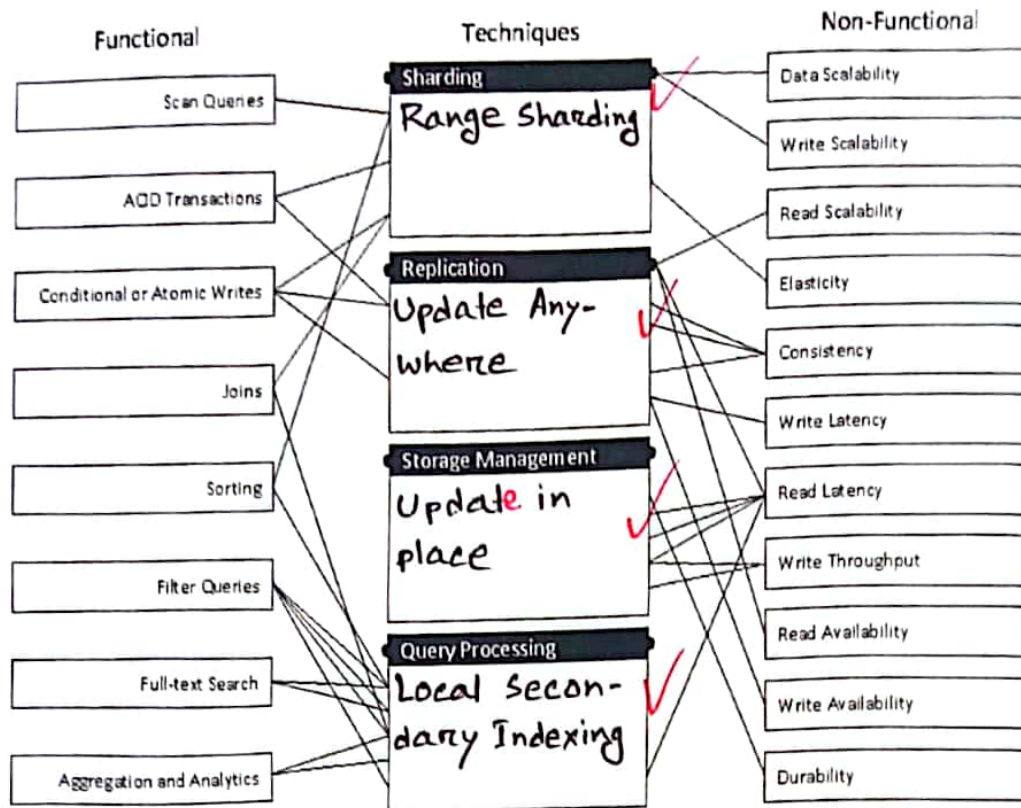


The following figure (adapted from the paper *NoSQL and decision guidance* by Felix Gessert et al.) shows the so-called NoSQL techniques that connect the functional and non-functional system properties are missing.



④

- (a) (4 points) Provide one example for each technique and write them in the empty boxes of the figure above.

③

- (b) (6 points) A clinic wants to build a data lake. In the process of requirements engineering, you start to map the different data formats, data types and requirements. You collect this information using a table for four different kinds of data. Please complete the table below with the missing classifications.

Data set	Code			
Patient data (Names and Medication)	1			
Medical images	2			
Payment / Transactions	3			
Laboratory results	4			
Data set code	1	2	3	4
Structured (S) or Unstructured (U)?	S	U	S	<del>U</del>
Availability: High (H) or Medium (M)?	<del>M</del>	M	H	M
Query Pattern: Ad-Hoc (A) or Complex Analytics? (C)	A	<del>A/C</del>	<del>A/C</del>	A/C
Consistency: Strong (S) or Eventual (E)?	<del>E</del>	E	S	S/E

## 2. Database Design Scenario

- (a) (10 points) The owner of a car rental comes to a software company in which you started working not a long time ago. She asks you to create a database for her business. The task description is the following: The car rental rents out cars to its customers. Every car has a license plate, model, color and mileage. Every customer has a first name, last name and gender. It is also important for the car rental to know for how long the customer has his or her drivers license, from when until when the car is rented and how much she has to pay. One customer can rent only one car.

Your colleagues proposed the following database design:

Customer

<u>f.n.</u>	<u>l.n.</u>	gender	driving_exp
Matias	Johnson	m	13
Mark	Smith	m	16
Anna	Taylor	f	29

Car

<u>licence_plate</u>	model	mileage
db234e	Mercedes	2100
hn455j	BMW	123000
gh789e	BMW	34523

Rentals

<u>licence_plate</u>	<u>f.n.</u>	<u>l.n.</u>	model	price	from	till
db234e	Anna	Taylor	Mercedes	300	03.04.2020	16.04.2020
db234e	Mark	Smith	Mercedes	40	05.06.2020	10.06.2020
gh789j	Matias	Johnson	BMW	150	25.06.2020	02.07.2020

Prove your colleagues that the provided database design is not optimal. Use the database design criteria discussed during the lectures. Name at least 3 criteria and why are they violated.

