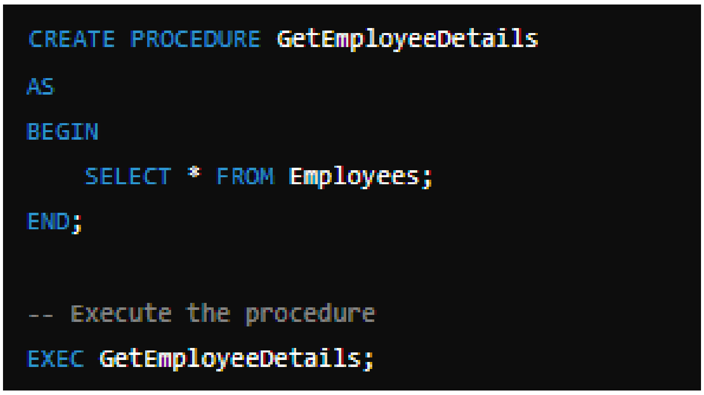
Comcast Data Engineer Interview Guide – Experienced 3+

1. Stored Procedure Syntax and Execution



```
CREATE PROCEDURE GetEmployeeDetails
AS
BEGIN
    SELECT * FROM Employees;
END;

-- Execute the procedure
EXEC GetEmployeeDetails;
```

Explanation:

- **Stored Procedures** are reusable SQL code blocks.
- EXEC or EXECUTE keyword is used to run the procedure.

2. Indexing – Types and Benefits

- **Types**:

  Clustered Index: Sorts and stores data rows in a table based on key columns. One per table.

  Non-clustered Index: Maintains a separate structure for index entries. Multiple per table.

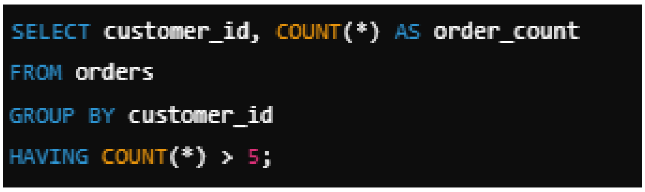
  Unique Index: Ensures all values in the indexed column are unique.

  Full-text Index: Optimized for text searches.

Benefits:

- Speeds up data retrieval.
- Reduces I/O operations by using a smaller subset of data.

3. List Customers with More Than 5 Orders

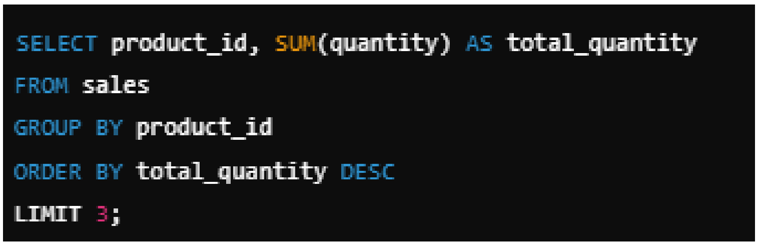


```
SELECT customer_id, COUNT(*) AS order_count
FROM orders
GROUP BY customer_id
HAVING COUNT(*) > 5;
```

Explanation:

- GROUP BY groups rows, and HAVING filters based on aggregation results.

4. Find Top 3 Products Sold Based on Total Quantity

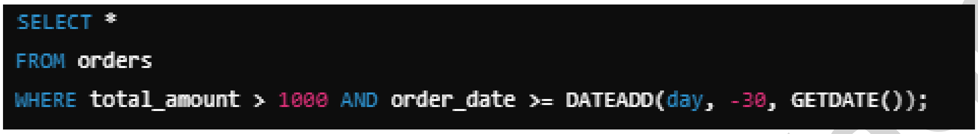


```sql
SELECT product_id, SUM(quantity) AS total_quantity
FROM sales
GROUP BY product_id
ORDER BY total_quantity DESC
LIMIT 3;
```

Explanation:

- LIMIT restricts the number of results.

5. Find Orders Exceeding $1,000 in the Last 30 Days



```sql
SELECT *
FROM orders
WHERE total_amount > 1000 AND order_date >= DATEADD(day, -30, GETDATE());
```

Explanation:

- DATEADD and GETDATE() calculate the date range dynamically.

6. Azure Fabric in Cloud Architecture

Azure Service Fabric is a platform for building distributed systems with microservices.
Uses:

- Provides high availability and scalability for services.
- Manages stateful and stateless services.
- Used in IoT applications and cloud-scale solutions.

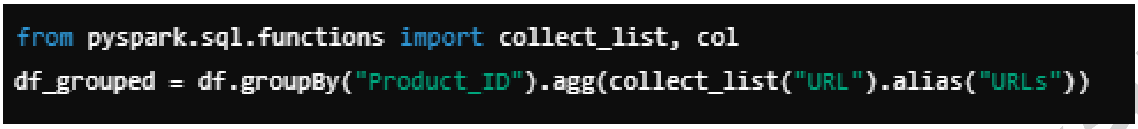
7. Azure Functions vs. Logic Apps

Azure Functions	Logic Apps
Serverless compute service for custom code execution.	Workflow automation using pre-built connectors.
Triggers based on events (HTTP, queue).	Integrates multiple systems and processes.

