School of Information Technology

Semester 1, 2018

Tutorial 4

Scalable Data Analytics: The role of indexes and data partitioning

Question 1: Storage Layer of a DBMS

Discuss the following questions about data on external storage in a DBMS:

a) Why does a DBMS store data on external storage?

b) Why are I/O costs important in a DBMS?

c) What is the role of the buffer cache in a DBMS?

Question 2: System Catalog Exploration

a) What information is stored in the database system catalogs? Explore which tables you have in your local schema and which indexes are already in place.

SELECT * FROM PG_INDEXES

Question 3: Indexing Tuning with PostgreSQL

Download and run the SQL statements in the file DATA2001_Tutorial_Wk5.sql. This will create three relationships (driver, vehicle, triplog) and some demo data for a car hire company which loggs every trip by a driver.

a) Check whether there are already some indexes available on your schema. To do so, issue the SELECT statement on PG INDEXES as given above, and then explore the details on columns and indexes shown in the result. Can you explain, what you see?

b) Test then the following queries, recording the execution times:

```
SELECT * FROM Driver WHEHoiRE given_name = 'Eric';
SELECT * FROM Driver WHERE family_name = 'Andrews';
SELECT COUNT(*) FROM Vehicle WHERE year = 2005;
SELECT COUNT(*) FROM Vehicle WHERE year < 2005;
SELECT * FROM TripLog WHERE distance > 200;
SELECT * FROM TripLog WHERE distance > 200 AND distance < 250 ;
SELECT * FROM TripLog WHERE distance > 200 AND
                (end_time - start_time) < (interval '1 hours')  ;
```

c) Try these a few times to see how the execution time varies. Why is it not the same each time?
d) Do all three queries take about the same time? If not, why?
e) Hint: You can get data on how the queries are executed by putting EXPLAIN ANALYZE before the rest of the query.
f) Create an index for the TripLog relation for the distance, car_id and vehicle_id attributes:

```
CREATE INDEX distance_triplog_ind ON TripLog (distance);
CREATE INDEX car_id_triplog_ind ON TripLog (car_id);
CREATE INDEX vehicle_id_triplog_ind ON TripLog (vehicle_id);
CREATE INDEX vehicle_year_ind ON Vehicle (year);
```

(You may find it useful to run VACUUM ANALYZE to update the query planner's statistics.)

g) Rerun the above queries with the indexes and compare the estimates to see if the query is affected.

Question 4: The Federal Government has set an Australia-wide speed limit of 110 km/h. You have been hired to find out who exceeded this limit during their trip.

Consider the following questions:

a) How many vehicles went over the maximum speed during their trip?
b) What is the minimal number tables to find this information?
c) We wish to send a penalty letter to all the drivers that violated the maximum speed. Construct a query to give the addresses of drivers who exceeded the speed limit. Your result should contain, driver address, given and family names, vehicle model and description, trip start time and speed.
d) How many tables must you access?
e) How many joins are required?
f) What types of indexes are required?
g) On which tables and on which columns should these indexes be built?

Question 5: Partitioning Data

Discuss how portioning (horizontal vs vertical) the above car hire company data could help in scalability and accessibility of the data.