

PR 2
Database Design

Alejandro Pérez Bueno

Dec 30, 2025


Table of Contents

1	Data Loading	2
1.1	Exercise 1a	2
1.2	Exercise 1b	2
1.3	Exercise 1c	4
1.4	Exercise 1d	5
1.5	Exercise 1e	6
2	SQL	7
2.1	Exercise 2a	7
2.2	Exercise 2b	8
2.3	Exercise 2c	10
2.4	Exercise 2d	11
3	DDL	13
3.1	Exercise 3a	13
3.2	Exercise 3b	14
3.3	Exercise 3c	15
3.4	Exercise 3d	16
4	Indexes and Views	17
4.1	Exercise 4a	17
4.2	Exercise 4b	18
4.3	Question 4c	20
4.4	Question 4d	21
4.5	Exercise 4e	23
5	Model review and optimization	28
5.1	Question 5a	28
5.2	Exercise 5b	29
5.3	Exercise 5c	30
5.4	Exercise 5d	30
5.5	Exercise 5e	31

Chapter 1

Data Loading

1.1 Exercise 1a



The screenshot shows a PostgreSQL query interface. The query bar contains the following SQL statement:

```
1 SELECT version() VER, system_identifier SI FROM pg_control_system();
```

Below the query bar, there are tabs for "Data Output", "Messages", and "Notifications". The "Data Output" tab is active, showing a table with two columns: "ver" (text) and "si" (bigint). The table contains one row of data.

	ver	si
1	PostgreSQL 18.1 (Ubuntu 18.1-1.pgdg24.04+2) on x86_64-pc-linux-gnu, compiled by gcc (Ubuntu 13.3.0-6ubuntu2~24.04) 13.3.0, 64-bit	7589554545532578310

Figure 1.1: Version and System Identifier

1.2 Exercise 1b

Warning

We are asked to create a schema named `S_Q254`, but everywhere else in the subject the schema is actually called `S_Q25`. I created the `S_Q25` schema instead, as otherwise I would get unexpected empty queries and errors.

1.2.1 Missing attribute in Invoice data

The data definition and the UML diagram contains 4 attributes, but the `invoice.csv` file contains 5. Its second column refers to an `incidentID`, which as mentioned in the subject, is what incidents are related to. This field should be a foreign key to the Incident class.

1.2.2 Missing attribute in InvoiceLine data

As before, there is a mismatch between the table of the attributes and what is actually shown in the `invoiceline.csv` file. The first column in it is missing in the table. The diagram shows a 1 to 1..* relationship between `Invoice` and `InvoiceLine`. It even includes the note “`InvoiceLine` is weak respect `Invoice`”. This means an `InvoiceLine` cannot exist without an `Invoice`. Therefore, we must add an `invoiceCode` column as a foreign key.

Furthermore, since the entity is weak, the primary key of `InvoiceLine` should not only be the `Line Number`, but rather a composition of (`invoiceCode`, `lineNumber`)

1.2.3 SQL table creation

1.2.3.1 Invoice table

```
CREATE TABLE invoice (  
    code VARCHAR(50) NOT NULL,  
    incident_id VARCHAR(50) NOT NULL,  
    invoiceDate DATE NOT NULL,  
    amount NUMERIC(12,2) NOT NULL,  
    amountWithVAT NUMERIC(12,2) NOT NULL,  
  
    CONSTRAINT pk_invoice PRIMARY KEY (code),  
    CONSTRAINT fk_invoice_incident FOREIGN KEY (incident_id)  
        REFERENCES incident(code)  
        ON UPDATE CASCADE  
        ON DELETE RESTRICT  
);
```

1.2.3.2 InvoiceLine table

```
CREATE TABLE invoiceline (  
    invoice_code VARCHAR(50) NOT NULL,  
    lineNumber INTEGER NOT NULL,  
    concept VARCHAR(200) NOT NULL,  
    amount NUMERIC(12,2) NOT NULL,  
    amountWithVAT NUMERIC(12,2) NOT NULL,  
  
    CONSTRAINT pk_invoiceline PRIMARY KEY (invoice_code, lineNumber),  
    CONSTRAINT fk_invoiceline_invoice FOREIGN KEY (invoice_code)  
        REFERENCES invoice (code)  
        ON UPDATE CASCADE  
        ON DELETE CASCADE  
);
```

1.3 Exercise 1c

Tip

In order to run the `COPY` command from the editor, I first had to copy the CSV files to a location the `postgres` user had access to. Thus, I copied them to `/tmp` and changed their permissions and owner as follows:

```
cp invoice.csv invoiceline.csv /tmp/  
chmod 644 /tmp/invoice.csv  
chmod 644 /tmp/invoiceline.csv  
sudo chown postgres:postgres /tmp/invoice.csv  
sudo chown postgres:postgres /tmp/invoiceline.csv
```

1.3.1 Invoice table

```
COPY invoice (code, incident_id, invoiceDate, amount, amountWithVAT)  
FROM '/tmp/invoice.csv'  
WITH (  
    FORMAT CSV,  
    DELIMITER ',',  
    ENCODING 'UTF8'  
);
```

1.3.2 InvoiceLine table

```
COPY invoiceline (invoice_code, lineNumber, concept, amount, amountWithVAT)  
FROM '/tmp/invoiceline.csv'  
WITH (  
    FORMAT CSV,  
    DELIMITER ',',  
    ENCODING 'UTF8'  
);
```

Warning

I initially had issues copying rows to the `invoiceline` table. After some research, I found it had to do with encoding between Linux and Windows. The CSV files were likely created from windows, and when importing them to Linux some special characters appeared in the code, making the invoice codes to mismatch and thus fail with errors similar to the following:

```
ERROR: insert or update on table "invoiceline" violates foreign key constraint
↪ "fk_invoiceline_invoice"
Key (invoice_code)=(INV0467) is not present in table "invoice".
```

I trimmed special characters to fix the issue with the following commands:

```
UPDATE invoice SET code = TRIM(both ' \r\n\t' from code);
UPDATE invoice SET code = regexp_replace(code, '[^a-zA-Z0-9]', '', 'g');
```

After this all rows were properly copied to the invoiceline table.

