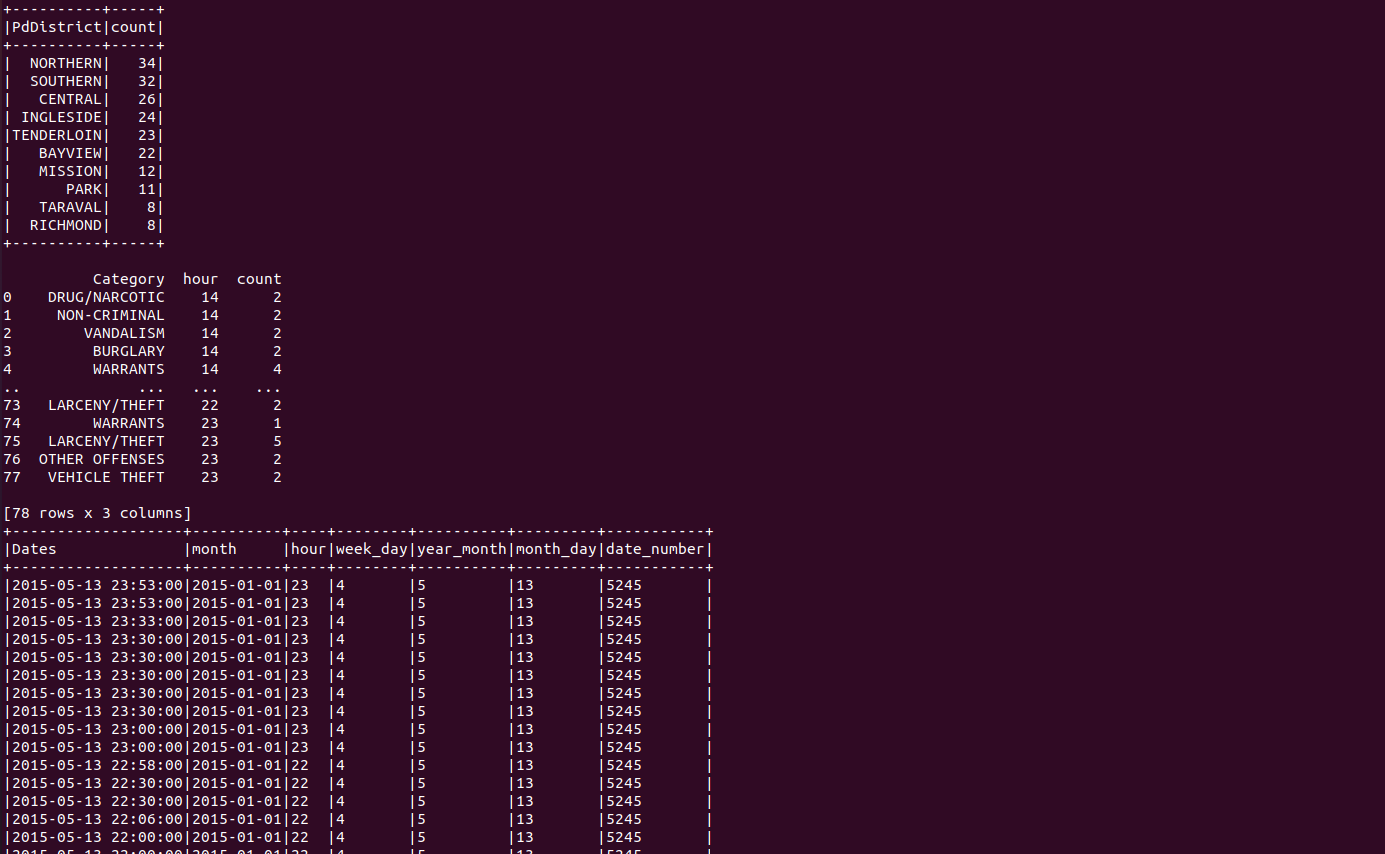
window size 200 Total time-6:05:21

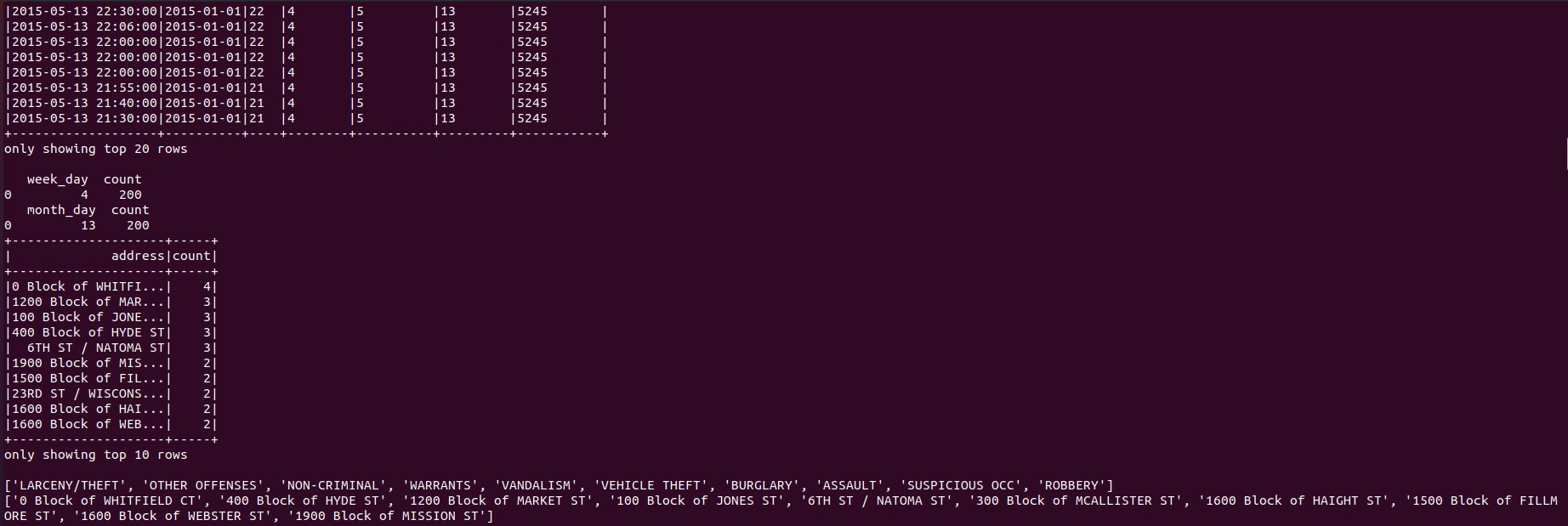
