

Exam notes

Functional dependencies

Trivial FD: $X \rightarrow Y$ is trivial if: $Y \subset X$.

Non-trivial FD: see above.

Superkey: Superset of a candidate key.

Candidate key: Minimal set of a key.

Primary key: Candidate key chosen by designer.

Prime attributes: Attributes contained in candidate key.

Closure of FDs: Set of all dependencies that include F and all those that can be inferred from F . Denoted as F^+ .

Closure of set of attributes: Maximum set of attributes that can be inferred from the provided set of attributes.

Normalization

Database design principle to

1. *reduce redundancy*
2. *avoid complexities*
3. *organize data in a consistent way.*
4. *eliminate duplicates*
but loses out on

- query efficiency

1NF: No multivalued attributes, i.e. only simple values.

2NF: Parts of the candidate key must not be functionally dependent on non-prime attributes.

- **Example:** $\{\mathbf{A}, \mathbf{B}\} : \overbrace{\mathbf{A} \rightarrow C, D, E}^{\text{illegal}}$
 \Rightarrow No arrows from part of candidate key!

Decompose: Make new table containing LHS and RHS. Remove RHS from origin table.

3NF: Non-prime attributes must not be functionally determined by other non-prime attributes.

- Equivalent to no transitive dependencies: $A \rightarrow \overbrace{B \rightarrow C}^{\text{illegal}}$
 \Rightarrow No arrows starting from non-prime attributes!

Decompose: Make new table containing LHS and RHS. Remove RHS from origin table.

BCNF: No dependencies from non-key attributes to key attributes.

- Example: $\{A, B\} \rightarrow \overbrace{C \rightarrow A}^{\text{illegal}}$
 \Rightarrow No arrows pointing towards prime attributes!

Decompose: Remove one of the prime-attributes!

NB: BCNF is lossless, but it is not dependency preserving!

Losslessness: Decomposition is lossless if we can recover initial table by performing multiple joins.

Dependency preservation: We do not lose dependencies. True if all dependencies can be inferred from the current set.

Triggers

- Can not modify the database schema; solely DML (*Data Manipulation Language*) for data-level operations.

Invoked *automatically*.

Defined in terms of the event *that invokes it*, and the *action it performs*.

May operate either BEFORE or AFTER the execution of the event that invokes it.

Storage

Main memory vs. disk:

- Data access from disk is typically 2 orders of magnitude slower.
- Data in main memory is **volatile**, while **non-volatile** for disk.
- Disk is solely mechanical moving parts, while main memory is completely electronic.

Each time we read from disk, we retrieve a *block* of records.

Size of these *blocks* is fixed, but depends on the OS.

Indexes

Index: data structure that facilitates quicker access to a data.

- *can be* used for both main memory and disk!
- *stored in a data file*.

Primary index: Indexes on the primary key (ordered on key).

- --> **Dense index**: Has exactly one index entry for each search key value.
- --> **Sparse index**: Has fewer index entries than search key values.

Clustering index: Indexes on a non-key field (ordered on field).

- --> One index entry of each (distinct!) value of the field.
- --> Each index points to first data block of records for search key.

Secondary index: Not ordered on the index's search key (purely a mess!)

Multi-level index: Structure of index on index until all entries of the top-level structure fit into 1 disk block.

- --> Pin top-level index in main memory (RAM).

B^+ -trees

- *multi-leveled* indexing structure
- tailored for disk-based data organization: *aligns with disk block sizes, so very efficient for disk storage and access.*
- grows horizontally by splitting the root.

↓ Operation	Average	Worst case
Search	$O(\log n)$	$O(\log n)$
Insert	$O(\log n)$	$O(\log n)$
Delete	$O(\log n)$	$O(\log n)$

Balanced: all paths from root to a leaf have the same length.

- --> guarantees good search performance!

Authorization

Preparing SQL statements they are compiled **only once**, but **executed multiple times**!

