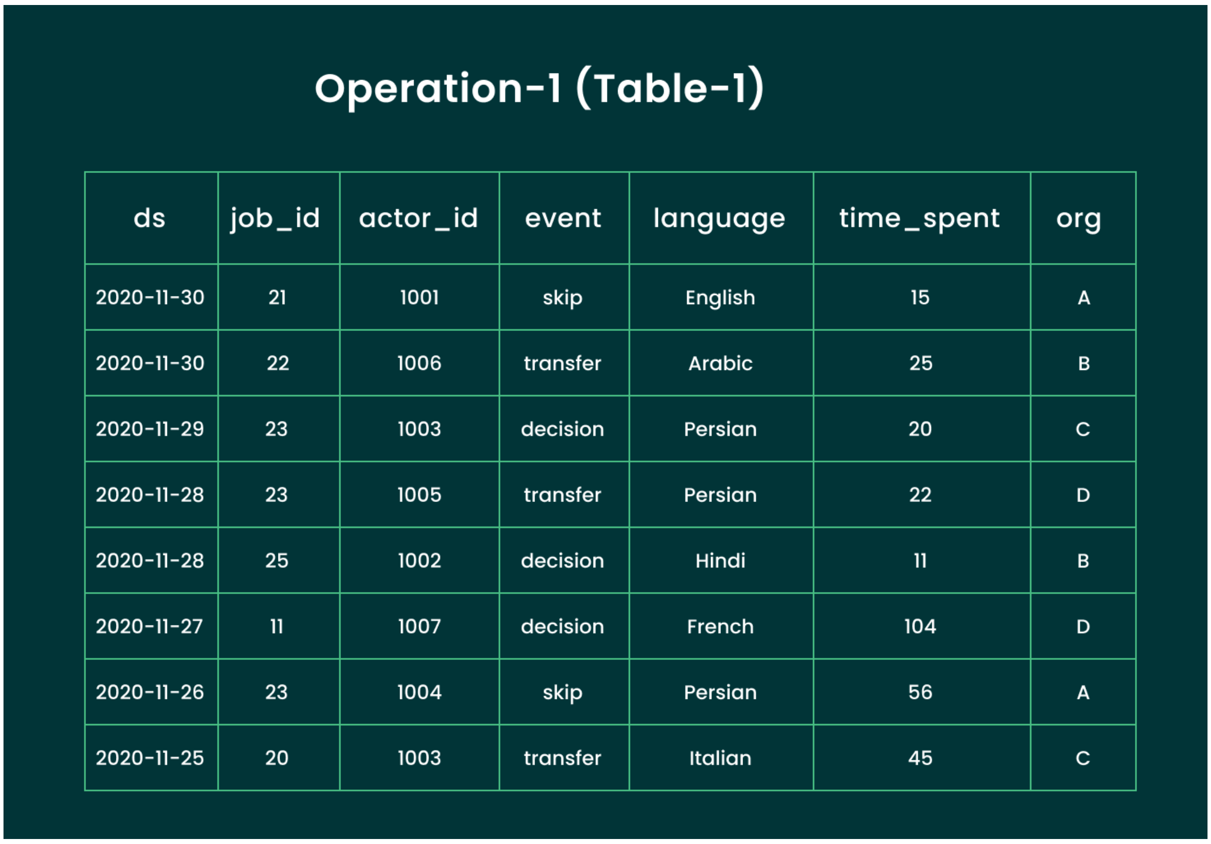
Operation Analytics and Investigating Metric Spike

Case Study 1 (Job Data)

Below is the structure of the table with the definition of each column that you must work on:

- **Table-1:** job_data



Operation-1 (Table-1)

ds	job_id	actor_id	event	language	time_spent	org
2020-11-30	21	1001	skip	English	15	A
2020-11-30	22	1006	transfer	Arabic	25	B
2020-11-29	23	1003	decision	Persian	20	C
2020-11-28	23	1005	transfer	Persian	22	D
2020-11-28	25	1002	decision	Hindi	11	B
2020-11-27	11	1007	decision	French	104	D
2020-11-26	23	1004	skip	Persian	56	A
2020-11-25	20	1003	transfer	Italian	45	C

