EGA Backend Developer Challenges Part 2

Exercise 2

Approach and solution SQL.

Directly below is the sql (and a link to that sql online) that I created to solve the sql query on the table file_processing_events.

Planning Time: 0.649 ms
Execution Time: 200.227 ms (saw
365 rows.

Solution SQL

This latest version of this sql is available online at:

https://github.com/KenBuckley/imdb/blob/master/part_2/part2.sql

```
WITH daily_events AS (
    SELECT
        event_date,
        num_entries_per_day,
        total_bytes_processed,
        average_processing_time,
        file_list as files,
        SUM(total_bytes_processed) OVER (ORDER BY event_date) AS
total_bytes_cumulative
    FROM (
        SELECT
            date_trunc('day', event_time) AS event_date,
            COUNT(*) AS num_entries_per_day,
            SUM(bytes_processed) AS total_bytes_processed,
            AVG(processing_time_ms) as average_processing_time,
            array_agg(DISTINCT file_name) AS file_list
        FROM file_processing_events
        GROUP BY event_date
    ) as day_stats
    ORDER BY event_date
),
-- pre-calculate weights for the slowest entries
slowest_entries_weights AS (
```

```
    SELECT
        event_id,
        date_trunc('day', event_time) as event_date,
        (SELECT AVG(weight::numeric)
         FROM jsonb_array_elements_text(metadata->'weights') AS weight) AS
avg_weight,
        (SELECT SUM(weight::numeric)
         FROM jsonb_array_elements_text(metadata->'weights') AS weight) AS
sum_weight
    FROM file_processing_events
    WHERE metadata->>'source' = 'HTSGET'
    AND metadata ? 'weights'
    AND jsonb_typeof(metadata->'weights') = 'array' -- neccessary
),
 daily_tags as(
      SELECT
       date_trunc('day', event_time) AS event_date,
       array_agg(DISTINCT tag) FILTER (WHERE tag IS NOT NULL) AS all_tags
    FROM file_processing_events_1
    LEFT JOIN LATERAL unnest(tags) AS tag on true
    GROUP BY event_date
),
 slowest_daily_htsget as (
    select
        date_trunc('day', event_time) AS event_date,
            MIN(processing_time_ms) FILTER (WHERE metadata->>'source' =
'HTSGET') as min_time_ms,
          -- extra:get the event_id of the row with the
min(processing_time_ms).
          (array_agg(event_id ORDER BY processing_time_ms)
          FILTER (WHERE metadata->>'source' = 'HTSGET'))[1] as min_event_id
            FROM file_processing_events
        WHERE metadata->>'source' = 'HTSGET'
        AND metadata ? 'weights'   --todo check if required
        AND jsonb_typeof(metadata->'weights') = 'array'  --required or will
get error
        GROUP BY event_date
    )
SELECT
    de.event_date,
    de.num_entries_per_day,
    de.total_bytes_processed,
    de.average_processing_time,
```

```
    de.files,
    de.total_bytes_cumulative,
    dt.all_tags as tags,
    sew.avg_weight as average_weight_from_slowest_entry,
    sew.sum_weight as total_weight_from_slowest_entry
FROM daily_events de,slowest_entries_weights sew,slowest_daily_htsget sdh
,daily_tags dt
where sew.event_id = sdh.min_event_id
and dt.event_date= de.event_date
and de.event_date = sdh.event_date
ORDER BY de.event_date;
```

Questions

Q.Did you need to do any assumptions because of any ambiguity or any edge cases.

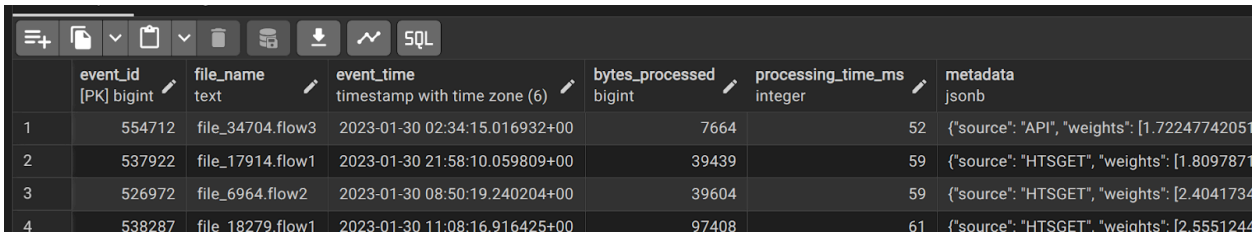
Yes,

1.There seems to be a problem with the aggregate counts -arising from where tag is null.
Ideally we would replace all the nulls with empty []

```
update file_processing_events
set tags ='{}'
where tags is null;
```

also some tags contain [Null] -which seems incorrect, these could also be removed in preprocessing.