Results









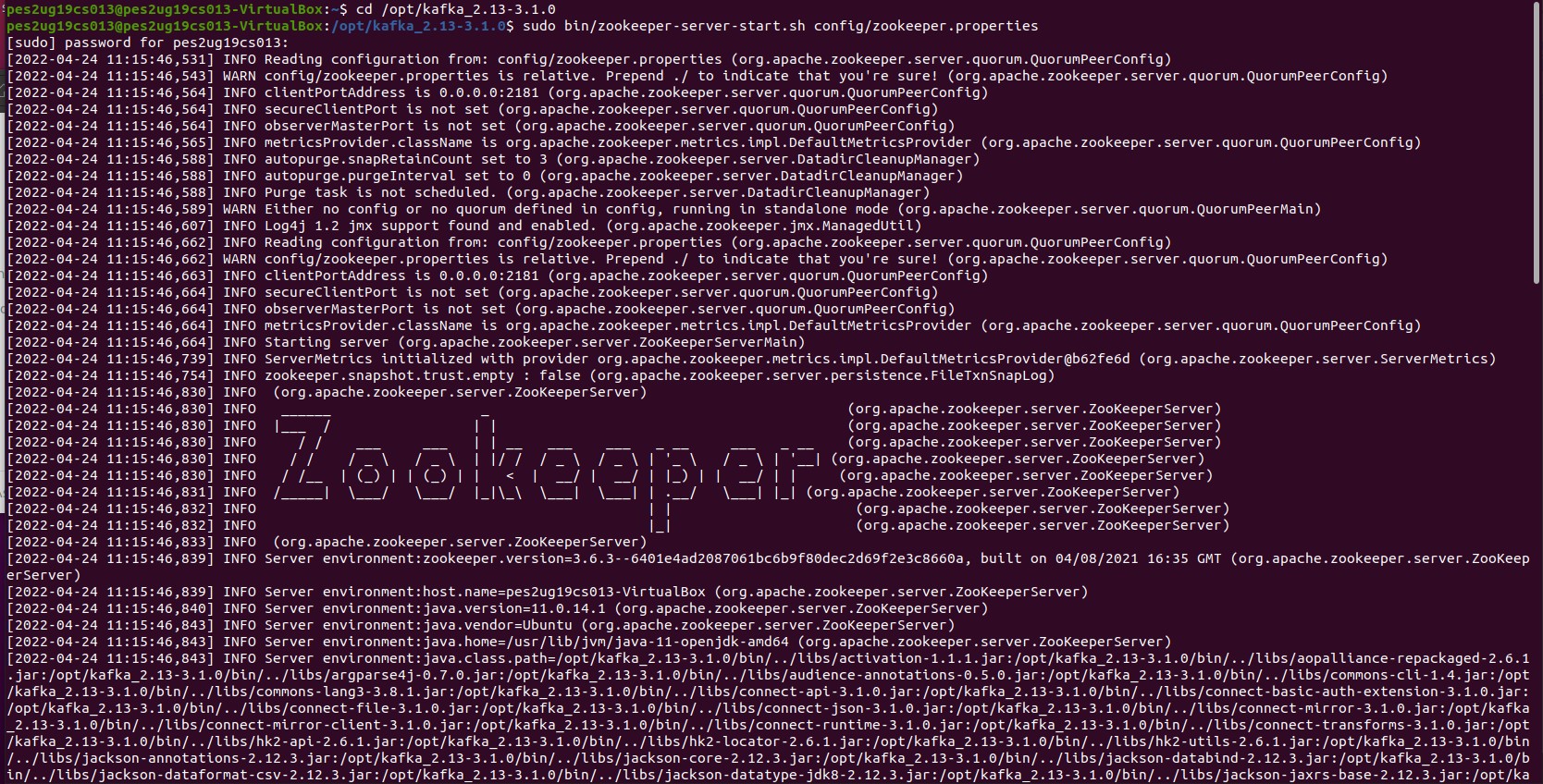


Kafka producer and consumer

Start Zookeeper

1)cd /opt/kafka\_2.13-3.1.0

