THE UNIVERSITY OF WESTERN AUSTRALIA

Data Warehousing and Data Mining

Lecture 12 Unit Review

CITS3401
CITS5504

Dr. Sirui Li

Computer Science and Software Engineering

School of Physics, Mathematics and Computing

THE UNIVERSITY OF WESTERN AUSTRALIA

DESK No. | | |

FAMILY NAME: _____

GIVEN NAMES: _____

SIGNATURE: _____

STUDENT NUMBER: | | | | | | | | |

Semester 1, 2024 EXAMINATIONS

School of Physics, Mathematics and Computing

CITS3401/CITS5504
Data Warehousing
Examination Duration: 2 hours

This is an examination WITH permitted materials

Provided by the University
1 x 18 Page Answer Booklet

Supplied by the Student
1 x double-sided A4 page of notes (printed or handwritten)

Calculator
UWA approved calculators with stickers are permitted

Instructions to Students
This exam contains 5 questions worth 20 marks each. Total mark: 100

The exam contributes to 50% of the total assessments of this unit.

Candidates must attempt ALL questions.

All questions should be answered in the provided Answer Booklet. Clearly indicate which question you are answering. Answers in the Question Booklet will NOT be considered.

Examination candidates may only bring authorised materials into the examination room. If you are found with unauthorised material, disciplinary action will be taken against you. This action may result in you being deprived of any credit for this examination or even, in some cases, for the whole unit. This will apply regardless of whether the material has been used at the time it is found. Any candidate who has brought unauthorised material into the examination room should declare it to the supervisor immediately. Candidates who are uncertain whether any material is authorised should ask the supervisor for clarification. Question papers and answer booklets must not be removed from the examination room.

Examination Cover Sheet

THE UNIVERSITY OF WESTERN AUSTRALIA

Face to Face (in person) close book exam.

2 hours, 50% of total assessment, total of 100 marks.

Attempt all questions

Answer in the provided Answer Booklet.

Better not to use pencils.

- 1 double-sided A4 page of notes (either printed or handwritten) is permitted
- UWA approved calculator is permitted

Examination Structure

- **Given a business scenario that describes the business and data needs, we have**
- **A total of 5 questions (total 100 marks)**
  - Q1: Data Warehouse Design (6+5+6+3 marks)
  - Q2: Dimension Modelling (5+6+5+4 marks)
  - Q3: Multi-dimensional cube and Association Rule Mining (4+4+4+8 marks)
  - Q4: Graph Database Design (5+5+10 marks)
  - Q5: Graph Data Science and Advanced Queries (15+5 marks)

Examination Structure (Continued)

- **A total of 5 questions (total 100 marks)**
  - **No calculations** only simple counting
  - **No** Python, Atoti and Docker
  - There are cypher coding required
  - Key Concepts
    - OLTP and OLAP, fact tables, dimension tables, business queries, data warehouse schemas, concept hierarchies, starnet
    - slowly changing dimensions, types of cells and cubes.
    - data cube, data mart, meta data.
    - Different types of patterns, association rules, Apriori algorithm, Measures (support, lift, confidence).
    - Graph data modelling for the given scenario (when to use node and when to use property)
    - Write cypher queries
    - Understand how to use path finding algorithms to solve search problems.

- **Why Data Warehouse and Data Mining?**
  - **Explosive Growth of Data: from terabytes to petabytes**
  - **We are drowning in data, but starving for knowledge.**
- **What is Data Warehouse and Data Mining?**
  - A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile, collection of data in support of management's decision-making.
  - Extraction of interesting (**non-trivial**, **implicit**, **previously unknown** and **potentially useful**) patterns or knowledge from huge amount of data.
- **OLAP and OLTP**
  - OLTP: Major task of traditional relational DBMS
  - OLAP: Major task of data warehouse system

Lecture 2—Data Warehouse Design

THE UNIVERSITY OF WESTERN AUSTRALIA

- Storing Data in Data Warehouse
- Fact Tables and Dimension Tables
- Schema of a Data Warehouse
  - Star, Snowflakes, Fact Constellations (Galaxy)
- OLAP Operations
  - Roll up, Drill down, Slice & Dice, Pivot

Lecture 3 Data Cube Technologies

THE UNIVERSITY OF WESTERN AUSTRALIA

- **Hierarchies, StarNet with footprint**
- **Data Warehouse design**
- **Data Cube**
  - Cuboids
  - Types of Cells
  - Types of Cubes
- **Answering Queries with Data Cube**

