INFS3200/7907 Advanced Database Systems

Semester 2, 2020

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Course Review

Dr Pingfu Chao

+ INFS3200/7907 Course Structure

Distributed DB

- Design
- Query Processing
- Transaction Management

Data Preprocessing

- Integration
- Quality Management

Fundamental Issues

- Data Privacy

Data Warehouse

- Design
- Implementation

Beyond Relational Data Type

- Enhanced Data Models

+ INFS3200/7907 Modules

- M1 – Distributed Databases
  - DDB Concepts
  - DDB Design
  - DDB Query Processing
  - DDB Transaction Management
- M2 – Data Warehousing
  - DW Design
  - DW Implementation
- M3 – Advanced Topics
  - Data Integration and Linkage
  - Data Quality Management
  - Data Privacy
  - Enhanced Data Models

+ 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

+ Final Exam

- Exam Time
  - Thursday, 05/11/2020 8am (Check SI-net for final time & location)
  - Online exam (ProctorU invigilated)
    - 120+30 minutes, **closed-book** Blackboard Test
  - On-campus exam
    - 120 writing time + 10 reading time
  - Calculators with UQ Labels approved
- Exam Scope
  - Focus on the lecture notes and recordings
    - Tutorial questions are highly relevant, no practical questions
  - Short answers only

+ Final Exam

- Exam Specification
  - 8 questions in total, 60 points
    - No graph/tree drawing
    - Problem-solving + simple calculation + concepts
  - Same format as past exams (2016-2020)
    - Guaranteed to pass the exam if you fully understand all past exam questions
  - Course contents
    - Mentioned in course review
      - Black: testable in final exam
      - Grey: non-testable in final exam
    - Not mentioned in course review: may appear as conceptual questions (very rare) for grade differentiation

+ Consultation

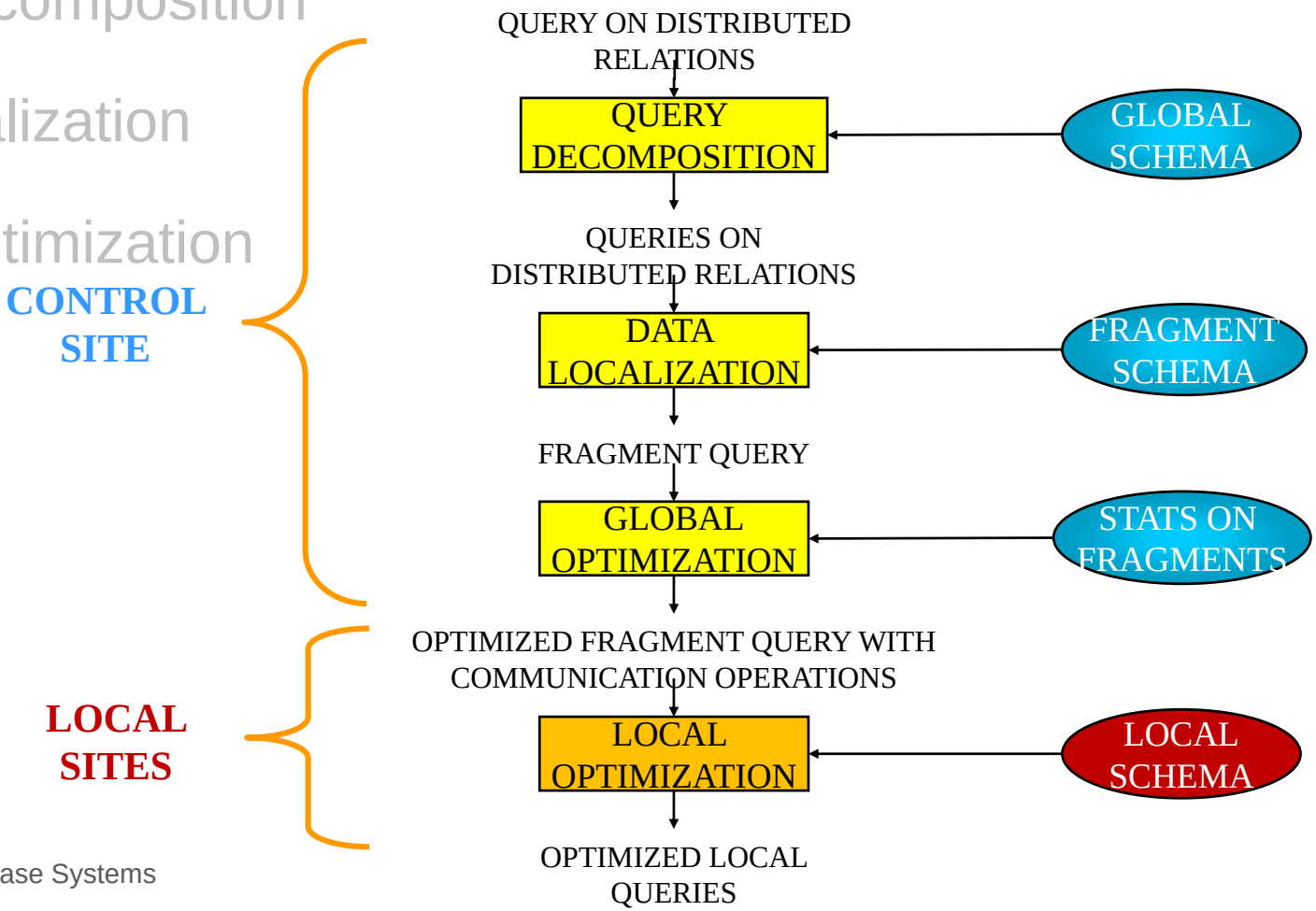
- Two consultation sessions
  - Monday, Nov 2nd, 7pm-8pm
  - Wednesday, Nov 4th, 4pm-5pm
  - Online Zoom sessions. Session ID: 970-070-9405
  - Both sessions have at least 2 tutors attending for live Q&A, no course walkthrough

