COMP207
Database Development

Review of Topics &
Exam Information

- EvaSys:
  - Low response rate so far
- Today: topic walkthrough & possible exam question types

Mock test exam

- There are 2 mock exams (on Vital)
- Each with 20 questions
  - Consider trying to give yourself half the time you will have at the exam, which is one hour for most of you
- Consider using one early in your revision and the other a few days before, so that you can read up on any aspects you might have missed

Exam

- 2 hours, 80% of final module mark
- Time and place:
  - Not known yet, sorry
- Questions:
  - 40 multiple choice questions
  - Full spectrum:
    - Bookwork: definitions or particular pieces of knowledge
    - …to problem solving: solve a problem
- Topics:
  - Most everything covered during this module

Read these slides carefully to be able to answer all questions

- Lecture 2
  - No direct question, but knowledge about SQL is needed for many questions
- Lectures 3-11
  - But not slides 15-18 lecture 9, i.e. the ones on **Nonquiescent Checkpoints**
- Lectures 12-16
  - But not slides 5-12 lecture 14, i.e the ones on **External Memory Mergesort**
- Lectures 17-18
  - (Lecture 19, on **Map Reduce**, is not necessary to read)
- Lectures 20-26
  - But not slides 23 and forward in lecture 20, i.e. the ones on **XMLSchema**
  - Also not slides 13 and forward in lecture 24 and not lecture 25 on **NoSQL databases**
- Lectures 27-29
- This lecture might be helpful as well…

Topics

Transaction Processing

Query Processing

Distributed Databases

Semi-Structured Data

Object-Relational Databases

Data Warehousing & Data Mining

Transaction Processing

- Expect questions on **all aspects but Nonquiescent Checkpoints**
- Some notable aspects (not exhaustive):
  - Basic concepts
    - Transactions: the concept & important properties
    - Schedules: purpose, execution, types & differences between them
  - Related tools
    - Precedence graphs
    - Locking, (strict) two-phase locking
    - Various timestamp-based methods
    - Logging
  - Relevant components of a DBMS and their responsibilities: concurrency control/manager, recovery control/manager

Example Question 1

Suppose an airline reservation system uses a database with a relation Seats(flightID, seat, passenger) to keep track of seat allocations. Assume two transactions T_1 and T_2 execute on the database as follows:

Time	Event
1	T_1 queries the database to make sure that seat '14B' is still available on flight '123'.
2	T_2 queries the database to make sure that seat '14B' is still available on flight '123'.
3	T_1 allocates seat '14B' to 'Anna' on flight '123'.
4	T_2 allocates seat '14B' to 'Ben' on flight '123'.

Assuming the DBMS does not prevent these transactions from executing and there is no system failure or any other interference, which of the following ACID properties would be violated?

☐ **A.** Consistency only

☐ **B.** Atomicity and Consistency

☐ **C.** Atomicity only

☐ **D.** Durability only

☐ **E.** Atomicity, Consistency, and Durability

Example Question 2

Consider the following schedule:

$$r_2(X);\ r_1(Y);\ w_2(X);\ r_3(X);\ w_1(Y);\ w_3(X);\ r_2(Y);\ w_2(Y)$$

This schedule is conflict-equivalent to which of the following serial schedules?

- ☐ **A.** $r_3(X);\ w_3(X);\ r_1(Y);\ w_1(Y);\ r_2(X);\ w_2(X);\ r_2(Y);\ w_2(Y)$
- ☐ **B.** $r_1(Y);\ w_1(Y);\ r_2(X);\ w_2(X);\ r_2(Y);\ w_2(Y);\ r_3(X);\ w_3(X)$
- ☐ **C.** $r_3(X);\ w_3(X);\ r_2(X);\ w_2(X);\ r_2(Y);\ w_2(Y);\ r_1(Y);\ w_1(Y)$
- ☐ **D.** $r_2(X);\ w_2(X);\ r_2(Y);\ w_2(Y);\ r_3(X);\ w_3(X);\ r_1(Y);\ w_1(Y)$
- ☐ **E.** $r_2(X);\ w_2(X);\ r_2(Y);\ w_2(Y);\ r_1(Y);\ w_1(Y);\ r_3(X);\ w_3(X)$

Example Question 3

A simple checkpoint is:

☐ **A.** A time point when all transactions commit.

☐ **B.** A time point when all transactions roll back.

☐ **C.** A time point when the DBMS is tested for efficiency

☐ **D.** Point of synchronisation between database and log file. All buffers are flushed to secondary storage

☐ **E.** Point of synchronisation between database and log file. All buffers are deleted.

