SQL-Mongo Project

Spatial Data of US Wildfires

By

Sinduja Senthil Kumar

Activity	Status
Prepared Data Model and Created Physical DB	✔
Loaded Data into Database	✔
Wrote SQL Queries	✔
Prepared Mongo Database	✔
Loaded data into Mongo DB	✔
Wrote Mongo Queries	✔
Prepared Report	✔
Reviewed Report	✔

Contents

Data Model

Assumptions/Notes About Data Entities and Relationships

Include assumptions about data entities and their relationships with each other.

Assumptions about data entities and their relationships with each other:

1. There were 3 source system types (type of the source database) and for each type there were multiple source systems (Name/Identifier of the source database) that were linked. Hence separate tables as SOURCE_SYSTEM_TYPES and SOURCE_ SYSTEMS were used with generated primary key and corresponding relationship was given between the tables
2. Similarly, one source system was linked to many source reporting units (agency preparing the fire reports) and hence one to many relationships was given
3. The FIRE_SIZE_CLASS field can be calculated using the FIRE_SIZE field while querying and hence it is dropped
4. The Julian dates in the FIRES table under DISCOVERY_DATE and CONT_DATE is converted and stored as normal dates as it would be easy for understanding as well as for querying
5. The county and the states column from the FIRES table are made as separate tables and are not added to owner's table because it is assumed that a owner can stay in a different place other than the fire location and manage the land
6. It is assumed that the state and county is not an redundant data though the latitude and longitude of the fire is mentioned separately because although the latitude and longitude gives the fire location, they may not necessarily give the point of origin of fire which might be given by the state and county
7. Separate table for FIPS name and FIPS code was generated because one FIPS name contained many FIPS code and one FIPS code contained many FIPS name resulting in the bridging table with one to one relationship
8. The parent field in the NWCG table didn't contain any values and hence it was dropped
9. Though the SOURCE_REPORTING_UNIT_NAME is identical as the NWCG_REPORTING_UNIT_NAME field and can be mapped one to one, there were certain values which were different. These differences were assumed to be manual typing errors and hence they were linked with one-to -one relationship
10. It is assumed that fields related to fire like the DISCOVERY DATE/DAY/TIME, CONTROLLED DATE/DAY/TIME, FIRE SIZE, LATITUDE, LONGITUDE etc., are the most important details that are required when querying about the fire and also this information seems mandatory to understand about the fire and hence they are grouped in a single table with the Global identifier as the primary key

The data model is in 3NF because of the following reasons:

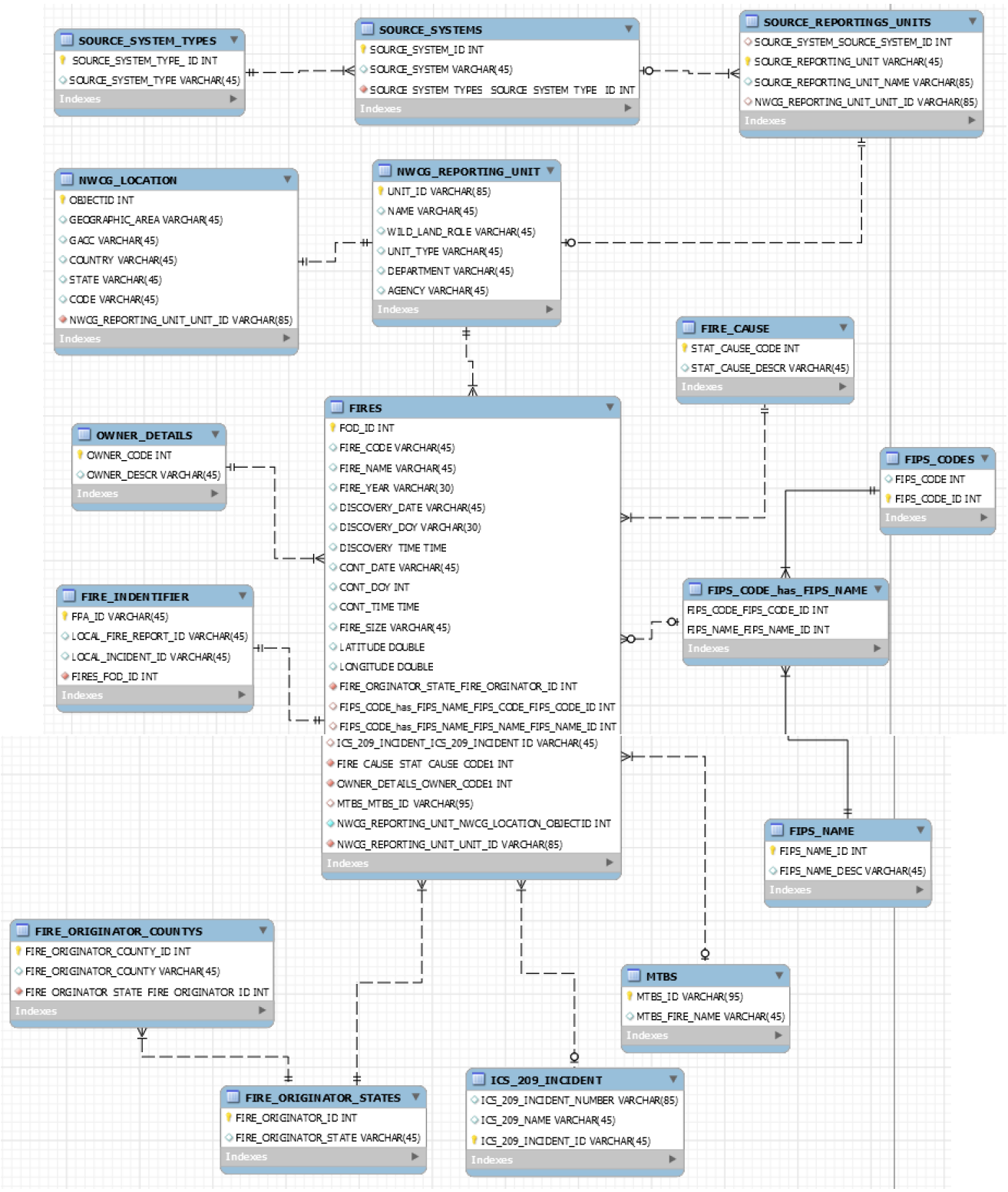
1. Every table has its own primary key. The primary key was generated for tables like SOURCE_SYSTEM_TYPES, FIRE_ORIGINATOR_COUNTY etc., where there were no unique key columns. It is ensured that the values in the primary key column are unique and are not null.

2. It is ensured that there are no multi part or multi valued fields in any of the tables.
   Example: In table SOURCE_REPORTING_UNITS the attribute SOURCE_REPORTING_UNIT_NAME had multi valued fields which were corrected
   Due to these reasons all tables are said to be in first Normal Form

3. It is ensured that every column other than the primary key is independent on the both the primary key columns when composite primary key is used. Thus, there is no partial dependency. Due to this reason all tables are said to be in second Normal Form

4. It is ensured that none of the tables have transitive dependency. That is, all the non key columns were purely made to depend on the primary key field, and they can cannot be derived or calculated using any other field. Example: The FIRE_SIZE_CLASS field is eliminated as it can be derived from the FIRE_SIZE field which is non-primary key column
   Hence all tables are in third Normal form

Entity-Relationship Diagram



Physical Database

Assumptions/Notes About Data Set

Include any assumptions made about data such as empty fields, sparse data, bad data, etc.

FPA_FOD_20170508 Database:

1. Though this database contains many tables, only the Fires and NWCG_UnitIDActive_20170109 tables was taken into consideration because,

- Tables such as Elementary Geometries, idx_Fires_Shape, KNN and Spatial Index were virtual tables and hence did not contain any information
- Tables such as geometry_columns, geometry_columns_auth, geometry_columns_field_infos, geometry_columns_statistics, geometry_columns_time, conatin very less information ie only one row is present per table which is not revelant
- Tables such as idx_Fires_shape_Node, idx_Fires_shape_parent, idx_Fires_shape_Rowid contain details about parent node columns which was also present in the NWCG table. However the parent column in the NWCG table itself is completely NULL which indicates that these data here will not be useful
- Tables such as spatial_ref_sys, spatial_ref_sys_aux, spatialite_history contains several rows with enough information. However, these tables are not related with Fires table and NWCG table which are our main tables. And so, we are not considering these tables
- Tables like views_geometry_columns, views_geometry_columns_auth, views_geometry_columns_feild_infos, views_geometry_columns_statistics contain no data. Therefore, we are not considering these tables

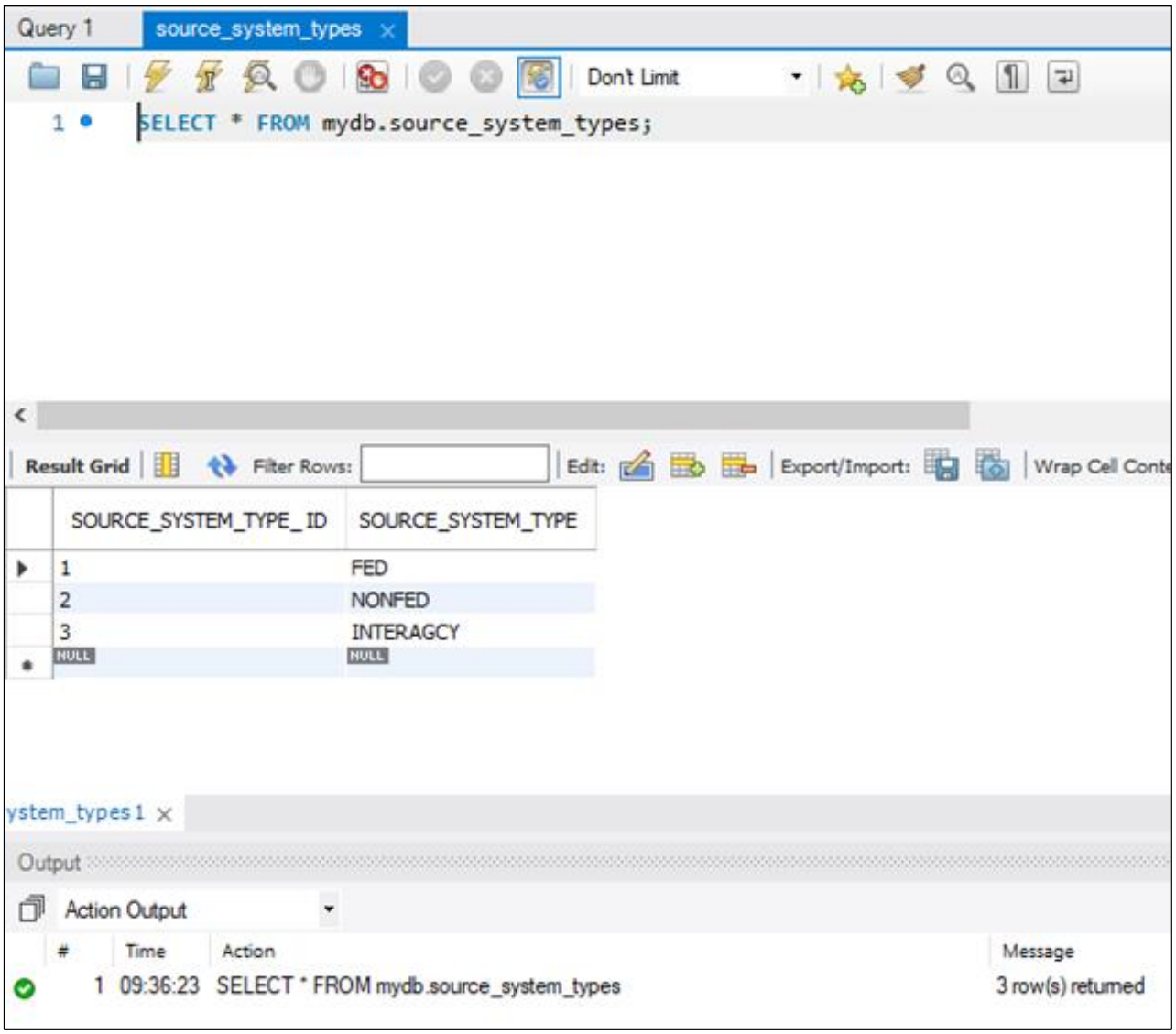
2. Since the data dictionary indicates that NWCG_UnitIDActive_20170109 is used as the source for populating the following fields in the Fires table: NWCG_REPORTING_AGENCY, NWCG_REPORTING_UNIT_ID, and NWCG_REPORTING_UNIT_NAME, we are tempted to believe that the values from these columns present in the FIRES table would be a subset of NWCG_UnitIDActive_20170109 table. But however, there are values in NWCG_REPORTING_UNIT_NAME tables which were not present in the main table

It is assumed that it could have happened because the NWCG_UnitIDActive_20170109 table contains all NWCG identifiers for agency units that were active/valid as of 9 January 2017. So, the identifiers for agency units which was present in the fires table could have been present before 2017

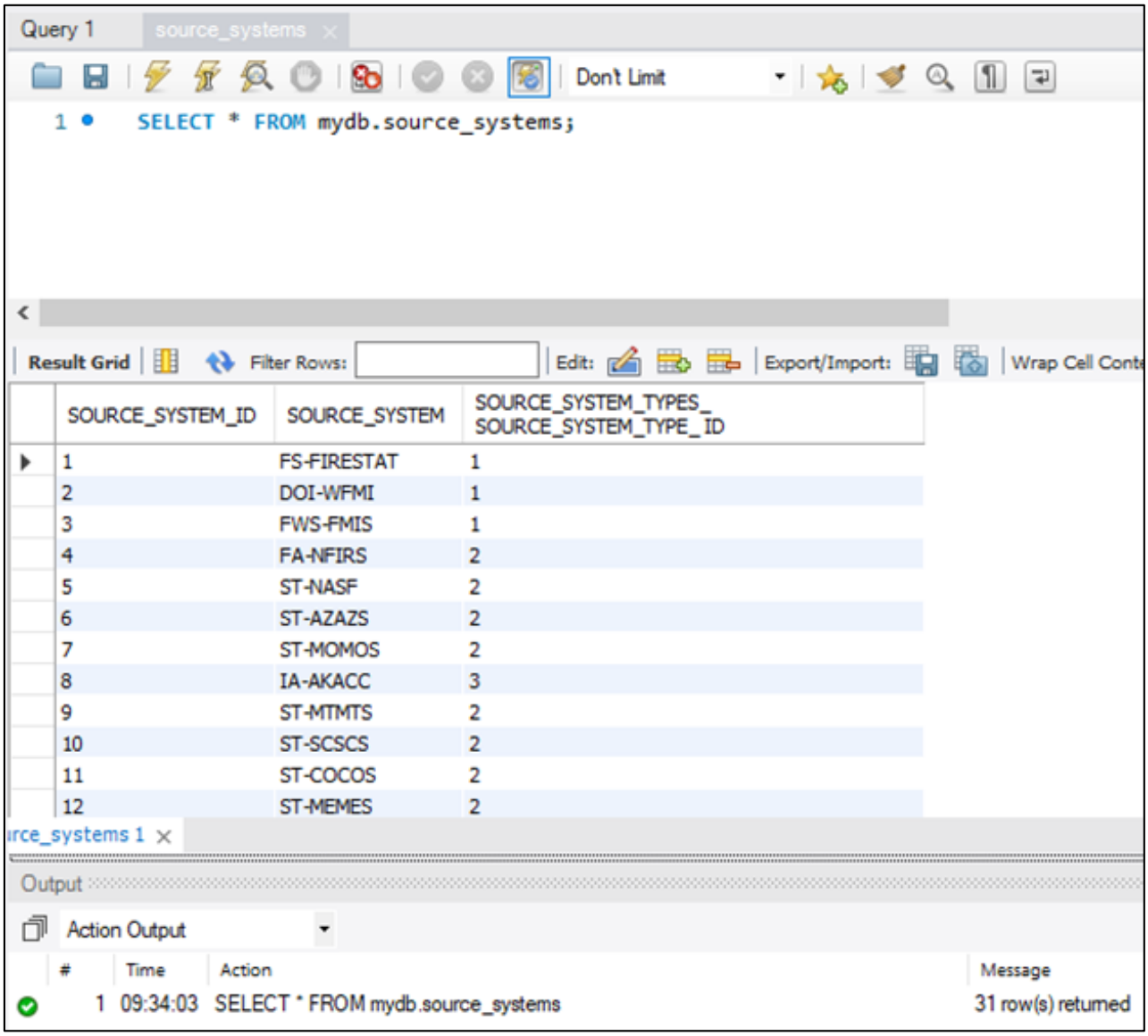
3. The FOD_ID = Global unique identifier, though contained unique values to be used as a primary key , it skipped a value in between. (Which was 74)

