SYSTEMS AND METHODS FOR BIG AND UNSTRUCTURED DATA

Prof. Marco Brambilla

POLITECNICO MILANO 1863

August 30, 2022

Last Name _____ First Name _____ Codice Persona _____

A travelling agency stores data about its customers. Each customer is defined by name, surname, birthdate, address, personal id number, email, and mobile phone number. Customers pay the agency to organise their travels. Each travel request is described by a unique identifier, maximum budget of the customer (the fee own to the agency is not considered here), starting date, end date, number of people, and general description provided by the customer about the trip they would like to do (e.g., “I would like to visit maritime cities in south Italy”). Each travel is assigned to one of the workers of the agency. Each worker is described by name, surname, birthdate, address, personal id number, email, mobile phone number, and an IBAN. The worker then arranges all the things needed for the travel proposal, namely they take care of:

- Transportation – described by departure place, departure date, departure time, arrival place, arrival date, arrival time, price, means of transportation
- Stay – described by name of the structure, address, arrival date, departure date, check-in time, check-out time, price, list of services provided by the structure (e.g., wifi, air conditioning, etc.), rating out of 5 stars (only if the structure is a hotel), customers rating, description.
- Events – described by start date, start time, end date, end time, description, price, address

When the travel proposal is complete, the agency contacts the customer and sends them the preview. Each proposal has a status attribute (with default value “WIP”) and a unique identifier. If the customer accepts the proposal, its status is updated to “accepted”. If the customer refuses the proposal, its status is updated to “to be revised”. As soon as the proposal is accepted, the agency takes care of providing the customer with their invoice described by unique identifier, date, the data of the customer, and the final price (which includes the budget for the travel and the money owed to the agency). For simplicity, assume that each travel request can only have one travel proposal which is edited whenever it is rejected.

Exercise 1 (3 PT)

1.1. Describe the conceptual model of the data using an Entity-Relationship model. Focus on conceptualising the problem and avoiding redundancy. Add appropriate entities, relationships, cardinalities, and attributes (2 PT).

ER model:

1.2. In the table below, identify which parts of the model you would implement in different database solutions (relational or non-relational, specifying the type of non-relational). Briefly motivate the choices. (1 PT)

#	ENTITIES / RELATIONSHIPS	DB TYPE	MOTIVATION
1			
2			
3			

Exercise 2 (5 PT)
Consider the entities customer, travel request, travel proposal, and agency worker (and their relationships) from the ER model and suppose you want to store the respective data instances in a graph database. Sketch a graph model example describing the nodes, main attributes, and edges. Either show an example graph or a graph with types. (1 PT)

2.2. Write a Cypher query to extract the agency workers who are currently working on more than 3 travel proposals that were requested by customers born after 12/07/1979. (2 PT)

2.3. Write a Cypher query to extract the list of all the travel request for which the budget is greater than 20'000 € whose agency worker has at least 5 accepted travelling proposals. (2 PT)

SYSTEMS AND METHODS FOR BIG AND UNSTRUCTURED DATA
Prof. Marco Brambilla

POLITECNICO MILANO 1863

August 30, 2022

Last Name _____ First Name _____ Codice Persona _____

Exercise 3 (5 PT)
Suppose you store in a documental database (MongoDB) the lists of travel proposals, transportations, stay, and events. Provide a simple documental representation. How many collections would you define? Provide a simple documental representation. How would you implement the relations between the concepts? [For writing the queries, provide a name to your collections and be coherent with your ER diagram.] (1 PT)

3.1. Write a query to extract all the travel proposals with at least one transportation departing from “Rome” on the “01/01/2022”. (1 PT)

3.2. Write a query to count the number of events with start time later than 9:00 PM and end time earlier than 11:30 PM. (1 PT)

3.3. Write the query to find the sum of the prices of the stays with check-in time earlier than 10:00 AM and rating greater or equal than 4 stars. (2 PT)

Exercise 4 (3 PT)
Suppose you store an Elasticsearch index of the agency invoices, coherently with your ER diagram.

4.1. Provide the complete mapping of the index (i.e., field name, field type, the structure of the mapping, etc.) (1 PT)

PUT …

4.2. Write the complete query to extract all the invoices with final price greater than 2’000 €, whose customer name is “Niccolò”, prioritising those whose surname is “Macchiavelli”.