Duration: 3 hours

Exam Advanced Database and Big Data
II.2414

June 2023

Before starting, carefully read the following instructions:

- This is a **closed-book exam**, meaning that you cannot use additional resources or notes (**No Internet, no Devices**).
- You **MUST return this document** along with your answers and it is strictly forbidden to leave the classroom with any document (original or copy)!
- Read questions carefully and answer all parts as completely as possible.
- This exam has a duration of **three hours.**
- You must fill answer in appropriate **empty blocks after each answer**. If there is not enough space, you can add additional sheets.
- **Write your name** in the footer of each page

Duration: 3 hours

Part I: Advanced DB (15 points)

Lesson questions:

1. **What does ACID mean?**

2. **What is a Transaction in database?**

3. **What is a view, indicate a main situation where views are useful?**

4. **How does Transaction handles concurrency?**

5. **Is the response time (in seconds) a good metric to analyse queries? Why?**

6. **What are the differences between roles and permission in database access control?**

7. **What is a lock in a database?**

8. Exercise: SQL Queries (5 points)

Here is the context for this exercise: A student may have study several courses. A course may have been studied by several students. The type of a course can be "Science", "Languages", "Cultural", "Sport"...

Here is our data model:	• Student (sid: integer, sname: string, city: string, age: real) • Study (sid: integer, cid: integer, endDate: date) • Course (cid: integer, cname: string, type: string, hours: integer)

8.1. Write an SQL query: Find students younger than 10 that have studied a "Languages" course (No NATURAL JOIN allowed!! (And Explain why it is not recommended to use NATURAL JOIN).

8.2. Write an SQL query: Display how many courses have been studied by each students and in the same query display only students who have studied at least 10 courses.

8.3. **Write an SQL query: Find students that are CURRENTLY STUDYING a Sport course and also a Science one - In SQL, date of today is “NOW()”.**

8.4. **Write an SQL query: Find all students that have studied all courses.**

8.5. **Write an SQL query: For each course, display the top 3 city the most represented.**

Duration: 3 hours

9. Exercise: Normalize the following method applying normalization techniques

We propose the following relation 1st FN (First Normal form):

Full Name	**Physical Address**	**Movies Rented**	**Salutation**
Janet Jones	First street Plot No 4	Pirates of the Caribbean, Clash of the Titans	Ms.
Robert Phil	3rd Street 34	Forgetting Sarah Marshal, Daddy’s Little Girls	Mr.
Robert Phil	5TH Avenue	Clash of the Titans	Mr.

The functional dependencies are the following:

- Full Name => Physical Address
- Full Name => Salutation
- Full Name => Movie Rented

According to the normalization process, **decompose this relation in 3rd FN (third Normal Form).**

Duration: 3 hours

10. Exercise: How to optimize this SQL query?

We have 2 tables of clients and sales of a clothes store company:

- Client (userId: string, firstname: string, lastname: string, birthday: date, address: string)
- Sale (salesId: string, userId: string, amount: int, productId: string, soldAt: date)

In the database, there is currently 100 million sales for last 10 years an 100K clients.

We want to find the name of the top 100 buyers in March 2022.

You should **write the query and suggest several ways to optimize the query.**

Duration: 3 hours

Part II: Big Data (20 points)

11. **What are the types of NoSQL databases. For each type, mention a tool and a use case?**

12. **While dealing with Cassandra tool, what is the difference between cluster key / Partition Key and primary key?**

13. **What does C.A.P Theorem mean?**

14. What is the difference between Data Lake and Data Warehouse?

15. What is the difference between Elastic and Kibana?

16. What are good practices for resources optimization in Data Lake

17. **What is the interest of orchestration with Airflow?**

18. **Explain the process of a Map Reduce (cite the role of Map, Shuffle and Reduce functions)?**

19. **What is the difference between a stage, an application, a task and a job in Spark?**

20. What is the difference between Transformation and Action in Spark?

21. What is a DAG in a Graph Database?

22. What is the difference between ETL and ELT in Data Lake?

23. Data Lake: What are good layers in a Data Lake?

24. What is the Java library behind the scenes used in Elastic?

25. What is difference between shard and index in Elastic?

26. Spark: What is wrong in the following code? (Assuming anyTransformMethod exist)

```
df = spark.read.csv(“data.csv”)
df.map(anyTransformMethod)
print(df.count())
df.write.parquet(“output_file.parquet”)
```

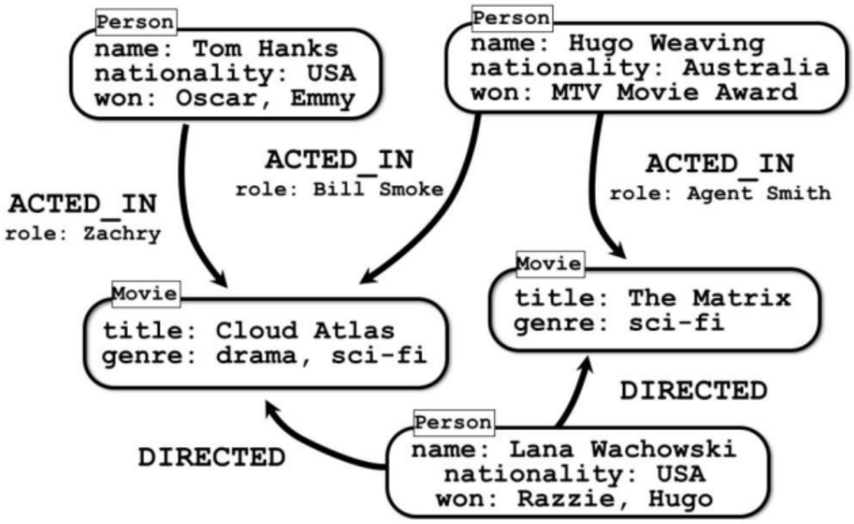
27. Cassandra: Why did the following query returned <u>‘If you want to execute this query, use ALLOW FILTERING’</u>?

```
SELECT * FROM table3 WHERE table3.country = ‘FRANCE’
```

Duration: 3 hours

28. Exercise: Querying with Cypher

We work with the movies graph database described in the following illustration:



Write Cypher queries:

- **Q1:** Display actors who played the movie “The Matrix”

- **Q2:** Actors who worked with “Tom Hanks”