Screen shot of Physical Database objects

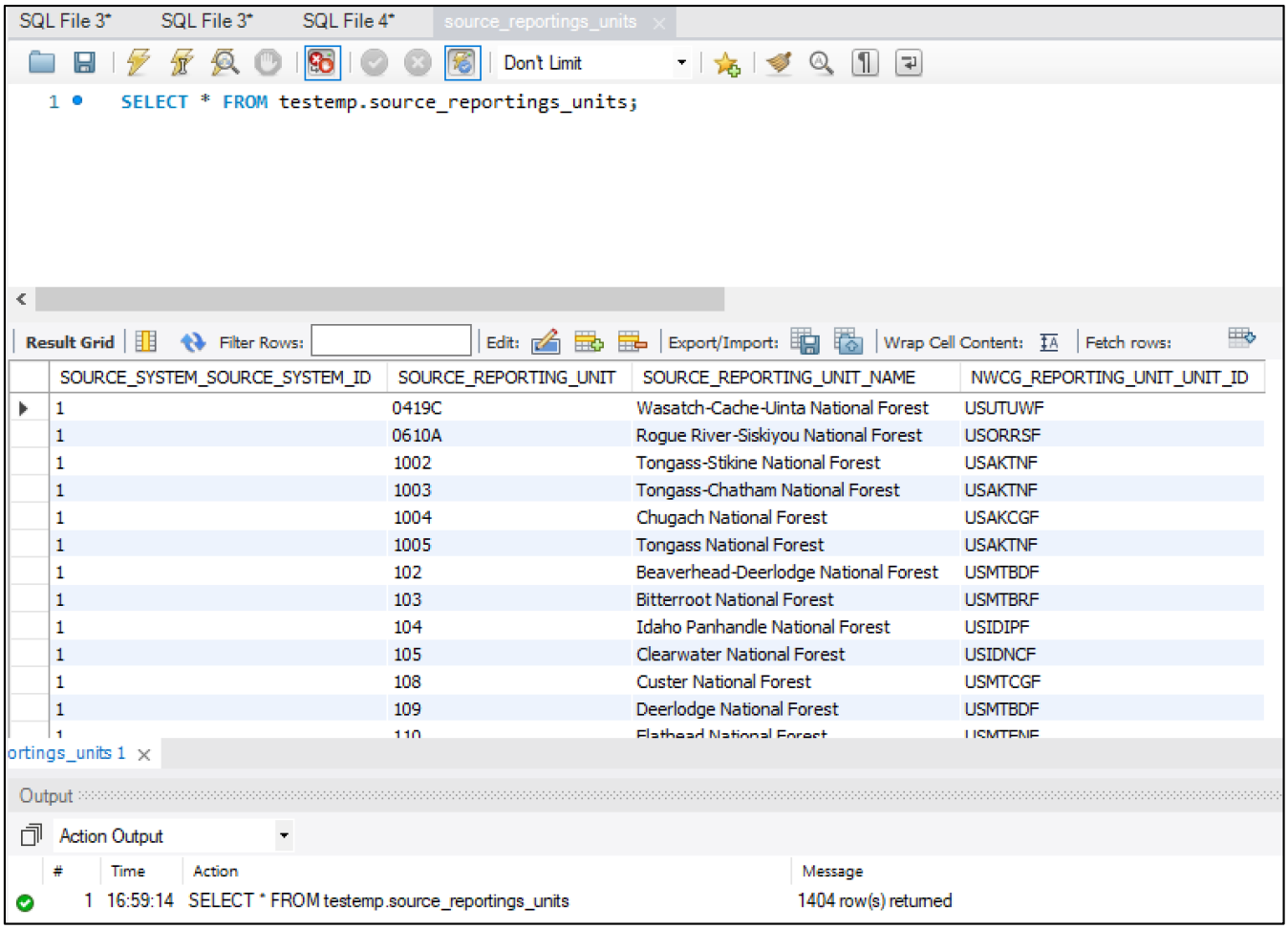
1. SOURCE_SYSTEM_TYPES



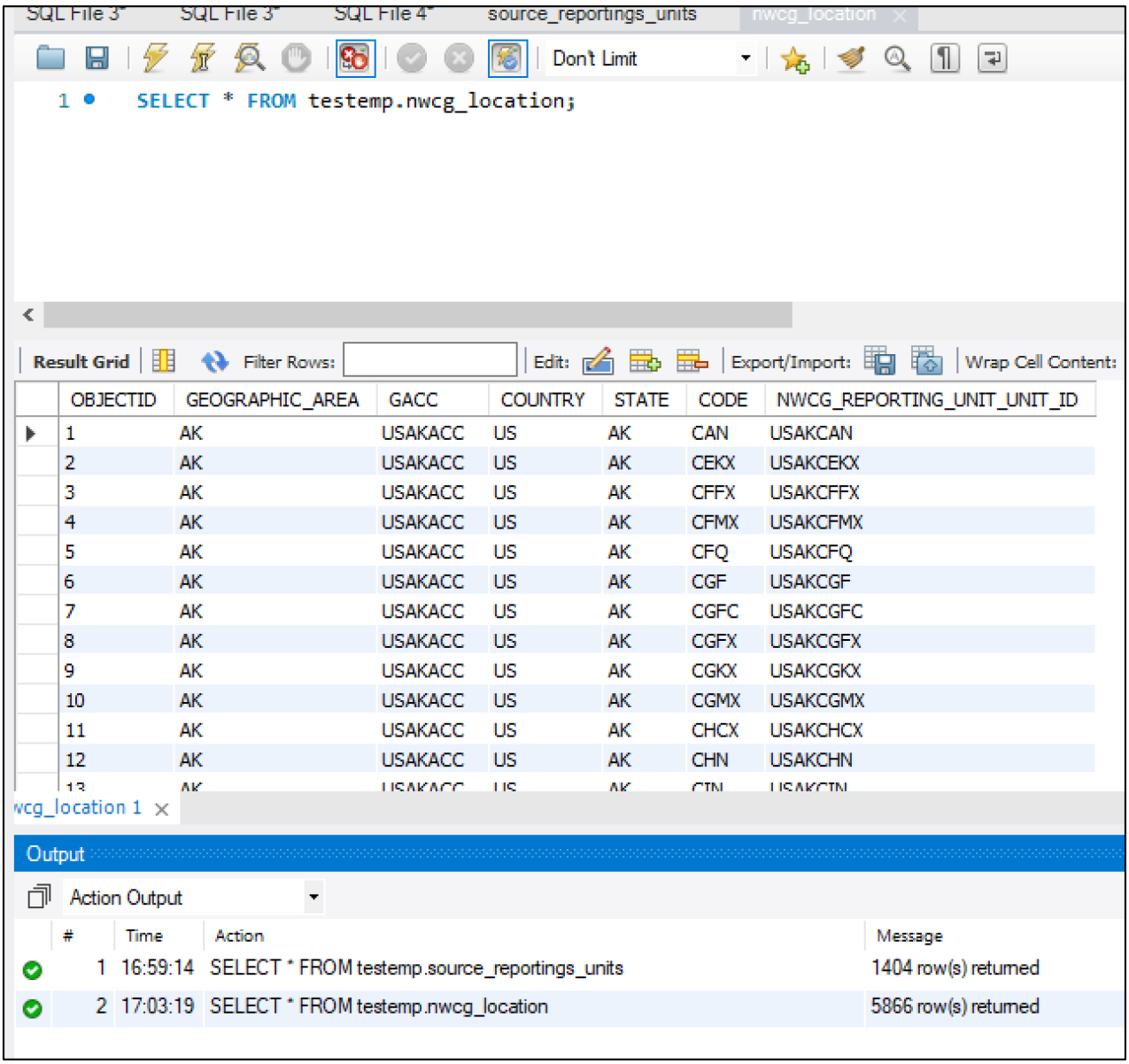
2. SOURCE_SYSTEMS



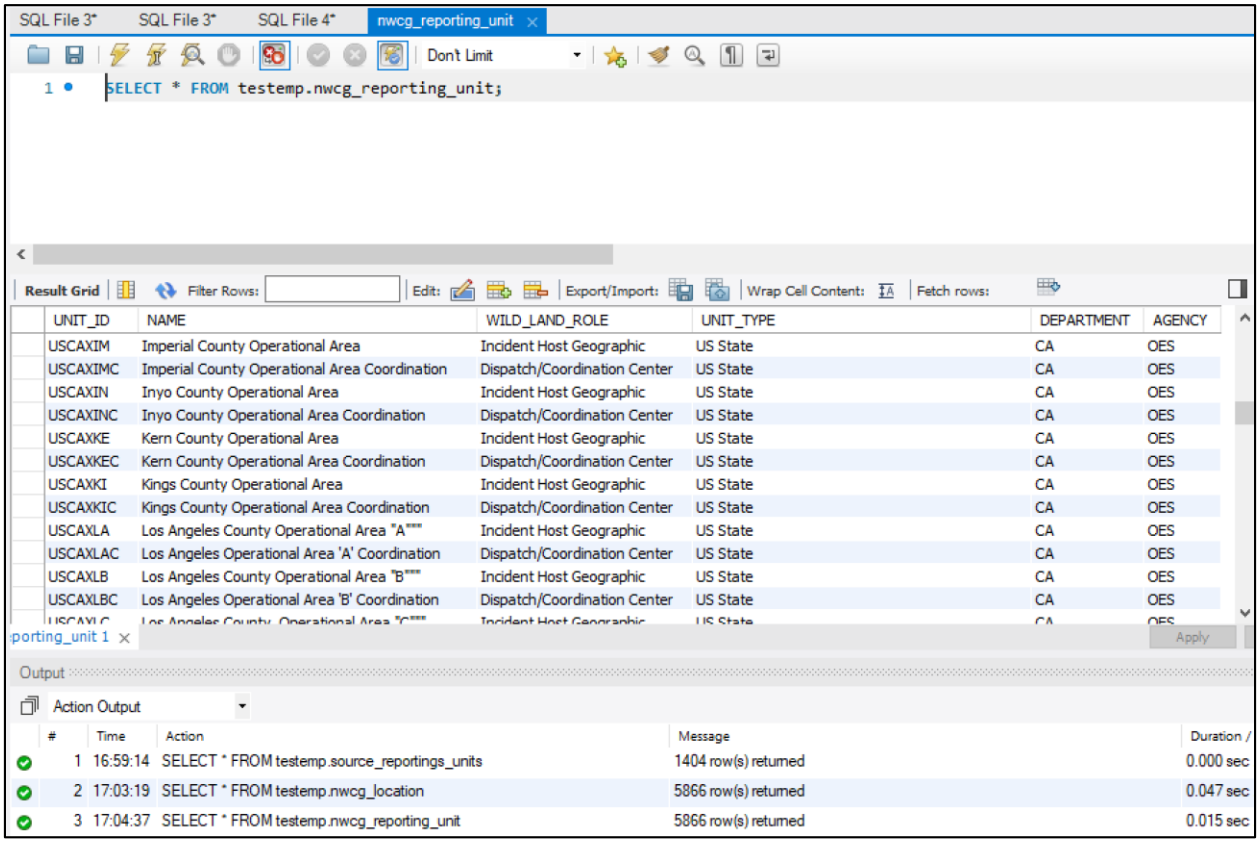
3. SOURCE_REPORTING_UNITS



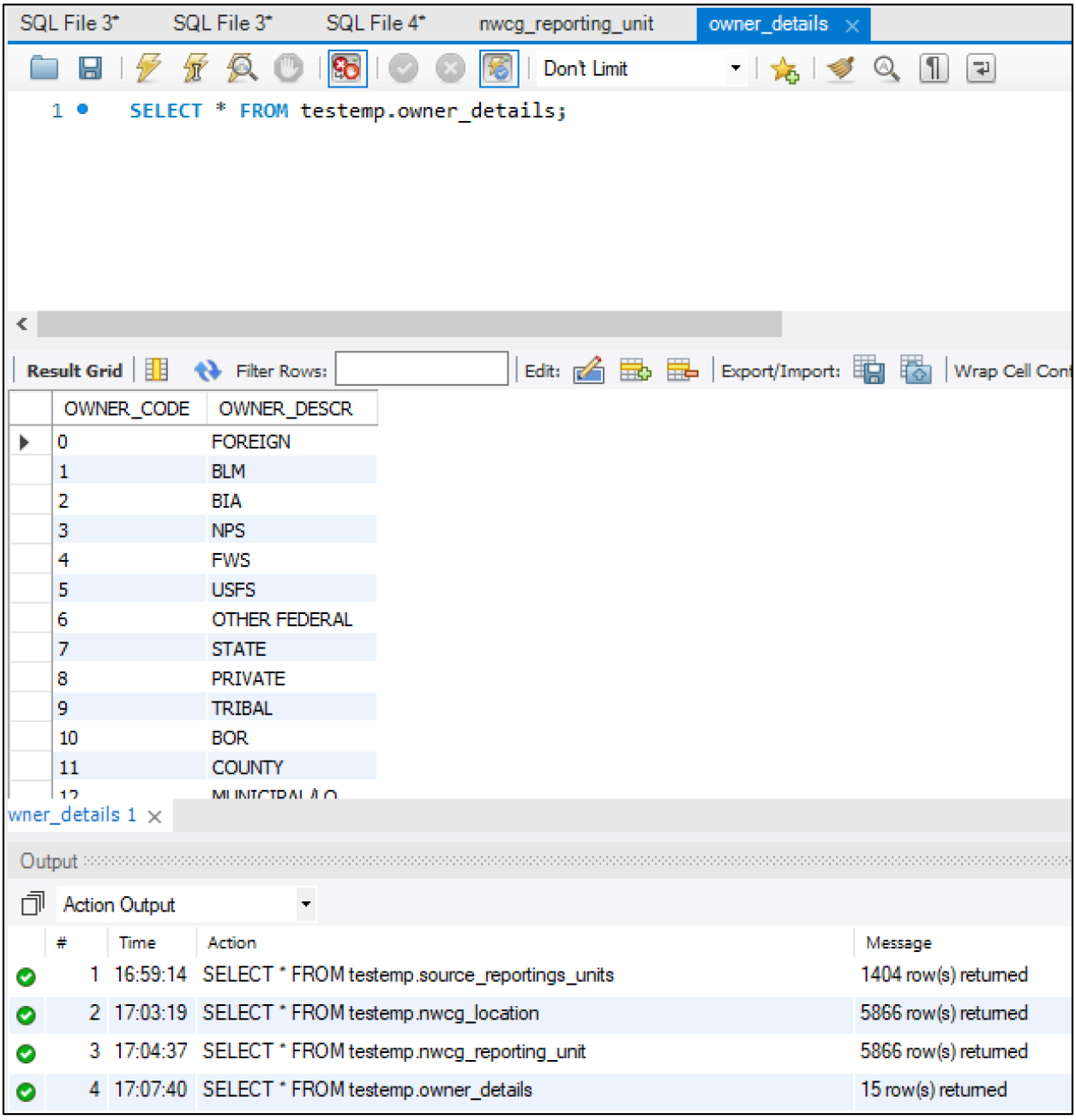
4. NWGC_LOCATION



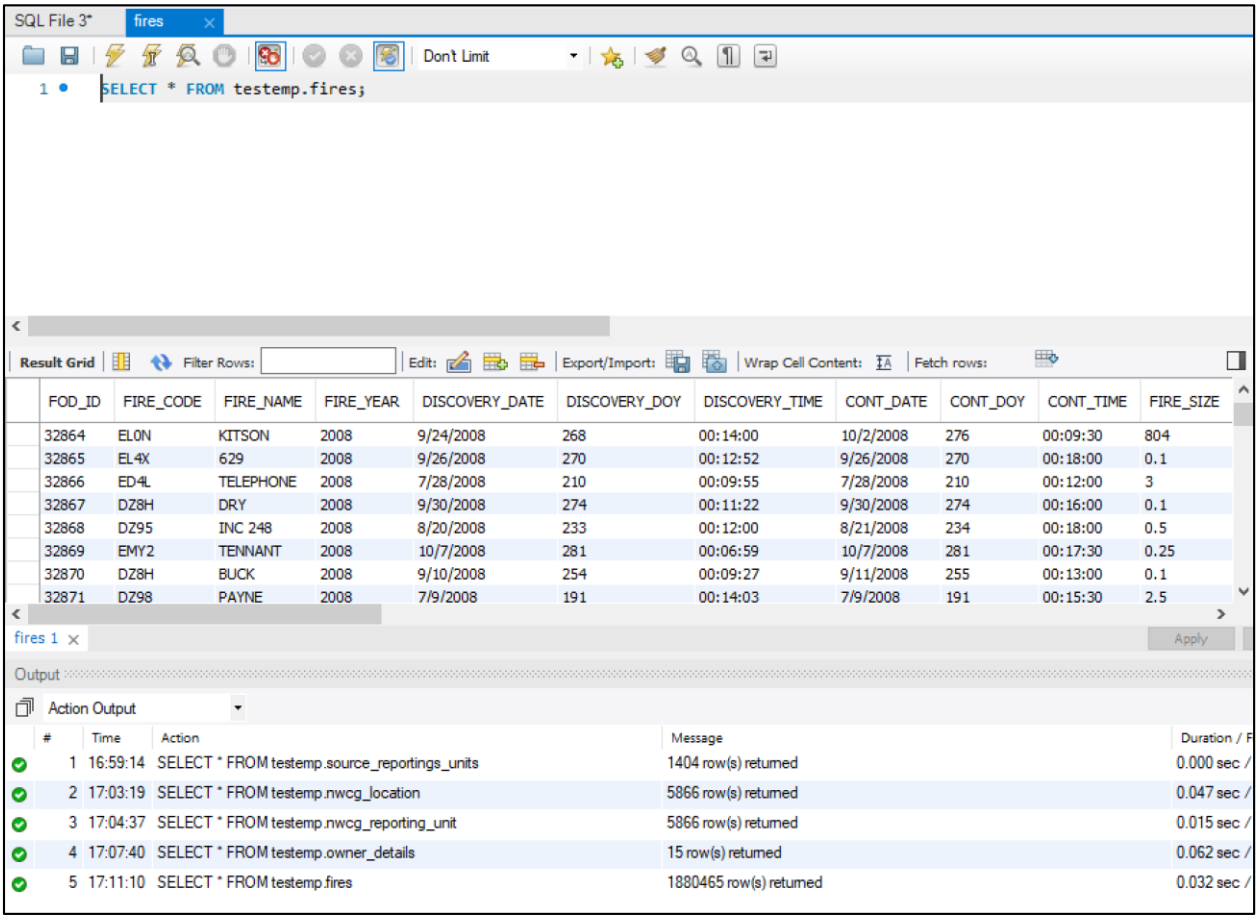
5. NWGC_REPORTING_UNIT



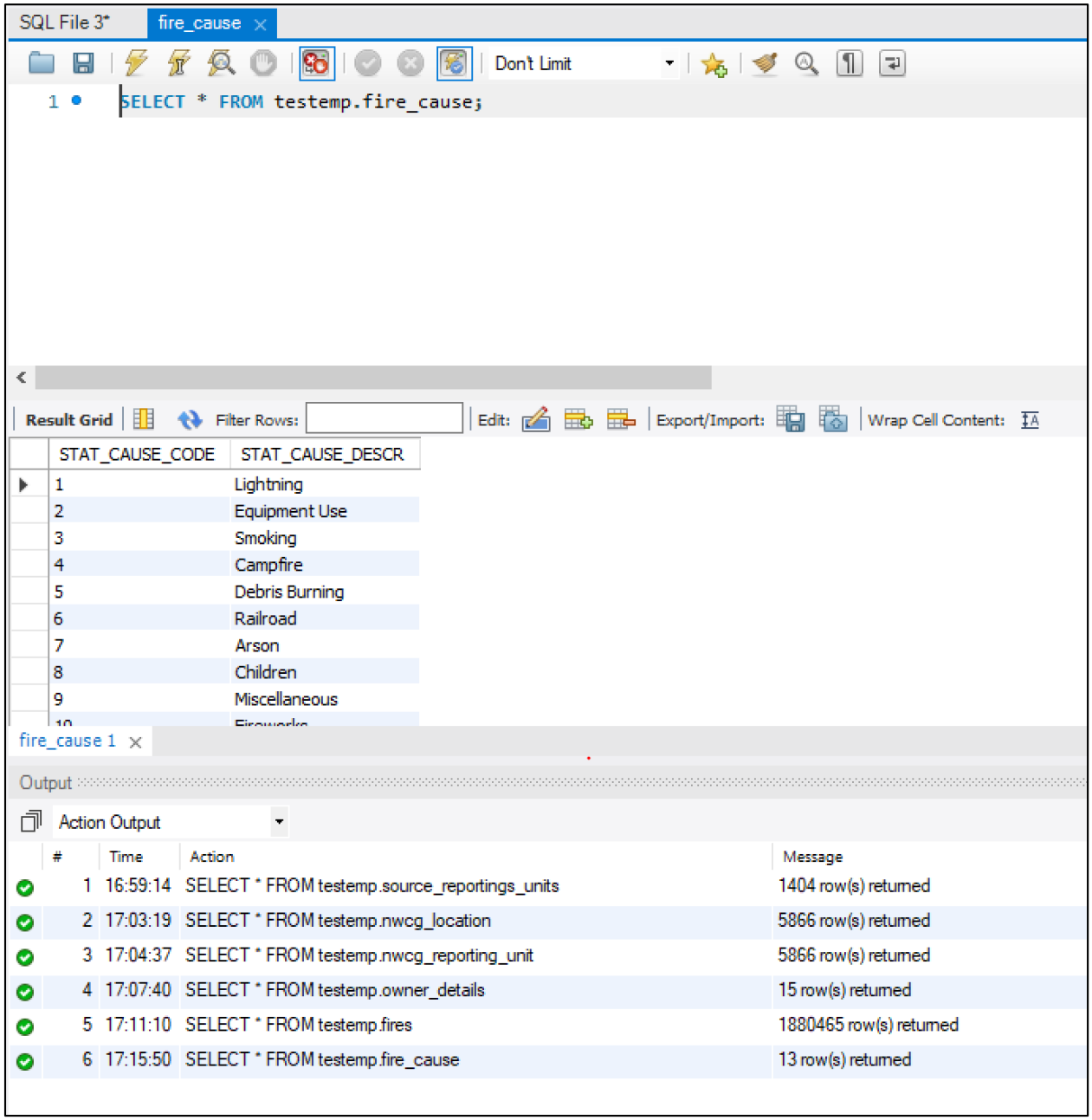
6. OWNER_DETAILS



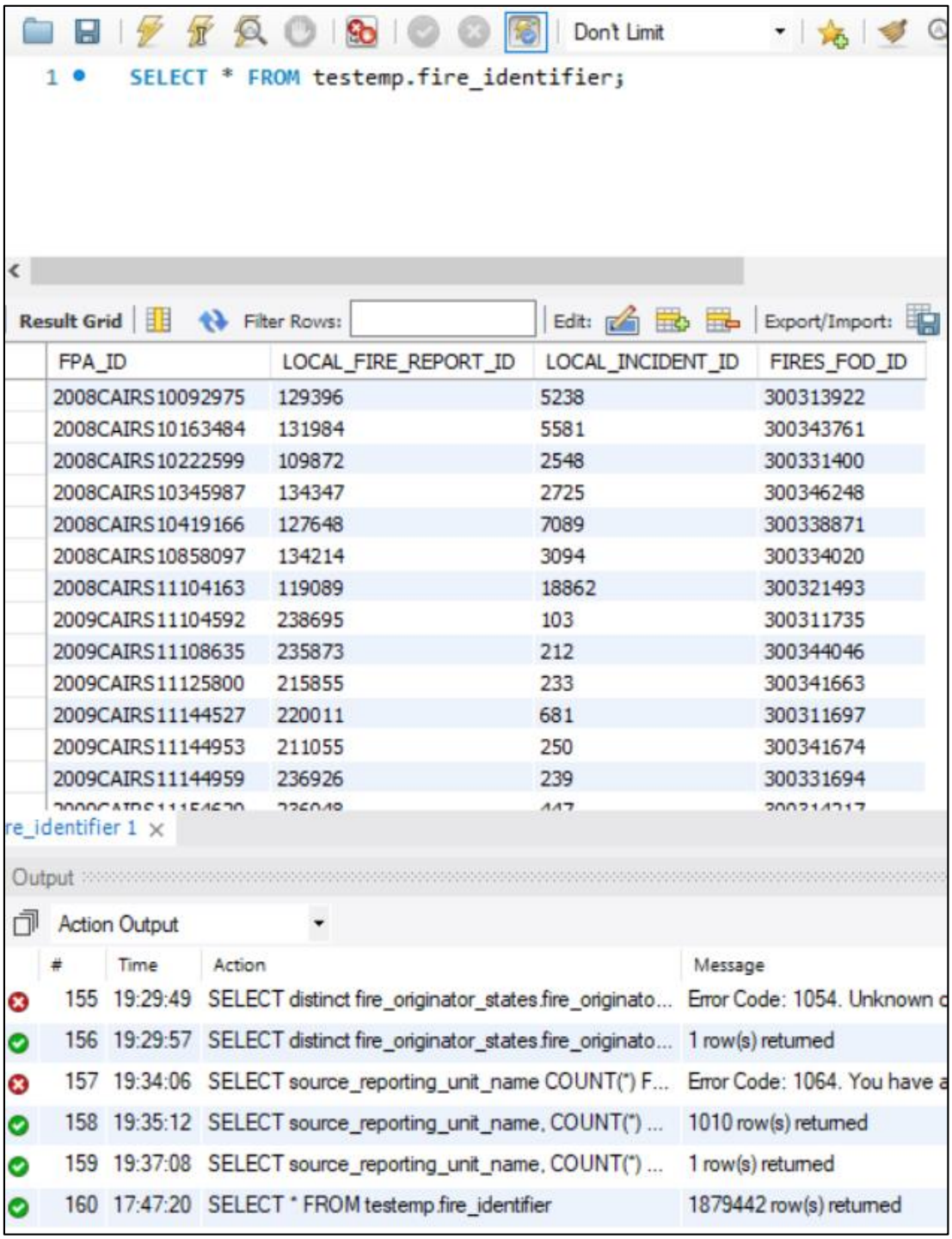
7. FIRES


SQL File 3*
fires
Don't Limit
1 SELECT * FROM testemp.fires;
Result Grid
Filter Rows:
Edit:
Export/Import:
Wrap Cell Content:
Fetch rows:
FOD_ID FIRE_CODE FIRE_NAME FIRE_YEAR DISCOVERY_DATE DISCOVERY_DOY DISCOVERY_TIME CONT_DATE CONT_DOY CONT_TIME FIRE_SIZE
32864 EL0N KITSON 2008 9/24/2008 268 00:14:00 10/2/2008 276 00:09:30 804
32865 EL4X 629 2008 9/26/2008 270 00:12:52 9/26/2008 270 00:18:00 0.1
32866 ED4L TELEPHONE 2008 7/28/2008 210 00:09:55 7/28/2008 210 00:12:00 3
32867 DZ8H DRY 2008 9/30/2008 274 00:11:22 9/30/2008 274 00:16:00 0.1
32868 DZ95 INC 248 2008 8/20/2008 233 00:12:00 8/21/2008 234 00:18:00 0.5
32869 EMY2 TENNANT 2008 10/7/2008 281 00:06:59 10/7/2008 281 00:17:30 0.25
32870 DZ8H BUCK 2008 9/10/2008 254 00:09:27 9/11/2008 255 00:13:00 0.1
32871 DZ98 PAYNE 2008 7/9/2008 191 00:14:03 7/9/2008 191 00:15:30 2.5