1.4 Exercise 1d

Query
Query History

```

1  SELECT
2  pg_class.relname AS objectname, pg_class.relkind AS objecttype,
3  pg_class.reltuples, pg_size_pretty(pg_class.relpages::bigint*8*1024) AS size
4  FROM pg_class, pg_catalog.pg_statio_user_tables
5  WHERE
6  pg_catalog.pg_statio_user_tables.relid = pg_class.OID AND
7  SchemaName = 's_q25'
8  ORDER BY relpages DESC;

```

Data Output
Messages
Notifications

+

📄

▼

📋

▼

🗑️

🗄️

⬇️

📈

SQL

	objectname name	objecttype "char"	reltuples real	size text
1	invoiceline	r	24520	3776 kB
2	repair	r	30861	2112 kB
3	invoice	r	9999	2000 kB
4	incident	r	10500	960 kB
5	person	r	10000	848 kB
6	bikinguser	r	9585	568 kB
7	repairshop	r	495	160 kB
8	bicycle	r	2300	128 kB
9	employee	r	1264	72 kB
10	base	r	1000	64 kB
11	maintenancemanag...	r	744	32 kB
12	mechanic	r	502	24 kB

Figure 1.2: Summary

1.5 Exercise 1e

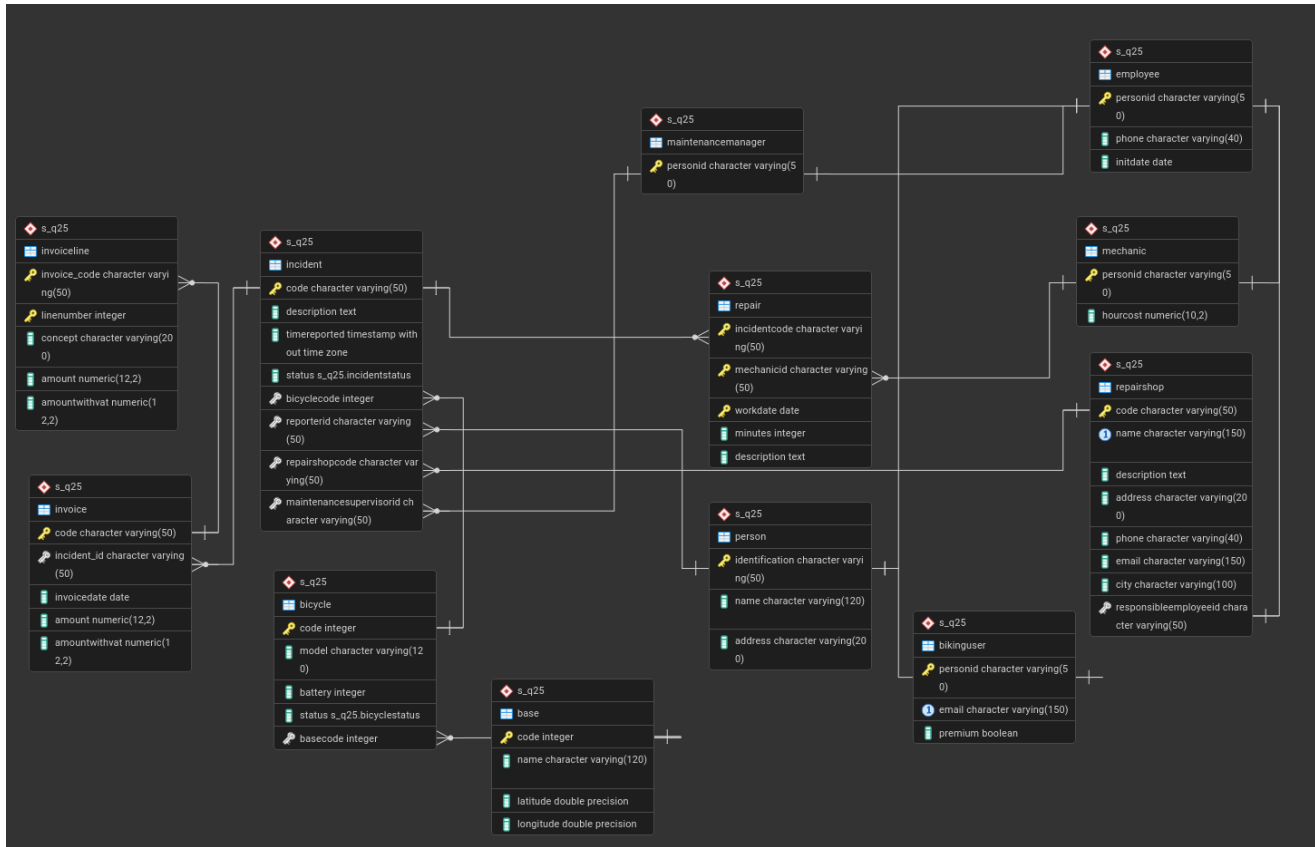


Figure 1.3: ERD Database Diagram

Chapter 2

SQL

2.1 Exercise 2a

2.1.1 Executed SQL command

```
SELECT T2.name AS "Name",  
       T1.phone AS "Phone"  
FROM employee AS T1  
JOIN person AS T2 ON T1.personid = T2.identification  
WHERE EXTRACT(YEAR FROM T1.initdate) = 2021  
      AND EXTRACT(MONTH FROM T1.initdate) = 1;
```


Query Query History

```

1 SELECT T2.name AS "Name",
2       T1.phone AS "Phone"
3 FROM employee AS T1
4 JOIN person AS T2 ON T1.personid = T2.identification
5 WHERE EXTRACT(YEAR FROM T1.initdate) = 2021
6       AND EXTRACT(MONTH FROM T1.initdate) = 1;

```

Data Output Messages Notifications

	Name character varying (120)	Phone character varying (40)
1	Martina Pérez Rodríguez	632345006
2	Alberto Soler Delgado	616001215
3	Isaac Varela Ríos	742000281
4	Javier Fuentes Benítez	673007204
5	Patricia Palacios Cuesta	785000382
6	Víctor Rubio Navarro	616009194
7	Daniel Parra Crespo	742006101
8	Nora Pérez Miralles	785000829
9	Sofía Parra Gil	616005725
10	María Piña Molero	632000435
11	Helena Soto Fuentes	722000227
12	Joel Gallego Ledesma	642000336

Figure 2.1: Screenshot of the result of the SQL command

2.2 Exercise 2b

2.2.1 Executed SQL command

```

SELECT T1.personid,
       T1.hourcost,
       COUNT(T2.mechanicid) AS totalrepairs
FROM mechanic AS T1
JOIN repair AS T2 ON T1.personid = T2.mechanicid
GROUP BY T1.personid, T1.hourcost
ORDER BY totalrepairs DESC
LIMIT 10;

