Lecture Notes Week 13

25 May 2021

INFS3200 Advanced Database Systems

Semester 1, 2021

Course Review & Preparation of Final Exam

Professor Xue Li

The University Of Queensland
AUSTRALIA

Notes on SETutor & SECaTs Deadlines

- The deadline of the **SETutor** survey is extended to **11.59pm Sunday 30 May** (previously scheduled to close at 11.59pm Sunday 23 May 2021).
- For connectivity issues, please refresh the browser or try opening the survey portal at https://eval.uq.edu.au/ (UQ s-number login required) in a different browser, (e.g., *Chrome, Firefox, or Safari* on a desktop, or an *Android* or *iPhone mobile browser.* You may also need to clear the cache using "*Ctrl + F5" (Windows), "Ctrl + Shift + R" (Linux)*, or "*Command + Shift + R" (Mac).*
- The **SECaTs** (closes at **11.59pm Friday 04 June**).

+ Course Objectives

- "To provide an understanding of the issues involved in designing and implementing a **large scale information system**, beyond the RDBMS".
- "To equip the students with sufficient conceptual and practical knowledge, to be able to recognise the challenges, analyse the appropriateness of the technology and understand the design and implementation complexities."

After successfully completing this course **you should be able to**:

1. Distinguish and classify the main challenges and architectures **in large scale database** systems including centralised, distributed and multi-databases.
2. Develop insights into issues and challenges in **data integration** including **data quality** control, data cleansing, and design and construction of **data warehouses**.
3. Relate your conceptual and practical knowledge in advanced database systems to past knowledge as well as **emerging trends** including implications of **big data**.
4. Understand and analyse post-relational database models and advanced issues related to data integration, **data quality,** and **data privacy**.

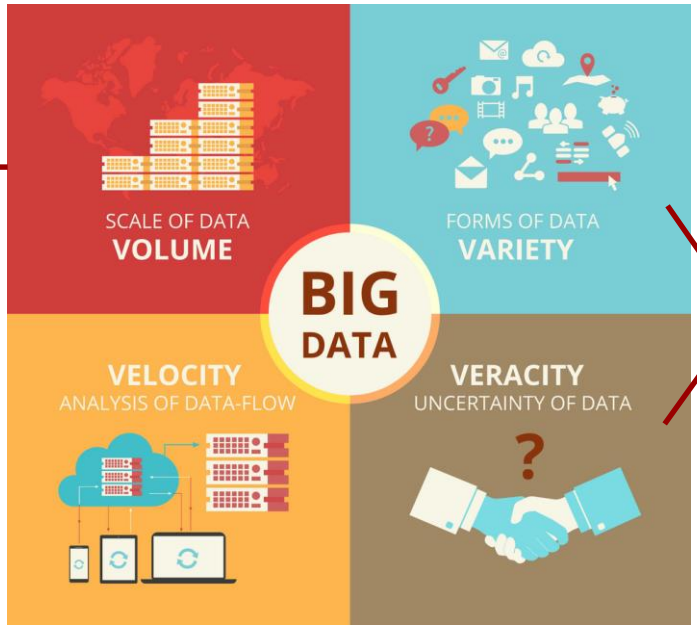
- EPC of INFS3200

https://course-profiles.uq.edu.au/student_section_loader/print/112166?print_section_1=1&print_section_2=1&print_section_3=1&print_section_4=1&print_section_5=1&print_section_6=1&print_section_7=1

+ INFS3200 Course Structure

Distributed DB

- Design
- Query Processing
- Transaction Management



Data Preprocessing

- Integration & Linkage
- Quality Management

Fundamental Issues

- Data Privacy

Data Warehouse

- Design
- Implementation

Advanced Issues

- Big Data Applications
- Network Science

+ INFS3200 Modules

- Distributed Databases
  - DDB Concepts
  - DDB Design
  - DDB Query Processing
  - DDB Transaction Management
- Data Warehousing
  - DW Design
  - DW Implementation
- Data Integration and Linkage
- Data Quality Management
- Data Privacy
- Advanced Database Applications

+ Final Exam Information

- **Exam Time**
  - Offline on Campus invigilated (Option 1)
    - 120+10 minutes, closed-book
  - Online ProctorU invigilated (Option 2)
    - 120+30 minutes, closed-book Blackboard Test
  - The same questions for both Online/Offline Invigilated Exam.
  - No Calculators
  - Short Answers Only
  - Check SI-net for the official exam time & other information
- **Exam Scope (60%)**
  - Covering the entire course materials
  - Focusing on the Lecture Notes and Tutorial Questions
    - Examples in tutorials and lectures are important
    - No practicals
- **Survive Guide**
  - Reasoning + Questioning
  - Tutorial Questions
  - Past Exam Papers
  - Questions/Samples discussed during the lectures

+ Final Exam Preparation

- The total number of individual questions: 29 (28+1)
  - There are no multiple-choice questions
  - Mainly for Brief discussions
  - Short answers
  - No graph/tree drawing
  - Problem-solving + simple calculation + concepts
- The same format as past exams (2018-2020)
  - Past exam questions could re-appear
  - You can pass the exam if you have tried past exam questions without much difficulties
- There will be online ZOOM consultations to be announced (TBA on Course Website)

