Operation Analytics and Investigating Metric Spike

Project Description:

Operation Analytics is a crucial process that involves analyzing a company's end to end operations. This analysis helps identify areas for improvement within the company. As a Data Analyst, I will work closely with various teams, such as operations, support and marketing, helping them derive valuable insights from the data collected.

One of the key aspects of Operation Analytics is investigating metric spikes. This involves understanding and explaining sudden changes in **key metrics**, such as dip in daily user engagement or a drop in sales. As a Data Analyst these questions needs to be answered daily, making it crucial to understand how to investigate these **metric spikes**.

In this project, I will work as a Lead Data Analyst at a company like Microsoft, where I'll be provided with various datasets and tables, and my task will be to derive insights from the data to answer questions posed by different departments within the company. My goal is to use my advanced SQL skills to analyze the data and provide valuable insights that can help improve the company's operations and understand sudden changes in key metrics.

Approach:

For the project, the data is collected from various departments and combined into various tables, ensuring data integrity and quality. For the **Case Study-1**, we have a small dataset so, it can directly be imported using the **Table Data Import Wizard**, however, for **Case Study-2**, since we have a large dataset. So, we need to make use of **load data infile** to import the dataset. After importing the dataset, we see that the column containing the dates are stored as text and not datetime, so we need to alter them and change their datatype.

After this, through SQL queries, we will explore and analyze the patterns, trends in the data. We will find out the key performance metrics and their change with time. Based on the findings, the insights and recommendations will be made for various departments to make data-driven informed decisions for process improvement, increase in sales, user engagement and growth of the company.

Tech-Stack Used:

For the project, **MYSQL Workbench 8.0** which is owned by Oracle was used as this is the most widely used database software in the industry. It has a user-friendly GUI tool for MYSQL which allows to analyze the data and work on it easily. Also, it allows to virtually create physical database design models that can easily be transitioned into MYSQL databases.

CASE STUDY-1: Job Data Analysis

Task 1: Jobs Reviewed Over Time

```sql
select * from job_data;
SELECT
    DATE_FORMAT(ds, '%d-%M-%Y') AS Day,
    round(COUNT(job_id) / sum(time_spent)*3600) AS No_of_Jobs_Reveiwed
FROM
    job_data
WHERE
    ds BETWEEN '2020-11-01' AND '2020-11-30'
GROUP BY 1
ORDER BY 1;
```

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

Day	No_of_Jobs_Reveiwed
25-November-2020	80
26-November-2020	64
27-November-2020	35
28-November-2020	218
29-November-2020	180
30-November-2020	180

Here, we see the total number of jobs reviewed per hour for each day in November 2020. The maximum jobs reviewed in a day was on **28-November-2020** which is **218**. The number of jobs being reviewed has the lowest count of **35** and highest being **218.** We need to find out the reason for these differences and work on it to get a consistent and better result, thus improving overall performance.

Task 2: Throughput Analysis

```sql
select round((count(event)/sum(time_spent)),2) as Weekly_throughput
from job_data;

select ds,
       throughput,
       avg(throughput)over(order by ds rows between 6 preceding and current row) as Daily_throughput
from(
select ds, count(event)/sum(time_spent)  as throughput from job_data
where ds between '2020-11-24' and '2020-11-30'group by ds order by ds ) as a;
```

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

Weekly_throughput
0.03

ds	throughput	Daily_throughput
2020-11-25	0.0222	0.02220000
2020-11-26	0.0179	0.02005000
2020-11-27	0.0096	0.01656667
2020-11-28	0.0606	0.02757500
2020-11-29	0.0500	0.03206000
2020-11-30	0.0500	0.03505000

Here, we find out the 7-day rolling average of throughput (number of events per second). This provides us to analyze the trends in the metrics on a weekly basis. Also, found out the throughput on a daily basis.

It is preferable to calculate the throughput on a weekly basis as this gives us the clearer picture of the performance trends without getting into the complexity of analyzing trend on a daily basis.

