

Operation Analytics and Investigating Metric Spike (project-3)

By- Abhishek Adyani

Project description:

Operational Analytics is a crucial process that involves analyzing a company's end-to-end operations. This analysis helps identify areas for improvement within the company. One of the key aspects of Operational Analytics is investigating metric spikes. This involves understanding and explaining sudden changes in key metrics, such as a dip in daily user engagement or a drop in sales.

In this project, with various datasets and tables, the task is to derive insights from the given data to answer questions posed by different departments within the company.

Approach:

As we have different datasets, I have first studied them thoroughly and understand the terminologies that are being used in it, then after checking for any null values and cleaned the data if needed. I solved the different questions using my SQL knowledge to derive different insights for the management team.

Tech-stack used:

I have used MYSQL workbench 8.0.36 community version by connecting to my local server with the given database, I chose this software as it is very user friendly and a great tool to analyze the data

Execution:

Case study 1 : Job Data Analysis

- a) **Jobs Reviewed Over Time** : Calculate the number of jobs reviewed per hour for each day in November 2020.

Query code used:

```
SELECT
    ds AS date,
    ROUND((COUNT(job_id) / SUM(time_spent) * 3600, 2) AS job_review_per_hour
FROM
    job_data
WHERE
    ds BETWEEN '2020-11-01' AND '2020-11-30'
GROUP BY ds
ORDER BY ds;
```

- b) **Throughput Analysis**: Calculate the 7-day rolling average of throughput (number of events per second).

Query code used:

```
SELECT
    ROUND((COUNT(event) / SUM(time_spent)), 2) AS weekly_throughput
FROM
    job_data;
```

and for daily throughput :

```
SELECT
    ds AS date,
    ROUND((COUNT(event) / SUM(time_spent)), 2) AS daily_throughput
FROM
    job_data
GROUP BY ds;
```

- c) **Language Share Analysis:** Calculate the percentage share of each language in the last 30 days.

Query code used:

```
SELECT
    language,
    COUNT(language) AS language_used,
    (COUNT(language) / (SELECT COUNT(*) FROM job_data)) * 100 AS perc_share_language
FROM
    job_data
WHERE
    ds BETWEEN '2020-11-01' AND '2020-11-30'
GROUP BY language
ORDER BY language DESC;
```

- d) **Duplicate Rows Detection:** Identify duplicate rows in the data.

Query code used:

```
SELECT *
FROM
(
    SELECT *,
    row_number() over(partition by job_id) as rownum
    FROM job_data
)a
WHERE rownum>1
union
SELECT * from
(
    SELECT *,
    row_number() over(partition by actor_id) as rownum
    FROM job_data
)a
WHERE rownum>1;
```

Case study 2 : Investigating Metric Spike

a) **Weekly User Engagement:** Measure the activeness of users on a weekly basis.

Query code used:

```
SELECT
    EXTRACT(WEEK FROM occurred_at) AS week_num,
    COUNT(DISTINCT user_id) AS users_engage
FROM
    events
WHERE
    event_type = 'engagement'
GROUP BY week_num
ORDER BY week_num;
```

b) **User Growth Analysis:** Analyze the growth of users over time for a product

Query code used:

```
SELECT quarter_num, year, active_users,
    SUM(active_users) OVER (ORDER BY year, quarter_num rows between unbounded
preceding and current row) AS growth_of_users
FROM (
    SELECT EXTRACT(QUARTER FROM activated_at) AS quarter_num,
        EXTRACT (YEAR FROM activated_at) AS year,
        EXTRACT (DISTINCT user_id) AS active_users
    FROM users
    GROUP BY quarter_num, year
    ORDER BY quarter_num, year
)AS a;
```

- c) **Weekly Retention Analysis:** Analyze the retention of users on a weekly basis after signing up for a product.

Query code used:

```
SELECT
    EXTRACT(WEEK FROM occurred_at) AS week_num,
    COUNT(DISTINCT user_id) AS users_retained
FROM
    events
WHERE
    event_name = 'complete_signup'
    AND event_type = 'signup_flow'
GROUP BY week_num
ORDER BY week_num;
```

- d) **Weekly Engagement Per Device:** Measure the activeness of users on a weekly basis per device.

Query code used:

```
SELECT
    device,
    EXTRACT(WEEK FROM occurred_at) AS week_num,
    COUNT(DISTINCT user_id) AS users_engage
FROM
    events
WHERE
    event_type = 'engagement'
GROUP BY week_num , device
ORDER BY week_num;
```

e) **Email Engagement Analysis:** Analyze how users are engaging with the email service.

Query code used:

```
SELECT
  (SUM(CASE WHEN email_category = 'opened' THEN 1 ELSE 0 END) / SUM(CASE WHEN
email_category = 'email_sent' THEN 1 ELSE 0 END)) * 100 AS percentage_email_opened ,

  (SUM(CASE WHEN email_category = 'clickthrough' THEN 1 ELSE 0 END) / SUM(CASE WHEN
email_category = 'email_sent' THEN 1 ELSE 0 END)) * 100 AS percentage_email_clickthrough

FROM
  (SELECT *,
    CASE
      WHEN action IN ('email_clickthrough') THEN ('clickthrough')
      WHEN action IN ('email_open') THEN ('opened')
      WHEN action IN ('sent_weekly_digest' , 'sent_reengagement_email') THEN ('email_sent')
    END AS email_category
  FROM
    email_events) a;
```

Insights:

Case study 1 :

- Number of jobs reviewed per hour for each day in November 2020 are varied from **64.29** jobs/hour (lowest) on **26-11-2020** to **180** jobs/hour (highest) on **30-11-2020**.
- I used 7-day rolling throughput or weekly throughput as it directly gives us the insight on day 1st to day 7th. 7-day rolling throughput average was **0.03**
- Percentage share of each language for the past 30days are Persian (**37.5%**) being most used and every other language is **12.5%**
- There are 3 duplicate 2 from job_id and 1 from actor_id when we partition it by job_id, but if we look the overall data it's all unique.

Case study 2:

- In the weekly user engagement we saw a trend of increasing user engagement from 18th week to 30th week, and saw a gradual decrease from 31st week to 31st week, the highest number of user engagement was on 30th week being **1467**.
- There are total of **9381** number of active users from 1st quarter of 2013 to 3rd quarter of 2014. The highest growth rate of users was on 2nd quarter of 2014.
- From weekly retention analysis, the number of users on a weekly basis after signing up for a product retained **238 on 30th week** being the highest and the lowest went to **18 on 35th week**.
- As from the weekly engagement per device we saw a trend where users who were using **MacBook pro** and **Lenovo ThinkPad** were engaging more than others.
- The opening rate for email is around **34%** and the clickthrough rate is around **15%**. Means **15%** of the customers are engaging with the company through emails and **34%** of the customers are only opening the emails.

Results:

In this project, I learned how to perform operational analysis through my SQL knowledge where I used a bunch of different concepts like window functions, nested queries, aggregate functions etc. It helped me to vast my SQL knowledge. This project helped me know what kind of right questions to ask under different circumstances. It helped me gain knowledge on investigating metric spike which is the key concept of operational analysis.