HBase Assignment – Los Angeles Crime Data

Yacine Ammi, Touazi Aimen

NOSQL Module Project

Introduction

This project involved analyzing a sample of Los Angeles crime data using Apache HBase. The primary objectives were to practice loading a large dataset into HBase, designing an appropriate data model (rowkeys and column families), querying the data via the HBase Shell and Python (using happybase), and deploying the environment using Docker.

Technologies Used

- Docker / Docker Compose
- Apache HBase
- Python 3.x (with pandas, happybase, matplotlib, seaborn)
- GitHub

Environment Setup

The HBase environment was launched using Docker Compose with a docker-compose.yml file defining the HBase service. The HBase shell was accessed via docker exec -it hbase /bin/bash followed by hbase shell. Python tasks were performed in Jupyter Notebooks using a virtual environment.

Data Exploration (EDA)

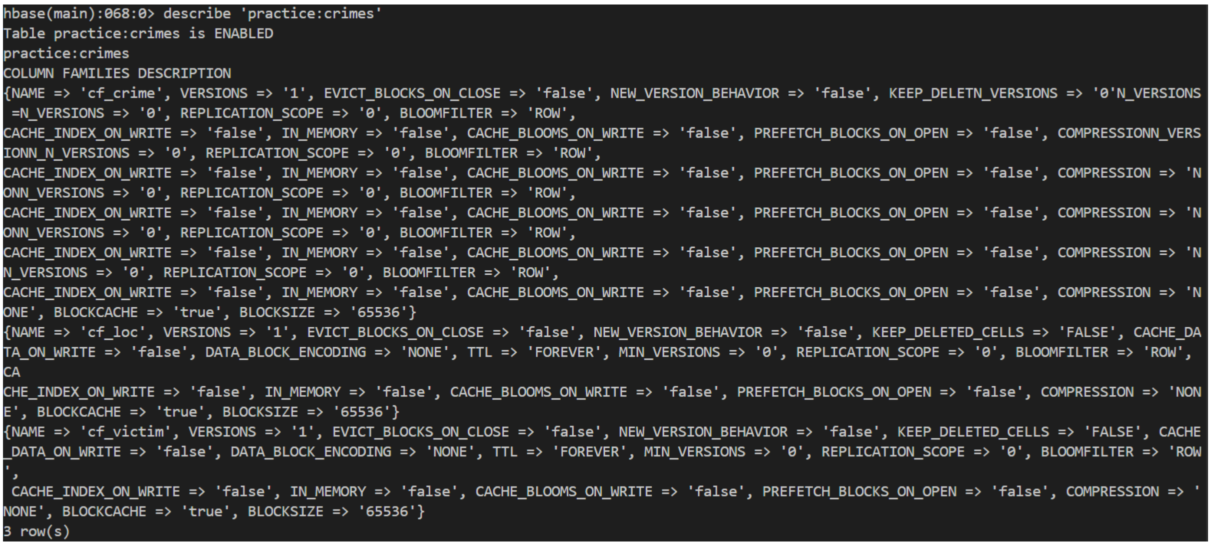
The provided sample_crimes_data.csv dataset, containing approximately 1 million rows and 28 columns, was loaded using Pandas.

- **Missing Values:** Significant missing data was noted in columns such as Weapon Used Cd (~67%), Weapon Desc (~67%), Crm Cd 2/3/4 (93-99%), and Cross Street (~85%). Columns like DR_NO, Date Rptd, DATE OCC, and AREA NAME had no missing values.
- **Key Categories:**
  - Top crime categories included "VEHICLE - STOLEN".
  - "Central" was the most frequent crime area.
- **Temporal Trends:**
  - [Briefly describe one key finding from your temporal charts, e.g., "The data spans from YYYY to YYYY, with the highest number of crimes reported in YYYY."]
  - [Optional: Embed one small, key chart here, e.g., "Crimes per Year"]
- **EDA Insights for Design:** The EDA highlighted the importance of temporal fields (DATE OCC) and area information (AREA NAME) for querying, guiding the rowkey design. The prevalence of missing values confirmed the suitability of HBase's sparse data handling.

HBase Data Model

- **Namespace:** A namespace practice was created using create_namespace 'practice'.
- **Table:** practice:crimes
- **Column Families:**
  - cf_loc: Stores location-specific data (e.g., area_name, lat, lon, premis_desc).
  - cf_crime: Contains details of the crime event (e.g., crm_cd_desc, weapon_desc, status_desc, date_rptd).
  - cf_victim: Holds victim-related information (e.g., vict_age, vict_sex, vict_descent).
  - *Design Rationale:* Column families were chosen to group data that is often accessed together, improving read efficiency and logical organization.
- **Rowkey Design:** SALT_YEAR_MONTH_AREACODE_DRNO
  - SALT: 2-digit DR_NO % 100 (zero-padded) for write distribution, preventing hotspotting.
  - YEAR: 4-digit year from DATE OCC for temporal scans.
  - MONTH: 2-digit month from DATE OCC (zero-padded) for finer-grained temporal scans.
  - AREACODE: 2-digit AREA number (zero-padded) for area-specific queries.
  - DRNO: Original DR_NO for uniqueness.
  - *Design Rationale:* This structure optimizes for common query patterns involving time ranges and specific areas, while salting ensures good data distribution across regions.

HBase Table Description:



