SQL-Mongo Project – Spatial Data of US Wildfires

BUAN 6320

Chirag Hamirani
Sudheer Akkirala
Vinay Krishna Vuppala
R Vidya Reddy

Group 5

Activity	Chirag	Sudheer	Vinay	Vidya
Prepared Data Model and Created Physical DB	x	x	x	x
Loaded Data into Database		x		
Write SQL Queries	x	x		
Prepared Mongo Database	x			
Loaded data into Mongo DB	x			
Wrote Mongo Queries	x			
Prepared Report	x		x	x
Reviewed Report	x			x

Contents

Data Model

Assumptions/Notes About Data Entities and Relationships

Functional Requirements:
Functionally our project has two modules:

- Incidents being the Fires
- Incident reporting being the Reporting

Technical Requirements:
The data model has four main entities: the fires, the owners on whose land the fire occurred, the report on the fire incident and the agency who reported it.

The detailed information about the 4 main entities are as follows:

1.) Key Entity 1: Fires
The fires table contains the details about the incident like
- Discovery date and time
- Containment date and time
- Latitude and longitude
- State and County
- FIPS code and name

Relationships:
- One or many Fires can occur in one Owner's property
- One or many Fires can occur under each FireClass
- One or many Fires can happen due the same cause
- One fire can be reported only once

Assumptions:
- Multiple Fires can occur in the same property
- Multiple Fires can occur at the same time in different properties
- Each cause can have many fires categorized under it

2.) Key Entity 2: Owner
The Owner table has
- Details about the primary owner at the time of the fire
- Details about the person managing the property at the time of the fire

Relationships:
- One or many Fires can occur in the property owned by/ managed by the same person

3.) Key Entity 3: Reporting
The Reporting incidents have
- Source system and type
- Local fire report ID
- Local incident ID
- Fire code and name
- Reporting agency who reported the incident

Relationships:
- One to many reports can be there under one MTBS
- One to many reports can be there under one ICS identifier
- There can only to one report for one fire

Assumptions:
- Many fire reports can come under one complex
- The agency name is local but its recognized on a globally to back track the fire reporting

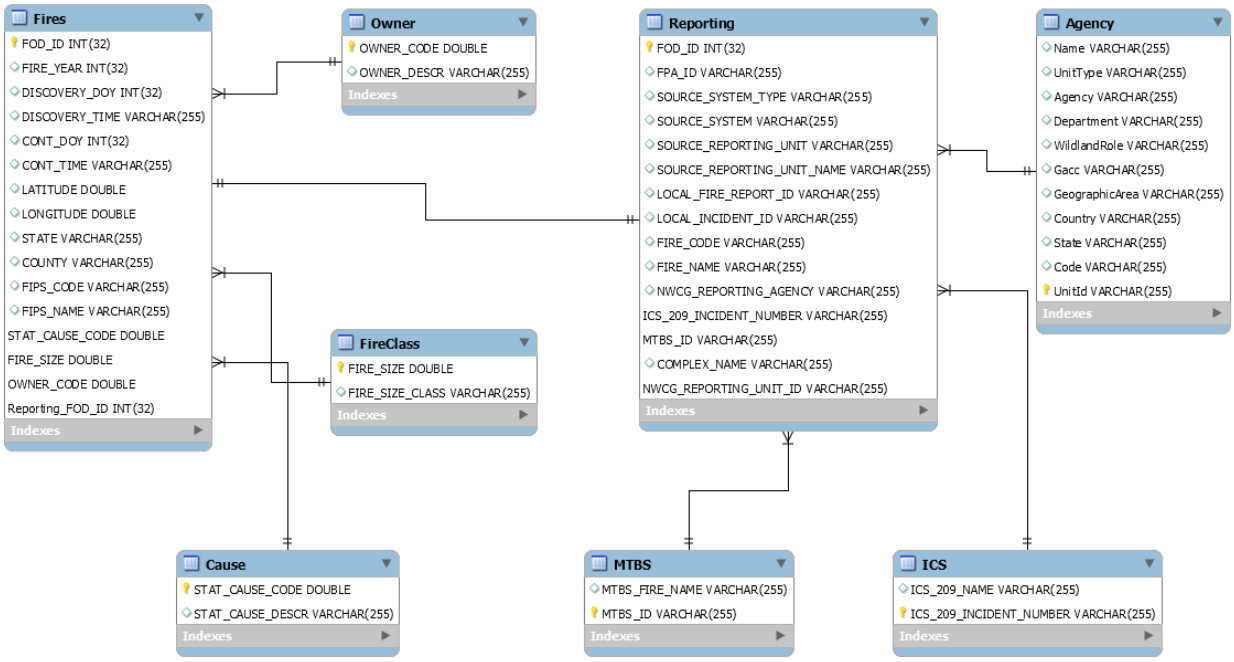
4.) Key Entity 4: Agency
Agency table has
- Agency name
- Department
- Wild and Role
- Geographic area where the unit is location
- County and State

Relationships:
- Agency can have one to many reports
- Agency can report one or many incidents

Other than the four main entities, the data model has four more tables which are:

1.) FireClass: FireClass table has details about the size of the fire and what class it comes under depending on the perimeter of the fire.
2.) Cause: Cause table has details about the cause code of the fire and the cause description.
3.) ICS: ICS table has incident identifier name and number that is present in the report.
4.) MTBS: MTBS table has details about the trends in severity of the fires which are obtained from various aspects obtained from the wildlife fire reports.

Entity-Relationship Diagram



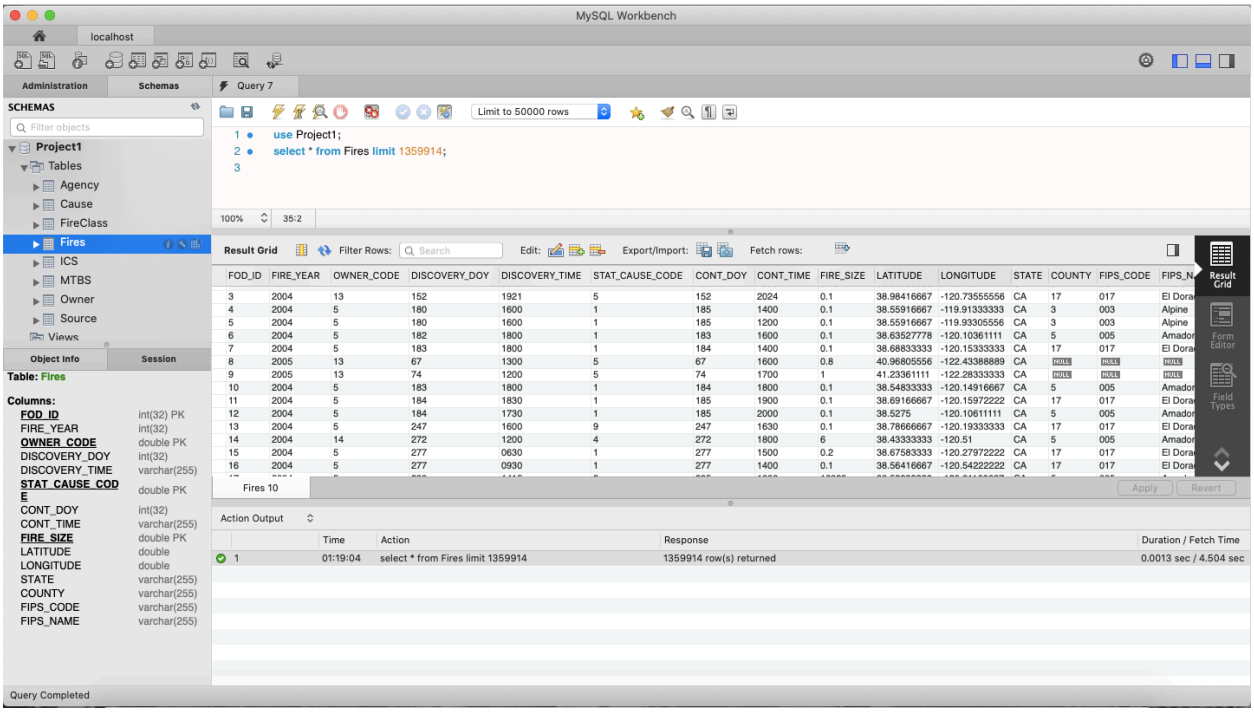
Physical Database

Assumptions/Notes About Data Set

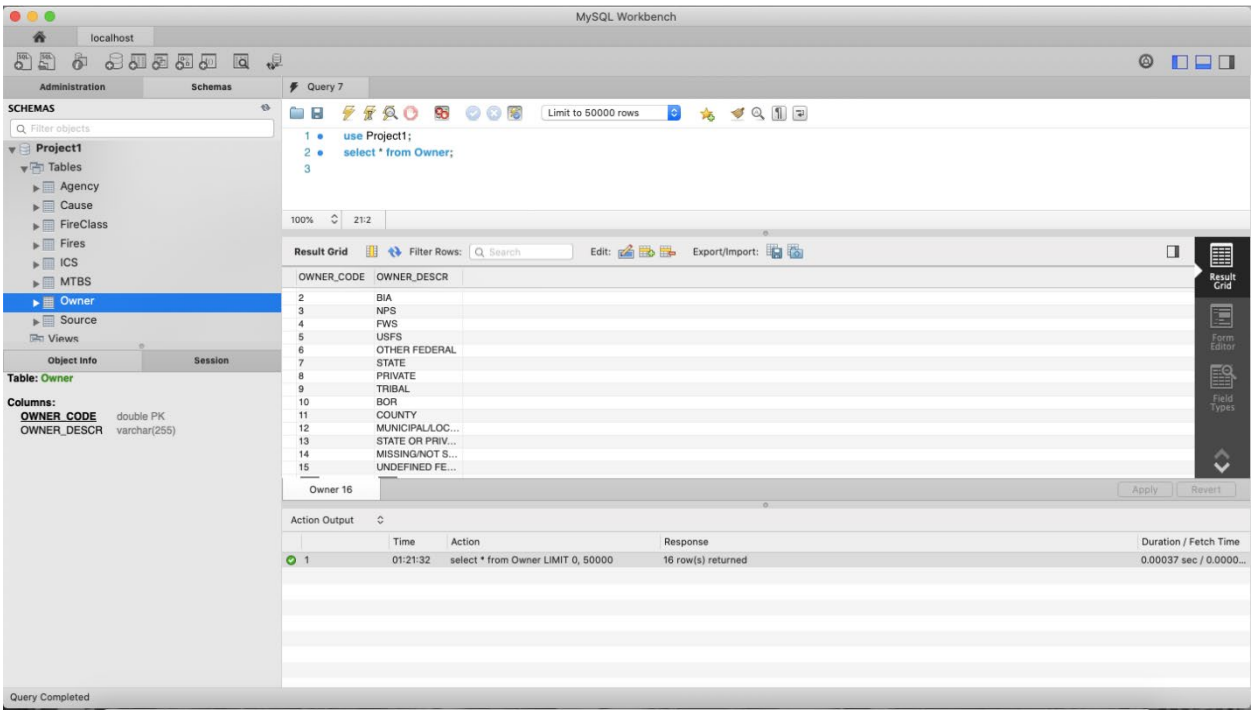
- The shapes column was not used in our data model as it has only null values and it doesn't impose any significance.
- FIPS Name and County is the same.
- FIPS Code and County Code is the same.