Final Exam Content:
There are eight parts to be examined (60%)

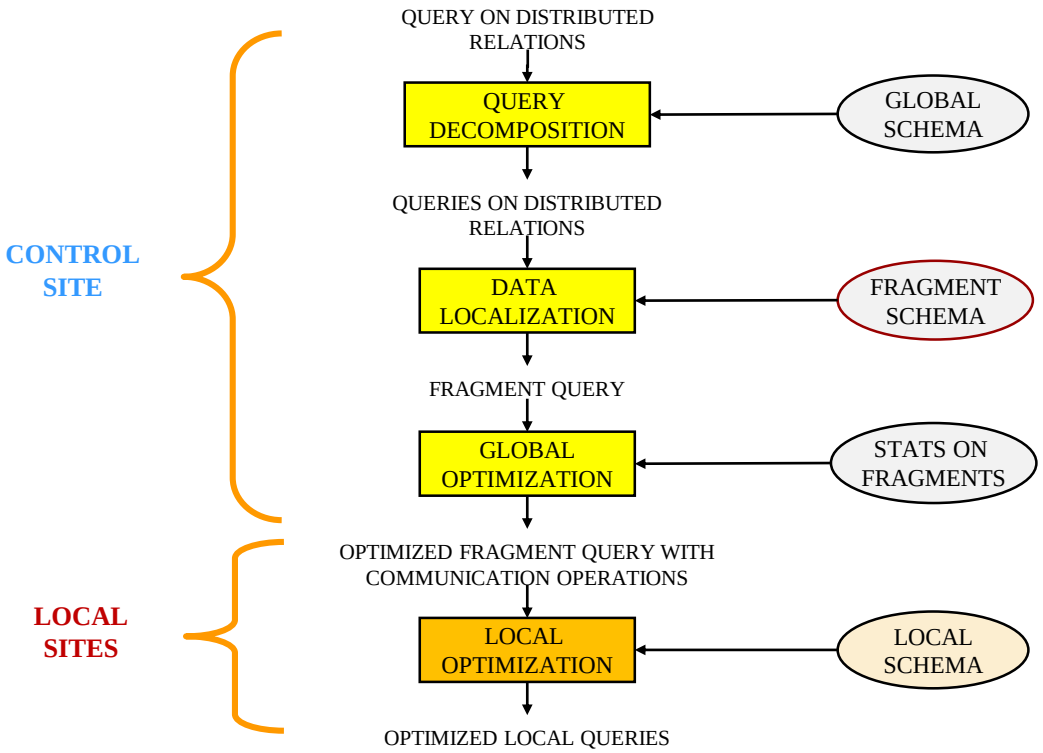
- Part 1: Distributed Databases (12 Marks)
- Part 2: Distributed Transaction Management (8 Marks)
- Part 3: Data Warehouse Design (9 Marks)
- Part 4: Data Warehouse Implementation (5 Marks)
- Part 5: Database Integration and Data Linkage (10 Marks)
- Part 6: Data Quality Management (6 Marks)
- Part 7: Data Privacy (4 Marks)
- Part 8: Advanced Topics (6 Marks)

+ Part 1: Distributed Databases (12%)

- Fragmentation, replication and allocation
  - Horizontal, vertical and mixed fragmentation
  - Primary and derived horizontal fragmentation
- Fragmentation property (design criteria)
  - Completeness
  - Disjointness
  - Reconstruction
- Tools for fragmentation
  - *Minterm* predicates and their application in horizontal fragmentation process

+ Distributed DB Query Processing

- Optimization objectives
- Simplification
- *Semi-join* operation with multiple sites and its detailed process



+ Part 2: Distributed Transaction Management (8 Marks)

- TM in RDBMS
  - *ACID* properties
  - Conflict operations
  - Serial schedule and serializability
  - Interleaving transactions, *WAL, Checkpoint, Rollback, Abort, Commit, Redo, Undo*
  - 2PL
- DDB TM
  - Data replication
    - Synchronous replication: voting, read-any-write-all
    - Asynchronous replication: primary site, P2P
  - Multi-site transactions
    - 2PC

+ Part 3: Data Warehouse Design (9 Marks)

- Motivation and Requirements
  - Volume & velocity vs. value
  - RDBMS vs. DW
- Multidimensional Data Model
  - Facts and dimensions
  - *Star schema* vs. *snowflake schema*
  - *Fact Constellation* (a variation of Star Schema)
  - The differences between these types of schemas
- OLAP Operations
  - Main differences between OLTP and OLAP?
  - Example and the meanings of *Drill-down/Roll-up, Slicing/Dicing, Pivoting, CUBE queries*
  - What are the results of OLAP operations?
  - Applicability of OLAP operations

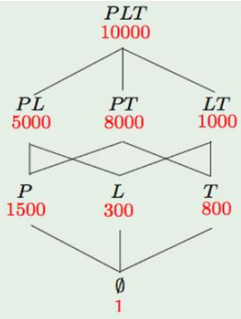
+ Part 4: Data Warehouse Implementation (5 Marks)