SQL injection attacks: Using dynamic SQL queries with user inputs, people can inject malicious statements!

- **Fix**: Bind inputs to variables and then **sanitize** them!

Sanitize: Filters matching the input to patterns and stripping it from invalid patterns.

The Database Administrator (**DBA**) can **grant permissions**:

```
GRANT *privileges* ON *object* TO *identity* WITH GRANT OPTION;
```

Transitive granting: Allows users to grant other users privileges: `WITH GRANT OPTION`.

There are the following *privileges*:

SELECT :

INSERT :

DELETE :

UPDATE :

REFERENCES : *the right to include an attribute in a foreign key.*

Revoking privileges can be done using

```
REVOKE *privileges* ON *object* FROM *identities* <CASCADE/RESTRICT>
```

where `<CASCADE/RESTRICT>` revolves around *transitive granted privileges*.

To **create roles**

```
CREATE ROLE Paymaster;  
GRANT UPDATE(salary) ON Payroll TO Paymaster;  
GRANT Paymaster TO amoeller WITH GRANT OPTION;
```

NoSQL

≡ *Not Only SQL*.

Motivation

- Huge amounts of data
- High performance (response time)
- Consistency less important
- Scalability (relational model *too restrictive!*)
 - ⇒ Often done using **distributed databases**!

Data-model: "*Master-slave replication or master-master replication and sharding.*"

Commonly *hashing* and *range partitions* is used to localize data.

Categories of systems:

1. Document-based: **MongoDB**

- --> **Normalized**: *Decompose into several documents of similar structure and content*
- --> **Denormalized**: *All information in one document*

1. Key-value stores: **Voldemort** (builds on **DynamoDB**)
2. Column-based / wide-column: **Apache HBase**
3. Graph-based: **Neo4j**

CRUD: Create, Read, Update, Delete (+ Search) operations provided through API.

CAP properties: Goals when replicating data across distributed network:

- **C**: Consistency (same value among replicas)
- **A**: Availability (successful read/writes)
- **P**: Partition tolerance (robust under network failure)

Note: In DBMS *consistency* was integrity constraints, now we only want identical replicas!

CAP theorem: Impossible to guarantee all three: C, A, P.

- --> NoSQL systems slack on **C**: consistency.

Keywords:

- *No schema required*, instead *semi-structured* data such as JSON, XML!
- *Less powerful query languages!*
- May provide *multiple version storage*.
- *Vertically* scalable (not horizontally)

Recovery strategies

Idempotent: *If a system crashes during the recovery stage, the new recovery must still give correct results.*

Cascading rollback: *A transaction fails or aborts, and as a result, all other transactions that read data modified by the failed transaction must also be rolled back.*

Checkpointing: process where the system periodically writes all modified (*dirty*) pages from memory to disk.

- improves **recovery** efficiency
- ensures a **consistent** state can be restored after unexpected crash.

Shadow-paging: copy-on-write technique for avoiding in-place updates of pages

- when a page is modified (*dirty*), the system writes changes to a **new (shadow) page** instead of overwriting the old.
- upon commit the page table pointer is switched to the new page (**atomic!**).
≈ no-undo / no-redo

⇒ *possible to use both techniques, but often redundant because both handle recovery well.*

Write ahead logging (WAL): maintain “before image” in log (main-memory!), flush to disk before overwritten with “after image” on disk

Undo: rollback changes of *uncommitted* transactions

Redo: reapply changes of *committed* transactions after a crash.

Steal: *uncommitted* dirty pages can be written, so you need to **undo** them.

- e.g. to save space in main memory you flush the dirty pages more often.

Force: *committed* pages are immediately written (forced!) to disk, so you do not need to repeat them (**no-redo**).

- generally not favorable because it leads to a ton of continuously costly I/Os

	Steal	No-steal
Force	Undo / no-redo	No-undo / no-redo
No-force	Undo / redo	No-undo / redo

Undo/redo:

1. Undo all transactions that has a log entry of “start” but no “commit”.
2. Redo all transactions that has a long entry of “start” and “commit”.

