ARVIND KUMAR

Infosys

Round - 1

Experience: 0-5 Year

1. OLTP vs OLAP

What they asked me

Explain the difference between OLTP and OLAP. Give a real-time example.

What I said

OLTP handles real-time transactional processing, while OLAP handles analytical workloads.

Scenario:
In an e-commerce project, the order database in MySQL was OLTP.
For analytics, daily order data was loaded into Snowflake (OLAP) to generate dashboards.

Tips

- Always add a business scenario.
- Mention one OLTP tool and one OLAP tool.

2. What is ETL? Explain with a real use case

What they asked me

What is ETL? Explain with an example.

What I said

ETL stands for Extract, Transform, Load.

Scenario:
I extracted customer data from Excel, cleaned null values and duplicates using PySpark, and loaded the final dataset into Snowflake for reporting.

Tips

- Keep ETL steps very clear.
- Mention tools like ADF, Glue, Informatica.

3. Types of SQL Joins

What they asked me

Explain all join types with examples.

What I said

I explained inner, left, right, full, and cross joins.

Example:
Left join used to fetch all customers even if they have no orders.

```
SELECT c.customer_id, o.order_id
FROM customer c
LEFT JOIN orders o
ON c.customer_id = o.customer_id;
```

Tips

Focus on left join and inner join. They are the most frequently asked.

4. Partitioning vs Bucketing

What they asked me

How is partitioning different from bucketing in Hive or Spark?

What I said

- Partitioning divides data based on column values, improving filter queries.
- Bucketing distributes data into uniform buckets, improving joins.

Scenario:
A 100M sales table partitioned by year and bucketed by customer_id.

Tips

Use large dataset scenarios; Infosys interviewers like practical optimization answers.

5. Window Functions

What they asked me

What are window functions? Write SQL for running total.

What I said

Window functions allow calculations across rows without grouping.

```
SELECT order_id, amount,
SUM(amount) OVER (ORDER BY order_date) AS running_total
FROM orders;
```

Tips

Mention two examples: running total and ranking.

6. Kafka Basics

What they asked me

What is Kafka? Why is it used?

What I said

Kafka is a distributed streaming system used for building real-time data pipelines.
It uses producers, consumers, topics, and partitions.

Scenario:
Kafka ingested app event logs which Spark Streaming processed and loaded into Snowflake.

Tips

Keep the explanation simple and avoid deep architecture unless asked.

7. RDD vs DataFrame vs Dataset in Spark

What they asked me

What is the difference between RDD, DataFrame, and Dataset?

What I said

- RDD: Low-level, unstructured, slower.
- DataFrame: Column-based, optimized by Catalyst optimizer.
- Dataset: Type-safe, structured, compile-time checks.

Tips

Always mention Catalyst Optimizer and Tungsten engine.

8. PySpark Code to Remove Duplicates

What they asked me

Write PySpark code to remove duplicates.

What I said

```
df2 = df.dropDuplicates(["customer_id"])
```

Tips

Add one-line explanation: dropDuplicates() is a transform, not an action.

9. Explain SCD Type 1 and Type 2

What they asked me

What is Slowly Changing Dimension? Explain Type 1 and Type 2.

What I said

- Type 1 overwrites old data.
- Type 2 maintains history using start_date, end_date, and active flag.

Scenario:
Customer address change stored using Type 2 to maintain history.

Tips

Use a customer dimension as your example always.

10. Data Lake vs Data Warehouse

What they asked me

What is the difference between a Data Lake and a Data Warehouse?

What I said

- Data Lake stores raw data with schema-on-read.
- Data Warehouse stores structured, cleaned data with schema-on-write.

Tips

Give a simple scenario:
Application logs to Data Lake → curated business tables to Data Warehouse.