# Operation Analytics and Investigating Metric Spike

VICTOR SHAH

#### Project Description

- As a Data Analyst collaborating with the Microsoft Product team, my primary responsibility is about identifying areas for improvement within the company.
- The aspect of my role is to extract meaningful insights from the user data. These insights will offer actionable information for various teams within the business. •
- The Operational Analytics is a crucial process that involves analysing a company's end-to-end operations. One of the key aspects being Investigating metric spikes.
- The goal is to empower the product manager and the entire team with actionable insights that will shape the future development and user experience in the application.

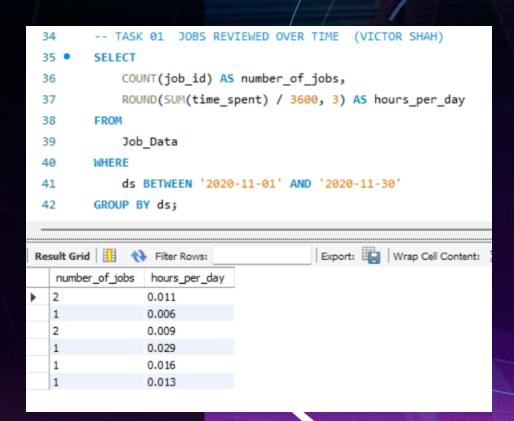
#### Approach

- Importing the Dataset: The first step is maintaining a file path to export data (.csv) into the SQL workbench.
- Understanding the Schema: The next step is to examine the structure of the table holding the data (
  job\_data, events, etc).
- Identifying the Key tables: Identification of the primary key from each of the tables of job\_data, email\_events, events, users etc.
- Checking for null values: Before the analysis, it is necessary to check for null values in the given tables
- Visually Appealing: The SQL Queries need to be properly formatted so that they can be understood by any user.

## CASE STUDY 1: Job Data Analysis

#### A. Jobs Reviewed Over Time:

Calculate the number of jobs reviewed per hour for each day in November 2020.



#### B. Throughput Analysis:

Calculate the 7-day rolling average of throughput

(number of events per second).

```
-- TASK 2: THROUGHPUT ANALYSIS (VICTOR SHAH)
45
       SELECT
       AVG(number of events) OVER(ROWS BETWEEN 6 PRECEDING AND CURRENT ROW) AS 7 day rolling avg
        FROM
       (SELECT
            COUNT(DISTINCT event) AS number of events
50
        FROM
51
           Job Data
       GROUP BY ds) AS sub;
                                          Export: Wrap Cell Content: TA
 7_day_rolling_avg
 1.0000
 1.0000
 1.0000
 1.2500
 1.2000
 1.3333
```

## C. Language Share Analysis:

Calculate the percentage share of each language

in the last 30 days.

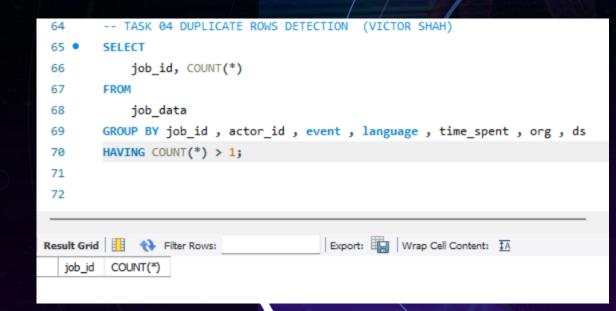
The language Persian has the most share out of the other languages in the past 30 days.

```
-- TASK 03 LANGUAGE SHARE ANALYSIS
54
                                                (VICTOR SHAH)
        SELECT language,
55 0
             ROUND((COUNT(language) / (SELECT COUNT(*) FROM job_data)) * 100,2)
56
57
                      AS language share
                 job_data
58
59
        WHERE
                  ds > (SELECT
60
                      MAX(ds) - INTERVAL 30 DAY
61
                                   job_data)
62
         GROUP BY language;
                                            Export: Wrap Cell Content: TA
Result Grid
              Filter Rows:
            language_share
   language
  English
            12.50
            12.50
  Arabic
            37.50
  Persian
            12.50
            12.50
  French
  Italian
            12.50
```

#### D. Duplicate Rows Detection:

Identify duplicate rows in the data.

	job_id	actor_id	event	language	time_spent	org	ds
•	21	1001	skip	English	15	Α	2020-11-30 00:00:00
	22	1006	transfer	Arabic	25	В	2020-11-30 00:00:00
	23	1003	decision	Persian	20	С	2020-11-29 00:00:00
	23	1005	transfer	Persian	22	D	2020-11-28 00:00:00
	25	1002	decision	Hindi	11	В	2020-11-28 00:00:00
	11	1007	decision	French	104	D	2020-11-27 00:00:00
	23	1004	skip	Persian	56	Α	2020-11-26 00:00:00
	20	1003	transfer	Italian	45	С	2020-11-25 00:00:00



The given table job\_data has no duplicate rows.

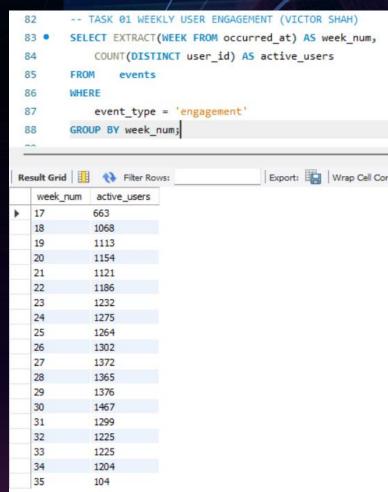
Hence the query returns an empty output.