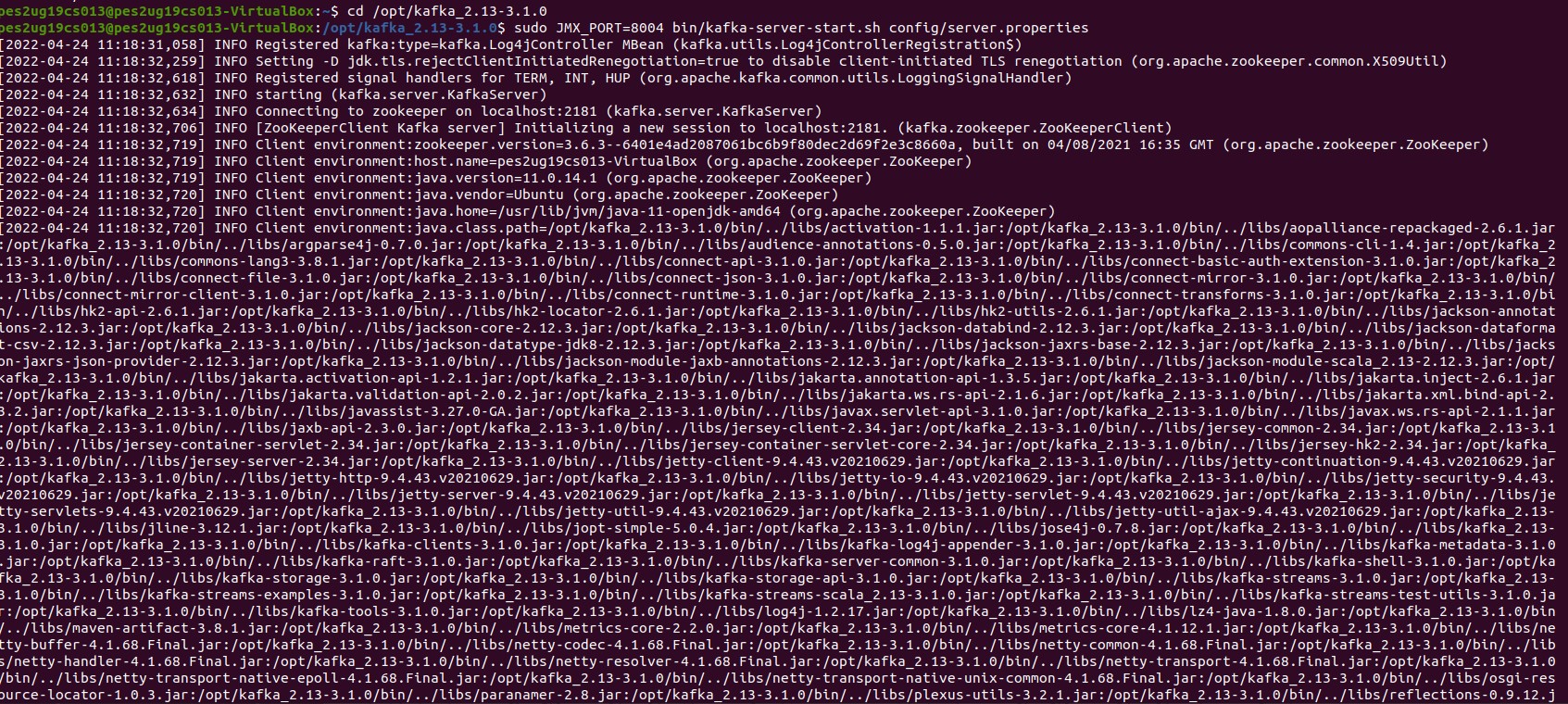
2)sudo bin/zookeeper-server-start.sh config/zookeeper.properties