```

Query

Query History

```
1 SELECT T1.personid,
2       T1.hourcost,
3       COUNT(T2.mechanicid) AS totalrepairs
4 FROM mechanic AS T1
5 JOIN repair AS T2 ON T1.personid = T2.mechanicid
6 GROUP BY T1.personid, T1.hourcost
7 ORDER BY totalrepairs DESC
8 LIMIT 10;
```

Data Output

Messages

Notifications

≡+

📄

▼

📋

▼

🗑️

🗄️

⬇️

📈

SQL

	personid [PK] character varying (50)	hourcost numeric (10,2)	totalrepairs bigint
1	P0621	17.40	97
2	P0179	17.37	96
3	P0264	18.19	92
4	P0967	17.61	89
5	P1318	18.66	88
6	P5278	19.53	88
7	P0004	19.15	88
8	P8176	21.60	88
9	P0702	20.23	87
10	P0331	21.32	87

Figure 2.2: Screenshot of the result of the SQL command

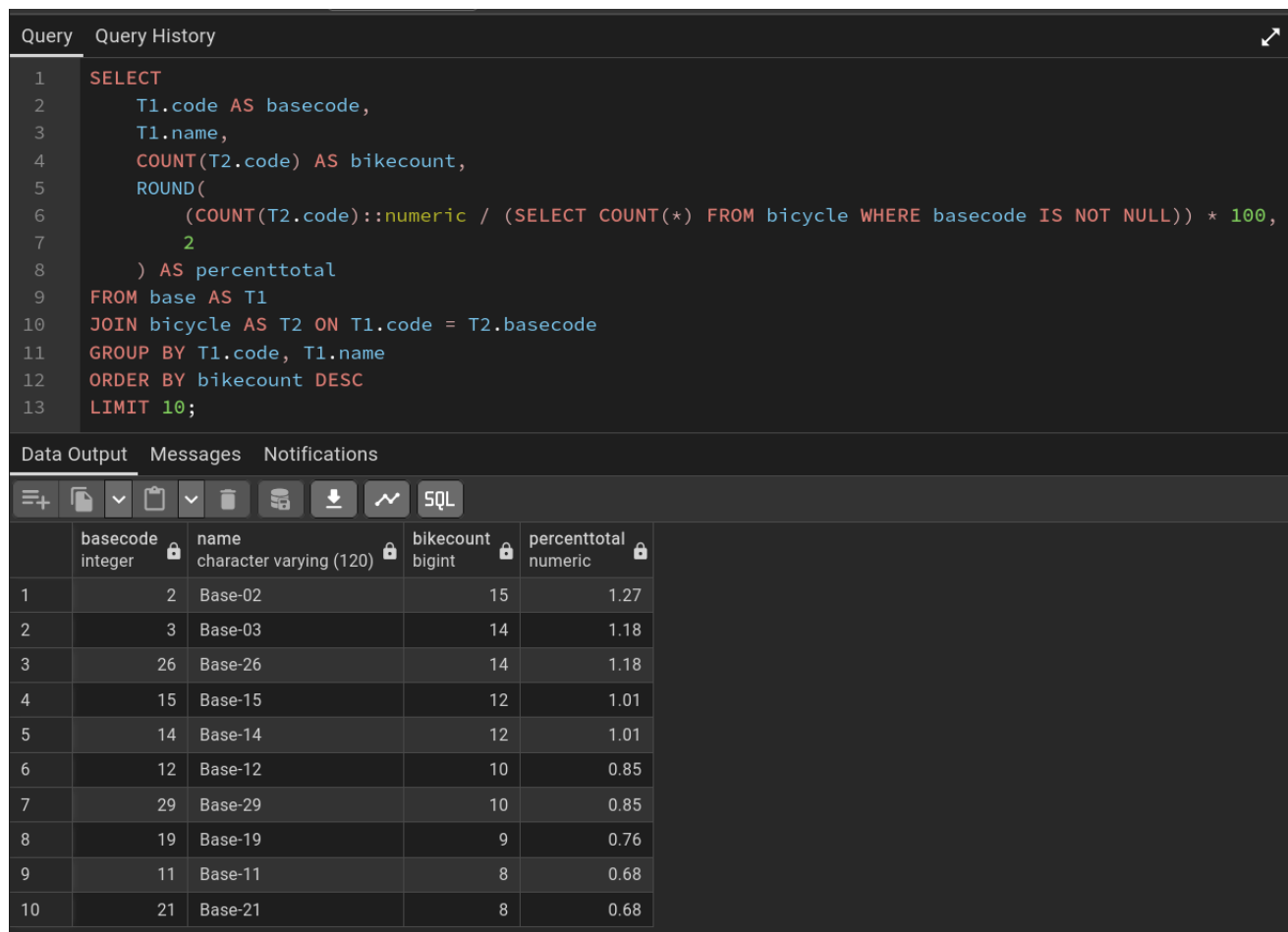
2.3 Exercise 2c

2.3.1 Executed SQL command

```

SELECT
    T1.code AS basecode,
    T1.name,
    COUNT(T2.code) AS bikecount,
    ROUND(
        (COUNT(T2.code)::numeric / (SELECT COUNT(*) FROM bicycle WHERE basecode IS NOT NULL))
        * 100,
        2
    ) AS percenttotal
FROM base AS T1
JOIN bicycle AS T2 ON T1.code = T2.basecode
GROUP BY T1.code, T1.name
ORDER BY bikecount DESC
LIMIT 10;

```



Query Query History

```

1 SELECT
2     T1.code AS basecode,
3     T1.name,
4     COUNT(T2.code) AS bikecount,
5     ROUND(
6         (COUNT(T2.code)::numeric / (SELECT COUNT(*) FROM bicycle WHERE basecode IS NOT NULL)) * 100,
7         2
8     ) AS percenttotal
9 FROM base AS T1
10 JOIN bicycle AS T2 ON T1.code = T2.basecode
11 GROUP BY T1.code, T1.name
12 ORDER BY bikecount DESC
13 LIMIT 10;

```

Data Output Messages Notifications

	basecode integer	name character varying (120)	bikecount bigint	percenttotal numeric
1	2	Base-02	15	1.27
2	3	Base-03	14	1.18
3	26	Base-26	14	1.18
4	15	Base-15	12	1.01
5	14	Base-14	12	1.01
6	12	Base-12	10	0.85
7	29	Base-29	10	0.85
8	19	Base-19	9	0.76
9	11	Base-11	8	0.68
10	21	Base-21	8	0.68