## CASE STUDY 2: Investigating Metric Spike

#### A. Weekly User Engagement:

Measure the activeness of users on a weekly basis.

We can see that the user's engagement is measured upon events like login, home\_page, like\_message, view\_inbox.



#### B. User Growth Analysis:

Analyze the growth of users over time for a product.

The cumulative frequency is used to determine the number of observations that lie above a particular value in a dataset.

```
-- TASK 02 USER GROWTH ANALYSIS (VICTOR SHAH)
with tab1 as (select extract(year from created at) as years,
extract(month from created at) as months, count(*) as freq from users
group by years, months)
select years, months, sum(freq) over(order by years, months) as cum_freq,
freq as user growth from tab1;
                              Export: Wrap Cell Content: TA
        cum_freq user_growth
                  160
        320
        470
                 150
        651
                 181
                 214
                 213
        1078
        1362
        1678
        2398
        2797
                 399
        3283
        4360
        4975
        5701
                 726
        6480
                 779
        7353
        8350
                 1031
        9381
```

#### C. Weekly Retention Analysis:

Analyze the retention of users on a weekly basis after signing up for a product.

Week by week we can see the number of active users start to deteriorate.

```
-- TASK 03 WEEKLY RETENTION ANALYSIS (VICTOR SHAH)
        with retention as (
        select e.user id, extract(week from created at) as create week no,
100
        min(case when event type = 'engagement' then extract(week from occurred at) end) as login week
101
         from users u join events e on u.user id = e.user id
102
          group by e.user_id, create_week_no),
         week_retention as (
103
104
          select *, login week - create week no as weeks retained from retention order by weeks retained DESC)
105
          select weeks retained, count(user id) as no of users from week retention group by weeks retained order by weeks retained;
                                     Export: Wrap Cell Content: TA
                no_of_users
```

## D. Weekly Engagement Per Device:

Measure the activeness of users on a weekly basis per device.

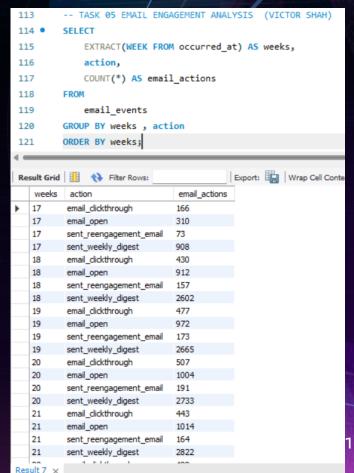
The users are differentiated weekly by the devices they use concurrently. The count of these devices are measured.

```
-- TASK 04 WEEKLY ENGAGEMENT PER DEVICE (VICTOR SHAH)
109
         select extract(week from occurred at) as weeks, device,
110
         count(distinct user_id) as occurrence from events where event_type = 'engagement'
111
         group by weeks, device order by weeks, device;
112
113
         -- TASK 05 EMAIL ENGAGEMENT ANALYSIS (VICTOR SHAH)
114
115 •
         SELECT
116
             EXTRACT(WEEK FROM occurred at) AS weeks,
          Filter Rows:
                                            Export: Wrap Cell Content: IA
                             occurrence
          acer aspire desktop
          hp pavilion desktop
          ipad air
          ipad mini
          iphone 4s
          iphone 5
          iphone 5s
          kindle fire
          lenovo thinkpad
          macbook air
          macbook pro
          nexus 10
          nexus 5
```

## E. Email Engagement Analysis:

Analyze how users are engaging with the email service.

We are able to figure out the events occurred in the email like email\_clickthrough, email\_open, sent\_weekly\_digest and many more.



#### Tech-Stack Used

- MySQL Workbench(8.0.34): This is the primary interactive development environment for SQL queries. It enables efficient query building, execution and debugging for data analysis
- SQL commands:
  - 1. DDL commands: These commands were used for the creation of the database and the multiple tables such as users, job\_data, email\_events, events.
  - DML commands: These commands were used for the insertion of the data into the records of the table.
  - DQL commands: The select query with where, order by, group by clauses helped for the further analysis of the data from the table.
- Window Functions: Does calculationn across a set of rows that are related to the current row. These functions are used when we want to calculate Average Running Price, Running Total Orders, Running Sum Sales, Rank and Percentile.

## Insights

- While analysing the tables we were able to figure the jobs reviewed per hour for each day in November 2020.
- The identification of duplicate rows from the table.
- We were able to understand how different users engage with email events in the application.
- The span of every language share in the last 30 days.
- To be able to analyze the retention of users on a weekly basis after signing up for a product...

#### Result

- Remembering to adapt these queries on specific database schemas.
- These learned insights helped me understand specific business questions which were addressed by SQL queries.
- Learning about the SQL clauses such as the join clauses and sub-queries. The importance of order by and group by and many more.
- We were able to import the dataset from a .CSV file into the SQL workbench for performing analysis.
- Achieving the ability to learn and write SQL queries to execute different business questions.
- Solving business related problems using Windows functions of SQL.

## Thankyou 18