

Metatron Discovery



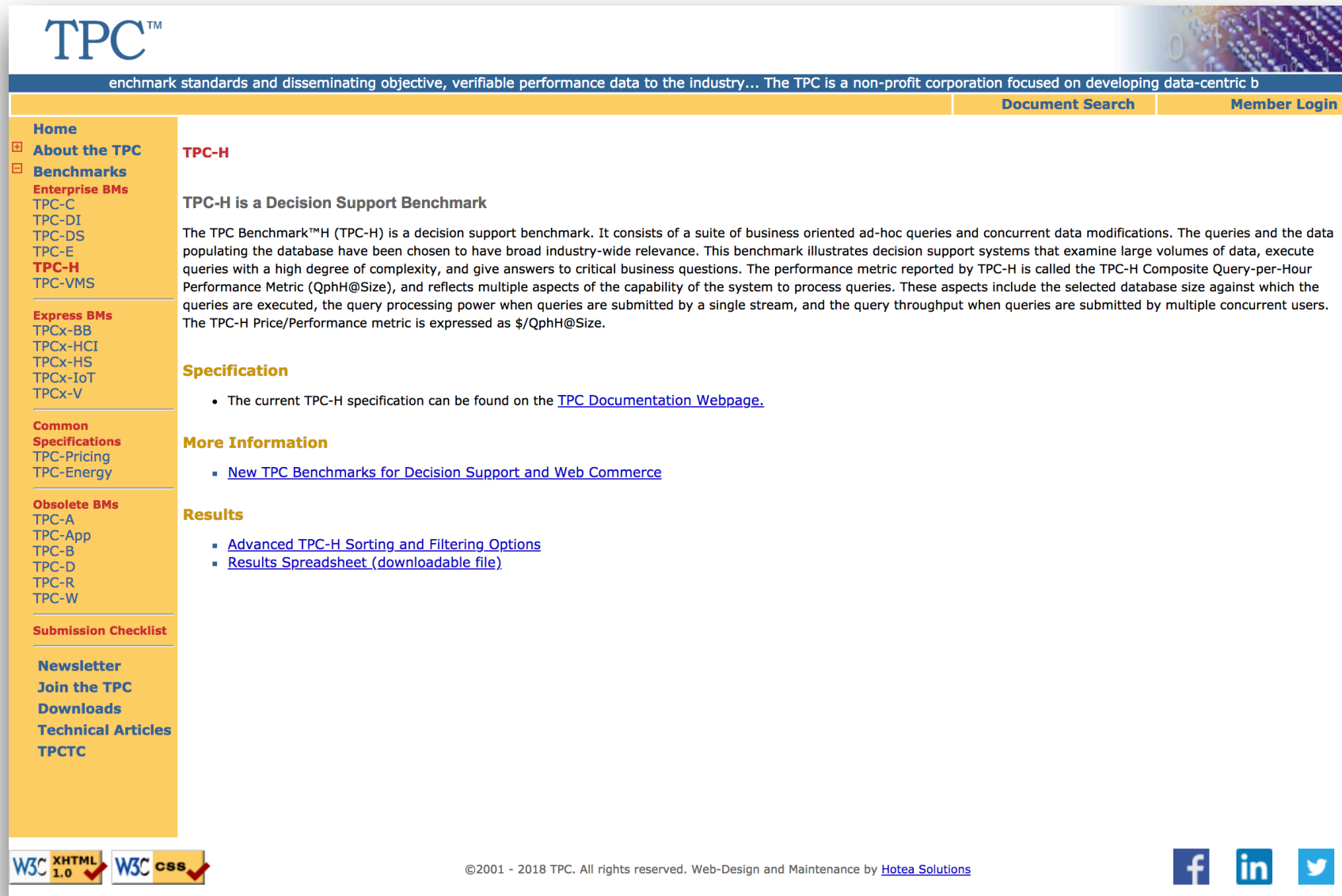
Benchmark Data

>>> Using TPC-H Benchmark



TPC-H?