Start kafka server

cd /opt/kafka\_2.13-3.1.0

sudo JMX\_PORT=8004 bin/kafka-server-start.sh config/server.properties

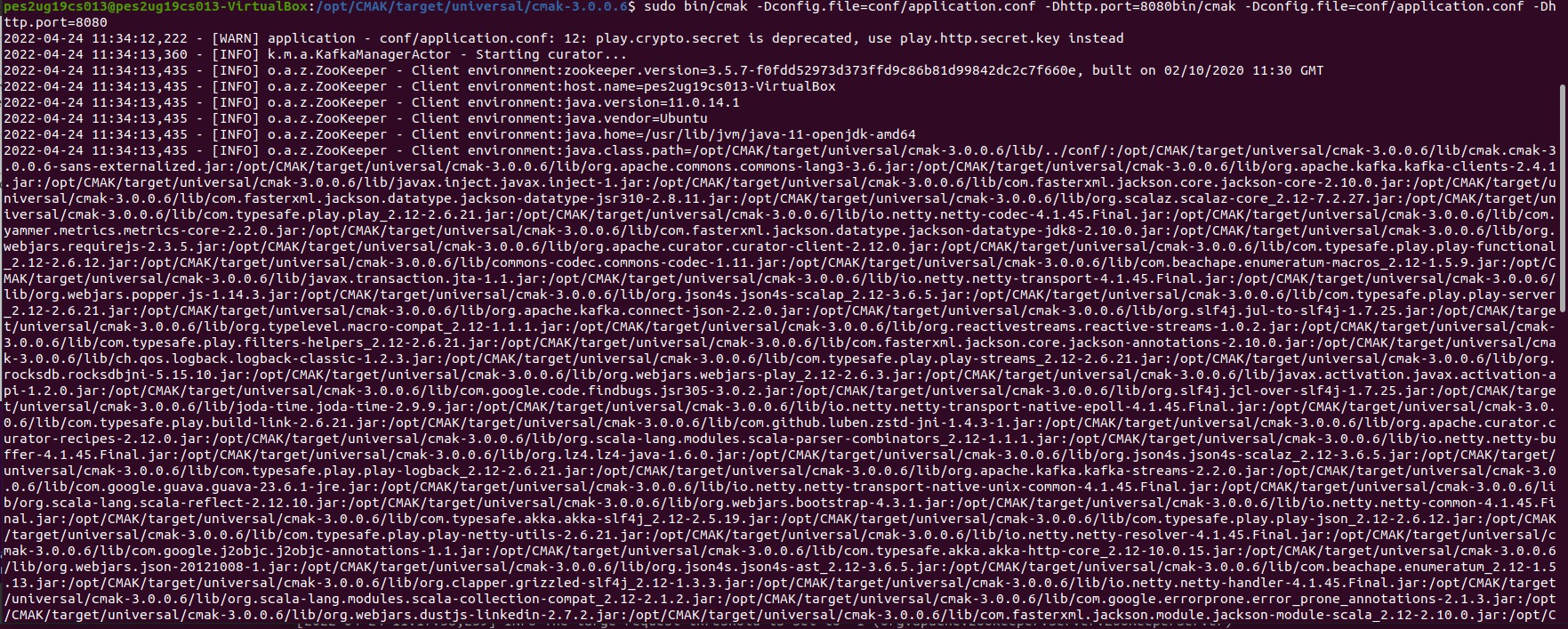


Start kafka configuration manager(CMAK) cd /opt/CMAK/target/universal/cmak-3.0.0.6

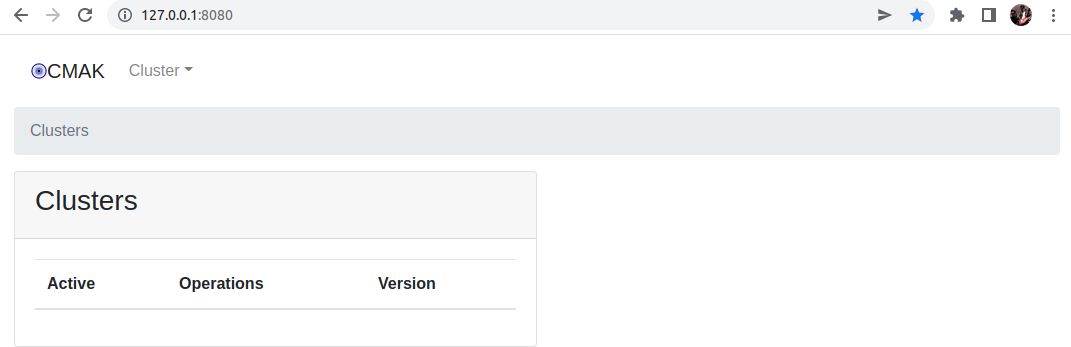
sudo bin/cmak -Dconfig.file=conf/application.conf