fires 1
Apply
Output
Action Output
Time Action Message Duration / F
1 16:59:14 SELECT * FROM testemp.source_reportings_units 1404 row(s) returned 0.000 sec /
2 17:03:19 SELECT * FROM testemp.nwcg_location 5866 row(s) returned 0.047 sec /
3 17:04:37 SELECT * FROM testemp.nwcg_reporting_unit 5866 row(s) returned 0.015 sec /
4 17:07:40 SELECT * FROM testemp.owner_details 15 row(s) returned 0.062 sec /
5 17:11:10 SELECT * FROM testemp.fires 1880465 row(s) returned 0.032 sec /

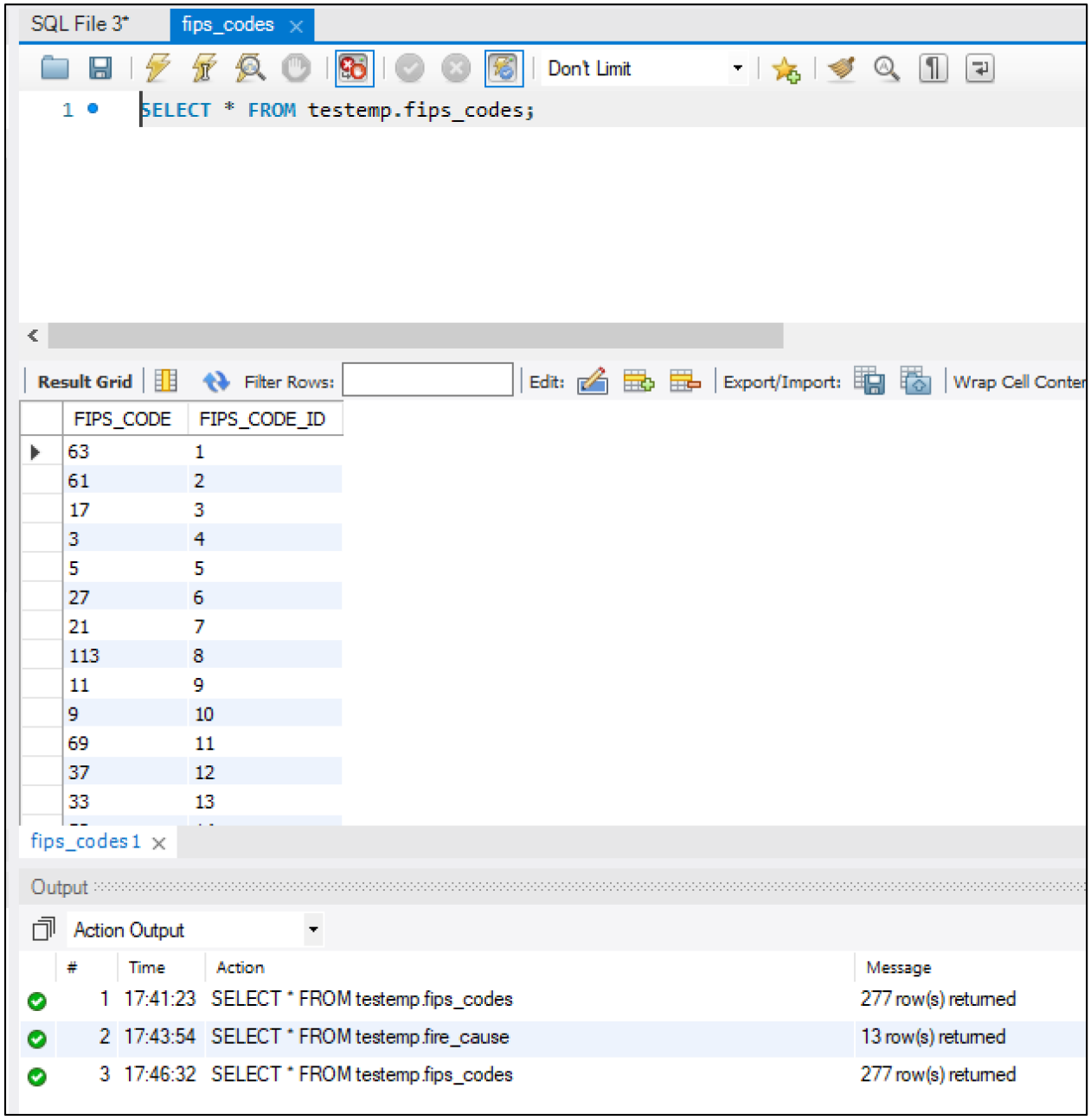
8. FIRE_CAUSE


SQL File 3*
fire_cause
Don't Limit
1 SELECT * FROM testemp.fire_cause;
Result Grid
Filter Rows:
Edit:
Export/Import:
Wrap Cell Content:
STAT_CAUSE_CODE STAT_CAUSE_DESCR
1 Lightning
2 Equipment Use
3 Smoking
4 Campfire
5 Debris Burning
6 Railroad
7 Arson
8 Children
9 Miscellaneous
fire_cause 1
Output
Action Output
Time Action Message
1 16:59:14 SELECT * FROM testemp.source_reportings_units 1404 row(s) returned
2 17:03:19 SELECT * FROM testemp.nwcg_location 5866 row(s) returned
3 17:04:37 SELECT * FROM testemp.nwcg_reporting_unit 5866 row(s) returned
4 17:07:40 SELECT * FROM testemp.owner_details 15 row(s) returned
5 17:11:10 SELECT * FROM testemp.fires 1880465 row(s) returned
6 17:15:50 SELECT * FROM testemp.fire_cause 13 row(s) returned

