DS Group Assignment 4: Query Processing and Optimization

Deadline - 11:55 PM, 23rd October 2020

Question A

Given a relational database schema $(R_1, R_2, R_3, \ldots, R_k)$ and that there are atleast enough appropriate attributes such that all k relations can be joined like so

$$R_1 \bowtie R_2 \bowtie \ldots \bowtie R_k$$

Note that more joins are possible. For example, $R_1 \bowtie R_k$ or $R_3 \bowtie R_5$ might be possible, but at the very least there will be attributes to allow the following join

$$R_1 \bowtie R_2 \bowtie \ldots \bowtie R_k$$

And given a **batch** of independent queries $\mathcal{Q}$ where we define a **batch** of queries to be a set of one or more than one queries which are executed simultaneously

What are the different issues or considerations for a batch of queries $\mathcal{Q}$ to be executed together?

Question B:

Come up with 3 disjoint batches of queries $(\mathcal{Q}_1, \mathcal{Q}_2, \mathcal{Q}_3)$ such that $\mathcal{Q}_1$ has 2 queries, $\mathcal{Q}_2$ has 5 queries and $\mathcal{Q}_3$ has 10 queries for the COMPANY database as shown below

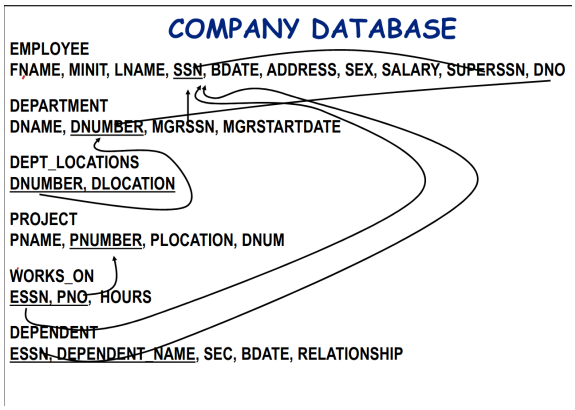

Figure 5.6
One possible database state for the COMPANY relational database schema.

EMPLOYEE

Fname	Minit	Lname	Ssn	Bdate	Address	Sex	Salary	Super_ssn	Dno
John	B	Smith	123456789	1965-01-09	731 Fondren, Houston, TX	M	30000	333445555	5
Franklin	T	Wong	333445555	1955-12-08	638 Voss, Houston, TX	M	40000	888665555	5
Alicia	J	Zelaya	999887777	1968-01-19	3321 Castle, Spring, TX	F	25000	987654321	4
Jennifer	S	Wallace	987654321	1941-06-20	291 Berry, Bellaire, TX	F	43000	888665555	4
Ramesh	K	Narayan	666884444	1962-09-15	975 Fire Oak, Humble, TX	M	38000	333445555	5
Joyce	A	English	453453453	1972-07-31	5631 Rice, Houston, TX	F	25000	333445555	5
Ahmad	V	Jabbar	987987987	1969-03-29	980 Dallas, Houston, TX	M	25000	987654321	4
James	E	Borg	888665555	1937-11-10	450 Stone, Houston, TX	M	55000	NULL	1

DEPARTMENT

Dname	Dnumber	Mgr_ssn	Mgr_start_date
Research	5	333445555	1988-05-22
Administration	4	987654321	1995-01-01
Headquarters	1	888665555	1981-06-19

DEPT_LOCATIONS

Dnumber	Dlocation
1	Houston
4	Stafford
5	Bellaire
5	Sugarland
5	Houston

WORKS_ON

Essn	Pno	Hours
123456789	1	32.5
123456789	2	7.5
666884444	3	40.0
453453453	1	20.0
453453453	2	20.0
333445555	2	10.0
333445555	3	10.0
333445555	10	10.0
333445555	20	10.0
999887777	30	30.0
999887777	10	10.0
987987987	10	35.0
987987987	30	5.0
987654321	30	20.0
987654321	20	15.0
888665555	20	NULL

PROJECT

Pname	Pnumber	Plocation	Dnum
ProductX	1	Bellaire	5
ProductY	2	Sugarland	5
ProductZ	3	Houston	5
Computerization	10	Stafford	4
Reorganization	20	Houston	1
Newbenefits	30	Stafford	4

DEPENDENT

Essn	Dependent_name	Sex	Bdate	Relationship
333445555	Alice	F	1986-04-05	Daughter
333445555	Theodore	M	1983-10-25	Son
333445555	Joy	F	1958-05-03	Spouse
987654321	Abner	M	1942-02-28	Spouse
123456789	Michael	M	1988-01-04	Son
123456789	Alice	F	1988-12-30	Daughter
123456789	Elizabeth	F	1967-05-05	Spouse

In total you have to come up with 17 different queries i.e. there should exist no query that simultaneously belongs to any 2 batches or

$$\mathcal{Q}_1 \cap \mathcal{Q}_2 = \mathcal{Q}_2 \cap \mathcal{Q}_3 = \mathcal{Q}_3 \cap \mathcal{Q}_1 = \phi$$

Naturally, you would observe that as the number of queries in a batch increase, there would ideally exist overlaps between these queries

Question C

Part 1:

Given 2 queries q_1 and q_2 such that both q_1 and q_2 belong to the same batch $\mathcal{Q}$, we define operators $< op >$s to be **overlapping** or **related** wrt these queries if the $< op >$ operate on the same set of relations and need to access common rows of the relations to produce an output.

Operators can be Unary i.e. they operate on one relation. For example, `SELECT` (σ) and `PROJECT` (Π) or Binary i.e. they operator on 2 relations. For example, `JOIN` ($\bowtie$), `NATURAL_JOIN` ($*$), `UNION` ($\cup$), `INTERSECTION` ($\cap$)

For example, wrt the following queries $q_1 : \Pi_{Name}(R)$, $q_2 : \Pi_{SSN}(R)$ and $q_3 : \sigma_{SSN=12345678}(R)$
or in SQL

```
q1: SELECT Name FROM R
q2: SELECT SSN FROM R
q3: SELECT * FROM R WHERE SSN = 12345678
```

The 3 operators Π_{Name}, Π_{SSN} and $\sigma_{SSN=12345678}$ all operate on the relation R and all access common rows. So, if all 3 queries belong to one batch, it is possible to execute these queries simultaneously by scanning relation R just once. Hence we say the 3 operators are *overlapping* or *related*

In this first part, you have to find these overlapping operators within each of the 3 batches of queries you came up with in question B.

Part 2

In part 2, you have to try to minimize the number of times relations are accessed within a batch of queries. You have to use the overlapping operator information that you derived in the previous part

Question D

Come up with a mechanism to select indices on relations for a batch of queries $\mathcal{Q}$ such that these indices help in executing at least one or more queries from a batch faster.

The format for the answer should look like so

Index Description	Queries it helps
Primary index on attribute SSN on table Employee	$\mathcal{Q}_3$ - q_1, q_4
...	...

Question E

Part 1
What would be the best case scenario for a batch of queries such that all queries in a batch are distinct i.e. no 2 queries in the batch theoretically yield the same output? Give a sample batch of 5 queries to demonstrate.

Part 2
What would be the worst case scenario for a batch of queries? Give a sample batch of 5 queries to demonstrate.

Question F

Explain briefly (2-4 sentences) what you learnt from this group assignment. Also compare how your answers from questions C and D relate to what was asked in question A.

Submission Guidelines

- Deadline - 11:55 PM, 23rd October
- You can submit a handwritten or typed out document (as announced in class)

Link to class recording - link

You will be penalized for plagiarizing online material