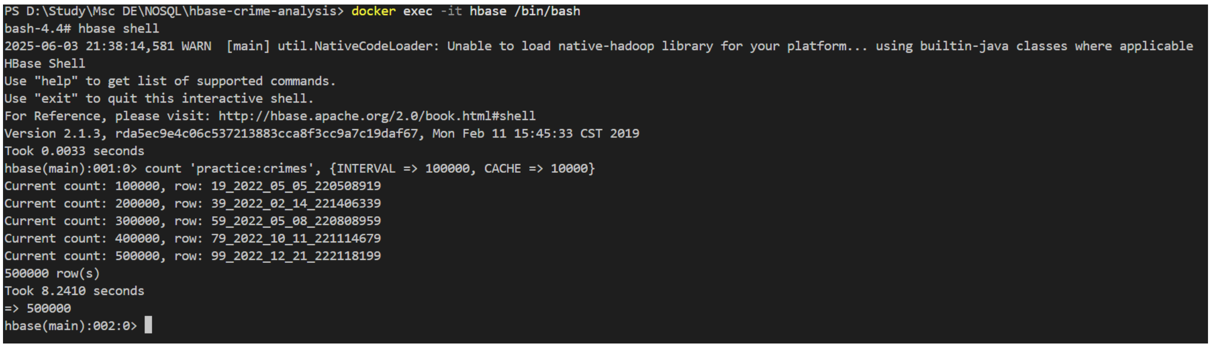
```
hbase(main):068:0> describe 'practice:crimes'
Table practice:crimes is ENABLED
practice:crimes
COLUMN FAMILIES DESCRIPTION
{NAME => 'cf_crime', VERSIONS => '1', EVICT_BLOCKS_ON_CLOSE => 'false', NEW_VERSION_BEHAVIOR => 'false', KEEP_DELETN_VERSIONS => '0'N_VERSIONS
 =N_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW',
CACHE_INDEX_ON_WRITE => 'false', IN_MEMORY => 'false', CACHE_BLOOMS_ON_WRITE => 'false', PREFETCH_BLOCKS_ON_OPEN => 'false', COMPRESSIONN_VERS
IONN_N_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW',
CACHE_INDEX_ON_WRITE => 'false', IN_MEMORY => 'false', CACHE_BLOOMS_ON_WRITE => 'false', PREFETCH_BLOCKS_ON_OPEN => 'false', COMPRESSION => 'N
ONN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW',
CACHE_INDEX_ON_WRITE => 'false', IN_MEMORY => 'false', CACHE_BLOOMS_ON_WRITE => 'false', PREFETCH_BLOCKS_ON_OPEN => 'false', COMPRESSION => 'N
ONN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW',
CACHE_INDEX_ON_WRITE => 'false', IN_MEMORY => 'false', CACHE_BLOOMS_ON_WRITE => 'false', PREFETCH_BLOCKS_ON_OPEN => 'false', COMPRESSION => 'N
N_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW',
CACHE_INDEX_ON_WRITE => 'false', IN_MEMORY => 'false', CACHE_BLOOMS_ON_WRITE => 'false', PREFETCH_BLOCKS_ON_OPEN => 'false', COMPRESSION => 'N
ONE', BLOCKCACHE => 'true', BLOCKSIZE => '65536'}
{NAME => 'cf_loc', VERSIONS => '1', EVICT_BLOCKS_ON_CLOSE => 'false', NEW_VERSION_BEHAVIOR => 'false', KEEP_DELETED_CELLS => 'FALSE', CACHE_DA
TA_ON_WRITE => 'false', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', MIN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW',
CA
CHE_INDEX_ON_WRITE => 'false', IN_MEMORY => 'false', CACHE_BLOOMS_ON_WRITE => 'false', PREFETCH_BLOCKS_ON_OPEN => 'false', COMPRESSION => 'NON
E', BLOCKCACHE => 'true', BLOCKSIZE => '65536'}
{NAME => 'cf_victim', VERSIONS => '1', EVICT_BLOCKS_ON_CLOSE => 'false', NEW_VERSION_BEHAVIOR => 'false', KEEP_DELETED_CELLS => 'FALSE', CACHE
_DATA_ON_WRITE => 'false', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', MIN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW
',
 CACHE_INDEX_ON_WRITE => 'false', IN_MEMORY => 'false', CACHE_BLOOMS_ON_WRITE => 'false', PREFETCH_BLOCKS_ON_OPEN => 'false', COMPRESSION => '
NONE', BLOCKCACHE => 'true', BLOCKSIZE => '65536'}
3 row(s)
```

Data Insertion

- **Data Cleaning:**
  - Column names were standardized (lowercase, underscores).
  - DATE OCC and TIME OCC were processed to extract components for the rowkey.
  - The rowkey was generated in Python:
  - Cells with NA values were not inserted to leverage HBase's sparse storage.
- **Loading:** The first 500,000 rows from the CSV were successfully pushed to practice:crimes using happybase with batching.

