School of Computing
ST1501 Data Engineering CA2
AY2021/2022 Semester 1

A. Instructions and Guidelines

1. This is a group assignment with individual components. It requires students to design and set up a Data Warehouse, and demonstrate competency in designing a Data Warehouse and integrating data from various sources.

2. You are to work in a group of 3 (recommended) or 4 members, with a representative appointed and a name for the team.

3. This assignment will account for 40% of the module grade.

4. The deadline of this assignment is **6 Aug 2021 Friday by 11 pm**. Submission should be made in Blackboard by the stated deadline.

5. Your submission should include:

   For the team representative:

   - A zip file for the group submission with the file-naming convention: "ST1501-Class-GroupName.zip"
   - A zip file for the individual submission with the file-naming convention: "ST1501-Class-StudentID-Name.zip"

   For team member:

   a) A zip file for the individual submission with the file-naming convention: "ST1501-Class-StudentID-Name.zip"

6. The group submission should consist of the following:

   - The database diagram of the Data Warehouse in PDF, JPEG or PNG format
   - All SQL scripts used to setup the OLTP database in Microsoft SQL server
   - All SQL scripts used to setup the Data Warehouse in Microsoft SQL server
   - A Microsoft Word document that captures the five meaningful queries, along with its query results and insights or recommendations to the business owner
   - A Microsoft Powerpoint document that explains the following:
     - Problem(s) encountered when importing data into the OLTP database
     - The ETL process of transferring data from the OLTP to the Data Warehouse
     - Why the Star or Snowflake schema is adopted
     - The reasons for the various Dimension tables created
     - The metrics that are included in the Fact table
     - The creation of Time dimension and the attributes it contains
     - Explain the Surrogate keys and how they are linked to the foreign keys

7. The individual submission should consist of the following:

    - All SQL scripts used to setup the group Data Warehouse in Databricks platform.
    - A Microsoft Word document that provides the three meaningful queries, along with the query results and the insights or recommendations to the business owner.

8. The team is required to present the assignment. The presentation should include:

    - The explanation of the Data Warehouse design
    - The explanation of the process of setting up the OLTP database and the Data Warehouse
    - The explanation of the meaningful queries to the business owners

    Each team member is required to demonstrate his/her ability to explain the Data Warehouse, and the questions fielded by the module tutor during the presentation.

9. Student who is absent from the presentation will be given zero mark for the assignment.

10. No marks will be awarded, if the work is copied or you have allowed others to copy your work.

11. 50% of the marks will be deducted for assignments that are received within ONE (1) calendar day after the submission deadline. No marks will be given thereafter. Exceptions to this policy will be given to students with valid LOA on medical or compassionate grounds. Students in such cases will need to inform the module tutor as soon as reasonably possible. Students are not to assume on their own that their deadline has been extended.

B. The Business Scenario

Rolling Music Store is a retailer of music records based in Canada, specializing in various types of music. They have been using an OLTP system for their day-to-day business operations since 2009.

