SYSTEMS AND METHODS FOR BIG AND UNSTRUCTURED DATA

Prof. Marco Brambilla

POLITECNICO
MILANO 1863

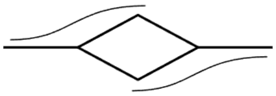
January 17, 2022

Last Name ___________________ First Name __________ Codice Persona _________

Suppose you need a data storage solution for supporting the information system of a media production company. The company needs to track the production process of media products (shows of various types: TV programs, series, movies). Each show has a category (comedy, drama, etc.). The company keeps track of all its shows and of the people on its payroll. Each person has their personal data, bio, and detailed history of their activities with the company and other companies: the shows they worked in, their role, their paychecks, the period of work covered (date from-to). Each show has a title, description, production year, list of reviews, list of locations where the show was shot and is linked to all the people working at it. Both shows and people may win some awards. These are recorded too (type and date). Shows are also described by the various expenses recorded (equipment, services, rentals, and so on), with date, type of expense, invoice, amount and name of the company providing the service or product. For each show, all events are recorded in detail to keep track of the advancement of the production (each step of the writing, casting, shooting, contracts signed, and so on). Events may amount to tens per day. Shows have a script, which consists of the complete writing of the show's organisation (i.e., what it's expected to happen, and how; which actor says what and who does what; and so on). The company also includes staff workers not associated with shows. Information about them is also available.

Exercise 1 (3 PT)

1.1. Describe the conceptual model of the data using an Entity-Relationship model. Focus on the conceptualisation of the problem and avoid redundancy (when possible). To avoid confusion, highlight the read direction for the relationships by adding the arrow's tip to the lines in the picture on the right (1,5 PT).



1.2. Highlight 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. (0,5 PT)

PART	MAIN ELEMENTS	MOTIVATION
part 1		
part 2		
part 3		

1.1. ER MODEL: (make sure it’s readable and tidy)

1.3 Represent a graph schema for at least four different nodes, their respective relations and their attributes that you would implement in a graph DB. Be coherent with the names you provided in the ER diagram. (0,5 PT)

1.4 Provide an example of a document for the *Show* entity and the entities it contains. For each entity, provide its attributes at least once within the document. Be coherent with the name you provided in your ER diagram and the data type. (0,5 PT)

Exercise 2 (4 PT)

Suppose you store in a graph database in Neo4j the lists of people, shows, and their connections. We advise you to provide a simple representation of the graph database if you haven’t already done it during the 1st Exercise. Be coherent with your ER diagram.

2.1. Write a Cypher query that finds the pairs of actors who acted (together) in the same show and have the same age. (1 PT)

2.2. Write a Cypher query that finds the pairs of shows produced by the same producer and were produced in the same year. (1,5 PT)

2.3. Write a Cypher query that finds the people who received an award that acted in a show that received at least 2 awards in 2021 and in which at most one shooting production is included. (1,5 PT)

Last Name ___________________ First Name __________ Codice Persona _________

Exercise 3 (3 PT)

Suppose you store in a documental database in MongoDB the lists of shows, awards, locations and reviews and their connections. We advise you to provide a simple document representation if you haven't already done it during the 1st Exercise. Write the whole query, naming your collection. Be coherent with your ER diagram.

3.1. Write a query to collect 5 shows produced in 2021 in which at least one of the expenses is of type "services". (1,5 PT)

3.2. Write a query to collect the number of awards received by each show whose average review vote/rating is greater than 4. (1,5 PT)

Exercise 4 (3 PT)
Suppose you store an Elasticsearch index of the reviews. Include an attribute representing the show's title (i.e., a sort of an “external key”). Be coherent with your ER diagram. Write the whole query for each of the cases below, naming your index.

4.1. Provide the complete mapping of the reviews index (i.e., field name, field type, the structure of the mapping, etc.) (0,5 PT)

4.2. Write a query to extract the count of positive reviews (i.e., vote/rating greater or equal to 4) received by each show (1 PT)

4.3. Write a query to extract the list of all the reviews whose text contains the word “good”, ordered by Elasticsearch return score. (1,5 PT)

Exercise 5 (3 PT)

The media production company has employed you to manage a Cassandra database. A colleague of yours provides you with a Cassandra script (i.e., a series of operations) to set up the database. The operations included in the script are

- the creation of a keyspace named “staff_management_system”,
- the use of the keyspace named “staff_management_system”,
- the creation of a table named “staff_members”,
- the manual insertion of a tuple to test the database,
- the upload of a list of tuples from a file named “staff_members.txt”,
- execution of two queries as a final test.

Your task is to check that the flow of operations has been properly coded. If **any** mistake is found, mark it and provide the solution in the space provided by re-writing the command or describing how to fix it. There could be no errors, or one or multiple errors in each step. Errors can be of multiple types.

```
CREATE KEYSPACE “staff_management_system” WITH replication =
{‘class’: ‘SimpleStrategy’, ‘replication_factor’: 15};
```

```
USE staff_management_system;
```

```
CREATE TABLE staff_members (
      name string,
      surname string,
      personal_id string,
      age varint,
      PRIMARY KEY (personal_id)
);
```

```
INSERT INTO staff_members (name, personal_id, surname, age) VALUES
(“Francesco”, FRNRSS95E12F675T, “Rossi”, 19)
```

```
UPLOAD “D:/Program Files/Cassandra/Data/staff_members.txt”;
```

```
SELECT name, surname
FROM staff_management_system
WHERE personal_id = “FRNTCC95E12F675T”
```

```
SELECT COUNT(*)
FROM staff_management_system
WHERE name = “Francesco”
```

……