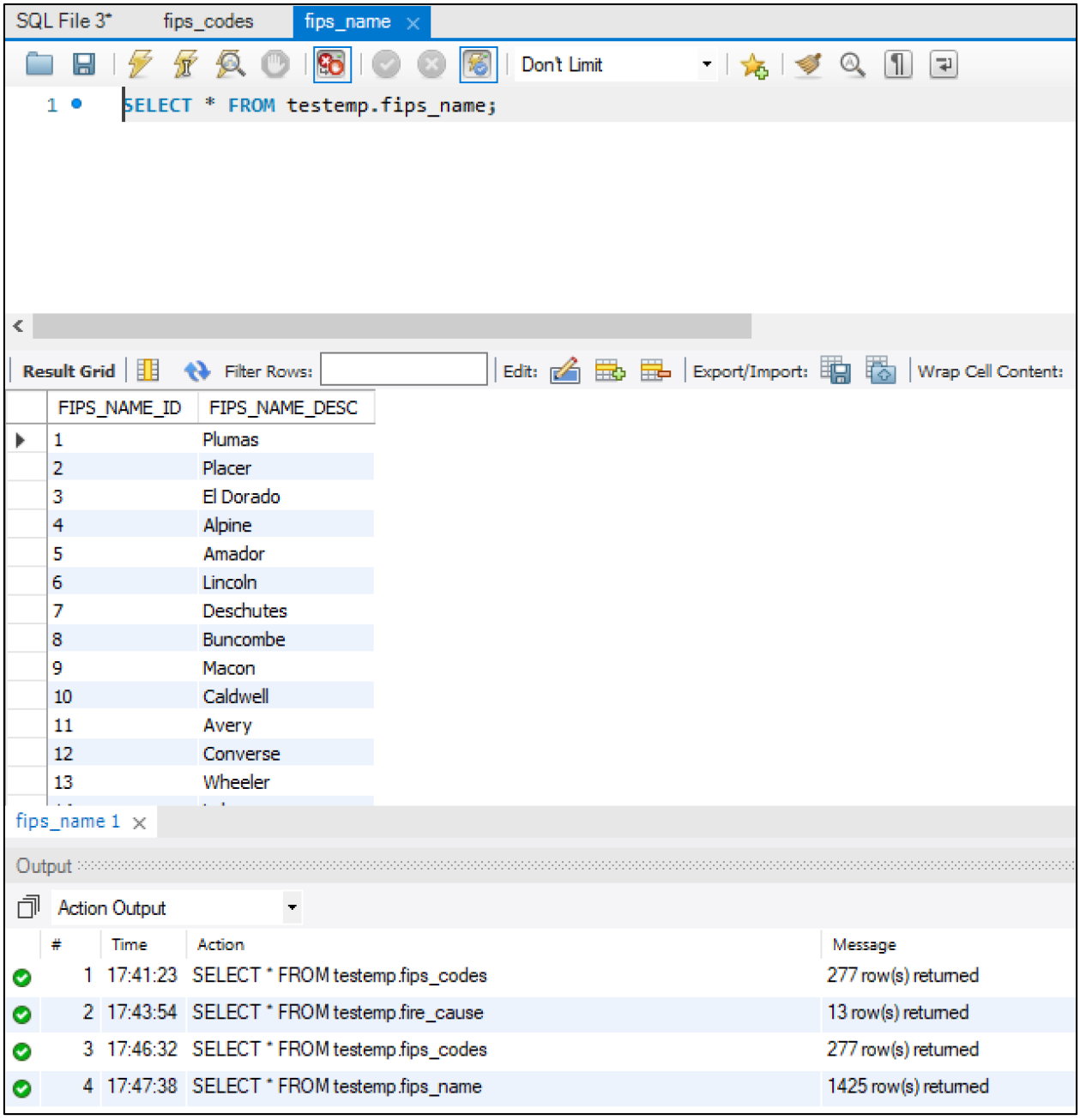
9. FIRE_IDENTIFIER



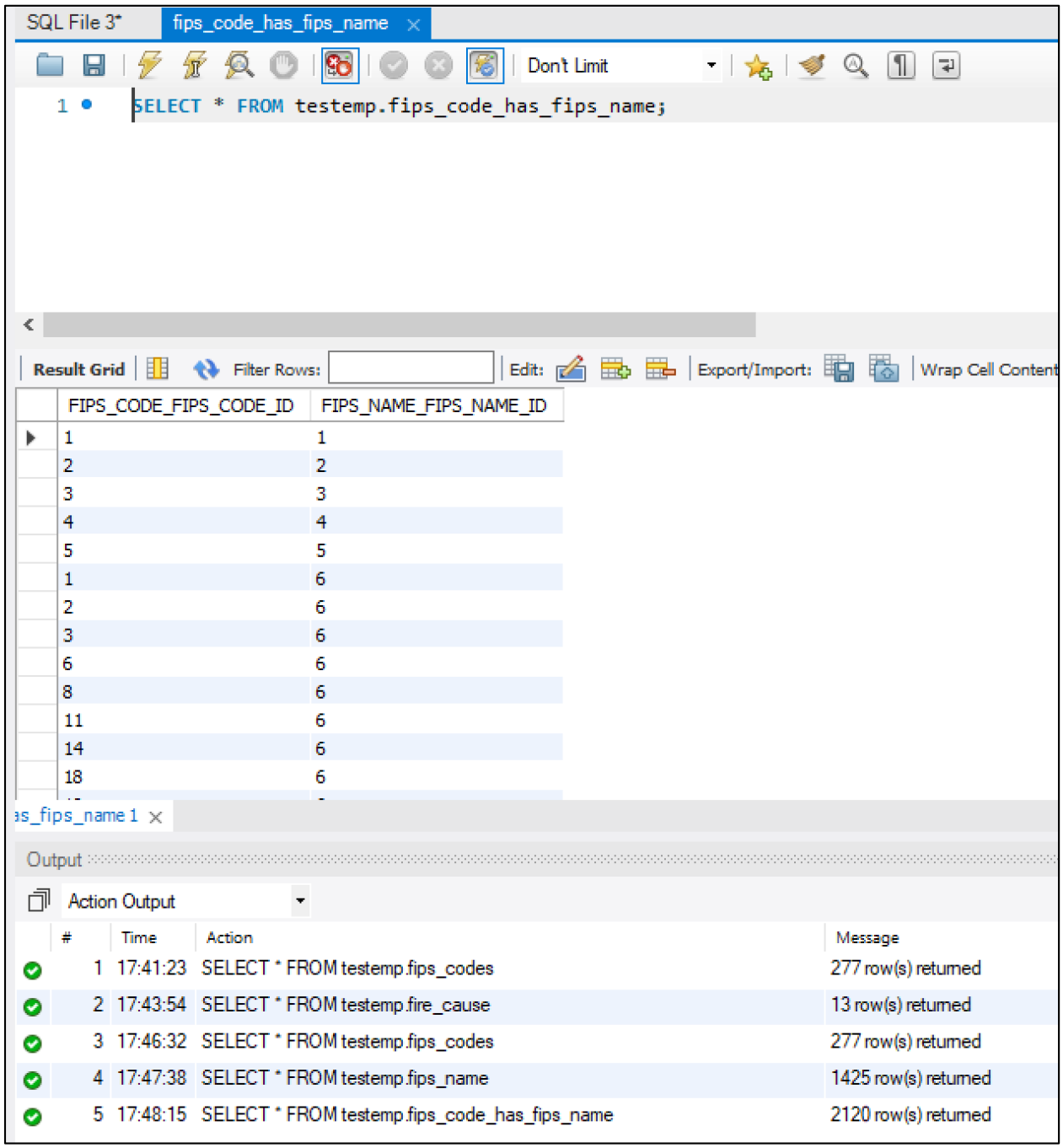
10. FIPS_CODES



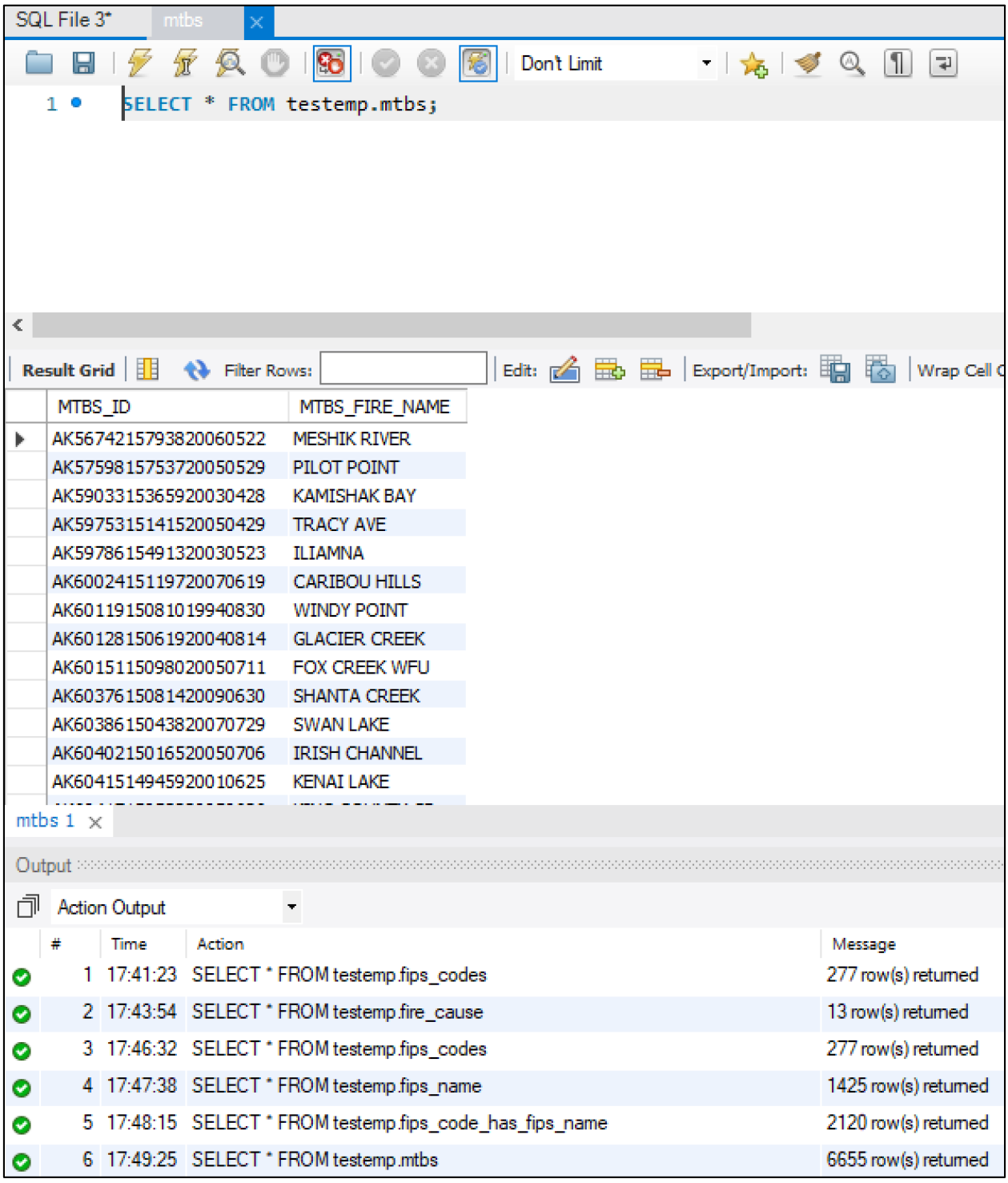
11. FIPS_NAME


SQL File 3*
fips_codes
fips_name
Don't Limit
1 SELECT * FROM testemp.fips_name;
Result Grid
Filter Rows:
Edit:
Export/Import:
Wrap Cell Content:
FIPS_NAME_ID FIPS_NAME_DESC
1 Plumas
2 Placer
3 El Dorado
4 Alpine
5 Amador
6 Lincoln
7 Deschutes
8 Buncombe
9 Macon
10 Caldwell
11 Avery
12 Converse
13 Wheeler
fips_name 1
Output
Action Output
Time Action Message
1 17:41:23 SELECT * FROM testemp.fips_codes 277 row(s) returned
2 17:43:54 SELECT * FROM testemp.fire_cause 13 row(s) returned
3 17:46:32 SELECT * FROM testemp.fips_codes 277 row(s) returned
4 17:47:38 SELECT * FROM testemp.fips_name 1425 row(s) returned

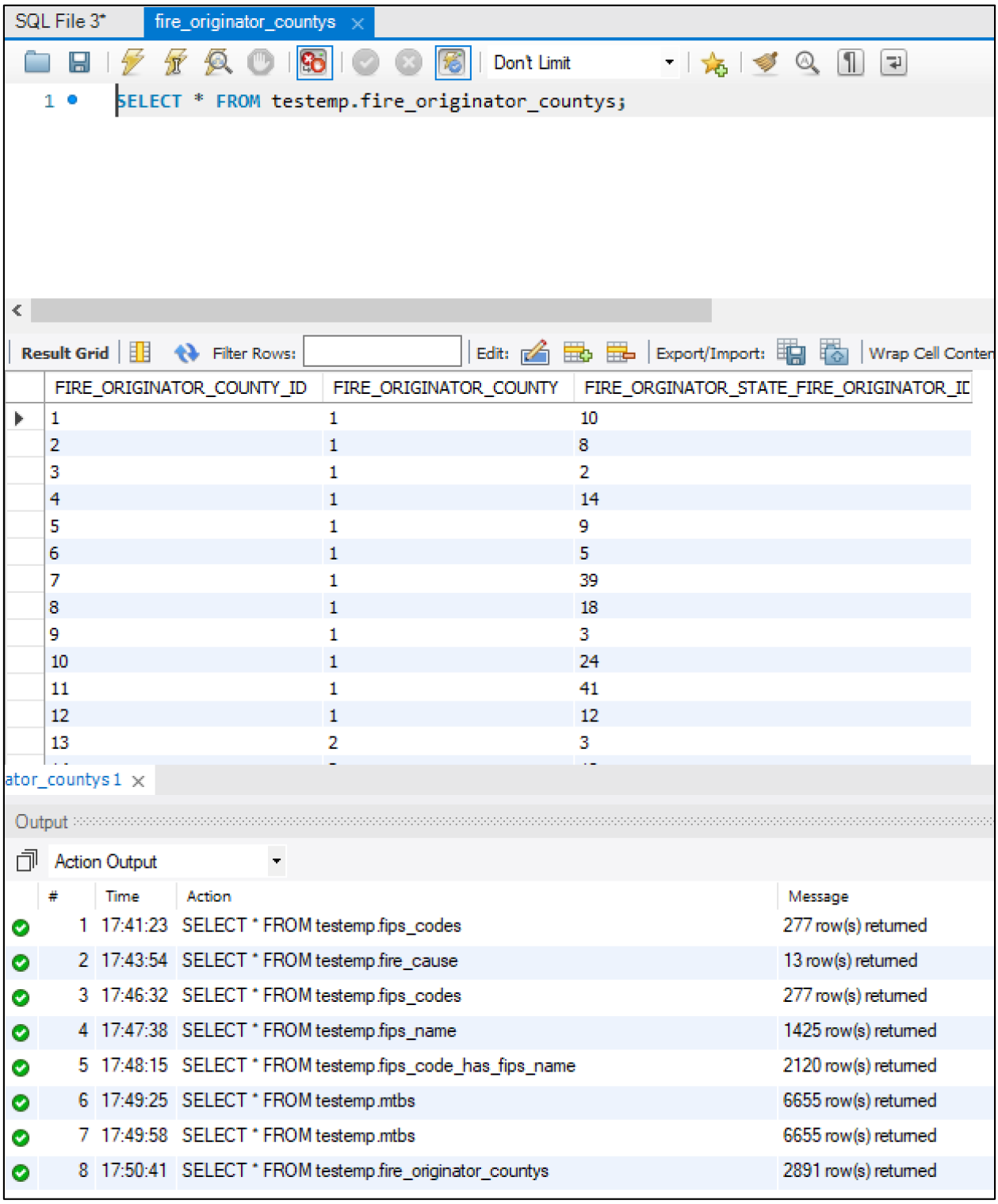
12. FIPS_CODE_has_FIPS_NAME



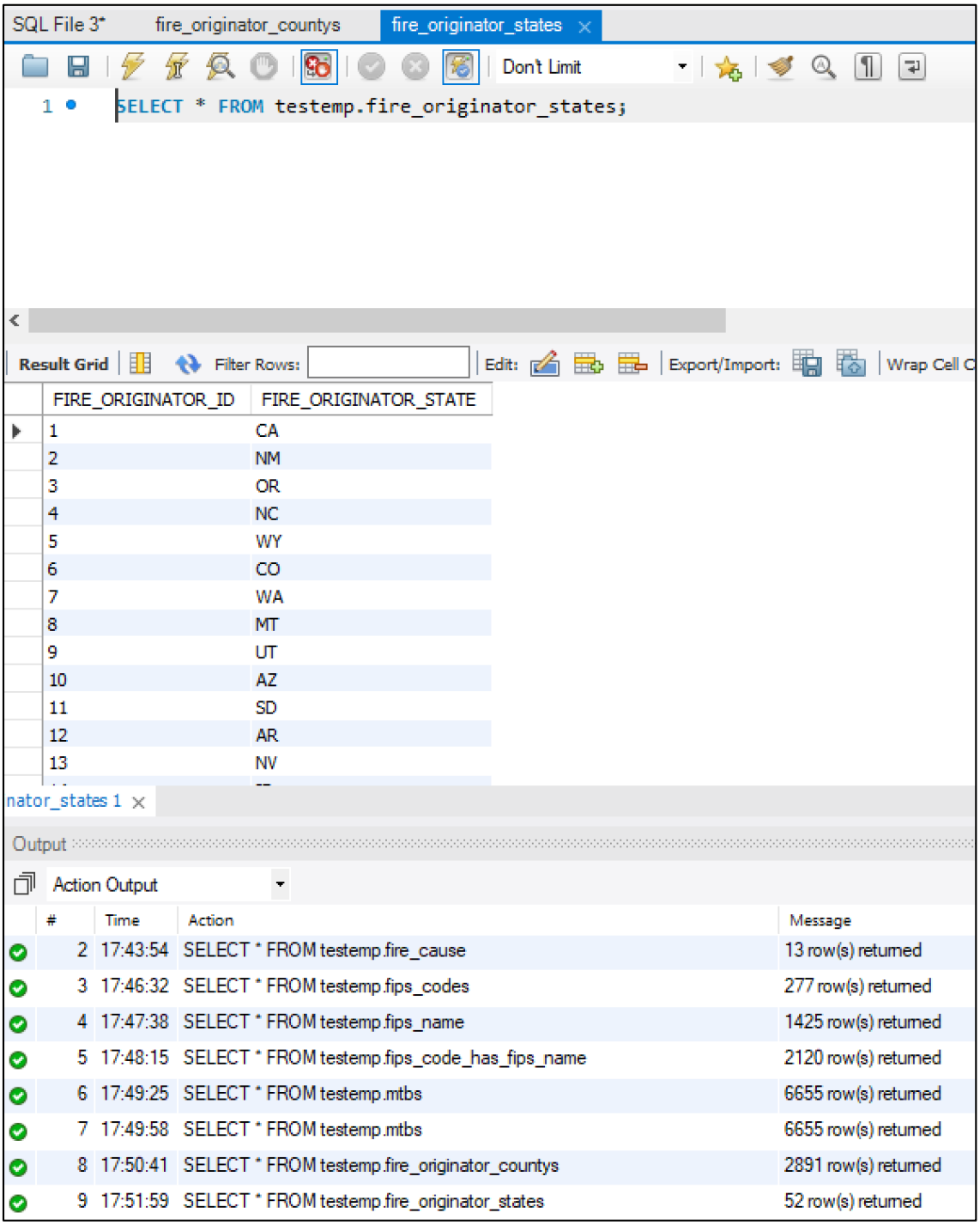
13. MTBS


SQL File 3*
mtbs
Don't Limit
1 SELECT * FROM testemp.mtbs;
Result Grid
Filter Rows:
Edit:
Export/Import:
Wrap Cell C
MTBS_ID
MTBS_FIRE_NAME
AK567421579382006052 2
MESHIK RIVER
AK575981575372005052 9
PILOT POINT
AK590331536592003042 8
KAMISHAK BAY
AK597531514152005042 9
TRACY AVE
AK597861549132003052 3
ILIAMNA
AK600241511972007061 9
CARIBOU HILLS
AK601191508101994083 0
WINDY POINT
AK601281506192004081 4
GLACIER CREEK
AK601511509802005071 1
FOX CREEK WFU
AK603761508142009063 0
SHANTA CREEK
AK603861504382007072 9
SWAN LAKE
AK604021501652005070 6
IRISH CHANNEL
AK604151494592001062 5
KENAI LAKE
mtbs 1
Output
Action Output
Time Action Message
1 17:41:23 SELECT * FROM testemp.fips_codes 277 row(s) returned
2 17:43:54 SELECT * FROM testemp.fire_cause 13 row(s) returned
3 17:46:32 SELECT * FROM testemp.fips_codes 277 row(s) returned
4 17:47:38 SELECT * FROM testemp.fips_name 1425 row(s) returned
5 17:48:15 SELECT * FROM testemp.fips_code_has_fips_name 2120 row(s) returned
6 17:49:25 SELECT * FROM testemp.mtbs 6655 row(s) returned

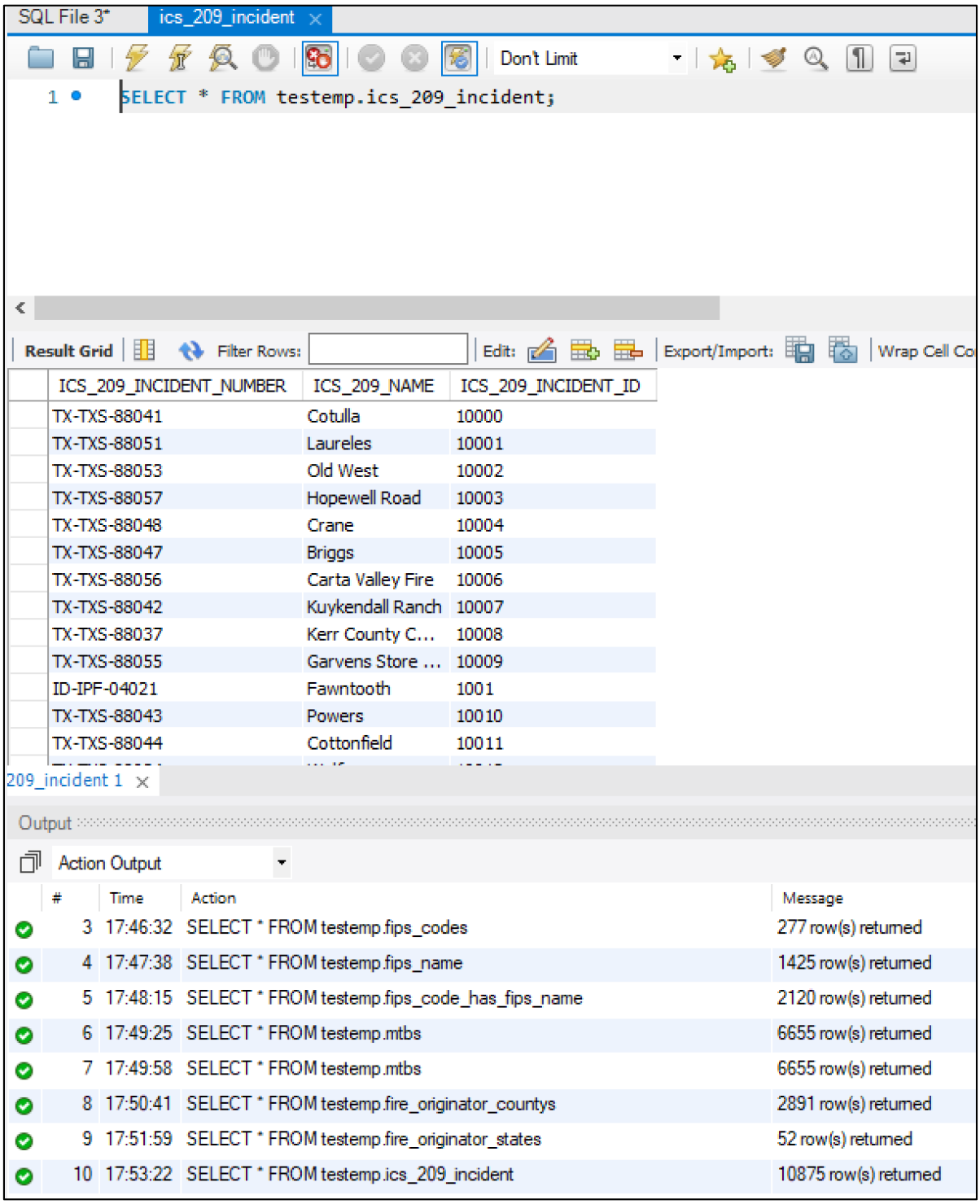
14. FIRE_ORIGINATOR_COUNTYS



15. FIRE_ORIGINATOR_STATES



16. ICS_209_INCIDENT



Data in the Database

SNO.	Table Name	Primary Key	Foreign Key	# of Rows in Table
1	SOURCE_SYSTEM_TYPES	SOURCE_SYSTEM_TYPE_ ID	-	3
2	SOURCE_SYSTEMS	SOURCE_SYSTEM_ID	SOURCE_SYSTEM_TYPE_ ID	31
3	SOURCE_REPORTING_UNITS	SOURCE_REPORTING_UNIT	SOURCE_SYSTEM_ID, UNIT_ID(NWGC_REPORTING_UNIT)	1404
4	NWGC_LOCATION	OBJECTID	UNIT_ID(NWGC_REPORTING_UNIT)-	5866
5	NWGC_REPORTING_UNIT	UNIT_ID		5866
6	OWNER_DETAILS	OWNER_CODE	-	15
7	FIRES	FOD_ID	FIRE_ORIGINATOR_ID, FIPS_CODE_ID, FIPS_NAME_ID, ICS_209_INCIDENT_ID, STAT_CAUSE_CODE, OWNER_CODE, MTBS_ID, OBJECTID (NWGC_LOCATION) UNIT_ID (NWGC_REPORTING_UNIT)	1880465

8	FIRE_CAUSE	STAT_CAUSE_CODE	-	13
9	FIRE_IDENTIFIER	FPA_ID	FOD_ID	1879442
10	FIPS_CODE	FIPS_CODE_ID	-	277
11	FIPS_NAME	FIPS_NAME_ID	-	1425
12	FIPS_CODE_has_FIPS_NAME	FIPS_CODE_ID FIPS_NAME_ID		2120
13	MTBS	MTBS_ID	-	6655
14	FIRE_ORIGINATOR_COUNTYS	FIRE_ORIGINATOR_COUNTY_ID	FIRE_ORIGINATOR_ID (From FIRE_ORIGINATOR_STATE)	2891
15	FIRE_ORIGINATOR_STATE	FIRE_ORIGINATOR_ID	-	52
16	ICS_209_INCIDENT	ICS_209_INCIDENT_ID	-	10875

SQL Queries

Query 1

Question:

1. 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 state has the least chance to win a share of the fund?

