Q2 - Data Reliability & Debugging

What Do I Know?

- On **July 9, 2024**, a spike was observed in the **purchaser user rate** metric.
- Both the Product and Marketing teams confirmed that they did not take any actions, and an investigation was requested to determine whether this spike resulted from a data issue.
- The dataset only covers data for the period of **July 7 - July 14, 2024**.

How Can I Test This?

- Since the dataset is not very large, I can aggregate daily values and create a table that includes the following metrics:
  - **Daily Active Users**
  - **Daily Unique Purchasers**
  - **Daily Total Purchases**
  - **Daily Average Purchases (Per Purchaser)**
  - **Daily Purchaser Rate**
  - **Daily Purchase Rate**
  - **Daily iOS Purchases**
  - **Daily Android Purchases**
  - **Daily Unique iOS Purchasers**
  - **Daily Unique Android Purchasers**
- By listing these metrics, I can easily detect any anomalies that occurred on **July 9th**.

Based on this approach, I created the following query:

```
WITH DailyMetrics AS (
  SELECT
    DATE(event_date) as date,
    COUNT(DISTINCT user_id) as daily_active_users,
```

```
    COUNT(DISTINCT CASE WHEN event_name = 'purchase' THEN user_id END) as total_unique_purchas
ers,
    COUNTIF(event_name = 'purchase') as total_purchase_count
  FROM `data-sciene-for-business-imp.app_analytics.data_reliability`
  GROUP BY DATE(event_date)
),
PurchaseByOS AS (
  SELECT
    DATE(event_date) as date,
    COUNT(DISTINCT CASE WHEN operating_system = 'ios' THEN user_id END) as ios_unique_purchase
rs,
    COUNT(DISTINCT CASE WHEN operating_system = 'android' THEN user_id END) as android_unique_
purchasers,
    COUNTIF(operating_system = 'ios') as ios_purchase_count,
    COUNTIF(operating_system = 'android') as android_purchase_count
  FROM `data-sciene-for-business-imp.app_analytics.data_reliability`
  WHERE event_name = 'purchase'
  GROUP BY DATE(event_date)
)

SELECT
  dm.date,
  dm.daily_active_users,
  dm.total_unique_purchasers,
  dm.total_purchase_count,
  ROUND(SAFE_DIVIDE(dm.total_purchase_count, dm.total_unique_purchasers), 2) as avg_purchase_p
er_purchaser,
  ROUND(SAFE_DIVIDE(dm.total_unique_purchasers, dm.daily_active_users) * 100, 2) as purchaser_
rate,
  ROUND(SAFE_DIVIDE(dm.total_purchase_count, dm.daily_active_users) * 100, 2) as purchase_rat
e,
  COALESCE(pos.ios_unique_purchasers, 0) as ios_unique_purchasers,
  COALESCE(pos.ios_purchase_count, 0) as ios_purchase_count,
  COALESCE(pos.android_unique_purchasers, 0) as android_unique_purchasers,
  COALESCE(pos.android_purchase_count, 0) as android_purchase_count
FROM DailyMetrics dm
LEFT JOIN PurchaseByOS pos
  ON dm.date = pos.date
ORDER BY dm.date;
```

As a result of the query above, I arrive at the following table:

date	daily_active_users	total_unique_purchasers	total_purchase_count	avg_purchase_per_purchaser	purchaser_rate	
2024-07-07	28369	484	484	1	1.71	
2024-07-08	31832	563	563	1	1.77	
2024-07-09	42369	796	1075	1.35	1.88	
2024-07-10	50637	956	956	1	1.89	
2024-07-11	51892	983	983	1	1.89	
2024-07-12	50627	872	872	1	1.72	
2024-07-13	38578	628	628	1	1.63	
2024-07-14	32427	558	558	1	1.72	

Findings:

1. **On regular days:**
   - Purchaser rate is around **1.7%-1.9%**.
   - The number of purchases per user is exactly **1.0**.
   - On both iOS and Android, the numbers for unique purchasers and purchase counts are **equal**.
2. **On July 9th:**
   - Purchaser rate increased to **2.54%**.
   - The number of purchases per user rose to **1.35**.
   - On **Android**, everything appears normal (517 unique purchasers, 517 purchases).
   - On **iOS**, there are duplicate records (279 unique purchasers, 558 purchases).

Response:

a) **Yes, there is definitely a data issue.**

On the iOS platform, every purchase event was logged **twice** on July 9th. This has led to:

- An artificial increase in the purchase rate from **1.88% to 2.54%**.
- The issue is **exclusive to the iOS platform**.
- It occurred **only on July 9th**.
- There are no such anomalies on other dates.

b) **Percentage of affected users:**

- All purchase records of the **279 unique users** on iOS were duplicated.
- This accounts for approximately **0.66%** of the total **42,369 active users**.
- From a metrics perspective, the duplicate data inflated the purchase rate by **35%**.

Potential Cause:

The issue seems to stem from a **technical problem in the iOS event tracking system** on July 9th, which caused purchase events to be logged as duplicates.

To address this:

1. **Clean up duplicate records** on the iOS platform for July 9th.
2. **Update the event tracking system** to include duplicate detection and prevention.
3. Implement a **monitoring system** to avoid similar issues in the future.

```
SELECT
    user_id,
    event_date,
    event_time,
    operating_system
  FROM `data-sciene-for-business-imp.app_analytics.data_reliability`
  WHERE DATE(event_date) = '2024-07-09'
    AND event_name = 'purchase'
    AND operating_system = 'ios'
ORDER BY  event_time;
```

The query above is evidence of the data duplication issue that occurred on **July 9th**.