Data processing engine들을 benchmark 할때 많이 사용하는 데이터



The screenshot shows the TPC-H website. The header includes the TPC logo and a navigation bar with links for Document Search and Member Login. The main content area is titled "TPC-H" and describes it as a Decision Support Benchmark. It explains that the benchmark consists of a suite of business-oriented ad-hoc queries and concurrent data modifications. The performance metric reported is the TPC-H Composite Query-per-Hour Performance Metric (QphH@Size). The website also provides links to the TPC-H specification, more information, and results. The footer includes a copyright notice for 2001-2018 TPC, a link to the website design and maintenance by Hotea Solutions, and social media icons for Facebook, LinkedIn, and Twitter.

TPC-H

TPC-H is a Decision Support Benchmark

The TPC Benchmark™H (TPC-H) is a decision support benchmark. It consists of a suite of business oriented ad-hoc queries and concurrent data modifications. The queries and the data populating the database have been chosen to have broad industry-wide relevance. This benchmark illustrates decision support systems that examine large volumes of data, execute queries with a high degree of complexity, and give answers to critical business questions. The performance metric reported by TPC-H is called the TPC-H Composite Query-per-Hour Performance Metric (QphH@Size), and reflects multiple aspects of the capability of the system to process queries. These aspects include the selected database size against which the queries are executed, the query processing power when queries are submitted by a single stream, and the query throughput when queries are submitted by multiple concurrent users. The TPC-H Price/Performance metric is expressed as \$/QphH@Size.

Specification

- The current TPC-H specification can be found on the [TPC Documentation Webpage](#).

More Information

- [New TPC Benchmarks for Decision Support and Web Commerce](#)