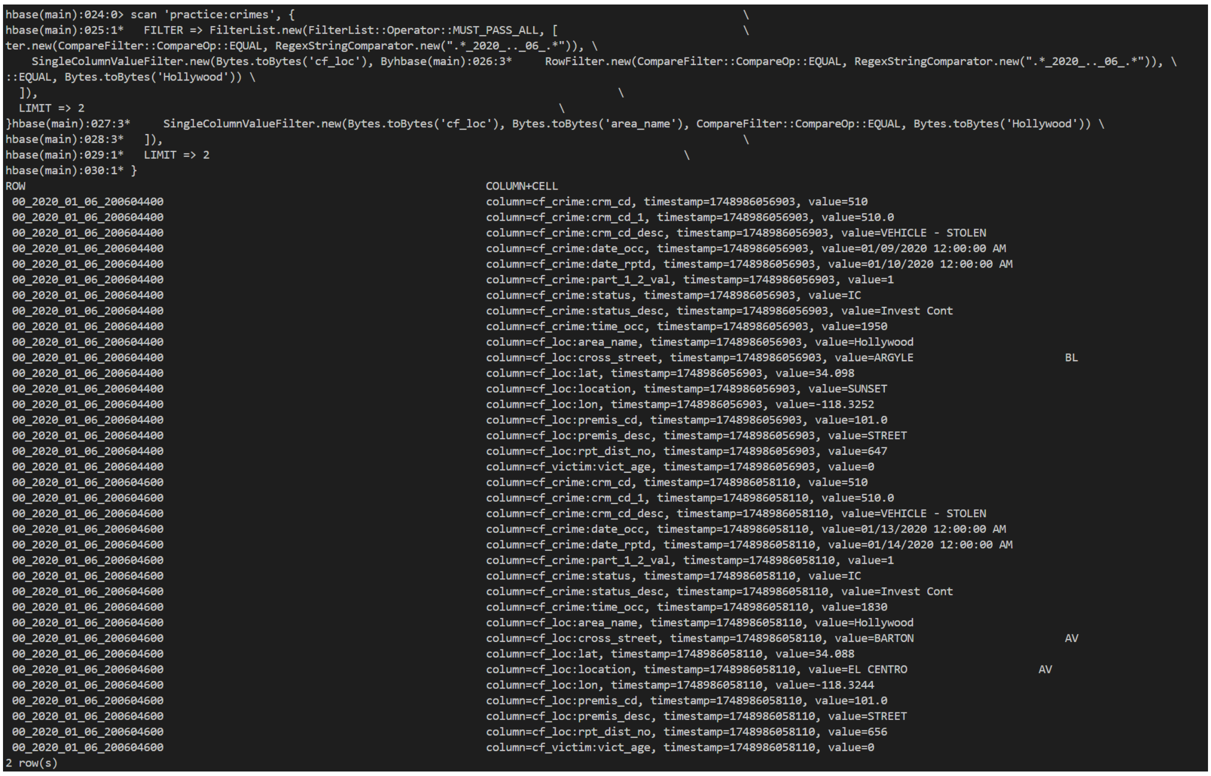
HBase Shell Queries & Results

1. Count the number of rows of our practice:crimes table



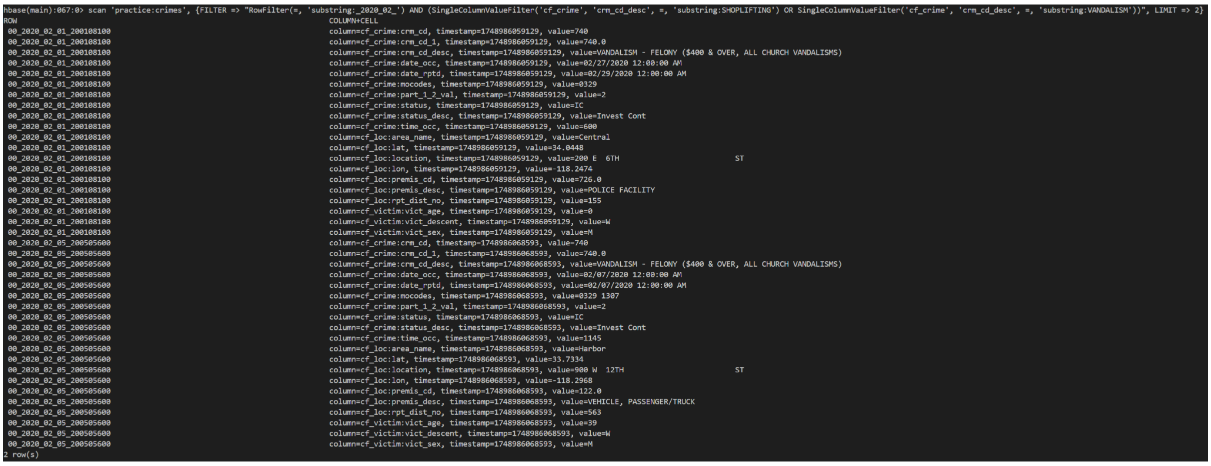
```
PS D:\Study\Msc DE\NOSQL\hbase-crime-analysis> docker exec -it hbase /bin/bash
bash-4.4# hbase shell
2025-06-03 21:38:14,581 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
For Reference, please visit: http://hbase.apache.org/2.0/book.html#shell
Version 2.1.3, rda5ec9e4c06c537213883cca8f3cc9a7c19daf67, Mon Feb 11 15:45:33 CST 2019
Took 0.0033 seconds
hbase(main):001:0> count 'practice:crimes', {INTERVAL => 100000, CACHE => 10000}
Current count: 100000, row: 19_2022_05_05_220508919
Current count: 200000, row: 39_2022_02_14_221406339
Current count: 300000, row: 59_2022_05_08_220808959
Current count: 400000, row: 79_2022_10_11_221114679
Current count: 500000, row: 99_2022_12_21_222118199
500000 row(s)
Took 8.2410 seconds
=> 500000
hbase(main):002:0>
```

2. All crimes in Hollywood in 2020 (first 2)



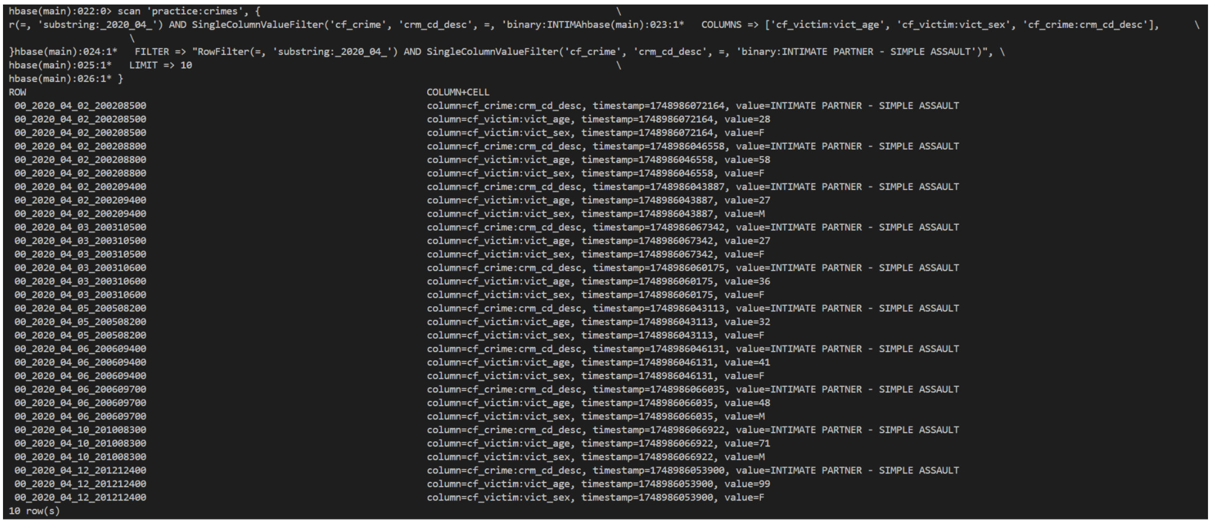
```
hbase(main):024:0> scan 'practice:crimes', {                                                          \
hbase(main):025:1*   FILTER => FilterList.new(FilterList::Operator::MUST_PASS_ALL, [                  \
ter.new(CompareFilter::CompareOp::EQUAL, RegexStringComparator.new(".*_2020_.._06_.*")), \
    SingleColumnValueFilter.new(Bytes.toBytes('cf_loc'), Byhbase(main):026:3*     RowFilter.new(CompareFilter::CompareOp::EQUAL, RegexStringComparator.new(".*_2020_.._06_.*")), \
::EQUAL, Bytes.toBytes('Hollywood')) \
  ]),                                                                                   \
  LIMIT => 2                                                                    \
}hbase(main):027:3*     SingleColumnValueFilter.new(Bytes.toBytes('cf_loc'), Bytes.toBytes('area_name'), CompareFilter::CompareOp::EQUAL, Bytes.toBytes('Hollywood')) \
hbase(main):028:3*   ]),                                                                      \
hbase(main):029:1*   LIMIT => 2                                                               \
hbase(main):030:1* }
ROW                                                COLUMN+CELL
 00_2020_01_06_200604400                           column=cf_crime:crm_cd, timestamp=1748986056903, value=510
 00_2020_01_06_200604400                           column=cf_crime:crm_cd_1, timestamp=1748986056903, value=510.0
 00_2020_01_06_200604400                           column=cf_crime:crm_cd_desc, timestamp=1748986056903, value=VEHICLE - STOLEN
 00_2020_01_06_200604400                           column=cf_crime:date_occ, timestamp=1748986056903, value=01/09/2020 12:00:00 AM
 00_2020_01_06_200604400                           column=cf_crime:date_rptd, timestamp=1748986056903, value=01/10/2020 12:00:00 AM
 00_2020_01_06_200604400                           column=cf_crime:part_1_2_val, timestamp=1748986056903, value=1
 00_2020_01_06_200604400                           column=cf_crime:status, timestamp=1748986056903, value=IC
 00_2020_01_06_200604400                           column=cf_crime:status_desc, timestamp=1748986056903, value=Invest Cont
 00_2020_01_06_200604400                           column=cf_crime:time_occ, timestamp=1748986056903, value=1950