Figure 2.3: Screenshot of the result of the SQL command

2.4 Exercise 2d

2.4.1 Executed SQL command

```
SELECT CAST(EXTRACT(YEAR FROM invoiceDate) AS INTEGER) AS "year",  
       CAST(EXTRACT(MONTH FROM invoiceDate) AS INTEGER) AS "month",  
       SUM(amountWithVAT) AS "totalAmount"  
FROM invoice  
WHERE EXTRACT(YEAR FROM invoiceDate) = 2025  
GROUP BY EXTRACT(YEAR FROM invoiceDate), EXTRACT(MONTH FROM invoiceDate)  
ORDER BY "month";
```


Chapter 3

DDL

3.1 Exercise 3a

3.1.1 Default ‘pending’ value for ‘status’

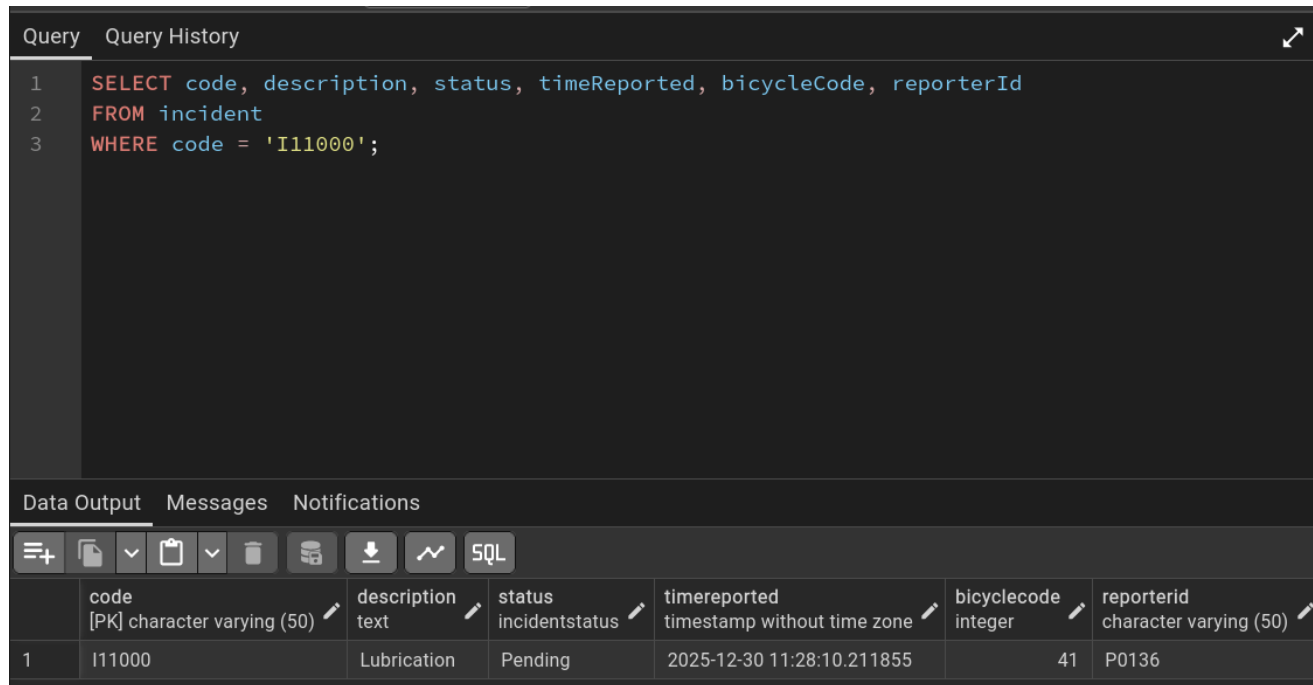
```
ALTER TABLE incident
ALTER COLUMN status SET DEFAULT 'Pending';
```

3.1.2 Add new row

```
INSERT INTO incident (code, description, timeReported, bicycleCode, reporterId)
VALUES ('I11000', 'Lubrication', NOW(), 41, 'P0136');
```

3.1.3 Query table to show status as ‘pending’ by default

```
SELECT code, description, status, timeReported, bicycleCode, reporterId
FROM incident
WHERE code = 'I11000';
```



The screenshot shows a database query interface. At the top, there are tabs for 'Query' and 'Query History'. The 'Query' tab is active, displaying a SQL query:

```
1 SELECT code, description, status, timeReported, bicycleCode, reporterId
2 FROM incident
3 WHERE code = 'I11000';
```

Below the query editor, there are tabs for 'Data Output', 'Messages', and 'Notifications'. The 'Data Output' tab is active, showing a table of results. The table has 7 columns: code, description, status, timeReported, bicyclecode, and reporterid. The data is as follows:

	code [PK] character varying (50)	description text	status incidentstatus	timereported timestamp without time zone	bicyclecode integer	reporterid character varying (50)
1	I11000	Lubrication	Pending	2025-12-30 11:28:10.211855	41	P0136

Figure 3.1: Query result

3.2 Exercise 3b

3.2.1 Make ‘battery’ for ‘bicycle’ to use 2 decimals instead of int

```
ALTER TABLE bicycle
ALTER COLUMN battery TYPE NUMERIC(5,2);
```

3.2.2 Add new row with the ‘battery’ as a float

Warning

The subject says the code should be 301, but that one already exists. I created one with the code 2301, since that is the next available code. It was probably a typo.

```
INSERT INTO bicycle (code, model, battery, status, basecode)
VALUES (2301, 'Model-01', 75.50, 'Available', 4);
```

3.2.3 Query table to show ‘battery’ as float

```
SELECT code, model, battery, status, basecode
FROM bicycle
WHERE code = 2301;
```

Query

Query History

1

2

3

SELECT

code,

model,

battery,

status,

basecode

FROM

bicycle

WHERE

code

=

2301

;

Data Output

Messages

Notifications

≡+

📄

▼

📋

▼

🗑️

🗄️

⬇️

📈

SQL

	<div>code</div> <div>[PK] integer</div>	<div>model</div> <div>character varying (120)</div>	<div>battery</div> <div>numeric (5,2)</div>	<div>status</div> <div>bicyclestatus</div>	<div>basecode</div> <div>integer</div>
1	2301	Model-01	75.50	Available	4

Figure 3.2: Query result

3.3 Exercise 3c

⚠️ Warning

What is requested here was already done as part of Section 1.2.3.

However, this would be the way to insert the constraint:

```
ALTER TABLE invoiceline
ADD CONSTRAINT fk_invoiceline_invoice
FOREIGN KEY (invoice_code)
REFERENCES invoice (code)
ON UPDATE CASCADE
ON DELETE CASCADE;
```


3.4 Exercise 3d

Adding the `UNIQUE` constraint to a column prevents duplicates in its table.

