STADVDB Release Date: Sep 15-16, 2025

MCO1: Query Processing in a Data Warehouse Due Date: Oct 13-18, 2025

Project Description

The course project on *Query Processing in a Data Warehouse* is a venue for students to integrate their knowledge and skills to achieve the following learning competencies:

1. **Build a data warehouse**. Design the data warehouse following either the *Star* or the *Snowflake* schema; Include restructuring of tables and denormalization as needed.
2. **Setup an ETL script** to load the data from source to the warehouse. Perform data wrangling tasks, specifically, cleaning, splitting, merging, aggregating, and transforming when necessary.
3. **Develop an OLAP application**. Formulate complex and parameterized queries to generate analytical reports that require the use of OLAP operations, namely roll-up, drill-down, slice and dice. Design a web-based interface to display the generated reports in an organized manner.
4. **Apply query processing and optimization strategies** and evaluate their impact on the performance of the application, specifically, the speed in generating reports from complex queries.

Students will use a public dataset (to be indicated by your teacher) to design a dimensional model for their data warehouse; extract, transform and load the data from this dataset to their model using an ETL script; and formulate queries showing roll-up, drill-down, dice and slice operations. Strategies in improving query processing speed through proper query formulation, the use of indexes, and database restructuring will be evaluated to determine how these affect the response time of the application.

All decisions involved in building the data warehouse, formulating complex queries that utilize OLAP operations, and optimizing query processing speed, as well as the method, results and findings from validating that the generated reports are correct and evaluating the query performance will be documented in a technical report to be presented as part of the output.

Methodology

Students are to form teams with **3 - 4 members** who should be under the same instructor. Each team will work on a public dataset (to be indicated by your teacher). To proceed, each team should do the following:

Step 1. Download and import the dataset to a local MySQL or Postgres Database

Teams must download the source dataset to their local machine and replicate it to their local environment as the SOURCE DATABASE.

Step 2. Build the Data Warehouse

Design the dimensional model using a *star* or *snowflake* schema containing **at least one (1) fact table** and **three (3) dimension tables**. The schema is to be implemented in MySQL or Postgres.

Step 3. Setup an ETL Script

Code an ETL Script to extract, transform and load data from your source MySQL or Postgres database to your data warehouse. Perform data wrangling tasks as needed, specifically, cleaning, splitting, merging, aggregating, fixing the null values, changing the data types, among others.

Step 4. OLAP Application

Formulate complex queries that utilize OLAP operations (roll-up, drill-down, slice, dice, pivot) to generate relevant analytical reports (as opposed to transaction reports in CCINFOM). The queries should show evidence of appreciation for and full understanding of advanced SQL constructs and OLAP.

Aside from the default OLAP operations, find and study one statistical technique that is applicable to the data set and use SQL to implement said technique with respect to the dataset provided. Examples of which are standard deviation, collaborative filtering, Pearson's correlation, chi-square test, etc.

The queries are to be packaged in a web-based application to be developed and deployed in your chosen platform. The interface must be organized such that users can specify query parameters or filters and can easily view the reports at varying levels of granularity and details.

Step 5. Prepare and Execute the Test Script (Functional Testing)

Perform functional testing on your queries by running each query multiple times on different input values to check that the resulting reports are correct. Remember from your Software Engineering class that test scripts should be designed to facilitate an efficient and effective testing process.

Step 6. Optimize the Performance of the Queries (Performance Testing)

To prepare for Step 6, review your answers to ***Exercise 3 Query Optimization*** which will comprise part of your *MCO1 Technical Report*.

1) Run each of your queries multiple times on different input values, database structure, and/or hardware setup. Record the execution time.
2) Evaluate the performance of the queries. Identify and explain the performance speed.
3) Apply optimization strategies and validate if there are any performance gains. Strategies include:
   - Creating secondary indexes for candidate keys and/or frequently queried column(s);
   - Rewriting query statements based on properties of relational algebra operations; and
   - Redesigning the tables, e.g., splitting tables with > 20 columns to multiple tables, or denormalizing the tables.

Document the query optimization strategies that you employed. Compare the execution times of the queries based on the original implementation (database design, SQL statements) and the implementation of optimization strategies (use of indexes, reformulated SQL statements, database restructuring). Analyze and explain the causes for the improvement or non-improvement in performance and include these in your technical report.

Step 7. Prepare the Final Report

From all the data you collected by performing Steps 1 - 6, write your Technical Report using the prescribed template.

Technical Report

Use the given template and follow the outline provided below to prepare your Technical Report.

1. Introduction.
   - Give a brief description of the dataset, and an overview of your data warehouse, the OLAP application that you built, their intended usage, and target users.
   - Cite related literature for your definition of terms.