We see that the weekly throughput for the data is **0.03.**

Task 3: Language Share Analysis

```sql
SELECT
    language,
    COUNT(job_id) AS jobs,
    ROUND(100 * COUNT(job_id) / (SELECT
                    COUNT(job_id)
                FROM
                    job_data),
            2) AS percentage_share
FROM
    job_data
GROUP BY language;
```

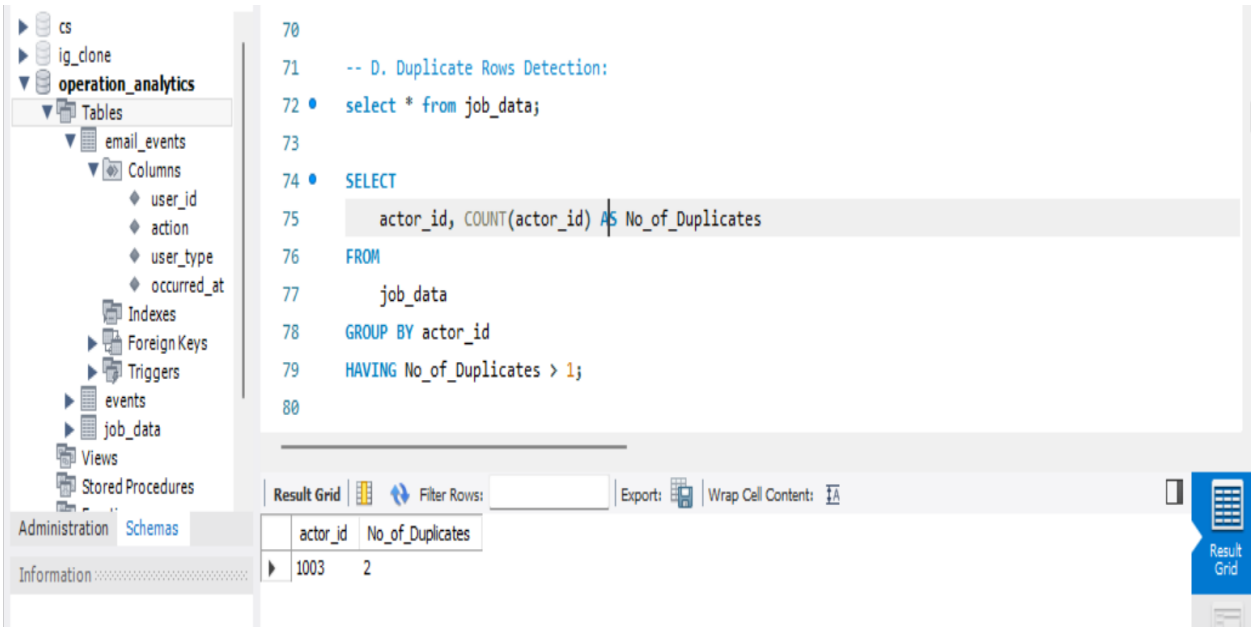
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

language	jobs	percentage_share
English	1	12.50
Arabic	1	12.50
Persian	3	37.50
Hindi	1	12.50
French	1	12.50
Italian	1	12.50

In this, we find out the percentage share of each language in the last 30 days.

Based upon the findings, we conclude that **Persian** is the most used language with a share of **37.50%.** Also, we see that all the other languages have the same amount of share which is **12.50%**. We can include more language specific content / features to increase the overall user engagement.