### 3. Hash-based distribution of data

- (a) (3 points) Your colleague Dave explains the benefits of hash-based data distribution as follows:

Hash-based data distribution provides efficient data retrieval by creating a one-to-one mapping between keys and database locations, allowing direct access to data without searching. In addition, hash-based data distribution can handle complex queries involving joins, range-based access or sorting.

As often, he mixes up things and you need to correct this statement. Explain to him why.

Start your sentence with "Sorry Dave, but ..."


(b) (4 points) Given the following database table structure and data:

OrderID	CustomerName	Product	Quantity
1	John	Apples	20
2	Jane	Oranges	15
3	Mary	Bananas	10
4	John	Oranges	5
5	Jane	Apples	25
6	Mary	Bananas	30

Design a simple hash function that will distribute the data based on the 'CustomerName' field on three available buckets that uses the ASCII letter codes of the string. Discuss how your hash function ensures a uniform distribution of data.



- (e) (3 points) Your team has noticed that under heavy write load, the application's write latency increases. Suspecting that this could be related to memtable flushes, your team decides to modify the memtable cleanup threshold. Explain how changing this threshold could affect the application's performance, and discuss any potential risks associated with this change.

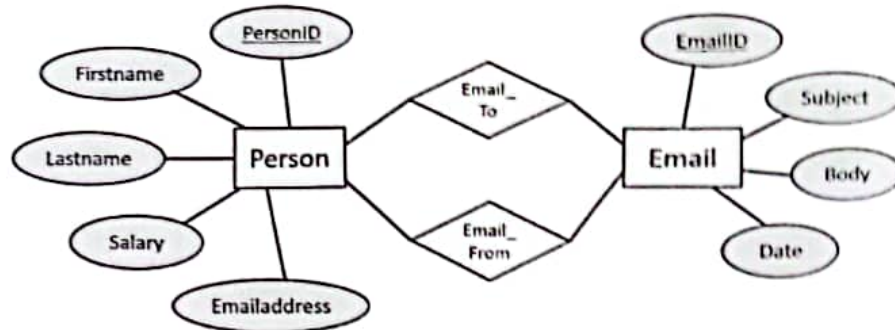


- (f) (1 point) Tick all new requirements that are supported by Extensible record stores.

- ☐ Data structure complexity
- ☒ Schema independence
- ☐ Sparseness
- ☐ Self-descriptiveness
- ☒ Scalability
- ☐ Variability

#### 4. Query Language.

Consider that we have designed a conceptual schema for a company called "Eron". It has two entities: Email and Person with the properties shown in the ER schema below.



For a case study, the same schema has been implemented in different database management systems. Therefore, we use different query languages to execute queries on the data. Consider the query below for the following questions:

```
db. Person . aggregate ([ { $lookup : {
    from : "Email",
    localField : "Emailaddress",
    foreignField : "Email_From",
    as : "emails"
  } }, { $unwind : "$emails"},
{ $project: {"Firstname":1, "emails.Body":1 } } ] )
```



(a) (1 point) What is the name of the query language?

(b) (1 point) In which database management system can it be executed?

(c) (1 point) Does this query return the "Emailaddress" of the sender of the emails?

(d) (2 points) Describe shortly the result of the query.

- (e) (3 points) Your team has noticed that under heavy write load, the application's write latency increases. Suspecting that this could be related to memtable flushes, your team decides to modify the memtable cleanup threshold. Explain how changing this threshold could affect the application's performance, and discuss any potential risks associated with this change.



- (f) (1 point) Tick all new requirements that are supported by Extensible record stores.

- ☐ Data structure complexity
- ☒ Schema independence
- ☐ Sparseness
- ☐ Self-descriptiveness
- ☒ Scalability
- ☐ Variability

### 5. Extensible Record Stores

(a) (1 point) What is the main role of a memtable in Apache Cassandra?

- ☒ Storing data on disk
- ☐ Reading data from sorted data files
- ☐ Serving as a cache for recently accessed data
- ☐ Temporarily storing write operations in memory

(b) (1 point) What triggers a flush operation, moving data from a memtable to the sorted data file?

- ☐ When a data record is updated
- ☐ When a data record is deleted
- ☐ When a read operation is performed
- ☒ When the memtable reaches a certain size limit

(c) (1 point) What happens to the data in the memtable when it is flushed to the sorted data file (SDF)?