3.4.1 Advantages

1. Reliable data rules: the database itself enforces uniqueness instead of relying on the app to do it. That guards against race conditions when two requests try to insert the same email or username at the same time and ensures duplicates are always rejected.
2. Faster lookups on the constrained column: most relational databases implement `UNIQUE` with a unique index. Tasks like finding a user by email can use that index instead of scanning the whole table, so reads on that field are much quicker.

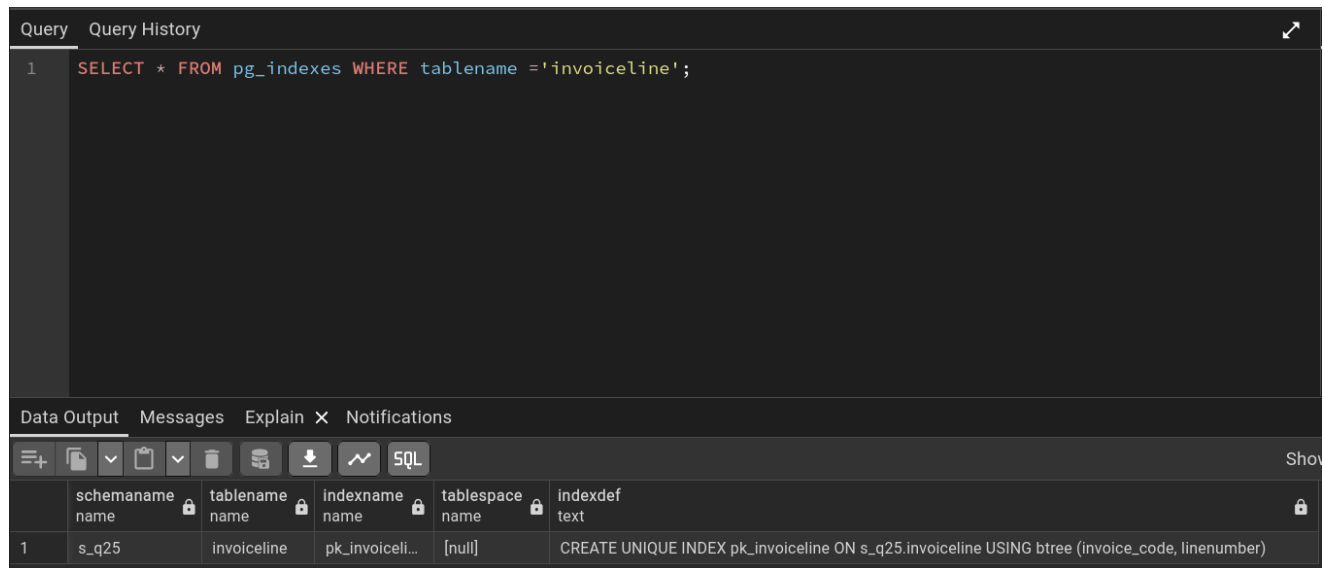
3.4.2 Disadvantages

1. Extra cost on writes: Every `INSERT` or `UPDATE` that affects the unique column must check and maintain the index. That extra work could add latency, which can be a problem in some systems.
2. Less flexibility if requirements change: declaring a value unique could be a rigid rule. If the business later allows duplicates dropping or changing the constraint on a large table can require index rebuilds, which may affect the runtime of the application relying on the DB.

Chapter 4

Indexes and Views

4.1 Exercise 4a



The screenshot shows a database query interface with a dark theme. The top bar has tabs for 'Query' and 'Query History'. The main area contains a SQL query: `SELECT * FROM pg_indexes WHERE tablename = 'invoiceline';`. Below the query editor is a toolbar with icons for various actions like copy, paste, and execute. The bottom section displays the query results in a table format. The table has five columns: `schemaname`, `tablename`, `indexname`, `tablespace`, and `indexdef`. The first row of data shows the index `pk_invoiceline` on the `invoiceline` table in the `s_q25` schema, using the `btree` tablespace. The `indexdef` column contains the full `CREATE INDEX` statement.

	schemaname name	tablename name	indexname name	tablespace name	indexdef text
1	s_q25	invoiceline	pk_invoiceli...	[null]	CREATE UNIQUE INDEX pk_invoiceline ON s_q25.invoiceline USING btree (invoice_code, linenumber)

Figure 4.1: Query result

In `pg_indexes`, the `indexdef` column stores the exact `CREATE INDEX` statement used to build each index.

The command is

```
CREATE UNIQUE INDEX pk_invoiceline ON invoiceline USING btree (invoice_code, linenumber)
```

From this we can tell:

- It is `UNIQUE`, so the pair (`invoice_code`, `linenumber`) can't repeat. In practice, this is the index backing the table's primary key
- The index name is `pk_invoiceline`.

- It uses the **btree** (Balanced Tree) method, PostgreSQL's default, which works well for equality lookups and range queries.
- The key is composite: **invoice_code** + **linenumber**. A line number may repeat across invoices, but not within the same invoice, showing that an invoiceline is identified by its parent invoice together with the line number.

4.2 Exercise 4b

4.2.1 Execution plan

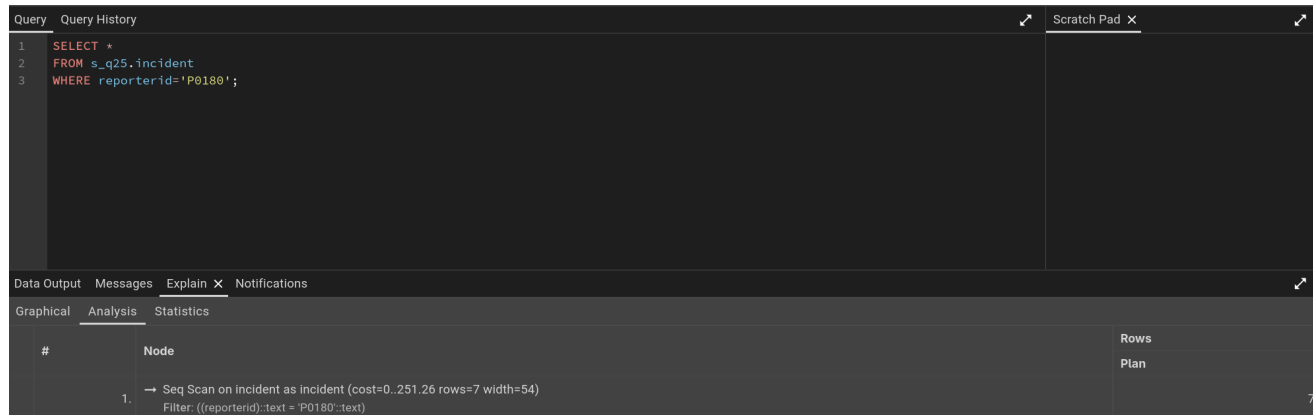


Figure 4.2: Query execution plan

4.2.2 Index creation

```
CREATE INDEX idx_incident_reporterid ON incident (reporterid);
```



Tip

Re-run analyze:

```
ANALYZE incident;
```

4.2.3 New execution plan

The screenshot shows a database interface with a query editor at the top containing the following SQL:

```

1 SELECT *
2 FROM s_q25.incident
3 WHERE reporterid='P0180';