Task 4: Duplicate Rows Detection

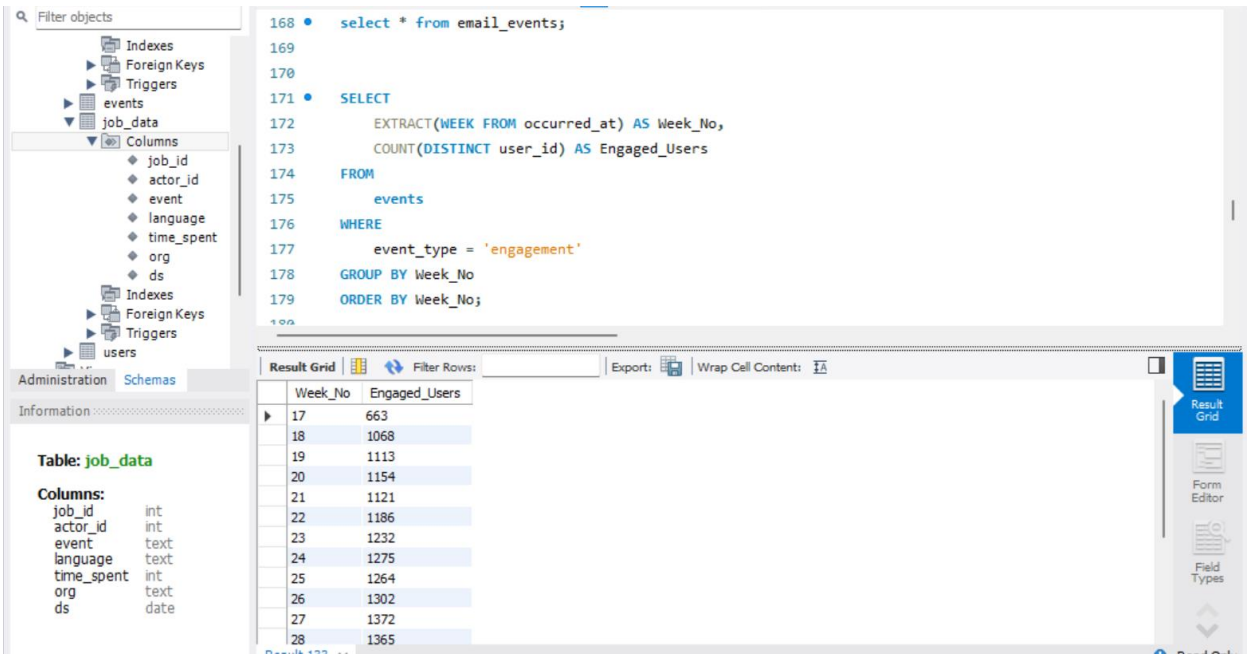


In this, we checked for any duplicate rows in the data. For this, we used **actor_id** column.

We see that there is one duplicate row present in the row. Data validation methods can be used here to ensure there are no duplicate data in future to get an accurate analysis and result.

