DE Evaluation Questions

1. How many years of experience do you have working in Data Engineering?
2. Who are your stakeholders?
3. How often do you use window functions in your day-to-day use of SQL?
4. What are Common Table Expressions (CTEs) in SQL, and how do they differ from subqueries? Provide examples of when to use CTEs.
5. What techniques you have used for dimensional data modelling, if ever worked on data modelling?
6. What pillars you evaluate while designing ETL pipelines?
7. What open table formats you are familiar with? How do you decide which one to opt for scalable performance?
8. How can you describe cloud native data platform? What components should be part of it?
9. Describe a challenging data engineering problem you've encountered and how you solved it.
10. Describe your approach to identify and resolve performance bottlenecks in data systems
11. What data quality checks you should have in place by default for any dataset?
12. How do you audit health for your data platform?
13. Explain the steps that you will take to identify performance challenges in a SQL Query. With a specific example
14. When do you consider Batch vs Streaming Ingestion?

**\*\*PySpark Coding Exercise:\*\***

Suppose you have a dataset containing user activity logs, with the following schema:

```

+---------------+----------+-----+-----+-----+

| user\_id | event\_type|value|timestamp|device|

+---------------+----------+-----+-----+-----+

| 1001| login| 1|123456789| iOS|

| 1001| logout| 1|123457789| iOS|

| 1002| login| 1|123456799| iOS|

| 1002| purchase| 100|123457899|Android|

| 1003| login| 1|123456909|Android|

| 1003| login| 1|123457009|Android|

| 1003| logout| 1|123457109|Android|

+---------------+----------+-----+-----+-----+

```

Tasks:

1. Write a PySpark function to calculate the total login duration for each user.

2. Implement a PySpark function to find the top 3 users with the highest total purchase value.

3. Describe how you would handle if the data is not refreshed in this dataset using PySpark.

**\*\*SQL Coding Exercise:\*\***

Suppose you have the following table schema for an e-commerce database:

Table MERCHANTS   
  
SAMPLE DATA  
+-------------+-------------+-------------+--------------+-------------------+  
| merchant\_id | marketplace | launch\_date | product\_desc | country\_of\_origin |  
+-------------+-------------+-------------+--------------+-------------------+  
| 12 | DE | 1/1/2016 | BOOKS | DE |  
| 12 | FR | 6/1/2016 | BOOKS | DE |  
| 34 | FR | 6/1/2016 | ELECTRONICS | |  
| 34 | UK | 3/1/2016 | WIRELESS | FR |  
| 56 | DE | 4/1/2015 | BOOKS | UK |  
| 56 | IT | 1/1/2015 | SOFTLINES | |  
| 56 | UK | 2/1/2015 | SOFTLINES | UK |  
| 78 | ES | 12/1/2019 | ELECTRONICS | ES |  
+-------------+-------------+-------------+--------------+-------------------+  
  
Table ORDERS   
  
SAMPLE DATA  
+-------------+-------------+----------+--------+-------+----------------+  
| merchant\_id | marketplace | order\_id | sales | units | order\_date |  
+-------------+-------------+----------+--------+-------+----------------+  
| 12 | DE | BTHDOF | 100.50 | 6 | 1/1/2017 0:45 |  
| 12 | DE | BAFTHF | 120.32 | 8 | 1/1/2017 7:12 |  
| 43 | FR | GTHDAE | 18.5 | 1 | 1/1/2017 14:37 |  
| 12 | DE | RTZHGD | 90.3 | 3 | 1/2/2017 10:35 |  
| 56 | IT | ASDQER | 180.2 | 9 | 1/2/2017 22:01 |  
| 56 | UK | YCCCOD | 4000 | 400 | 1/3/2017 17:01 |  
+-------------+-------------+----------+--------+-------+----------------+

Q. Write a query to identify Merchant that got churned out for Feb 2017?

Also, how would you write in dynamic way to identify churn for any given month and year?

Definition of Churn – Merchant having at least 1 order for 2016 and 0 order for 2017

Q. Write a query to report WOW (Week over Week) sales for each product category for trailing 6 weeks?