Project Title: Operation Analytics and Metric Investigation

Project Overview:

Operation Analytics is a vital aspect of modern business management, providing insights into the end-to-end operations of a company. As a Data Analyst Lead at a company similar to Microsoft, my role involves analyzing various datasets and tables to gain valuable insights that aid in decision-making and process improvement across departments. This analysis not only enhances operational efficiency but also fosters collaboration among cross-functional teams.

The primary objective of this project is to extract meaningful insights from the provided data sets, addressing critical questions from different departments. The key metrics we are focusing on include the number of jobs reviewed, throughput, percentage share of each language, identification of duplicate rows, user engagement, user growth, weekly retention, and email engagement.

Approach:

Understanding the data and its structure is the initial step. This involves comprehending the significance of columns such as job_id, actor_id, and event. I have established a dedicated database named "operation_analytics" and structured tables based on the provided data schema. The subsequent phase involves data analysis using SQL queries to uncover valuable insights for the company.

Execution:

Case Study 1 (Job Data):

A. Number of Jobs Reviewed:

```
SELECT
   COUNT(DISTINCT job_id) / (30 * 24) AS num_jobs_reviewed
FROM
   job_data
WHERE
   ds BETWEEN '2020-11-01' AND '2020-11-30';
```

B. Throughput:

```
SELECT
   ds,
   jobs_reviewed,
   AVG(jobs_reviewed) OVER (ORDER BY ds ROWS BETWEEN 6 PRECEDING AND
CURRENT ROW) AS throughput_7_rolling_avg
FROM
   (
      SELECT
         ds,
         COUNT(DISTINCT job_id) AS jobs_reviewed
      FROM
         job_data
      WHERE
         ds BETWEEN '2020-11-01' AND '2020-11-30'
      GROUP BY
         ds
      ORDER BY
         ds
   ) a;
```

C. Percentage Share of Each Language:

```
SELECT
  language,
  num_jobs,
  (100.0 * num_jobs) / total_jobs AS pct_share_jobs
FROM
  (
    SELECT
      language,
      COUNT(DISTINCT job_id) AS num_jobs
    FROM
      job_data
    GROUP BY
      language
  ) a
CROSS JOIN
  (
    SELECT
      COUNT(DISTINCT job_id) AS total_jobs
    FROM
      job_data
  ) b;
```

D. Duplicate Rows:

```
SELECT *
FROM
  (
    SELECT *,
        ROW_NUMBER() OVER (PARTITION BY job_id) AS rownum
    FROM
      job_data
  ) a
WHERE
  rownum > 1;
```

Case Study 2 (Investigating Metric Spike):

A. User Engagement:

```
SELECT
   EXTRACT(WEEK FROM occurred_at) AS num_week,
   COUNT(DISTINCT user_id) AS no_of_distinct_user
FROM
   tutorial.yammer_events
GROUP BY
   num_week;
```

B. User Growth:

```
SELECT
   year,
   num_week,
   num_active_users,
   SUM(num_active_users) OVER (ORDER BY year, num_week ROWS BETWEEN
UNBOUNDED PRECEDING AND CURRENT ROW) AS cumm_active_users
FROM
   (
      SELECT
         EXTRACT(YEAR FROM a.activated_at) AS year,
         EXTRACT(WEEK FROM a.activated_at) AS num_week,
         COUNT(DISTINCT user_id) AS num_active_users
      FROM
         tutorial.yammer_users a
      WHERE
         state = 'active'
      GROUP BY
         year, num_week
      ORDER BY
         year, num_week
   ) a;
```

C. Weekly Retention:

```
SELECT
  COUNT(user_id),
  SUM(CASE WHEN retention_week = 1 THEN 1 ELSE 0 END) AS
per_week_retention
FROM
  (
    SELECT
      a.user_id,
      a.sign_up_week,
      b.engagement_week,
      b.engagement_week - a.sign_up_week AS retention_week
FROM
      (
        SELECT DISTINCT user_id, EXTRACT(WEEK FROM occurred_at) AS
sign_up_week
        FROM tutorial.yammer_events
        WHERE event_type = 'signup_flow'
            AND event_name = 'complete_signup'
            AND EXTRACT(WEEK FROM occurred_at) = 18
      ) a
    LEFT JOIN
      (
        SELECT DISTINCT user_id, EXTRACT(WEEK FROM occurred_at) AS
engagement_week
        FROM tutorial.yammer_events
        WHERE event_type = 'engagement'
      ) b
    ON
      a.user_id = b.user_id
  )
GROUP BY
  user_id
ORDER BY
  user_id;
```

D. Weekly Engagement:

```
SELECT
  EXTRACT(YEAR FROM occurred_at) AS year_num,
  EXTRACT(WEEK FROM occurred_at) AS week_num,
  device,
  COUNT(DISTINCT user_id) AS no_of_users
FROM
  tutorial.yammer_events
WHERE
  event_type = 'engagement'
GROUP BY
  1, 2, 3
ORDER BY
  1, 2, 3;
```

E. Email Engagement:

```
SELECT
  100.0 * SUM(CASE WHEN email_cat = 'email_opened' THEN 1 ELSE 0 END) /
SUM(CASE WHEN email_cat = 'email_sent' THEN 1 ELSE 0 END) AS
email_opening_rate,
  100.0 * SUM(CASE WHEN email_cat = 'email_clicked' THEN 1 ELSE 0 END) /
SUM(CASE WHEN email_cat = 'email_sent' THEN 1 ELSE 0 END) AS
email_clicking_rate
FROM
  (
    SELECT *,
        CASE
          WHEN action IN ('sent_weekly_digest', 'sent_reengagement_email')
THEN 'email_sent'
          WHEN action IN ('email_open') THEN 'email_opened'
          WHEN action IN ('email_clickthrough') THEN 'email_clicked'
        END AS email_cat
    FROM
      tutorial.yammer_events
  ) a;
```

Insights:

Case Study 1 (Job Data):

- The average number of distinct jobs reviewed per hour per day for November 2020 was 83%.
- A 7-day rolling average for throughput was used as it provides a more comprehensive view of data trends compared to daily metrics.
- The highest percentage share among languages in the last 30 days was observed for the Persian language at 37.5%.
- Duplicate rows were identified based on the 'job_id' column, revealing two duplicate rows in the dataset.

Case Study 2 (Investigating Metric Spike):

- Weekly user engagement showed an initial increase from week 18th to week 31st but later declined, indicating a change in user perception or product quality.
- The cumulative count of active users over time from 2013 to 2014 reached 9,381.
- The highest weekly engagement per device was observed for MacBook and iPhone users.
- Email engagement metrics indicated a healthy email opening rate of around 34% and an email clicking rate of approximately 15%, demonstrating positive user engagement with the email service.

Result:

This project has equipped me with advanced SQL skills, particularly in utilizing Window Functions and complex queries. It has provided valuable insights into real-world industry practices, emphasizing the importance of asking the right questions, data exploration, and collaboration across departments. Investigating metric spikes and providing actionable insights to drive business growth has been a key focus, contributing to my mastery of SQL concepts and data analysis techniques.

SQL Queries Link

Case Study 1: Aman-Patel-Github-Repo
Case Study 2: Aman-Patel-Github-Repo