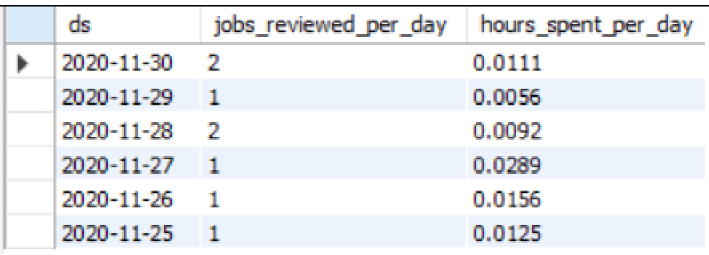
  - **job_id:** unique identifier of jobs
  - **actor_id:** unique identifier of actor
  - **event:** decision/skip/transfer
  - **language:** language of the content
  - **time_spent:** time spent to review the job in seconds
  - **org:** organization of the actor
  - **ds:** date in the yyyy/mm/dd format. It is stored in the form of text and we use presto to run. no need for date function

Use the dataset attached in the Dataset section below the project images then answer the questions that follows

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 ds,count(job_id) as jobs_reviewed_per_day,
sum(time_spent)/3600 as hours_spent_per_day
from job_data
```

```
where ds >='2020-11-01'  and ds <='2020-11-30'
group by ds ;
```

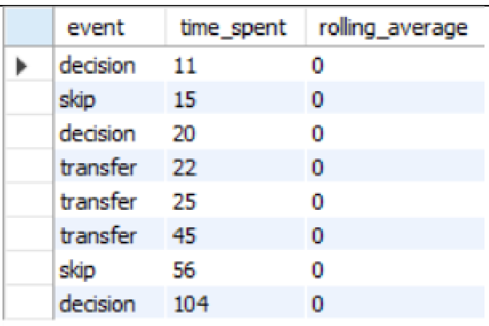


ds	jobs_reviewed_per_day	hours_spent_per_day
2020-11-30	2	0.0111
2020-11-29	1	0.0056
2020-11-28	2	0.0092
2020-11-27	1	0.0289
2020-11-26	1	0.0156
2020-11-25	1	0.0125

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?

I prefer 7-day moving average over the daily metric as they are used to discover trends in both short-term and long-term trend or cycle. Using time periods, we can compare the short, medium and long term momentum.

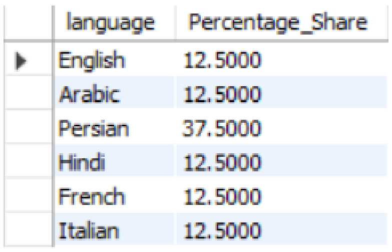
```
SELECT event, time_spent,
AVG(event)
OVER(ORDER BY time_spent ROWS BETWEEN 6 PRECEDING AND CURRENT ROW)
AS rolling_average
FROM job_data;
```



event	time_spent	rolling_average
decision	11	0
skip	15	0
decision	20	0
transfer	22	0
transfer	25	0
transfer	45	0
skip	56	0
decision	104	0

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?

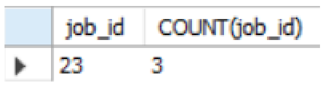
```
Select language, (Count(language)* 100 / (Select Count(*) From
job_data)) as Percentage_Share
From job_data
Group By language;
```



language	Percentage_Share
English	12.5000
Arabic	12.5000
Persian	37.5000
Hindi	12.5000
French	12.5000
Italian	12.5000

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 job_id, COUNT(job_id)
FROM job_data
GROUP BY job_id
HAVING COUNT(job_id) > 1;
```

