Big Data Systems - CS4545/CS6545
Hands-on 2
Due: February 11, 2021 at 11 am

In this hands-on you'll become familiar with a parallel relational database by working with Stado (previously know as, GridSQL).

INSTRUCTIONS: How to start the Stado parallel database and run queries there

Follow steps 1 through 5 below.

1. Remotely connect to the FCS lab (see *RemoteDesktopToALabMachine.pdf*)

2. Launch and login into all 4 BigDataSystems VMs (Master, Slave 1, Slave 2 and Slave 3).

3. Start single PostgreSQL database instances on all 4 servers by running the command on a Terminal:

```
$ pg_ctl start
```

4. On BigDataSystemsMaster VM, start the parallel database server:

```
$ cd /home/bigdata/stado/stado/bin
$ gs-server.sh -d testdb
```

5. On BigDataSystemsMaster VM, start a Stado SQL client to connect to the parallel database server:

```
$ gs-cmdline.sh -d testdb -u admin -p admin -z
```

Then show the tables

```
Stado -> show tables;
```

6. (optional) You can download this file by the typing the command in a Terminal:

```
$ wget http://www.cs.unb.ca/~sray/teaching/bds/handson/bds_handson2.pdf
```

INSTRUCTIONS: Task1 (deliverable)

Run the necessary commands on Stado SQL client for this task.

1.a Create a partitioned table *mytable1* with the partitioning key *col1*. The schema of *mytable1* is shown below:

mytable1

Column name	**Type**
col1	int
col2	char

1.b Insert the following tuples (with given values) into *mytable1*. Use Stado -> command prompt.

(1, 'A')
(2, 'B')
(3, 'C')
(4, 'D')
(5, 'E')

2.a Create a replicated table *mytable2*. The schema of *mytable2* is shown below:

mytable2

Column name	**Type**
fld1	int
fld2	char(2)

2.b Insert the following records into *mytable2*

(101, 'NL')
(102, 'PE')
(103, 'NS')
(104, 'NB')

INSTRUCTIONS: How to run queries on a single instance PostgreSQL on the Master VM

1. Assuming the database was already started with the pg_ctl start command, you can launch a SQL client

```
$ psql testdb
```

2. You can run SQL queries using the SQL prompt. To enable timing run: `\timing`

INSTRUCTIONS: Task2 (deliverable)

1. Run the 3 queries, Q3, Q4 and Q14, from TPC-H benchmark (see next page) via **Stado** SQL client. Run each query three times. You can ignore the first run (called "cold"). Make a note of the query execution times of the subsequent 3 (called "warm") runs.

2. Then on BigDataSystemsMaster VM, open a single instance **PostgreSQL** SQL client and run the same 3 TPC-H queries on PostgreSQL SQL client. Run each query three times. You can ignore the first run. Make a note of the query execution times of the subsequent 2 runs.

3. Calculate the speedup of TPC-H queries Q3, Q4 and Q14 with the parallel execution of Stado over that of single instance PostgreSQL execution.

4. To calculate speedup use the formula discussed in the class. Note, take an average of the 2 warm runs. Do not use the cold runs.

INSTRUCTIONS: How to shutdown the parallel database and PostgreSQL

Finally, on BigDataSystemsMaster VM, shutdown the Stado client and stop the parallel database server:

```
gs-dbstop.sh -d testdb -u admin -p admin
```

To stop single PostgreSQL instances, run the command on all machines: `pg_ctl stop`

TPC-H BENCHMARK QUERIES

Q3.

```
select l_orderkey, sum(l_extendedprice * (1 - l_discount)) as revenue, o_orderdate, o_shippriority from customer, orders, lineitem where c_mktsegment IN ('AUTOMOBILE') and c_custkey = o_custkey and l_orderkey = o_orderkey and o_orderdate < date '1995-03-19' and l_shipdate > date '1995-03-19' group by l_orderkey, o_orderdate, o_shippriority order by revenue desc, o_orderdate;
```

Q4.

```
select o_orderpriority, count(*) as order_count from orders, lineitem where o_orderdate >= date '1994-01-01' and o_orderdate < date '1994-04-01'  and  l_orderkey = o_orderkey and l_commitdate < l_receiptdate  group by o_orderpriority order by o_orderpriority;
```

Q14.

```
select 100.00 * sum(case when p_type like 'PROMO%' then l_extendedprice * (1 - l_discount) else 0 end) / sum(l_extendedprice * (1 - l_discount)) as promo_revenue from lineitem, part where l_partkey = p_partkey and l_shipdate >= date '1995-03-01' and l_shipdate < (date '1995-03-01' + interval '1 month');
```

SUBMISSION INSTRUCTIONS :

1) Submit a .pdf file via Desire To Learn (D2L) with the following:

a. From Task1:

i) Schema of your table mytable1 (output of: show table mytable1)
ii) Output of the query: SELECT * FROM mytable1
iii) Schema of your table mytable2 (output of: show table mytable2)
iv) Output of the query: SELECT * FROM mytable2

b. From Task2:

i) Average execution time of the warm runs of Q3, Q4 and Q14 with Stado
ii) Average execution time of the warm runs of Q3, Q4 and Q14 with single instance PostgreSQL
iii) Speedup of Q3, Q4 and Q14 achieved with Stado against single instance PostgreSQL

2) Mention the following on the top of your submitted file: your name and hands-on#. Hands-on not submitted electronically via D2L or submitted after the due date will NOT be marked.

3) Work must be done individually.