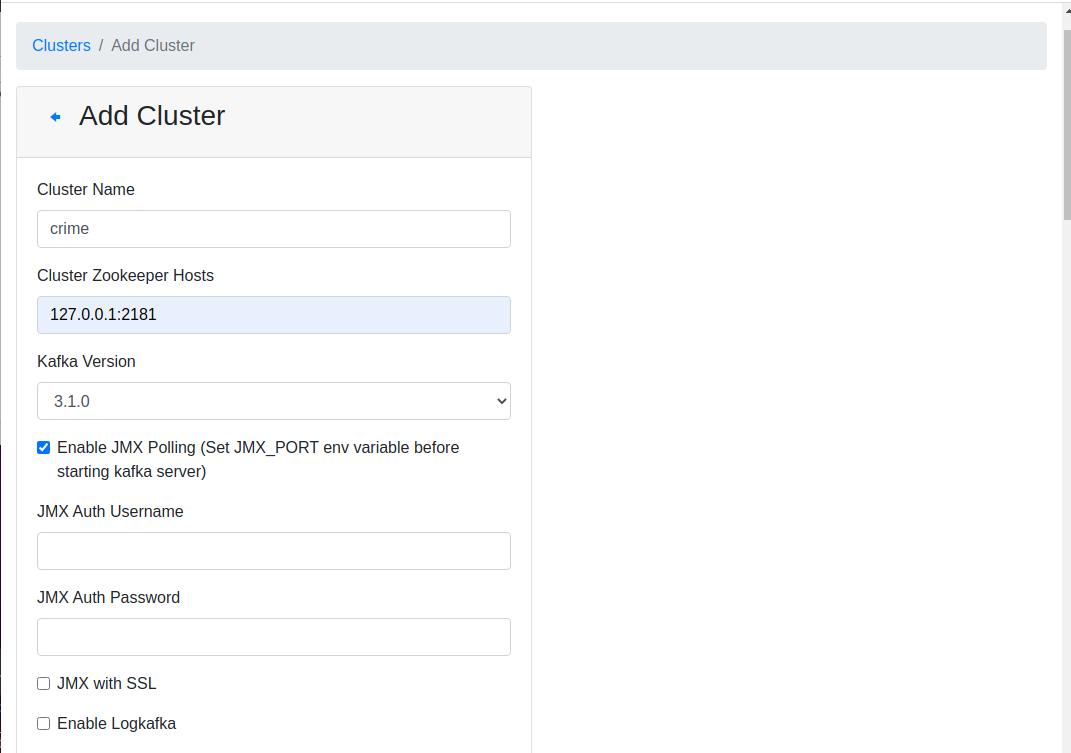
-Dhttp.port=8080bin/cmak

-Dconfig.file=conf/application.conf -Dhttp.port=8080



http://127.0.0.1:8080





run producer

python3 kafka\_server.py

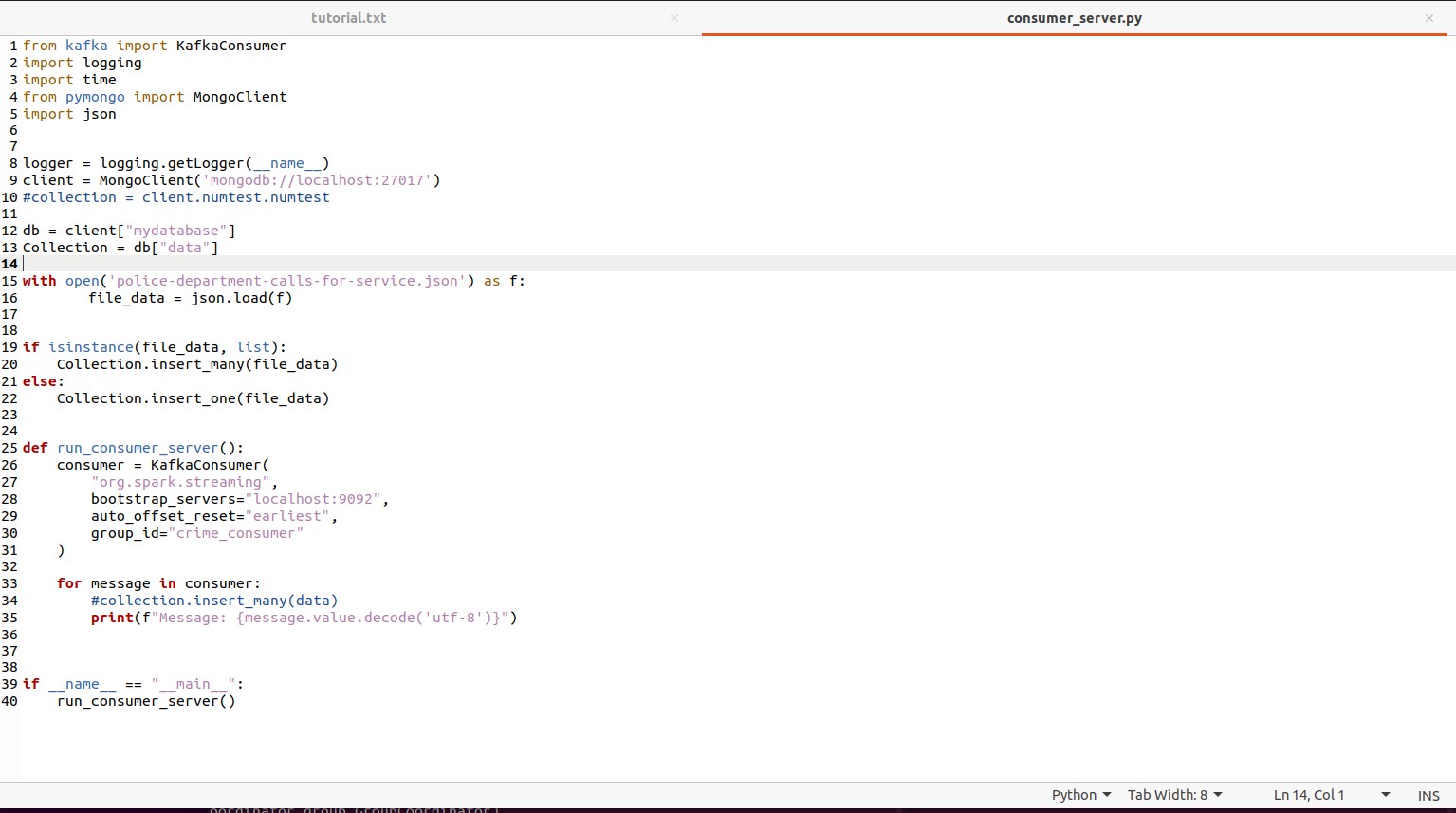