ARIES: *Recovery algorithm implemented in many IBM related databases.*

Based on 3 key ideas:

1. Write Ahead Logging: *Log changes **before** writing them to disk.*
2. Repeat history during redo: **Reapply all actions, including uncommitted, to **reconstruct state****
3. Logging changes during undo: *When undoing, **log each undo action**, so crash during recovery is safe (**idempotent**)*

Using a 3-phase recovery:

1. Analysis: *Identify dirty pages and active transactions.*
2. Redo: *Repeat history.*
3. Undo: *Roll back any uncommitted transactions using the log.*

Transactions

- Multiuser DBMS systems on single-threaded CPUs are interleaved.
- Provides mechanism for logical units of database processing.

Database access operations: Done by reading database item into a *program variable*, and then writing the *program variable value* to a database item.

Def.= (Transaction): A unit of work defined by a `BEGIN TRANSACTION` and `END TRANSACTION`.

Isolation level: Defines degree to which transactions are *isolated*!

1. **SERIALIZABLE** : guarantees transactions behave as though they were serial.
2. **REPEATABLE READ** : read and write locks on rows (not on ranges!)
--> no dirty reads!
--> *phantoms*: new rows can still be inserted
3. **READ COMMITTED** : read and write locks on rows, but read locks released immediately after reading.
--> no dirty reads, but because read locks are released you can get different values.
4. **READ UNCOMMITTED** : no read locks!
--> very risky!

(!) Isolation levels are a *personal choice*, which affects how *you* see the database state!

Isolation level (highest to lowest!)	Dirty reads allowed?	Unrepeatable reads allowed?	Phantom reads allowed?	Concurrency
SERIALIZABLE	No	No	No	Low
REPEATABLE READ	No	No	Yes	Medium
READ COMMITTED	No	Yes	Yes	High
READ UNCOMMITTED	Yes	Yes	Yes	Very High

Overview of problems:

1. **Dirty reads**:
 - A transaction updates a database item and then fails for some reason.
 - Updated item is accessed by another transaction before it is changed back to its original value.
2. **Unrepeatable reads**:
 - A transaction reads the same item twice.
 - Another transaction changes the value between first and second read
3. **Lost update**:
 - Two transactions access same item rendering its value incorrect.
4. **Phantom reads**:
 - A transaction queries for an item twice.

- Another transaction inserts an item matching the query between first and second read.

Concurrency control

Schedule: Order in which transactions or operations are performed.

- --> *Example:* for

$$\begin{array}{l} T_1 : R_1(X), W_1(X) \\ T_2 : R_2(X) \end{array}$$

can be interleaved in many different ways!

$$\begin{array}{ll} 1 : R_1(X), W_1(X), R_2(X) & \Rightarrow \text{Serial} \\ 2 : R_2(X), R_1(X), W_1(X) & \Rightarrow \text{Serial} \\ 3 : R_1(X), R_2(X), W_1(X) & \Rightarrow \text{Concurrent} \end{array}$$

- ---> Relevant for: *consistency* and *isolation*!

Conflict: Occurs when you have

1. actions from at least 2 transactions
2. one of them is a write
3. they are on the same attribute

Conflict equivalent: If transactions I and J do not conflict, we may swap their order to produce a new schedule S' . We then denote S and S' are **conflict equivalent**.

- *Note:* conflicting operations must maintain their relative order!

Serializable: A transaction that is **(conflict) equivalent** to a serial schedule.

Precedence graphs:

4. For every transaction make a node.
5. Draw directed edge from i to j if

$$\begin{array}{l} RW : R_i(X), \dots, W_j(X) \\ WR : W_i(X), \dots, R_j(X) \\ WW : W_i(X), \dots, W_j(X) \end{array}$$

3. If there is a cycle (despite labeling of attributes!) then it is NOT serializable.
4. To obtain the serializable schedule, follow the arrows from end to end. Note; there may be several.