```

Below the query editor, the 'Explain' tab is active, displaying the execution plan. The plan consists of two nodes:

#	Node	Rows Rows X	Actual	Plan	Loops
1.	→ Bitmap Heap Scan on incident as incident (cost=4.34..27.35 rows=7 width=54) (rows=7 loops=1) Recheck Cond: ((reporterid)::text = 'P0180'::text) Heap Blocks: exact=7	1	7	7	1
2.	→ Bitmap Index Scan using idx_incident_reporterid (cost=0.434 rows=7 width=0) (rows=7 loops=1) Index Cond: ((reporterid)::text = 'P0180'::text)	1	7	7	1

Figure 4.3: Query execution plan after index

4.2.4 Performance analysis

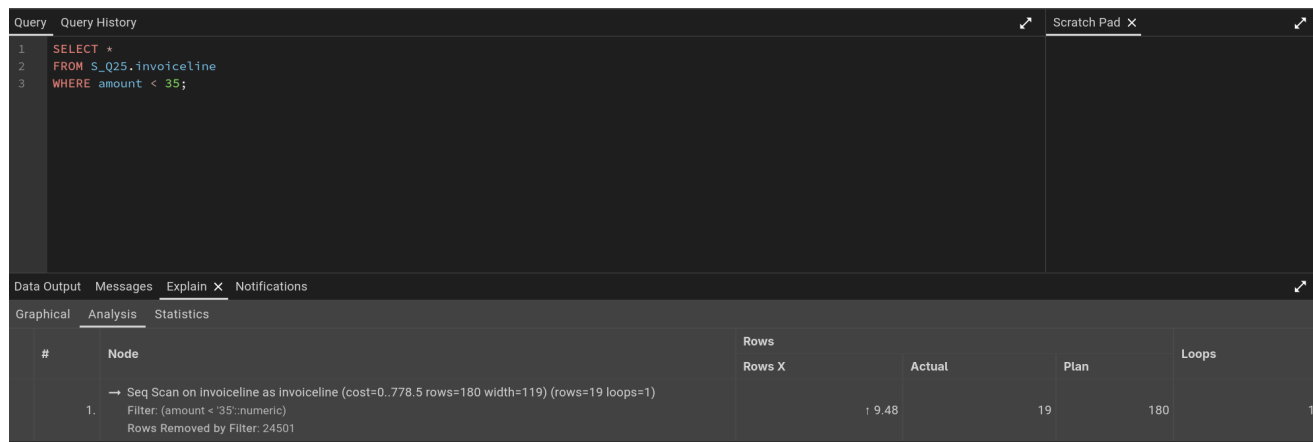
Creating the index clearly sped up the target query. Before, the planner chose a sequential scan with an estimated total cost of 251.26 because every row had to be inspected. Afterward, it used a bitmap index scan with an estimated total cost of 27.35, which is roughly a **90% reduction**.

Regarding the writes, the index in question is on the incident table (`idx_incident_reporterid` on `incident`). Index maintenance happens only when rows in that same table change. Inserts, updates, or deletes on the `Property` table do not touch this index and therefore suffer no overhead from it.

Therefore, keeping this index is the right decision to make. It meaningfully improves read performance for the query on incident, and it has no negative effect on write activity in `Property`.

4.3 Question 4c

4.3.1 Execution plan



The screenshot shows a database interface with a query editor and an execution plan. The query is:

```
1 SELECT *
2 FROM S_Q25.invoiceline
3 WHERE amount < 35;
```

The execution plan is displayed in a table with the following columns: #, Node, Rows, Actual, Plan, and Loops. The plan shows a single node (1) representing a Seq Scan on invoiceline as invoiceline, with a cost of 0..778.5, 180 rows, and 1 loop. The actual execution time was 9.48 seconds, and 180 rows were scanned. The plan also indicates that 24501 rows were removed by the filter.

#	Node	Rows	Actual	Plan	Loops
1.	→ Seq Scan on invoiceline as invoiceline (cost=0..778.5 rows=180 width=119) (rows=19 loops=1) Filter: (amount < 35::numeric) Rows Removed by Filter: 24501	180	9.48	19	1

Figure 4.4: Query execution plan

4.3.2 Index choice

We should use a B-tree index. It's standard in PostgreSQL and shines for range and comparison operators:

- Range lookups: `amount < 35` lets the engine seek to the key near 35 and then scan in order, which is precisely how a B-tree is laid out.
- Numeric column: `amount` is numeric and B-trees are built for ordered types (numbers, timestamps, text) and operators like `<`, `<=`, `=`, `>=`, `>`.
- Discarded indexes:
 - Hash: supports equality only, so it is of no use for ranges.
 - GiST/GIN: aimed at things such as full-text, arrays, or geometry, not plain numeric comparisons.

4.3.3 Index creation

```
CREATE INDEX idx_invoiceline_amount
ON invoiceline USING btree (amount);
```

```
ANALYZE invoiceline;
```