Lecture 4: Dimension Modelling – Practical Concerns

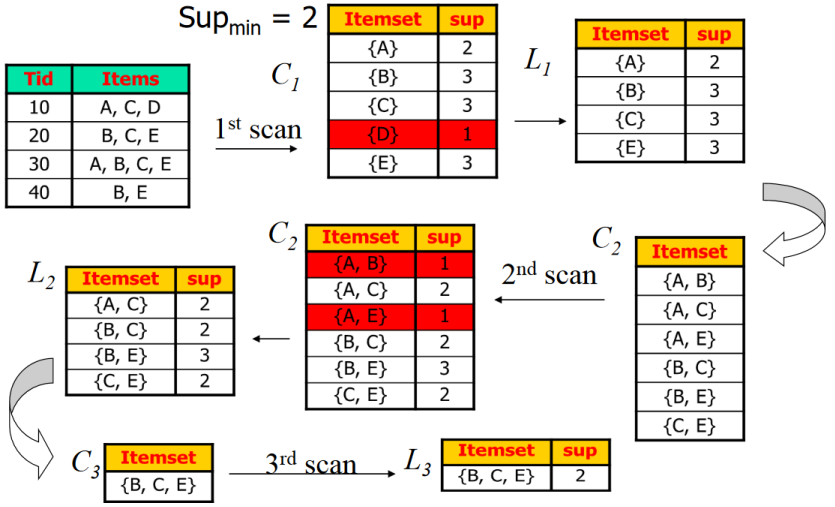
THE UNIVERSITY OF WESTERN AUSTRALIA

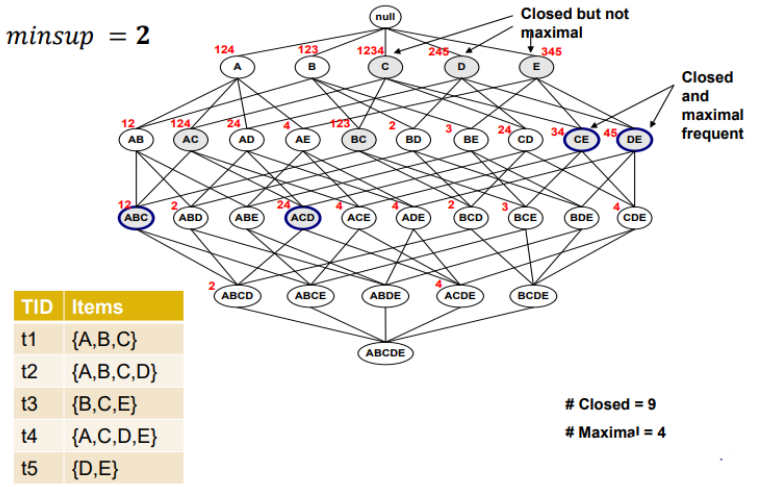
- **Dimension Topics**
  - How many dimensions?
  - Date/Time Dimensions
  - Surrogate Keys
- **Fact Topics**
  - Semi-additive facts
  - "Factless" fact tables
- **Slowly Changing Dimensions**
  - Overwrite history, preserve history, hybrid schemes
- **More dimension topics**
  - Dimension roles
  - Junk dimension
- **More fact topics**
  - Multiple currencies

Lecture 5: Association Rule Mining

- **Concepts**
  - Support
  - Confidence
  - Lift

- **Frequent itemsets and association rules.**
  - Frequent Patterns, Closed Patterns and Max-Patterns
  - The Apriori Algorithm
  - How to generate the association rules





Lecture 6: Meta Data and Data Mart

THE UNIVERSITY OF WESTERN AUSTRALIA

- What's Meta Data?
  - Store descriptive information about data model to store and share
  - e.g. Data Definition Metadata
- Why meta data are useful
  - Operational
  - Technical
  - Business
- Data Mard and Data Lake

- Why graph and graph databases?
- Graphs are excellent tools for modelling highly-connected data.
- Graph technology can be separated into graph databases and graph compute engines.
- In the property graph model, graphs consist of nodes, relationships, properties, and labels.

Not examinable

- **Relational DB vs. Graph DB**
  - Graph DBs, excel at dealing with highly connected data
  - Graph DBs do not need to store null values
  - Graph DBs are agile, meaning that new properties, new node and relation types can be added on the fly, without designing a schema apriori.
  - Graphs DBs are better at handing recursive queries and queries across multiple entities.
- **Graph modelling**
  - On whiteboard (using the Arrows tool)
  - Directly from relational tables (table to entity types, rows to nodes, column names to properties, cells to property values, a join or a foreign key is a relationship)
  - Decide on whether the information should be modelled as a node or a property value.

- **Demonstrate the process of graph data modelling for a given scenario.**

Lecture 9: Introduction to Cypher

THE UNIVERSITY OF WESTERN AUSTRALIA

- Cypher
  - Cypher is a query language for Neo4j that allows us to
  - Create, Read, Update and Delete data.
  - Clauses

- **Aggregation and other useful Cypher clauses**
- **Awesome Procedures on Cypher (APOC)**
  - Importing data in APOC
  - APOC text functions
  - Path Expansion in Cypher & APOC
  - Virtual Graph

- Have a basic understanding of what APOC are capable of that are not easily achievable in Cypher
- Good grasp of aggregation, path expansion queries and be able to apply to write queries

Lecture 11: Path finding algorithms

- **Why graph projections**
- **Breath/Depth First Search**
- **A* Search**
- **Minimum Spanning Tree**
- **Random Walk**

- Have a good understanding of the search algorithms
- **No** need to remember Cypher syntax.