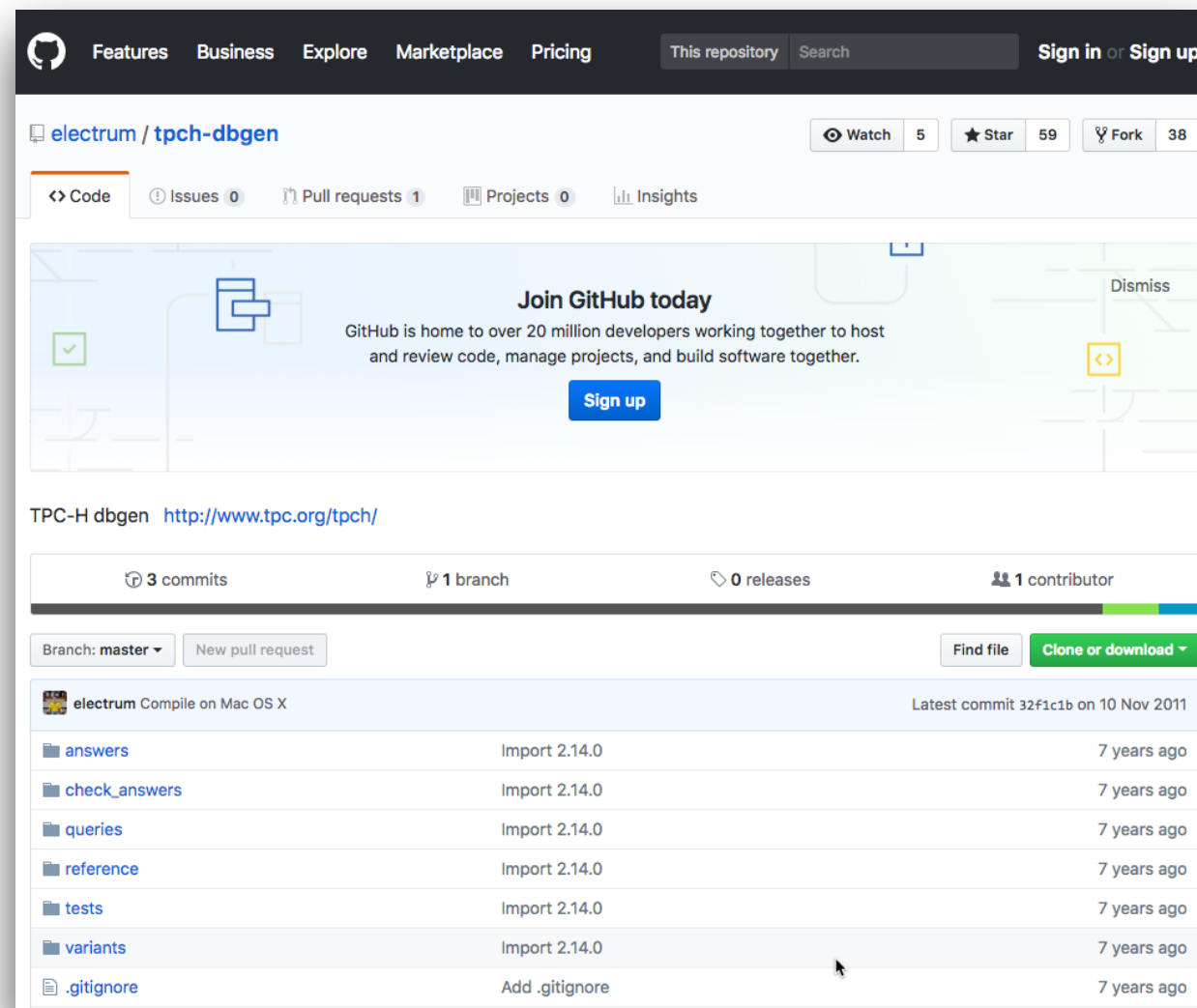
Results

- [Advanced TPC-H Sorting and Filtering Options](#)
- [Results Spreadsheet \(downloadable file\)](#)

메타트론을 다른 툴과 비교하기 위해 TPC 테스트를 해보자!

