Intel Data Engineering Interview Q&A - 2025

1. What is the difference between ETL and ELT?

ETL (Extract, Transform, Load): Transforms data before loading into the data warehouse.

ELT (Extract, Load, Transform): Loads data first, then transforms using data engines.

Use ETL when source systems are limited, ELT with cloud-based warehouses.

2. Python program to read CSV and store in PostgreSQL:

```
import pandas as pd
import psycopg2

df = pd.read_csv('sales_data.csv')
conn = psycopg2.connect(database='intel_db', user='user', password='pass', host='localhost')
cur = conn.cursor()

for i, row in df.iterrows():
    cur.execute("INSERT INTO sales (id, product, price) VALUES (%s, %s, %s)", tuple(row))

conn.commit()
cur.close()
conn.close()
```

3. What is a Slowly Changing Dimension (SCD)? Types?

SCD handles changes in dimension data over time:

- Type 1: Overwrites old data

- Type 2: Keeps history with new rows

- Type 3: Adds new columns for old/new values

4. SQL: Get top 3 highest salaries per department

```
SELECT *
FROM (
  SELECT name, department, salary,
         RANK() OVER (PARTITION BY department ORDER BY salary DESC) as rnk
  FROM employees
) ranked
WHERE rnk <= 3;
```

5. Kafka vs RabbitMQ

Kafka: Distributed log, great for high-throughput streaming.

RabbitMQ: Message broker, good for transactional real-time messaging.

6. Spark job optimization techniques

- Use persist/cache
- Prefer DataFrame API
- Filter early
- Use broadcast joins wisely

7. PySpark: Convert JSON to Parquet

```
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName('JSONToParquet').getOrCreate()
df = spark.read.json('logs.json')
```

```
df.write.parquet('output/logs_parquet')
```

8. Data Lake vs Data Warehouse

Data Lake: Raw, unstructured/semi-structured data (e.g., S3).

Data Warehouse: Structured, optimized for querying (e.g., Snowflake).

9. Handling duplicate records in large datasets

SQL:

```
DELETE FROM employees
WHERE id NOT IN (
  SELECT MIN(id)
  FROM employees
  GROUP BY email
);
```

10. Describe a data pipeline you built

Example:

- Ingestion: Kafka
- Processing: Spark
- Storage: S3/Redshift
- Orchestration: Airflow with Grafana