UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Data WareHouse Case Study

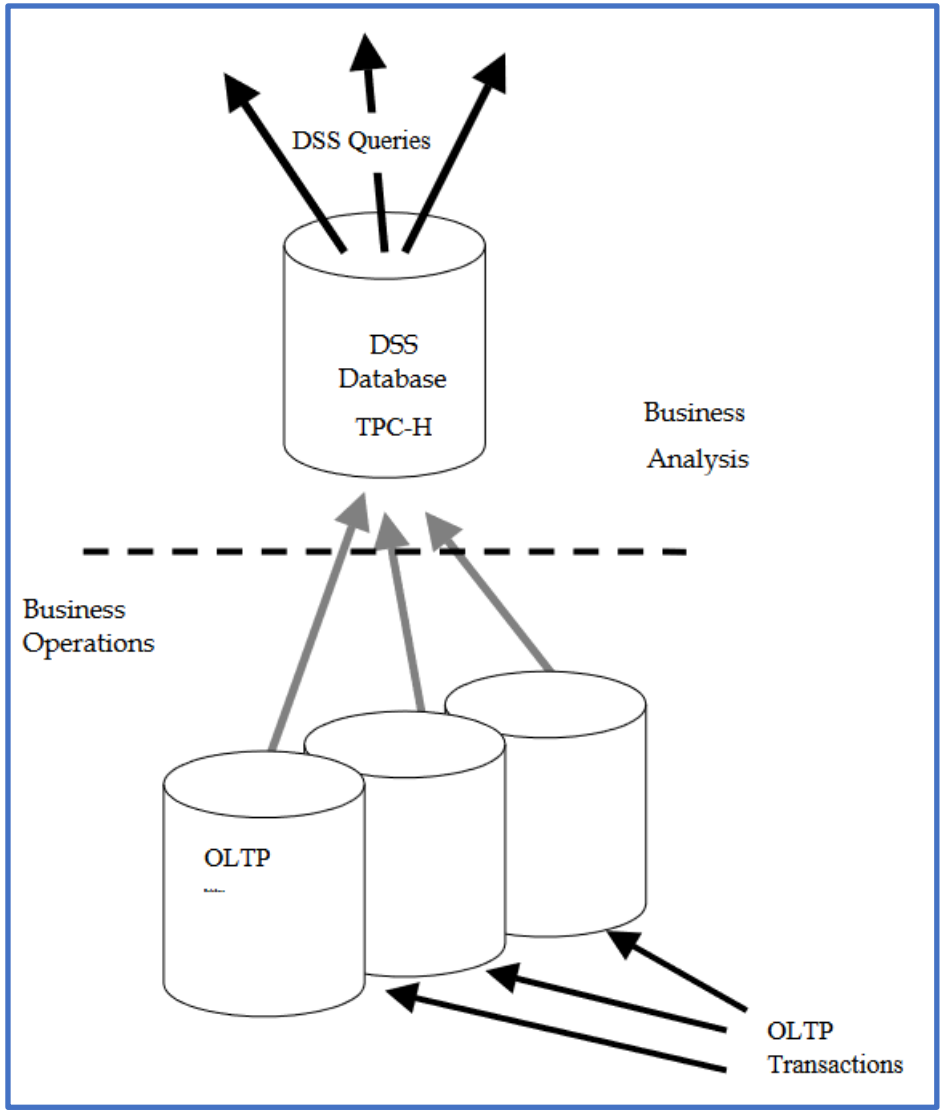
Prof. A. Peron

TPC-H

- A benchmark for decision support.
- www.tpc.org
- 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 while maintaining a sufficient degree of ease of implementation.
- This benchmark illustrates decision support systems that:
  - Examine large volumes of data;
  - Execute queries with a high degree of complexity;
  - Give answers to critical business questions.

TPC-H

- TPC Benchmark™ H is comprised of a set of business queries designed to exercise system functionalities in a manner representative of complex business analysis applications.
- These queries have been given a realistic context, portraying the activity of a wholesale supplier.
- TPC-H does not represent the activity of any particular business segment, but rather any industry which must manage sell, or distribute a product worldwide (e.g., car rental, food distribution, parts, suppliers, etc.).
- It includes:
  - a logical schema
  - a set of queries
  - a scalable set of data

