The Koodoo Data Engineering Challenge

Introduction

You are working for online wine retailer Native Wines and you have just received a data-set from a new supplier whose wines the company is considering stocking. Before they add them to the store the product team want to do some analysis of the wines and so the data engineering manager has asked you to load them into the data warehouse.

NW's Data Warehouse is a SQLite database. Your task is to load the data into a staging table in the database and then transform that data into simple data mart ready for further analysis. We'll call this *The Wine Mart*

For the data engineering tasks you may use whatever programming language you wish. Feel free to use different languages for different tasks if you think it appropriate. The data engineering languages we use at Koodoo are predominantly SQL, Python, GoLang and some shell scripting (e.g. Bash or PowerShell depending on OS).

Please explain your choice of language for each task, providing the code you used along with any other thoughts you had.

The objective is to complete the tasks as efficiently as possible.

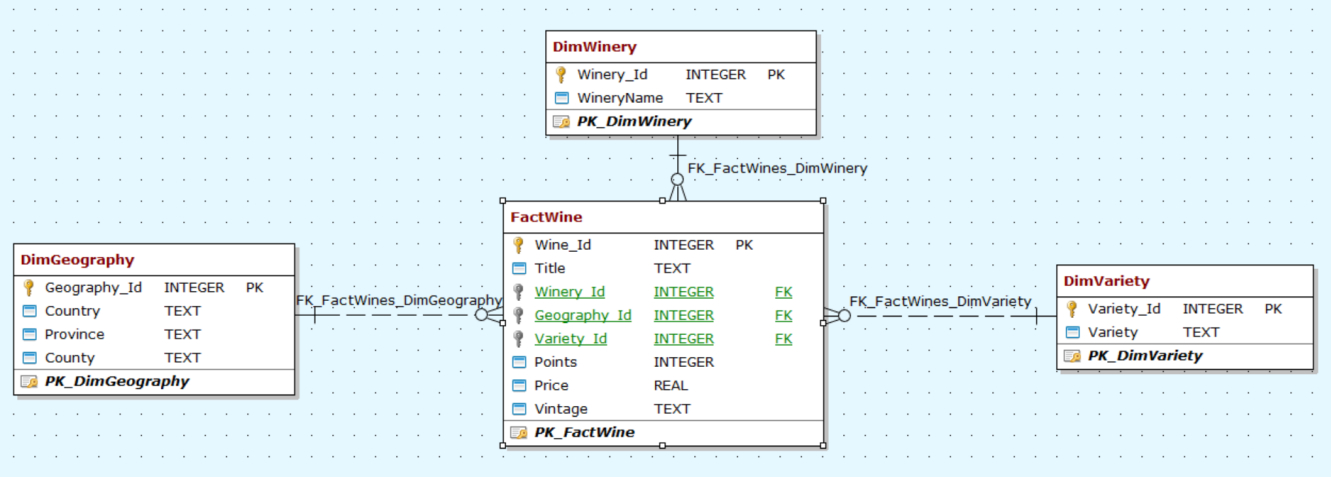
Database

The project will use SQLite as a database.

You may install SQLite on Windows Mac or Linux using the instructions in the link below.

https://www.servermania.com/kb/articles/install-sqlite/

The data model below describes the database structure for *The Wine Mart*.



Task One

Stage the data

Import the supplied CSV file into a SQLite database called Wine into a table called staging_wines.

When using SQLite in Windows, file paths need to be escaped. So for instance, c:\temp\data would become c:\\temp\\data

How many rows are in the staging_wines table?

What is the average price of a bottle of wine?

How much is the most expensive wine?

Task Two

Load the staged data into *The Wine Mart*

Create the four tables shown in the data model diagram above, that is:

- DimWinery
- DimGeography
- DimVariety
- FactWine

Write a script or program to load the contents of the staging table into the four tables above.

How many rows are there in the FactWine table?

What would happen if you deleted a record from the DimWinery table?

How could you prevent that deletion from occurring?

Finally

The product team have decided that they will be stocking wines from this new supplier and the supplier has agreed to supply a weekly file containing all their wines. They will place this file on their SFTP server every Sunday night.

What sort of process do you imagine could be used to automate the weekly task of updating *The Wine Mart* from the new file?

What sort of software tools would be most appropriate for this process?

What changes could you make to improve the *The Wine Mart* design to provide an audit trail of when different wines are added and/or updated?