Screen shot of Physical Database objects

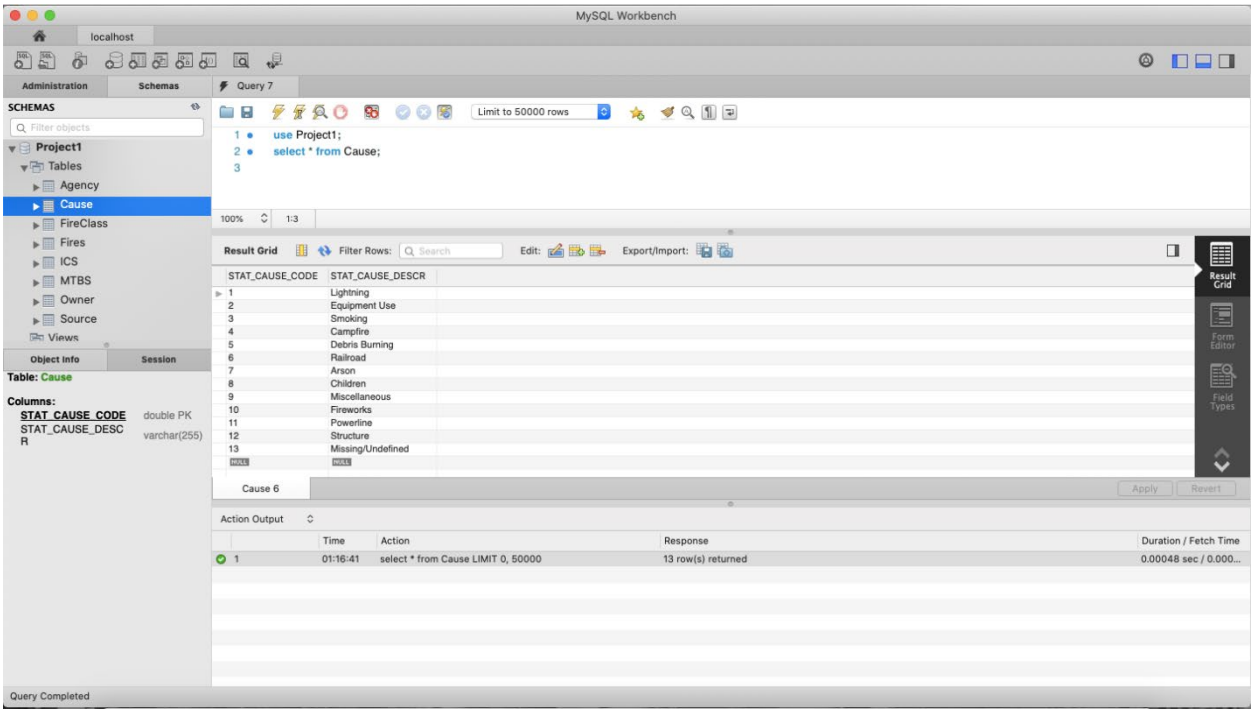
Fires Table



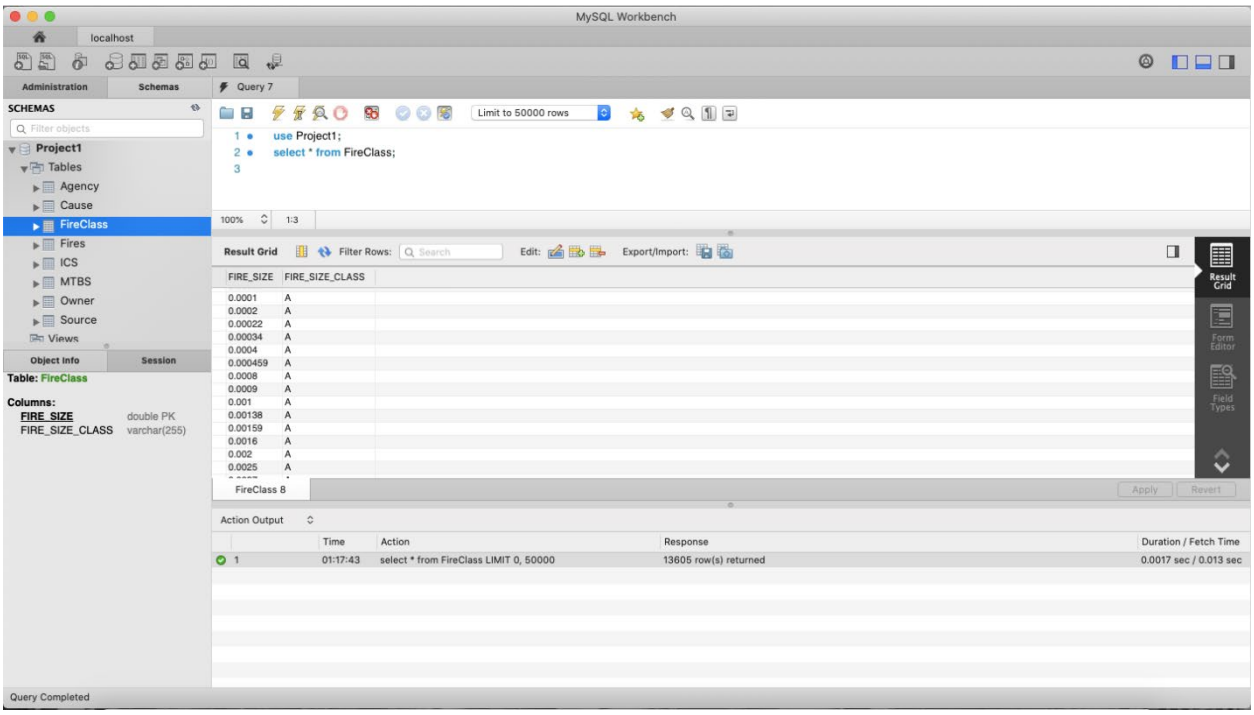
Owner Table



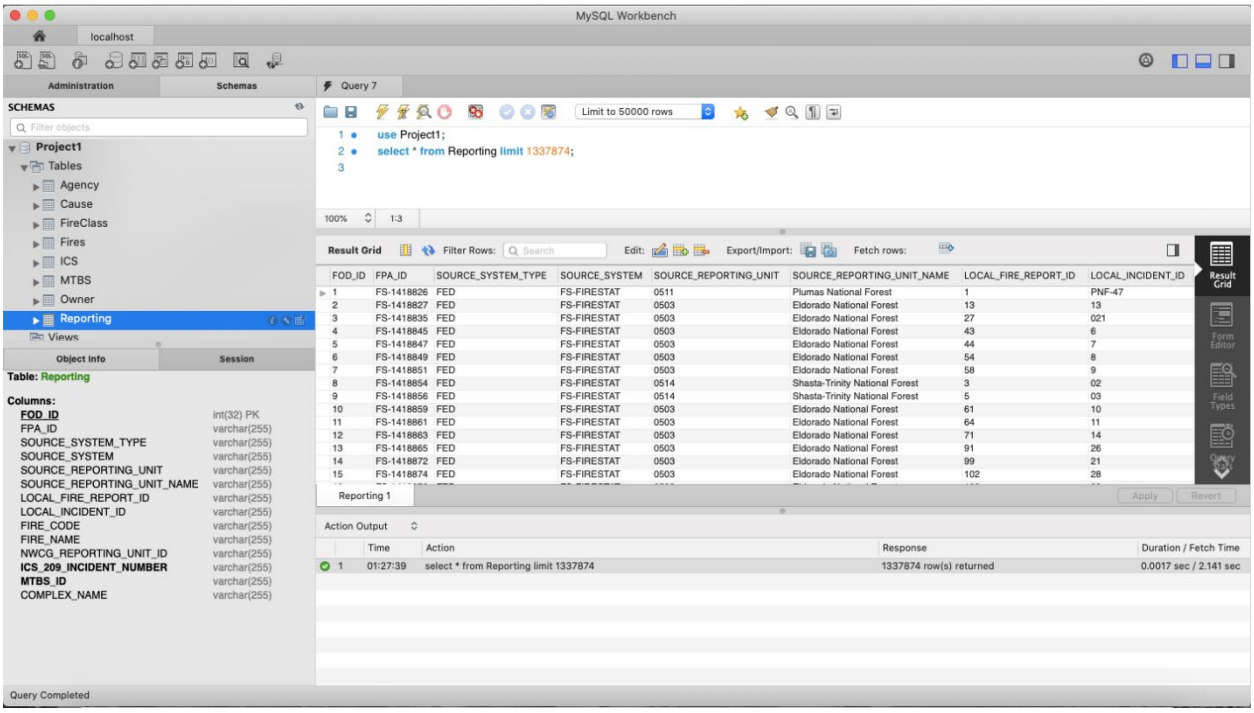
Cause Table



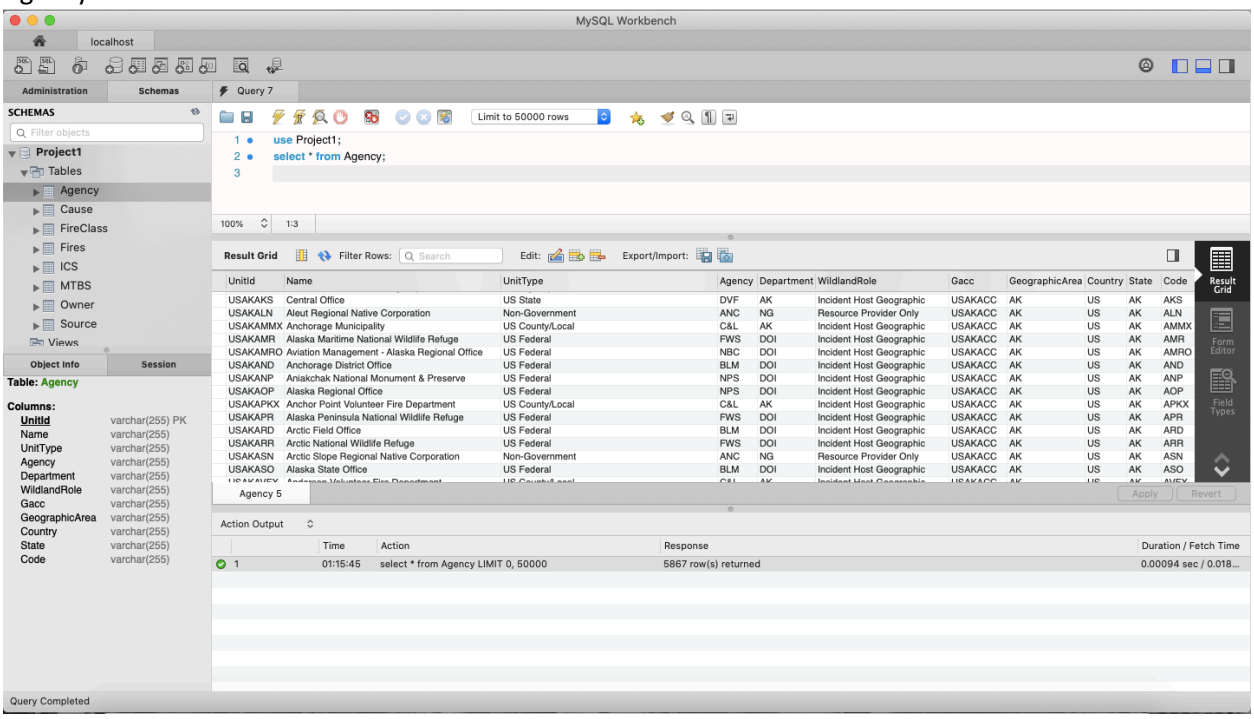
FireClass Table



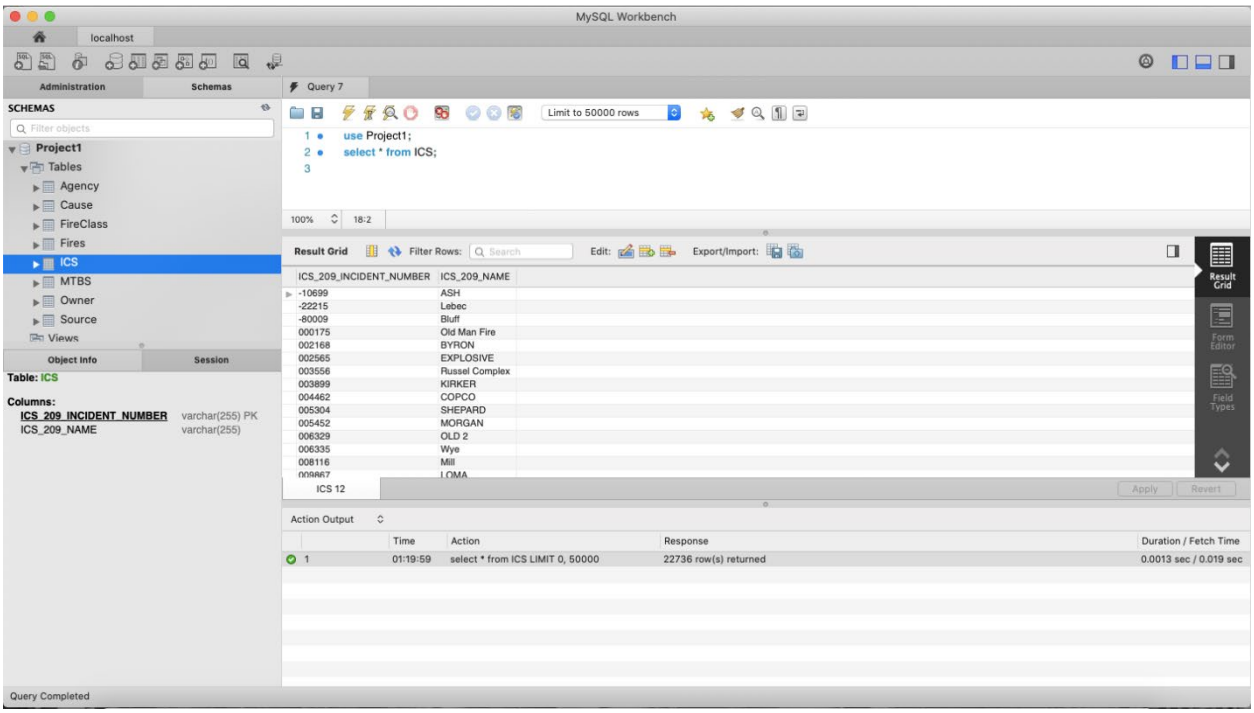
Reporting Table



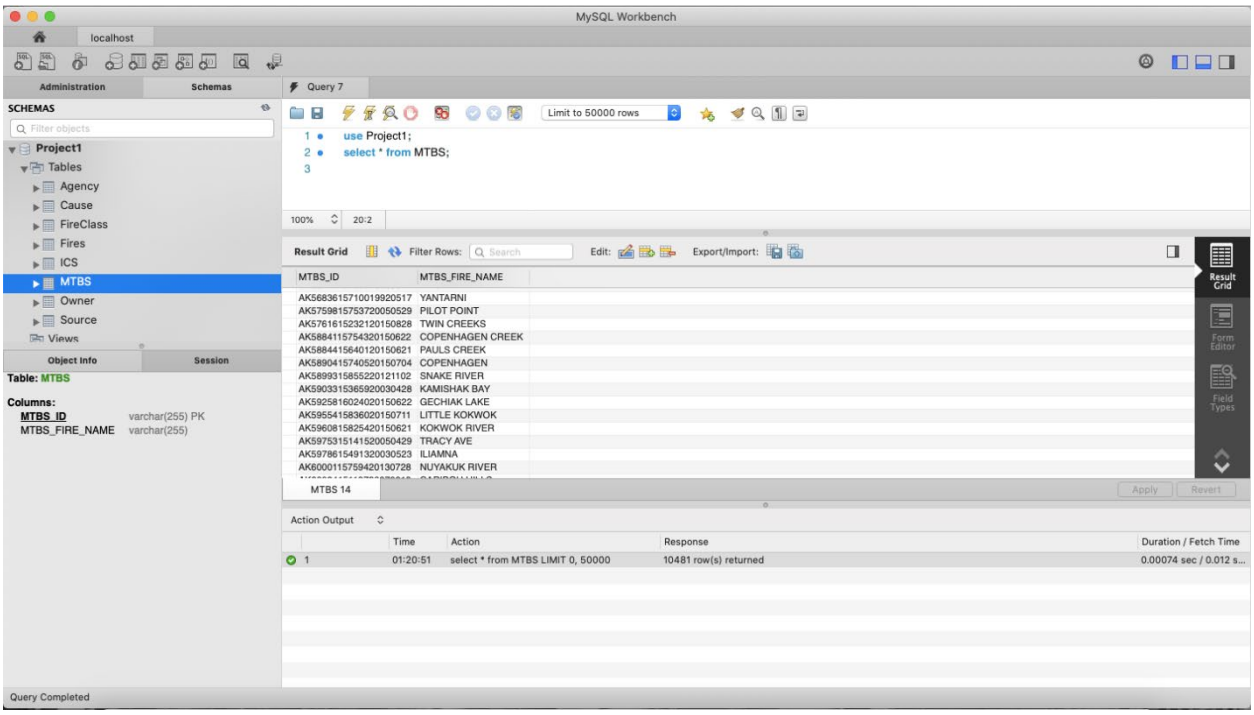
Agency Table



ICS Table



MTBS Table



Data in the Database

Table Name	Primary Key	Foreign Key	# of Rows in Table
Fires	FOD_ID	STAT_CAUSE_CODE FIRE_SIZE OWNER_CODE REPORTING_FOD_ID	1359914
Owner	OWNER_CODE		16
FireClass	FIRE_SIZE		13605
Cause	STAT_CAUSE_CODE		13
Reporting	FOD_ID	ICS_209_INCIDENT_NUMBER MTBS_ID NWCG_REPORTING_UNIT_ID	1337874
Agency	UnitID		5867
MTBS	MTBS_ID		10481
ICS	ICS_209_INCIDENT_NUMBER		22736

SQL Queries

Query 1

Question : A leading beverage company has announced a billion-dollar fund for removing debris from forests, rivers and mountains in the US. All states are interested. Which 2 states have the least chance to win a share of the fund?

Notes:

# of rows in result: 2

Translation: States with less count of fires caused by debris burning have less chance of the fund.

