

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-ugONISA6wil1hw1LzlSpe0-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

/SQL workbench

The screenshot displays the SQL Workbench interface. The left sidebar shows a tree view of database objects, including tables like 'auth_group' and 'auth_group_permissions'. The main editor window contains the following SQL script:

```
2 • use project03;
3 • create table job_data(job_id int,actors_id int,event varchar(255),
4   language varchar(255),time_spent int,org varchar(255),ds date);
5   #Data Insert into Table Query:
6 • insert into job_data (job_id,
7   actors_id, event, language, time_spent, org, ds)
8   values
9   (21,1001,'skip','English',15,'A','2020-11-30'),
10  (22,1006,'transfer','Arabic',25,'B','2020-11-30'),
11  (23,1003,'decision','Persian',20,'C','2020-11-29'),
12  (24,1005,'transfer','Persian',22,'D','2020-11-28'),
13  (25,1002,'decision','Hindi',11,'B','2020-11-28'),
14  (26,1007,'decision','French',104,'D','2020-11-27'),
15  (27,1004,'skip','Persian',56,'A','2020-11-26'),
16  (28,1008,'transfer','Italian',45,'C','2020-11-25'),
17  (28,1008,'transfer','Italian',45,'C','2020-11-25');
18 • SELECT job_id,actors_id,event,language,time_spent,org,ds,COUNT(*) AS CNT
19   from project03.job_data
20   GROUP BY job_id,actors_id,event,language,time_spent,org,ds
21   HAVING COUNT(*)>1;
```

Below the script, the 'Result Grid' is displayed, showing the output of the query. The grid has columns for job_id, actors_id, event, language, time_spent, org, ds, and CNT. The data is as follows:

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?

```
24 1. Number of jobs reviewed: Amount of jobs reviewed over time.
25 Your Task: Calculate the number of jobs reviewed per hour per day for November 2020?
26 */
27 • select
28 count(distinct job_id)/(30*24) as reviewed_per_hour_per
29 from project03.job_data where ds between
30 2020-11-01 and 2020-11-30;
31
```

Result Grid

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?

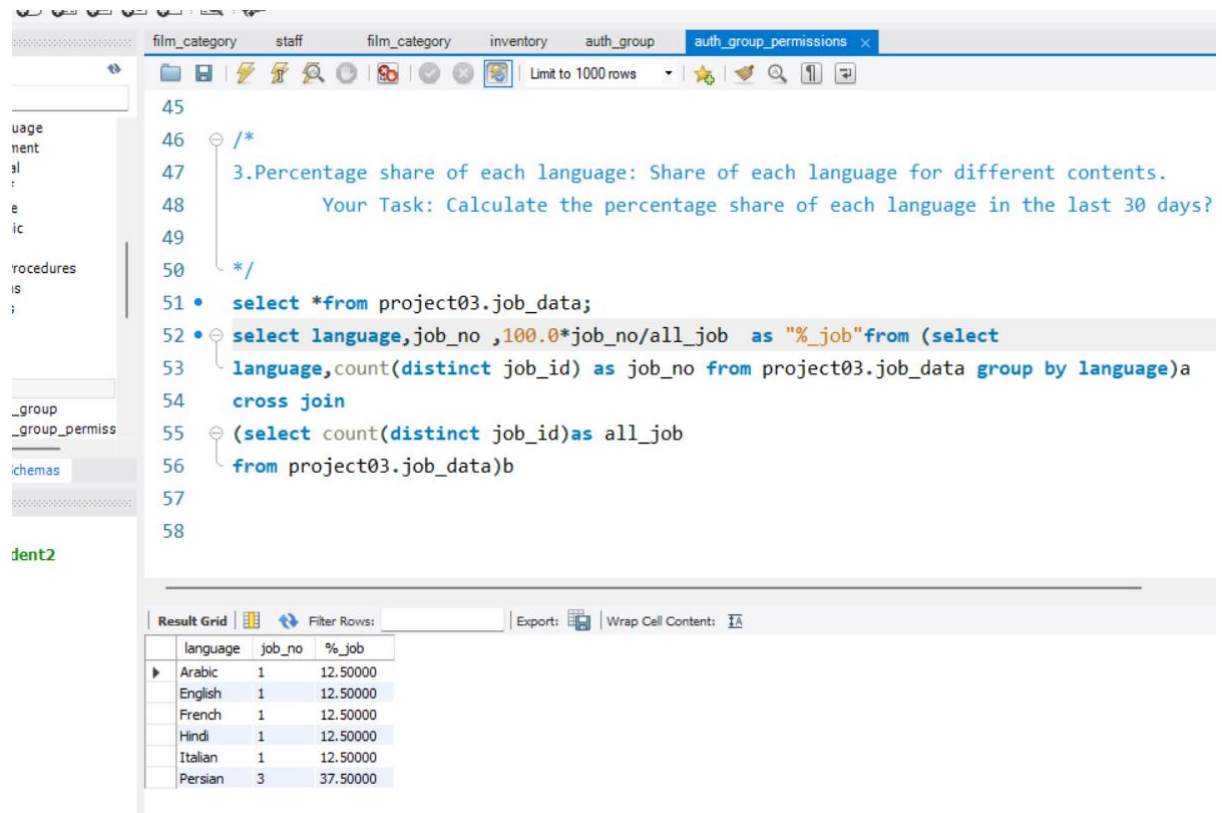
```
37 */
38 • select ds,job_review,avg(job_review) over
39 (order by ds rows between 5 preceding and current row)as throughput
40 from (select ds,count(distinct job_id) as job_review
41 from project03.job_data
42 where ds between "2020-11-01" and"2020-11-30"
43 group by ds order by ds)a;
44
```

Result Grid

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. 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?