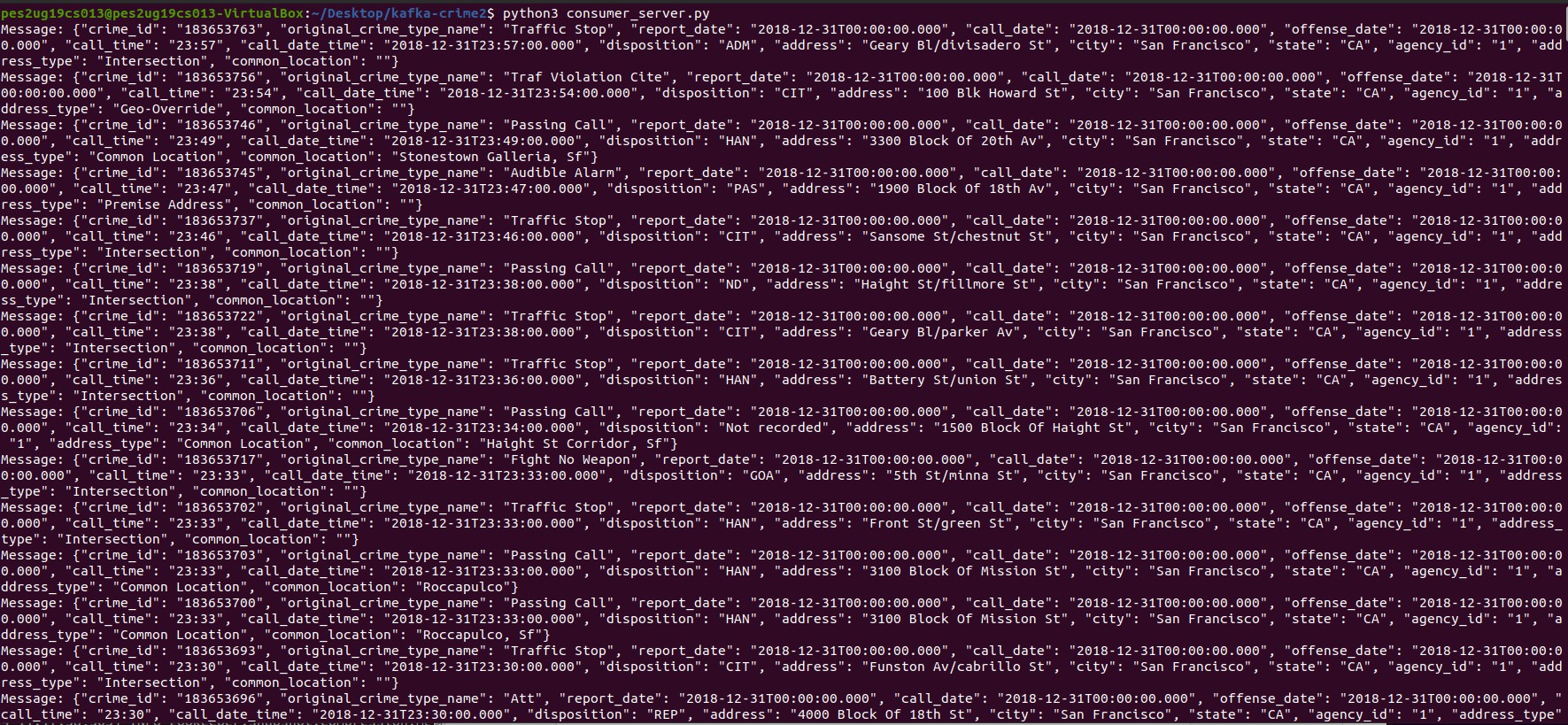
producer\_server



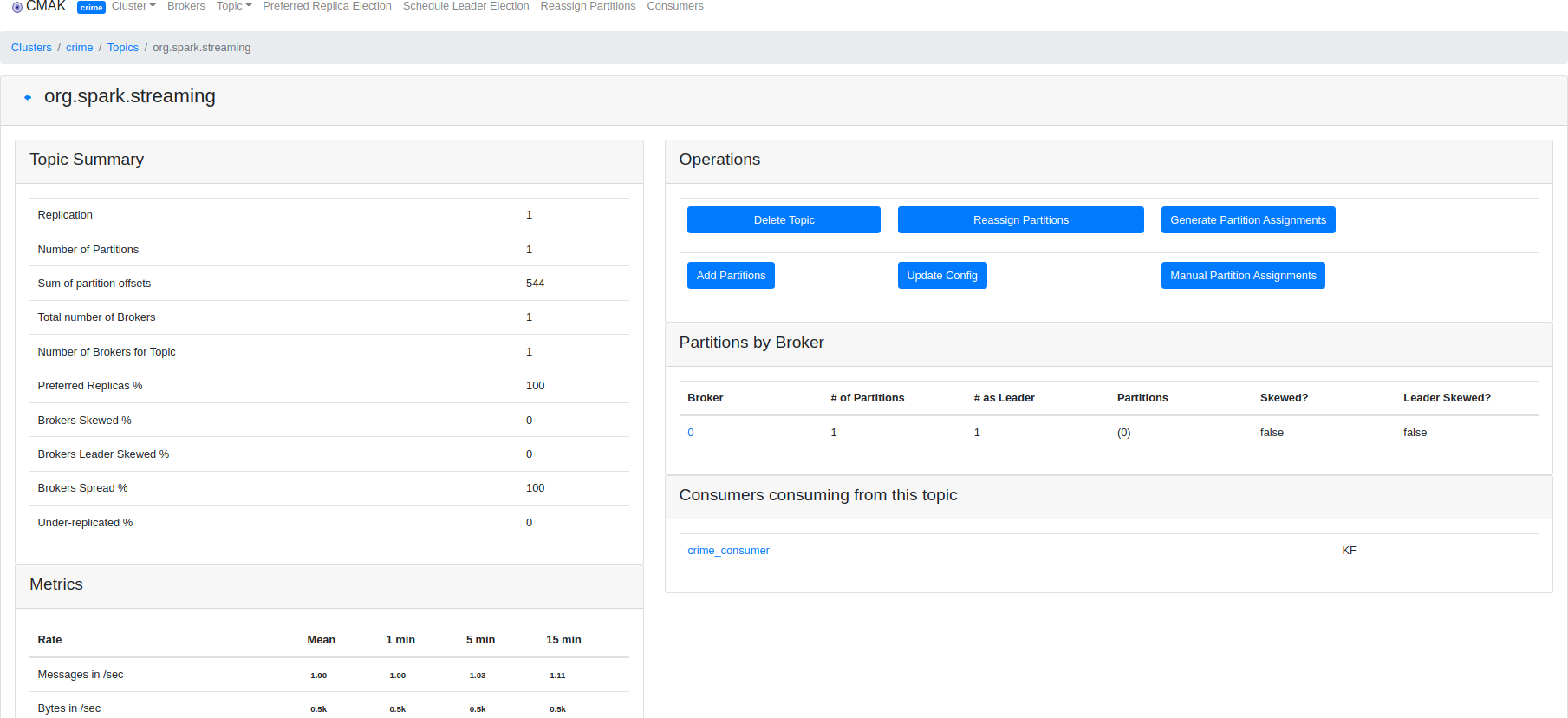
run consumer

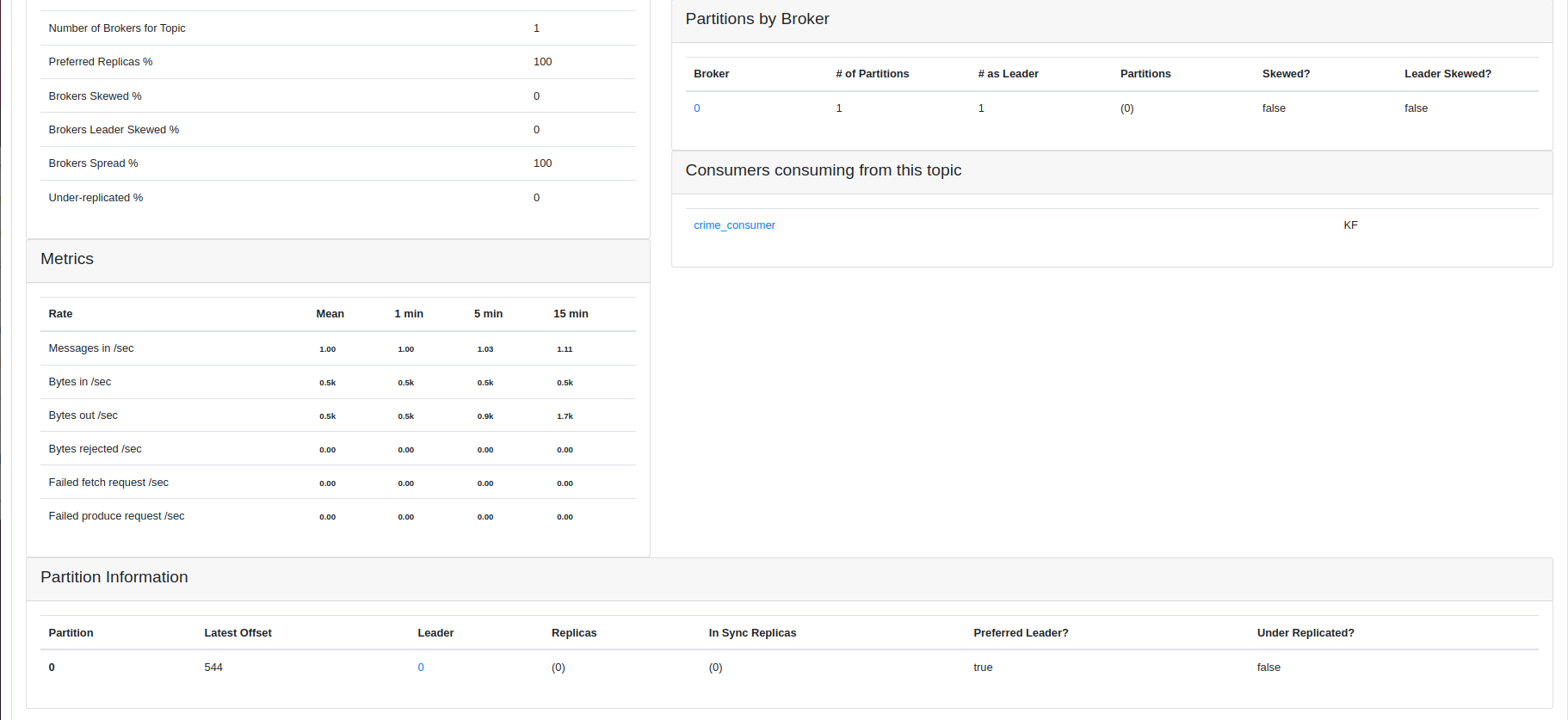
python3 consumer\_server.py

