**Unstructured and Streaming Data Engineering 2020-21 04.02.2021**

|  |  |
| --- | --- |
| **Surname:** | **I** |
| **Name:** |
| **Codice persona:** |

Exercise 1. Key-value DB (5 points + 0.5)

A maintenance and service management company supports public administrations in the management of public properties (parks, streets, and so on). The company needs to record the notification of maintenance needs and how they are addressed. Citizens can submit issues in the system which can be of different types (urgent maintenance, cleaning services, damage repair), and are described by a title, city, address, and description, possibly including photos.

City administrators can review the notifications and approve them. Approved ones need to be addressed, therefore the company assigns a maintenance team to the issue and assigns a schedule for the operations. The team may be composed of employees, including a team leader plus one or more team members, plus a set of vehicles and of devices (cleaning machines, lawn mowers, and so on). At the end, a report on the work done is filed into the system. Information about the employees, vehicles and devices is stored in the system.

1) Describe the data at high level showing a conceptual model, for instance an ER model. (2)

2) Explain if and how you could use a key-value DB for the problem. Which parts/elements could be covered by a key-value solution? Describe which key you would use for each element. (2)

3) Write a command that adds a new citizen in the DB and a command that retrieves its data from the DB (you may assume you can use any key-value technology you like). (1)

4) Would you classify the interaction with the citizens as crowdsourcing? why? (bonus: 0.5)

Exercise 2. Graph DB (6 points)

Suppose you want to re-design the problem of exercise 1 as a graph-database implementation.

1) How would you structure the graph? Show it by drawing a model including examples of all the relevant concepts and relations. Remember to add types/labels to every item, together with important attributes (2)

2) Write a graph query using Cypher that finds employees that have worked both in tasks, that are either of type cleaning or repair. [more precisely, in at least one task of type cleaning and at least one of type repair] (1.5)

3) Write a graph query using Cypher that finds employees that worked for tasks in more than 10 different cities (1.5)

4) Write a command that adds a new issue and connects it to an existing citizen. (1)

Exercise 3. Documental DB (6 points)

Suppose you want to record in a document-based storage the reports of works done (as described in exercise 1). The report must include: the reference to the issue, address, date, start time, end time, the team structure (as a subdocument including the leader and the members), the list of consumables used (with name, brand, quantity used), a description of the issue and work done, and possible request for follow-up interventions. Assume you can use a DB technology like MongoDB.

1. Represent as a JSON an example document that supports these needs. (1.5)

2. Write the following query: Find the documents that describe interventions in the city of Milano on January 29. (1.5)

3. Write the following query: Find the documents that describe interventions that consumed more than 5 consumables (meaning, total quantity of consumables being at least 5, counting all the consumables in the intervention). (1.5)

4. Write a command that inserts a new report in the collection. The report must contain at least 3 features, 2 team members, and 1 consumable. (1.5)

Exercise 4. Streaming (10 points)

Suppose you want to monitor with a stream processing engine a group of robots used for picking and placing goods in an Industry 4.0 storehouse.

Each robotic arm sends events reporting its status: ready to pick the good, good grasped, moving the good, and placing the good. Several Force-Sensing Resistors measure the stress levels of the robotic arm. If the stress level is between 0 and 6, the robot is safely operating. If it is between 7 and 8, a controller should raise a warning. If it is above 9, a controller should stop the robot.

In a continuous query language of your choice among EPL, flux, KSQL, and Spark Structured Streaming propose:

* How to model the streaming data generated by the robotic arms. (1)
* A continuous query that emits the max stress for each arm. (1)
* A continuous query that emits the average stress level between a pick and a place. (2)
* A continuous query that returns the robotic arms that: (4)
  + in less than 10 second
  + picked a good while safely operating,
  + moved it while the controller was raising a warning, and
  + placed it while safely operating again.
* A continuous query that monitors the results of the previous one and counts how many times each robotic arm is present in the stream over a tumbling window of 1 min (2).

Explain your choices in detail.

Exercise 5. EPL (4 points)

Suppose you receive the following stream of events: A1@0,C1@1,B1@2,B2@3,A2@4,A3@5,B3@6,A4@6,B4@10.

Note that A3@6 denotes an event of type A identified by the number 3 that is received at time 6.

Given the pattern: every((A -> B and not C) where timer:within(3 sec))

Which are the events that trigger the matching? (0.5) Why? (1.5)

Which are the events that do not trigger the matching? (0.5) Why? (1.5)