4.3.4 New execution plan

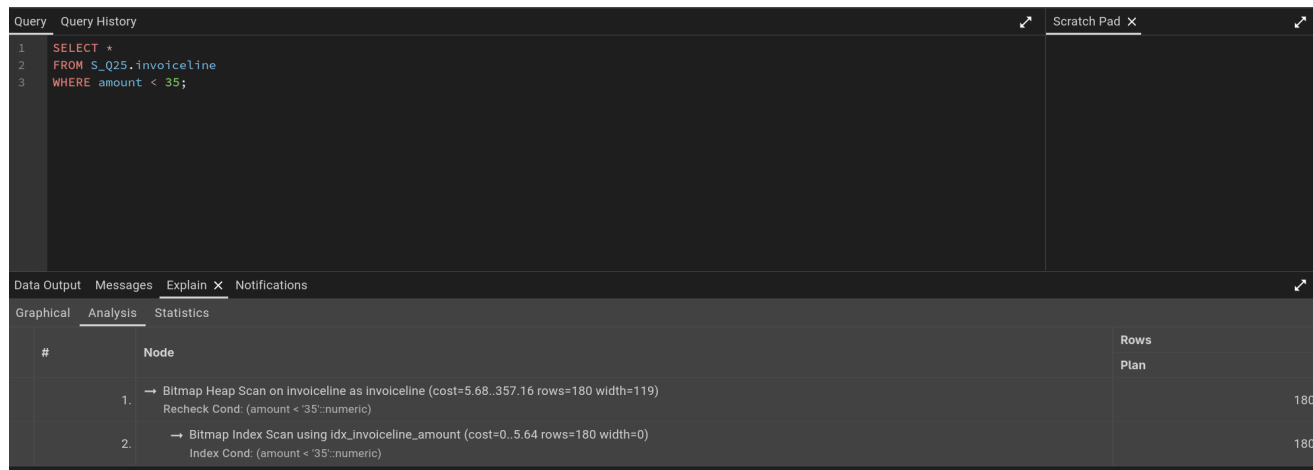


Figure 4.5: Query execution plan after index

4.3.5 Performance analysis

4.3.5.1 Before indexing (sequential scan)

- Estimated cost: 778.5
- The planner chose a full table scan, checking every row for `amount < 35`.
- Filtered out: 24,501 rows to keep 19 actual matches.

4.3.5.2 After indexing

- New estimated cost: 357.16
- The B-tree index was used to find the matching row pointers first (Bitmap Index Scan) and then only those pages were fetched.

4.3.5.3 Verdict

My recommendation is therefore to keep the B-tree the index. Range filters like `amount < 35` are a great fit for this index type, the reduction is well over 50% and the small overhead on inserts is a fair trade for much faster reads on this query pattern.

4.4 Question 4d


4.4.1 View SQL create code

```
CREATE OR REPLACE VIEW VW_IncidentsDetail AS
SELECT
    i.code AS "Incident Code",
    i.timereported AS "Reported Date/Time",
    i.description AS "Description",
```

```
i.repairshopcode AS "Assigned Workshop",
i.maintenancesupervisorid AS "Assigned Supervisor",
b.model AS "Bicycle Model",
i.status AS "Status"
FROM
  incident i
JOIN
  bicycle b ON i.bicyclecode = b.code;
```

4.4.2 View SQL query

```
SELECT
  "Incident Code",
  "Reported Date/Time",
  "Bicycle Model"
FROM
  VW_IncidentsDetail
WHERE
  "Status" = 'Resolved'
ORDER BY
  "Reported Date/Time" DESC
LIMIT 5;
```











Query Query History 

```

1  SELECT
2      "Incident Code",
3      "Reported Date/Time",
4      "Bicycle Model"
5  FROM
6      VW_IncidentsDetail
7  WHERE
8      "Status" = 'Resolved'
9  ORDER BY
10     "Reported Date/Time" DESC
11  LIMIT 5;

```

Data Output Messages Explain X Notifications




	Incident Code character varying (50) 	Reported Date/Time timestamp without time zone 	Bicycle Model character varying (120) 
1	I0325	2025-12-31 15:32:42	Model-13
2	I9147	2025-12-31 06:02:58	Model-14
3	I7763	2025-12-30 23:22:40	Model-11
4	I0393	2025-12-29 14:45:47	Model-04
5	I7081	2025-12-29 03:59:05	Model-11

Figure 4.6: View SQL query Screenshot

4.5 Exercise 4e

4.5.1 View SQL create code

```

CREATE OR REPLACE VIEW VW_CountIncidents2025 AS
SELECT
    rs.city AS "City",

```



```
    COUNT(*) AS "Total Incidents"
FROM
    VW_IncidentsDetail v
JOIN
    repairshop rs ON v."Assigned Workshop" = rs.code
WHERE
    EXTRACT(YEAR FROM v."Reported Date/Time") = 2025
GROUP BY
    rs.city
ORDER BY
    "Total Incidents" DESC,
    "City" ASC;
```

4.5.2 Query to view 10 first records

```
SELECT *
FROM VW_CountIncidents2025
LIMIT 10;
```

Query

Query History

1

2

3

SELECT *

FROM VW_CountIncidents2025

LIMIT 10;

Data Output

Messages

Explain X

Notifications

SQL

	City character varying (100)	Total Incidents bigint
1	Valencia	118
2	Sevilla	102
3	Valladolid	100
4	Bilbao	96
5	Murcia	93
6	La Coruña	91
7	Madrid	87
8	Cadiz	86
9	Santander	83
10	Salamanca	82

Figure 4.7: View SQL query Screenshot

4.5.3 Materialized View

Warning

We were asked to create the Materialized View of the view `VW_CountIncidentsByCity2025`, but it is not referenced before this statement. Therefore, will use the previously created `VW_CountIncidents2025` view.

```
CREATE MATERIALIZED VIEW V_VW_CountIncidents2025 AS
SELECT *
FROM VW_CountIncidents2025;
```

4.5.4 Performance analysis

4.5.4.1 Standard view (`VW_CountIncidents2025`)

- Estimated cost: 379.69
- 16-step plan:
 - Sequential scans of incident (step 11), bicycle (step 14), and repairshop (step 16)
 - Several hash joins to connect those tables (steps 7 and 9)
 - Sorting and a GroupAggregate (step 3) to count incidents by city
- A view is just a stored query, so every `SELECT` reruns all joins, filters, and aggregations over the base tables.

4.5.4.2 Materialized view (`V_VW_CountIncidents2025`)

- Estimated cost: 13.20
- One sequential scan of the physical table `v_vw_countincidents2025`
- The materialized view stores the query's result on disk; reading it doesn't touch incident, bicycle, or repairshop or recompute joins/counts.

All in all, querying the materialized view is almost 30x cheaper because the expensive work—joining and counting—was done beforehand and only needs to be redone when you refresh it.

4.5.5 Advantages and Disadvantages

Given that 2025 is still underway, here's a concise take on using a materialized view for this data.

4.5.5.1 Advantages

- The plan cost drops quite significantly, so year-to-date dashboards open quickly without redoing heavy joins.
- The analytical reads shift to the view, so inserts/updates on incident aren't competing with reporting queries.

4.5.5.2 Disadvantages

- A materialized view is a snapshot. Anything added after the last refresh won't appear, so counts drift until you refresh.
- You need a refresh schedule. Refreshing too often eats into performance; refreshing rarely makes the numbers too old to trust.

