Distributed execution*

Distributed Data Processing Environments

Lab Guide 6

This session aims at using a distributed cloud-based SQL tool to run data processing tasks. It uses the Free Tier[1] and should not consume credits.

Steps

1. Explore existing data:
   - Go to BigQuery > Studio.
   - Select "Add Data" in the Explorer pane, then "Star a project by name" and type *bigquery-public-data*.
   - Select *bigquery-public-data* from the Explorer and filter/navigate to datasets *thelook_ecommerce*.
   - View Schema, Details, and Preview for each table.
2. Execute a query and inspect results:
   - Try query:

```
SELECT p.name, COUNT(*) AS c
FROM `bigquery-public-data.thelook_ecommerce.order_items` oi
JOIN `bigquery-public-data.thelook_ecommerce.products` p
ON oi.product_id = p.id
GROUP BY p.id
ORDER BY c
LIMIT 10
```

   - View Execution details and Execution graph panes. Explain the purpose of each operator.
3. Formulate additional queries for:
   - Who are the top spenders?
   - What is the min/average/max order delivery time for each product category?
   - And for each destination country?

Learning Outcomes Describe the purpose of each relational operator in a query plan. Identify factors that impact the choice of query plan.

*Use of AI tools is encouraged in step 3.

[1] https://cloud.google.com/free/docs/free-cloud-features#free-tier