2. Data Warehouse. Discuss the following:
   - Present your dimensional model.
   - What are the contents of your fact table?
   - What are your dimensions? Describe the hierarchy per dimension.
   - Justify your choice of dimensions and facts.
   - What issues did you encounter in your model / schema design? How did you address these?
   - Use figures and tables accordingly.
   - Cite related literature to support your design decisions.

3. ETL Script. Discuss the following:
   - Discuss issues pertaining to loading the volume of data to the warehouse
   - Describe the process of extraction, transformation and loading.
     - For extraction, describe your data sources and their relevance to the model.
     - For transformation, present the rules or functions you applied on the extracted data to perform data wrangling as needed. What made you decide to apply these rules and functions?
     - For loading, discuss any additional constraints in the database schema.
   - Present the issues you encountered during ETL. How did you address these issues?
   - Use figures and tables accordingly.
   - Cite related literature to support your design decisions.

4. OLAP Application. Discuss the following:
   - State the main purpose of your application.
   - What decision-making or analytical task(s) is the application intended for?
   - Discuss each of your analytical reports:
     - Give the corresponding SQL statement.
     - What advanced SQL constructs were utilized and for what purpose?
     - How were OLAP operations (roll-up, drill-down, dice, slice) utilized in the queries?
   - Provide relevant sample output to complement and enrich your discussion.
   - Use figures and tables accordingly.
   - Cite related literature to support your design decisions.

5. Query Processing and Optimization. For each of your queries:
   - Reference *Exercise 3 Query Optimization* that contains guide questions to help you understand a DBMS's query processing and optimization strategies

- What is query optimization? Why is it necessary?
- What query optimization strategies are available?
- What strategies did you employ to try to improve the performance of the query? Why these strategies?
  - Correct database design and normalization
  - Use of indexes – primary and secondary indexes
  - Query restructuring
  - Optimize at the hardware level
- Provide illustrations – diagrams, SQL codes, schema – to illustrate the optimization strategy.
- Cite related literature to support your design decisions.

6. Results and Analysis. Discuss your validation process:

   - What are the different types of testing that you conducted?
   - For each type, discuss the rationale or purpose for doing the test (what do you intend to validate by doing this test?), the test process (how did you conduct the test?), test data (what data did you use?), and the test results (what are your results?).
   - Guide questions to discuss *Function Testing*:
     - How did you validate the correctness of your ETL script? Describe the procedure and test cases.
     - How did you validate the correctness of your OLAP operations? Describe the procedure and test cases.
   - Guide questions to discuss *Performance testing*.
     - How did you evaluate the performance of your queries? Describe the procedure and test cases.
       - How many times did you run each of the queries?
       - What are your hardware specifications?
       - What is the size of the input data?
       - What is the size, in terms of rows and columns, of the query results?
     - What are the performance results?
       - How did you measure the query performance?
       - How long did the query take to execute for the given test conditions (hardware, input size)?
     - Analyze your results by comparing the execution times of the queries based on the original implementation (database design, SQL statements) and the implementation of optimization strategies (use of indexes, reformulated SQL statements, database restructuring).
       - What affected the performance of your queries? The next questions below can help you write this part of your technical report.
       - The first strategy presented by MySQL to reduce query execution time is to ensure correct design of the database. How does your table structure affect the speed of your database access? Specifically, explain why would "*applications that perform frequent updates often have many tables with few columns (normalized tables), while applications that perform queries over large amounts of data often have few tables with many columns (denormalized tables)*"?
       - Were primary indexes available? Did you use these? How did these affect the query speed?
       - Did you use secondary indexes? For which queries? How did these affect the query speed?
       - Is it recommended to create an index for every possible column used in a query? Cite an instance when the use of indexes does not help improve the execution time of queries.
       - Did you restructure any queries that contain join clauses? How did this affect the query speed?
       - Did you employ any hardware-level optimization? How did these affect the query speed?

- What is the impact of query operations to the performance of your application?

- Cite related literature to support your design decisions.

7. Conclusion.
   - Summarize what you did in this project.
   - After doing this project, what are your learnings that are related to Database concepts (data warehouse, OLAP, ETL, and query processing)? For example:
     - What is the importance of building and maintaining a data warehouse?
     - How does ETL keep your warehouse updated?
     - What is the primary reason for doing OLAP aside from the usual OLTP you learned in CCINFOM?
     - Why is there a need to optimize queries? What strategies are available?
     - Under what condition would you need to create your own indexes aside from those automatically generated by MySQL?
   - What are the relevant contributions of your work and findings to society (end users) and other database developers?

