FIT3003 Major Assignment - Sem 2/2021 (Weight = 20%)
Due date: Week 10, Wednesday 6-October-2021, 11:55pm

Version: 1.0 – 26/08/2021

Learning Outcomes:

LO1. Design multi-dimensional databases and data warehouses.
LO2. Use fact and dimensional modelling.
LO3. Implement online analytical processing (OLAP) queries.
LO4. Explain the roles of data warehousing architecture and the concepts of granularity in data warehousing.
LO5. Propose business intelligence reports using data warehouses and OLAP.

A. General Information and Submission

- This is a group assignment. One group consists of **3 students** from the **same tutorial**. You need to register your group composition through the Major Assignment Group Selection Form as soon as possible.
- *Submission method*: Submission is online through Moodle.
- *Penalty for late submission*: 10% deduction for each day.
- *Oracle account details*: You will need to supply an Oracle username and password used for this assignment.
- *Assignment Coversheet*: You will need to sign the assignment coversheet.
- *Contribution Form*: All members must complete the contribution form, and please sign (e-signature is acceptable) the form as an agreement between members.
- *Assignment FAQ*: There is a Major Assignment Frequently Asked Questions page set up for the Major Assignment on EdStem Forum.

B. Problem Description

Monash University acquired an entertainment corporation that owned more than 20 cinemas in different states of Australia and changed its name to MonC.

The company has an existing operational database that maintains and stores all of the business transaction information (such as ticket sales, movies, customers, etc.) required for the management's daily operation. However, with the acquisition of the company, the management of MonC company has decided to hire your team of Data Warehouse Engineers to design, develop, and quickly generate reports from a Data Warehouse to improve the work efficiency. The management at MonC wants to generate reports to keep track of the cinema revenues and related information (e.g. calculating statistics of tickets sold and movies offered), which can later be used for forecasting various trends and making predictions about the customers' interest.

The operational database tables can be found at the MonCinema account. You can, for example, execute the following query:

select * from MonCinema.<table_name>;

The data definition of each table in MonCinema is as follows:

Table Name	Attributes and Data Types		Notes
BOOKING_MODE	MODE_ID	NUMBER	This table stores the booking mode information.
	MODE_DESCRIPTION	VARCHAR	
CINEMA	CINEMA_ID	NUMBER	This table stores the cinema information.
	CINEMA_NAME	VARCHAR	
	CINEMA_ADDRESS	VARCHAR	
	CINEMA_SUBURB	VARCHAR	
	CINEMA_POSTCODE	CHAR	
	CINEMA_STATE	CHAR	
CINEMA_RATING	RATING_ID	NUMBER	This table stores the cinema rating information. The rating was given anonymously.
	RATING_SCORE	NUMBER	
	RATING_DATE	DATE	
	CINEMA_ID	NUMBER	
CUSTOMER	CUST_ID	NUMBER	This table stores the information of the customers.
	CUST_NAME	VARCHAR	
	CUST_ADDRESS	VARCHAR	
	CUST_SUBURB	VARCHAR	
	CUST_POSTCODE	CHAR	

	CUST_STATE	CHAR	
	CUST_CONTACT_NUMBER	VARCHAR	
	CUST_DOB	DATE	
	CUST_GENDER	CHAR	
GENRE	GENRE_ID	NUMBER	This table stores the information of the movie genre.
	GENRE_DESCRIPTION	VARCHAR	
MOVIE	MOVIE_ID	NUMBER	This table stores the information about the movie.
	MOVIE_NAME	VARCHAR	
	MOVIE_RELEASE_DATE	DATE	
	MOVIE_SPOKEN_LANGUAGE	VARCHAR	
	MOVIE_RUNTIME	NUMBER	
MOVIE_CINEMA	MOVIE_ID	NUMBER	This table stores the information of the specific movie showing in a specific cinema.
	CINEMA_ID	NUMBER	
MOVIE_COMPANY	MOVIE_ID	NUMBER	This table stores the information of the specific movie produced by a specific company.
	COMPANY_ID	NUMBER	
MOVIE_GENRE	MOVIE_ID	NUMBER	This table stores the information on the genre of a specific movie.
	GENRE_ID	NUMBER	
PRODUCTION_COMPANY	COMPANY_ID	NUMBER	This table stores the information of the movie production company.
	COMPANY_NAME	VARCHAR	
	COMPANY_ADDRESS	VARCHAR	

