Project 03

Operation Analytics and Investigating Metric Spike

NAME :- NIRAJ INGOLE

INTERNSHIP PROJECT

USING : SQL Fundamentals

Description:

Operation Analytics is the analysis done for the complete end to end operations of a company. With the help of this, the company then finds the areas on which it must improve upon. You work closely with the ops team, support team, marketing team, etc and help them derive insights out of the data they collect.Being one of the most important parts of a company, this kind of analysis is further used to predict the overall growth or decline of a company's fortune. It means better automation, better understanding between cross-functional teams, and more effective workflows.

Investigating metric spike is also an important part of operation analytics as being a Data Analyst you must be able to understand or make other teams understand questions like- Why is there a dip in daily engagement? Why have sales taken a dip? Etc. Questions like these must be answered daily and for that its very important to investigate metric spike.You are working for a company like Microsoft designated as Data Analyst Lead and is provided with different data sets, tables from which you must derive certain insights out of it and answer the questions asked by different departments.

I will be using SQL to derive at solution for the problem statements.

You are required to provide a detailed report answering the questions below of Two Case Study:

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
  - **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

dataset link:-

https://drive.google.com/drive/folders/1bB-uqONISA6wiI1hw1LzISpe0-kHg0Nx

1. **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?

2. **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?

3. **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?

4. **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?

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

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?
B. **User Growth**: Amount of users growing over time for a product.
   Your task: Calculate the user growth for product?
C. **Weekly Retention:** Users getting retained weekly after signing-up for a product.
   Your task: Calculate the weekly retention of users-sign up cohort?

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

How to do this Project?

- Create the Database and Tables: You are supposed to create a database and then the tables using the structure and links provided.
- Perform Analysis: Use SQL to perform your entire analysis answering the questions asked above.
- Submit a Report: Make a report (PDF/PPT) to be presented to the leadership team. The report should/can contain the following details:

Case Study 1 (Job Data)

Dataset

```
use project03;
create table job_data(job_id int,actors_id int,event varchar(255),
language varchar(255),time_spent int,org varchar(255),ds date);
 #Data Insert into Table Query:
insert into job_data (job_id,
actors_id, event, language, time_spent, org, ds)
values
(21,1001,'skip','English',15,'A','2020-11-30'),
(22,1006,'transfer', 'Arabic',25,'B','2020-11-30'),
(23,1003,'decision','Persian',20,'C','2020-11-29'),
(24,1005,'transfer','Persian',22,'D','2020-11-28'),
(25,1002,'decision','Hindi', 11,'B','2020-11-28'),
(26,1007, 'decision', 'French', 104,'D', '2020-11-27'),
(27, 1004, 'skip', 'Persian', 56, 'A', '2020-11-26'),
(28,1008, 'transfer', 'Italian', 45,'C','2020-11-25'),
(28, 1008, 'transfer', 'Italian', 45,'C','2020-11-25');
SELECT job_id,actors_id,event,language,time_spent,org,ds,COUNT(*) AS CNT
from project03.job_data
GROUP BY job_id,actors_id,event,language,time_spent,org,ds
HAVING COUNT(*)>1;
```

job_id	actors_id	event	language	time_spent	org	ds	CNT
21	1001	skip	English	15	A	2020-11-30	2
22	1006	transfer	Arabic	25	B	2020-11-30	2
23	1003	decision	Persian	20	C	2020-11-29	2
24	1005	transfer	Persian	22	D	2020-11-28	2
25	1002	decision	Hindi	11	B	2020-11-28	2
26	1007	decision	French	104	D	2020-11-27	2
27	1004	skip	Persian	56	A	2020-11-26	2
28	1008	transfer	Italian	45	C	2020-11-25	4

1. **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?

```
/
1.  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
count(distinct job_id)/(30*24) as reviewed_per_hour_per
from project03.job_data  where ds  between
2020-11-01 and 2020-11-30;
```

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

reviewed_per_hour_per
0.0000

2. **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 ds,job_review,avg(job_review) over
(order by ds rows between 5 preceding and current row)as throughput
from (select ds,count(distinct job_id) as job_review
from project03.job_data
where ds between "2020-11-01" and"2020-11-30"
group by ds order by ds)a;
```

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

ds	job_review	throughput
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

3.**<u>Percentage share of each language</u>**: Share of each language for different contents.

Your Task: Calculate the percentage share of each language in the last 30 days?

```
/*
3.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 *from project03.job_data;
select language,job_no ,100.0*job_no/all_job  as "%_job"from (select
language,count(distinct job_id) as job_no from project03.job_data group by language)a
cross join
(select count(distinct job_id)as all_job
from project03.job_data)b
```

language	job_no	%_job
Arabic	1	12.50000
English	1	12.50000
French	1	12.50000
Hindi	1	12.50000
Italian	1	12.50000
Persian	3	37.50000

4. **<u>Duplicate rows:</u>** that have the same value present in them.

Your :Let’s say you see some duplicate rows in the data. How will you display duplicates from the table?