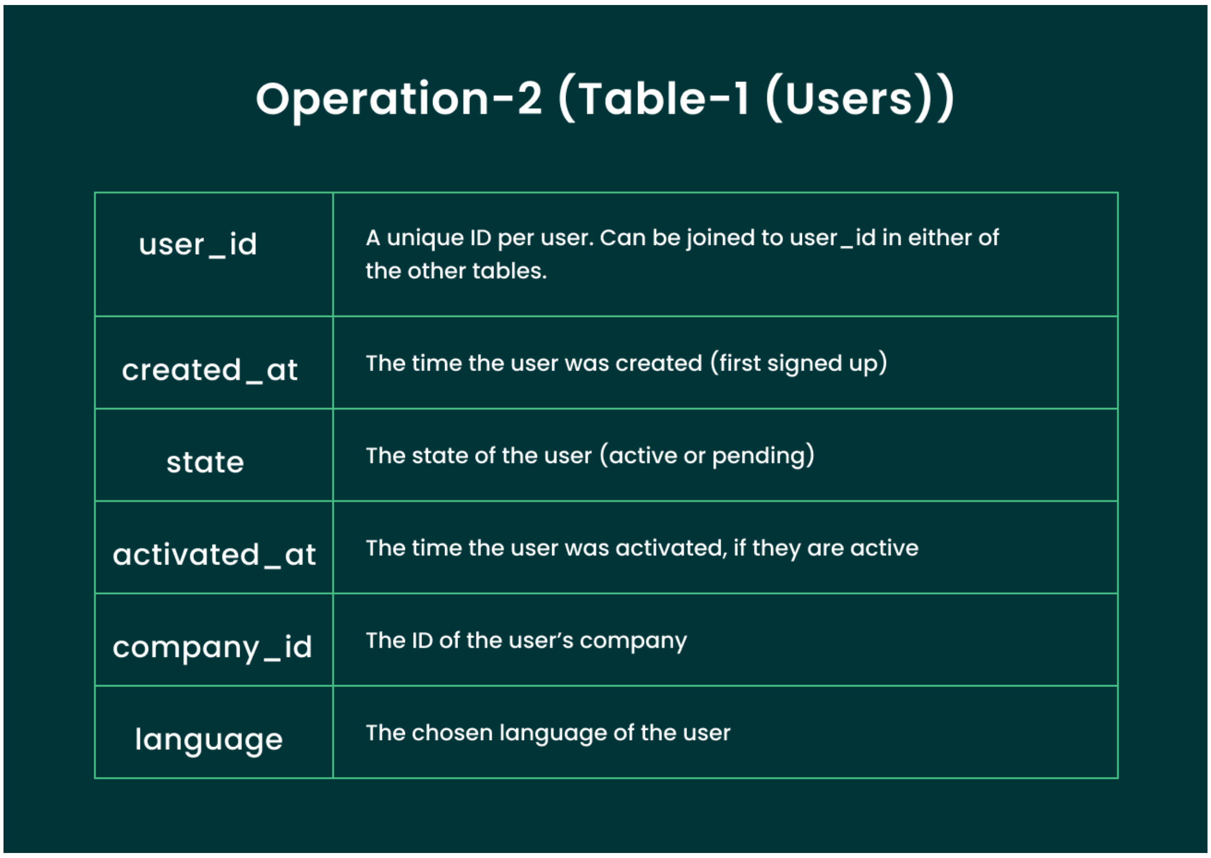


job_id	COUNT(job_id)
23	3

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

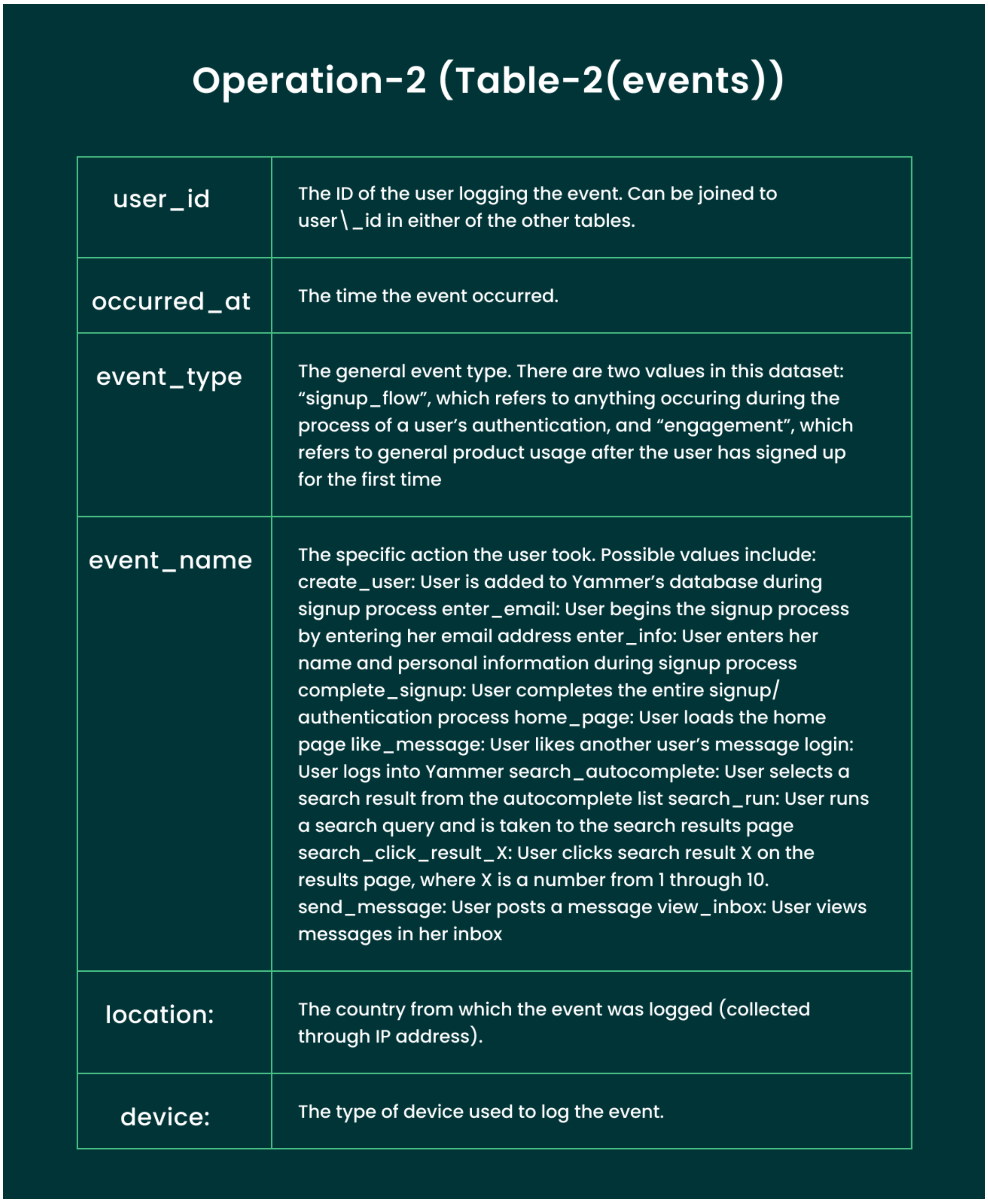
- **Table-1:** users
  This table includes one row per user, with descriptive information about that user’s account.



Operation-2 (Table-1 (Users))

user_id	A unique ID per user. Can be joined to user_id in either of the other tables.
created_at	The time the user was created (first signed up)
state	The state of the user (active or pending)
activated_at	The time the user was activated, if they are active
company_id	The ID of the user’s company
language	The chosen language of the user