 00_2020_01_06_200604400                           column=cf_loc:area_name, timestamp=1748986056903, value=Hollywood
 00_2020_01_06_200604400                           column=cf_loc:cross_street, timestamp=1748986056903, value=ARGYLE                      BL
 00_2020_01_06_200604400                           column=cf_loc:lat, timestamp=1748986056903, value=34.098
 00_2020_01_06_200604400                           column=cf_loc:location, timestamp=1748986056903, value=SUNSET
 00_2020_01_06_200604400                           column=cf_loc:lon, timestamp=1748986056903, value=-118.3252
 00_2020_01_06_200604400                           column=cf_loc:premis_cd, timestamp=1748986056903, value=101.0
 00_2020_01_06_200604400                           column=cf_loc:premis_desc, timestamp=1748986056903, value=STREET
 00_2020_01_06_200604400                           column=cf_loc:rpt_dist_no, timestamp=1748986056903, value=647
 00_2020_01_06_200604400                           column=cf_victim:vict_age, timestamp=1748986056903, value=0
 00_2020_01_06_200604600                           column=cf_crime:crm_cd, timestamp=1748986058110, value=510
 00_2020_01_06_200604600                           column=cf_crime:crm_cd_1, timestamp=1748986058110, value=510.0
 00_2020_01_06_200604600                           column=cf_crime:crm_cd_desc, timestamp=1748986058110, value=VEHICLE - STOLEN
 00_2020_01_06_200604600                           column=cf_crime:date_occ, timestamp=1748986058110, value=01/13/2020 12:00:00 AM
 00_2020_01_06_200604600                           column=cf_crime:date_rptd, timestamp=1748986058110, value=01/14/2020 12:00:00 AM
 00_2020_01_06_200604600                           column=cf_crime:part_1_2_val, timestamp=1748986058110, value=1
 00_2020_01_06_200604600                           column=cf_crime:status, timestamp=1748986058110, value=IC
 00_2020_01_06_200604600                           column=cf_crime:status_desc, timestamp=1748986058110, value=Invest Cont
 00_2020_01_06_200604600                           column=cf_crime:time_occ, timestamp=1748986058110, value=1830
 00_2020_01_06_200604600                           column=cf_loc:area_name, timestamp=1748986058110, value=Hollywood
 00_2020_01_06_200604600                           column=cf_loc:cross_street, timestamp=1748986058110, value=BARTON                      AV
 00_2020_01_06_200604600                           column=cf_loc:lat, timestamp=1748986058110, value=34.088
 00_2020_01_06_200604600                           column=cf_loc:location, timestamp=1748986058110, value=EL CENTRO                   AV
 00_2020_01_06_200604600                           column=cf_loc:lon, timestamp=1748986058110, value=-118.3244
 00_2020_01_06_200604600                           column=cf_loc:premis_cd, timestamp=1748986058110, value=101.0
 00_2020_01_06_200604600                           column=cf_loc:premis_desc, timestamp=1748986058110, value=STREET
 00_2020_01_06_200604600                           column=cf_loc:rpt_dist_no, timestamp=1748986058110, value=656
 00_2020_01_06_200604600                           column=cf_victim:vict_age, timestamp=1748986058110, value=0
2 row(s)
```