- ☐ The data is deleted from the memtable and moved to the SDF
- ☐ The data is duplicated in the SDF and stays in the memtable
- ☒ The data is moved to the SDF and the memtable size is reduced
- ☐ The data is encrypted in the SDF and the original data stays in the memtable

(d) (3 points) Imagine you are a database engineer working on a web application that uses Apache Cassandra as its database. The application is a simple blogging platform where each record is a blog post. Each blog post has a unique identifier, the author's user ID, the post title, the post content, and a timestamp indicating when it was created.

Describe how Apache Cassandra handles write operations when a new blog post is created. In your answer, specify the roles of the memtable and SSTables.



In another database system, the following query was used:

```
MATCH (n:Email) - [:EMAIL_TO] - (p:Person{firstname: "Joe"})  
WHERE n.email_date > "2001-09-01 00:00:00+0000"  
and p.salary>1000 RETURN n,p
```

(e) (1 point) What is the name of this query language?

(f) (1 point) In which database management system can it be executed?

(g) (3 points) Would this query return all the emails sent by Joe after the date "2001-09-01 00:00:00+0000"? If not, what should be changed to return the desired result?

## 6. General Knowledge

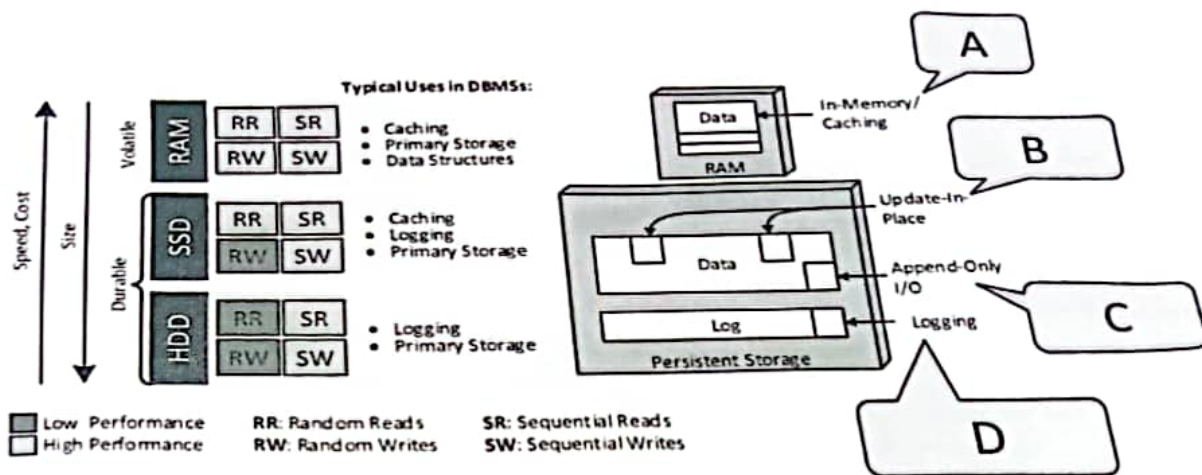
- (a) (4 points) During the requirements analysis, customer interviews give you different statements about what they expect from the system. Please identify and name which property of the database system it refers to and fill the table below. Please use the term we introduced in the lecture.

Customer Statement	Database Property / Support for
"Is there a way to make sure my database won't crash or become sluggish as more users use my application?"	
"My data has a lot of blank spaces, and I need a database that won't store them all as it's taking up too much storage."	
"How can I ensure my database won't become a bottleneck for my application's performance as it gets more popular?"	
"What happens if my data exceeds the database's capacity? Is there a way to prevent that?"	
"I'm concerned about future changes to my data schema. Can the database adapt easily?"	
"I'm looking for a database that includes metadata to describe the meaning of the data stored."	
"I'm looking for a database that can handle data with changing formats or attributes."	
"Is there a way to make sure my data is protected from any failures or errors in the database?"	

- (b) (2 points) Explain the differences between the consistency definition in CAP and in

(c) (4 points) The following figure (from a talk of Wolfram Wingerath) gives an architectural overview on typical uses in database management systems. Techniques like in-memory data management or caching, Update-in-place, Append-Only-I/O, and Logging are typically combined to improve the overall performance. To highlight the specific benefits of these four techniques, we'd like to add short statements.

## Storage Management



Which statements could be added to this figure to summarize the benefits of these specific techniques?

Statement A (for In-Memory/Caching):

Statement B (for Update-in-Place):

Statement C (for Append-Only I/O):

## 7. Graph Databases

- (a) (3 points) Define a property graph and describe its main components.

---

---

---

---

---

---

---

---

---

---

- (b) (3 points) Suppose you have a property graph representing a social network, where nodes represent users and edges represent friendships. Each user node has properties such as "name," "age," and "location." Write a Cypher query to find all users who are 30 years old and are located in New York City.

- (c) (4 points) Give two cases where graph databases outperform traditional relational databases and explain why.

---

---

---

---

---

---

---

---

---

---