Project 1
Data warehousing

Politecnico di Torino
Advanced data bases

The project consists of (1) designing and querying a data warehouse based on the problem specification and (2) visualizing query results.

1. Problem specifications

Data analysts of a food delivery company are interested in analyzing the deliveries.

The company has a database with the delivery logs. The OLTP database tables of the company are the following (primary keys are underlined).

```
Rider (RiderSSN, Surname, Name, DateBirth)
Transport (trasportId, trasportName)
PaymentMode (paymentModeId, paymentMode, availableDiscount)
Restaurants (restaurantId, restaurantname, restaurantCategory, restaurantAddress)
Delivery (RiderSSN,  DeliveryDate,  DeliveryHour,  DeliveryMinute, restaurantId,
trasportId, paymentModeId, amount, distance, deliveryDuration)
```

For each delivery, the rider and the mode of transport, the restaurant, the payment information and the amount spent, the start delivery time (date, hour, minute) are known. The system also stores the delivery duration (the delivery time for the delivery, in minutes) and the distance from the restaurant to the delivery address.
Table 1 shows a detailed description of the OLTP tables and the field data types.

The data analysts want to analyze efficiently the information about the delivery, the *average revenue* for delivery and the *average delivery time* (in minutes).
In particular, the data warehouse must be designed to efficiently analyze the following information.

- The restaurant name and city, province and region of the restaurant. The restaurant is located in a specific city.
- The category of the restaurant. A restaurant belongs to only one category. There are 5 possible categories (“Indian”, “Italian”, “Pizzeria”, “Chinese/Japanese”, “Other”).
- The date, day of the week, if the day is a holiday or not, month, semester and year of the delivery
- The payment modality. There are four payment modalities (“Bancomat”, “Credit card”, “Cash” and “Satispay”)
- The transport mode (“bike”, “scooter”, “car”).

In particular, the analysts want to analyze the following situations:

- Daily, monthly, and yearly revenue, average delivery time (derived from attribute deliveryDuration) and the number of deliveries for each restaurant.
- Daily, monthly, and yearly revenue, the average delivery time and the number of deliveries for each transport mode.
- The total revenue, the number of deliveries and the average delivery time for payment method.
- The total revenue and number of deliveries for each day of the week and restaurant.

Table	Description
Rider (fiscalCodeRider VARCHAR(20) NOT NULL, Surname VARCHAR(20) NOT NULL, Name VARCHAR(20) NOT NULL, BirthDate DATE NOT NULL, PRIMARY KEY(RiderSSN))	
Trasport (trasportId INT NOT NULL, trasportName VARCHAR(20) NOT NULL PRIMARY KEY(trasportId));	Different transport mode cardinality(trasportName) = 4
Restaurants (restaurantId INT NOT NULL, restaurantName VARCHAR(20) NOT NULL restaurantCategory VARCHAR(20) NOT NULL, restaurantAddress VARCHAR(20) NOT NULL, PRIMARY KEY(restaurantId));	Restaurants
PaymentMode (paymentModeId INT NOT NULL, paymentMode VARCHAR(20) NOT NULL availableDiscount FLOAT NOT NULL, PRIMARY KEY(trasportId));	
Delivery (fiscalCodeRider VARCHAR(20) NOT NULL, DeliverDate DATE NOT NULL, DeliverHour INT NOT NULL, DeliverMinute INT NOT NULL, restaurantId VARCHAR(20) NOT NULL, trasportId INT NOT NULL, paymentModeId INT NOT NULL, amount FLOAT NOT NULL, distance FLOAT NOT NULL, deliveryDuration INT NOT NULL, PRIMARY KEY(RiderSSN,DeliverDate,DeliverHour, DeliverMinute), FOREIGN KEY(RiderSSN) REFERENCES Rider(RiderSSN) ON DELETE CASCADE, FOREIGN KEY(trasportId) REFERENCES Trasport(trasportId) ON DELETE CASCADE, FOREIGN KEY(paymentModeId) REFERENCES PaymentMode(paymentModeId) ON DELETE CASCADE, FOREIGN KEY(restaurantId) REFERENCES Restaurant(restaurantId) ON DELETE CASCADE);	Deliveries

Table 1 - OLTP - Source data base with single delivery information

2. Design

Design the data warehouse to address the specifications and to efficiently answer the provided frequent queries. Draw the conceptual schema of the data warehouse and the logical schema (fact and dimension tables).

3. Querying the data warehouse

Create and populate the tables according to the designed logical schema with sample data.
Use these tables as sources for the following queries.

a. For each day, select the total revenue and the average revenue per delivery. Sort the result by date.
b. Select the yearly revenue and the total number of deliveries for each restaurant. Sort the results by descending yearly revenue.
c. Separately for each transport mode and year, select the total number of deliveries and the average time for delivery.
d. Consider only the deliveries with “bike” as transport mode. Separately for each month and restaurant, select the total revenue and the average delivery time.
e. Separately for date and transport mode, select the total revenue and the maximum delivery time.
f. Separately for each month, select the total revenue and the average daily revenue.

4. Reporting and visualization

We want to visualize some of the results of the previous analysis.

Analyze the data by creating the following visualization, deriving from query *a*-*f* of Section 3. Then, explore and create new visualizations to find interesting insights on your own.

A. For each day, select the total revenue. Sort the result by date. Visualize the result as a time series.
B. Visualize the yearly revenue for each restaurant with a bar plot representation. Sort the results by descending yearly revenue.
C. Separately for each transport mode and year, visualize the total number of deliveries with a bar plot representation.
D. For queries *d* to *f*, explore and define new visualizations to find interesting insights.

5. Project assignment

Report format.
Write a report up to 4 pages with your proposed design of the data warehouse, the SQL queries and their visualization.
The report must be structured using the following sections and subsections:
1. Problem overview
2. Proposed design of the data warehouse
   (a) conceptual schema
   (b) logical schema
with a discussion of the design choices.
3. Querying the data warehouse
Report the SQL queries *a* to *f* in Section 3.
4. Analysis of the results
Describe the visualizations *A* to *C.* For the additional visualizations of interest (point D), you should motivate the choice of the visualization and describe the concept(s) you want to highlight.

Visualization generation - Software.
Any material used for the visualization must be submitted or linked.
In the case of Google Data Studio, integrate the link to the Google Data Studio in the report.
Analogously, if visualizations are generated through Google Colab and python (e.g. matplotlib, plotly libraries), you must integrate the link in the report.
Finally, if any other tool (e.g. python script) is used, please generate a zip file and submit it along with the report.

Constraints on the report.
The report must comply with the following constraints:

- The report must be generated using the standard IEEE conference template, available in LATEX and Word format. You are strongly encouraged to use the LaTeX version (either locally, or on Overleaf).
- The report must follow the division in the listed sections and subsections.
- The report must be, at most, 4 pages long.

Submission
Upload the report in pdf format. The archive must be uploaded to the "Portale della Didattica", under the Homework section (“Elaborati”). Please use as file name “Project_1_DW.pdf”.
If Google Data Studio or Google Colab/python are not used for the visualization, upload the files generating the visualizations as a zip file as “Project_1_DW_additional_material.zip”.

Deadline. 29th March 2021

For any doubts or problems, please contact Eliana Pastor (at eliana.pastor@polito.it)