3. Count & show the first 2 rows of: All SHOPLIFTING and VANDALISM crimes (if the label of the crime contains it) in February 2020



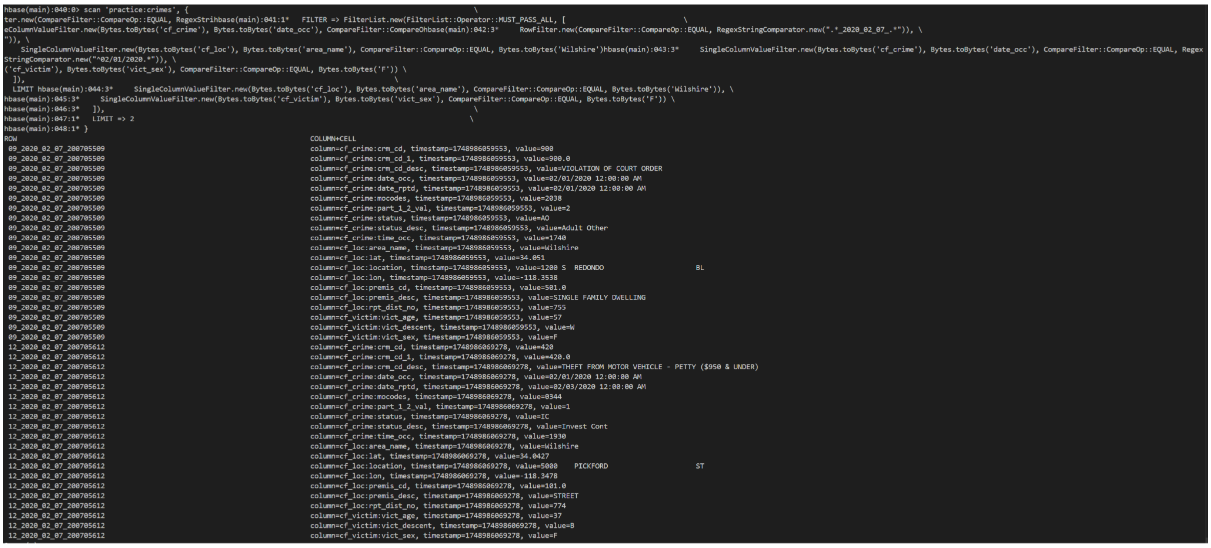
```
hbase(main):067:0> scan 'practice:crimes', {FILTER => "RowFilter(=, 'substring:_2020_02_') AND (SingleColumnValueFilter('cf_crime', 'crm_cd_desc', =, 'substring:SHOPLIFTING') OR SingleColumnValueFilter('cf_crime', 'crm_cd_desc', =, 'substring:VANDALISM'))", LIMIT => 2}
ROW                                 COLUMN+CELL
 00_2020_02_01_200108100            column=cf_crime:crm_cd, timestamp=1748986059129, value=740
 00_2020_02_01_200108100            column=cf_crime:crm_cd_1, timestamp=1748986059129, value=740.0
 00_2020_02_01_200108100            column=cf_crime:crm_cd_desc, timestamp=1748986059129, value=VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)
 00_2020_02_01_200108100            column=cf_crime:date_occ, timestamp=1748986059129, value=02/27/2020 12:00:00 AM
 00_2020_02_01_200108100            column=cf_crime:date_rptd, timestamp=1748986059129, value=02/29/2020 12:00:00 AM
 00_2020_02_01_200108100            column=cf_crime:mocodes, timestamp=1748986059129, value=0329
 00_2020_02_01_200108100            column=cf_crime:part_1_2_val, timestamp=1748986059129, value=2
 00_2020_02_01_200108100            column=cf_crime:status, timestamp=1748986059129, value=IC
 00_2020_02_01_200108100            column=cf_crime:status_desc, timestamp=1748986059129, value=Invest Cont
 00_2020_02_01_200108100            column=cf_crime:time_occ, timestamp=1748986059129, value=600
 00_2020_02_01_200108100            column=cf_loc:area_name, timestamp=1748986059129, value=Central
 00_2020_02_01_200108100            column=cf_loc:lat, timestamp=1748986059129, value=34.0448
 00_2020_02_01_200108100            column=cf_loc:location, timestamp=1748986059129, value=200 E  6TH                      ST
 00_2020_02_01_200108100            column=cf_loc:lon, timestamp=1748986059129, value=-118.2474
 00_2020_02_01_200108100            column=cf_loc:premis_cd, timestamp=1748986059129, value=726.0
 00_2020_02_01_200108100            column=cf_loc:premis_desc, timestamp=1748986059129, value=POLICE FACILITY
 00_2020_02_01_200108100            column=cf_loc:rpt_dist_no, timestamp=1748986059129, value=155
 00_2020_02_01_200108100            column=cf_victim:vict_age, timestamp=1748986059129, value=0
 00_2020_02_01_200108100            column=cf_victim:vict_descent, timestamp=1748986059129, value=W
 00_2020_02_01_200108100            column=cf_victim:vict_sex, timestamp=1748986059129, value=M
 00_2020_02_05_200505600            column=cf_crime:crm_cd, timestamp=1748986068593, value=740
 00_2020_02_05_200505600            column=cf_crime:crm_cd_1, timestamp=1748986068593, value=740.0
 00_2020_02_05_200505600            column=cf_crime:crm_cd_desc, timestamp=1748986068593, value=VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)
 00_2020_02_05_200505600            column=cf_crime:date_occ, timestamp=1748986068593, value=02/07/2020 12:00:00 AM
 00_2020_02_05_200505600            column=cf_crime:date_rptd, timestamp=1748986068593, value=02/07/2020 12:00:00 AM
 00_2020_02_05_200505600            column=cf_crime:mocodes, timestamp=1748986068593, value=0329 1307
 00_2020_02_05_200505600            column=cf_crime:part_1_2_val, timestamp=1748986068593, value=2
 00_2020_02_05_200505600            column=cf_crime:status, timestamp=1748986068593, value=IC
 00_2020_02_05_200505600            column=cf_crime:status_desc, timestamp=1748986068593, value=Invest Cont
 00_2020_02_05_200505600            column=cf_crime:time_occ, timestamp=1748986068593, value=1145
 00_2020_02_05_200505600            column=cf_loc:area_name, timestamp=1748986068593, value=Harbor
 00_2020_02_05_200505600            column=cf_loc:lat, timestamp=1748986068593, value=33.7334
 00_2020_02_05_200505600            column=cf_loc:location, timestamp=1748986068593, value=900 W  12TH                     ST
 00_2020_02_05_200505600            column=cf_loc:lon, timestamp=1748986068593, value=-118.2968
 00_2020_02_05_200505600            column=cf_loc:premis_cd, timestamp=1748986068593, value=122.0
 00_2020_02_05_200505600            column=cf_loc:premis_desc, timestamp=1748986068593, value=VEHICLE, PASSENGER/TRUCK
 00_2020_02_05_200505600            column=cf_loc:rpt_dist_no, timestamp=1748986068593, value=563
 00_2020_02_05_200505600            column=cf_victim:vict_age, timestamp=1748986068593, value=39
 00_2020_02_05_200505600            column=cf_victim:vict_descent, timestamp=1748986068593, value=W
 00_2020_02_05_200505600            column=cf_victim:vict_sex, timestamp=1748986068593, value=M
2 row(s)
```

