Cairo University Law and Consultation

Faculty of Engineering

Benchmark Assignment

Benchmark name: TPC-H Benchmark

Link:

website: https://www.tpc.org/tpch/

Document: https://www.tpc.org/TPC_Documents_Current_Versions/pdf/TPC-H_v3.0.1.pdf

Name:

احمد هانى فاروق توفيق

Ahmed Hany Farouk Tawfik

Mail:

ahmed.tawfik011@eng-st.cu.edu.eg

1. Preface

The TPC-H Benchmark is managed by the Transaction Processing Performance Council (TPC), a non-profit organization dedicated to developing industry-standard benchmarks for evaluating the performance of transaction processing and database systems. The TPC-H Benchmark specifically focuses on decision support systems, measuring the performance of database management systems (DBMS) in data warehousing scenarios. My choice, which is TPC-H, is a benchmark for TPC-H evaluates the performance of various decision support systems by the execution of sets of queries against a standard database under controlled conditions. Moreover, the TPC-H Benchmark is designed to be universally acceptable. To achieve this, all companies participating in the benchmarking process must adhere to a set of agreed-upon rules and conditions. This ensures a level playing field where different organizations can confidently compare the performance of their systems without concerns about biases or unfair advantages.

2. Introduction

The TPC-H benchmark, conducted by the Transaction Processing Performance Council (TPC), a non-profit organization renowned for defining transaction processing and database benchmarks, serves as a standardized performance measurement tool. This decision support benchmark comprises business-oriented ad-hoc queries and concurrent data modifications chosen for broad industry relevance. The TPC-H benchmark exemplifies decision support systems capable of handling large volumes of data, executing complex queries, and providing critical business insights.

Originating in the 1980s with a focus on OLTP systems, the TPC evolved to address the surge in Decision Support Systems (DSS) popularity in the 1990s. To adapt to advancements in technology and maintain relevance, the TPC introduced V3.0 of the TPC-H Benchmark. This version continues to be actively used, reflecting the TPC-H benchmark's ongoing applicability and non-obsolescence.

The TPC-H benchmark evaluates performance through the Composite Query-per-Hour Performance Metric (QphH@Size). This metric considers database size, query processing power for single-stream submissions, and query throughput for multiple concurrent users. The benchmark's purpose is paramount — it provides a reliable, industry-standard method for evaluating and comparing the performance of various Database Management Systems (DBMSs) and hardware architectures in a standardized manner.

3. Benchmark Methodology

Benchmark Execution:

TPC-H employs crafted ad-hoc queries and concurrent data modifications to replicate real-world decision support workloads. The benchmark assesses system performance by measuring the number of queries executed per hour (QphH).

Query Structure and Database Schema:

TPC-H queries are intricate and business-oriented, tailored to the demands of decision support systems. The benchmark utilizes a predefined and standardized database schema, ensuring consistency across diverse systems.

Algorithmic Overview:

1. Database Population: The first step in the TPC-H benchmark involves populating the database with a synthetic dataset generated by the DBGEN data generation program.

2. Query Execution: The benchmark consists of executing a suite of twenty-two (22) complex queries against the populated database. These queries are designed to represent real-world business questions.

3. Data Modification: In addition to the queries, the benchmark includes two concurrent data modification operations, or "refresh functions," which insert and delete data from the database.

4. Performance Measurement: The performance of a system under the TPC-H benchmark is measured using the TPC-H Composite Query-per-Hour Performance Metric (QphH@Size). This metric reflects multiple aspects of the system's ability to process queries, including 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.

5. Result Reporting: The results of the benchmark are reported in terms of the QphH@Size metric and the TPC-H Price/Performance metric. The latter is expressed as Price/QphH@Size for Version 2 and Price/kQphH@Size for Version 31.

Hardware Configuration:

TPC-H avoids specifying a fixed hardware configuration but mandates public disclosure. Transparency in hardware specifications, including processors, memory, storage, and network architecture, ensures interpretable and comparable benchmark results.

Execution Workflow:

TPC-H queries are executed sequentially, simulating a decision support workload. The benchmark measures response time for each query and calculates throughput in queries per hour. Concurrent data modifications, simulating updates, assess the system's capability for simultaneous read and write operations.