4.5.6 Updating view

Materialized views do not update automatically. One must regularly refresh it with the following:

```
REFRESH MATERIALIZED VIEW V_VW_CountIncidents2025;
```

Chapter 5

Model review and optimization

5.1 Question 5a

5.1.1 hourCost SQL query

```
SELECT
    'hourCost' AS column_name,
    'Mechanic' AS table_name,
    COUNT(*) AS number_of_rows,
    COUNT(DISTINCT hourcost) AS distinct_values,
    AVG(pg_column_size(hourcost)) AS average_occupation
FROM mechanic;
```

5.1.2 Space occupied by tables

Query run:

```
SELECT
pg_class.relname AS objectname, pg_class.relkind AS objecttype,
pg_class.reltuples, pg_size_pretty(pg_class.relpages::bigint*8*1024) AS size
FROM pg_class, pg_catalog.pg_statio_user_tables
WHERE
pg_catalog.pg_statio_user_tables.relid = pg_class.OID AND
SchemaName = 's_q25'
```

Query

Query History

1

SELECT

2

pg_class.relname AS objectname, pg_class.relkind AS objecttype,

3

pg_class.reltuples, pg_size_pretty(pg_class.relpages::bigint*8*1024) AS size

4

FROM pg_class, pg_catalog.pg_statio_user_tables

5

WHERE

6

pg_catalog.pg_statio_user_tables.relid = pg_class.OID AND

7

SchemaName = 's_q25'

8

ORDER BY relpages DESC;

Data Output

Messages

Explain X

Notifications

</

Figure 5.1: Query result

5.2 Exercise 5b

The most appropriate column to move to a separate table is **concept from the InvoiceLine table** due to the following:

- Very low cardinality: 11 unique values across 24,520 rows, so the same strings repeat thousands of times.
- The strings are fairly large (about 95 bytes on average). If those 11 concepts live in a small lookup table (ConceptID, Description), InvoiceLine would keep only a 4-byte foreign key.

How the others compare:

- **hourCost** (Mechanic): 342 distinct out of 502 rows. High variety, little duplication—normalizing won't help

much.

- **description** (Incident): 20 distinct in 10,501 rows, but the text is short (~15 bytes), so the payoff is limited.
- **description** (RepairShop): Only 3 distinct and fairly long (~92 bytes), so it's a reasonable candidate, but with just 495 rows the total gain is small next to InvoiceLine.

5.3 Exercise 5c

```
CREATE TABLE invoice_concept (  
  id SERIAL PRIMARY KEY,  
  description VARCHAR(200) NOT NULL,  
  
  CONSTRAINT uq_invoice_concept_description UNIQUE (description)  
);
```

Added the `uq_invoice_concept_description` constraint to ensure descriptions are unique so normalization works properly.

5.4 Exercise 5d

5.4.1 Populate the table

```
INSERT INTO invoice_concept (description)  
SELECT DISTINCT concept  
FROM invoiceline;
```

5.4.2 Query the table

```
SELECT * FROM invoice_concept;
```

Query

Query History

1

SELECT * FROM invoice_concept;

Data Output

Messages

Explain

X

Notifications

Figure 5.2: Query result

5.5 Exercise 5e

- Add a new column to store the identifier

```
ALTER TABLE invoiceline
ADD COLUMN concept_id INTEGER;
```

- Update the values of the new column

```
UPDATE invoiceline il
SET concept_id = ic.id
FROM invoice_concept ic
WHERE il.concept = ic.description;
```

- Delete the original column


```
ALTER TABLE invoiceline  
DROP COLUMN concept;
```

- Define the foreign key constraint

```
ALTER TABLE invoiceline  
ADD CONSTRAINT fk_invoiceline_concept  
FOREIGN KEY (concept_id)  
REFERENCES invoice_concept(id);
```

- Verification: Show the combined values (First 10 records)

```
SELECT  
    il.invoice_code,  
    il.linenummer,  
    ic.description AS concept,  
    il.amount,  
    il.amountwithvat  
FROM  
    invoiceline il  
JOIN  
    invoice_concept ic ON il.concept_id = ic.id  
LIMIT 10;
```

The screenshot shows a database query interface. At the top, there are tabs for 'Query' and 'Query History'. Below the tabs, a SQL query is entered: `SELECT * FROM invoice_concept;`. Below the query editor, there are tabs for 'Data Output', 'Messages', 'Explain', and 'Notifications'. The 'Data Output' tab is selected, showing a table with two columns: 'id' (integer, primary key) and 'description' (character varying (200)). The table contains 11 rows of data.

	id [PK] integer	description character varying (200)
1	1	To fix a flat tire remove the wheel check for objects replace or patch the tube then reinstall and inflate.
2	2	Inspect the bike frame and connections for cracks corrosion or looseness after impacts for safety.
3	3	Lubricate the bike chain and moving parts to reduce wear and ensure smooth operation.
4	4	Brake systems need cleaning pad alignment and wear checks to keep stopping power and safety optimal.
5	5	Routine bicycle repair starts with inspecting for loose bolts worn brake pads or flat tires to ensure a safe ride.
6	6	Realign the wheels with a truing stand to fix wobbles caused by uneven spokes.
7	7	After repairs take a test ride to confirm effectiveness and adjust for reliable bike performance.
8	8	Smooth gear shifts depend on derailleur alignment cable tension and lubrication to prevent slipping.
9	9	Replace worn-out brake pads and cables to restore full stopping power.
10	10	Clean derailleurs and gears with a brush and degreaser to prolong drivetrain life.
11	11	A slipped or broken chain requires reattachment using a chain tool or quick link for smooth gear movement.

Figure 5.3: Query result

5.5.1 Final comparison

Here's how the table changed.

- Space used by `invoiceline`
 - Before normalization:
 - * The concept text averaged about 95 bytes per row.
 - * With roughly 24,520 rows, that single column was ~2.3 MB on its own ($24,520 \times 95$ bytes).
 - * Once you include other columns and storage overhead, the table would land around 3–4 MB.
 - After normalization:
 - * `invoiceline` now shows about 1,448 kB (~1.4 MB).
 - * `concept_id` is an int (4 bytes), so every row swapped ~95 bytes of text for 4 bytes.
- Net effect: a reduction over 50%.

5.5.1.1 Trade-offs

- Pros
 - Much smaller footprint, which saves disk and helps the table fit in memory caches.
 - Single place to maintain each concept
 - Quicker scans on `invoiceline` (sums, date filters) because there are fewer pages to read.
- Cons
 - Queries are less direct: to see the concept text you now join with `invoice_concept`.
 - Reconstructing the full view costs a join, which is heavier than reading a single wide row.