REVIEW	REVIEW_ID	NUMBER	The table stores the information of movie reviews.
	REVIEW_SCORE	NUMBER	
	REVIEW_DATE	DATE	
	MOVIE_ID	NUMBER	
SALE	SALE_ID	NUMBER	This table stores the information about the ticket sales.
	SALE_DATE	DATE	
	SALE_NUMBER_OF_TICKETS	NUMBER	
	SALE_UNIT_PRICE	NUMBER	
	SALE_TOTAL_PRICE	NUMBER	
	CUST_ID	NUMBER	
	MOVIE_ID	NUMBER	
	CINEMA_ID	NUMBER	
	MODE_ID	NUMBER	
	STAFF_NO	NUMBER	
STAFF	STAFF_NO	NUMBER	This table stores the information of the staff.
	STAFF_NAME	VARCHAR	
	STAFF_GENDER	CHAR	

C. Tasks

The assignment is divided into **FOUR** main tasks:

1. **Design a data warehouse for the above MonC database.**
   You are required to create a data warehouse for the MonC database.
   The management is especially interested in the following fact measures:

   - Total sales revenue
   - Number of tickets sold
   - Number of cinemas
   - Number of movies
   - Average sales revenue

   The following show some possible dimension attributes that you should need in your data warehouse:

   - Month, year
   - Season
   - Customer's age group [Child: 0-16 years old; Young adults: 17-30 years old, Middle-aged adults: 31-45 years old, Old-aged adults: Over 45 years old]
   - Movie genre
   - Production company
   - Movie runtime category: Short (less than 50 minutes), Medium (between 50 and 100 minutes), Long (longer than 100 minutes)
   - Booking mode
   - Cinema's location
   - Cinema's rating scores
   - Movie's review scores

   For each attribute, you may apply your own design decisions on specifying a range or a group, but make sure to specify them in your submission.

- **Preparation stage.**
  Before you start designing the data warehouse, you have to ensure that you have explored the operational database and have done sufficient data cleaning. Once you have done the data cleaning process, you are required to explain what strategies you have taken to explore and clean the data.

The outputs of this task are:

a) The E/R diagram of the operational database,

b) If you have done the data cleaning process, explain the strategies you used in this process (you need to show the SQL to explore the operational database and SQL of the data cleaning, as well as the screenshot of data *before* and *after* data cleaning),

- **Designing the data warehouse by drawing star/snowflake schema.**

The star schema for this data warehouse contains multi-facts. You need to identify the fact measures, dimensions, and attributes of the star/snowflake schema. The following queries might help you to determine the fact measures and dimensions:

- How many tickets were sold in October 2018?
- Which season has the highest sales revenue?
- How many short movies were released in 1990?
- How many 5 stars rating cinemas are there in Melbourne?
- What is the most popular booking mode by young adults (by the number of sales)?
- How much revenue was generated by short comedy movies in Winter 2020?
- What was the average sales revenue for 3 stars reviewed movies?
- What were the total revenues for movies produced by NEW Century?

You should pay attention to the granularity of your fact tables. You are required to create **two versions** of star/snowflake based on different levels of aggregation.
The two versions of the star/snowflake represent different levels of aggregation. Version-1 should be at the highest level of aggregation. Version-2 should be in level 0, which means no aggregation. To make it simple, you can assume that the highest aggregation for this assignment is Level-2.

Version Name	Level
Version-1	High aggregation (Level 2)
Version-2	No aggregation (Level 0)

The star/snowflake schema of both versions you created might contain **Bridge Table** and **Temporal**. If a bridge table is needed, you will need to include GroupList and WeightFactor attributes in the relevant dimension. If a temporal dimension is needed, you can use any suitable temporal data warehousing techniques for the temporal dimension and provide the reasons for your choice.

The outputs of this task are:

c) Two versions of star/snowflake schema diagrams,

d) The reasons for the choice of SCD type for temporal dimension, if any,

e) A short explanation of the difference between the two versions of the star/snowflake schema.

2. **Implement the two versions of the star/snowflake schema using SQL.**
You are required to implement the star/snowflake schema for the two versions you have drawn in Task 1. This implies that you need to create the different fact and dimension tables for two versions in SQL and populate these tables accordingly.

