TPC-DS query logs

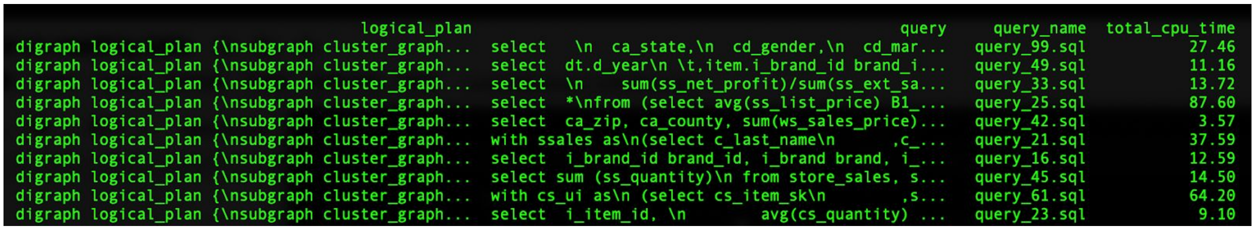
What to expect

- We are releasing **5,586 TPC-DS hive query** logs. All logs are obtained from a total of **83** publicly available TPC-DS templates.

- Each entry is a successfully executed query to our Presto engine, in which we record down the total CPU time taken for the execution and the resultant query plan. This timing is normalized using min-max normalization.

- A snippet of the Pandas.DataFrame is as such:

Columns	Description
logical_plan	Logical plan after running Presto command explain (format graphviz) <query>
query	Raw query string
query_name	TPC-DS query template name
total_cpu_time	Recorded total CPU timing in minutes



```
                                    logical_plan                                              query     query_name  total_cpu_time
digraph logical_plan {\nsubgraph cluster_graph...  select   \n  ca_state,\n  cd_gender,\n  cd_mar...  query_99.sql           27.46
digraph logical_plan {\nsubgraph cluster_graph...  select  dt.d_year\n \t,item.i_brand_id brand_i...  query_49.sql           11.16
digraph logical_plan {\nsubgraph cluster_graph...  select  \n    sum(ss_net_profit)/sum(ss_ext_sa...  query_33.sql           13.72
digraph logical_plan {\nsubgraph cluster_graph...  select  *\nfrom (select avg(ss_list_price) B1_...  query_25.sql           87.60
digraph logical_plan {\nsubgraph cluster_graph...  select  ca_zip, ca_county, sum(ws_sales_price)...  query_42.sql            3.57
digraph logical_plan {\nsubgraph cluster_graph...  with ssales as\n(select c_last_name\n      ,c_...  query_21.sql           37.59
digraph logical_plan {\nsubgraph cluster_graph...  select  i_brand_id brand_id, i_brand brand, i_...  query_16.sql           12.59
digraph logical_plan {\nsubgraph cluster_graph...  select sum (ss_quantity)\n from store_sales, s...  query_45.sql           14.50
digraph logical_plan {\nsubgraph cluster_graph...  with cs_ui as\n (select cs_item_sk\n        ,s...  query_61.sql           64.20
digraph logical_plan {\nsubgraph cluster_graph...  select  i_item_id, \n        avg(cs_quantity) ...  query_23.sql            9.10
```

Tutorial on how to extract the data

Step 1: Reading a CSV to Pandas.Dataframe object

```
# Unzip dependencies classes to folder
unzip tpc-ds-plans.csv.zip -d .

# Spawn Python REPL
import pandas as pd

df = pd.read_csv("./tpc-ds-plans.csv")
```