**Case Study 1: Job Data Analysis**

**job\_data with the following columns:**

* **job\_id:**Unique identifier of jobs
* **actor\_id:**Unique identifier of actor
* **event:**The type of event (decision/skip/transfer).
* **language:**The Language of the content
* **time\_spent:**Time spent to review the job in seconds.
* **org:**The Organization of the actor
* **ds:**The date in the format yyyy/mm/dd (stored as text).

**Tasks:**

1. **Jobs Reviewed Over Time:**

(Calculating the number of jobs reviewed per hour for each day in November 2020)

SELECT

DATE\_TRUNC('hour', ds) AS review\_hour,

DATE\_TRUNC('day', ds) AS review\_date,

COUNT(\*) AS num\_jobs\_reviewed

FROM

job\_data

WHERE

DATE\_TRUNC('month', ds) = '2020-11-01'::DATE

GROUP BY

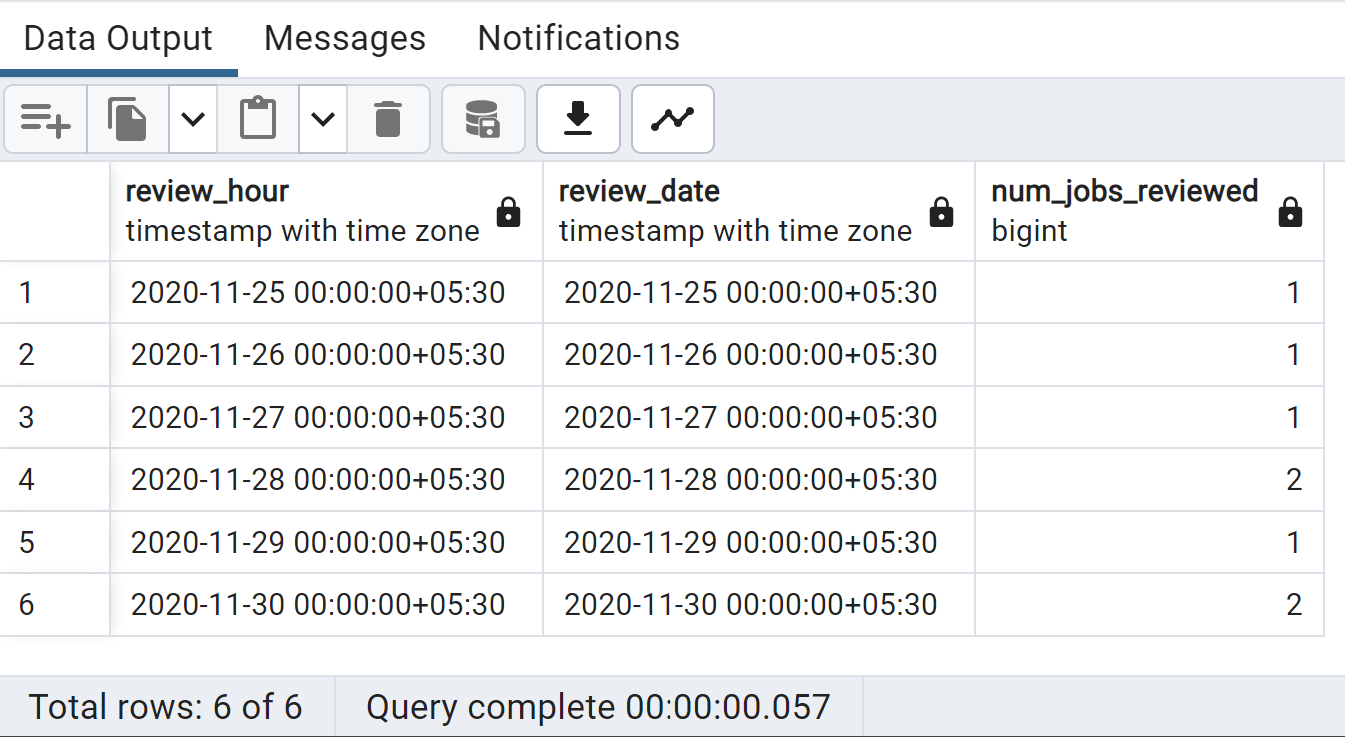
DATE\_TRUNC('hour', ds),

DATE\_TRUNC('day', ds)

ORDER BY

review\_date,

review\_hour;



**B.Throughput Analysis:**

Calculating the 7-day rolling average of throughput (number of events per second)

WITH DAILY\_METRIC AS (

SELECT

ds,

COUNT(job\_id) AS job\_review

FROM

job\_data

GROUP BY

ds

)

