Operation Analytics

22.07.2023

PRAJJWAL PANDEY

Case-1 JOB DATA

Project 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 in which it must improve upon.

The specific things that I have found in the analysis:

A. **Number of Jobs Reviewed:** Amount of jobs reviewed over time.

   My task - To calculate the number of jobs reviewed per hour per day for November 2020 ?

B. **Throughput:** It is no. of events happening per second.

   My task - Calculate a 7-day rolling average of throughput. For throughput, do you prefer daily metric or 7-day rolling and why?

C. **Throughput:** Share of each language for different contents.

   My task - To Calculate the percentage share of each language in the 30 days?

D. **Throughput:** Rows that have the same value present in them.

   My task - How will you display duplicates from the table?

Approach

The approach was to use SQL to query the Operation Analytics database. The following queries were used to answer the above questions:

- To find the number of jobs reviewed, the following query was used:

```
SELECT
  ds,
  hour,
  COUNT(*) AS num_jobs
FROM job_data
WHERE ds BETWEEN '2020-11-01' AND '2020-11-30'
GROUP BY ds, hour
ORDER BY ds, hour;
```

This query will return a table with the following columns:

ds: The date on which the jobs were reviewed.

Hour: The hour of the day on which the jobs were reviewed.

num_jobs: The number of jobs reviewed in the hour.

- The 7-day rolling average of throughput can be calculated using the following SQL query:

```
SELECT
  date,
  AVG(throughput) AS throughput_7d_avg
FROM (
  SELECT
    ds AS date,
    COUNT(*) AS throughput
  FROM job_data
  WHERE ds BETWEEN '2020-11-01' AND '2020-11-30
  GROUP BY ds
  ORDER BY ds
)
WINDOW win1 AS (
  PARTITION BY date
  ORDER BY ds
  ROWS BETWEEN 7 PRECEDING AND CURRENT ROW
);
```

● To find the percentage share of each language in last 30 days can be found using the following SQL query:

```sql
SELECT
  language,
  COUNT(*) AS num_jobs,
  ROUND(100 * COUNT(*) / SUM(COUNT(*)), 2) AS percentage
FROM job_data
WHERE ds BETWEEN '2020-10-31' AND '2020-11-30'
GROUP BY language
ORDER BY percentage DESC;
```

● The duplicate rows in the table can be displayed using, the following query was used:

```sql
SELECT
  *
FROM job_data
GROUP BY job_id, actor_id, event, language, time_spent, org, ds
HAVING COUNT(*) > 1;
```

Which metric do you prefer and why?

I prefer the 7-day rolling average of throughput over the daily metric. The 7-day rolling average is more stable than the daily metric, and it is less likely to be affected by outliers. This makes it a better measure of the overall trend in throughput.

Tech-Stack Used

The following tech stack was used to do this project:

- Database- MySQL
- Query Language- SQL
- Data Analysis Tool - Jupyter Notebook

RESULTS

The results of this project have provided valuable insights into Operation Analytics. This information can be used by management teams to make decisions about the future of the company.