Data Generator

데이터는 고정적이지 않으며 필요한 만큼 만들어서 사용해야 하기 때문에 data generator를 제공함



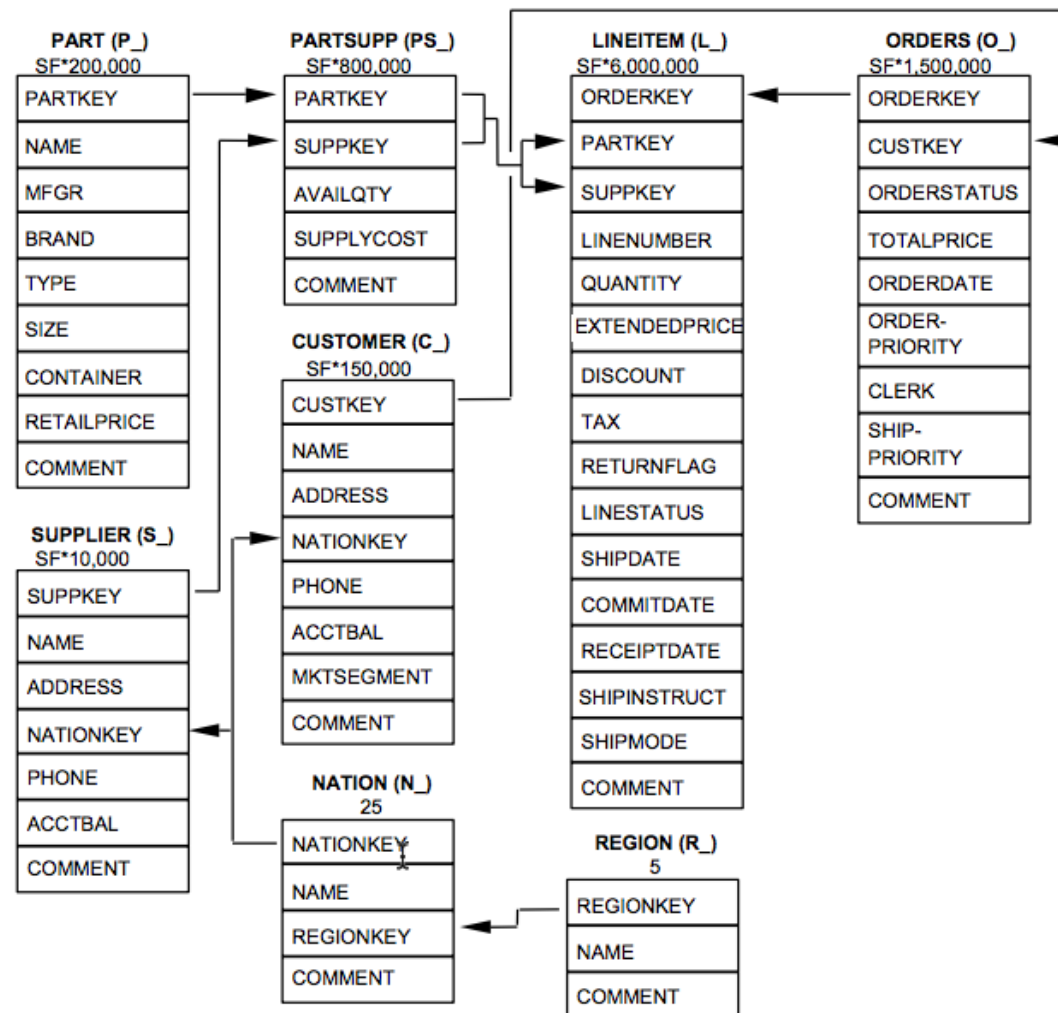
Makefile에서 DATABASE, MACHINE, WORKLOAD 설정하고 make 하여 ./dbgen 생성