Other things that may help

- Reviewing the following might also help:
  - Lab exercises for weeks 3-6
  - The first assessment
  - The mock exams

Query Processing

- All aspects are relevant (except for External Memory Mergesort)
- A selection:
  - Transformation of SQL into relational algebra
  - Algorithms for executing selections, projections, joins & their running times / required number of disk accesses
  - Indexes: what these are, how they help
  - B+ trees: construction, lookups, insertions, deletions
  - Query optimisation
    - How does it work?
    - Initial query plans
    - Calculation of blocks needed to store a relation

Example Question Type 1

- Possible type of question:
  - Given a particular B+ tree. What happens when we insert or delete a particular value-pointer pair into the tree? (See mock exam 1 for an example)
  - Or: how many disk blocks need to be accessed when performing a certain task with B+ trees

Example Question Type 2

- Given:
  - A query plan
  - Number of blocks of each input relation
  - Number of distinct values per column per relation
- Task:
  - Estimate the size of some/all intermediate results of the query plan

Other things that may help

- Reviewing the following might also help:
  - Lab exercises for weeks 7-9
  - The second assessment
  - The mock exams

Distributed Databases

- Expect questions on all aspects of distributed databases, including:
  - What's a distributed database?
  - Techniques: fragmentation & replication
  - Forms of transparency (e.g., replication transparency)
  - Query execution in distributed databases
    - E.g., what is a semi-join and where is it useful?
  - Two-phase and three-phase commit
- No questions on Map-Reduce

Example Question 1

An organisation uses a distributed database over two sites, A and B:

- Site A holds a relation $Passenger(passenger\_id, first\_name, last\_name)$. Values for each of the attributes, i.e. $passenger\_id$, $first\_name$ and $last\_name$ require 30 byte each.
- Site B holds a relation $Flights(flight\_id, time, seat, passenger\_id)$.

  Each value of attributes $flight\_id$ and $passenger\_id$ require 30 byte, each value of $seat$ require 15 byte, and each value for attribute $time$ requires 25 byte.

Assume the following:

- $|\pi_{\text{passenger\_id}}(\text{Flights})| = 100000$
- $|\pi_{\text{passenger\_id}}(\sigma_{\text{time='10/12/2018 at 10:00'}}(\text{Flights}))| = 1000$
- $|\text{Passenger} \ltimes \sigma_{\text{time='10/12/2018 at 10:00'}}(\text{Flights})| = 1000$
- $|\pi_{\text{flight\_id}}(\sigma_{\text{time='10/12/2018 at 10:00'}}(\text{Flights}) \ltimes \text{Passenger})| = 10$
- $|\text{Passenger}| = 100000$

To execute the query $\pi_{\text{first\_name,last\_name,flight\_id,time}}(\text{Passenger} \bowtie \sigma_{\text{time='10/12/2018 at 10:00'}}(\text{Flights}))$ at site B, how many bytes have to be transferred between A and B at a minimum?

☐ **A.** $300 + 600 = 90$ bytes

☐ **B.** $30000 + 60000 = 90000$ bytes

☐ **C.** $30000 + 90000 = 120000$ bytes

☐ **D.** $100000 + 60000 = 160000$ bytes

☐ **E.** 9000000 bytes

Example Question 2

What does fragmentation transparency refer to?

- ☐ **A.** The database may not use fragmentation
- ☐ **B.** The database must use fragmentation
- ☐ **C.** How the database is divided up over a distributed database does not matter for how to write queries for it
- ☐ **D.** Certain relations in the database are invisible
- ☐ **E.** None of the above

Other things that may help

- Reviewing the following might also help:
  - The mock exams

Semi-Structured Data & Object/Object-Relational Databases

- Expect questions on XML, XPath, Xquery, object databases, and object-relational databases
  - Characteristics of XML documents
  - Construction & evaluation of XPath/XQuery queries
  - Object-databases and impedance mismatch
  - Object-relational database features
  - **No questions on XMLSchema**
- No questions on NoSQL databases specifically, but note that some NoSQL database systems
  - store semi-structured data (there will be questions on this)
  - use techniques from relational databases and distributed databases (there will be questions on these as well)

Example Question

Given the following XQuery, what is the equivalent SQL statement?

```
FOR $v IN $doc//dvd
WHERE $v/year = 2015
RETURN $v/title
```

☐ **A.** `SELECT title FROM dvd WHERE year = 2015`

☐ **B.** `SELECT dvd FROM title WHERE year = 2015`

☐ **C.** `SELECT title, year FROM dvd WHERE year = 2015`

☐ **D.** `SELECT year FROM dvd WHERE title = 2015`

☐ **E.** `SELECT title FROM year WHERE year = 2015`

Another Question Types

- The exam paper could contain an XML document and ask you to
  - Tell what a given XPath/XQuery does on the document
  - Select an XPath/XQuery that performs a given task on the XML document
- Questions about how to define the structure of an XML document etc.

Other things that may help

- Reviewing the following might also help:
  - Labs week 10-11
  - The mock exams

Data Warehousing & Data Mining

- Expect questions on
  - Data warehouses, OLAP/OLTP
  - Applications/goals of data mining
    - Prediction/predictive modelling, segmentation, link analysis, etc.
  - Types of discovered knowledge
  - Frequent-itemset mining
- Possible question types:
  - General questions about data warehousing
  - General questions about data mining
  - **Questions on frequent-itemset mining: see exercises from lectures 28-29**

Other things that may help

- Reviewing the following might also help:
  - Labs week 12 (this week)
  - The mock exams

Good luck in the exams!