When naming the fact tables and dimension tables, you must provide an identical name for the two versions and end with the version number to differentiate them. For example, "MonC_fact_v1" for version-1 and "MonC_fact_v2" for version-2. If the dimension is the same between the two versions, you do not need to create them twice.

The output is a series of SQL statements to perform this task. You will also need to show that this task has been carried out successfully.

If your account is full, you will need to drop all of the tables you have previously created during the tutorials.

The outputs of this task are:

a) SQL statements (e.g. create table, insert into, etc) to create the star/snowflake schema Version-1

b) SQL statements (e.g. create table, insert into, etc) to create the star/snowflake schema Version-2

c) Screenshots of the tables you have created; this includes the contents of each table that you have created. If the table is very big, you can show only the first part of the data.

3. **Create the following reports using OLAP queries.**
You are required to generate the reports using both data warehouse versions, **version-1 (Level 2)** and **version-2 (Level 0 no aggregation)**, that you have implemented in Task 2. For each report, you ought to produce the SQL command and sample report output.

   a. Simple reports:

Produce **three** reports. Each report contains two attributes from two different dimensions and one fact measurement.

For the report itself, the first report must be about **Top *k***, the second report is **Top *n%***, and the third report is **Show All.**

The outputs of this task are:

(a) The query questions that are written in English,
(b) Your explanation on why such a query is necessary or valuable for the management,
(c) The SQL commands, and
(d) The screenshots of the query results (or part of the query results), including all attribute names.

b. *Reports with proper sub-totals:*

Produce **four** reports. These reports must include subtotals, using the Cube or Roll-up or Partial Cube/Roll-up operators.

REPORT 4 and REPORT 5: What are the subtotals and total sales revenue from each cinema's location, season, movie runtime category? (You must use the Cube and Partial Cube operator)

REPORT 6 and REPORT 7: Produce two other subtotals reports that are useful for management using Roll-up and Partial Roll-up

The outputs of this task are:

(a) The query questions that are written in English,
(b) Your explanation on why such a query is necessary or valuable for the management,
(c) The SQL commands that include subtotals, using the Cube or Roll-up or Partial Cube/Roll-up operators, and
(d) The screenshots of the query results (or part of the query results).

c. *Reports with moving and cumulative aggregates:*

Produce **three** reports containing moving and cumulative aggregates.

REPORT 8: What are the total sales and cumulative total sales of animation movies in each year?

REPORT 9 and REPORT 10: Produce two other moving/cumulative aggregate reports that are useful for management.

The outputs of this task are:

(a) The query questions that are written in English,
(b) Your explanation on why such a query is necessary or valuable for the management,

(c) The SQL commands that contain moving and cumulative aggregates, and
(d) The screenshots of the query results (or part of the query results).

d. Reports with Partitions:

Produce **two** reports that contain partitions.

REPORT 11: Show ranking of each movie genre based on the monthly total number of tickets sold and the ranking of each booking mode based on the monthly total number of tickets sold.

REPORT 12: Produce another partitioning report that is useful for management.

The outputs of this task are:

(a) The query questions that are written in English,
(b) Your explanation on why such a query is necessary or useful for the management,
(c) The SQL commands that contain partitions, and
(d) The screenshots of the query results (or part of the query results), including all attribute names.

Note: At the end of this task, you should have **24** reports in total: 12 reports using data warehouse **version-1** and 12 reports using data warehouse **version-2**.

4. Business Intelligence (BI) Reports.

Choose any **five** reports from Task 3, and change the presentation of these reports by representing these in a graph format. This new presentation should be more appealing to the management. You can use any visualisation tools (e.g. Oracle Report, PowerBI, Tableau) to show the graph reports. Additionally, in these new reports, you might want to include some selection buttons (for illustrative purposes), which may give users options on what criteria to choose so that the graph report will be more dynamic.

D. Checkpoints

There will be checkpoints in Week 7, 8, 9:

Checkpoint	Weight	Assessment	Deadline
Checkpoint 1	1%	ER Diagram Data Cleaning	Week 7 (during tutorial)
Checkpoint 2	1%	Star Schema v1	Week 8 (during tutorial)
Checkpoint 3	1%	Star Schema v2	Week 9 (during tutorial)

The Checkpoints will only be assessed during the allocated tutorial. Your group is required to complete the assessment for a given checkpoint in order to obtain the allocated mark. There are associated mark penalties for not meeting the checkpoint assessment on time to a satisfactory state.