Assumption:

It is assumed that the company is going to fund the states to remove debris in order to prevent fires in the future. Hence the company is looking for the states that have the highest amount of fires due to debris. The question is about which state is having the least chance to receive the funding, i.e, the state with the least number of fires caused by debris will have the least chance to win the fund. So, we are considering the states that had least fires caused by debris. Since many states have the same number of fires due to debris, we are randomly picking one as our answer.

Translation:

SELECT FIRE_ORIGINATOR_STATE from fires table
SUM of fire_size as firesize then
INNER JOIN NWCG_Reporting_Unit ON nwcg_reporting_unit.unit_id and
INNER JOIN source_reportings_units ON nwcg_reporting_unit.unit_id and
INNER JOIN fire_originator_states ON FIRE_ORGINATOR_STATE_FIRE_ORGINATOR_ID and
INNER JOIN fire_cause on fire_cause.stat_cause_code and
WHERE stat_cause_code= '5' then
GROUP BY source_reporting_unit_name and then
ORDER BY firesize ASC and limit 1

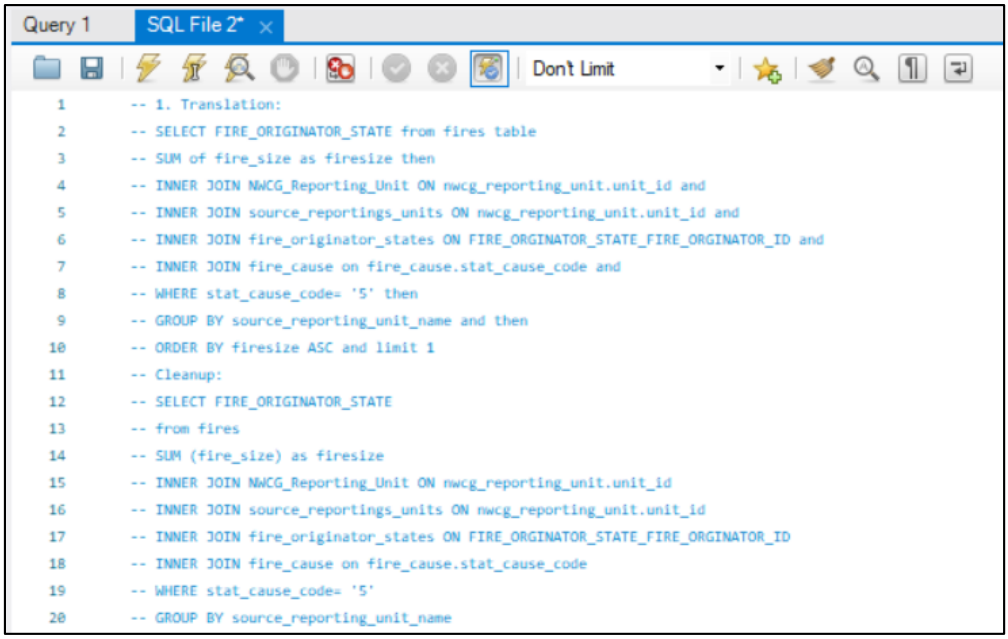
Cleanup:

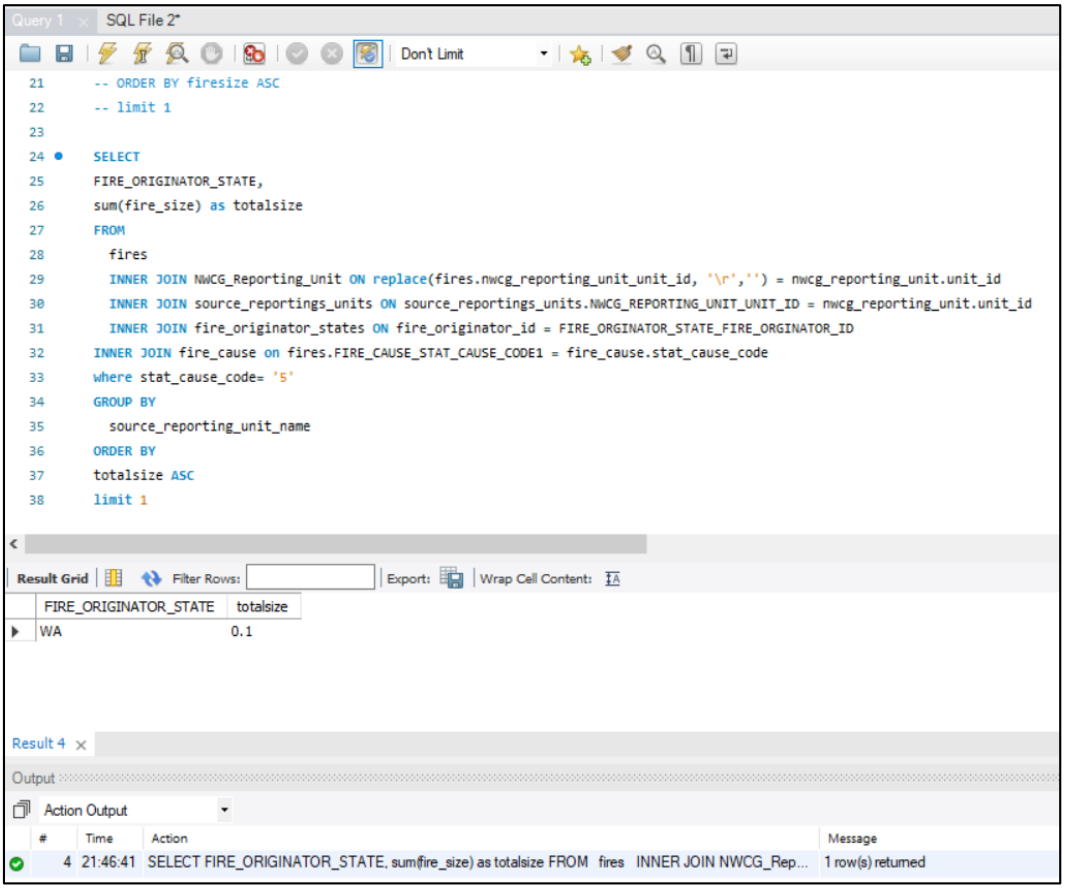
```
SELECT FIRE_ORIGINATOR_STATE from fires
INNER JOIN NWCG_Reporting_Unit ON nwcg_reporting_unit.unit_id
INNER JOIN source_reportings_units ON nwcg_reporting_unit.unit_id
INNER JOIN fire_originator_states ON FIRE_ORGINATOR_STATE_FIRE_ORGINATOR_ID
INNER JOIN fire_cause on fire_cause.stat_cause_code
where stat_cause_code= '5'
GROUP BY source_reporting_unit_name
ORDER BY COUNT(*) ASC
limit 1
```

Result:

Though there were many states (around 697 states) that had lower fires caused because of debris they all have the least chance of getting selected for the fund. However, we have randomly picked only one among the states that have the least chance of being selected

Screen Shot of SQL Query and Results





Query 2

Question:

2. One of the reporting agencies has suggested that children be banned from its forests unless there is one adult for every 3 children in a group visiting a forest. Name 3 forests where this would be the least appropriate.
(3 forests where fire due to children is least prevalent)

Notes/Comments About SQL Query and Results (Include # of Rows in Result)

Assumption:

We need to find 3 forests that have the least fires that are caused due to the activities of children, since the agencies are suggesting that children should be banned from the forests that have the highest rate of fires due to children. After running the query, there were more than 3 forests where this would apply, so we used the "limit 3" query to only show 3 forests where this is applicable. However, in figure 1 we have run the query to show that there were 63 forests which complies to this, but we are choosing only 3 randomly. Also, we are using the Source Reporting Unit Name as the Forest Name because there is no actual forest name that is available in the given dataset.

Translation:

Select source_reporting_unit_name from fires table and
Inner join NWCG_Reporting_Unit table on Unit_ID and
Inner join fire_cause table on STAT_CAUSE_CODE column
Where the stat_cause_code= '8' and SOURCE_REPORTING_UNIT_NAME like '%forest' then
Group by the source_reporting_unit_name and
Order by COUNT(*) in Asc and then
Limit 3 rows

Cleanup:

```
Select source_reporting_unit_name from fires
Inner join NWCG_Reporting_Unit on Unit_ID
Inner join fire_cause table on STAT_CAUSE_CODE
Where  stat_cause_code= '8' and SOURCE_REPORTING_UNIT_NAME like '%forest'
Group by source_reporting_unit_name
Order by COUNT(*) Asc
Limit 3
```

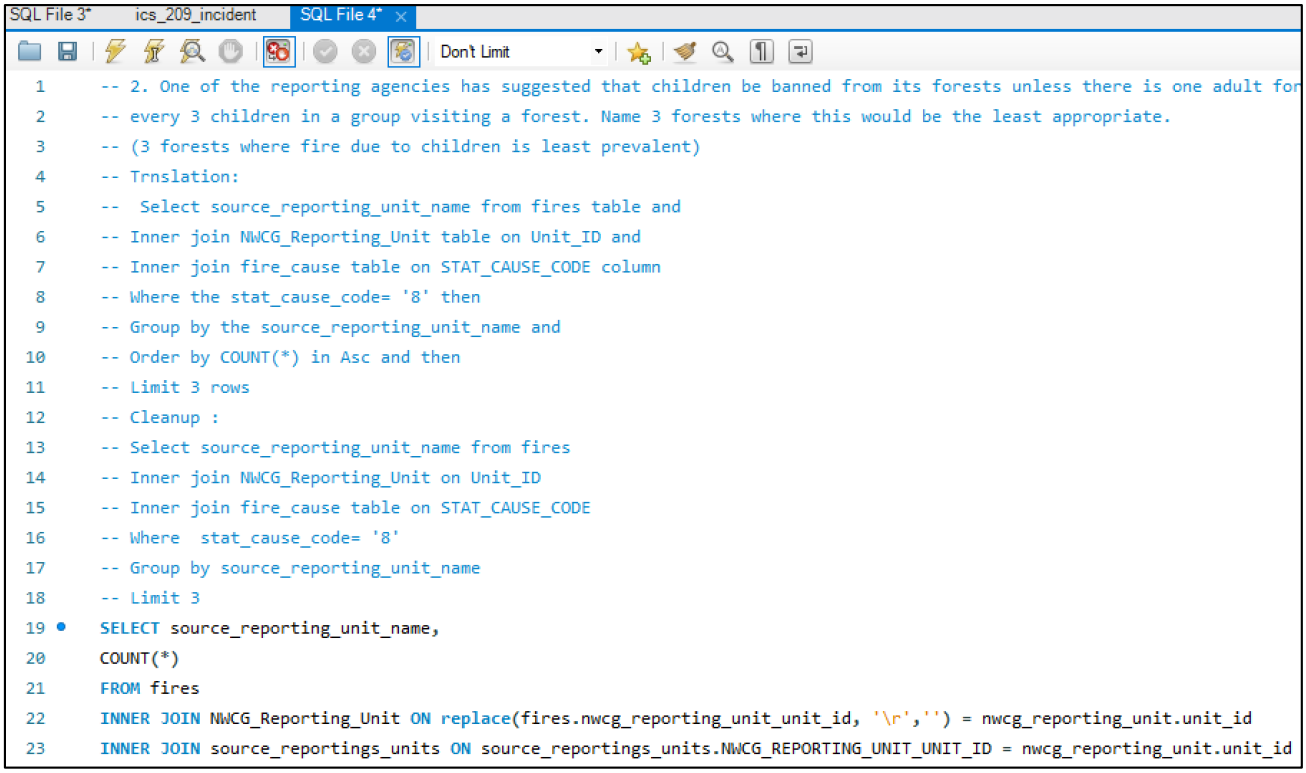
Result:

The 3 forests that were randomly selected among 63 are

1. Bridger-Teton National Forest
2. Pike San Isabel National Forest
3. Wallowa- Whitman National Forest

No. of rows = 3

Screen Shot of SQL Query and Results



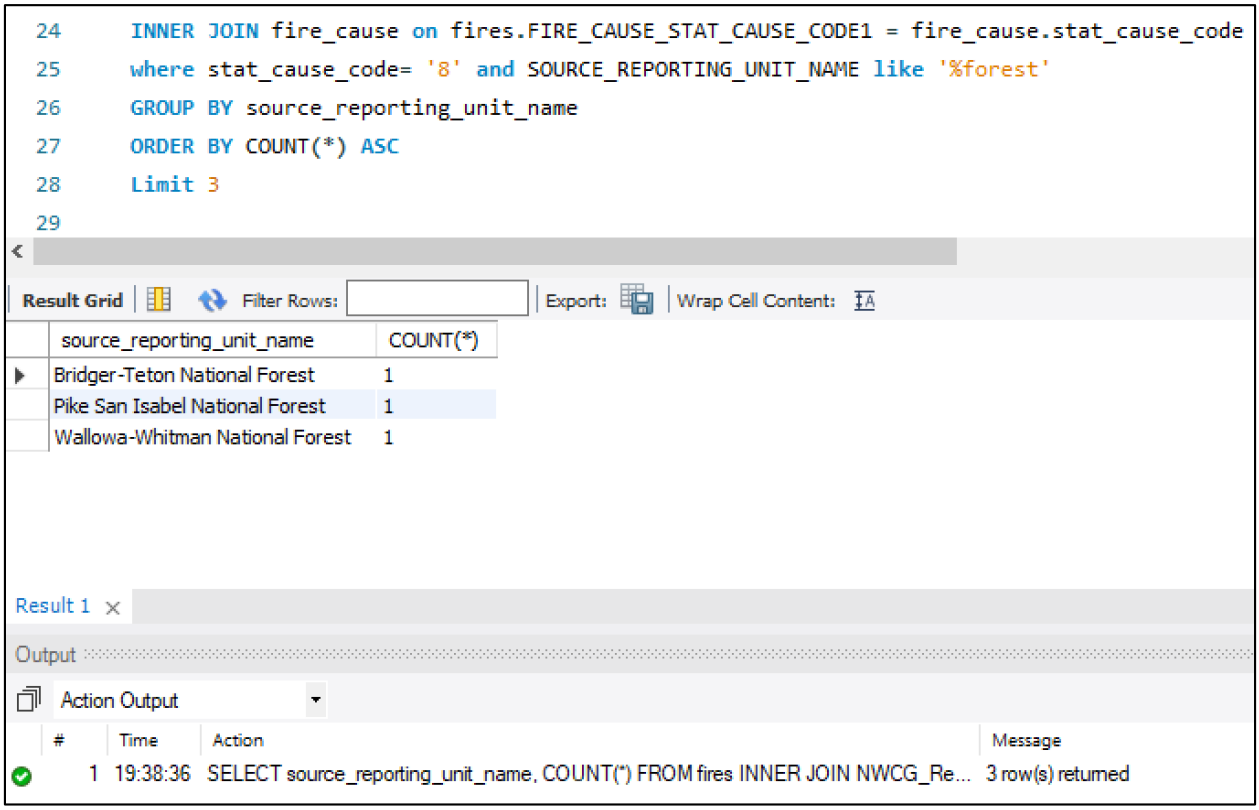
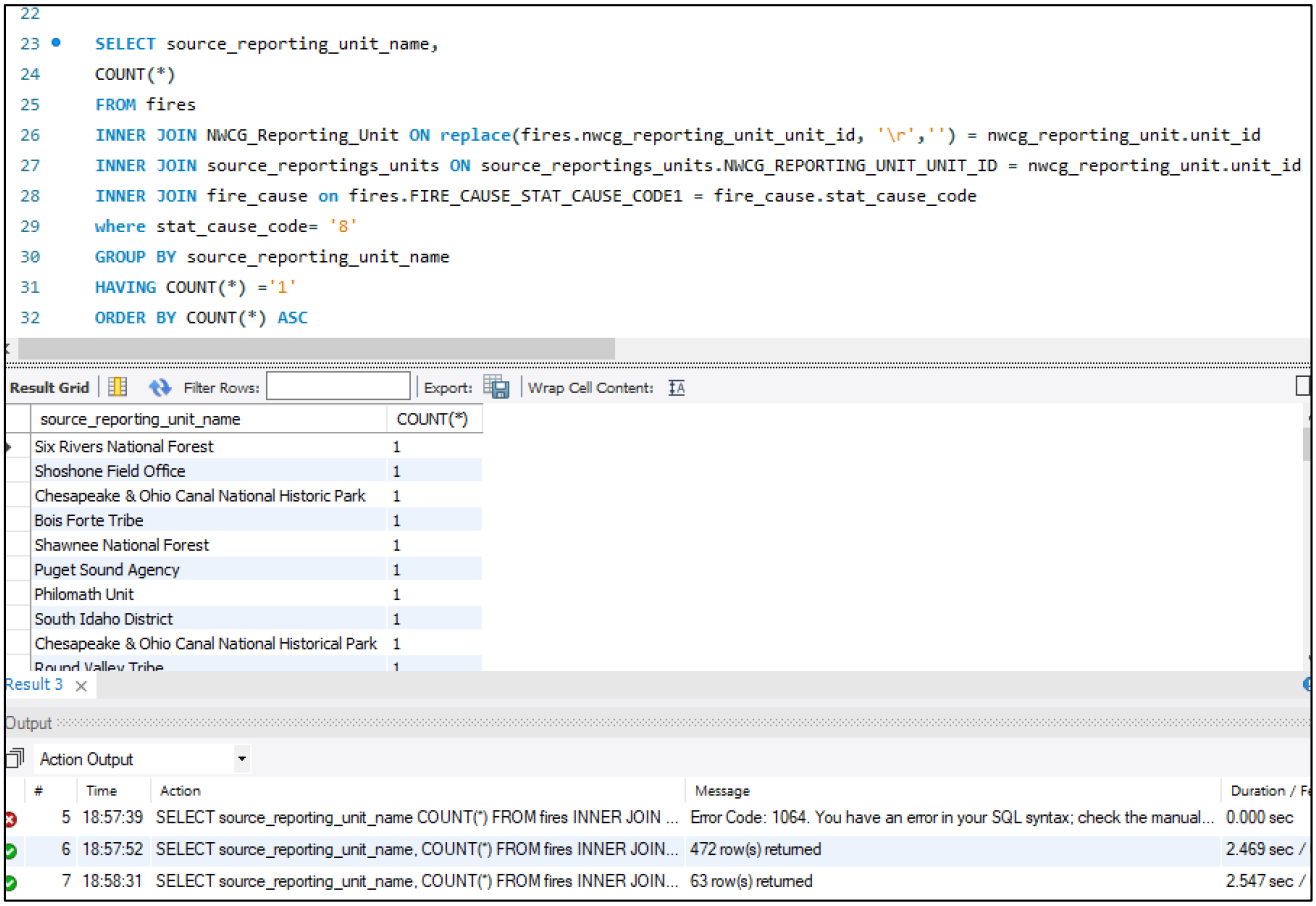