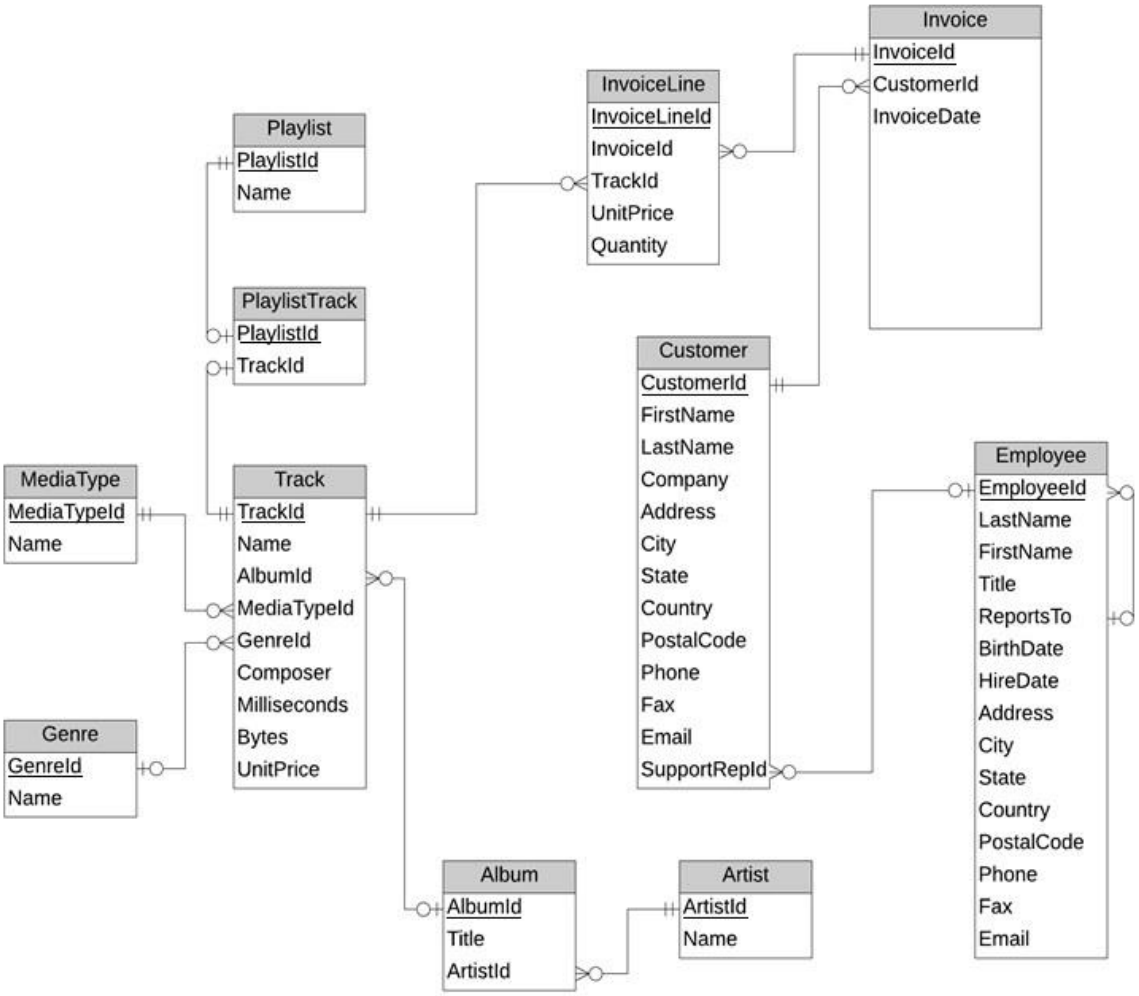
After 5 years of selling music, the business owner, Jackson Sam wanted to incorporate a Data Warehouse solution in his existing IT infrastructure after listening to a talk on Business Analytics.

Mr. Sam wants to have business queries on his business model to be answered from a Data Warehouse. He also wishes to keep as much details of time attributes so that he knows the peak and low seasons. Mr. Sam wants to know what can be done to improve the sales, to improve the operational efficiency, to increase their customers' loyalty etc.

There are 11 tables in the OLTP system as shown in page 4:

- employees table stores employee data such as employee id, last name, first name, etc. It also has a field named ReportsTo to specify who reports to whom.
- customers table stores customers data.
- invoices & invoice_items tables: these two tables store invoice data. The invoices table stores invoice header data and the invoice_items table stores the invoice line items data.
- artists table stores artists data. It is a simple table that contains only the artist id and name.
- albums table stores data about a list of tracks. Each album belongs to one artist. However, one artist may have multiple albums.
- media_types table stores media types such as MPEG audio and AAC audio files.
- genres table stores music types such as rock, jazz, metal, etc.
- tracks table stores the data of songs. Each track belongs to one album.
- playlists & playlist_track tables: playlists table store data about playlists. Each playlist contains a list of tracks. Each track may belong to multiple playlists. The relationship between the playlists table and tracks table is many-to-many. The playlist_track table is used to reflect this relationship.

OLTP Database Schema



Note:

The data files provided are encoded in Unicode. You are to use the Unicode data types in Microsoft SQL Server when creating the OLTP database. For instance, nchar or nvarchar(255).

C. Your Task

As a Data Specialist team, your team is asked to design and set up three databases: an OLTP database and a Data Warehouse using Microsoft SQL Server, and a Data Warehouse using Databricks (Community Edition).

Group Task

1. Design a Star or Snowflake schema for a Data Warehouse that will help answer various business questions based on the scenarios described in Section B. The team must seek your module tutor's feedback on the design, before proceeding to setup the Data Warehouse.