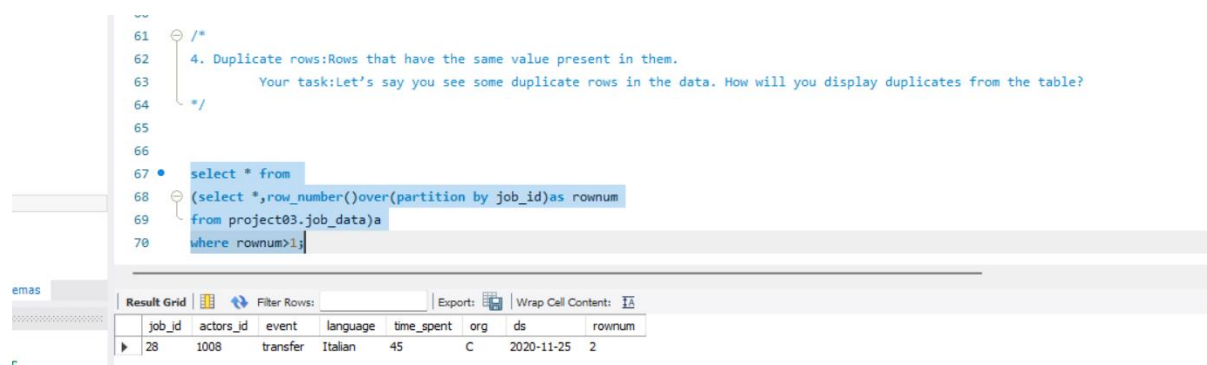


The screenshot shows a database IDE with a SQL editor and a result grid. The SQL editor contains a query to calculate the percentage share of each language in the last 30 days. The result grid displays the following data:

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. Duplicate rows: 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?



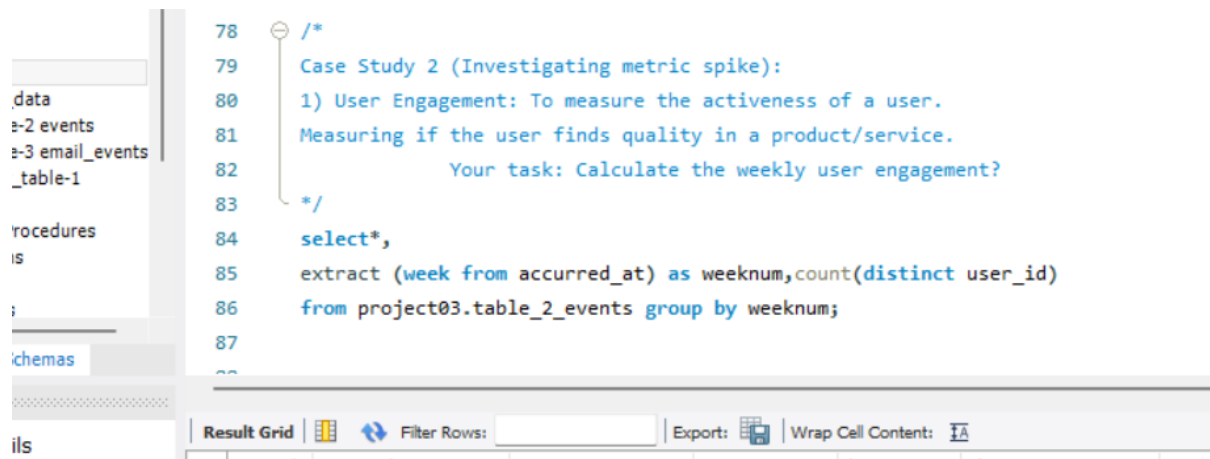
The screenshot shows a database IDE with a SQL editor and a result grid. The SQL editor contains a query to display duplicate rows. The result grid displays the following data:

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.