Figure 1:



```
22
23 •  SELECT source_reporting_unit_name,
24    COUNT(*)
25    FROM fires
26    INNER JOIN NWCG_Reporting_Unit ON replace(fires.nwcg_reporting_unit_unit_id, '\r','') = nwcg_reporting_unit.unit_id
27    INNER JOIN source_reportings_units ON source_reportings_units.NWCG_REPORTING_UNIT_UNIT_ID = nwcg_reporting_unit.unit_id
28    INNER JOIN fire_cause on fires.FIRE_CAUSE_STAT_CAUSE_CODE1 = fire_cause.stat_cause_code
29    where stat_cause_code= '8'
30    GROUP BY source_reporting_unit_name
31    HAVING COUNT(*) ='1'
32    ORDER BY COUNT(*) ASC
```

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

source_reporting_unit_name	COUNT(*)
Six Rivers National Forest	1
Shoshone Field Office	1
Chesapeake & Ohio Canal National Historic Park	1
Bois Forte Tribe	1
Shawnee National Forest	1
Puget Sound Agency	1
Philomath Unit	1
South Idaho District	1
Chesapeake & Ohio Canal National Historical Park	1
Round Valley Tribe	1

Result 3

Output

Action Output

#	Time	Action	Message	Duration / Fe
5	18:57:39	SELECT source_reporting_unit_name COUNT(*) FROM fires INNER JOIN ...	Error Code: 1064. You have an error in your SQL syntax; check the manual...	0.000 sec
6	18:57:52	SELECT source_reporting_unit_name, COUNT(*) FROM fires INNER JOIN...	472 row(s) returned	2.469 sec /
7	18:58:31	SELECT source_reporting_unit_name, COUNT(*) FROM fires INNER JOIN...	63 row(s) returned	2.547 sec /

Query 3

Question:

3. One advocacy group says human actions and not Nature is to blame for most wildfires. Write a query that supports this statement.

Assumption:

We need to check whether the assumption of the advocacy group that “human actions are reason for forest fires and not the nature” is right. In the fire cause table, there are several causes of fires that are given. We are considering the nature causes to be lightening and debris, and all other causes are due to human activities. We are also considering that the miscellaneous and undefined fields are also human activity. If the count of fires is more for human related than nature related, then the advocacy group is proven right.

Translation:

Select Case When stat_cause_code='1' or stat_cause_code = '5' then 'NATURE' ELSE as 'HUMAN' then
count(*) as Fire_count
from the fires table
Inner join fire_cause table on stat_cause_code then
Group by Cause_Type and then
order by Fire_count in desc

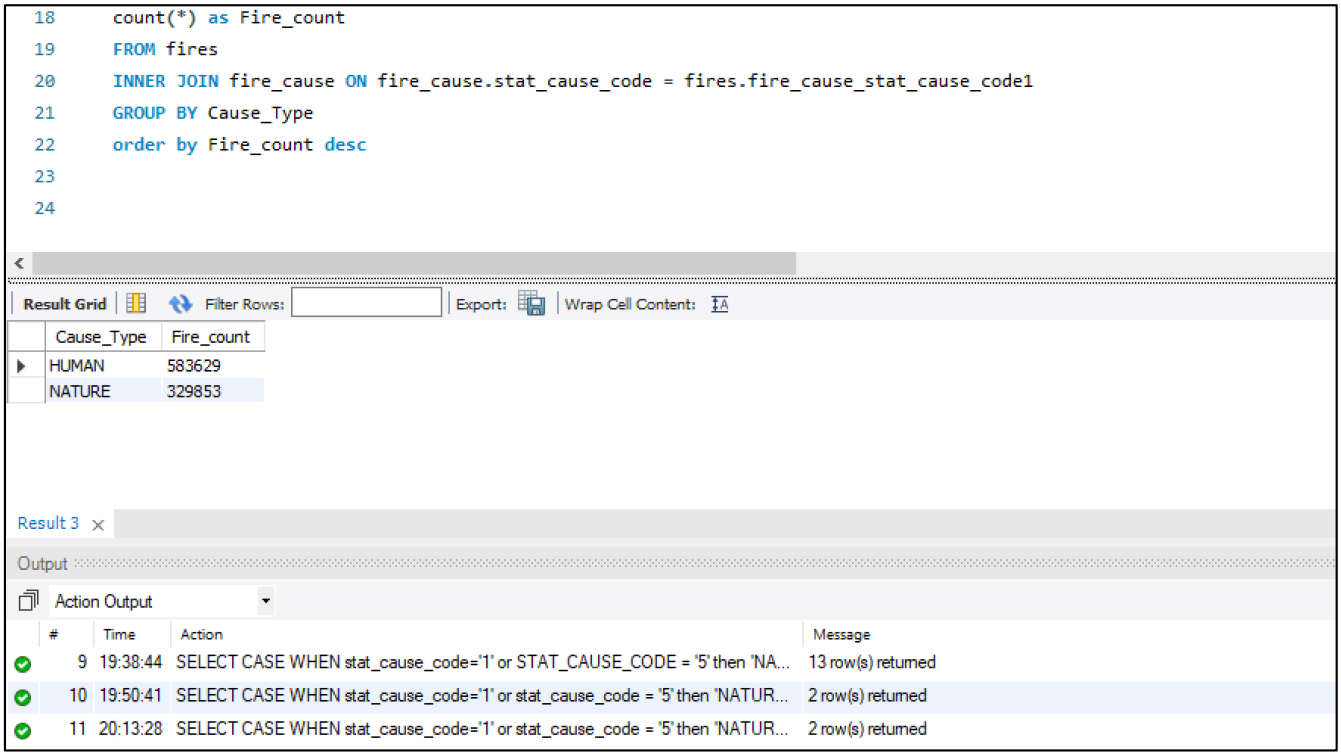
Cleanup:

Select Case When stat_cause_code='1' or stat_cause_code = '5' then 'NATURE' ELSE as 'HUMAN'
count(*) as Fire_count
from fires
Inner join fire_cause table on stat_cause_code
Group by Cause_Type
order by Fire_count desc

Result:

From the result we see that though the fires because of humans is more than the number of fires caused from nature, we cannot completely assume that all fires are caused by humans only and not by nature.
Fires Caused:
By humans = 583629
By Nature = 329853

Screen Shot of SQL Query and Results



Query 4

Question:

7. Which state had more fires in the first half of a calendar year than the second half of the calendar year?

Assumption:

We need to find which state had more fires in the first 182 days (first half) of the year when compared to the second 182 days (second half) of the year. In this case, we are assuming that 182 days is the exact half of the entire calendar year, so we are using the "day of year" field to accomplish this calculation.

Translation:

Select distinct fire_originator_state and sum(fist_half -sec_half) Difference
Then do subquery, where
Select FIRE_ORGINATOR_ID, sum(CASE when discovery_doy > 182 then '1' else '0' end) as FIST_HALF, then
sum(CASE when discovery_doy < 182 then '1' else '0' end) as sec_half
from fires table then
group by 1)
fires INNER JOIN fire_originator_states table on FIRE_ORGINATOR_ID
where fist_half > sec_half then
group by 1 and
order by Difference in descending and finally
limit 1

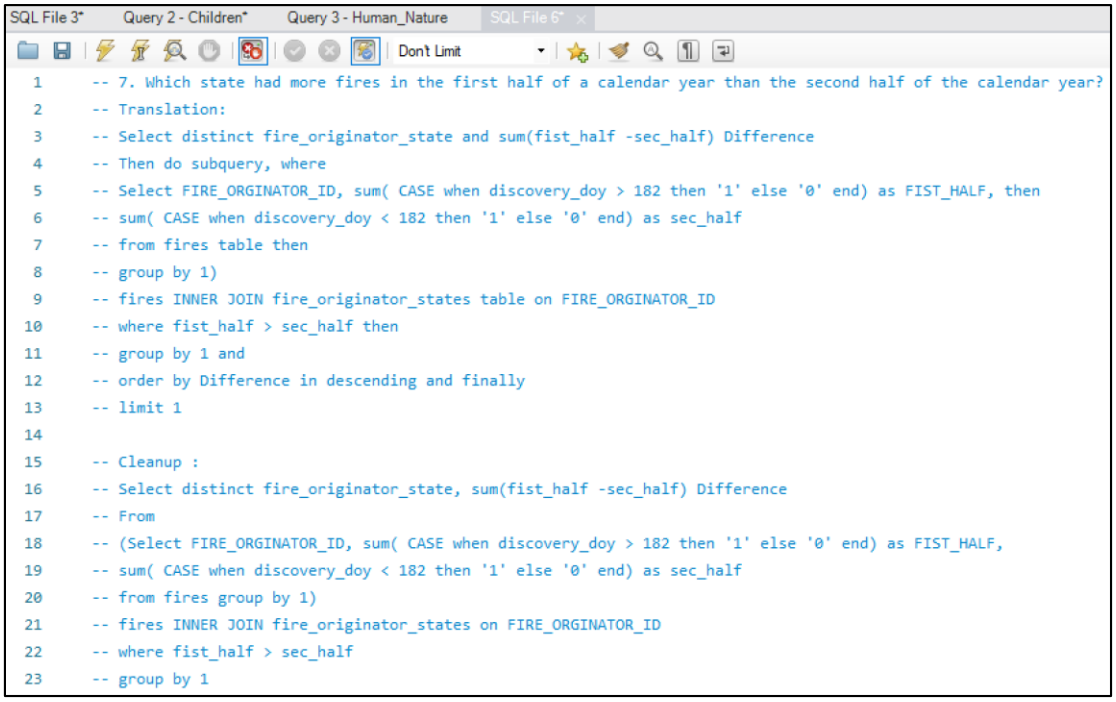
Cleanup:

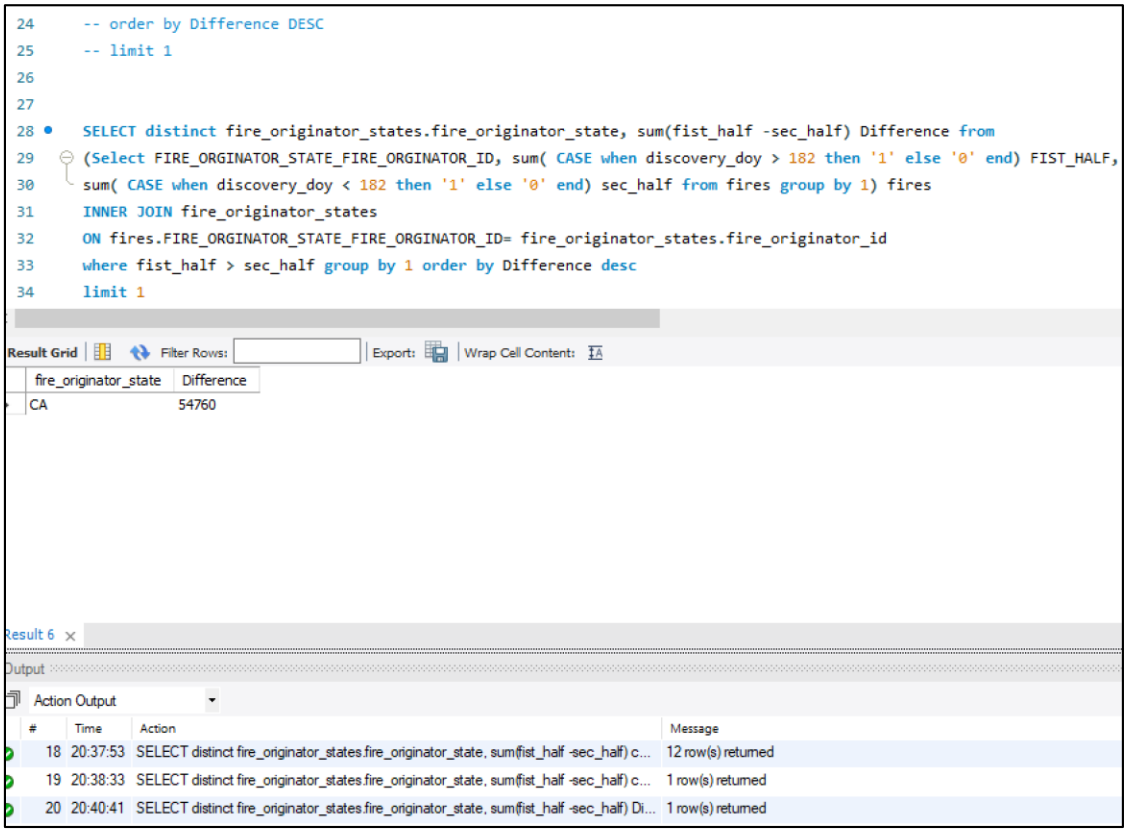
Select distinct fire_originator_state, sum(fist_half -sec_half) Difference
From
(Select FIRE_ORGINATOR_ID, sum(CASE when discovery_doy > 182 then '1' else '0' end) as FIST_HALF,
sum(CASE when discovery_doy < 182 then '1' else '0' end) as sec_half
from fires group by 1)
fires INNER JOIN fire_originator_states on FIRE_ORGINATOR_ID
where fist_half > sec_half
group by 1
order by Difference DESC
limit 1

Result:

After getting the states based on the condition that "fires in the first half of a calendar year greater than the second half of the calendar year", we sort them in descending order and then pick 1.
From the result we see that California has more no. of fires in the fires half of the calendar year than in the second half when compared to the other states.

Screen Shot of SQL Query and Results





Query 5

Question:

6. What were the forests that had no fires that lasted more than two days?

Assumption:

We need to find the forests that had fires that lasted 2 days or less. Therefore, we are assuming that this is applicable with the fires that are lasting for less than 3 days, but not 3 days. We are considering "discovery_doy" as the day that the fire was discovered and use that as the first day of the fire. We are using the "cont_doy" as the last day of the fire, so the difference between these two days will give the total length of the fire. We are using the Source Reporting Unit Name as the Forest Name because there is no actual forest name that is available in the given dataset

Translation:

Select distinct SOURCE_REPORTING_UNIT_NAME column
From
From here subquery:
(Select all from fires tables where cont_doy - discovery_doy <3)
Fires
INNER JOIN NWCG_Reporting_Unit table ON unit_id column and replace
(fires.nwcg_reporting_unit_unit_id, '\r','')
INNER JOIN source_reportings_units table ON unit_id

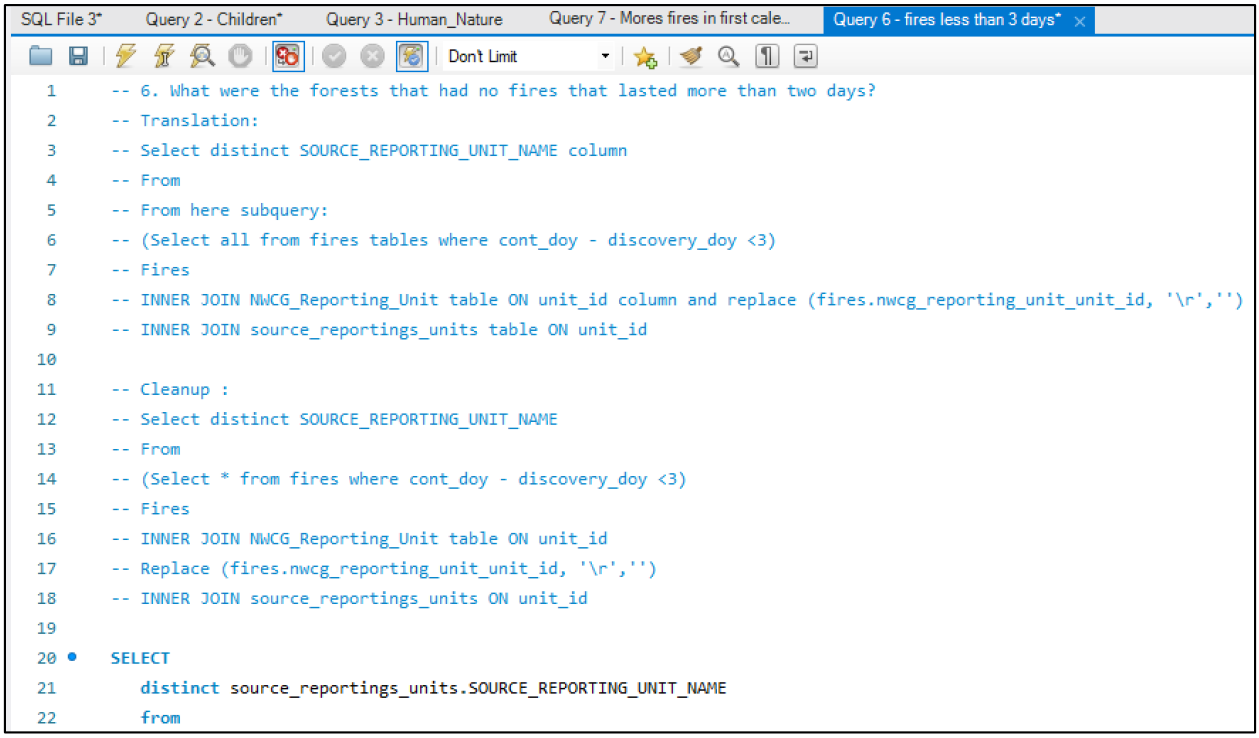
Cleanup:

