Practice 4: Asking Questions of the Cube

SQL. You need the database you built in Practice 3.

Where we are

You have a star schema: one fact table and four dimensions. Today you ask it questions. There is nothing new to build — the four operations below are ordinary SQL on the tables you already have.

Operation	What it means	In SQL
Slice	Fix one dimension, look at the rest	`WHERE y.year = 2024`
Dice	Fix several dimensions at once	`WHERE ... AND ... AND ...`
Roll-up	Go from detail to summary	`GROUP BY e.region`
Drill-down	Go from summary back to detail	`GROUP BY e.name`

Roll-up and drill-down are the same thing in opposite directions. You are choosing how coarse the `GROUP BY` is.

Part 1 — Slice and dice

Task 1.1 — Slice: one year

Fix the year, leave the countries free:

```
SELECT e.name, f.value / 1e9 AS gdp_bn
FROM fact_observation f
JOIN dim_entity    e ON e.entity_key    = f.entity_key
JOIN dim_indicator i ON i.indicator_key = f.indicator_key
JOIN dim_year      y ON y.year_key      = f.year_key
WHERE i.indicator_cd = 'GDP_CUR_USD'
  AND y.year = 2024
  AND e.entity_type = 'country'
ORDER BY f.value DESC
LIMIT 10;
```

Now change the year to 2015 and run it again. That is a slice: same question, different position on one axis.

Task 1.2 — Dice: several conditions at once

Unemployment, in high-income European countries, 2020 to 2024:

```
SELECT e.name, y.year, f.value AS unemployment
FROM fact_observation f
JOIN dim_entity    e ON e.entity_key    = f.entity_key
JOIN dim_indicator i ON i.indicator_key = f.indicator_key
JOIN dim_year      y ON y.year_key      = f.year_key
WHERE i.indicator_cd = 'UNEMPLOYMENT'
  AND e.region       = 'Europe & Central Asia'
  AND e.income_level = 'High income'
  AND y.year BETWEEN 2020 AND 2024
  AND e.entity_type  = 'country'
ORDER BY e.name, y.year;
```

A dice is just a slice on more than one dimension. There is no special syntax — it is the same `WHERE` clause with more conditions.

Task 1.3 — Your own dice

Write a query for inflation in upper-middle-income countries outside Europe, for 2022. How many rows come back?

Part 2 — Roll-up and drill-down

This is where the dimension columns earn their place. `dim_entity` stores each country's region and income level, so you can group by them without touching the fact table.

Task 2.1 — Roll up: country to region

Average unemployment per region in 2024:

```
SELECT e.region,
       COUNT(*)              AS countries,
       ROUND(AVG(f.value),1) AS avg_unemployment
FROM fact_observation f
JOIN dim_entity    e ON e.entity_key    = f.entity_key
JOIN dim_indicator i ON i.indicator_key = f.indicator_key
JOIN dim_year      y ON y.year_key      = f.year_key
WHERE i.indicator_cd = 'UNEMPLOYMENT'
  AND y.year = 2024
  AND e.entity_type = 'country'
  AND e.region IS NOT NULL
GROUP BY e.region
ORDER BY avg_unemployment DESC;
```

Seven rows. You have gone from 180-odd countries to seven numbers.

Task 2.2 — Drill down into one of them

South Asia averaged about 5.0%. Open it up:

```
SELECT e.name, ROUND(f.value,1) AS unemployment
FROM fact_observation f
JOIN dim_entity    e ON e.entity_key    = f.entity_key
JOIN dim_indicator i ON i.indicator_key = f.indicator_key
JOIN dim_year      y ON y.year_key      = f.year_key
WHERE i.indicator_cd = 'UNEMPLOYMENT'
  AND y.year = 2024
  AND e.region = 'South Asia'
  AND e.entity_type = 'country'
ORDER BY f.value DESC;
```

The regional average was 5.0%, but Nepal is 10.5% and Bhutan 3.3%. **An average hides its spread.** That is the whole reason drill-down exists: the summary tells you where to look, not what is true.

Task 2.3 — Roll up a different way

Change `GROUP BY e.region` to `GROUP BY e.income_level` and run it again. Four rows instead of seven. Same facts, different hierarchy.

Task 2.4 — Two levels at once

Group by region **and** income level together. How many rows? Why fewer than 7 × 4?

Part 3 — Two things that will catch you out

Averaging an average

`AVG(f.value)` over countries gives every country equal weight, so Nepal counts the same as India. That is a fair answer to “what is the typical country like?” but the wrong answer to “what is unemployment in South Asia?” — for that you would need to weight by labour force.

Neither is wrong. Know which question you asked.

Rows with nothing to group on

25 entities have facts but no region — the ones you met in Practice 2 (Taiwan, Kosovo, West Bank and Gaza). Without `AND e.region IS NOT NULL` they collect into a single row labelled empty. Drop that condition and see.

A roll-up silently loses anything whose grouping column is empty. Always check your group totals still add up to the whole.

Part 4 — Put it together

Task 4.1 — One question, four ways

Pick an indicator. Write four queries on it: a slice, a dice, a roll-up and a drill-down. Label each one. That is the whole vocabulary of cube analysis, and you now have all of it.

Next class

Visualisation. The same four queries, drawn as charts, on one page. A dashboard is these operations with buttons on top.

Warehouse and earlier sheets: dwhdemo.vedara.net