National Australia Bank Data Engineer Interview Guide – Experienced 3+

Technical Round 1 (14th November 2024)

1. Rank vs Dense Rank

Differences:

- **RANK()**: Skips the next rank for duplicates.
- **DENSE_RANK()**: Does not skip ranks for duplicates.
- **ROW_NUMBER()**: Assigns a unique number to each row.

Use Case:

- Use **DENSE_RANK** when consecutive ranking is required, e.g., reporting 1st, 2nd, 3rd places irrespective of ties.
- Use **RANK** when rank gaps are acceptable for ties.

Example Query:

```sql
SELECT
    ROW_NUMBER() OVER (ORDER BY salary DESC) AS RowNumber,
    RANK() OVER (ORDER BY salary DESC) AS RankId,
    DENSE_RANK() OVER (ORDER BY salary DESC) AS DenseRank
FROM Employee;
```

2. Indexing: When to Use and Avoid

When to Use Indexes:

- For columns with a wide range of values.
- Columns frequently used in WHERE, JOIN, or ORDER BY clauses.
- On columns with few null values.

When to Avoid Indexes:

- Small tables where scanning is faster.
- Columns not often used as conditions.
- Frequently updated columns (indexes slow down writes).

3. Query to Find Employees Without Managers

Tables:

1. **Employee**: Contains employee details.
2. **Manager**: Links employees to their managers.

```sql
SELECT emp.emp_id, emp.emp_name, emp.emp_salary
FROM employee emp
WHERE emp.emp_id NOT IN (SELECT emp_id FROM manager);
```

4. Materialized View

- A **Materialized View** stores query results physically, unlike regular views which compute results on the fly.

- Use cases:

  Frequently queried data.

  Data aggregation for reports.

5. Stored Procedure Optimization

- Analyze execution plans.

- Use **proper indexing**.

- Avoid **cursor-based loops**—use set-based operations.

- Minimize **network round trips** by combining queries.

6. Normalization vs Denormalization

- **Normalization**: Organizing data to reduce redundancy.

  o Use: OLTP systems (transactional).

- **Denormalization**: Combining tables for faster queries.

  o Use: OLAP systems (analytics).

7. Stuff Function for XML Usages

Used to concatenate column values into a single delimited string.

Example Query:

```sql
SELECT
    STUFF((SELECT ',' + emp_name
           FROM Employee
           FOR XML PATH('')), 1, 1, '') AS EmployeeList;
```

8. Data Warehouse vs Data Lake

- **Data Warehouse**: Structured, schema-based storage for analytics.

- **Data Lake**: Stores raw, unstructured, or semi-structured data for diverse use cases.

1. Garbage Collector in Python

- **Garbage Collection**: Automatically frees memory by removing unused objects.

- **Real-Time Example**: Clearing temporary variables during batch data processing in PySpark to manage memory.

2. Multiprocessing in Python

- **Definition**: Enables parallel execution across multiple CPUs.

- **Example**:

```python
from multiprocessing import Pool
def process_data(data):
    return data**2

with Pool(4) as p:
    result = p.map(process_data, [1, 2, 3, 4])
print(result)  # [1, 4, 9, 16]
```

Use Case: Parallel data processing for large datasets.

3. Docker

- **Purpose**: Containerize applications with all dependencies.

- **Handling Requirements**:
    - Dependencies defined in Dockerfile or requirements.txt.
    - Layered builds optimize storage and caching.

4. Databricks

- **Platform**: Unified analytics platform for big data and AI.

- **Use Cases**:
    - Collaborative notebooks.
    - Scalable Spark jobs.
    - Integrated ML workflows.

5. Libraries for Data Wrangling

- **Pandas**: Data manipulation.

- **NumPy**: Numerical operations.

- **Dask**: Distributed dataframes.

- **PySpark**: Big data processing.

6. GeoPandas

- **Definition**: Library for geospatial data analysis in Python.
- **Features**:
    - Works with shapefiles, GeoJSON.
    - Spatial joins and buffer operations.
- **Example**:

```python
import geopandas as gpd
world = gpd.read_file(gpd.datasets.get_path('naturalearth_lowres'))
africa = world[world['continent'] == 'Africa']
africa.plot()
```

Glassdoor National Australia Bank Review –

https://www.glassdoor.co.in/Reviews/NAB-National-Australia-Bank-Reviews-E4262.htm

National Australia Bank Careers –

https://www.nab.com.au/about-us/careers

Subscribe to my YouTube Channel for Free Data Engineering Content –

https://www.youtube.com/@shubhamwadekar27

Connect with me here –

https://bento.me/shubhamwadekar

Checkout more Interview Preparation Material on –

https://topmate.io/shubham_wadekar