Group PG 6 Phase 2

COMP 4380

**Summary of Data:**

* **What DBMS are you using?**

SQLite3 with SQLite Studio

* **What are the specs of the computer/cloud service on which you are running your** queries?

8 cores at 3.6ghz, RTX 3070, ~16 gb ram, 512gb SSD

* **How many tables do you have? How many rows/columns in each?**

1 Table, 41 columns, ~32 million rows

* **What is the file size of each table?**

~17 gb

**Query 1**

Find the top 10 complaints and find the percentage of how many of those complaints have been closed when city begins with B from the the first day of 2020

SELECT complaint\_type,

COUNT(\*) AS total\_complaints,

ROUND((COUNT(CASE WHEN closed\_date IS NOT NULL AND closed\_date <> '' THEN 1 END) \* 100.0 / COUNT(\*)), 2) AS closed\_percentage

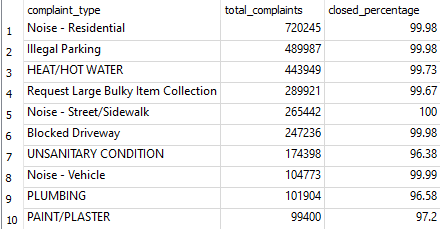
FROM requests

WHERE city LIKE 'B%' AND created\_date NOT LIKE '%/%/201%'

GROUP BY complaint\_type

ORDER BY total\_complaints DESC

LIMIT 10;



61.7s

Add an index on the closed date column to ensure it is not null or empty if not already taken care of by SQL and also add indexes for city starts with B and year is 2020 onwards.

**Query 2**

Find the number of requests from each year by month from NYPD that are greater than 2000

SELECT strftime('%Y', date) AS year, max(request\_count) AS max\_requests, month

FROM (

SELECT (substr(created\_date, 7, 4) || '-' || substr(created\_date, 4,2) || '-' || substr(created\_date, 1,2)) AS date, strftime('%d', (substr(created\_date, 7, 4) || '-' || substr(created\_date, 4,2) || '-' || substr(created\_date, 1,2))) AS month, count(\*) AS request\_count

FROM requests

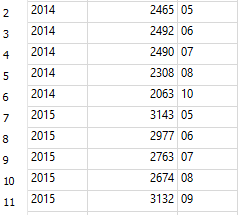
where agency = 'NYPD'

HAVING request\_count > 2000

GROUP BY date, month

) AS request\_counts

GROUP BY year, month;



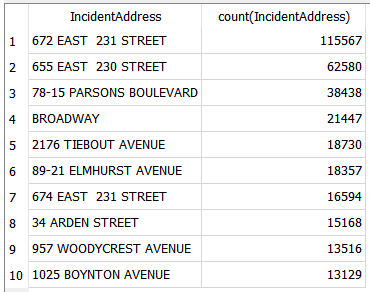
~67s

Update table to follow correct date format to not use SUBSTR() function calls and create index for NYPD

**Query 3**

Get the top 10 most frequent addresses to have a 311 incident.

select IncidentAddress, count(IncidentAddress) from "311\_Requests" group by IncidentAddress order by count(IncidentAddress) desc limit 10



486734 ms, so 486.7 seconds, roughly 8 minutes

Knowing high activity areas could be useful for city planning, police dispatch, ect., so a secondary table keeping track of the incident counts for locations could be helpful and greatly speed up the query.

**Query 4**

Find the Agency that has closed the most requests on Christmas that is not the NYPD

select Agency, AgencyName, count(Agency) as "Requests Closed" from "311\_Requests" where ClosedDate like "12/25/%" and Agency is not "NYPD" group by Agency order by count(Agency) desc limit 1



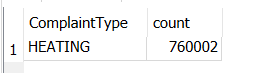
149299 ms, so 149.3 seconds, just over 2 minutes

Add an index for the ClosedDate column, unfortunately, unlike the CreatedDate column that is clustered based on the nature of the table, the ClosedDate column is not clustered, so an index’s functionality would be limited.

**Query 5**

Get the number of heat problems in Brooklyn

SELECT ComplaintType, count() as count FROM NewYork311 WHERE ComplaintType LIKE'%Heat%' AND ParkBorough LIKE '%BROOKLYN%'



Query finished in 1938.239 seconds.

Add an index on Park Borough for Brooklyn

**Query 6**

Find the count of partying grouped by city from occurrences in the description of the complaint and also starting with zip “10”

select city, descriptor, count(descriptor) as description

from requests\_test

where incident\_zip like '10%' and (descriptor like '%Party%' or descriptor like '%party%')

group by city

order by description desc

Ran in 451s

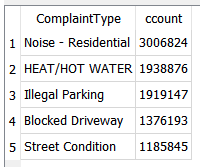


Index descriptions relating to partying, sort on zip

**Query 7**

List the 5 most common complaints in the table based on the complaint type and order by the count of each complaint type in descending order.

SELECT complaint\_type, count() as count FROM Complaint GROUP BY complaint\_type ORDER BY count DESC LIMIT 5;

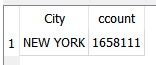


Ran in 356514ms

**Query 8**

Count all of the records where complaint type contains some relation to noise and group by the city.

SELECT city, count() from Complaint where complaint\_type LIKE ‘%Noise%’ GROUP BY city ORDER BY DESC LIMIT 1;



Ran in 234361ms

**Query 9**

Find the city with the least complaint types from the last 2 years

SELECT city, COUNT(\*) from Complaint where created\_date >= DATE\_SUB(NOW(), INTERVAL 2 YEAR) GROUP BY city ORDER BY ASC LIMIT 1;

**Query 10**

Count the number of complaints during hurricane sandy

SELECT COUNT(\*) as count, complaint\_type as type

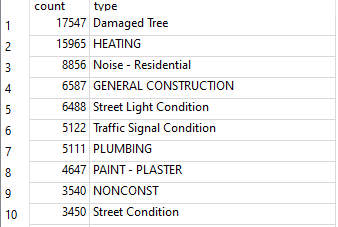
FROM requests\_test

WHERE julianday(strftime('%Y-%m-%d', substr(created\_date, 1, 10))) >= julianday('2012-10-10')

AND julianday(strftime('%Y-%m-%d', substr(created\_date, 1, 10))) <= julianday('2012-11-03')

group by type

order by count desc



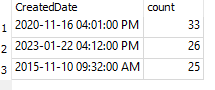
Runtime: ~380s

Index the days between

**Query 11**

Get top 3 days of noise complaints in Brooklyn

SELECT CreatedDate, count() as count FROM NewYork311 Where ComplaintType LIKE'%Noise%' GROUP BY CreatedDate ORDER BY count DESC LIMIT 3;



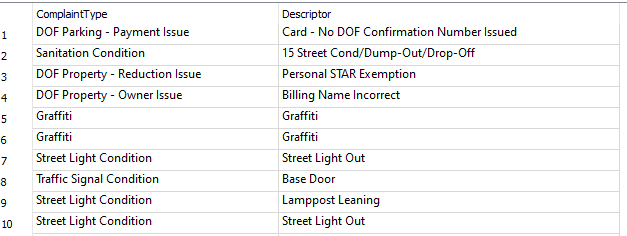
Query finished in 441.510 seconds.

Sort on CreatedDate and Index CreatedDate

**Query 12**

Get and list all complaints in New York created on the end of the world (December 21st, 2012)

SELECT ComplaintType, Descriptor FROM NewYork311 WHERE CreatedDate LIKE'%2012-12-21%';



Query finished in 244.943 seconds.

Sort on CreatedDate and Index CreatedDate