CASE STUDY-2: Investigating Metric Spike

Task 1: Weekly User Engagement

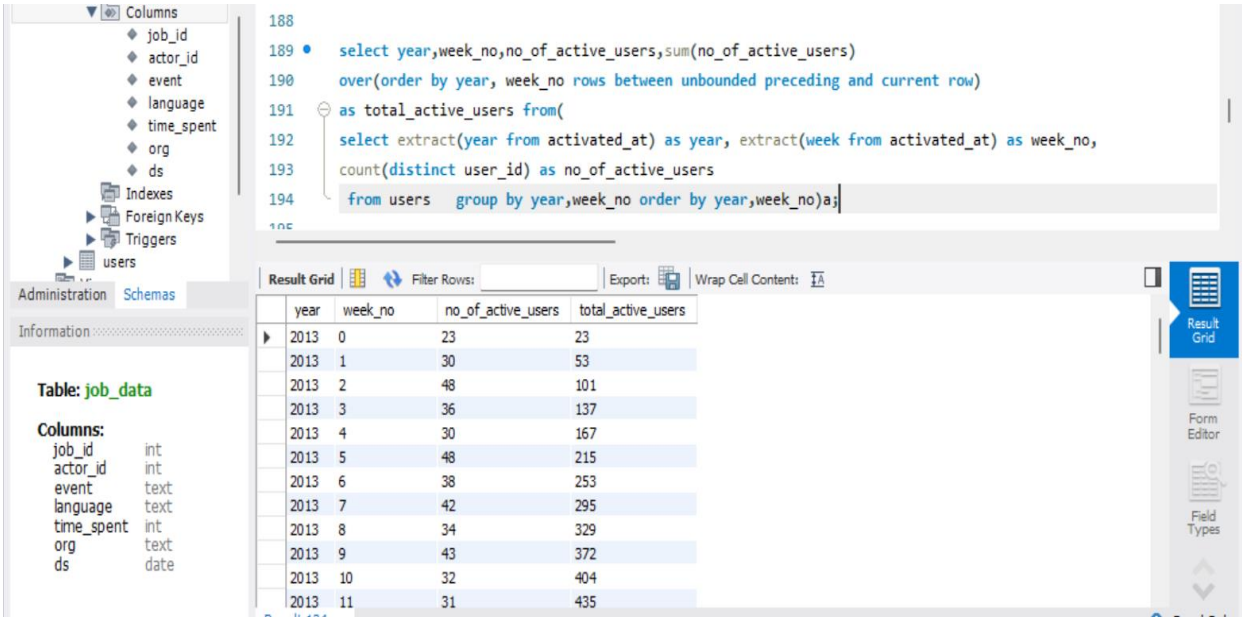


Here, we see the total number of users engaged per week which helps us understand the user's activeness on a weekly basis.

We see that for the **week no 30,** the user engagement was at its peak with a total count of **1467.**

We can check for the reason of this increase whether it is due to a marketing campaign, updates or any other events and then plan our future strategies to engage more users.

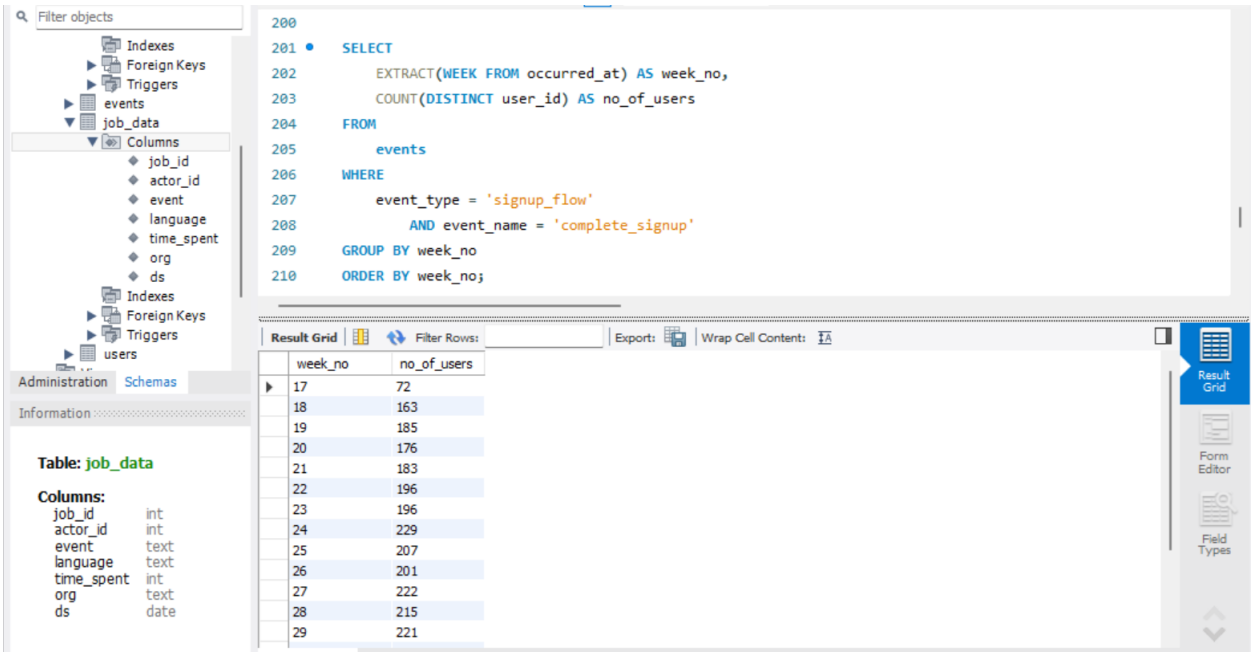
Task 2: User Growth Analysis



Here, we are analyzing the growth of users for a product over time.