Cleanup: select state, count of fires from fires join cause where cause description is 'Debris Burning' sorting by count of fires ascending order and limit 2.

```
8  use Project1;
9  select f.STATE
10 from Fires f join Cause c on f.STAT_CAUSE_CODE = c.STAT_CAUSE_CODE
11 where c.STAT_CAUSE_DESCR in ('Debris Burning')
12 group by f.STATE
13 ORDER BY count(f.FOD_ID) ASC
14 limit 2;
15
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

STATE
DC
RI

Result Grid | Form Editor | Result 2 | Read Only

Interpretations:

The 2 states which have the least chance to win a share of the fund are District of Columbia and Rhode Island as these have been the least successful in removing debris from forests, rivers and mountains among all the states.

Query 2

Question: One of the reporting agencies has suggested that children be banned from its forests unless there is one adult for every 4 children in a group visiting a forest. Name top 5 forests where this would be the least appropriate.

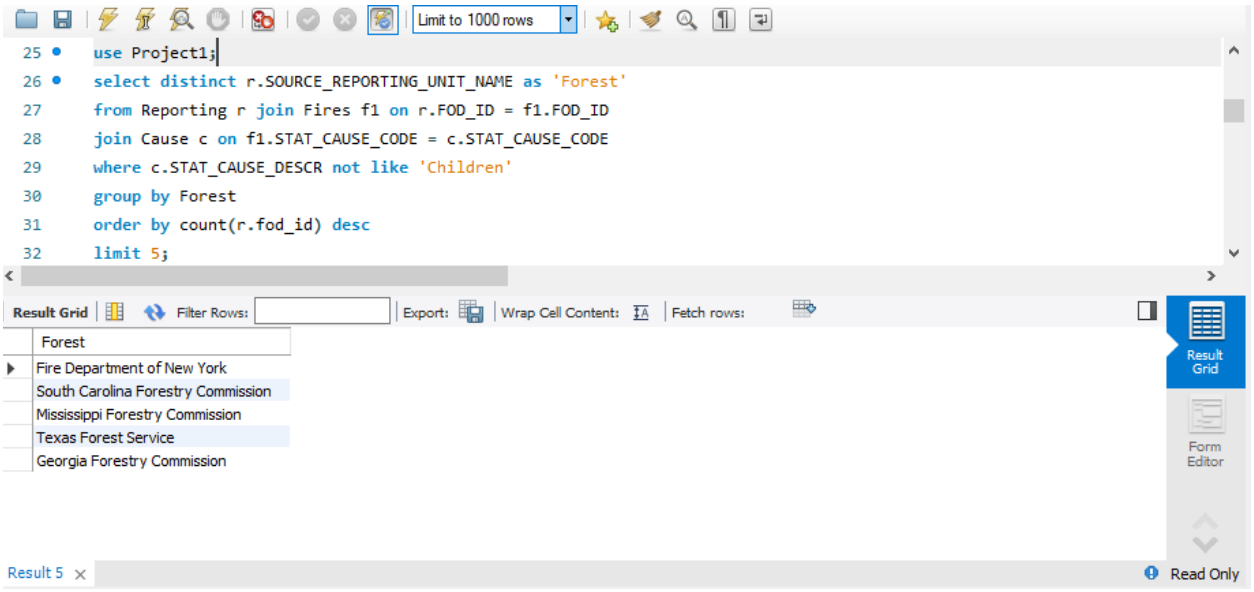
Notes:

- Assumption 1: forests with most fires caused by other causes apart from children will be inappropriate places.
- Assumption 2: higher the non-chidren fire count, less appropriate they are.

# of rows in result: 5

Translation: select the forest where the fires are not caused by children and does not require the ban.

Cleanup: select Forest from Reporting, Fires and Cause where cause description is not children group by forest order by count of fires limit 5.



Interpretations:

The top 5 forests where it is safe to allow children or in other words it would be least appropriate to ban children are the ones that are reported by Fire Department of New York, South Carolina Forestry Commission, Mississippi Forestry Commission, Texas Forest Service and Georgia Forestry Commission. These 5 forests are the ones that do not have any fires caused by children in the past in the fires database. Hence, these places are safe for children to hike, camp, etc.

Query 3

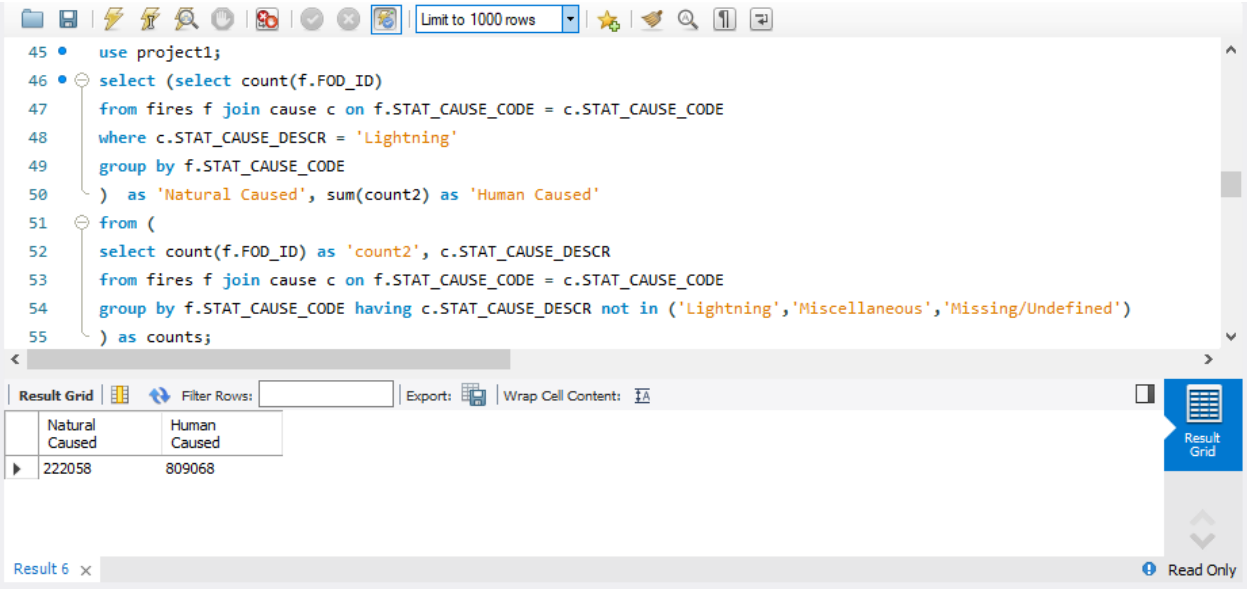
Question: One advocacy group says human actions and nature are equally to blame for most wildfires. Write a query that can help determine the truth of this statement.
Notes:

- Assumption 1: Natural cause is lightning alone.
- Assumption 2: Human cause is everything apart from lightening, misc and undefined.

# of rows in result: 1

Translation: Compare fires caused by lightning and other human reasons.

Clean up: select count of fires from fires join causes where cause description is 'Lightning' grouped by cause description displayed as 'Natural cause' and sum of counts as 'Human caused' from count, cause description from fires join cause group by cause description having cause description not in 'Lightning', 'Miscellaneous', 'Missing/Undefined'.