There are different types of schedules:

1. **Cascadeless schedule:** Mitigates cascading rollbacks

"Every transaction reads only items written by committed transactions"

2. **Strict schedule**: Mitigates cascading rollbacks

"Transaction can neither read or write item until the last transaction that wrote it has committed."

⇒ Transactions can only read committed values!

3. **Recoverable schedule**:

"No transaction commits until all transactions that have written an item that it read have committed"

Locking

Binary locks (*mutex locks*): Ensures safe access to data by having

- (1) Locked: *data is not accessible by other*
- (0) Unlocked: *data is available for access*

Important: Once locked, no other transaction can read or write the data item!

This is different than shared- and exclusive locks.

Description:

A lock needs to be acquired before any read or write, and must unlock again after completing all operations; it unlocks only items it currently has a lock on, and requests locks only for items it currently has no lock on.

Shared- and exclusive locks (*multiple-mode locking*):

- Shared lock (S): *Allows multiple transactions to read the data item.*
 - *Note:* Shared locks are unique for transactions and hence are not "shared"!
- Exclusive lock (X): *Allows one transaction to write (and also read, if needed).*
- Unlock (U): ...
- Lock-conversion: *A transaction can upgrade a shared- to an exclusive lock, or downgrade an exclusive- to a shared lock.*

Description:

A transaction must acquire a shared or exclusive lock prior to reading, and an exclusive lock prior to writing, and it must unlock again after all its operations are completed; it issues lock requests only for items it does not already hold a lock on, and it issues unlock requests only for items it holds a lock on.

2-phase locking (2PL) :

1. Phase 1 (*growing phase*) :

- Transaction *may request* locks

- Transactions *may not unlock* locks
- (Can convert (upgrade): $S(X) \rightarrow E(X)$)

2. Phase 2 (shrinking phase) :

- Transactions *may not request* locks
- Transactions *may unlock* locks
- (Can convert (downgrade): $E(X) \rightarrow S(X)$)

(!) When the first locks is released, the transaction moves from phase 1 to phase 2.

⇒

Deadlocks: Cycle of transactions all waiting for another to unlock a data item.

Validation (optimistic) concurrency control: Do the work on *local copies only*, before committing check if there is any issues; if so abort and restart, otherwise write changes to database.

1. **Read phase:**

- Read and write operations are made in local workspace (*copy of relevant data only!*)

2. **Validation phase:** (check serializability!)

- Assign timestamps when starting validation, check for R/W, W/W conflicts from *older* transactions.

3. **Write phase:**

- Local changes are written to database if validation is successful.

Validation is based on 3 steps: For all other recent: $TS(T_j) < TS(T_i)$

1. T_j executes all 3 phases before T_i begins (*serial execution!*)
2. T_j completes its write phase before T_i starts its write phase, and does not change any items read.

$$\text{write_set}(T_j) \cap \text{read_set}(T_i) = \emptyset$$

3. T_j completes its read phase before T_i starts its read phase, and does not change any item that is read or written to.

$$\text{write_set}(T_j) \cap \text{read_set}(T_i) = \emptyset$$

$$\text{write_set}(T_j) \cap \text{write_set}(T_i) = \emptyset$$

⇒ and one of these steps must be true!

==Multi-version 2-phase locking w. certify locks ==: Allows a transaction T' to read a data item X while it is write-locked by a conflicting transaction T .

- ⇒ 3 locks: *read*, *write* (now not exclusive, reads possible), and *certify* (fully exclusive).

1. When T wants to write X , it creates a second version X' after obtaining a *write-lock*.
2. Other transactions continue reading X .

3. When T is ready to commit, it obtains a *certify-lock* on X' .
4. Committed version X becomes X' .
5. T releases *certify-lock*.