+ DDB Design

- Fragmentation
  - Horizontal, vertical and mixed fragmentation
  - Primary and derived horizontal fragmentation
  - Fragmentation properties (design criteria)
    - Completeness
    - Disjointness
    - Reconstruction
- Minterm Predicates
- Allocation

DDB Query Processing

- Optimisation Objectives
- Query Decomposition
- Data Localization
- Global Optimization
- Semi-join



+ DDB Transaction Management

- TM in RDBMS
  - ACID
  - Conflict operations
  - Serial schedule and serializability
  - 2PL
- DDB TM
  - Data replication
    - Synchronous replication: voting, read-any-write-all
    - Asynchronous replication: primary site, P2P
  - Multi-site transactions
    - 2PC

+ Data Warehouse Design

- Motivation and Requirements
  - Volume & velocity vs. value
  - RDBMS vs. DW
- Multidimensional Data Model
  - Facts and dimensions
  - Star schema vs. snowflake schema
- OLAP Operations
  - Drill-down/roll-up, slicing/dicing, pivoting, CUBE
  - What is the result of an OLAP operation?
  - What OLAP operations to perform given a task at hand?

+ Data Warehouse implementation

- Indexing
  - Bitmap indexing
  - Join indexing
- View Materialization
  - Query with materialized views
  - Benefit of materialized views
    - Number of possible views to be materialized
    - Which queries can benefit from a certain view
  - Greedy algorithm

+ Data Integration and Linkage

- Database Integration
  - Motivation
  - Concept
  - Challenges
- Federated DB, multi-database and Interoperable Systems
- Schema Mapping
  - How to define and create Global Views given local views?

+ Data Integration and Linkage

- Data Linkage
  - Similarity measures
    - Edit distance (dynamic programming)
    - Q-gram and Jaccard coefficient
    - TF/IDF and cosine similarity
    - Numeric similarity
    - Phonetic similarity
  - Record matching
  - Group matching
  - Efficiency issues

		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

+ Data Quality

- Data Quality Dimensions
  - Concept only
- Data Quality Management
  - Different perspectives
  - Data Governance

+ Data Privacy

- Concept of Data Privacy
  - Data protection vs Data utility
  - Sensitive information
- Existing Solutions (concept only)
  - K-anonymity
  - L-diversity
  - T-closeness
  - Differential privacy
    - DP for data releasing

+ Enhanced Data Models

- Limitations of the Relational Model
- Spatial Databases
  - Using relational table
    - Data representation
    - Query processing
    - Problems
  - Using spatial data types
    - Spatial relationships/operations/query language
    - Spatial indexing
      - B-tree
      - Space-filling curves
      - R-tree
- Other Types of Data Models
- Object-Oriented Databases

Thanks, and All the Best!