# *Group Assignment: Data Management for Business*

Dr Nikolaos Korfiatis
(n.korfiatis@warwick.ac.uk)
NKResearch Analytics Consultants
Associate Professor of Business Analytics
Warwick Business School, MSc in Business Analytics

## Instructions

Please read the instructions carefully and discuss with your colleagues and provide an outline of your approach. Clarifications and/or any questions will be provided/answered **explicitly** through the module forum at **my.WBS**.

# 1. Overview and Pedagogical Goal

The goal of this assignment is to familiarize you with the complete process of *extracting*, *refining* and *delivering* datasets extracted from databases and unstructured data sources (e.g., unstructured files and the web). You are going to work in a group of five (5) in order to prepare datasets that can be used for further analysis. The assignment maps to level 7 qualification level and aims to establish the development of in-depth and original solutions to unpredictable problems and situations.

The assignment is structured in three (3) parts. The first part (Part A) covers structured data and in particular the design of databases and the extraction of data using structured query language (SQL). It aims to familiarize the students with the principles of relational databases and in particular: normalization, relational mapping, absorption of business rules on the relational view etc. The core of this assignment involves the translation of business requirements to data management solutions. The second and third parts (Part B and Part C) involves the management of unstructured data sourced by semi-structured data files such as spreadsheets (Part B) and the extraction of data from web sources (Part C).

# 2. Marking Criteria and Weights

The marking criteria for this assignment are as follows:

- Part A: 30% - Completeness of the solution, validation of the relational schema, normalization principles (adherence to the first normal form), definition of SQL queries (DDL), understanding of the business questions and translation to SQL.
- Part B: 20% - Solution validity (against the provided data structure)
- Part C: 20% - Solution validity. Efficiency of the solution.
- Part D: 10% – Solution validity, Efficiency of the solution.

The peer assessment component adds the remaining 20% of the mark as an individual component.

# 3. Feedback

Feedback will be provided in individual sessions upon request with points for further improvement.

# 4. Submission Instructions

Deadline for the assignment is the midday of **Friday 10/12/2020 (12:00).**

The assignment solutions should be submitted as one PDF-file document containing all code in

appendixes. No Zip files or additional files will be graded. Code should be formatted with R-markdown and any submissions where code is presented as an image will be penalized.

The file should be named as:

**group_number_X.pdf**

Where X is your given group number. Any failure to comply with the naming of the file will result to a 2% penalty.

# Part A: Structured Data

***Scenario***

*Scanco Hotel Group* is a rapidly growing hotel chain which is in need to create a system for optimizing its revenue management. The hotel provides accomodation services which comprises the main source of revenue as well as other channels for ancillary revenue including the rental of phone charging equipment for guests, the use of hotel facilities (which vary based on the hotel) for events such as : seminar rooms for training and meeting events, large rooms for events as well as banquet rooms for wedding and convention events. Each guest can book a room either directly from the hotel's booking system or through one of the 15 different channels that the hotel is active (e.g., Booking.com, Hotels.com, Tripadvisor.com, etc.). Each channel provider requires that at least two rooms are available for booking at anytime through the channel and charge a booking fee for each reservation. The booking fee is payable at the end of each month and the hotel needs to directly record the cost of each fee for its own accounts.

A guest can book a room and have additional services coupled in the same reservation. These can be breakfast, use of the mini-bar, restaurant meals etc. In addition, the use of facilities can be coupled with the reservation so when the guest checks out an invoice will be shown having a detailed breakdown of the costings incurred through the hotel.

There are many hotels in the chain. Each hotel has a name, a street address (which is made up of a street number, street name, city, state, and postal code), a home page URL (Web address), and a primary phone number.

Each hotel consists of a set of rooms arranged on various floors. Each room has an identifier which is unique within that hotel. Most of the time, rooms are numbered (e.g. 690), but they may be given a name (e.g. Presidential Suite) instead, so long as the name or number is unique within the hotel. Floors are numbered, and it's necessary, for each room, to know what floor it's on, since some customers prefer rooms on lower floors or higher floors. For simplicity, assume that each room is on only one floor. (Some real hotels have suites that span multiple floors.)