4. Victim age and sex for crimes of INTIMATE PARTNER - SIMPLE ASSAULT (exact match) in April 2020



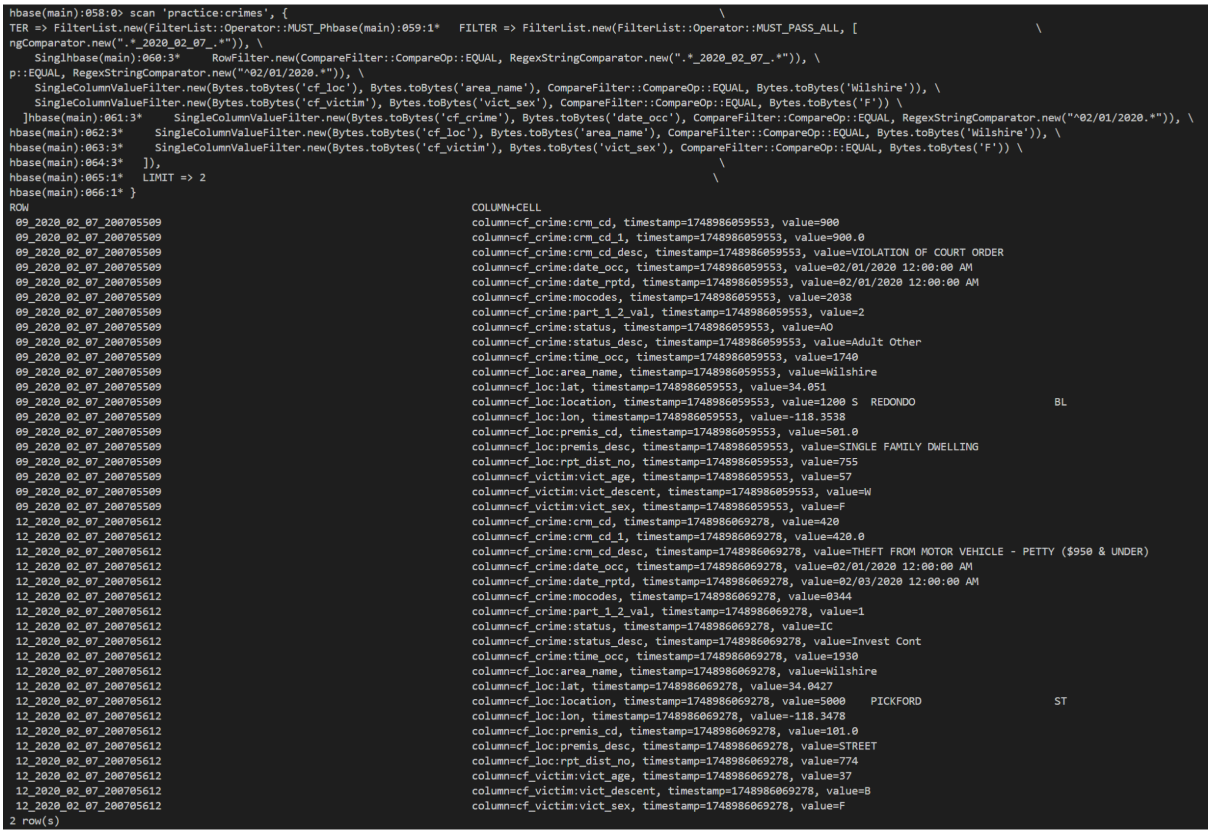
```
hbase(main):022:0> scan 'practice:crimes', {                                                          \
r(=, 'substring:_2020_04_') AND SingleColumnValueFilter('cf_crime', 'crm_cd_desc', =, 'binary:INTIMAhbase(main):023:1*   COLUMNS => ['cf_victim:vict_age', 'cf_victim:vict_sex', 'cf_crime:crm_cd_desc'],      \
                         \
}hbase(main):024:1*   FILTER => "RowFilter(=, 'substring:_2020_04_') AND SingleColumnValueFilter('cf_crime', 'crm_cd_desc', =, 'binary:INTIMATE PARTNER - SIMPLE ASSAULT')", \
hbase(main):025:1*   LIMIT => 10                                                                      \
hbase(main):026:1* }
ROW                                      COLUMN+CELL
 00_2020_04_02_200208500                 column=cf_crime:crm_cd_desc, timestamp=1748986072164, value=INTIMATE PARTNER - SIMPLE ASSAULT
 00_2020_04_02_200208500                 column=cf_victim:vict_age, timestamp=1748986072164, value=28
 00_2020_04_02_200208500                 column=cf_victim:vict_sex, timestamp=1748986072164, value=F
 00_2020_04_02_200208800                 column=cf_crime:crm_cd_desc, timestamp=1748986046558, value=INTIMATE PARTNER - SIMPLE ASSAULT
 00_2020_04_02_200208800                 column=cf_victim:vict_age, timestamp=1748986046558, value=58
 00_2020_04_02_200208800                 column=cf_victim:vict_sex, timestamp=1748986046558, value=F
 00_2020_04_02_200209400                 column=cf_crime:crm_cd_desc, timestamp=1748986043887, value=INTIMATE PARTNER - SIMPLE ASSAULT
 00_2020_04_02_200209400                 column=cf_victim:vict_age, timestamp=1748986043887, value=27
 00_2020_04_02_200209400                 column=cf_victim:vict_sex, timestamp=1748986043887, value=M
 00_2020_04_03_200310500                 column=cf_crime:crm_cd_desc, timestamp=1748986067342, value=INTIMATE PARTNER - SIMPLE ASSAULT
 00_2020_04_03_200310500                 column=cf_victim:vict_age, timestamp=1748986067342, value=27
 00_2020_04_03_200310500                 column=cf_victim:vict_sex, timestamp=1748986067342, value=F
 00_2020_04_03_200310600                 column=cf_crime:crm_cd_desc, timestamp=1748986060175, value=INTIMATE PARTNER - SIMPLE ASSAULT
 00_2020_04_03_200310600                 column=cf_victim:vict_age, timestamp=1748986060175, value=36
 00_2020_04_03_200310600                 column=cf_victim:vict_sex, timestamp=1748986060175, value=F
 00_2020_04_05_200508200                 column=cf_crime:crm_cd_desc, timestamp=1748986043113, value=INTIMATE PARTNER - SIMPLE ASSAULT
 00_2020_04_05_200508200                 column=cf_victim:vict_age, timestamp=1748986043113, value=32
 00_2020_04_05_200508200                 column=cf_victim:vict_sex, timestamp=1748986043113, value=F
 00_2020_04_06_200609400                 column=cf_crime:crm_cd_desc, timestamp=1748986046131, value=INTIMATE PARTNER - SIMPLE ASSAULT
 00_2020_04_06_200609400                 column=cf_victim:vict_age, timestamp=1748986046131, value=41
 00_2020_04_06_200609400                 column=cf_victim:vict_sex, timestamp=1748986046131, value=F
 00_2020_04_06_200609700                 column=cf_crime:crm_cd_desc, timestamp=1748986066035, value=INTIMATE PARTNER - SIMPLE ASSAULT
 00_2020_04_06_200609700                 column=cf_victim:vict_age, timestamp=1748986066035, value=48
 00_2020_04_06_200609700                 column=cf_victim:vict_sex, timestamp=1748986066035, value=M
 00_2020_04_10_201008300                 column=cf_crime:crm_cd_desc, timestamp=1748986066922, value=INTIMATE PARTNER - SIMPLE ASSAULT
 00_2020_04_10_201008300                 column=cf_victim:vict_age, timestamp=1748986066922, value=71
 00_2020_04_10_201008300                 column=cf_victim:vict_sex, timestamp=1748986066922, value=M
 00_2020_04_12_201212400                 column=cf_crime:crm_cd_desc, timestamp=1748986053900, value=INTIMATE PARTNER - SIMPLE ASSAULT
 00_2020_04_12_201212400                 column=cf_victim:vict_age, timestamp=1748986053900, value=99
 00_2020_04_12_201212400                 column=cf_victim:vict_sex, timestamp=1748986053900, value=F
10 row(s)
```

5. Crimes reported (this likely means DATE OCC for consistency with HBase rowkey, or Date Rptd if you want to query that specific column) in 03/12/2020 12:00:00 AM



