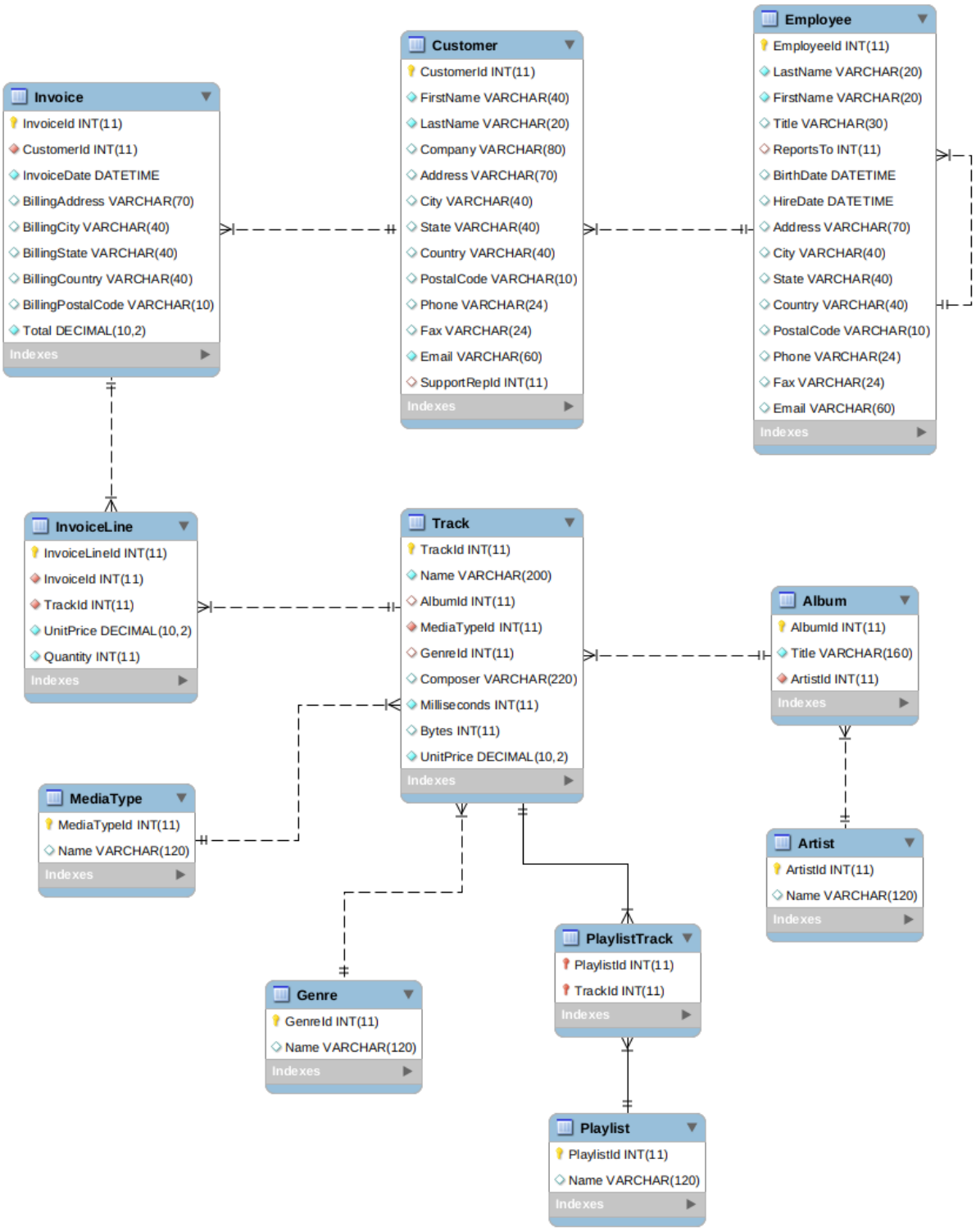
Introduction

In this project, you will be working with the Chinook database. This database refers to a digital media store, including tables for artists, albums, media tracks, invoices and customers. The following figure illustrates the database schema:



The script **Chinook.sql** contains the SQL instructions needed to create the Chinook database in MySQL. The diagram above was generated using MySQL Workbench, and the same tool can be used to explore the data and get more familiar with the database and its structure.

Challenges

In this project, you are asked to address the following challenges:

1. In the database, there is the concept of Customer and Employee. The company wants to create a new concept, called Contact, which may be either a customer or an employee. You are asked to develop a unified view over the Customer and Employee tables to provide this new concept. The view should use the common attributes between both tables.