For each room, it's also necessary to keep track of how many beds it has, as well as whether smoking is allowed in the room. This information is used to help match guests to rooms with desired characteristics.

When a guest plans to stay at a hotel in the future, he or she makes a room reservation at the desired hotel. Each reservation indicates information about the guest, the desired arrival and departure dates, as well as preferences that aid in selecting the right kind of room for that guest: whether the room should be smoking or non-smoking, whether the room should have one beds or two, and whether the room should be on a high floor or a low floor. These room preferences are optional and are not included with every reservation; some guests are willing to take any available room, while some only care about some preferences but not others.

Also required with each reservation is information about a credit card that's used to secure the reservation; credit cards are indicated by a credit card number (which is a sequence of up to 16 digits) and an expiration date (a month and a year, such as "January 2007").

At any given time, a guest may have multiple reservations; reservation information is removed from the database after the guest's reservation is used to put them into an actual room, or when the guest cancels the reservation prematurely. The database tracks historical information about every guest's stay in any room in any hotel. At minimum, it's necessary to know what day the stay began, what day it ended, what room it was, what hotel it was, and who the guest was.

Information about each guest of each hotel is tracked historically. For each guest who has ever reserved or stayed in a room, the database must store the guest's first, middle, and last names, street address, email address, and three phone numbers (home, work, cell). Email addresses and the phone numbers are optional, while the other information is required.

An invoice is generated during a guest's stay at the hotel, detailing the individual charges accrued by the guest. These charges include not only the regular room rate, but also applicable taxes, as well as charges at the hotel's restaurants, bars, spas, shops, and so on. An invoice is displayed — either in printed or Web-based form — as a sequence of line items, with each line item consisting of a description and an amount, such as "Hotel Cafe — £29.75". Note that the database does not keep track of, say, the costs of items on the restaurant's menu or the cost of renting each room at various times throughout the year; it is assumed that another software system provides this information to our database, since our system only handles reservations and billing.

When a guest pays his or her bill — or a portion of his or her bill — a line item is added to the invoice that indicates how much was paid, and in what form the payment was made (e.g. "Visa — £-500.00", in the case of a $500 payment made using a Visa credit card). At the bottom of each invoice is a total balance,

which is the sum of the amounts in each of the line items, including both charges and payments. An invoice is considered paid if the amount is £0.00.

***Tasks***

You need to provide a reflective report (max 2500 words) where you address the following:

- Identify the entities, their relationships, cardinality and attributes from the above text. *Any solution will be acceptable as long as you state your assumptions* for your modeling.
- An E-R diagram of the relationships as well as relationship flows for each pair.
- The SQL DDL for this database including setup of data types and key constraints
- A set of scenarios where SQL queries that satisfy business goals are provided. The following need to be included:
    - The total spent for the customer for a particular stay (checkout invoice).
    - The most valuable customers in (a) the last two months, (b) past year and (c) from the beginning of the records.
    - Which are the top countries where our customers come from ?
    - How much did the hotel pay in referral fees for each of the platforms that we have contracted with?
    - What is the utilization rate for each hotel (that is the average billable days of a hotel specified as the average utilization of room bookings for the last 12 months)
    - Calculate the Customer Value in terms of total spent for each customer before the current booking.
- Additional SQL Queries and Business Goals will be considered as a plus but will not exclude the attenuation of a full mark.

This should be written as a report, following the points requested above. Any additional material and code can be included in the report or the Appendix if excessive. Your solution will be graded on completeness, adherence to the assumptions that you have followed (e.g., cardinality related assumptions) as well as the general formatting and presentation of your diagrams and code.

# Part B: Semi-Structured Data (Files)

The goal of this part is to familiarize students with the use of unstructured data and the combination of various individual files exported from a database system to a single file which can be used for further analysis. The data format that we are dealing here is called an ***unbalanced panel***.