2. there are a few records that have the same minimum processing_time_ms, for example in this day we have two rows with processing_time_ms = 59. So which one should we choose?



	event_id [PK] bigint	file_name text	event_time timestamp with time zone (6)	bytes_processed bigint	processing_time_ms integer	metadata jsonb
1	554712	file_34704.flow3	2023-01-30 02:34:15.016932+00	7664	52	{"source": "API", "weights": [1.72247742051
2	537922	file_17914.flow1	2023-01-30 21:58:10.059809+00	39439	59	{"source": "HTSGET", "weights": [1.809787
3	526972	file_6964.flow2	2023-01-30 08:50:19.240204+00	39604	59	{"source": "HTSGET", "weights": [2.4041734
4	538287	file_18279.flow1	2023-01-30 11:08:16.916425+00	97408	61	{"source": "HTSGET", "weights": [2.5551244

All the days with multiple processing times are:
"2023-01-30"
"2023-04-19"
"2023-07-23"
"2023-10-16"

As it is, the selection of the minimum record is not deterministic . We need more information to determine which row to select. Also the data will not be guaranteed deterministic [the results might change on the next database release] unless we modify the query to make it deterministic: like [but this is still just a guess, really we need more information on which row to select].
(array_agg(event_id ORDER BY processing_time_ms, event_id)
 FILTER (WHERE metadata->>'source' = 'HTSGET'))[1] AS min_event_id

3.There are some records with missing metadata (the field weights is missing and the following code must be inserted:
jsonb_typeof(metadata->'weights') = 'array

also we need to include "AND metadata ? 'weights'"
so not all of the metadata is complete.

4 The date 2023-05-13 does not have an entry for HTSGET -all the sources are null for that day,
so there was some error in the pipeline. This is the code you can use to detect it.
detect days with no source of HTSGET:

```
SELECT DISTINCT CAST(event_time AS DATE) AS calendar_day
FROM public.file_processing_events
WHERE CAST(event_time AS DATE) NOT IN (
    SELECT DISTINCT CAST(event_time AS DATE)
    FROM public.file_processing_events
    WHERE metadata->>'source' = 'HTSGET'
)
```

```
ORDER BY calendar_day;
```

There is no clear technical solution here, we must manually intervene and remove the day or manually insert data.

Q. Did I encounter any errors, and how did I solve them

The errors I encountered were the bad data like " the date 2023-05-13 does not have an entry for HTSGET", I cannot solve this without directions.
There are obviously some records with missing metadata, so we have to do jsonb_typeof(metadata->'weights') = 'array

Q.If we were going to create a graph on the website with the event_date, how would we modify something?

We get 365 days of results which is quite a lot of data points to plot on the x-axis.
Also the data is spread over one calendar year 2023 to 2024.01.01 so we could aggregate the data up to months and show a point per month (12 data points obviously) -or we could go with weeks, it would be best to just export the data as a csv and plot the graph in excel and make a judgement from there.I would just make it a 2d graph in excel to keep it quick.

Q.would you create indexes on the table? Which ones and why.

Yes indexes always help in case of scans. An explain analyse will show the best areas to improve. Looking at my query my immediate thought would be to index the event_time,

This will not help our case here immediately but will help in paging and seeking.

CREATE INDEX idx_event ON file_processing_events (event_time);

Also due to the jsonb we should add:
 CREATE INDEX idx_metadata_gin ON file_processing_events USING gin (metadata);
(small decrease in runtime not noticeable with the data)

```
CREATE INDEX idx_file_events_tags_gin ON public.file_processing_events
USING gin (tags);
```

maybe - I tested and it did not improve things by much try:

```
CREATE INDEX idx_file_events_covering ON public.file_processing_events
USING btree (event_time)
INCLUDE (bytes_processed, processing_time_ms, file_name, tags, metadata);
```

If we could use materialized views it would be great (I know that this is not always possible if data is changing rapidly)
Perhaps use materialized view of the tablepublic.file_processing_events to add extra columns to the materialized view:
materialized view + ADD COLUMN event_day date (and then we can replace date_trunc('day', event_time) in the code);
materialized view + ADD COLUMN source_text text (and then we can replace (metadata->>'source') in the code.
)
now you can index on those two fields in the materialized view and change the query to work directly on event_day instead of date_trunc('day', event_time) and directly on the new field source_text instead of (metadata->>'source')

I would also consider a materialised view for the slowest_entries_weights table, just to avoid having to calculate this data all the time.

I ran out of time testing various indexes so I am going to finish the document here.