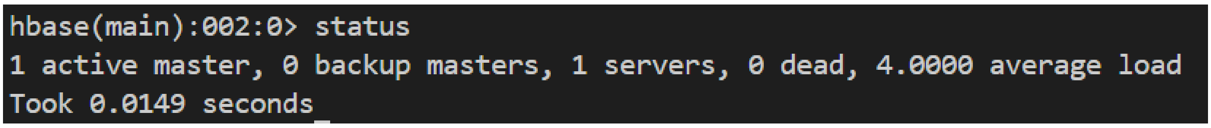
```
hbase(main):040:0> scan 'practice:crimes', {                                                                    \
ter.new(CompareFilter::CompareOp::EQUAL, RegexStrihbase(main):041:1*   FILTER => FilterList.new(FilterList::Operator::MUST_PASS_ALL, [                        \
eColumnValueFilter.new(Bytes.toBytes('cf_crime'), Bytes.toBytes('date_occ'), CompareFilter::CompareOhbase(main):042:3*     RowFilter.new(CompareFilter::CompareOp::EQUAL, RegexStringComparator.new(".*_2020_02_07_.*")), \
")), \
    SingleColumnValueFilter.new(Bytes.toBytes('cf_loc'), Bytes.toBytes('area_name'), CompareFilter::CompareOp::EQUAL, Bytes.toBytes('Wilshire')hbase(main):043:3*     SingleColumnValueFilter.new(Bytes.toBytes('cf_crime'), Bytes.toBytes('date_occ'), CompareFilter::CompareOp::EQUAL, Regex
StringComparator.new("^02/01/2020.*")), \
('cf_victim'), Bytes.toBytes('vict_sex'), CompareFilter::CompareOp::EQUAL, Bytes.toBytes('F')) \
  ]),                                                                                        \
  LIMIT hbase(main):044:3*     SingleColumnValueFilter.new(Bytes.toBytes('cf_loc'), Bytes.toBytes('area_name'), CompareFilter::CompareOp::EQUAL, Bytes.toBytes('Wilshire')), \
hbase(main):045:3*     SingleColumnValueFilter.new(Bytes.toBytes('cf_victim'), Bytes.toBytes('vict_sex'), CompareFilter::CompareOp::EQUAL, Bytes.toBytes('F')) \
hbase(main):046:3*   ]),                                                                        \
hbase(main):047:1*   LIMIT => 2                                                                \
hbase(main):048:1* }
ROW                                                         COLUMN+CELL
 09_2020_02_07_200705509                                    column=cf_crime:crm_cd, timestamp=1748986059553, value=900
 09_2020_02_07_200705509                                    column=cf_crime:crm_cd_1, timestamp=1748986059553, value=900.0
 09_2020_02_07_200705509                                    column=cf_crime:crm_cd_desc, timestamp=1748986059553, value=VIOLATION OF COURT ORDER
 09_2020_02_07_200705509                                    column=cf_crime:date_occ, timestamp=1748986059553, value=02/01/2020 12:00:00 AM
 09_2020_02_07_200705509                                    column=cf_crime:date_rptd, timestamp=1748986059553, value=02/01/2020 12:00:00 AM
 09_2020_02_07_200705509                                    column=cf_crime:mocodes, timestamp=1748986059553, value=2038
 09_2020_02_07_200705509                                    column=cf_crime:part_1_2_val, timestamp=1748986059553, value=2
 09_2020_02_07_200705509                                    column=cf_crime:status, timestamp=1748986059553, value=AO
 09_2020_02_07_200705509                                    column=cf_crime:status_desc, timestamp=1748986059553, value=Adult Other
 09_2020_02_07_200705509                                    column=cf_crime:time_occ, timestamp=1748986059553, value=1740
 09_2020_02_07_200705509                                    column=cf_loc:area_name, timestamp=1748986059553, value=Wilshire
 09_2020_02_07_200705509                                    column=cf_loc:lat, timestamp=1748986059553, value=34.051
 09_2020_02_07_200705509                                    column=cf_loc:location, timestamp=1748986059553, value=1200 S  REDONDO                    BL
 09_2020_02_07_200705509                                    column=cf_loc:lon, timestamp=1748986059553, value=-118.3538
 09_2020_02_07_200705509                                    column=cf_loc:premis_cd, timestamp=1748986059553, value=501.0
 09_2020_02_07_200705509                                    column=cf_loc:premis_desc, timestamp=1748986059553, value=SINGLE FAMILY DWELLING
 09_2020_02_07_200705509                                    column=cf_loc:rpt_dist_no, timestamp=1748986059553, value=755
 09_2020_02_07_200705509                                    column=cf_victim:vict_age, timestamp=1748986059553, value=57
 09_2020_02_07_200705509                                    column=cf_victim:vict_descent, timestamp=1748986059553, value=W
 09_2020_02_07_200705509                                    column=cf_victim:vict_sex, timestamp=1748986059553, value=F
 12_2020_02_07_200705612                                    column=cf_crime:crm_cd, timestamp=1748986069278, value=420
 12_2020_02_07_200705612                                    column=cf_crime:crm_cd_1, timestamp=1748986069278, value=420.0
 12_2020_02_07_200705612                                    column=cf_crime:crm_cd_desc, timestamp=1748986069278, value=THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)
 12_2020_02_07_200705612                                    column=cf_crime:date_occ, timestamp=1748986069278, value=02/01/2020 12:00:00 AM
 12_2020_02_07_200705612                                    column=cf_crime:date_rptd, timestamp=1748986069278, value=02/03/2020 12:00:00 AM
 12_2020_02_07_200705612                                    column=cf_crime:mocodes, timestamp=1748986069278, value=0344
 12_2020_02_07_200705612                                    column=cf_crime:part_1_2_val, timestamp=1748986069278, value=1
 12_2020_02_07_200705612                                    column=cf_crime:status, timestamp=1748986069278, value=IC
 12_2020_02_07_200705612                                    column=cf_crime:status_desc, timestamp=1748986069278, value=Invest Cont
 12_2020_02_07_200705612                                    column=cf_crime:time_occ, timestamp=1748986069278, value=1930
 12_2020_02_07_200705612                                    column=cf_loc:area_name, timestamp=1748986069278, value=Wilshire
 12_2020_02_07_200705612                                    column=cf_loc:lat, timestamp=1748986069278, value=34.0427
 12_2020_02_07_200705612                                    column=cf_loc:location, timestamp=1748986069278, value=5000    PICKFORD                     ST
 12_2020_02_07_200705612                                    column=cf_loc:lon, timestamp=1748986069278, value=-118.3478
 12_2020_02_07_200705612                                    column=cf_loc:premis_cd, timestamp=1748986069278, value=101.0
 12_2020_02_07_200705612                                    column=cf_loc:premis_desc, timestamp=1748986069278, value=STREET
 12_2020_02_07_200705612                                    column=cf_loc:rpt_dist_no, timestamp=1748986069278, value=774
 12_2020_02_07_200705612                                    column=cf_victim:vict_age, timestamp=1748986069278, value=37
 12_2020_02_07_200705612                                    column=cf_victim:vict_descent, timestamp=1748986069278, value=B
 12_2020_02_07_200705612                                    column=cf_victim:vict_sex, timestamp=1748986069278, value=F
```