   NOTE: Stating that the project is difficult is not a valid conclusion. The intent of this paper is to share with others what you did – your design decisions, the approaches you employed, the challenges you encountered, how you addressed the challenges, and your findings – as evidence of learning the required concepts.

8. References. Reviewing literature on data warehouse, OLAP, query processing and optimization strategies, and database design (normalization, indexes) should be conducted to help you in doing your project and writing your technical paper.

9. Declarations**.** Academic honesty reflects your true character and personal values. In this part of the paper, please provide the following.

   9.1 **Declaration of Generative AI in Scientific Writing**. "*During the preparation of this work the author(s) used [NAME TOOL / SERVICE] to assist with the following tasks: [LIST OF TASKS]. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the article*."

   9.2 **Record of Contribution**. Indicate the contributions in the software development and paper writing of each member of the team.

Grading

The following criteria will be used in grading the Technical Report:

- Appropriateness of the Dimensional Model
- Completeness of the ETL Process
- Quality of the OLAP queries with evident use cases for drill-down, roll-up, slice, and dice
- Critical application and evaluation of the Query Optimization strategies
- Efficiency and effectiveness of the Testing Process -- Function testing, Performance testing
- Clarity of the discussion of Results and Analysis
- Evidence of critical thinking in analyzing the test results

- Relevance of the Discussion and Conclusion, specifically the learnings derived from the conduct of the project that show an understanding of the key concepts of Data Warehouse, OLAP, and Query Processing and Optimization
- Use of proper citations and references to support the design decisions
- Overall document presentation, e.g., format (title page, page numbers, sections, tables and figures) and language (spelling, choice of words and grammar)

Final Deliverables

The following final deliverables are to be submitted at **date between Oct 13 – 18 and time (to be indicated by your teacher)** through AnimoSpace:

1. Database Schema for the Data Warehouse (filename: *STADVDB-MCO1-GroupN-Schema.xxx* where *N* is your group number and *xxx* is the extension code of the app you are using)
2. Source Code for the OLAP application (filename: *STADVDB-MCO1-GroupN-OLAP.xxx* where *N* is your group number and *xxx* is the extension code of the app you are using)
3. Processing Pipeline of the ETL script (filename: *STADVDB-MCO1-GroupN-ETL.xxx* where *N* is your group number and *xxx* can be *PDF, JPG,* or *PNG*)
4. Technical Report following ACM publication format (filename: *STADVDB-MCO1-GroupN-Report.PDF* where *N* is your group number)

- Late submissions will receive 10 points deduction per day late.
- No submission will be accepted after 18 October 2025.
- Prepare for a 15-to-20 minutes live presentation of your design decisions, validation procedure, and test results and analysis. Specifically, these include the data warehouse and dimensional model, ETL script, OLAP queries and reports, query optimization strategies.
    - Project presentations will be from October 13 – 19, 2025. The actual schedule will be provided by your respective teacher.
    - Be ready to show any aspect of your design, including your database schema, SQL query statements, and ETL pipeline.

Plagiarized works – works copied from others and AI generated content – will automatically be given a grade of 0.0 for the course.

Grading Rubric

DIMENSIONAL MODEL Design - fact and dimensional tables [15 pts]			
Critical thinking	[5] Clearly explained the **rationale for the design vis-à-vis OLAP queries**, showing deep understanding and insights	[3] Explanation lacks depth or has minor flaw, showing ample understanding of the DB schema and application	[1] **No clear use case** Evident lack of insight and understanding of the DB and its application
Thoroughness	[5] Presented a **schema design** that is complete with all the necessary and contingent fact and dimensional tables	[3] Presented schema design has minor flaw but has all the necessary fact and dimensional tables	[1] Incomplete or incorrect fact and dimensional tables
Organization	[5] Logical, optimized, proper star or snowflake schema; **Correct representation of pkeys, fkeys, and multiplicity constraints**	[3] Logical proper star or snowflake schema; Incomplete or missing some constraints, such as pkeys, fkeys, and multiplicity	[1] Minimal restructuring from source dataset; Fact table is a full or partial copy restructuring of the original data source

ETL SCRIPT [10 pts]			
Performance	[5] Reliable, resilient, reusable, maintainable, and/or well-performing; **Considerations in coding the ETL script were explained in detail**	[3] Reliable and well-performing, but lacks depth in explaining the ETL process and considerations with respect to data quality	[1] Cannot articulate the significance of the ETL process and considerations made with respect to data quality
Data Quality	[5] Performed appropriate **data wrangling tasks**; Process of **validating the data** has been explained	[3] No data wrangling tasks performed; Process of validating the data has been explained	[1] Static and raw data with no evident validation procedure

