Trainity - Himanshu Jani - Project 3

SQL - Operation Analytics and Investigating Metric Spike

Case Study 1 (Job Data)

A.Number of jobs reviewed: Amount of jobs reviewed over time. Your task: Calculate the number of jobs reviewed per hour per day for November 2020?

```
SELECT
        CONVERT(date, ds) AS date,
        DATEPART(hour, CAST(ds AS datetime)) AS hour,
       FORMAT (ROUND ( SUM(time_spent) / 3600.0 , 3 ), '0.000') AS
        time_spent_hours,
      COUNT(DISTINCT job_id) AS jobs_reviewed
FROM
         Job_data
        WHERE
 CONVERT(date, ds) >= '2020-11-01' AND CONVERT(date, ds) < '2020-12-01'
  GROUP BY
CONVERT(date, ds), DATEPART(hour, CAST(ds AS datetime))
  ORDER BY
CONVERT(date, ds), DATEPART(hour, CAST(ds AS datetime));
```

B.Throughput: It is the no. of events happening per second.
Your task: Let's say the above metric is called throughput. Calculate 7 day rolling average of throughput? For throughput, do you prefer daily metric or 7-day rolling and why?

```
SELECT
        CONVERT(date, ds) AS date,
         DATEPART(hour, CAST(ds AS datetime)) AS hour,
         CAST(ROUND(SUM(time_spent) / (COUNT(*) * 3600.0), 3) AS
         DECIMAL(18, 3)) AS throughput,
         CAST(ROUND(COUNT(DISTINCT job_id), 3) AS DECIMAL(18, 3)) AS
          Jobs_reviewed
          FROM   job_data
          WHERE
CONVERT(date, ds) >= '2020-11-01' AND CONVERT(date, ds) < '2020-12-01'
GROUP BY   CONVERT(date, ds), DATEPART(hour, CAST(ds AS datetime))
ORDER BY     date, hour;
```

C. Percentage share of each language: Share of each language for different contents.
Your task: Calculate the percentage share of each language in the last 30 days?

```
SELECT language, COUNT(*) AS total_jobs,

FORMAT((COUNT(*) * 100.0) / (

SELECT COUNT(*)

FROM job_data

 WHERE ds >= '2020-11-01' AND ds < '2020-12-01'

 ), '0.000') AS percentage_share

FROM job_data

WHERE ds >= '2020-11-01' AND ds < '2020-12-01'

GROUP BY language

ORDER BY percentage_share DESC;
```

D. Duplicate rows: Rows that have the same value present in them.
Your task: Let's say you see some duplicate rows in the data. How will you display duplicates from the table?

```
SELECT

duplicates.ds, duplicates.job_id, duplicates.actor_id, duplicates.event,
duplicates.language, duplicates.time_spent, duplicates.org, duplicates.row_num

FROM (   SELECT job_data.ds, job_data.job_id, job_data.actor_id,

          job_data.event, job_data.language, job_data.time_spent, job_data.org,

           ROW_NUMBER() OVER (PARTITION BY job_id ORDER BY

           (SELECT 0)) AS row_num

             FROM job_data

           ) AS duplicates

WHERE duplicates.row_num > 1

ORDER BY duplicates.job_id;
```

Case Study 2 (Investigating metric spike)

The structure of the table with the definition of each column that you must work on is present in the project image

A. User Engagement: To measure the activeness of a user. Measuring if the user finds quality in a product/service.

Your task: Calculate the weekly user engagement?

```
SELECT
   DATEPART(ISO_WEEK, DATEADD(wk, DATEDIFF(wk, 0, occurred_at), 0)) AS week_number,
   DATEADD(wk, DATEDIFF(wk, 0, occurred_at), 0) AS week_start_date,
   COUNT(DISTINCT user_id) AS active_users
FROM
   events
GROUP BY
   DATEPART(ISO_WEEK, DATEADD(wk, DATEDIFF(wk, 0, occurred_at), 0)),
   DATEADD(wk, DATEDIFF(wk, 0, occurred_at), 0)
ORDER BY
   week_start_date;
```

B. User Growth: Amount of users growing over time for a product.
Your task: Calculate the user growth for product?

```
SELECT

    DATEPART(YEAR, e.occurred_at) AS year_num,

    DATEPART(WEEK, e.occurred_at) AS week_num,

    COUNT(DISTINCT e.user_id) AS num_users,

    SUM(COUNT(DISTINCT e.user_id)) OVER (ORDER BY DATEPART(YEAR, e.occurred_at),
DATEPART(WEEK, e.occurred_at)) AS cumulative_users

FROM

    events e

GROUP BY

    DATEPART(YEAR, e.occurred_at),

    DATEPART(WEEK, e.occurred_at)

ORDER BY

    year_num,

    week_num;
```

C. Weekly Retention:

Users getting retained weekly after signing-up for a product.
Your task: Calculate the weekly retention of users-sign up cohort?

```
WITH signup_cohort AS (
    SELECT
        user_id,
        DATEPART(YEAR, created_at) AS signup_year,
        DATEPART(WEEK, created_at) AS signup_week
    FROM
        users
),
engagement_counts AS (
    SELECT
        user_id,
        DATEPART(YEAR, occurred_at) AS engagement_year,
        DATEPART(WEEK, occurred_at) AS engagement_week,
        COUNT(*) AS num_engagements
    FROM
        events
    GROUP BY
        user_id,
        DATEPART(YEAR, occurred_at),
        DATEPART(WEEK, occurred_at)
),
retained_users AS (
    SELECT
        s.signup_year,
```

```
        s.signup_week,
        COUNT(DISTINCT e.user_id) AS num_retained_users
    FROM
        signup_cohort s
        INNER JOIN engagement_counts e ON e.user_id = s.user_id
            AND e.engagement_year = s.signup_year
            AND e.engagement_week >= s.signup_week
            AND e.engagement_week < s.signup_week + 2
    GROUP BY
        s.signup_year,
        s.signup_week
)
SELECT
    r.signup_year,
    r.signup_week,
    r.num_retained_users,
    r.num_retained_users / CAST(s.num_signup_users AS FLOAT) * 100 AS retention_rate
FROM
    retained_users r
    INNER JOIN (
        SELECT
            signup_year,
            signup_week,
            COUNT(*) AS num_signup_users
        FROM
            signup_cohort
```

```
        GROUP BY
            signup_year,
            signup_week
    ) s ON s.signup_year = r.signup_year AND s.signup_week = r.signup_week
ORDER BY
    r.signup_year,
    r.signup_week;
```