LINEITEM 테이블 생성

여러 테이블 중 가장 큰 lineitem 선택하여 추가

Figure 2: The TPC-H Schema



-s : scale factor : 데이터의 크기를 의미
 -T : target data : 생성할 데이터 종류
 -v : verbose

`./dbgen -s 10 -T L -v` 입력하여 10의 크기로 lineitem.tbl 생성

Hive Table 생성

디렉토리 만들고 메타트론 워크벤치에서 테이블 생성 질의

Hdfs 디렉토리 생성

```
hadoop dfs -mkdir /sample/tpch_10
hadoop dfs -mkdir /sample/tpch_10/lineitem
hadoop dfs -put lineitem.tbl.1 /sample/tpch_10/lineitem
```

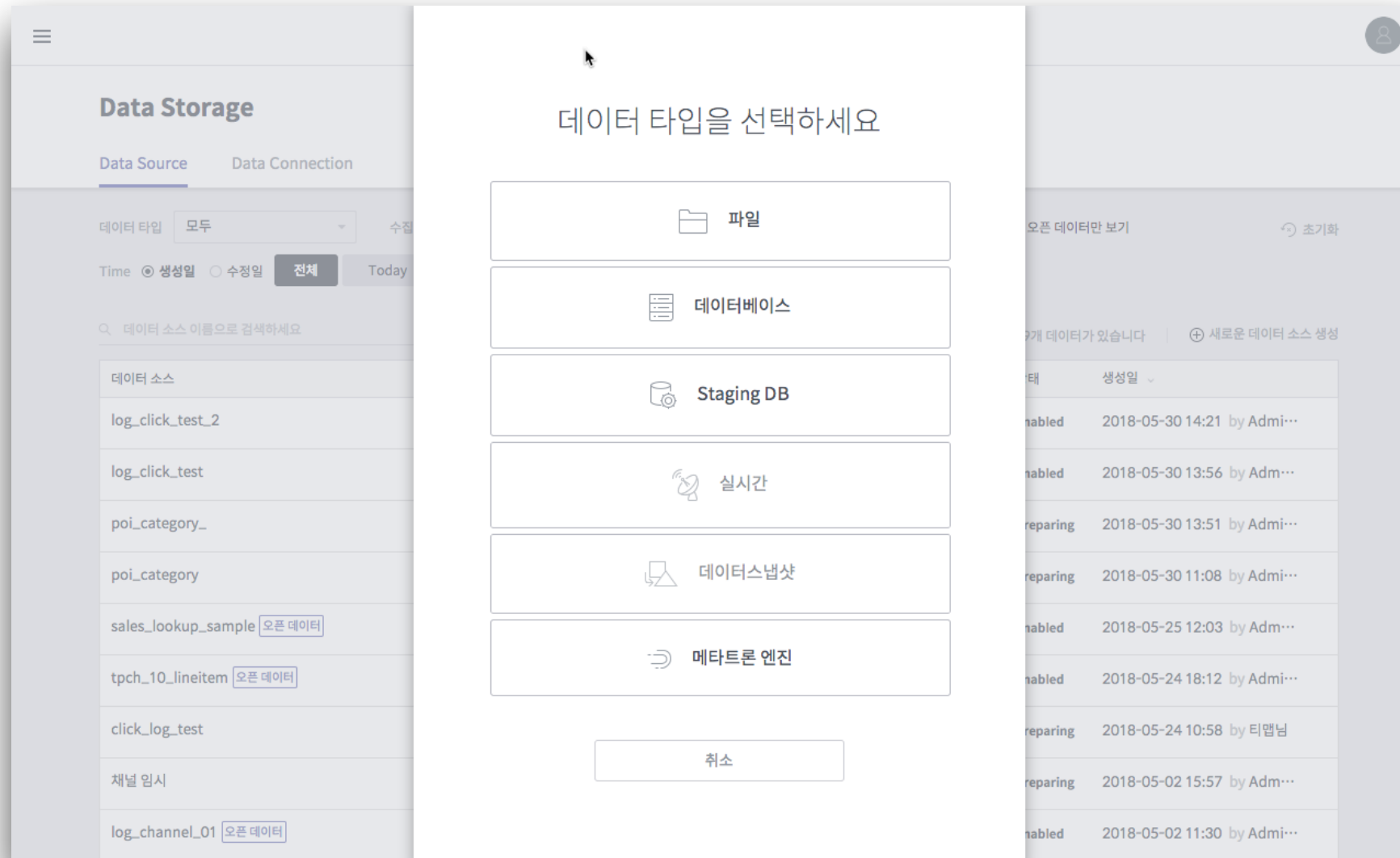
Hive table 생성

The screenshot shows the Metron Workbench interface. On the left, a sidebar displays a tree view with 'tpch_10' and 'lineitem'. The main area shows a SQL query for creating the 'tpch_10.lineitem' table. The query is as follows:

```
1 CREATE DATABASE tpch_10;
2
3
4
5 DROP TABLE tpch_10.lineitem;
6
7
8 CREATE EXTERNAL TABLE tpch_10.lineitem
9 (L_ORDERKEY BIGINT,
10 L_PARTKEY BIGINT,
11 L_SUPPKEY BIGINT,
12 L_LINENUMBER INT,
13 L_QUANTITY DOUBLE,
14 L_EXTENDEDPRICE DOUBLE,
15 L_DISCOUNT DOUBLE,
16 L_TAX DOUBLE,
17 L_RETURNFLAG STRING,
18 L_LINestatus STRING,
19 L_SHIPDATE STRING,
20 L_COMMITDATE STRING,
21 L_RECEIPTDATE STRING,
22 L_SHIPINSTRUCT STRING,
23 L_SHIPMODE STRING,
24 L_COMMENT STRING)
25 row format delimited fields terminated by '|' stored as textfile
26 location '/sample/tpch_10/lineitem';
27
28
29
30
31 SELECT * FROM tpch_10.lineitem;
```

Staging DB에서 데이터 적재

메타트론에서 데이터를 사용하기 위해 데이터 소스 생성



이후 과정은 druid ingestion 과정에서 다룸