MEIC

2018/2019

Data Analysis and Integration

1st semester

Project

Introduction

In this project, you will be working with the AdventureWorks database. This database was originally created by Microsoft for demonstration purposes. The database contains the sales data for *Adventure Works Bicycles Inc.*, a fictional bicycle wholesaler that sells and manufactures bicycles and sells clothing and cycling accessories to retailers around the world.

In this project, you are asked to perform two main tasks:

- Create an integrated view of the employee data that exists in the AdventureWorks database and in the "employees" database (employees.sql) that you used in the labs.
- Create a data warehouse from the AdventureWorks database, and write queries for analysis and reporting purposes.

The next sections describe in more detail what you should do in each of these tasks.

The AdventureWorks database

The script **AdventureWorks.sql** contains the SQL instructions needed to create the AdventureWorks database in MySQL. The database schema is quite large, but in this project we will be using only a subset of it. The file **AdventureWorks.pdf** provides a detailed view of the database schema.

In this project, we will be focusing mainly on data about **employees**, **customers**, **sales orders**, and **products**. Note that each order belongs to a single customer, but it may contain multiple products. Products are organized into categories and subcategories. Customers are organized into territories with geographical data associated to them, such as region, country, and continent.

Integrating the employee data

In the labs, you have worked with an "employees" database (employees.sql) that contained information about employees and departments. In the AdventureWorks database, you will also find information about employees and departments. Consider the following mediated schema:

```
all_employees(emp_no, birth_date, first_name, last_name, gender, hire_date, title, report_to)
all_departments(dept_no, dept_name)
all_dept_emp(emp_no, dept_no, from_date, to_date)
```

1. Present the schema matching between the AdventureWorks database and this mediated schema. The schema matching can be presented in the form of a table or diagram (as in the slides for this course).

2. Present the SQL views that define the schema mapping between the two data sources and the mediated schema.

3. Develop a PDI transformation to detect approximate duplicates between the job titles in both databases. The output of this transformation should be a list of pairs of potential duplicates.
   Take a screenshot of your transformation. Also, take a screenshot of the configuration window for each step. Finally, take a screenshot of the *Preview* window for the output step. You can use Alt+PrintScreen to capture these windows.

Creating the data warehouse

In the labs, you have created a data warehouse from the steelwheels database. This included developing an ETL process in Pentaho Data Integration (PDI), defining the OLAP cube in Pentaho Schema Workbench (PSW), querying the data warehouse in Saiku Analytics, and creating some reports in Pentaho Report Designer (PRD). In this project, you should follow a similar approach to create a data warehouse from the AdventureWorks database.

The data warehouse should have a star schema with the following dimensions:

- Territory with name, country region code, and group.
- Product with name, subcategory name and category name (slowly-changing dimension).
- Order date with day, month and year.

Use sales and quantity as measures. For sales, you can use the *LineTotal* that is already available.

4. Present the SQL instructions needed to create the data warehouse tables. The data warehouse should be called AdventureWorksDW. Ensure that all tables have appropriate primary and/or foreign keys. Present the SQL instructions in text format, but formatted and indented in a way that makes it easy to read (as in steelwheels_dw.sql).

5. Use PDI to develop a set of transformations and a job to implement an Extract-Transform-Load (ETL) process that populates the data warehouse with data coming from the AdventureWorks database. Take a screenshot of each transformation/job, followed by a screenshot of the configuration window for each step. You can use Alt+PrintScreen to capture these windows.

Exploring the data warehouse with SQL/OLAP

6. Write the following queries in SQL:
   a) What is the total quantity sold.
   b) What is the total quantity sold by country region code.
   c) What is the total quantity sold by country region code and by product category.

7. Write a single SQL/OLAP query (SQL with OLAP extensions) whose result is the same as the union of the results of the previous queries. The query should run on MySQL 5.7.[1]

Exploring the data warehouse with Saiku and MDX

8. Use PSW to create an XML definition of the data cube. The XML file should be called AdventureWorksDW.xml. Present the XML code in text format, but formatted and indented in a way that makes it easy to read.

9. Use Saiku to visualize the result of the following queries:
   a) Sales and quantity by country region code.
   b) Sales and quantity by country region code and product category.
   c) Sales and quantity by country region code and product category and year.

   Present a screenshot of the results for each query.

10. Write the following query in MDX:
    The quantity for each country region code and product category, for the years 2003 and 2004.

[1] See the SQL syntax supported by MySQL in: https://dev.mysql.com/doc/refman/5.7/en/

11. Use PRD to create a report based on the following query:
    Sales by product category, sorted by sales amount in decreasing order.
    The report should contain a list and a pie chart.
    Take two screenshots of the report in PRD: one in *Design* mode and another in *Preview* mode.

Submitting the project

After you complete the tasks above, use a word processor (e.g. LibreOffice or MS Word) to prepare a single document with all the requested materials (text and images).

At the top of the first page, write the number of your group, your name(s) and student number(s).

Save the document as a PDF file (without image compression, or with lossless compression) and submit it in Fénix until December 7, 2018.