We se that the user growth has been positive and the number of users has been increasing over time. We can check for various factors driving this goal whether it is a marketing campaign, any updates etc and replicate it for the overall growth of the company.

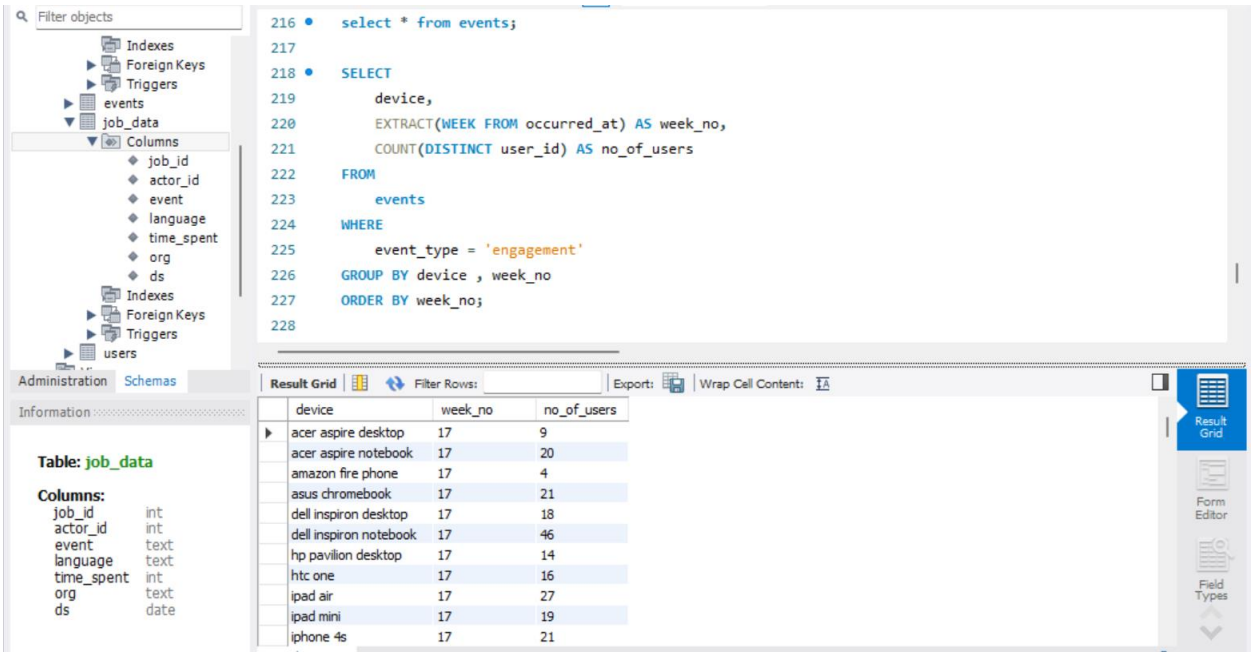
Task 3: Weekly Retention Analysis



Here, we are analyzing the weekly retention of users after signing up for a product.

We see from week **17** to **35**, a **rise** and **drop** in the retention of users. We can check for factors resulting in the drop and use retention strategies, check for points where the users are dropping off and improve user engagement and experience to have more users.