- RDBMS vs. Data Warehouse
- Indexing
  - Bitmap indexing
  - Join indexing
- View materialization
  - Query with materialized views
  - Concepts of *Cuboid, Lattice*
  - Benefit of materialized views
    - Number of possible views to be materialized
    - The number of *cuboids* that can be created by using GROUP BY clause.
    - Which queries are benefitted from a certain view
  - Greedy algorithm
    - Calculating the cost and benefit based on the Lattice of Materialized Views

View Materialization Problem:

Given the **fact table T**, the $T_G=\{T_{G1}, T_{G2}, \ldots, T_{Gt}\}$, an integer *k*, $k \le t$, **find a set *S* of *k* cuboids** ($S \subseteq T_G$) to be materialized **with the largest *benefit(S).***

Q	cost
{PLT}	10000
{PL}	10000
{PT}	10000
{LT}	1000
{P}	10000
{L}	1000
{T}	1000
∅	1000



How to study a course effectively: the **Line of Reasoning** (LoR) for a concept

e.g., why do we need "materialised views"?

- **Data analytics** uses **OLAP** operations.
- **OLAP** operations use SQL Statements with **GROUP_BY** Clauses
- **GROUP_BY** is an **aggregation operation** which is very **costly** and needs to be applied on the data which is **pre-calculated.**
- **Materialised views** have the **pre-calculated aggregation** operations that are the **pre-calculated GROUP_BY** clauses.
- For the given **n** attributes of a table, there are 2^n possible ways to calculate the **views for the materialization**.
- The different **materialised views** are denoted by **cuboids**.
- An **OLAP qu**ery which involves a **GROUP_BY** clause can be executed over a specific **cuboid**.
- So, the effective OLAP operations can be supported by **materialised views** in **data warehouses**.

Part 5: Data Integration and Linkage (10 Marks)

- Motivation and Requirements
  - Schema heterogeneity
  - Data type heterogeneity
  - Value heterogeneity
  - Semantic heterogeneity
- Federated Database (FDB), Multi-Databases (MDB) and Interoperable Information Systems
- Schema Mapping
  - How to define and create Global Views given local views?
- Database integration
  - Top-down vs bottom-up approaches
  - DDB/DW, to FDB, MDB and interoperable systems
  - Views are extensively used in DB integration
- Data linkage
  - The problem of data linkage (i.e., computing the same real-world entity)
  - Distance Measures: *Edit distance* (dynamic programming), *Q-gram (n-gram) and Jaccard coefficient, TF/IDF and cosine similarity, Numeric similarity, Phonetic similarity*
  - The applicability of different kind of similarity computations

Dynamic programming for calculating the edit distance

		j	o	h	n
	0	1	2	3	4
j	1	0	1	2	3
h	2	1	1	1	2
n	3	2	2	2	1

+ Part 6: Data Quality Management (6 Marks)

- Data Quality Dimensions
  - What is Data?
  - Governance vs. Management
  - Concepts of data quality dimensions (Integrity, Accuracy, Completeness, Currency, etc)
- Four Basic Steps of Data Governance
  - Recognize the ***problem***
  - Measure its ***costs***
  - Devise ***strategy*** for improvement
  - Aim towards *data governance* ***maturity***

Part 7: Data Privacy (4 Marks)

- Data Privacy Preservation/Protection Methods
  - k-anonymity
  - l-diversity
  - t-closeness
  - Differential privacy
- Advantages and Limitations of the above methods
  - Advantages of Differential Privacy approach over the k-anonymity method.
  - What changes will be performed on a database in order to apply the above-mentioned privacy preservation methods?
- Data Privacy Attacking Methods
  - Membership inference
  - Model inversion attack
  - Statistical attack
  - Poisoning attack
  - Evasion attack
  - Model poisoning attack

+ Part 8: Advanced Topics (6 Marks)

- Big Data is of Three Utilities:
  - Connecting Dots – "*From small to big*"
  - Discovering Specifics – "*From big to small"*
  - Data Inferencing – "*Knowing unknown"*
- Network Science is a Foundation of Big Data Studies
  - Scale-Free Networks (*Structure is independent from the size*)
  - Different types of networks (*Centralised, Decentralised, Distributed)*
  - Curse of Dimensionality
  - Computations of Network *Centrality*, *Modularity*, *Reachability*, etc
- Row Storage vs Column Storage in Database Systems
  - SQL vs. NOSQL
  - OLTP with row storage and OLAP with column storage
  - Frequent update operations with row storage and read-only data with column storage

+ A Note on Course Requirements

- Students must receive a passing grade on the **final exam** in order to pass this course (i.e., achieve at least half marks in the final exam paper, **0.5 * 60%**).
- If you fail the final exam, your **final mark** will be capped at 49 and your final grade will be capped at Grade 3.
- Total achieved percentage will be **rounded up** before grade cut-offs are applied.

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Thanks, and All the Best!