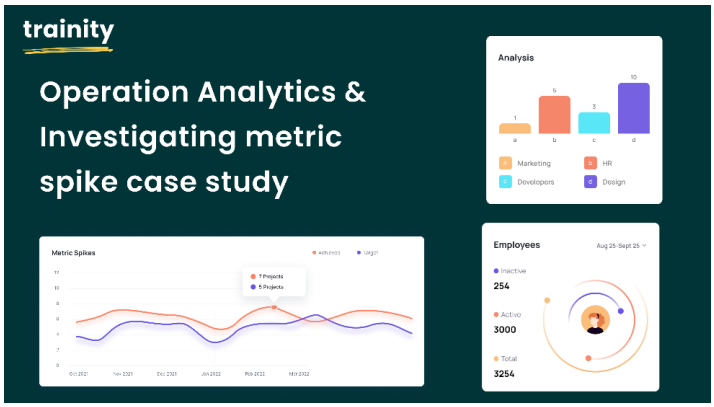
Operation Analytics and Investigating Metric Spike(Project-3)

Presented By: Srinivas Kamath



Project Description: Operation Analytics is the analysis of a company's whole end-to-end activities. The company can then use this to identify the areas where it needs to improve.

As one of the most important parts of a company, this type of analysis is also used to improve understanding among cross-functional teams and create more efficient workflows.

Investigating metric spikes is also an important element of operational analytics since, as Data Analysts, we must be able to grasp or make other teams understand queries such as "Why is there a drop in daily engagement?" Why have sales dropped? Etc. Such questions must be answered on a regular basis, and it is critical to study metric rise.

I work for a firm called Microsoft as a Data Analyst Lead and am given various data sets and tables from which I must derive specific insights and answer queries from various departments.

The following will be discovered through the projects:

- Number of jobs reviewed
- Throughput
- Percentage share of each language
- Duplicate rows
- User Engagement
- User Growth
- Weekly Retention
- Weekly Engagement
- Email Engagement

Approach: First, I took some time to understand the data/table provided. I answered my own questions, such as what the job_Id, Actor_Id, and event signify, and what factors to consider when examining the data. I use SQL to get various insights from the management team's dataset. I started by creating a database called "Operation_Analytics" and then the tables utilising the structure and linkages provided by the team. Then we conducted analysis to generate useful insights for the company.

Tech-Stack Used:

- MySQL Workbench (Version 8.0 CE): MySQL Workbench offers data modelling, SQL development, and other setup tools. It also has a graphical interface for working with databases in a systematic manner. It is simple and free to use MySQL to establish a database and conduct analysis in response to the questions posed in the description.
- Mode.com: It does advanced analytics quickly and provides useful insights. It does not necessitate any downloading or installation. We can connect Mode to our data warehouse. In Mode, I carried out Case Study 2 (investigating metric spike).

- Microsoft Word 2021: It is utilised to create a report (PDF) for the leadership team.

Execution:

Case Study 1 (Job Data):

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

➢
```
select
count(distinct job_id)/(30*24) as num_jobs_reviewed
from job_data
where
ds between '2020-11-01' and '2020-11-30';
```

Result Grid | Filter Rows:

num_jobs_reviewed
0.0083

B. **Throughput:** It is the no. of events happening per second.
My 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 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;
```

Result Grid | Filter Rows:

ds	jobs_reviewed	throughput_7
2020-11-25	1	1.0000
2020-11-26	1	1.0000
2020-11-27	1	1.0000
2020-11-28	2	1.2500
2020-11-29	1	1.2000
2020-11-30	2	1.3333

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

➢
```
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;
```

Result Grid | Filter Rows:

language	num_jobs	pct_share_lang
Arabic	1	16.66667
English	1	16.66667
French	1	16.66667
Hindi	1	16.66667
Italian	1	16.66667
Persian	1	16.66667

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

➢ select * from
(
select *,
row_number()over(partition by job_id) as rownum
from job_data
)a
where rownum>1;

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

job_id	actors_id	event	language	time_spent	org	ds	rownum
23	1005	transfer	Persian	22	D	2020-11-28	2
23	1004	skip	Persian	56	A	2020-11-26	3

Case Study 2 (Investigating metric spike):

A. **User Engagement:** To measure the activeness of a user. Measuring if the user finds quality in a product/service.
My task: Calculate the weekly 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;

Code:
https://app.mode.com/editor/srinivas_projects/reports/95543faa085f/queries/8c355dfcc05a

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

➢ 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;

Code:
https://app.mode.com/editor/srinivas_projects/reports/95543faa085f/queries/8c355dfcc05a

C. **Weekly Retention:** Users getting retained weekly after signing-up for a product.
My task: Calculate the weekly retention of users-sign up cohort?

- ```
  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 occured_at) as sign_up_week
  from tutorial.yammer_events
  where event_type = 'signup_flow'
  and event_name = 'complete_signup'
  and extract(week from occured_at)=18)a
  left join
  (select distinct user_id, extract(week from occured_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: To measure the activeness of a user. Measuring if the user finds quality in a product/service weekly.
My task: Calculate the weekly engagement per device?

- ```
  select extract(year from occured_at) as year_num,
  extract(week from occured_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: Users engaging with the email service.
My task: Calculate the email engagement metrics?

- ```
  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):

- For November 2020, the number of distinct jobs reviewed per hour per day is 83%.
- We utilised the 7-day rolling average of throughput because it shows the average for all days from day 1 to day 7, whereas the daily indicator just delivers the average for that day.
- Persian has the highest percentage share (37.5%).
- If we segment the data by job_id, we find two duplicate rows. However, if we look at the overall columns, we can see that each row is distinct.

Case Study 2 (Investigating metric spike):

- Weekly user involvement climbed from the 18th to the 31st week and then began to decline. This signifies that some users have expressed dissatisfaction with the product/service in recent weeks.
- From the first week of 2013 to the 35th week of 2014, there were a total of 9381 active users.
- MacBook and iPhone users have the highest total weekly involvement per device utilised.
- The email opening rate is approximately 34%, while the email clicking rate is approximately 15%. Users are engaging with the email service, which is beneficial to the company's growth.

Result: This project taught me how to use sophisticated SQL concepts such as Windows Functions, etc. I was aware of how the real-world industry operates. It aided me in understanding SQL concepts. Given the conditions, I learnt how to ask the appropriate questions. Which columns to evaluate from the given data and queries, and how to obtain significant insights that will help the firm grow? I discovered how the corporation investigates various aspects of the business in order to improve it further. I learned about metric spike investigation (why there is a boom and why there is a dip).