8. Transformation vs. Action in PySpark

- **Transformation**: Defines a **new RDD** without immediate execution (lazy). Example: map().
- **Action**: Triggers computation and returns a result. Example: collect().

9. Row-Level Records to Column Records



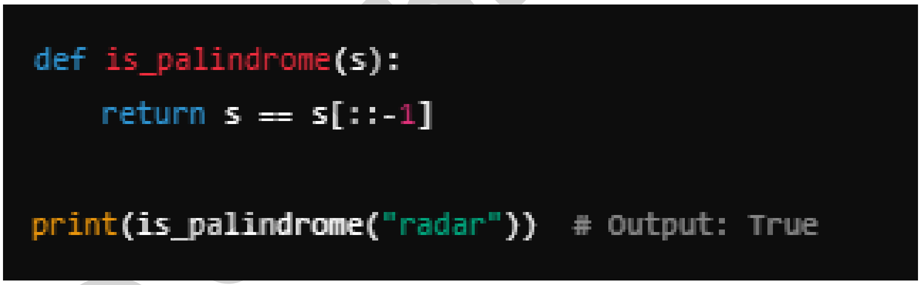
```python
from pyspark.sql.functions import collect_list, col
df_grouped = df.groupBy("Product_ID").agg(collect_list("URL").alias("URLs"))
```

10. Deployment Architecture

Explain a typical big data deployment:

- Data ingestion: Using tools like Kafka or Azure Data Factory.
- Processing: Databricks/Spark for transformation.
- Storage: Data lakes (ADLS) or warehouses (Synapse).
- Visualization: Power BI.

11. Palindrome Check in Python



```python
def is_palindrome(s):
    return s == s[::-1]

print(is_palindrome("radar"))  # Output: True
```

12. Database vs. Data Warehouse vs. Data Lakehouse

Database	**Data Warehouse**	**Data Lakehouse**
OLTP, structured data	OLAP, analytics-ready data	Combines warehouse and lake features
Normalized schema	Star/snowflake schema	Unified storage for both

13. ER Modeling vs. Dimensional Modeling

- ER Modeling: Focuses on entities and relationships. Used for OLTP systems.
- Dimensional Modeling: Uses facts and dimensions for analytics in data warehouses.

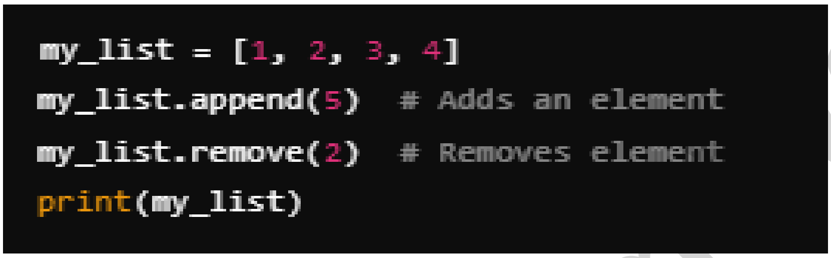
14. Data Warehouse for Grocery Store

- Star Schema:

  Fact Table: Sales (product_id, customer_id, amount, date).

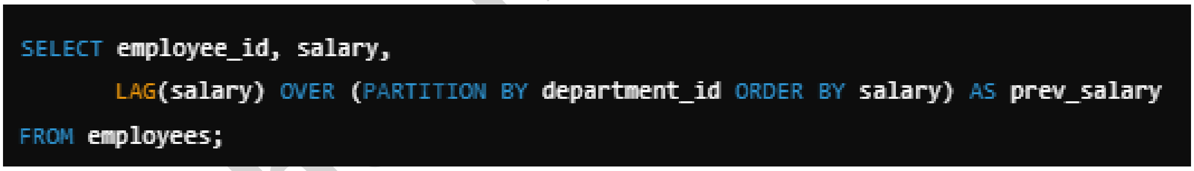
  Dimension Tables: Product, Customer, Date.

15. Python List Operations



```python
my_list = [1, 2, 3, 4]
my_list.append(5)  # Adds an element
my_list.remove(2)  # Removes element
print(my_list)
```

16. SQL Query with LAG Function



```sql
SELECT employee_id, salary,
       LAG(salary) OVER (PARTITION BY department_id ORDER BY salary) AS prev_salary
FROM employees;
```

17. Agile in Project Management

Agile: An iterative development methodology emphasizing collaboration, flexibility, and customer feedback. Used for continuous delivery of software.

18. Daily Tasks of a Data Engineer

- Building pipelines for data ingestion.
- Data cleaning and transformation.
- Optimizing performance of queries.

19. Databricks vs. PySpark

Databricks	**PySpark**
Managed service for Spark	Open-source Spark framework
Integrates with Azure	Requires setup and maintenance

20. Unix Scripting in Data Engineering

Used for automation of ETL processes, running batch jobs, and managing file systems.

Glassdoor Comcast Review –

https://www.glassdoor.co.in/Interview/Comcast-Data-Engineer-Interview-Questions-EI_IE1280.0,7_KO8,21.htm?filter.jobTitleFTS=Data+Engineer

Comcast Careers –

https://jobs.comcast.com/

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