```
/*
4. 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 * from
(select *,row_number()over(partition by job_id)as rownum
from project03.job_data)a
where rownum>1;
```

job_id	actors_id	event	language	time_spent	org	ds	rownum
28	1008	transfer	Italian	45	C	2020-11-25	2

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.
Your task: Calculate the weekly user engagement?

```
/*
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.
            Your task: Calculate the weekly user engagement?
*/
select*,
extract (week from accurred_at) as weeknum,count(distinct user_id)
from project03.table_2_events group by weeknum;
```

Output

weeknum	count
18	791
19	1244
20	1270
21	1341
22	1293
23	1366
24	1434
25	1462
26	1443
27	1477
28	1556
29	1556
30	1593
31	1685
32	1483
33	1438
34	1412
35	1442

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

```
B.  User Growth: Amount of users growing over time for a product.
Your task: Calculate the user growth for product?
*/
select year,weeknum,num_active_users,
sum(num_active_users) over(order by year,
weeknum rows between unbounded preceding
and current row)
cum_active_users fromm (select extract(year from a.activated_at)as year,
extract(week from a.activated_at) as weeknum,count(distinct user_id)num_active_users
from
project03.user_table-1 a
where state="active"
group by  weeknum,year)a
```

Output

100% View only

fx year

A	B	C	D
year	weeknum	num_active_use	cum_active_users
2013	1	67	67
2013	2	29	96
2013	3	47	143
2013	4	36	179
2013	5	30	209
2013	6	48	257
2013	7	41	298
2013	8	39	337
2013	9	33	370
2013	10	43	413
2013	11	33	446
2013	12	32	478
2013	13	33	511
2013	14	40	551
2013	15	35	586
2013	16	42	628
2013	17	48	676
2013	18	48	724
2013	19	45	769
2013	20	55	824
2013	21	41	865
2013	22	49	914
2013	23	51	965
2013	24	51	1016
2013	25	46	1062

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

```
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
count(user_id),sum(case when retention_week=1 then 1 else 0 end) as week_1
from(select a.user_id,a.sigmup_week,b.engagement_week,b.engagement_week-a.sigmup_week as retention_week
from
(select distinct user_id,extract(week from accurred_at)as signup_week
from project03.table-2_event
where event_type ="signup-flow" and event-name ="complete_signup"
and extract (week from accurred_at) =18)a
left jone(select distinct user_id,extract(week from accurred_at )as engagement_week
from project03.table-2_event where event_type ="engagement")b on a.user_id =b.user_id
order by a.user_id)a
```

output

	count	week_1
1	317	64

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?

```
/*
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 extract(year  from accurred_at)as year,
extract(week from accurred_at)as week,device,
count(distinct user_id)from  project03.table-2_events
where event_type="engagement"
group by 1,2,3
order by 1,2,3
```

output

100% View only

fx year

Ctrl + J)	B	C	D
	week	device	count
2014	18	acer aspire desk	10
2014	18	acer aspire notel	21
2014	18	amazon fire pho	4
2014	18	asus chromeboc	23
2014	18	dell inspiron des	21
2014	18	dell inspiron note	49
2014	18	hp pavilion deskt	15
2014	18	htc one	16
2014	18	ipad air	30
2014	18	ipad mini	21
2014	18	iphone 4s	21
2014	18	iphone 5	70
2014	18	iphone 5s	45
2014	18	kindle fire	6
2014	18	lenovo thinkpad	90
2014	18	macbook air	57
2014	18	macbook pro	154
2014	18	mac mini	8
2014	18	nexus 10	16
2014	18	nexus 5	43
2014	18	nexus 7	20
2014	18	nokia lumia 635	19
2014	18	samsumg galaxy	8
2014	18	samsung galaxy	7
2014	18	samsung galaxy	56

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

```
order by 1,2,3
/*
E. Email Engagement: Users engaging with the email service.
Your task: Calculate the email engagement metrics?

*/
sek=lect 100.0*sum(case when email_cat ="email_open"
then 1 else 0 end)/sum(case when email_cat ="email_sent"
then 1 else 0 end)as email_open_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_clicked_rate
from (select*,case when action in ("sent_weekly_digest","sent_reengagement_email")
then"email_open" when action in ("email_clickthrugh")
then "email-clicked" end as email_cat from project.table-3_email event)a
```

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

output

Output:

	email_open_rate	email_clicked_rate
1	33.5834	14.7899

Result

Operational analytics is the process of using data analysis and business intelligence to improve efficiency and streamline everyday operations in real time. A subset of business analytics, operational analytics is supported by data mining, artificial intelligence, and machine learning.