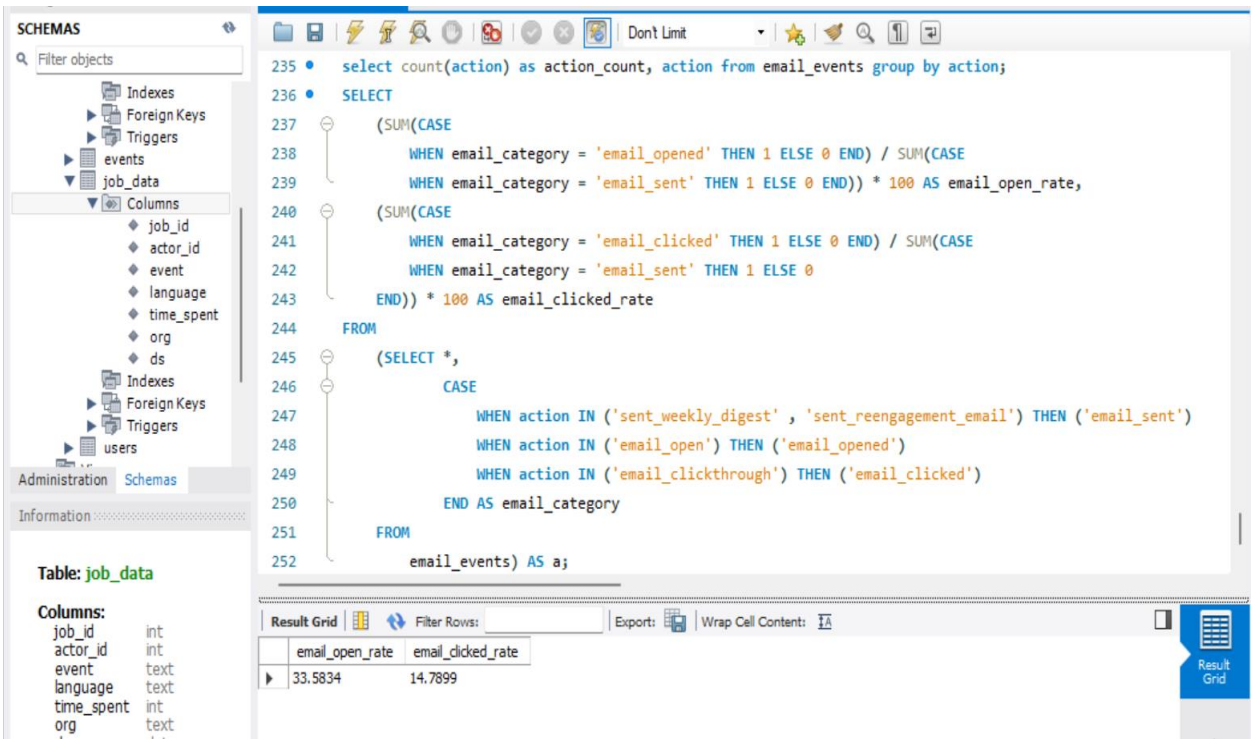
Task 4: Weekly Engagement Per Device



Here, we are checking for the activeness of users on a weekly basis per device.

We see that the user engagement varies across devices per week. Some devices have a higher engagement as compared to others. We can check for factors affecting the engagement and work more on device compatibility, optimizing user experience. Also, we need to monitor trends for such devices with lower user engagement and devise strategies to overcome the challenges that users are facing with the devices for more user engagement.

Task 5: Email Engagement Analysis



Here, we are analyzing how users are engaging with the email service and calculating the email engagement metrics.

In the table of **email_events**, we have **4** actions for emails which are: **sent_weekly_digest, email_open, email_clickthrough, sent_reengagement_email.**

We can use them to calculate the engagement metrics which are **email_open_rate** and **email_clicked_rate**.

After calculating, we see **email_open_rate = 33.5834 and email_clicked_rate = 14.7899.**

Now, we can check these metrics with the industry standard metrics and if found less, we can use more refined words in the email, change our subject line for the emails, targeting audience etc.

Atlast, regularly check and optimize our email content and campaigns to reach the industry standards.

Result:

While completing the project, I have gained knowledge on operational analytics, understanding its principle and methodologies and how important it is for a business or company to find a trend on the working and helping the business grow. Also, this project helped me on learning how to deal with large datasets and use aggregations, joins, subqueries to refine results to ensure accuracy, which gives us a correct and reliable analysis. In the project, I have worked with people from various departments and collaborated with them which enhanced my communication skills enabling me to share my findings and recommendations to the respective departments for data-driven decision making to improve and optimize overall user experience, user engagement, process engagement thus, contributing in the overall success of the project.