

CARNEGIE MELLON UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE
15-415/615 - DATABASE APPLICATIONS
FALOUTSOS AND PAVLO, SPRING 20XX

SAMPLE midterm examination *ANSWER:*

IMPORTANT

- You should have **5 non-empty pages, including this cover**
- **No aids allowed** - only an 8.5 x 11 page (both sides)

For your information:

- **80 minutes** duration (1:30-2:50pm)
- Graded out of **100** points
- Numbers in [square brackets] indicate points.
- Raise your hand, if you need clarifications.

LAST NAME (pls PRINT)	
First Name (pls PRINT)	
'andrew' login (pls PRINT)	

Contents

Q1. Formal Query Languages [20 pts]	2
Q2. SQL [25 pts]	2
Q3. Indexing - trees [20 pts]	3
Q4. File Organization [25 pts]	4
Q5. Sorting [10 pts]	5

Revision : 1.5

Q1. Formal Query Languages [20 pts]

Consider the schema $PERSON(pname, skill)$, $JOB(jobname, skill)$ and the extension in Figure 1. The meaning is that ‘Peter’ knows ‘C’ and ‘excel’, and that a ‘programmer’ position needs ‘C’, ‘Java’, and ‘Unix’. Answer the following queries, using relational algebra.

1. [4 pts] list *pnames* of all people that have at least one of the skills that ‘accountant’ requires.

ANSWER: $\pi_{pname}(PERSON \bowtie \sigma_{jobname=accountant}(JOB))$

2. [8 pts] list the *pnames* of all people that have all the skills that a ‘programmer’ position requires

ANSWER: $(PERSON \div \pi_{skill}(\sigma_{jobname=programmer}(JOB)))$

3. [8 pts] list the *pnames* of all people that can do all the listed jobs, that is, they have all the skills that any job requires

ANSWER: $(PERSON \div \pi_{skill}((JOB)))$

PERSON	pname	skill	JOB	jobname	skill
	Peter	excel		programmer	C
	Peter	C		programmer	Java
	John	math		programmer	Unix
	John	C		accountant	excel
	John	Java		accountant	bookkeeping
	John	Unix		accountant	math

Figure 1: Persons, Jobs and Skills

Q2. SQL [25 pts]

Consider the schema in Figure 2, with governors/rulers (prime ministers, presidents, etc), and information about their government period. Assume that ruler names are unique, and that each ruler ruled without interruption.

1. [5 pts] List the ruler names for the USA, in chronological order (smallest ‘start’, first).

ANSWER:

```
select name
from RULER
where country = 'USA'
order by start
```

2. [8 pts] List the pairs of rulers whose rule intervals overlap - eliminate mirror pairs, and self-pairs. *Hint:* the condition for intersection of intervals is

```
(start2 between start1 and end1) or  
(start1 between start2 and end2)
```

ANSWER:

```
select R1.name, R2.name  
from RULER as R1, RULER as R2  
where R1.name > R2.name  
and ( (R2.start between R1.start and R1.end)  
or  
(R1.start between R2.start and R2.end)  
)
```

3. [6 pts] List the names of rulers whose rule interval is included inside the rule interval of 'M. Thatcher'.

ANSWER:

```
select name  
from RULER as r1, RULER as r2  
where r2.name = 'Thatcher' and  
r1.start >= r2.start and  
r1.end <= r2.end  
You can add one more condition 'r1.name <> r2.name'.
```

4. [6 pts] List the names of rulers and their duration (=length of rule interval), sorted in reverse duration order (longest rule first).

ANSWER:

```
select name, end-start+1 as duration  
from RULER  
order by duration desc
```

Q3. Indexing - trees [20 pts]

Consider a B+ tree of order 2, that is, every node has at most 4 keys and at most 5 pointers. We do the standard key insertion algorithm, that is, one key at a time (i.e., no bulk-loading); and also we *do not* do key redistribution.

1. [6 pts] Draw the tree, after we insert the following keys in an initially empty tree:

1,2,3,4,5,6,7,8

RULER	name	country	start	end
	N. Mandela	S. Africa	1994	1999
	M. Thatcher	UK	1979	1990
	R. Reagan	USA	1980	1988
	W. Clinton	USA	1992	2000
	A. Lincoln	USA	1861	1865
	I. Gandhi	India	1966	1977
	...	...	...	...

Figure 2: Rulers

ANSWER: root: (3,5); leaves: (1,2), (3,4), (5,6,7,8)

We also accept answer: root: (3,6); leaves: (1,2,3), (4,5,6), (7,8)

2. [14 pts] In an initially empty tree, we insert the keys 1,2,3, ..., one at a time. Which key will make the tree to have height = 3?

ANSWER: 13

We also accept answer 17.

Q4. File Organization [25 pts]

Consider a file of supplier records with the schema $SUPPLIER(sname, address, volume)$. Assume the following:

- $N=100,000$ supplier records
- spanning $M=2,000$ blocks (pages)
- $v=100$ distinct volume values (1,2, ..., 100),
- $h=3$ the height of any B-tree index

Preliminary calculations, which will help you later:

- P1 [1 pts]** Assuming equal-size records, how many records fit in a page? (This is the capacity C of the page)

ANSWER: $C = N/M = 50$

- P2 [1 pts]** Assuming uniformity, how many records m match, that is, qualify for an exact match query on volume (e.g., 'volume=52')

ANSWER: $m=N/v = 1,000$

For each combination of file organization and operation, give your estimate for the number of disk (=page=block) accesses. (Ballpark estimates are fine: any answer within 20% of the correct one will get full points).

1. [1 pts] Heap file; sequential scan

ANSWER: $M=2,000$

2. [4 pts] Sorted file on 'volume' ; exact match query on 'volume' (eg., 'volume=10'). Assume binary search.

ANSWER: $\log M=11$ with binary search, plus the pages that the m records occupy: $11 + 1,000/50 = 11+20 = 31$

3. [3 pts] Clustered B+-tree index on 'volume'; exact match query on 'volume'

ANSWER: $h + m/C = 3 + 1,000/50 = 23$

4. [5 pts] Clustered B+-tree index on 'volume'; range query on 'volume' ($10 < \text{volume} \leq 20$)

ANSWER: $h + 10 * m/C = 3 + 10,000/50 = 203$

5. [5 pts] Non-clustered B+-tree index on 'volume'; exact match query on 'volume'

ANSWER: $h + m = 3 + 1,000 = 1,003$

6. [5 pts] Which file organization would you choose, if the majority of the queries are range queries on volume, with occasional insertions, deletions, and sequential scans. Choose one of:

- Heap
- Sorted on volume
- Clustered tree index on volume
- Non-clustered tree index on volume
- Clustered hash index on volume
- Non-clustered hash index on volume

ANSWER: Cluster tree index on volume

Q5. **Sorting [10 pts]**

Consider a file with $N=10,000$ pages. What is the smallest number of buffers B , so that external sorting will need only two passes? (Assume **quicksort** is used for in-memory sorting.)

ANSWER: $B=101$ - one point penalty for $B=100$.

END OF EXAM QUESTIONS _____GOOD LUCK!

ANSWER:

Total number of points: 90; section weights: 200 total count of minutes: 0