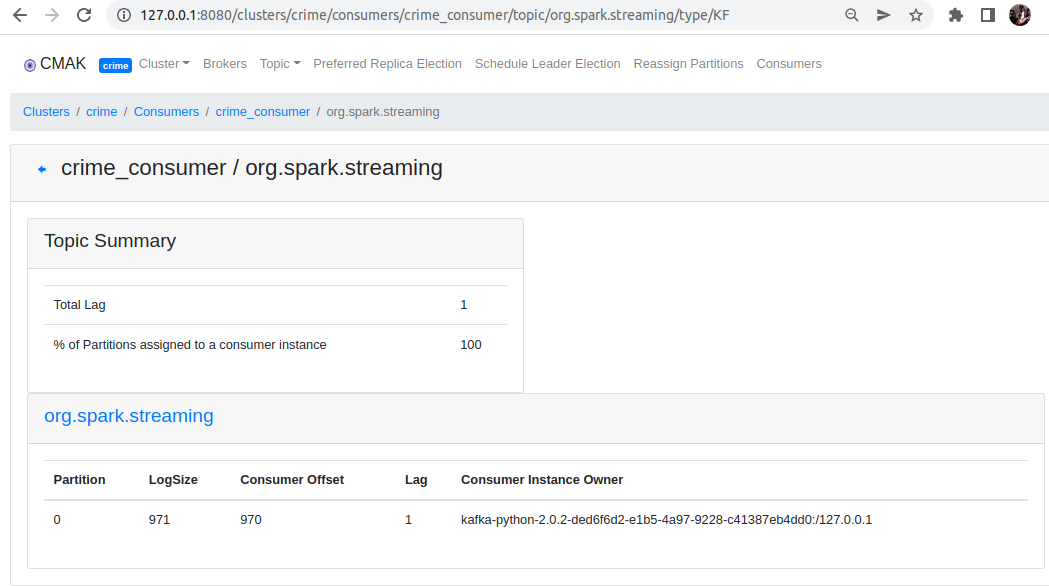




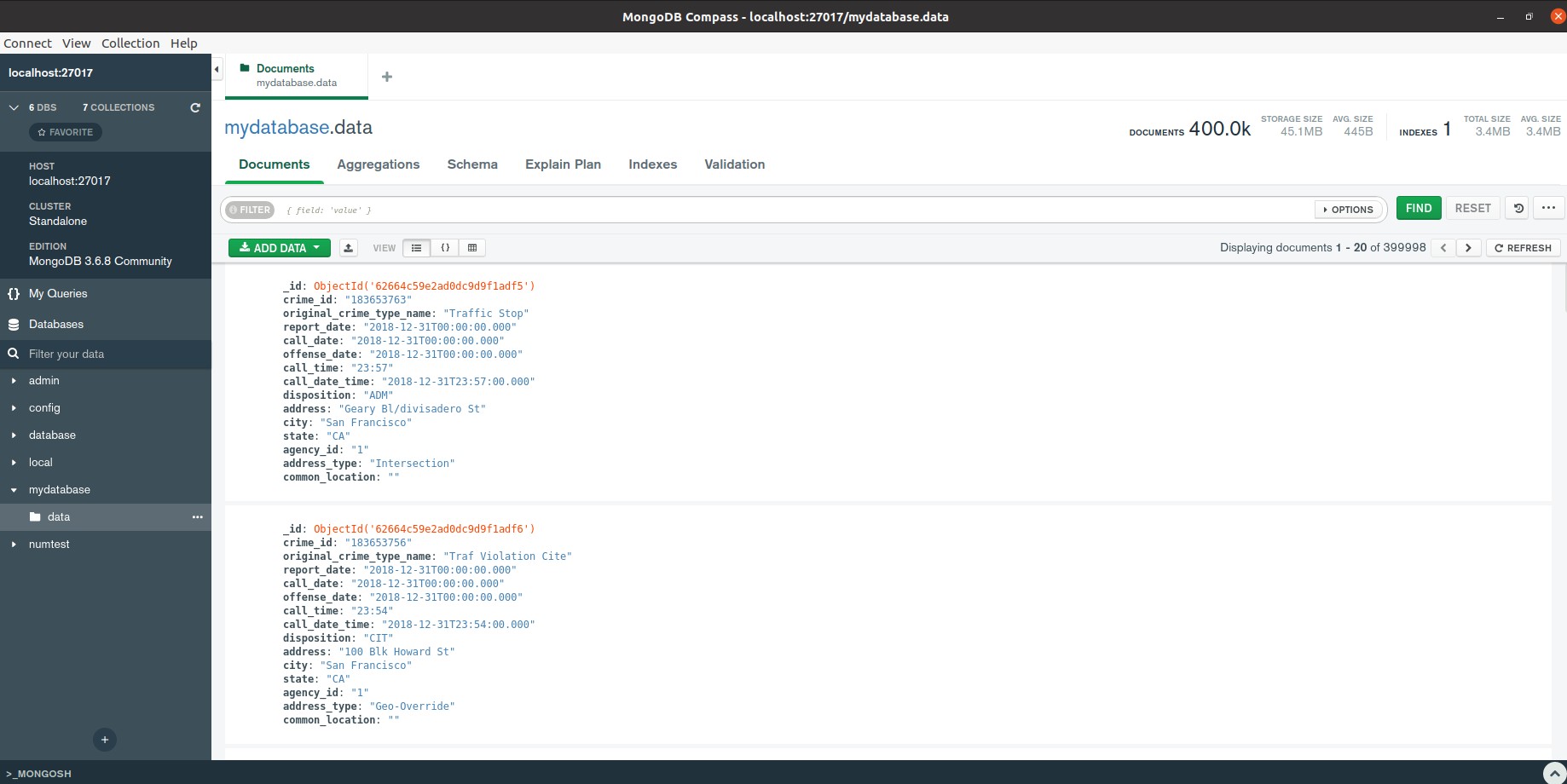
In CMAK topic org.spark.streaming is created and data is produced and consumed







Database-records added to mongodb



Conclusion

Data was streamed using spark streaming and spark queries were performed.

Data was also produced and consumed using kafka.

References <https://spark.apache.org/>

https://kafka.apache.org/documentation/#gettingStarted