2. Setup the OLTP database (i.e. create the tables, insert the data from the data files provided) using Microsoft SQL Server, based on the OLTP database schema shown in page 4:
   - The database should be named as MusicStoreXXXX, where XXXX is to be replaced by the team name.
   - The data to be loaded into the OLTP database come from various sources – refer to the zip file "Data.zip". The filenames are the same as the table names, as shown in the OLTP database schema in page 4. Please take note that the data are encoded in Unicode (not UTF8).
   - Some of the data files may contain "dirty data" like incorrect header, NULL data etc. Your team should identify the "dirty data" and rectify it, so that the data could be loaded successfully into the OLTP database.
   - All SQL scripts used to create the OLTP database and the data loading are to be submitted.

3. Create the Data Warehouse using Microsoft SQL Server, based on your team's Data Warehouse design that has been reviewed by your module tutor.
   - The database should be named as MusicStoreDWXXXX, where XXXX is to be replaced by the team name.
   - Implement Surrogate keys for all Dimension and Fact tables.
   - All SQL scripts used to create the Data Warehouse are to be submitted.

4. Load the data from the OLTP database (i.e. the source system) to the Data Warehouse.
   - Your team may wish to create a temporary Time Dimension using the codes from the Practical E1. And then create the real Time Dimension that stores only the relevant attributes for the assignment.
   - All SQL scripts used for loading the data into the Data Warehouse are to be submitted.

5. Provide five meaningful queries that can be supported by the Data Warehouse covering any of the following categories:
   - Sales/profits/discounts/revenue
   - Staff/store/demography
   - Seasons of Sales/time
   - Orders/customers
   - Albums/genres/playlists

The query should provide insightful findings to the business owner. It's recommended to incorporate Aggregate functions, Row-wise functions, Group-by, Sub-query etc in the SQL queries.

Remember, there is no right and wrong answers, only insightful or trivial observations. That will depend on how much you want to explore and your examination of the data. Stating the obvious will not eligible for high grade. Your module tutor can tell whether enough thoughts have been put into the queries presented.

Individual Task

6. Setup a similar database as your group Data Warehouse in the Databricks platform (Community Edition), using your own account:
    - The database should have all the tables from your group Data Warehouse, except the Time Dimension table. It is recommended to make use of SQL date functions, in the absence of the Time Dimension table.
    - There is no requirements to implement the Entity and Referential Integrity constraints.
    - All SPARK SQL scripts used to setup the Data Warehouse must be submitted.

7. Write the SPARK SQL queries:
    - Provide two queries as implemented in the group report on the Databricks platform.
    - Provide one addition query to be supported by the Data Warehouse on the Databricks platform.
    - All SPARK SQL statements, query results and the insights must be submitted.

D. Assessment Criteria

Components	Weightage
Data Warehouse Design • The design supports the described business scenario. • The chosen table names, field names and attributes are descriptive. • The explanation of the design is clear and concise.	15%
OLTP Database Setup in Microsoft SQL Server • The database implements the described design in Section B, with Entity and Referential Integrity constraints. • The database is loaded with all the given data from the sources. • The explanation and demonstration of OLTP setup are without errors during the presentation.	15%
Data Warehouse Setup in Microsoft SQL Server • The database implements the Star or Snowflake design, with Entity Integrity constraint, Referential Integrity constraint and Surrogate Keys. • The database is loaded with all the relevant data from OLTP Database using ETL process • The explanation and demonstration of Data Warehouse are without errors during the presentation.	15%

Five Insightful Queries
• Five insightful query that covers any of the following category (4% each):
  ○ Sales/profits/discounts/revenue
  ○ Staff/store/demography
  ○ Seasons of Sales/time
  ○ Orders/customers
  ○ Albums/genres/playlists
	20%

Data Warehouse Setup in Databricks (Individual) • The database implements the group's Data Warehouse in accordance to the specifications in Section C	15%
Three Insightful Queries in Databricks (Individual) • Two insightful queries as implemented in the group's Data Warehouse (5%) • One addition insightful query by the individual (5%)	10%
Demonstration & Interview • Questions are answered correctly during the presentation and demonstration	10%

** Up to 10 marks can be deducted for poor organization of the report