databricksAssignment

Spark Internals

Module 2 Assignment

☆ In this assignment you:

- Analyze the effects of caching data
- Speed up Spark queries by changing default configurations

For each **bold** question, input its answer in Coursera.

```
%run ../Includes/Classroom-Setup
```

Data mounted to /mnt/davis ...

OK

Make sure nothing is already cached by clearing the cache (**NOTE**: This will affect all users on this cluster).

```
CLEAR CACHE
```

OK

☆ Effects of Caching

Count the number of records in our `fireCalls` table.

Question 1

How many fire calls are in our table?

```
SHOW TABLES
```

	database	tableName	isTemporary
1	databricks	firecalls	false
2	databricks	fireincidents	false

Showing all 2 rows.

```
select COUNT(*) FROM firecalls
```

	count(1)
1	240613

Showing all 1 rows.

Now speed up your query by caching the data, then counting!

```
CACHE TABLE firecalls
```

OK

Look in the Spark UI, how many partitions is our data?

☆ Changing Spark Configurations

We cached the data to speed up our computation, but let's see if we can get it even faster by changing some default Spark Configurations.

Let's count the number of each type of unit. Group by the `Unit Type`, count the number of calls for each type, and display the unit types with the most calls first.

Question 2

Which "Unit Type" is the most common?

```
SELECT `Unit Type`,COUNT(*) FROM firecalls
GROUP BY `Unit Type`
```

	Unit Type	count(1)
1	AIRPORT	992
2	MEDIC	74219
3	CHIEF	17757
4	RESCUE SQUAD	4413
5	RESCUE CAPTAIN	8259
6	TRUCK	25995
7	INVESTIGATION	250
8	ENGINE	92828

Showing all 10 rows.

Question 3

What type of transformation, wide or narrow, did the `GROUP BY` and `ORDER BY` queries result in?

Let's change the `spark.sql.shuffle.partitions` configuration from its default value of `200` and set it to `2`.

```
-- answer: wide
SET spark.sql.shuffle.partitions=2
```

	key	value
1	spark.sql.shuffle.partitions	2

Showing all 1 rows.

Copy and run the code from earlier to get the unit type counts from earlier. Now how long does this query take?

```
SELECT `Unit Type`,COUNT(*) FROM firecalls
GROUP BY `Unit Type`
```

	Unit Type	count(1)
1	ENGINE	92828
2	MEDIC	74219
3	TRUCK	25995
4	AIRPORT	992
5	CHIEF	17757
6	PRIVATE	14793
7	RESCUE CAPTAIN	8259
8	SUPPORT	1107

Showing all 10 rows.

Why was it faster? Check the 2 stages in the UI and see how many tasks were launched. Which number of tasks corresponds to your shuffle partition number?

Question 4

How many tasks were in the last stage of the last job? -- 1