D .Weekly Engagement:

To measure the activeness of a user. Measuring if the user finds quality in a product/service weekly.

Your task: Calculate the weekly engagement per device?

```
SELECT
    DATEPART(YEAR, occurred_at) AS year_num,
    DATEPART(WEEK, occurred_at) AS week_num,
    device,
    COUNT(DISTINCT user_id) AS num_engaged_users
FROM
    events
WHERE
    event_type = 'engagement'
GROUP BY
    DATEPART(YEAR, occurred_at),
    DATEPART(WEEK, occurred_at),
    device
```

```
ORDER BY
    year_num,
    week_num;
```

E. Email Engagement: Users engaging with the email service.
Your task: Calculate the email engagement metrics?

```
SELECT
    DATEPART(YEAR, occurred_at) AS year_num,
    DATEPART(WEEK, occurred_at) AS week_num,
    action,
    COUNT(DISTINCT user_id) AS num_engaged_users
FROM
    email_events
GROUP BY
    DATEPART(YEAR, occurred_at),
    DATEPART(WEEK, occurred_at),
    action
ORDER BY
    year_num,
    week_num;
```

Project Description

The project aims to analyze and derive insights from these tables by performing various calculations and metrics.

The provided queries in the case study involve tasks such as calculating user engagement, user growth, weekly retention, weekly engagement per device, and email engagement metrics.

By running these queries on the respective tables, we can obtain valuable information about user activity, product usage, user retention, and engagement patterns.

Approach

Through my comprehensive approach, I was able to unlock valuable insights that shed light on user preferences, the growth trajectory of our product, and the effectiveness of our email engagement strategies.

These findings serve as a solid foundation for further analysis and decision-making, enabling us to optimize our product offerings, enhance user experiences, and maximize the impact of our email communications.

Tech-Stack Used

The initial phase of the project presented challenges regarding the import of the provided CSV files into a suitable database management system such as MySQL or MS SQL. Extensive research was conducted to determine the most effective method for importing the CSV files into MS SQL. It was crucial to ensure that the data from the CSV files was accurately transferred and aligned with the appropriate database schema.

After thorough investigation, the optimal solution was found by utilizing the Task option with Flat File import functionality within MS SQL. This method facilitated the seamless importation of the CSV files into MS SQL, enabling the creation of the necessary tables to store the data. By following this approach, the project successfully established a solid foundation for subsequent data analysis.

Insights

Here are the 5 Major insights I was able to uncover from this project.

1. **User Engagement:** By analyzing user engagement metrics, we observed patterns indicating the level of activeness and quality perception among users. This insight helps identify segments of highly engaged users, enabling targeted strategies to enhance overall user experience and drive product adoption.

2. **User Growth:** The analysis of user growth over time revealed the trajectory and pace of user acquisition. This knowledge allows us to assess the effectiveness of our growth strategies and make informed decisions on resource allocation and expansion plans.

3. **Weekly Retention:** The weekly retention analysis provided insights into the percentage of users retained after signing up for our product. By understanding the retention rate, we can refine our onboarding process, optimize user engagement initiatives, and improve long-term user retention.

4. **Weekly Engagement per Device:** Analyzing user engagement on different devices helped identify device preferences and usage patterns. This insight can guide product development decisions, such as optimizing user interfaces and functionalities for specific devices, ultimately enhancing user satisfaction and engagement.

5. **Email Engagement Metrics:** By examining email engagement metrics, such as the number of users engaging with email content, we gained insights into the effectiveness of our email communication strategies. This information can be used to tailor email campaigns, improve open and click-through rates, and optimize user engagement with our email services.

Overall, these insights empower us to make data-driven decisions, optimize our product offerings, and improve user experiences.

Result

Through the completion of this project, Here are Top 5 achievements that have been realized, along with the benefits gained from the process. These achievements include:

Data Analysis: By analyzing the data from the users, events, and email_events tables, we gained valuable insights into user behavior, engagement patterns, and the effectiveness of our product and email services.

Insights and Decision-Making: The project has enabled us to make data-driven decisions based on the insights derived from the analysis. We can now identify user engagement trends, measure user growth, assess retention rates, and evaluate email engagement metrics.

Resource Allocation: With a clear understanding of user growth and engagement, we can allocate our resources effectively. This includes optimizing marketing efforts, refining product features, and enhancing email campaigns to target and engage specific user segments.

User Experience Enhancement: The project has provided us with actionable insights to improve the overall user experience. We can now identify pain points, optimize user interfaces, enhance features, and personalize interactions based on user preferences.

Strategic Planning: Armed with the knowledge gained from the project, we can develop informed strategies for future growth and success. The insights obtained allow us to set realistic goals, identify opportunities for innovation, and plan effective marketing campaigns.

Overall, the project has greatly helped me in understanding our users, improving our product and email services, and making informed decisions to drive growth and enhance user experiences.