2. By looking at the Track table, the company suspects that there might be duplicate tracks in its database. You are asked to develop a transformation to identify approximate duplicates based on Name and Composer, for tracks where Composer is not NULL. Use the Jaro measure on both fields, and a combined threshold of 0.88.

3. The company is worried about the quality of data in the Customer table. Use a data profiling tool to conduct an analysis and provide your own conclusions with respect to completeness, string fields, referential integrity, and value distributions.

4. The company would like to build a data warehouse to analyze sales and quantity by track, album, artist, customer, geography and time. These should be organized into appropriate dimensions, namely:
   - a track dimension with track name, album title, artist name
   - a customer dimension with full name, city, country
   - a time dimension with day, month, year

   You should propose a star schema for this purpose, and provide the instructions for creating such data warehouse.

5. The company needs an ETL process to extract data from the Chinook database, transform it, and load it into the data warehouse. Develop this ETL process as a set of transformations implemented with Pentaho Data Integration (PDI).

6. Once the data warehouse has been created and populated, write two SQL queries to illustrate the types of analysis that can be carried out with pure SQL over the data warehouse. The queries should involve multiple dimensions. Using your own words, briefly explain what each query is doing, and show the query results (or an excerpt of the results, if they are too large).

7. Use Pentaho Schema Workbench (PSW) to define the data cube for the data warehouse.

8. Using Saiku Analytics, create two interesting queries by drag-and-drop and visualize their results. These queries should convince the company about the potential of using the data warehouse to analyze their data. For that purpose, the queries should involve multiple dimensions and measures, at different levels of aggregation.

9. Simplify as much as possible the MDX code that Saiku has generated for the previous queries, making them more easily readable or understandable by a human.

10. Using Pentaho Report Designer (PRD), prepare a report with data obtained from the data warehouse. The report should make use of an MDX query and a SQL query over the data warehouse, and it should have a listing and a chart. Using your own words, briefly explain what is being shown in the report.

Additional challenge

11. If you would like to try to obtain the maximum possible grade, turn the customer dimension into a slowly-changing dimension, so that if a customer changes city or country, a new version of that customer will be created. Describe the modifications that you had to do in order to implement this slowly-changing dimension, and present an example to show that it is working.

Results

With respect to each of the challenges above, you are asked to provide the following results:

1. Present the SQL code for the view.

2. Take a screenshot of the entire transformation, followed by screenshots of the configuration window and of the preview window for each step.

3. Take screenshots of the analysis results, and list your conclusions.

4. Present the SQL instructions needed to create the data warehouse. You can use steelwheels_dw.sql as an example. Once you create the data warehouse, use MySQL Workbench to generate a diagram for the data warehouse schema. In MySQL Workbench, you can use the menu option Database > Reverse Engineer for this purpose. Present the diagram.

5. For each transformation that you develop, do the following: take a screenshot of the entire transformation, followed by screenshots of the configuration window and of the preview window for each step.

6. Present the SQL code for each query, together with a brief description and the query results.

7. Present the XML code for the cube definition. The XML should be formatted and indented in a way that makes it easy to read for a human.

8. For each query, present a screenshot of the Saiku user interface, showing the measures, columns, rows and filters used in the query, together with the query results.

9. Present the MDX queries in their original form and in their simplified form.

10. Take screenshots of the report in Design mode and in Preview mode, and provide a brief description of the results.

11. Use screenshots and text to describe the modifications and the example.

To take screenshots, you can use Alt+PrintScreen to capture the active window inside the VM. There is also a Screenshot application in the VM that lets you capture a selected region on the screen. If you prefer, you can also use the screen-capture facilities of your host OS.

Questions about the project

The challenges are purposefully under-specified to create a more realistic scenario for the project. Whenever possible, you should make use of your own judgment rather than asking for additional details or requirements.

In case there is some issue you would like to clarify, please post your question on Slack, for example in the #project channel, so that everyone can see the question and answers, and possibly contribute to the discussion as well.

Submitting the project

Once you complete the challenges, use a presentation software (e.g. Powerpoint or Impress), to prepare a set of slides with the requested results (code, screenshots, etc).

On the first slide, write the number of your group, your names and student numbers.

Save the presentation as a PDF file (without image compression, or with lossless compression, to keep the image quality) and submit it in Fénix until the deadline (December 4, 2020).

Evaluation

Just for your information, we plan to evaluate the project as follows:

Challenge	Points
1	1
2	2
3	2
4	2
5	3
6	2
7	2
8	1
9	1
10	2
11	2
Total	20

Anyway, don’t worry too much about the grade, use this project as a learning experience!

Good luck!