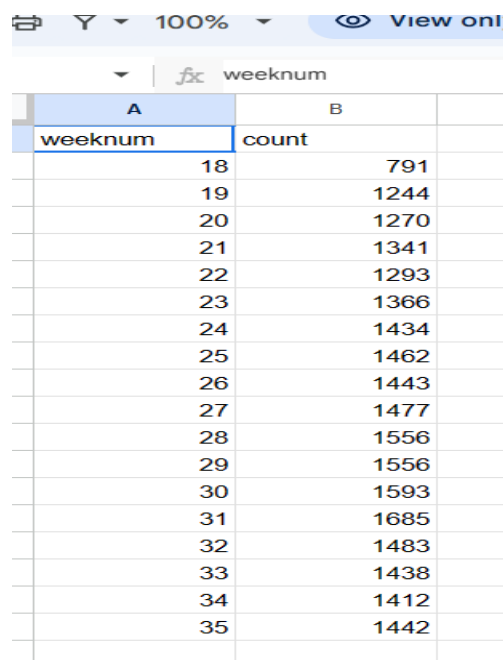
Your task: Calculate the weekly user engagement?



```
78  /*
79  Case Study 2 (Investigating metric spike):
80  1) User Engagement: To measure the activeness of a user.
81  Measuring if the user finds quality in a product/service.
82      Your task: Calculate the weekly user engagement?
83  */
84  select*,
85  extract (week from occurred_at) as weeknum,count(distinct user_id)
86  from project03.table_2_events group by weeknum;
87
```

The screenshot shows a SQL IDE interface. On the left, there's a sidebar with a tree view containing 'data', 'table-2 events', 'table-3 email_events', and 'table-1'. Below this, there are sections for 'procedures' and 'schemas'. The main area displays a SQL query. The query is a SELECT statement with a comment block at the top. The comment block describes the case study and the task. The query itself selects all columns from 'project03.table_2_events', grouped by 'weeknum' (extracted from 'occurred_at'), and counts the distinct 'user_id' for each week. The bottom of the screenshot shows a toolbar with 'Result Grid', 'Filter Rows', 'Export', and 'Wrap Cell Content' buttons.

Output



weeknum	
A	B
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

The screenshot shows a table viewer interface. At the top, there's a toolbar with a refresh icon, a dropdown menu set to '100%', and a 'View only' button. Below the toolbar, there's a table with two columns: 'weeknum' (labeled 'A') and 'count' (labeled 'B'). The table contains 18 rows of data, with 'weeknum' values ranging from 18 to 35 and 'count' values ranging from 791 to 1685. The table is displayed in a standard grid format with alternating row colors.

B. User Growth: Amount of users growing over time for a product.

Your task: Calculate the user growth for product?

```
90  B. User Growth: Amount of users growing over time for a product.
91  Your task: Calculate the user growth for product?
92  */
93  select year,weeknum,num_active_users,
94  sum(num_active_users) over(order by year,
95  weeknum rows between unbounded preceding
96  and current row)
97  cum_active_users fromm (select extract(year from a.activated_at)as year,
98  extract(week from a.activated_at) as weeknum,count(distinct user_id)num_active_users
99  from
100  project03.user_table-1 a
101  where state="active"
102  group by weeknum,year)a
103
104
```

Output

View only				
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?

```

104 C. Weekly Retention: Users getting retained weekly after signing-up for a product.
105 Your task: Calculate the weekly retention of users-sign up cohort?
106
107 */
108 select
109 count(user_id),sum(case when retention_week=1 then 1 else 0 end) as week_1
110 from(select a.user_id,a.signup_week,b.engagement_week,b.engagement_week-a.signup_week as retention_week
111 from
112 (select distinct user_id,extract(week from occurred_at)as signup_week
113 from project03.table-2_event
114 where event_type ="signup-flow" and event-name ="complete_signup"
115 and extract (week from occurred_at) =18)a
116 left join(select distinct user_id,extract(week from occurred_at )as engagement_week
117 from project03.table-2_event where event_type ="engagement")b on a.user_id =b.user_id
118 order by a.user_id)a
119

```

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?

```

121 /*
122 D. Weekly Engagement: To measure the activeness of a user. Measuring if the user finds quality in a product/service weekly.
123 Your task: Calculate the weekly engagement per device?
124 */
125 select extract(year from occurred_at)as year,
126 extract(week from occurred_at)as week,device,
127 count(distinct user_id)from project03.table-2_events
128 where event_type="engagement"
129 group by 1,2,3
130 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 note	21
2014	18	amazon fire pho	4
2014	18	asus chromebo	23
2014	18	dell inspiron des	21
2014	18	dell inspiron note	49
2014	18	hp pavilion desk	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	samsung 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?

```

130     order by 1,2,3
131  /*
132   E. Email Engagement: Users engaging with the email service.
133   Your task: Calculate the email engagement metrics?
134  */
135  */
136  sek=lect 100.0*sum(case when email_cat ="email_open"
137  then 1 else 0 end)/sum(case when email_cat ="email_sent"
138  then 1 else 0 end)as email_open_rate,100.0*sum(case when email_cat ="email_clicked"
139  then 1 else 0 end)/sum(case when email_cat ="email_sent"
140  then 1 else 0 end)as email_clicked_rate
141  from (select*,case when action in ("sent_weekly_digest","sent_reengagement_email")
142  then"email_open" when action in ("email_clickthrough")
143  then "email-clicked" end as email_cat from project.table-3_email event)a
144

```

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

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.