6. Crimes occurring between 02/01/2020 12:00:00 AM and 02/02/2020 12:00:00 AM (i.e., all of 02/01/2020), in Wilshire on female victims (first 2).



```
hbase(main):058:0> scan 'practice:crimes', {                                                                    \
TER => FilterList.new(FilterList::Operator::MUST_Phbase(main):059:1*   FILTER => FilterList.new(FilterList::Operator::MUST_PASS_ALL, [                        \
ngComparator.new(".*_2020_02_07_.*")), \
    Singlhbase(main):060:3*     RowFilter.new(CompareFilter::CompareOp::EQUAL, RegexStringComparator.new(".*_2020_02_07_.*")), \
p::EQUAL, RegexStringComparator.new("^02/01/2020.*")), \
    SingleColumnValueFilter.new(Bytes.toBytes('cf_loc'), Bytes.toBytes('area_name'), CompareFilter::CompareOp::EQUAL, Bytes.toBytes('Wilshire')), \
    SingleColumnValueFilter.new(Bytes.toBytes('cf_victim'), Bytes.toBytes('vict_sex'), CompareFilter::CompareOp::EQUAL, Bytes.toBytes('F')) \
  ]hbase(main):061:3*     SingleColumnValueFilter.new(Bytes.toBytes('cf_crime'), Bytes.toBytes('date_occ'), CompareFilter::CompareOp::EQUAL, RegexStringComparator.new("^02/01/2020.*")), \
hbase(main):062:3*     SingleColumnValueFilter.new(Bytes.toBytes('cf_loc'), Bytes.toBytes('area_name'), CompareFilter::CompareOp::EQUAL, Bytes.toBytes('Wilshire')), \
hbase(main):063:3*     SingleColumnValueFilter.new(Bytes.toBytes('cf_victim'), Bytes.toBytes('vict_sex'), CompareFilter::CompareOp::EQUAL, Bytes.toBytes('F')) \
hbase(main):064:3*   ]),                                                                        \
hbase(main):065:1*   LIMIT => 2                                                                \
hbase(main):066:1* }
ROW                                                         COLUMN+CELL
 09_2020_02_07_200705509                                    column=cf_crime:crm_cd, timestamp=1748986059553, value=900
 09_2020_02_07_200705509                                    column=cf_crime:crm_cd_1, timestamp=1748986059553, value=900.0
 09_2020_02_07_200705509                                    column=cf_crime:crm_cd_desc, timestamp=1748986059553, value=VIOLATION OF COURT ORDER
 09_2020_02_07_200705509                                    column=cf_crime:date_occ, timestamp=1748986059553, value=02/01/2020 12:00:00 AM
 09_2020_02_07_200705509                                    column=cf_crime:date_rptd, timestamp=1748986059553, value=02/01/2020 12:00:00 AM
 09_2020_02_07_200705509                                    column=cf_crime:mocodes, timestamp=1748986059553, value=2038
 09_2020_02_07_200705509                                    column=cf_crime:part_1_2_val, timestamp=1748986059553, value=2
 09_2020_02_07_200705509                                    column=cf_crime:status, timestamp=1748986059553, value=AO
 09_2020_02_07_200705509                                    column=cf_crime:status_desc, timestamp=1748986059553, value=Adult Other
 09_2020_02_07_200705509                                    column=cf_crime:time_occ, timestamp=1748986059553, value=1740
 09_2020_02_07_200705509                                    column=cf_loc:area_name, timestamp=1748986059553, value=Wilshire
 09_2020_02_07_200705509                                    column=cf_loc:lat, timestamp=1748986059553, value=34.051
 09_2020_02_07_200705509                                    column=cf_loc:location, timestamp=1748986059553, value=1200 S  REDONDO                    BL
 09_2020_02_07_200705509                                    column=cf_loc:lon, timestamp=1748986059553, value=-118.3538
 09_2020_02_07_200705509                                    column=cf_loc:premis_cd, timestamp=1748986059553, value=501.0
 09_2020_02_07_200705509                                    column=cf_loc:premis_desc, timestamp=1748986059553, value=SINGLE FAMILY DWELLING
 09_2020_02_07_200705509                                    column=cf_loc:rpt_dist_no, timestamp=1748986059553, value=755
 09_2020_02_07_200705509                                    column=cf_victim:vict_age, timestamp=1748986059553, value=57
 09_2020_02_07_200705509                                    column=cf_victim:vict_descent, timestamp=1748986059553, value=W
 09_2020_02_07_200705509                                    column=cf_victim:vict_sex, timestamp=1748986059553, value=F
 12_2020_02_07_200705612                                    column=cf_crime:crm_cd, timestamp=1748986069278, value=420
 12_2020_02_07_200705612                                    column=cf_crime:crm_cd_1, timestamp=1748986069278, value=420.0
 12_2020_02_07_200705612                                    column=cf_crime:crm_cd_desc, timestamp=1748986069278, value=THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)
 12_2020_02_07_200705612                                    column=cf_crime:date_occ, timestamp=1748986069278, value=02/01/2020 12:00:00 AM
 12_2020_02_07_200705612                                    column=cf_crime:date_rptd, timestamp=1748986069278, value=02/03/2020 12:00:00 AM
 12_2020_02_07_200705612                                    column=cf_crime:mocodes, timestamp=1748986069278, value=0344
 12_2020_02_07_200705612                                    column=cf_crime:part_1_2_val, timestamp=1748986069278, value=1
 12_2020_02_07_200705612                                    column=cf_crime:status, timestamp=1748986069278, value=IC
 12_2020_02_07_200705612                                    column=cf_crime:status_desc, timestamp=1748986069278, value=Invest Cont
 12_2020_02_07_200705612                                    column=cf_crime:time_occ, timestamp=1748986069278, value=1930
 12_2020_02_07_200705612                                    column=cf_loc:area_name, timestamp=1748986069278, value=Wilshire
 12_2020_02_07_200705612                                    column=cf_loc:lat, timestamp=1748986069278, value=34.0427
 12_2020_02_07_200705612                                    column=cf_loc:location, timestamp=1748986069278, value=5000    PICKFORD                     ST
 12_2020_02_07_200705612                                    column=cf_loc:lon, timestamp=1748986069278, value=-118.3478
 12_2020_02_07_200705612                                    column=cf_loc:premis_cd, timestamp=1748986069278, value=101.0
 12_2020_02_07_200705612                                    column=cf_loc:premis_desc, timestamp=1748986069278, value=STREET
 12_2020_02_07_200705612                                    column=cf_loc:rpt_dist_no, timestamp=1748986069278, value=774
 12_2020_02_07_200705612                                    column=cf_victim:vict_age, timestamp=1748986069278, value=37
 12_2020_02_07_200705612                                    column=cf_victim:vict_descent, timestamp=1748986069278, value=B
 12_2020_02_07_200705612                                    column=cf_victim:vict_sex, timestamp=1748986069278, value=F
2 row(s)
```

7. Check number of regions and region servers



```
hbase(main):002:0> status
1 active master, 0 backup masters, 1 servers, 0 dead, 4.0000 average load
Took 0.0149 seconds
```

Python Data Retrieval

Data was retrieved from HBase using happybase in Python.

Data retrieved from HBase using Python was consistent with expectations from the HBase shell queries. A full quantitative comparison with the original CSV (e.g., matching exact row counts after complex filtering) would require aligning query limits and potentially more complex Pandas filtering logic to mirror HBase's scan behavior. Conceptually, HBase is suited for large-scale data and targeted key-based retrievals, while Pandas excels at in-memory analytics.

Conclusion

This project provided practical experience in using Apache HBase for managing and querying a large dataset. Key takeaways include the importance of careful data modeling, especially rowkey design for efficient querying, and understanding HBase's sparse data capabilities. The process involved setting up a Dockerized HBase environment, performing EDA to inform design, inserting data using Python with happybase, and querying data through both the HBase shell and Python.

A notable challenge was ensuring sufficient resources for the HBase Docker container within the WSL2 environment, which was resolved by adjusting WSL2 configuration. Mastering the HBase shell filter syntax also required careful attention.

Overall, the assignment successfully demonstrated the workflow of an HBase project from setup to data analysis.