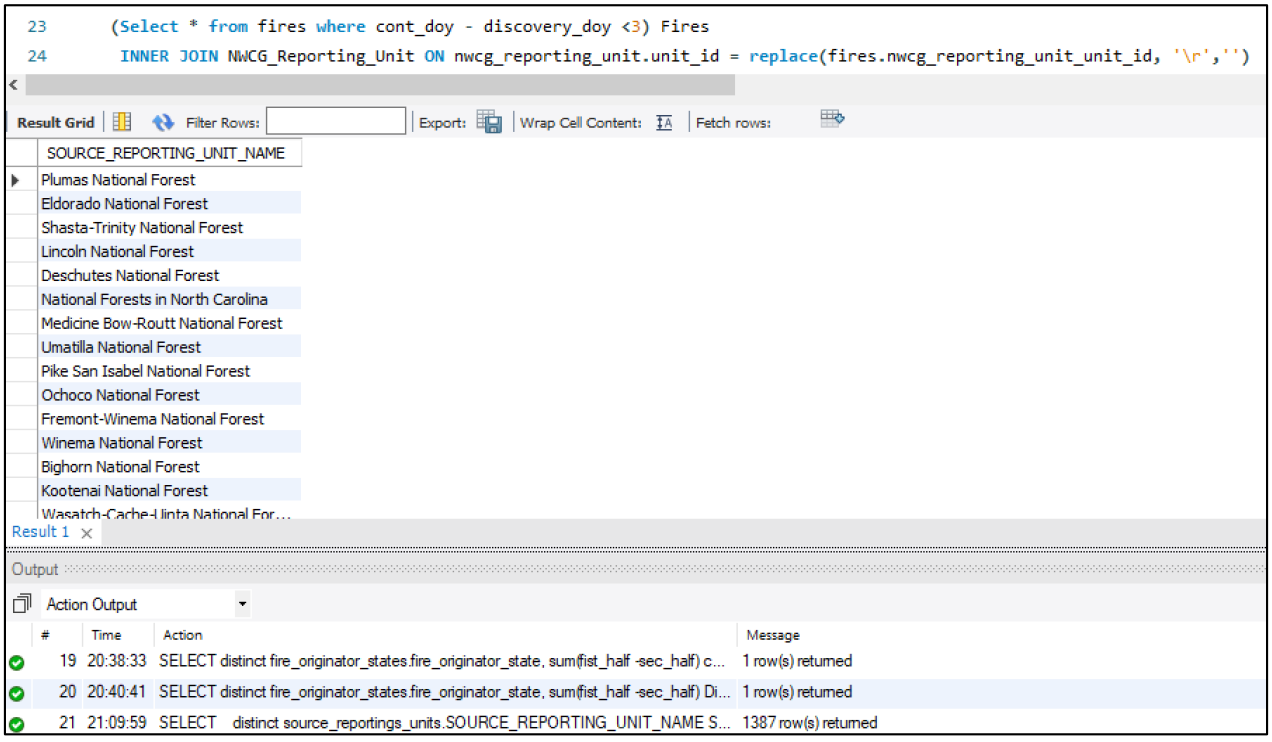
Select distinct SOURCE_REPORTING_UNIT_NAME
From
(Select * from fires where cont_doy - discovery_doy <3)
Fires
INNER JOIN NWCG_Reporting_Unit table ON unit_id
Replace (fires.nwcg_reporting_unit_unit_id, '\r','')
INNER JOIN source_reportings_units ON unit_id

Result:

From the result we can see that there were 1387 forests had fires which lasted for less than 3 days

Screen Shot of SQL Query and Results





Query 6

Question:

8. Which forest had the least number of fires?

Assumption:

We need to find the one forest that had the least number of fires. After running the query, there were 82 rows (forests) with the count as 1 fire as shown in figure 2. So, we had to limit the result to one forest, and got the result. We are using the Source Reporting Unit Name as the Forest Name because there is no actual forest name that is available in the given dataset.

Translation:

Select source_reporting_unit_name and COUNT(*)
FROM the fires table
INNER JOIN NWCG_Reporting_Unit table ON unit_id as nwcg_reporting_unit.unit_id
INNER JOIN source_reportings_units table ON UNIT_ID then
GROUP BY the source_reporting_unit_name
ORDER BY COUNT(FOD_ID) ASC then
Limit it by 1

Cleanup:

Select source_reporting_unit_name, COUNT(*)
FROM fires
INNER JOIN NWCG_Reporting_Unit table ON unit_id as nwcg_reporting_unit.unit_id
INNER JOIN source_reportings_units table ON UNIT_ID
GROUP BY source_reporting_unit_name
ORDER BY COUNT(FOD_ID) ASC
limit 1

Result:

From the 82 forests that had only one fire we have randomly selected one for our query

Screen Shot of SQL Query and Results

SQL File 3* | Query 8 - Forests with least fires

```
1   -- 8. Which forest had the least number of fires?
2   -- Translation:
3   -- Select source_reporting_unit_name and COUNT(*)
4   -- FROM the fires table
5   -- INNER JOIN NWCG_Reporting_Unit table ON unit_id as nwcg_reporting_unit.unit_id
6   -- INNER JOIN source_reportings_units table ON UNIT_ID then
7   -- GROUP BY the source_reporting_unit_name
8   -- ORDER BY COUNT(FOD_ID) ASC then
9   -- Limit it by  1
10
11  -- Cleanup:
12  -- Select source_reporting_unit_name, COUNT(*)
13  -- FROM fires
14  -- INNER JOIN NWCG_Reporting_Unit table ON unit_id as nwcg_reporting_unit.unit_id
15  -- INNER JOIN source_reportings_units table ON UNIT_ID
16  -- GROUP BY source_reporting_unit_name
17  -- ORDER BY COUNT(FOD_ID) ASC
18  -- limit 1
19
20  SELECT source_reporting_unit_name, COUNT(FOD_ID)
21  FROM fires
22  INNER JOIN NWCG_Reporting_Unit ON replace(fires.nwcg_reporting_unit_unit_id, '\r','') = nwcg_reporting_unit.unit_id
23  INNER JOIN source_reportings_units ON source_reportings_units.NWCG_REPORTING_UNIT_UNIT_ID = nwcg_reporting_unit.unit_id
24  GROUP BY source_reporting_unit_name
25  ORDER BY COUNT(FOD_ID) ASC
26  Limit 1
27
28
```

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

source_reporting_unit_name	COUNT(FOD_ID)
Russell Cave National Monument	1

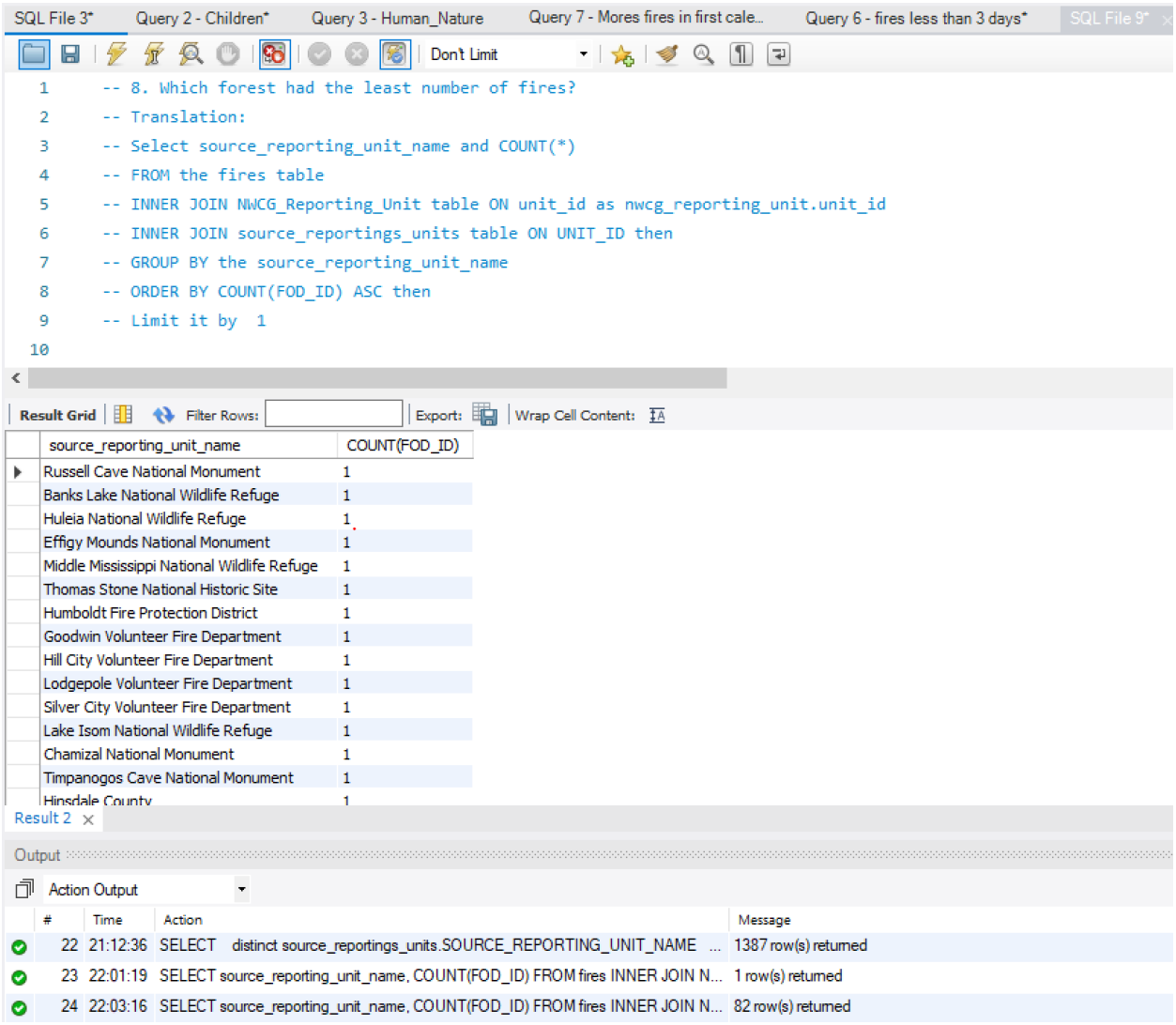
Result 3

Output

Action Output

#	Time	Action	Message
23	22:01:19	SELECT source_reporting_unit_name, COUNT(FOD_ID) FROM fires INNER JOIN N...	1 row(s) returned
24	22:03:16	SELECT source_reporting_unit_name, COUNT(FOD_ID) FROM fires INNER JOIN N...	82 row(s) returned
25	22:06:56	SELECT source_reporting_unit_name, COUNT(FOD_ID) FROM fires INNER JOIN N...	1 row(s) returned

Figure 2:


SQL File 3*
Query 2 - Children*
Query 3 - Human_Nature
Query 7 - Mores fires in first cale...
Query 6 - fires less than 3 days*
SQL File 9*
Don't Limit
1 -- 8. Which forest had the least number of fires?
2 -- Translation:
3 -- Select source_reporting_unit_name and COUNT(*)
4 -- FROM the fires table
5 -- INNER JOIN NWCG_Reporting_Unit table ON unit_id as nwcg_reporting_unit.unit_id
6 -- INNER JOIN source_reportings_units table ON UNIT_ID then
7 -- GROUP BY the source_reporting_unit_name
8 -- ORDER BY COUNT(FOD_ID) ASC then
9 -- Limit it by 1
10
Result Grid
Filter Rows:
Export:
Wrap Cell Content:
source_reporting_unit_name
COUNT(FOD_ID)
Russell Cave National Monument 1
Banks Lake National Wildlife Refuge 1
Huleia National Wildlife Refuge 1
Effigy Mounds National Monument 1
Middle Mississippi National Wildlife Refuge 1
Thomas Stone National Historic Site 1
Humboldt Fire Protection District 1
Goodwin Volunteer Fire Department 1
Hill City Volunteer Fire Department 1
Lodgepole Volunteer Fire Department 1
Silver City Volunteer Fire Department 1
Lake Isom National Wildlife Refuge 1
Chamizal National Monument 1
Timpanogos Cave National Monument 1
Hinsdale County 1
Result 2
Output
Action Output
Time Action Message
22 21:12:36 SELECT distinct source_reportings_units.SOURCE_REPORTING_UNIT_NAME ... 1387 row(s) returned
23 22:01:19 SELECT source_reporting_unit_name, COUNT(FOD_ID) FROM fires INNER JOIN N... 1 row(s) returned
24 22:03:16 SELECT source_reporting_unit_name, COUNT(FOD_ID) FROM fires INNER JOIN N... 82 row(s) returned

Data Review for MongoDB

Assumptions/Notes About Data Collections, Attributes and Relationships between Collections

In mongo DB, the database contains Collections and documents. The database name used here is "Wildfires". The collections are similar to tables in MySQL. This database contains only two tables and they are "fires" and "nwcg". The documents are similar to rows. The collection name "fires" contains 1880465 rows and the collection named nwcg contains 5867 rows.

Since MongoDB is a document-oriented database, no schema or structure is mandatory. Hence the tables are not normalized, and no relationship is created between them.

Physical Mongo Database

Assumptions/Notes About Data Set

The dataset "FPA_FOD_20170508" extracted from SQL Lite had many tables and were not in normalized form. The various tables that were present are Fires, NWCG, Elementary Geometries, idx_Fires_Shape, KNN, Spatial
Index geometry_columns, geometry_columns_auth, geometry_columns_field_infos, geometry_columns_statistics, geometry_columns_time, idx_Fires_shape_Node, idx_Fires_shape_parent, idx_Fires_shape_Rowid , spatial_ref_sys, spatial_ref_sys_aux, spatialite_history, views_geometry_columns, views_geometry_columns_auth, views_geometry_columns_feild_infos, views_geometry_columns_statistics. But we have only exported the Fires and NWCG tables from SQL Lite to csv or json and imported it into MongoDB because of the following reasons.

- Elementary Geometries, idx_Fires_Shape, KNN and Spatial Index were virtual tables and hence did not contain any information
- Tables as geometry_columns, geometry_columns_auth, geometry_columns_field_infos, geometry_columns_statistics, geometry_columns_time contain very less information i.e. only one row is present per table which is not relevant
- Tables such as idx_Fires_shape_Node, idx_Fires_shape_parent, idx_Fires_shape_Rowid contain details about parent node columns which was also present in the NWCG table. However, the parent column in the NWCG table itself is completely NULL which indicates that these data here will not be useful
- Tables such as spatial_ref_sys, spatial_ref_sys_aux, spatialite_history contains several rows with enough information. However, these tables are not related with Fires table and NWCG table which are our main tables. And so, we are not considering these tables
- Tables like views_geometry_columns, views_geometry_columns_auth, views_geometry_columns_feild_infos, views_geometry_columns_statistics contain no data. Therefore, we are not considering these tables

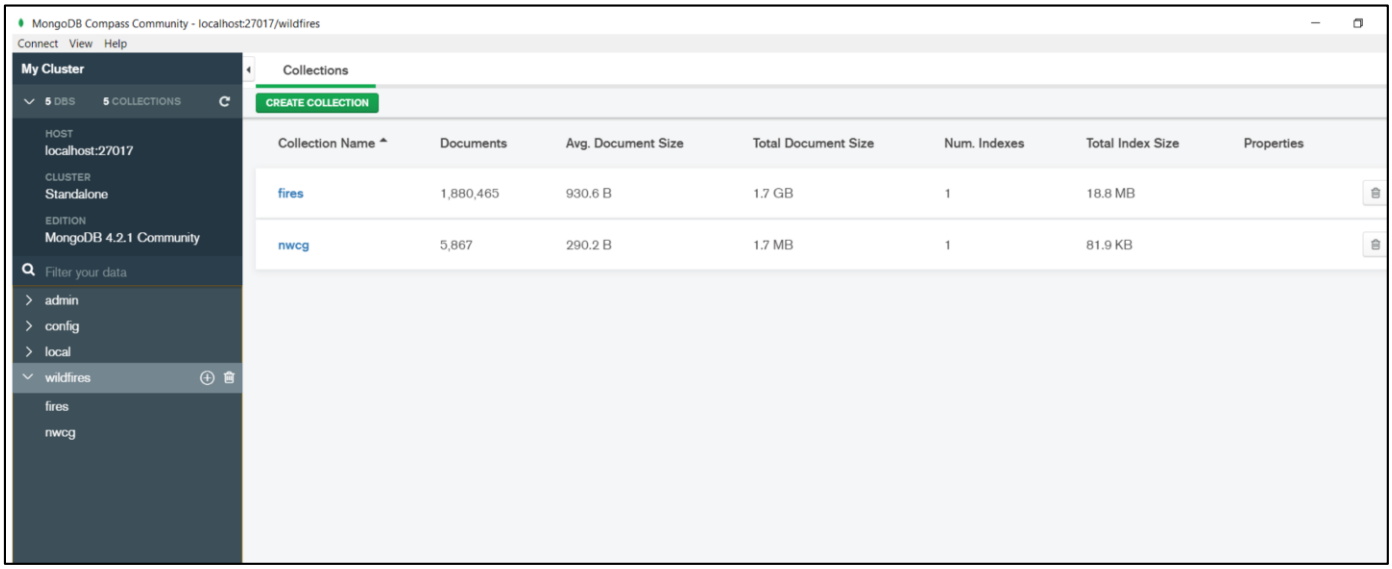
The two table Fires and NWCG were imported into Mongo DB directly without into normalization as Mongo DB doesn't required fixed structure or schema.