Interpretations:

We can see from the results that the fires caused by human actions are almost four times the fires caused by nature. The advocacy group was wrong to say that human actions and nature are equally to blame for wildfires. Human actions are more responsible for causing wildfires and necessary precautions have to be taken to avoid more forest fires in the future.

Query 4

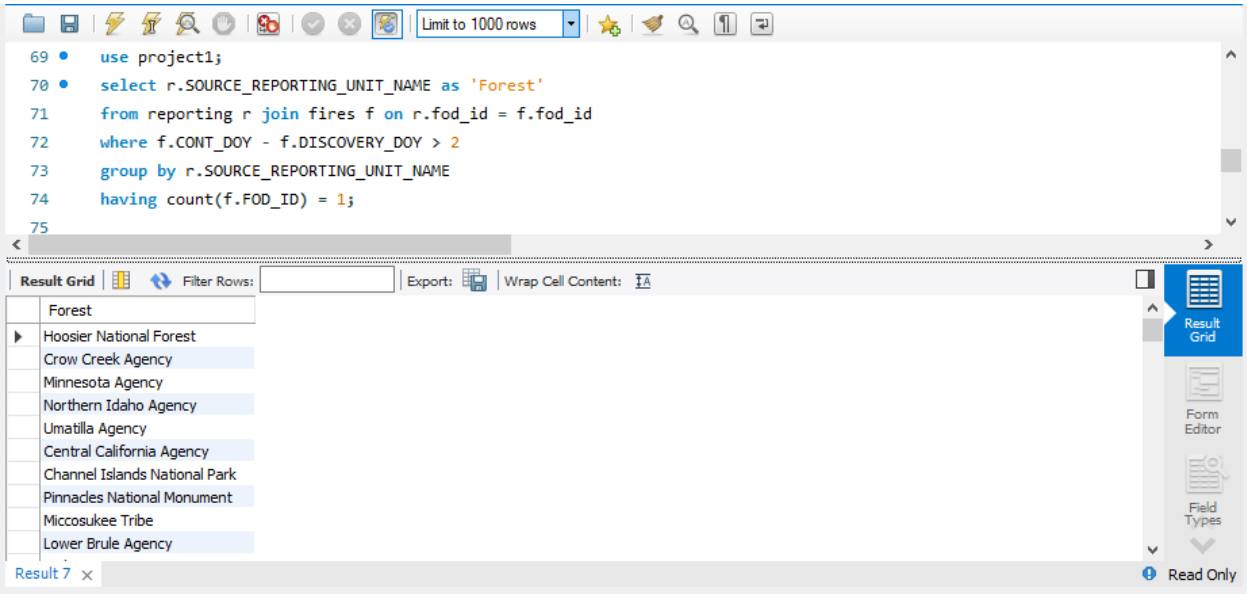
Question: What were the forests that had only one fire that lasted more than two days?

Notes:

of rows in result: 132

Translation: select forest from reporting join fires where contained day minus discovered day is greater than 2 group by forest having count = 1.

Clean up: select source reporting unit name from reporting join fires where contained day - discovered day > 2 group by source reporting unit name having count = 1.



Interpretations:

The forests that had only one fire that lasted more than 2 days are the ones listed in the results.

Query 5

Question: Which state had fires only in the second half of the calendar years?

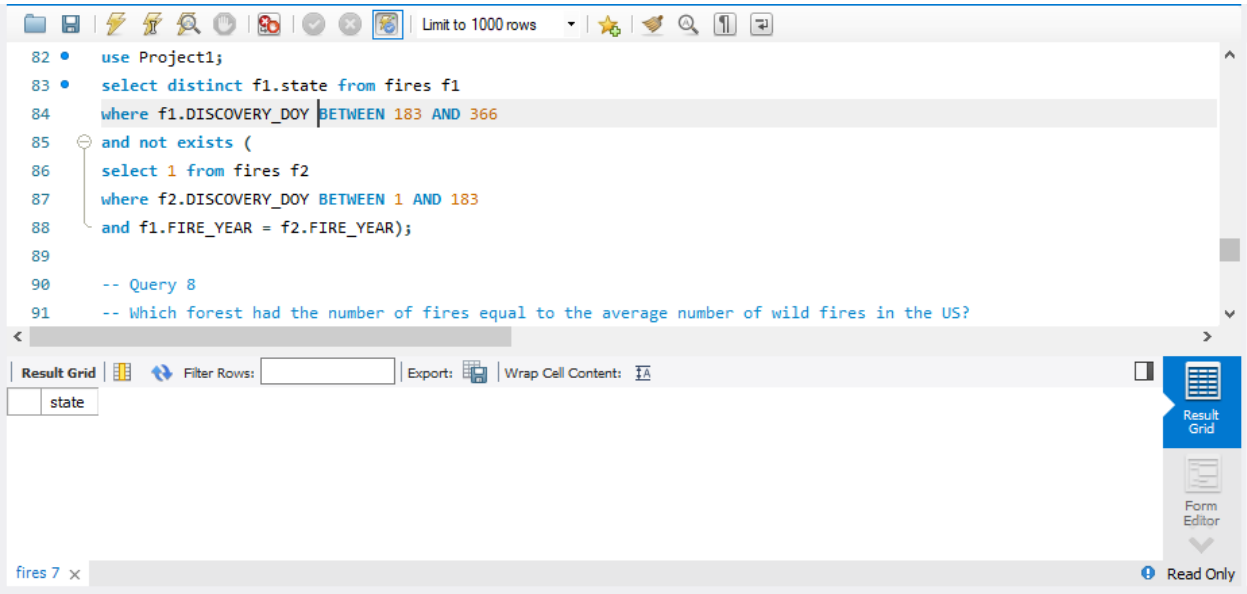
Notes:

- Assumption 1: For non-leap years the second half of the year is from 183 to 365.
- Assumption2: We have to check for states with fires in second half of the year, where they dint have fires in same year's first half

# of rows in result: 0

Translation: select states from Fires with discovery date in second half of year only.

Cleanup: select state from Fires where Discovery Day of Year is between 183 and 366 and state is not in select state from fires where discovery say of the year is between 1 and 183.



Interpretations:

There are no states which have had fires only in the second half of the calendar year.

Query 6

Question: Which forest had the number of fires equal to the average number of wild fires in the US?

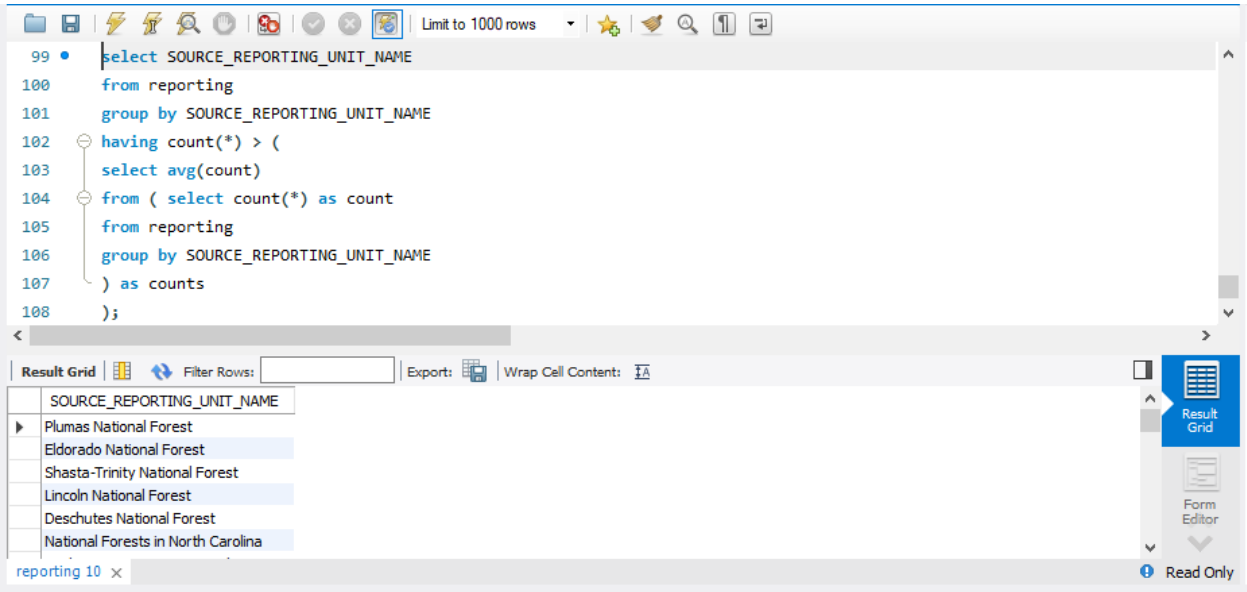
Notes:

- Assumption 1: All fires in given table are from US.
- Assumption 2: Since there so many records, we will not find an exact match for average value.
- Assumption 3: For business, let's find forests with above average count.

# of rows in result: 360

Translation: Select forests in source reporting unit name from reporting with sum of fires reported equal to average number of fires in the us.

Clean up: Select source reporting unit name from reporting where sum of fires reported equal to select average of fires from reporting.



Interpretations:

The forests that have fires which are above the average number of fires occurring in the US are the ones listed above in the results. These forests have a higher danger of catching fire or have had higher occurrences of fires than the other forests in the US which means these forests are not as safe to visit. It is advisable to not pay visits to these and find alternative forests instead. It is also advisable to evacuate any animals that might be in danger or take precautions to provide them safe habitats in case these forests have high animal population.

Data Review for MongoDB

Attributes and Relationships between Collections

While functional and technical requirements have been kept same for mongo database as well, there are few notes about data migration which will explain the collections and data in it.

Notes About Collections & Data Migration

1. Collections in mongo are constructed with same structure as that of sql relational database for ease of understanding
2. Every table had data extracted from sql into csv files, which were then imported in mongo using mongo compass
3. All the relational tables are made into collections with same functional value
4. While migrating, fires and reporting tables had over 13 million records, these were split into over 20 files each to facilitate data load
5. All the data load happened flawlessly, except one file in reporting collection, during which connection to database was lost. The file was then reimported, which may have led to certain duplicate records. However, querying is done to avoid any such duplicates affecting the results
6. Since the collections and tables have same functional values, the results are expected to reflect similar conclusions. Any difference can open opportunity to discussion of ways in which both software function and error in querying.

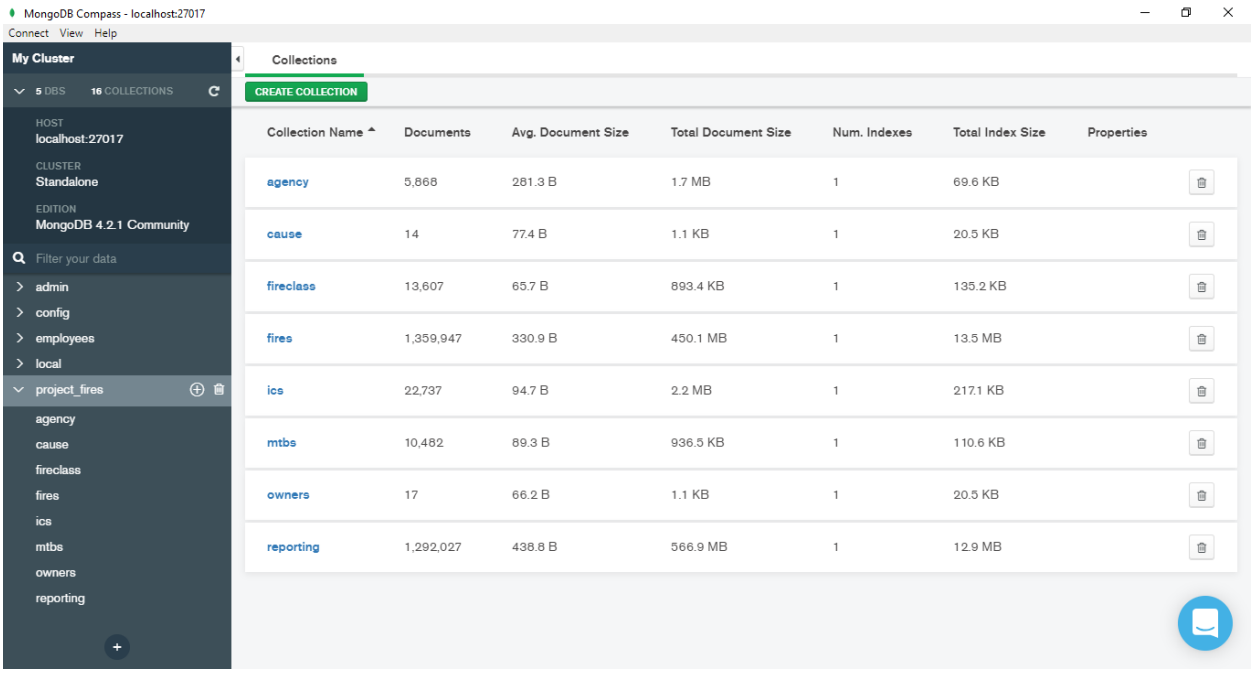
Physical Mongo Database

Assumptions/Notes About Data Set

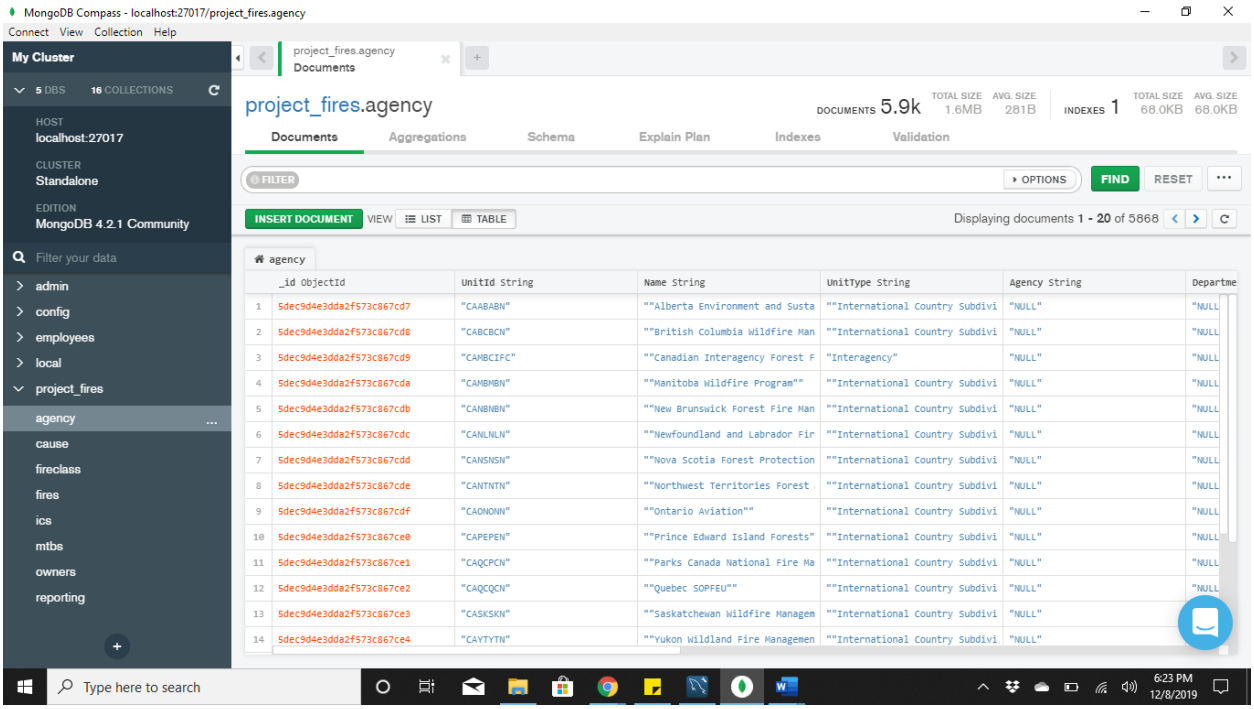
- The shapes field was not used as it has only null values and it doesn't impose any significance.
- FIPS Name field and County field are the same.
- FIPS Code field and County field Code are the same.

Screen shot of Physical Database objects (Database, Collections and Attributes)

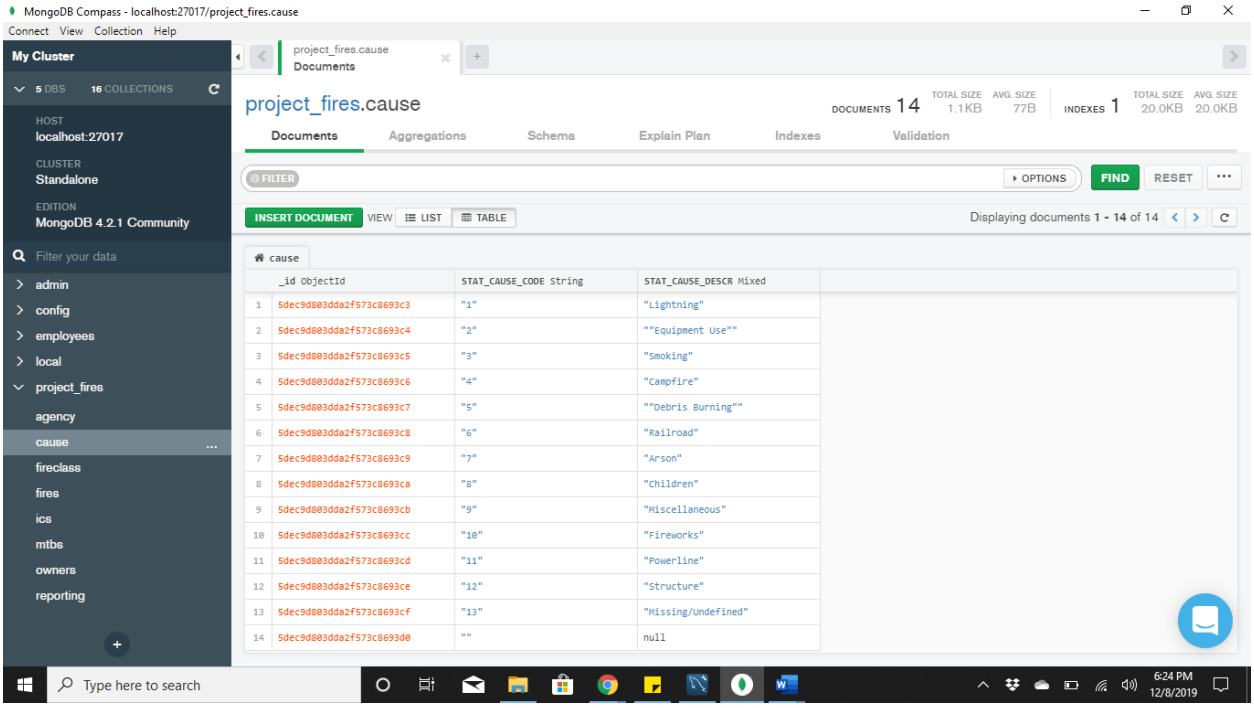
Project Database



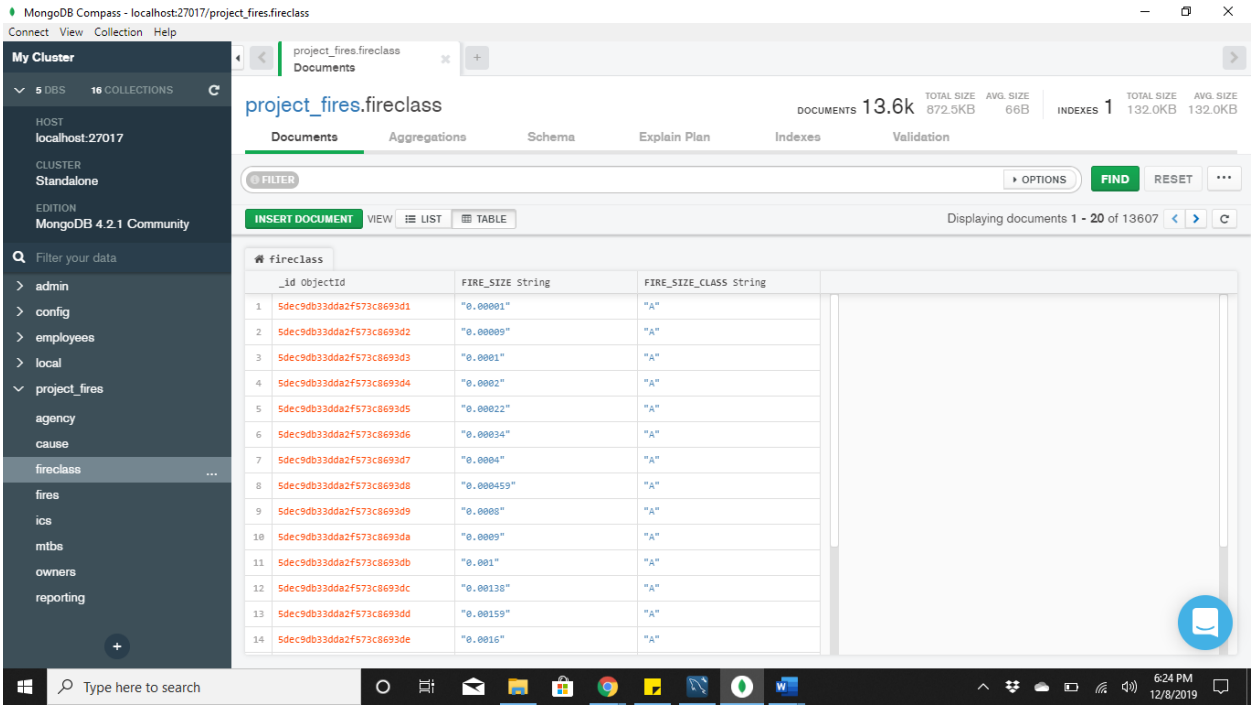
Agency Collection



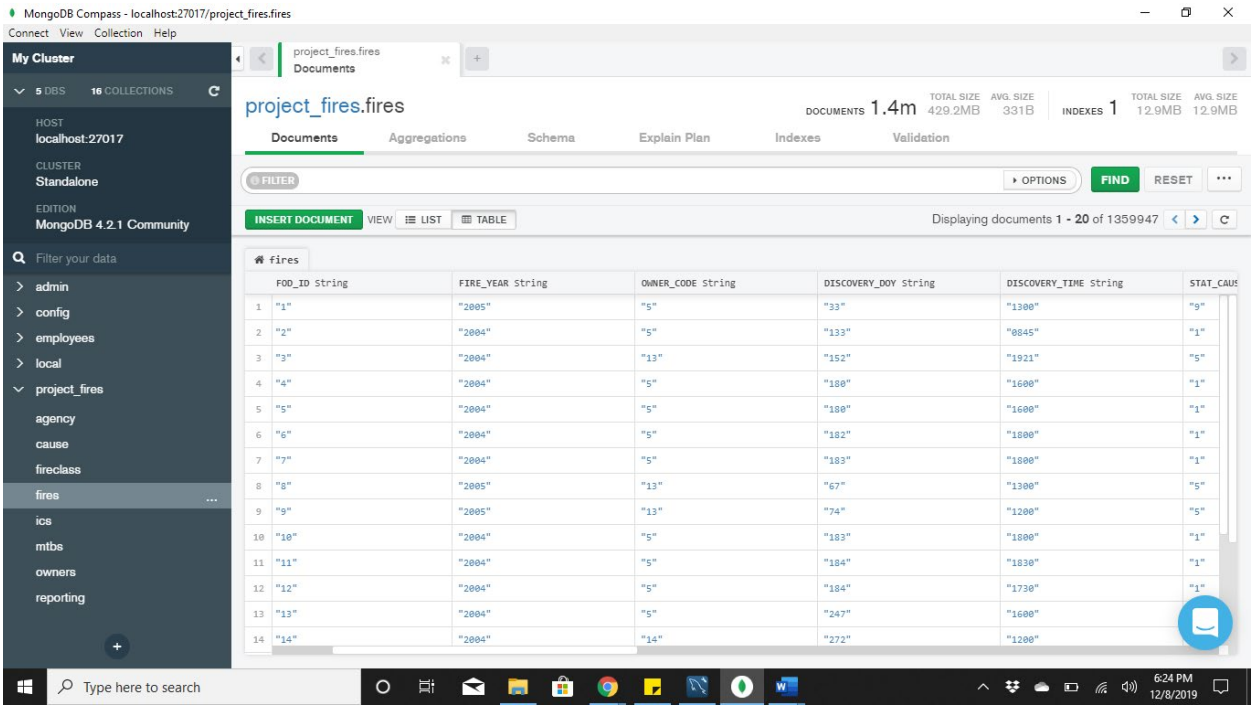
Cause Collection



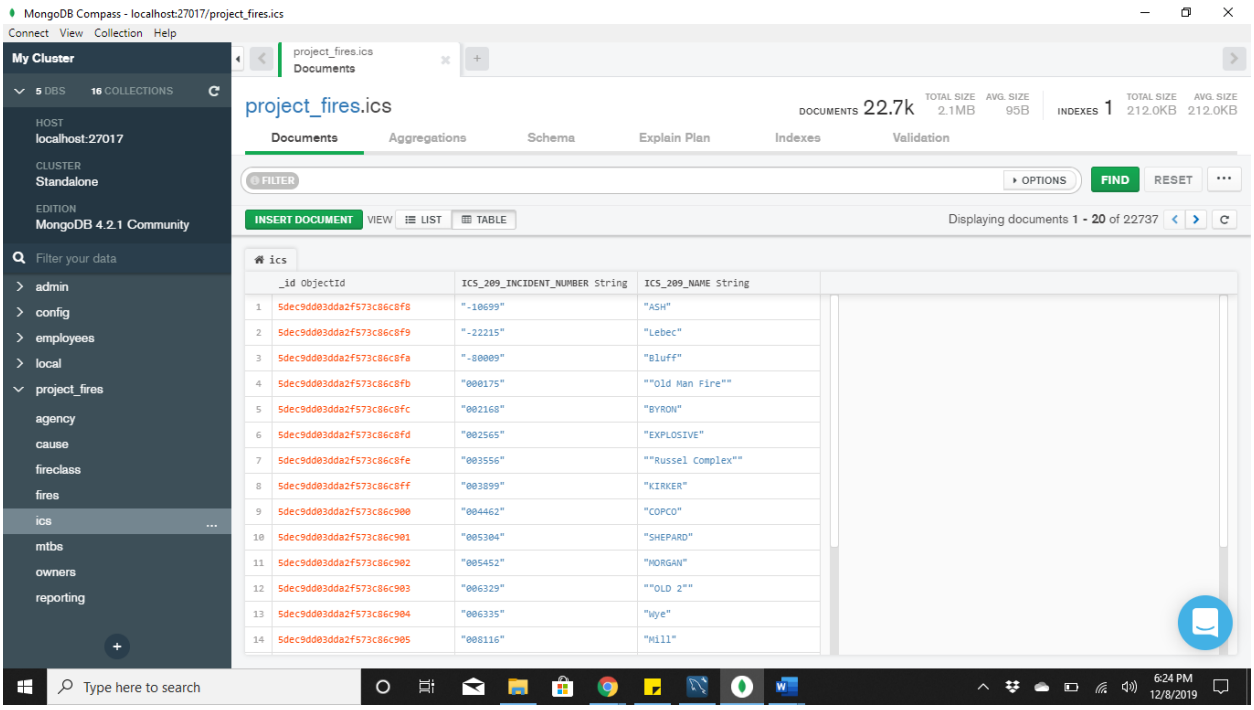
Fireclass Collection



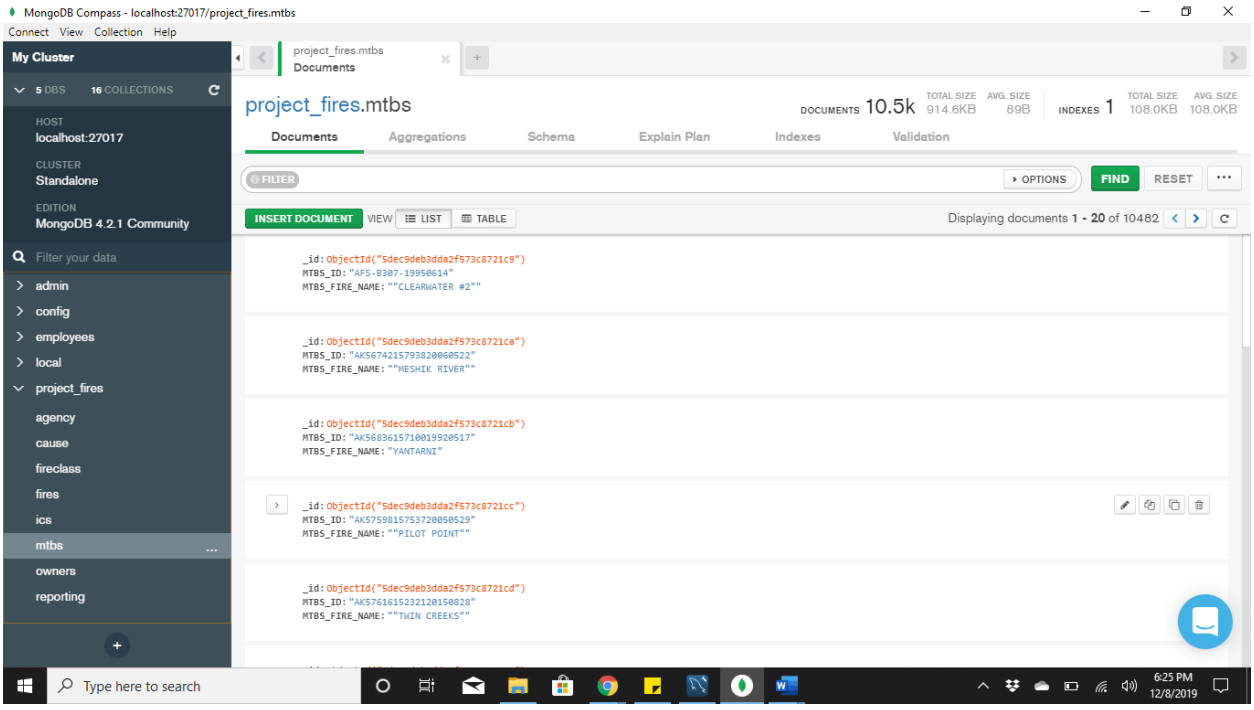
Fires Collection



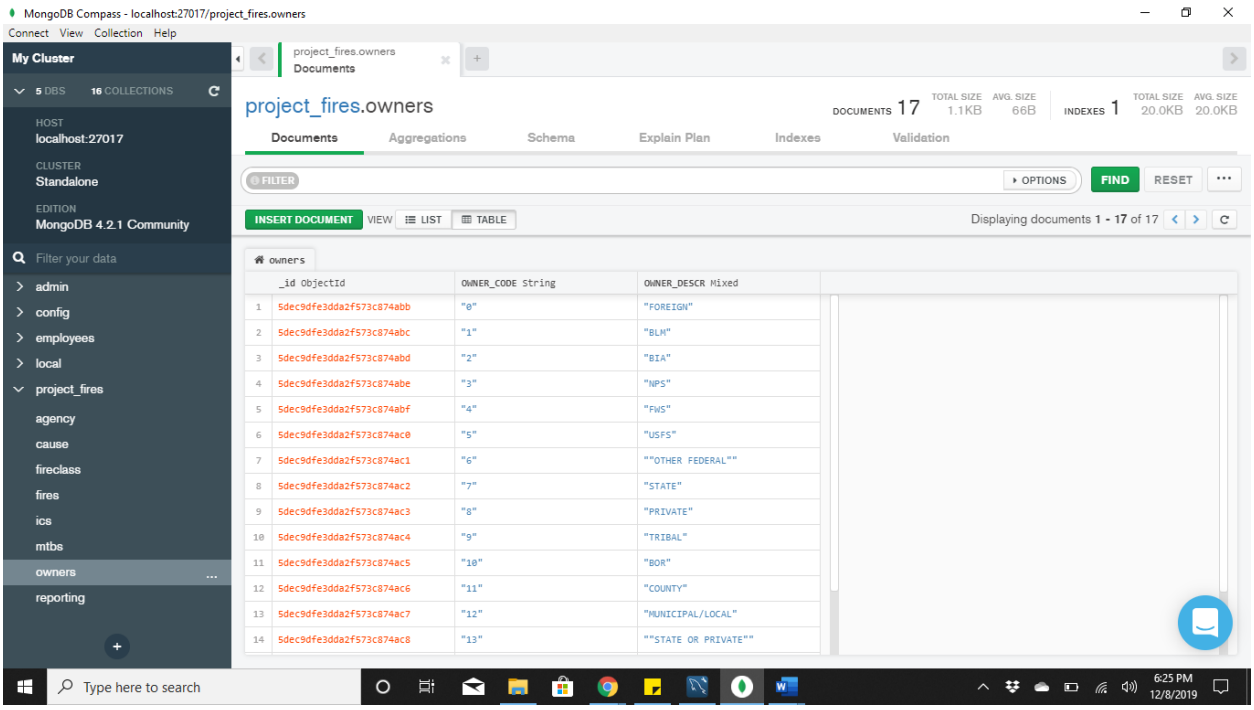
ICS collection



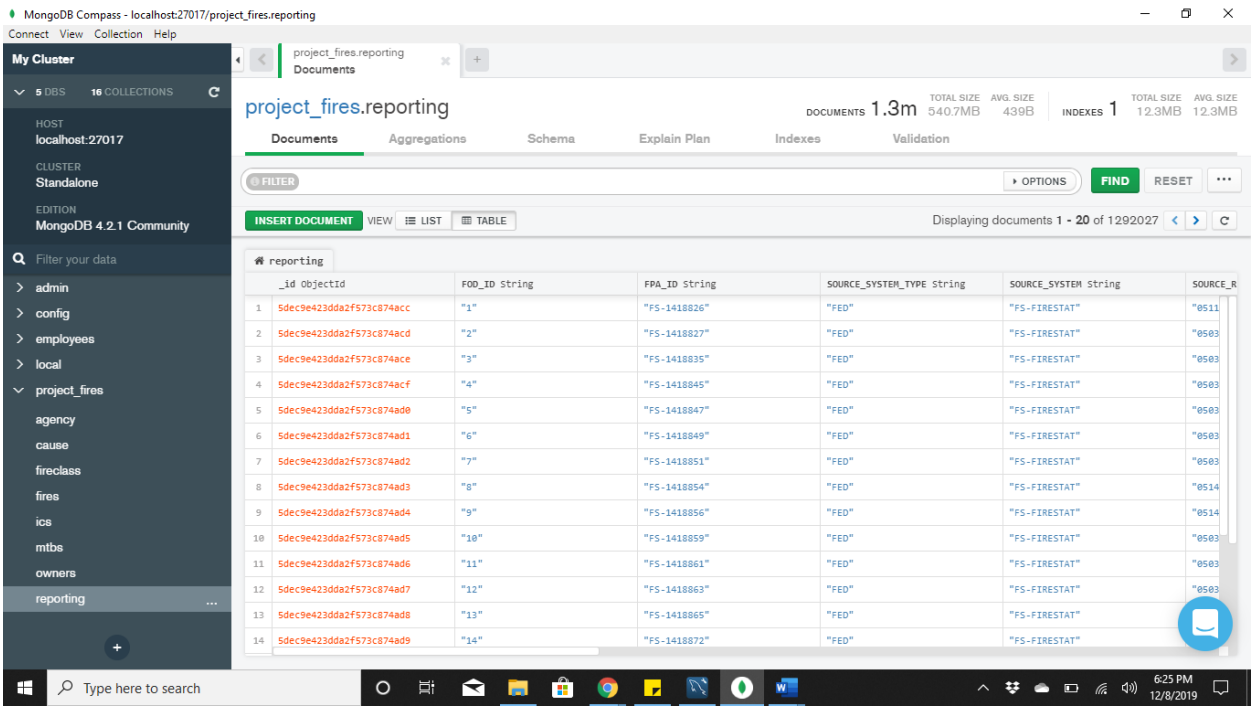
MTBS Collection



Owners Collection



Reporting Collection



Data in the Database

