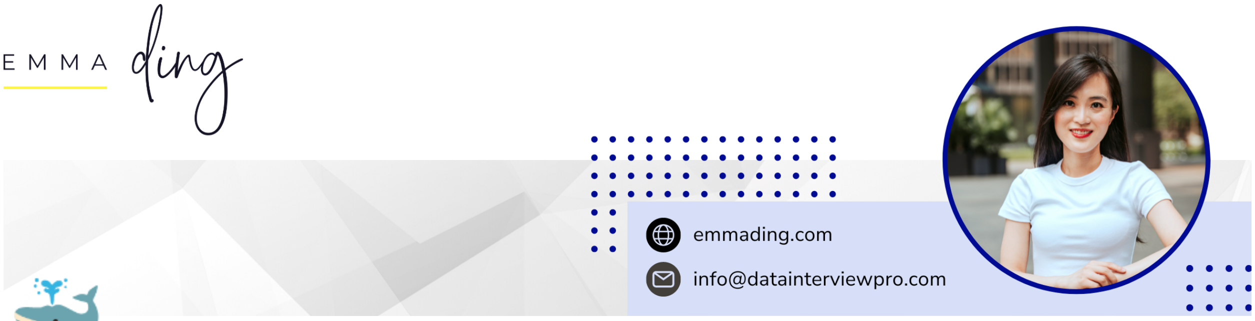
EMMA ding



emmading.com

info@datainterviewpro.com

Top N

Lesson Content

What is Top N Problem?
Top N Records
Top N Per Category
Top N Per Category With Ties

▼ What is Top N Problem?

📌 **Sample Questions**

- What are the Top 5 highest-rated movies?
- What are the Top 3 highest paid employees per department?
- What are the Top 3 highest paid employees per department when there're ties?

▼ Top N Records

- Query the 5th largest value in the table t.
- Assume the values are unique, and there are more than 5 values in the table.

Table t:

value
10
3
…
50

- Although we are not returning all top 5 values, we still consider this as the Top N problem.
- The only difference is that we exclude the Top N - 1 from the result.

▼ `LIMIT` and `OFFSET`

MySQL

```
SELECT value
FROM t
ORDER BY value DESC
LIMIT 1
OFFSET 4;
```

- Select values from **table t** and sort values in descending order.
- Sort the numbers, use `OFFSET`, and `LIMIT` to return the 5th row.

MS SQL Server

- In MS SQL server, the syntax is a little different.

```
SELECT value
FROM t
ORDER BY value DESC
OFFSET 4 ROWS
FETCH NEXT 1 ROWS ONLY;
```

- There is no `LIMIT` keyword, Use the `FETCH` keyword to specify how many rows to return.

▼ Window Functions

```
SELECT value
FROM (
    SELECT value,
    ROW_NUMBER() OVER(ORDER BY value DESC) AS row
    FROM t
) AS rk_table
WHERE row = 5;
```

> 💡 When using window functions, we cannot apply filters on the result generated by the window function directly → create a subquery to filter results.

value	ROW_NUMBER	**DENSE_RANK**	**RANK**
5	1	1	1
4.9	2	2	2
4.9	3	2	2
4.8	4	3	4

- The rank of a row is determined by one plus the number of ranks that come before it.

Top N Per Category

🌿 Cannot use `LIMIT` and `OFFSET`, the window function is a better choice.

Use any of `ROW_NUMBER()`, `RANK()`, and `DENSE_RANK()`.

Example: Query the 5 highest-rated restaurants in each city.

- highest-rated refers to the highest **average** rating.
- If two restaurants have the same average ratings, return either restaurant.

Table rating:

I.D.	Name	City	Rating
10010	Kim's Kitchen	New York	4
10011	Super Dragon	San Francisco	3
…	…	…	…
12010	Tom's Seafood	Tokyo	2

💡 **Idea:**
1. Compute average ratings for all the restaurants.
2. Sort ratings.
3. Select the top 5.

1. Compute average ratings for all the restaurants.

```
SELECT
    name,
    city,
    AVG(rating * 1.0) AS ave_rating
FROM rating
GROUP BY name, city;
```

- Since the ratings are integers, multiply by 1.0 to avoid integer division.
- Put this query in a `WITH CTE` and name it avg_ratings.

2. Sort ratings.

```
WITH avg_ratings AS (
  SELECT
      name, city,
      AVG(rating * 1.0) AS avg_rating
    FROM rating
    GROUP BY name, city
)
```

```
SELECT
    name, city, avg_rating,
    ROW_NUMBER() OVER(PARTITION BY city ORDER BY avg_rating DESC) as row
FROM avg_ratings;
```

- Since we need only 5 restaurants per city, and the ties can be broken arbitrarily.
- Put this query in another `WITH CTE` and name it **rating_rank**.

3. Select the top 5.

```
WITH avg_ratings AS (
  SELECT
    name, city,
    AVG(rating * 1.0) AS avg_rating
  FROM rating
  GROUP BY name, city
),
rating_rank AS (
  SELECT
      name, city, rating,
      ROW_NUMBER() OVER(PARTITION BY city ORDER BY avg_rating DESC) as row
  FROM avg_ratings
)

SELECT
    name, city, rating
FROM rating_rank
WHERE row <= 5;
```

- Filter the row as less or equal to 5 → select only 5 top-rated restaurants per city.

Top N Per Category With Ties

> ❓ What if there are **ties** in the ranks, and we want to get all the restaurants with the top 5 ratings per city? How do we modify the query?

- If the restaurants have the same average ratings, return all restaurants with the same ratings.
  - Number of restaurants per city ≥ 5.
- Change the ranking function from `ROW_NUMBER` to `DENSE_RANK` .

value	ROW_NUMBER	DENSE_RANK	**RANK**
5	1	1	1
4.9	2	2	2
4.9	3	2	2
4.8	4	3	4

```
WITH avg_ratings AS (
  SELECT
```

```
      name, city,
      AVG(rating * 1.0) AS avg_rating
  FROM rating
  GROUP BY name, city
),
rating_rank AS (
  SELECT
      name, city, avg_rating,
      DENSE_RANK() OVER(PARTITION BY city ORDER BY avg_rating DESC) as rk
  FROM avg_ratings
)

SELECT
    name, city, avg_rating
FROM rating_rank
WHERE rk <= 5;
```

🔥 **During interviews:**

• Clarify the logic - whether to output top N records, or all records (≥ N) that match the top N scores.