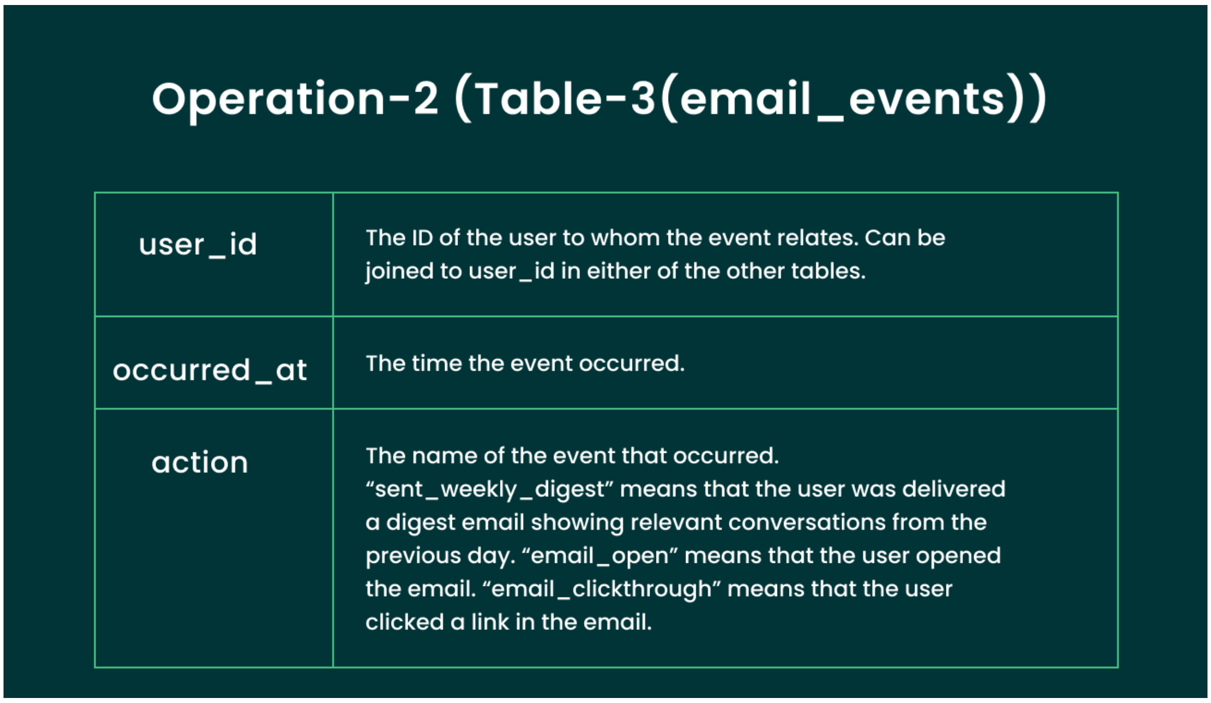
- **Table-2:** events
  This table includes one row per event, where an event is an action that a user has taken. These events include login events, messaging events, search events, events logged as users progress through a signup funnel, events around received emails.



Operation-2 (Table-2(events))

user_id	The ID of the user logging the event. Can be joined to user_id in either of the other tables.
occurred_at	The time the event occurred.
event_type	The general event type. There are two values in this dataset: "signup_flow", which refers to anything occuring during the process of a user's authentication, and "engagement", which refers to general product usage after the user has signed up for the first time
event_name	The specific action the user took. Possible values include: create_user: User is added to Yammer's database during signup process enter_email: User begins the signup process by entering her email address enter_info: User enters her name and personal information during signup process complete_signup: User completes the entire signup/ authentication process home_page: User loads the home page like_message: User likes another user's message login: User logs into Yammer search_autocomplete: User selects a search result from the autocomplete list search_run: User runs a search query and is taken to the search results page search_click_result_X: User clicks search result X on the results page, where X is a number from 1 through 10. send_message: User posts a message view_inbox: User views messages in her inbox
location:	The country from which the event was logged (collected through IP address).
device:	The type of device used to log the event.

- **Table-3:** email_events
  This table contains events specific to the sending of emails. It is similar in structure to the events table above.



Operation-2 (Table-3(email_events))

user_id	The ID of the user to whom the event relates. Can be joined to user_id in either of the other tables.
occurred_at	The time the event occurred.
action	The name of the event that occurred. "sent_weekly_digest" means that the user was delivered a digest email showing relevant conversations from the previous day. "email_open" means that the user opened the email. "email_clickthrough" means that the user clicked a link in the email.

Use the dataset attached in the Dataset section below the project images then answer the questions that follows

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 user_id,created_at
FROM users
where date(created_at) BETWEEN (NOW() - INTERVAL 7 DAY) AND NOW();
```

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

```
SELECT user_id, company_id,
LEAD(company_id) OVER(ORDER BY user_id) as user_growth
FROM users
where state = 'active';
```

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

```
select *
from events
left join events as future_retention on
  events.user_id = users.user_id
  and events.occured_at = users.activated_at - interval '1 day'
```

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 user_id,occurred_at,device
FROM events
where date(occurred_at) BETWEEN (NOW() - INTERVAL 7 DAY) AND NOW();
```

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

```
select *
FROM email_events
where date(occurred_at) BETWEEN (NOW() - INTERVAL 1 DAY) AND NOW();
```