Note that the Final Report and Code are worth 17%.

E. Submission Checklist

1. One **combined pdf file** containing all tasks mentioned above:
    - □ Cover page
    - □ A signed coversheet
    - □ Details of your ORACLE accounts
    - □ A contribution declaration form:

      Each student must state the parts of the assignment that the student did. An example is as follows:

      Percentage of contribution:
      1. Name: Adam, ID: 210008, Contribution: 50%
      2. Name: Ben, ID: 230933, Contribution: 30%
      3. Name: Clara, ID: 240024, Contribution: 20%

      List of parts that each student did:
      1. Adam: list the parts that Adam did
      2. Ben: list the parts that Ben did
      3. Clara: list the parts that Clara did

- Task C.1 (outputs a, b, c, d, e)
- Task C.2 (outputs a, b, c)
- Task C.3 Simple Reports (outputs a, b, c, d)
- Task C.3 Reports with Subtotals (outputs a, b, c, d)
- Task C.3 Reports with Moving and Cumulative Aggregates (outputs a, b, c, d)
- Task C.3 Reports with Partitions (outputs a, b, c, d)
- Task C.4 (five graphs)

2. **.sql files** for the following task:
    - Task C.1 (SQL command as required by output *b*)
    - Task C.2 Implement Star Schemas (SQL command as required by output a and b)
    - Task C.3 Simple Reports (SQL command as required by output c)
    - Task C.3 Reports with Subtotals (SQL command as required by output c)
    - Task C.3 Reports with Moving and Cumulative Aggregates (SQL command as required by output c)
    - Task C.3 Reports with Partitions (SQL command as required by output c)

    All of the above SQL files must be runnable in Oracle.

3. Zip all the SQL files from #2, and name the ZIP folder as **MA_SQL.zip**.

Submission Method:

1. Upload the **PDF file** from Checklist #1 and the **ZIP file** from Checklist #3 to Moodle by the due date: **Wednesday, 6 October 2021, 11:55pm**.
    - The submission of this assignment must be in the form of **a single PDF file AND a single ZIP file**. No other forms will be accepted.
    - One member of your group can upload the submission. However, **please note that all group members must click the submit button and accept the submission statement** (failure to do so will mean your assignment will not be submitted and will incur late penalties).

- You must ensure that you have all the files listed in this checklist before submitting your assignment to Moodle. Failure to submit a complete list of files will lead to mark penalties.

2. Penalty for late submission: 10% deduction for each day, including weekends.

3. Submission Cut-off time: **Wednesday, 13 October 2021, 11:55 pm** (Submission link will be unavailable after the cut-off date).

Getting help and support:

What can you get help for?

- ***Consultations with the Teaching Team***
  Talk to the Teaching Team:
  https://lms.monash.edu/course/view.php?id=115942&section=2
- ***English language skills***
  Talk to English Connect: https://www.monash.edu/english-connect
- ***Study skills***
  Talk to a learning skills advisor: https://www.monash.edu/library/skills/contacts
- ***Counselling***
  Talk to a counsellor: https://www.monash.edu/health/counselling/appointments

Extensions:

If you are experiencing difficulties that you think will impact your ability to meet this deadline, you may apply for an assignment extension. You must apply **no later than two University working days after the due date** of this assignment.

The extension application can be found on *Moodle > Assessments > How to Apply for an Extension*. Please allow **two business days** for your application to be processed.

Please ensure your application is supported by appropriate documentation. You can find more information about assignment extensions at the Special Consideration website.

Special Considerations:

Students should carefully read the Special Consideration website, especially the details about what formal documentation is required.

All special consideration requests should be made using the Special Consideration Application.

Please do not assume that submission of a Special Consideration application guarantees that it will be granted – you must receive an official confirmation that it has been granted.

Late Penalty:

Late assignments submitted without an approved extension may be accepted (up to a maximum of **seven days**) with the approval of the Chief Examiner and/or Lecturer but will be **penalised at the rate of 10% per day (including weekends and public holidays)**. Assignments submitted more than seven days after the due date will receive a zero mark for that assignment and may **not receive any feedback**.

Plagiarism and Collusion:

Monash University is committed to upholding standards and academic integrity and honesty. Please take the time to view these links.

Academic Integrity Module

Student Academic Integrity Policy

Test your knowledge, collusion (FIT No Collusion Module)

END OF MAJOR ASSIGNMENT