The data folder is available on the my.WBS homepage and is exported from OECD. It provides a table of extended energy balances of OECD countries and each folder contains an excel file that its contents look like this:

**Dataset: Extended Energy Balances (OECD Countries)**

| Flow | Production | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Unit | TJ | | | | | | | | | | | |
| Country | Germany | | | | | | | | | | | |
| Product | Additives/blending components | Anthracite | Aviation gasoline | BKB | Biodiesels | Biogases | Biogasoline | Bitumen | Blast furnace gas | Brown coal (if no detail) | Charcoal | Coal tar |
| Time | | | | | | | | | | | | |
| 1960 | .. | .. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 804180 | 0 | 0 |
| 1961 | .. | .. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 812640 | 0 | 0 |
| 1962 | .. | .. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 845308 | 0 | 0 |
| 1963 | .. | .. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 890351 | 0 | 0 |
| 1964 | .. | .. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 925752 | 0 | 0 |
| 1965 | .. | .. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 850478 | 0 | 0 |
| 1966 | .. | .. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 814429 | 0 | 0 |
| 1967 | .. | .. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 801365 | 0 | 0 |
| 1968 | .. | .. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 839884 | 0 | 0 |
| 1969 | .. | .. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 887783 | 0 | 0 |
| 1970 | .. | .. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3048447 | 0 | 0 |
| 1971 | 0 | .. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2989454 | 0 | 0 |
| 1972 | 0 | .. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2957319 | 0 | 0 |
| 1973 | 0 | .. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3006752 | 0 | 0 |
| 1974 | 0 | .. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3045140 | 0 | 0 |
| 1975 | 0 | .. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3047372 | 0 | 0 |
| 1976 | 0 | .. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3139394 | 0 | 0 |
| 1977 | 0 | .. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3099669 | 0 | 0 |
| 1978 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | .. | 0 | 0 |
| 1979 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | .. | 0 | 0 |

Your goal is to use your R knowledge from the lectures and provide a dataset that conforms with the following structure like this:

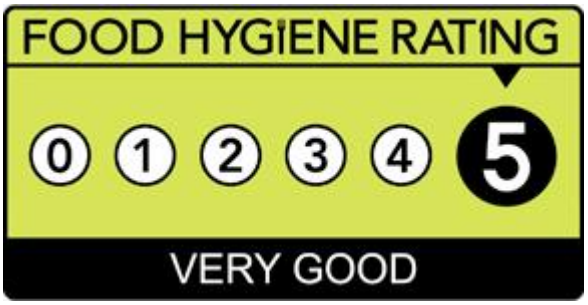
| country | year | flow | product | value |
|---|---|---|---|---|
| **Germany** | 1960 | Production | Additives/blending components | NA |
| **..** | .. | | .. | .. |
| **Germany** | 1960 | Production | Brown coal (if no detail) | 804180 |

The combination of country, year and flow and product should be unique, suggesting that there is only one particular value for that particular combination of the other 3 columns. You will have to provide a complete outline of the R code along with the output for each stage (Rmarkdown run). You also need to provide the

total number of records on the dataset and the total number of records for each product across countries across years.

## Part C: Semi-Structured Data (Web)

The UK Food Standards Agency runs the food hygiene rating scheme which aims to evaluate the standards of food hygiene found on the date of the inspection in a restaurant serving food by the local authority. The food hygiene rating sticker looks like this:




The UK government provides an open API in either JSON or XML to download the data and make them available under the following URL:

https://www.food.gov.uk/uk-food-hygiene-rating-data-api

Your job is to write an R script to fetch the ratings dataset from the government website and store it in a format that will enable further analysis. The resulting data frame should capture all XML defined fields from the website. You need to document and articulate every stage in your code and explain your steps clearly.

## Part D: Dashboard with R/Shiny

Using the food hygiene data, create a Shiny dashboard where you depict a navigation scenario for the ratings. You are free to select the scenario that you think that is more appropriate. Your solution needs to include the following:

- A map depicting the locations of the rated companies using the geocode information
- A graphical representation of the rating values as obtained from the parsing of the XML document.

Additional representations and visualizations will be considered as a plus but will not exclude the attenuation of a full mark.