Collection Name	Relationships with Other Collections (if any)	# of Documents in Collection
Fires	Owner collection (Owner_Code) FireClass collection (Fire_Size) Cause Collection (Stat_Cause) Reporting Collection (FOD_ID)	1359914
Owner	Fire Collection (Owner_Code)	16
FireClass	Fire Collection (Fire_Size)	13605
Cause	Fire Collection (Stat_Cause)	13
Reporting	Agency Collection (UnitId) MTBS Collection (MTBS_ID) ICS Collection (ICS_Incident_Number) Fire Collection (FOD_ID)	1337874
Agency	Reporting Collection (UnitId)	5867
MTBS	Reporting Collection (MTBS_ID)	10481
ICS	Reporting Collection (ICS_Incident_Number)	22736

MongoDB Queries/Code

Query 1

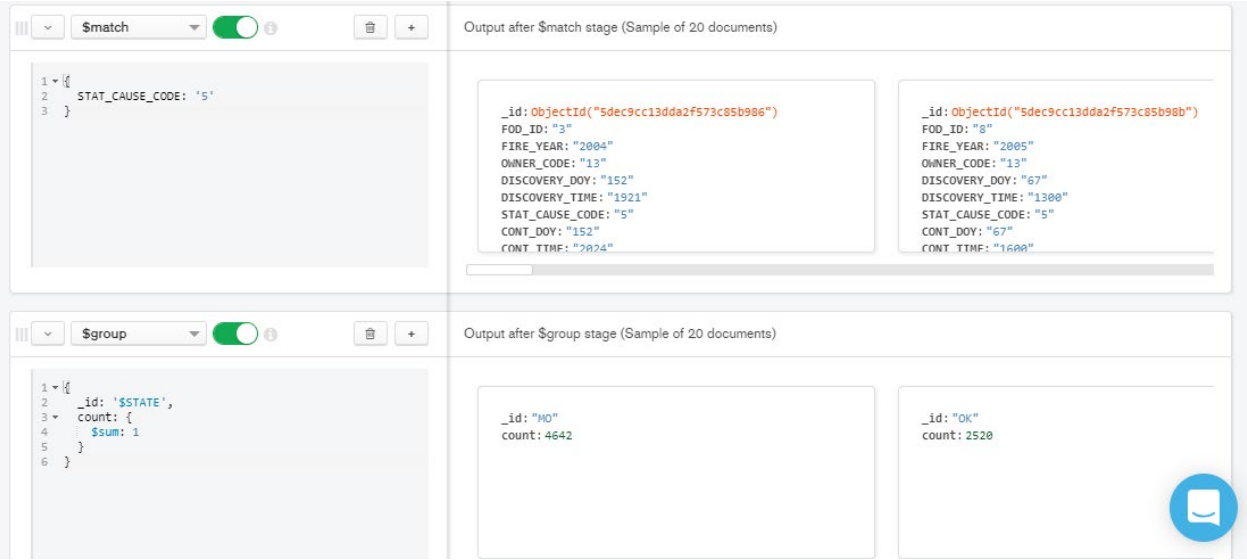
Question : A leading beverage company has announced a billion-dollar fund for removing debris from forests, rivers and mountains in the US. All states are interested. Which 2 states have the least chance to win a share of the fund?

Notes:

# of rows in result: 2

Translation: Match cause code to '5' for Debris, Group by State and generate count. Project state, sort by count, limit at 2

Screen Shot of MongoDB Query/Code and Results



Interpretations:

The 2 states which have the least chance to win a share of the fund are DC & PR as these have been the least successful in removing debris from forests, rivers and mountains among all the states.

Query 2

Question: One of the reporting agencies has suggested that children be banned from its forests unless there is one adult for every 4 children in a group visiting a forest. Name top 5 forests where this would be the least appropriate.

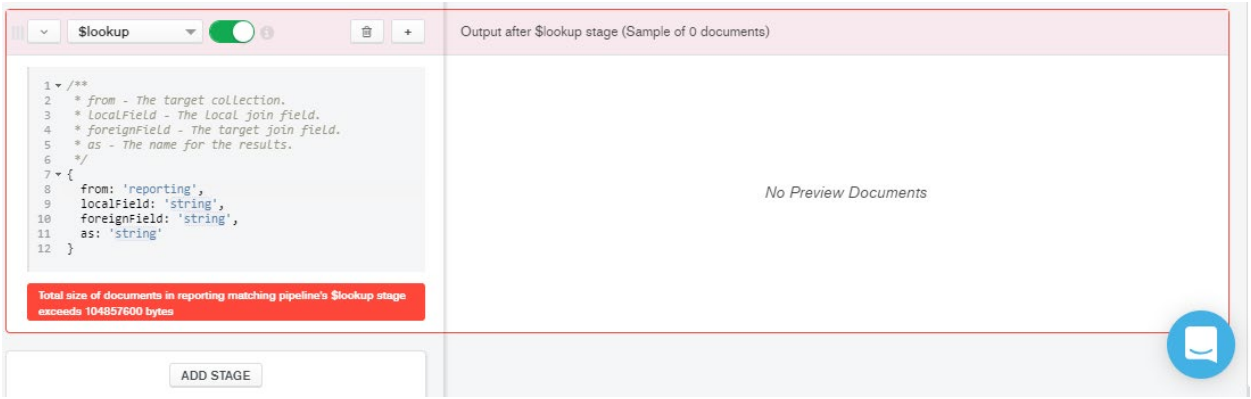
Notes:

- Assumption 1: forests with most fires caused by other causes apart from children will be inappropriate places.
- Assumption 2: higher the non-chidren fire count, lesser appropriate there are.

# of rows in result: NA

Translation: Match cause code as '8' for children, lookup reporting collection with FOD_ID, group by forests and generate count. Project Forest, sort by count desc, limit 5

Screen Shot of MongoDB Query/Code and Results



Interpretations:

Compass throws error for lookup between fires and reporting table, saying the aggregation pipeline exceeds 104mb. This has no possible work around without altering database structure.

Query 3

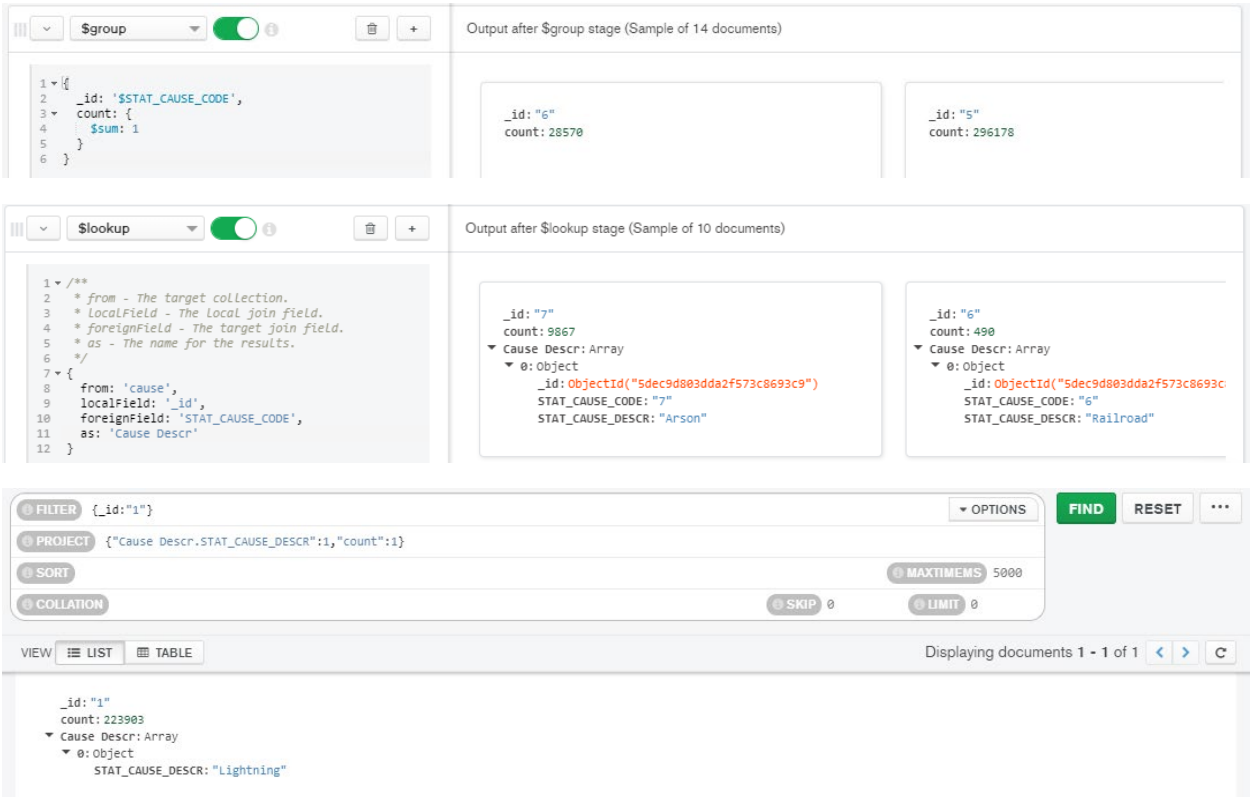
Question: One advocacy group says human actions and nature are equally to blame for most wildfires. Write a query that can help determine the truth of this statement.
Notes:

- Assumption 1: Natural cause is lightening alone.
- Assumption 2: Human cause is everything apart from lightening, misc and undefined.

# of rows in result:

Translation: 1