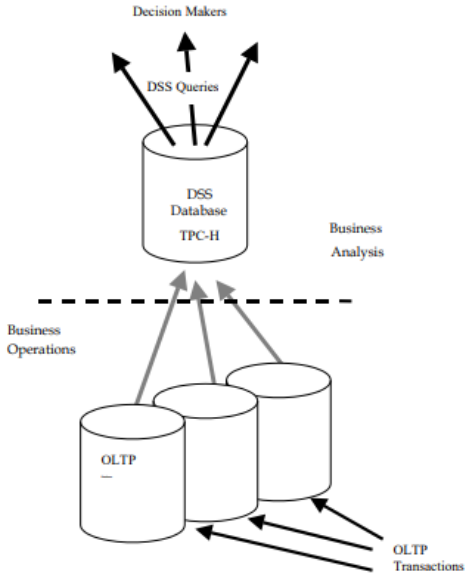
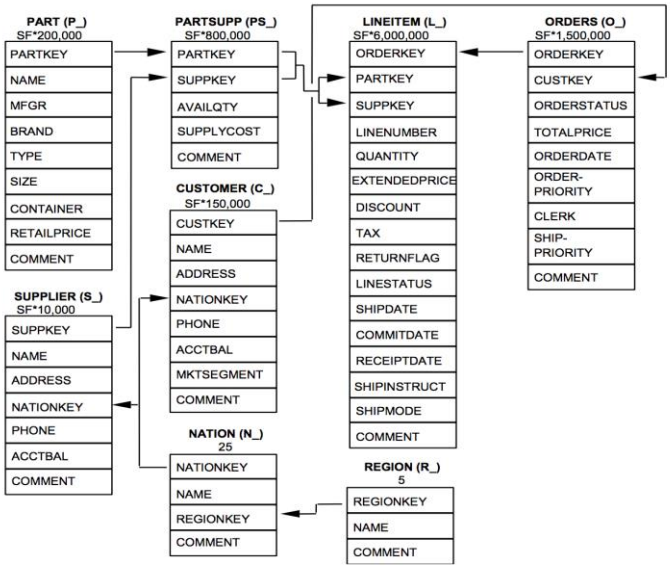


Figure 2: The TPC-H Schema



Legend:

- The parentheses following each table name contain the prefix of the column names for that table;
- The arrows point in the direction of the one-to-many relationships between tables;
- The number/formula below each table name represents the cardinality (number of rows) of the table. Some are factored by SF, the Scale Factor, to obtain the chosen database size. The cardinality for the LINEITEM table is approximate (see Clause 4.2.5).

4. SAMPLE RESULTS

100 GB Results

KTNF KR580S1: QphH = 43,903, Price/kQphH = 1,567,768.70 KRW, System Availability = 02/16/21, Database = Altibase 7.1, Operating System = Red Hat Enterprise Linux 7.9

1,000 GB Results

HPE ProLiant DL385 Gen11: QphH = 1,156,627, Price/kQphH = 265.09 USD, System Availability = 12/05/22, Database = Microsoft SQL Server 2022 Enterprise Edition 64-bit, Operating System = Microsoft Windows Server 2022 Datacenter Edition

Dell PowerEdge R7515: QphH = 979,335, Price/kQphH = 269.23 USD, System Availability = 05/03/21, Database = Microsoft SQL Server 2019 Enterprise Edition 64-bit, Operating System = Red Hat Enterprise Linux 8

3,000 GB Results

HPE ProLiant DL385 Gen11: QphH = 2,405,162, Price/kQphH = 490.02 USD, System Availability = 04/03/23, Database = Microsoft SQL Server 2022 Enterprise Edition 64-bit, Operating System = Microsoft Windows Server 2022 Datacenter Edition

Dell PowerEdge R7525: QphH = 1,542,560, Price/kQphH = 327.38 USD, System Availability = 07/05/21, Database = Microsoft SQL Server 2019 Enterprise Edition 64-bit, Operating System = Red Hat Enterprise Linux Server Release 8.3

10,000 GB Results

Dell PowerEdge R6525: QphH = 22,756,594, Price/kQphH = 68.79 USD, System Availability = 07/01/21, Database = EXASOL 7.1, Operating System = Ubuntu 20.04.2 LTS

HPE ProLiant DL380 Gen11: QphH = 2,028,444, Price/kQphH = 821.80 USD, System Availability = 05/01/23, Database = Microsoft SQL Server 2022 Enterprise Edition 64-bit, Operating System = Microsoft Windows Server 2022 Standard Edition

5. Conclusion

Importance:

1. **Standardized Testing**: TPC-H provides a standardized testing framework that allows for fair comparisons between different systems. This is crucial for businesses and researchers who need to evaluate the performance of various DBMSs and hardware configurations.
2. **Complex Queries**: The benchmark includes a suite of complex queries and concurrent modifications of a database. The queries are designed to answer critical business questions and are representative of real-world business data processing and analytics.
3. **Scalability**: TPC-H supports different scales of test data, making it possible to evaluate how well a system performs as the volume of data increases. This is particularly important in the era of big data, where systems must be able to manage increasingly large datasets.
4. **Comprehensive Metrics**: TPC-H measures performance in terms of both query processing speed and price-performance ratio. This allows businesses to consider both the raw performance and the cost-effectiveness of their systems.
5. **Industry Recognition**: TPC-H is widely recognized in the industry, and the results of TPC-H tests are often used in marketing materials. Therefore, achieving a good result on the TPC-H benchmark can be seen as a mark of quality.

As for future research:

1. **Benchmarking New Technologies**: As modern technologies and architectures emerge, it will be important to adapt the TPC-H benchmark to these new contexts. For example, how does the performance of traditional DBMSs compare with new NoSQL databases or in-memory databases?
2. **Real-Time Analytics**: With the rise of real-time analytics, it would be interesting to see how the TPC-H benchmark could be adapted to measure the performance of systems under real-time constraints.
3. **Big Data**: The TPC-H benchmark could be extended to better handle big data scenarios. This could involve increasing the scale of the test data or incorporating more complex queries that are representative of big data analytics.
4. **Machine Learning**: With the increasing integration of machine learning into data processing pipelines, future research could look at incorporating machine learning tasks into the TPC-H benchmark.