SELECT

ds,

job\_review,

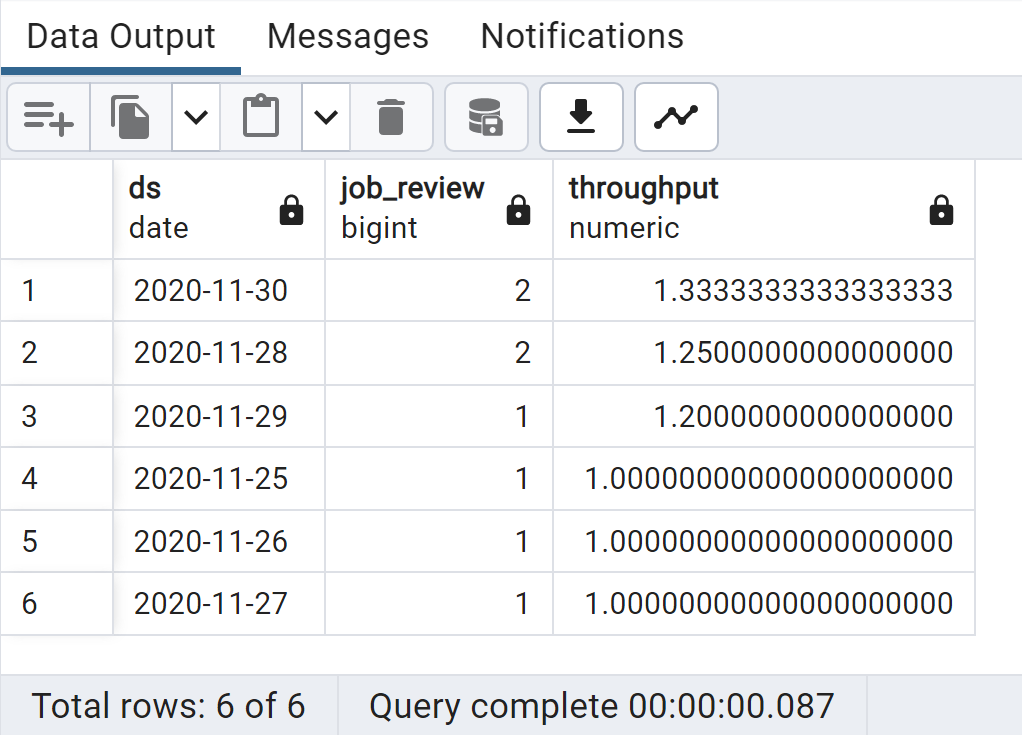
AVG(job\_review) OVER (ORDER BY ds ROWS BETWEEN 6 PRECEDING AND CURRENT ROW) AS throughput

FROM

DAILY\_METRIC

ORDER BY

throughput DESC;



**C.Language Share Analysis:**

 Calculating the percentage share of each language in the last 30 days.

SELECT

language,

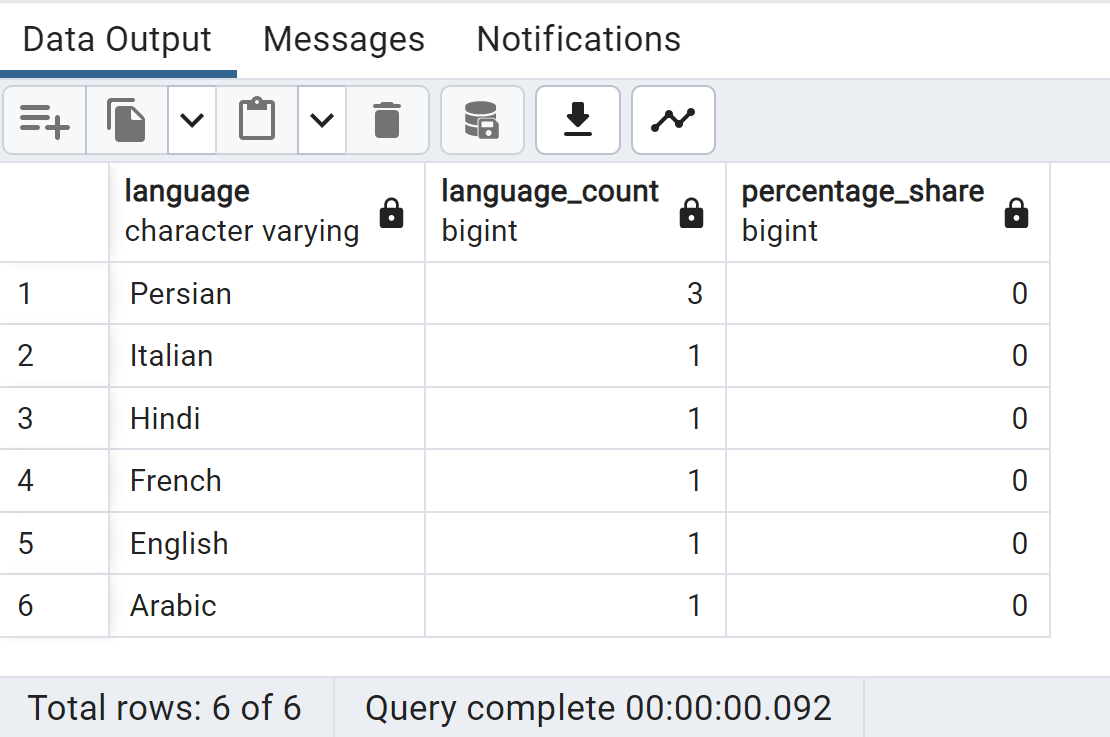
COUNT(language) AS language\_count,

(COUNT(language) / (SELECT COUNT(\*) FROM job\_data)) \* 100 AS percentage\_share

FROM job\_data

GROUP BY language

ORDER BY language DESC;



**D.Duplicate Rows Detection:**

 Identifying duplicate rows in the data.

SELECT \*

FROM (

SELECT \*,

ROW\_NUMBER() OVER (PARTITION BY JOB\_ID) AS DUPLICATE\_ROWS

FROM JOB\_DATA

) AS A\_R

WHERE DUPLICATE\_ROWS > 1;



© 2024 Paul M. Lazarescu. All rights reserved.

https://github.com/PMLAnalytics