Project Description:

The project involves analysing the operation analytics of a company using SQL. The data provided is in the form of tables, which include information about job data, user engagement, user growth, and email engagement. The aim of the project is to derive insights from the data and provide answers to various questions related to different operations of the company.

Approach

The first step was to create a database and tables using SQL. Then, the analysis was performed by running queries on the tables. The questions related to job data were analysed first, followed by the questions related to investigating metric spike.

Execution:

Case Study 1 (Job Data):

1). 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';
```

2) 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;
```

3). 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;
```

4). **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;
```

Case Study 2 (Investigating metric spike):

1). 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;
```

2). 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;
```

3). **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;
```

4). 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;
```

5). 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;
```

Tech-Stack Used

Mode.com: It perform advanced analytics quickly and deliver valuable insights. It does not required any download and installation. We can connect our data warehouse with Mode. I performed case study 2 (investigating metric spike) in Mode. MySQL Workbench (Version 8.0 CE): MySQL Workbench provides data modelling, SQL development, and various administration tools for configuration. It also offers a graphical interface to work with the databases in a structured way. It is easy and free to use MySQL to create a database and perform analysis answering the questions given in the description. Microsoft Word 2021: It is used to make a report (PDF) to be presented to the leadership team.

Insights

Case Study 1 (Job Data):

- November 2020: The number of distinct jobs reviewed per hour per day is 83%.
- As it gives the average for all the days right from day 1 to day 7, we used the 7-day rolling average of throughput whereas daily metric gives the average for only that particular day itself
- The most percentage share of Persian language is 37.5%.
- If we partition the data by job_id, there are two duplicate rows. But all the rows are unique if we look at the overall columns.

Case Study 2 (Investigating metric spike):

- From week 18th to week 31st, the weekly user engagement increased and then started declining from then onwards. This means that some of the users do not find much quality in the product/service in the last of the weeks.
- From the 1st week of 2013 to the 35th week of 2014, there are in total 9381 active users.
- The most overall count of weekly engagement per device used is for MacBook users and iPhone users.
- The users are engaging with the email service, which is good for the company to expand. The email opening rate is around 34%, and the email clicking rate is around 15%.

Result

The project helped in understanding the importance of operation analytics for a company and how data analysis can provide valuable insights for improving the performance of various operations. The insights derived from the analysis can be used by the leadership team to make data-driven decisions and improve the overall efficiency and growth of the company.