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

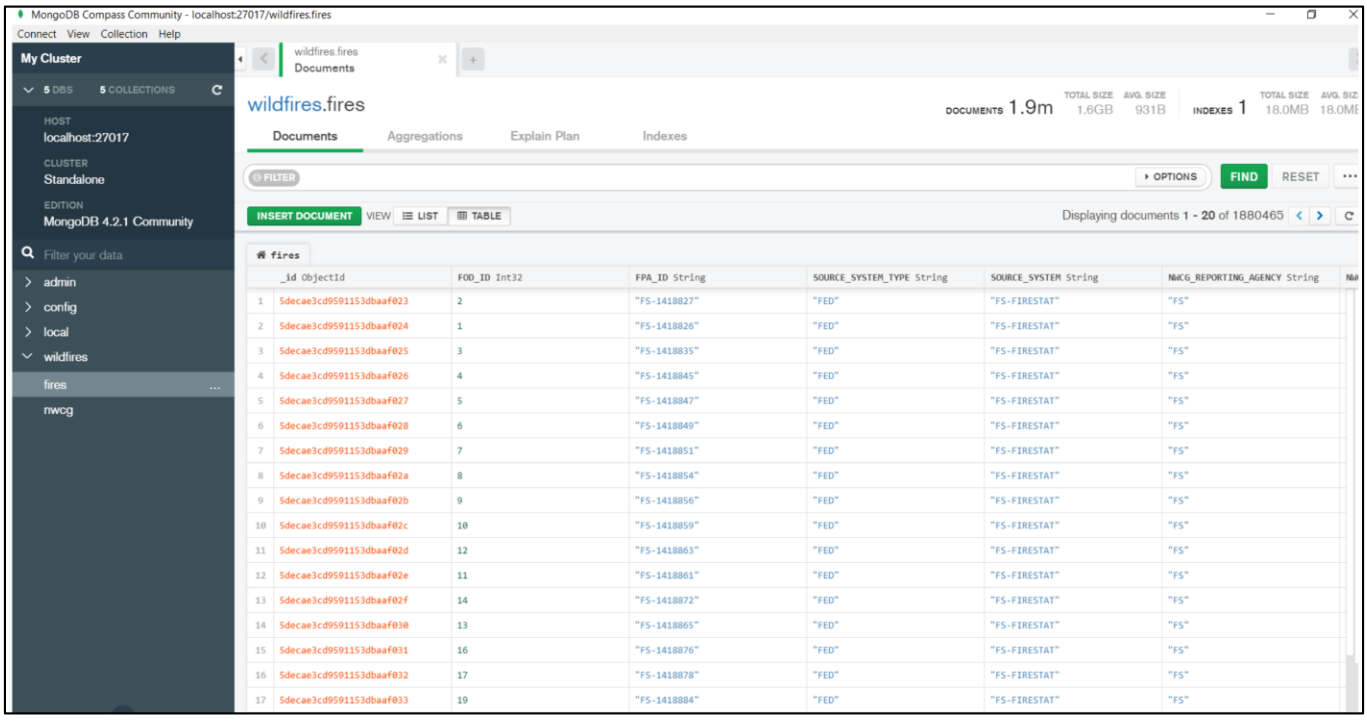
Database Name: Wildfires

Collections: Fires, NWCG

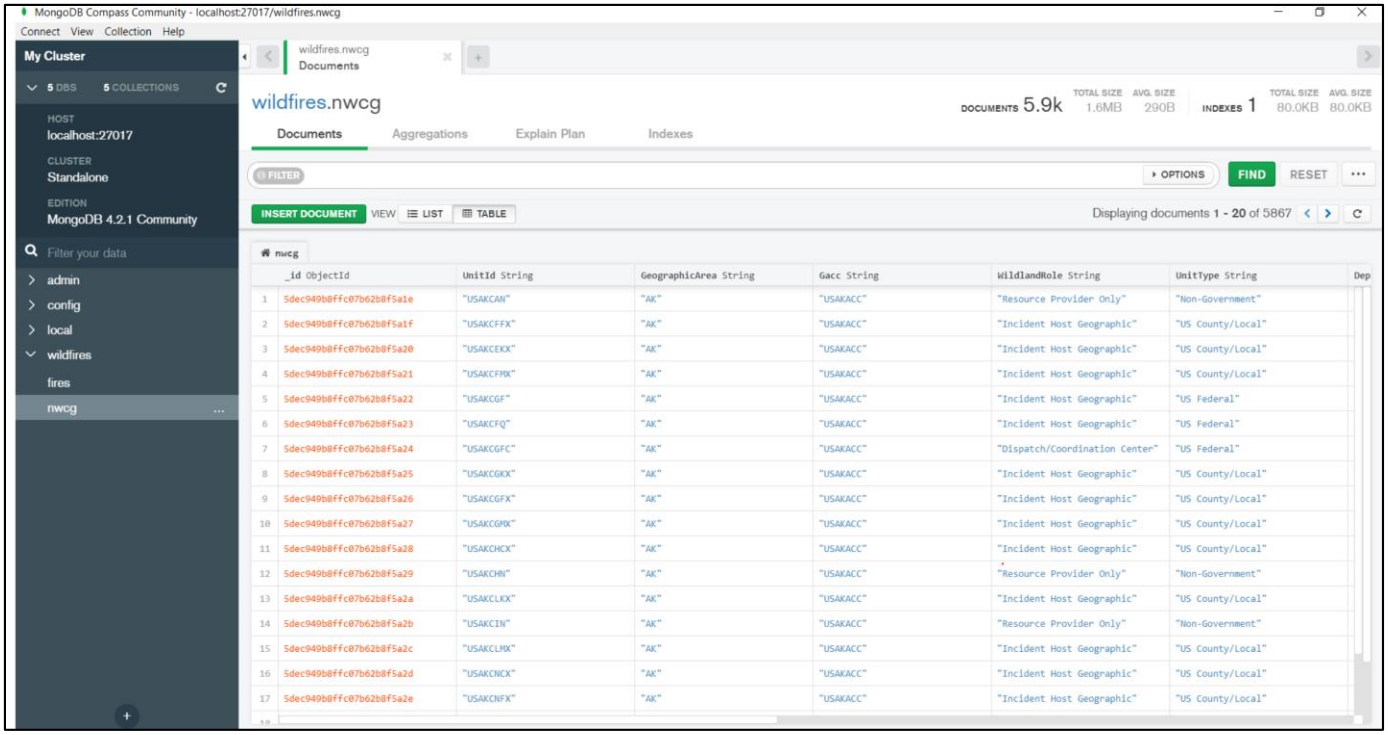
1. Screenshot of database with collections and attributes



2. Screenshot of collection "fires"



3. Screenshot of collection "NWCG"



Data in the Database

Collection Name	Relationships With Other Collections (if any)	# of Documents in Collection
Fires	-	1,880,465
NWCG	-	5867

MongoDB Queries/Code

Query 1

Question:

1. 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 state has the least chance to win a share of the fund?

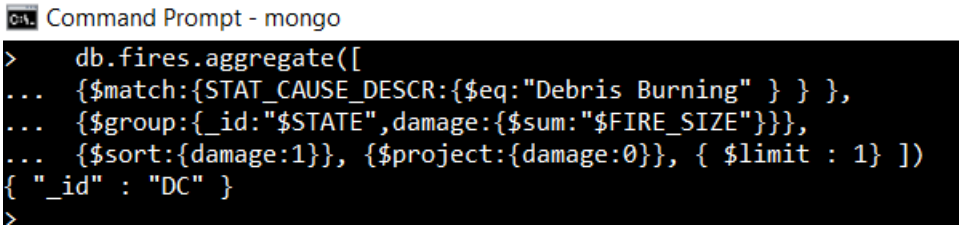
Translation

Select documents from the Fires collection and use the aggregate function such as $match
To match the condition Stat_Cause_Desc to Debris burning, then
Use $group to group the documents based on state
And then calculate the sum of fire size for each corresponding groups, and
then $sort to return the documents according to the sum of fire size in ascending order, and then
Set $limit as 1 to display 1 document

Cleanup:

Use db.fires.aggregate
$match Stat_Cause_Desc = “Debris Burning”
$group “$State” , damage: { $sum : “$fire size” }
$sort : damage :1
$project : damage :0
$limit 1

Screen Shot of MongoDB Query/Code and Results



Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result)

The leading beverage company is going to fund the states that has higher fires caused because of debris. Hence all states that had fires caused by debris are grouped and their fire size is calculated. The state which has the lease fire size will have the lowest chance to be picked.

Result: No. Of documents = 1

Assumption: Though there were many states that had least number of fires, one was picked randomly for the sake of answering to this query

Query 2

Question:

2. One of the reporting agencies has suggested that children be banned from its forests unless there is one adult for every 3 children in a group visiting a forest. Name 3 forests where this would be the least appropriate.

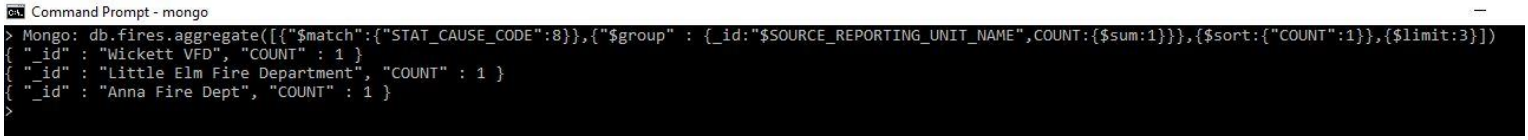
Translation:

Select documents from the Fires collection and use the aggregate function such as $match
To check the condition Stat_Cause_Code as 8, then
Use $group to group the documents based on reporting unit id as SOURCE_REPORTING_UNIT_NAME
And then calculate the count of fires for each corresponding groups, and
then $sort to return the documents according to the count of fires in ascending order, and then
Set $limit as 3 for produce 3 documents.

Cleanup:

Use db.fires.aggregate
$match Stat_Cause_Code= 8
$group SOURCE_REPORTING_UNIT_NAME
COUNT: $SUM,
$sort Firecount in ascending,
$limit 3

Screen Shot of MongoDB Query/Code and Results



Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result)

We need to find 3 forests that have the least fires that are caused due to the activities of children, since the agencies are suggesting that children should be banned from the forests that have the highest rate of fires due to children. Though there were many forests that satisfied the condition, since the question specifically asks us to pick only 3, we have randomly picked 3 forest from among all the forest that are in alignment with this condition.
The forests where this condition would be least appropriate are displayed in output

Result : No. of documents = 3

Assumption: We are using source reporting unit name as the forest name because there is no actual forest name that is available in the given dataset

Query 3

Question:

3. One advocacy group says human actions and not Nature is to blame for most wildfires. Write a query that supports this statement.

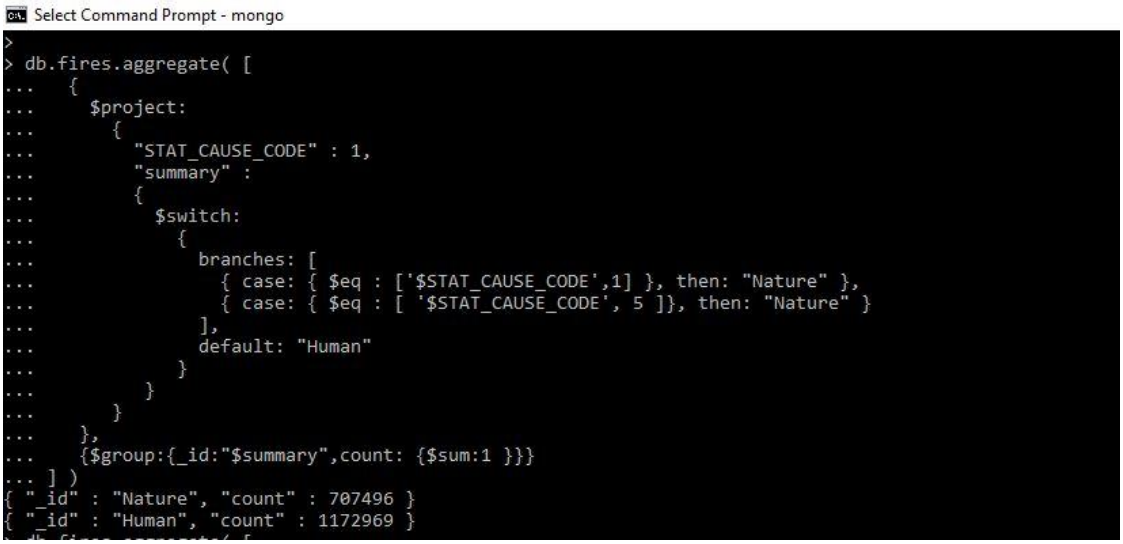
Translation:

Select documents from the Fires collection and use the aggregate function $project
To display the condition Stat_Cause_Code , then
Use $switch to group the documents based on branches Nature and human
And then use $group to display counts based on groups

Cleanup:

Use db.fires.aggregate
$project Stat_Cause_Code= 1, “summary”:
$switch branch ‘$STAT_CAUSE_CODE”,1 then “Nature”, ‘‘$STAT_CAUSE_CODE”,5 then “Nature”
Default :”Human”
$group: “$summary”, count: $sum:1

Screen Shot of MongoDB Query/Code and Results



Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result)

The total number of fires caused by humans is 1172969 and the total number of fires caused by nature is 707496. This proves that human cause more fires than nature and hence this query supports the statement made by the advocacy group

Assumptions: We are considering the nature causes to be lightening and debris, and all other causes are due to human activities. We are also considering that the miscellaneous and undefined fields are also human activity

Query 4

Question:

5. How many wildfires were not reported by more than one unit/agency? (wildfires that were reported by only one agency?)

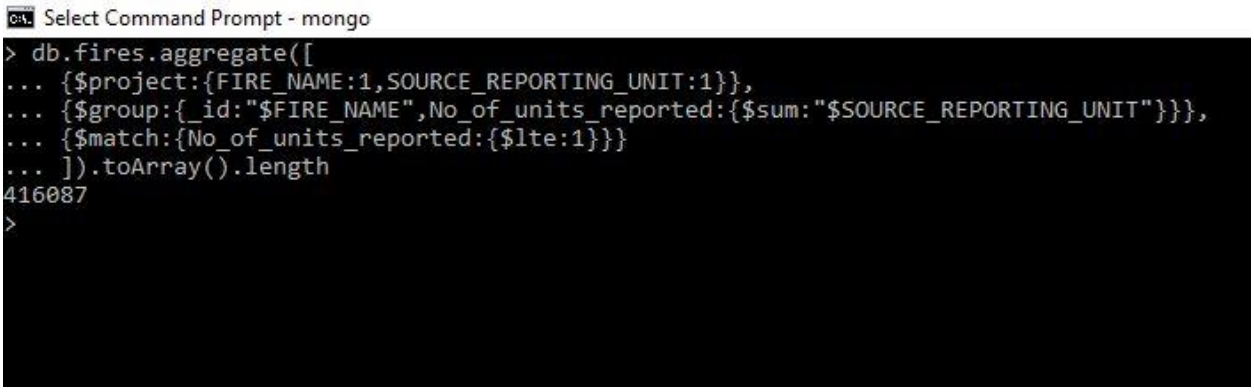
Translation:

Select the fires collection and use the aggregate function $project
To show the documents with FIRE_NAME and SOURCE_REPORTING_UNIT then we
Use $group to group the documents according to FIRE_NAME and sum of SOURCE_REPORTING_UNIT and then use $match for to set the condition, where, No._of_units_reported is less than or equal to 1, then we use the function toArray().length to find the count

Cleanup:

Use db.fires.aggregate
$project FIRE_NAME and SOURCE_REPORTING_UNIT ,
$group FIRE_NAME and total SOURCE_REPORTING_UNIT,
$match No_of_units_reported <= 1, toArray().length

Screen Shot of MongoDB Query/Code and Results



Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result)

The wildfires that were not reported by more than one agency or that was reported by only one agency were around 416087

Result : No. of documents = 416087

Query 5

Question:

8. Which forest had the least number of fires?

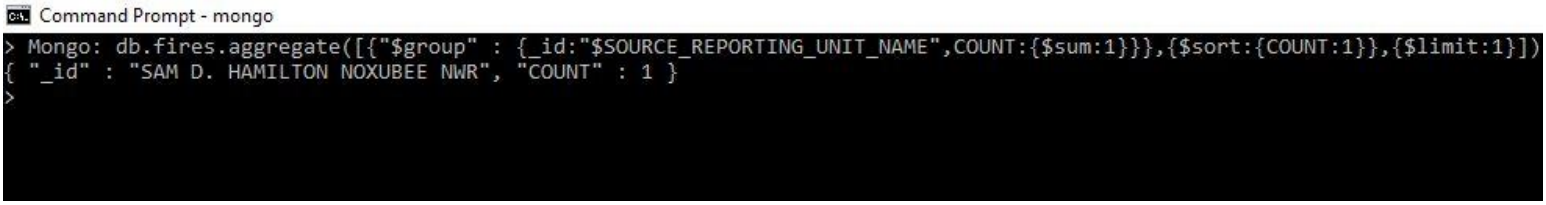
Translation:

Select documents from the fires collection and use the aggregate function $group
Use $group based on "Source_Reporting_Unit_Name" and count the no. of fires, then
Use $sort to arrange the documents based on count values in ascending order
And then use $limit to display forest with least no. of fires

Cleanup:

Use db.fires.aggregate
$group: "$Source_Reporting_Unit_Name", Count: $sum:1
$Sort: Count:1
$limit : 1

Screen Shot of MongoDB Query/Code and Results

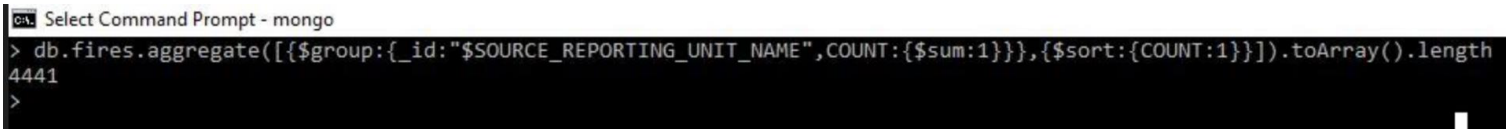


Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result)

There are many forests (around 4441 forests - as in figure A below) that had least no. of fires but for the sake of answering the query we have used the limit function to limit the answer to one

Assumption: We are using source reporting unit name as the forest name because there is no actual forest name that is available in the given dataset

Figure A



Query 6

Question:

6. What were the forests that had no fires that lasted more than two days?

Translation:

Select the documents from fires collection and and use the aggregate function $addfields,
Using this, we calculate date_diff which is obtained by subtracting Cont_Date and Discovery_Date
Then we use $match to match the condition where date_diff is less than or equal to 2
Then we use $project to show source_reporting_unit_name
Finally group the documents based on source_reporting_unit_name using $group and then find the total number of fires using Count function, then we use the function toArray().length to find the count

Cleanup:

Use db.fires.aggregate
$addFields : date_diff : {subtract : ["$CONT_DATE", "$DISCOVERY_DATE"]
$match : date_diff :$lt :3
$project : {Source_Reporting_Unit_Name":1}
$group: "$Source_Reporting_Unit_Name", Firescount: $sum:1
$project : {Source_Reporting_Unit_Name":1, Firescount:1 }
toArray().length

Screen Shot of MongoDB Query/Code and Results

Command Prompt - mongo

```
> db.fires.aggregate([
... {$addFields:{date_diff: { $subtract: ["$CONT_DATE", "$DISCOVERY_DATE"]}}},
... {$match:{date_diff: { $lt: 3}}},
... {$project:{SOURCE_REPORTING_UNIT_NAME : 1}},
...
... {$group: {_id:{SOURCE_REPORTING_UNIT_NAME: "$SOURCE_REPORTING_UNIT_NAME"}, Firescount:{$sum:1}}},
... {$project:{Firescount:1,SOURCE_REPORTING_UNIT_NAME : 1}}
... ]).toArray().length
4400
>
```

Notes/Comments About MongoDB Query/Code and Results (Include # of Documents in Result)

To identify how many days a fire existed we use the subtracted value of fire discovery date and fire contained date. Here the fire contained date is the day when the fire was controlled or put out.

Assumption: We are using source reporting unit name as the forest name because there is no actual forest name that is available in the given dataset

Result : No. of documents = 4400