Core idea:

Instead of just read and write (shared and exclusive, respectively), we now have three locks, read, write, certify (shared, shared with reads, exclusive, respectively). With this, we can allow some reads while an item is write locked, but transactions may have delay when waiting to certify and deadlocks may occur.

Snapshot isolation: Each transaction sees a snapshot of the database at the time it started.

- \Rightarrow No read locks, only write locks!

Granularity locking

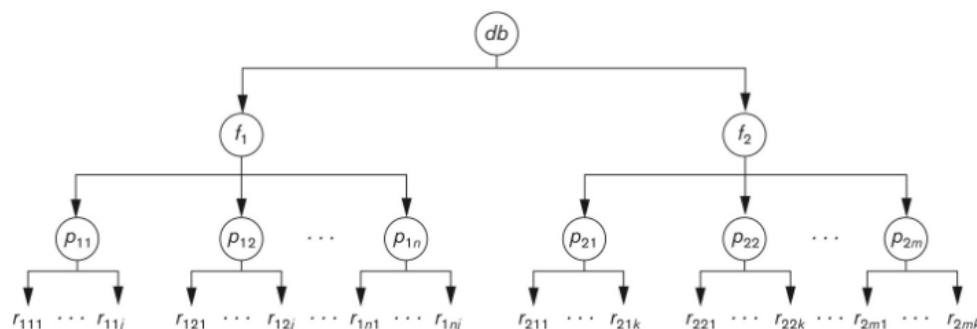
Def. (*Granularity*): Size of a lockable unit of data

- \rightarrow *coarse*: entire database
- \rightarrow *fine*: tuple or attribute of relation
- \Rightarrow significantly affects concurrency performance:
- low degree of concurrency for coarse granularity!
- high degree of concurrency for fine granularity!

Rule of thumb:

"We want to choose granularity to reflect the typical access size of transactions"

Granularity hierarchy: B-tree structure of coarse and finer granularity:



Core idea: Writing to B-tree family indexes could lock a lot of pages, as all index access starts at the root. A more optimistic approach holds shared locks on non-leaf nodes, and exclusive locks on the leaf, unless a split becomes necessary.

To manage this hierarchy we introduce 3 more locks:

- **Intention-shared (IS)**: indicates that a shared lock(s) will be requested on some descendant nodes(s)
- **Intention-exclusive (IX)**: indicates that an exclusive lock(s) will be requested on some descendant node(s)

- **Shared-intention-exclusive (SIX)**: indicates that the current node is locked in shared mode but an exclusive lock(s) will be requested on some descendant nodes(s)
 ⇒ typically when you want to read a node and want an exclusive lock on a descendant node.

Timestamping

- Every transaction gets a *timestamp*; $ST_1 \rightarrow TS(T_1) = 1$.
- Resolves at *what point in time* the transactions should see the state of information in.

Single version:

1. $R(X) : W_{TS}(X) \leq TS(T_i)$

- If the write-timestamp is younger than yours, then you should not be allowed to read.
 ⇒ if allowed: $R_{TS}(X) = \max(R_{TS}(X), TS(T_i))$
 ⇒ else abort and rollback.

2. $W(X) : W_{TS}(X) \leq TS(T_i) \text{ AND } R_{TS}(X) \leq TS(T_i)$

- Same as before, but now also if someone younger has written the value, then you should not be allowed to change it at an earlier time.
 ⇒ if allowed: $W_{TS}(X) = \max(W_{TS}(X), TS(T_i))$
 ⇒ else abort and rollback.

Multi version: Reads will always be okay, but not necessarily for writes.

1. $R(X)$: always allowed, but read version with highest W_{TS} that is $\leq TS(T_i)$

2. $W(X)$: find version with highest W_{TS} that is still $\leq TS(T_i)$.

If $R_{TS} \leq TS(T_i)$, then it is allowed. Then make a new version with $R_{TS} = W_{TS} = TS(T_i)$.

(!) Storage (RAM and disk) is required to maintain multiple versions!

Misc