OLAP QUERIES [17 pts]			
Report Quality	[6] Reports require formulation of complex queries that show evidence of appreciation for and full understanding of	[4] Queries can be further enhanced to support analysis with OLAP operations (roll-up, drill-down, slice, dice, pivot)	[2] Formulated queries show lack of appreciation and understanding of advanced SQL constructs and OLAP

	advanced SQL constructs and OLAP (roll-up, drill-down, slice, dice, pivot)		
Report Granularity	[6] **OLAP operations are parameter-driven** to support varying levels of dimensional analysis involving roll-up, drill-down, slice, dice, pivot (at least 4);	[4] Minimal support for dimensional analysis due to the presence of queries with hardcoded parameters that reduces the level of granularity of the reports	[2] Single-level OLAP operations for roll-up, drill-down, slice, dice
Interactivity and Usability	[5] Reports are presented in an interface with clear structure and layout, requiring minimal effort for users to specify query parameters and view reports	[3] Overall user interface is clean from clutter, but requires some effort to specify query parameters and view the reports	[1] Static tables are used for presentation of data; Interface design decisions are not evident; UI displays technical terms
QUERY OPTIMIZATION [18 pts]			
Quality	[6] Evident appreciation and full understanding of SQL capabilities through **properly formulated queries that utilized advanced clauses**, e.g., GROUP BY	[4] Shows ample appreciation and basic understanding of SQL capabilities	[2] Shows little to no appreciation or understanding of SQL capabilities
Strategies	[6] Correct application of **2 or more strategies** to improve query speed; Indexes are properly utilized; **Results of query processing speed tests are presented**	[4] Applied 1-2 strategies with partial improvement in performance; Reliance on the tool's optimization with no database restructuring	[2] Optimization strategies did not yield any improvement in query speed
Analysis	[6] Detailed analysis and discussion of how DB design and optimization strategies impacted the DB performance	[4] Presentation of optimizations strategies lacks depth and has minor flaw	[2] Cannot articulate the rationale for the (non-)improvement in performance

TECHNICAL REPORT [30 pts]			
Design	[6] Clear, concise and thorough discussion of design decisions **with supporting related work**, evident of critical and reflective thinking as well as active learning in the conduct of the project	[4] Clear discussion of design decisions evident of critical thinking but partially supported by related work; Insufficient supporting related work; Limited discussion of DW model design	[2] Superficial, vague and/or incomplete discussion of design decisions, lacking evidence of critical thinking and supporting related work
Methodology	[6] Clear and concise description of the **activities performed and challenges faced**, from creating the DW, scripting the ETL process, designing the dashboard, to query optimization with evidence of independent, resourceful learning and analytical thinking	[4] Activities performed and challenges faced were described but contain flaws or missing details; required some guidance in executing the methodology with partial evidence of independent learning and resourcefulness	[2] Incomplete and/or ambiguous description of activities; lack important details; reliant on the guidance of mentors and/or peers throughout the execution of the activities showing lack of independent learning and resourcefulness
Query Performance	[6] Clear and correct **optimization strategies** were employed and justified based on relevant theories and evaluation method; Validation of dashboard and query performance is efficient and effective	[4] Employed optimization strategies are acceptable but may not be necessary; Validation of dashboard and query performance is sufficient but not thorough, missing details of test cases	[2] Vague description of optimization strategies; Insufficient method in validating the performance of the dashboard and the queries
Analysis and Discussion of Results	[6] Careful thought and reflection is evident in presenting the **analysis and the insights** gained from the test results and conduct of the activity; Includes ETL data validation, app function testing and query performance testing	[4] Discussion lacks depth and can be further enriched by sharing insights and lessons learned from the conduct of the activity	[2] Incomplete and/or ambiguous discussion of insights and lessons gained from the results and the experience; **Missing test results**
Language and Format	[6] Evidence of careful choice of words and use of correct grammar; Properly follows ACM publication format; Proper citations and references	[4] Has minor flaws in the choice of words and/or grammar; Minor flaw in the use of prescribed template; Missing some citations and/or references	[2] Numerous issues in choice of words and grammar; Non-compliance with prescribed template; Missing important references and citations

PRESENTATION [10 pts]			
Presentation	[10] Concise presentation of design decisions, method, results and insights gained from the conduct of the study ❏ Dimensional Model ❏ ETL script ❏ Queries (SQL), processing speed, optimization strategies ❏ Test methodology, results ❏ Conclusion	[7] Presentation sufficiently explains the tasks that were carried out, but some parts of the discussion lack details	[3] Vague discussion of the activity that was conducted and the results and insights gained from the experiment