Screen Shot of MongoDB Query/Code and Results



Interpretations:

We can see from the results that the fires caused by human actions are almost four times the fires caused by nature. The advocacy group was wrong to say that human actions and nature are equally to blame for wildfires. Human actions are more responsible for causing wildfires and necessary precautions have to be taken to avoid more forest fires in the future.

Query 4

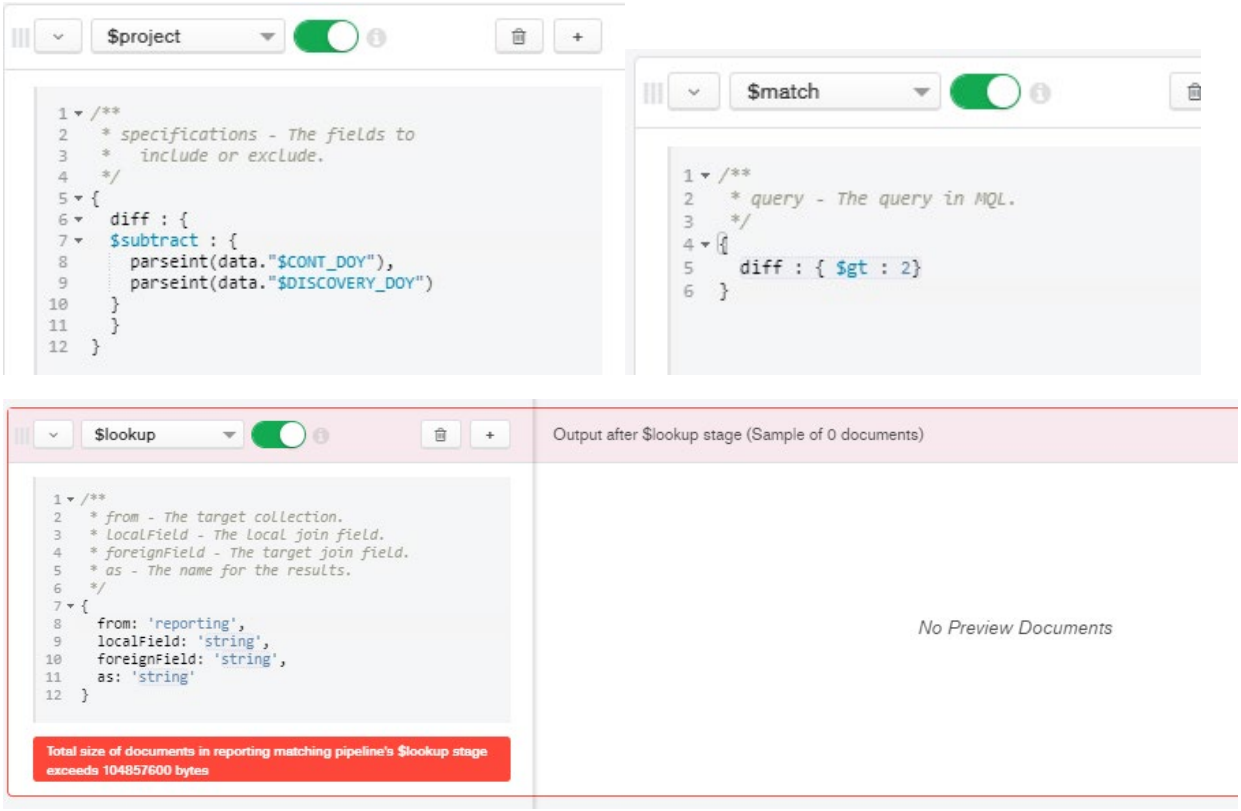
Question: What were the forests that had only one fire that lasted more than two days?

Notes:

of rows in result: NA

Translation: Project 'Diff' as subtract between int values of discovery and contained day. Match diff with greater than 2. Lookup fires FOD_IDs in reporting collection. Group by forests and generate count. Match count to 1 for output.

Screen Shot of MongoDB Query/Code and Results



Interpretations:

Compass throws error for lookup between fires and reporting table, saying the aggregation pipeline exceeds 104mb. This has no possible work around without altering database structure.

Query 5

Question: Which state had fires only in the second half of the calendar years?

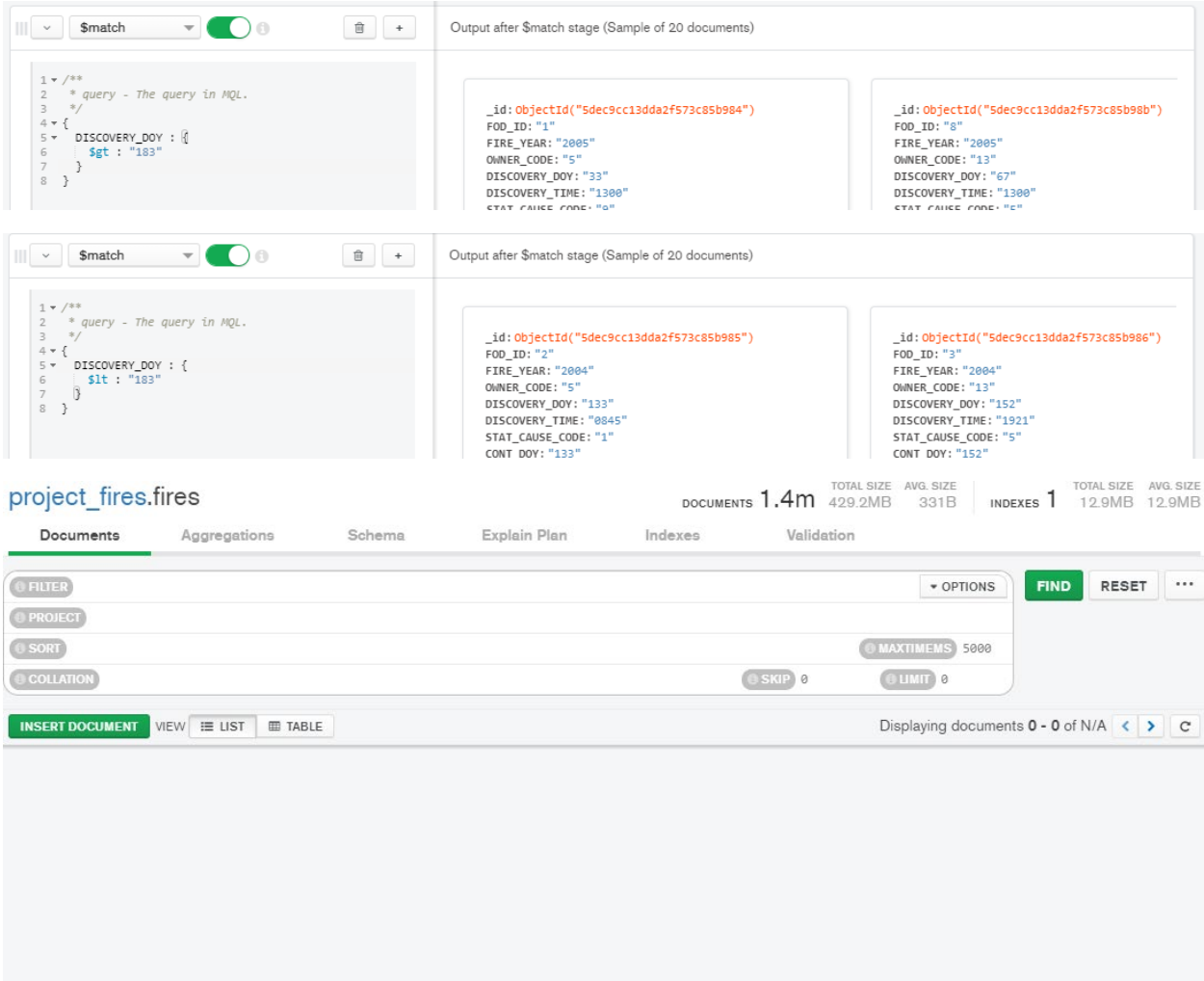
Notes:

- Assumption 1: For non-leap years the second half of the year is from 183 to 365.

# of rows in result: 0

Translation:

Screen Shot of MongoDB Query/Code and Results



Interpretations:

There are no states which have had fires only in the second half of the calendar year.

Query 6

Question: Which forest had the number of fires equal to the average number of wild fires in the US?

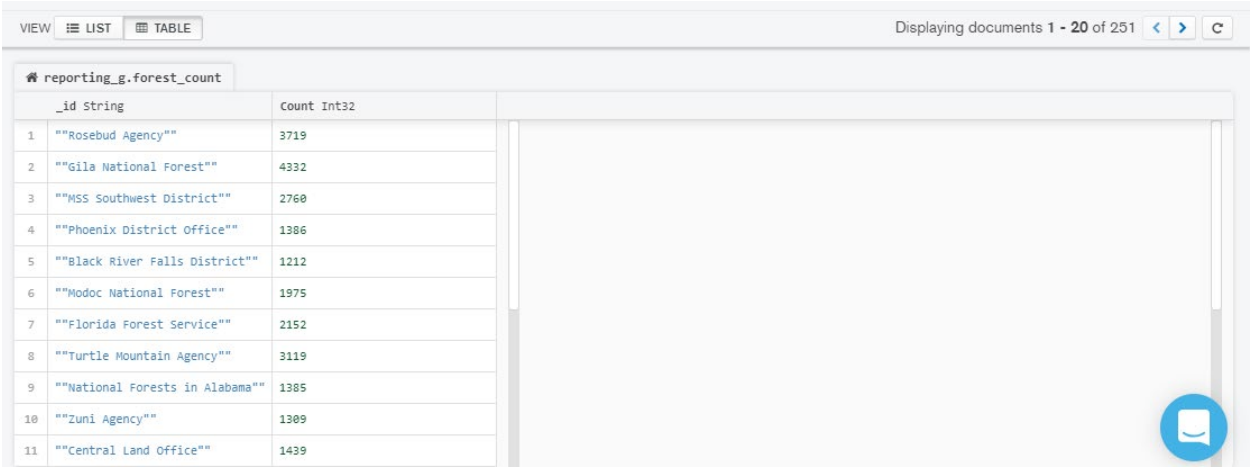
Notes:

- Assumption 1: All fires in given table are from US.
- Assumption 2: Since there so many records, we will not exact match for average value.
- Assumption 3: For business, let's find forests with above average count.

# of rows in result: 251

Translation: Group by forest and generate count. Group by null to generate average of count. Match by count greater than average.

Screen Shot of MongoDB Query/Code and Results



Interpretations:

The forests that have fires which are above the average number of fires occurring in the US are the ones listed above in the results. These forests have a higher danger of catching fire or have had higher occurrences of fires than the other forests in the US which means these forests are not as safe to visit. It is advisable to not pay visits to these and find alternative forests instead. It is also advisable to evacuate any animals that might be in danger or take precautions to provide them safe habitats in case these forests have high animal population.