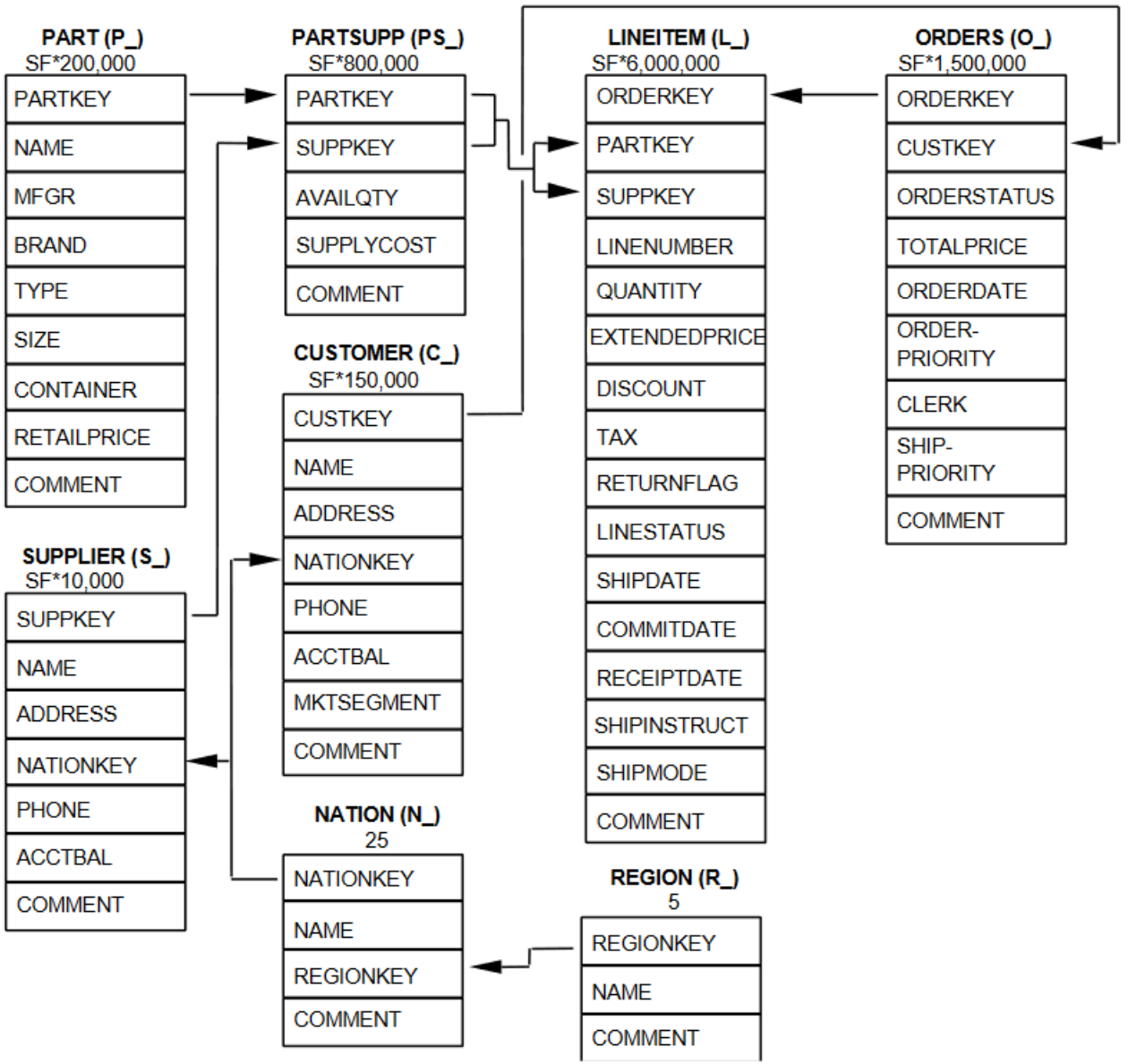


TPC-H - Schema

SF is Scale Factor * Number of rows

→

One-to-many association



TPC-H - Queries

- These selected queries provide answers to the following classes of business analysis:
  - Pricing and promotions;
  - Supply and demand management;
  - Profit and revenue management;
  - Customer satisfaction study;
  - Market share study;
  - Shipping management.
- The queries that have been selected exhibit the following characteristics:
  - They have a high degree of complexity;
  - They use a variety of access;
  - They are of an ad hoc nature;
  - They examine a large percentage of the available data;
  - They all differ from each other;
  - They contain query parameters that change across query executions.

TPC-H – Benchmark download

- The benchmark can be downloaded at site www.tpc.org
- Complete documentation of the benchmark
- https://www.tpc.org/tpc_documents_current_versions/pdf/tpc-h_v3.0.1.pdf
- The documentation contains:
- A complete specification of the logical schema (relational schemata and data types).
- A set of predefined parametric queries.
- The package for creating the set of data can be obtained at site
- https://www.tpc.org/tpc_documents_current_versions/current_specifications5.asp
- The package allows to generate data parametric with respect to a Scale Factor

TPC-H – Benchmark download

- **Metrics for a DB with scale factor 1**

Table Name	Cardinality (in rows)	Length (in bytes) of Typical[2] Row	Typical[2] Table Size (in MB)
SUPPLIER	10,000	159	2
PART	200,000	155	30
PARTSUPP	800,000	144	110
CUSTOMER	150,000	179	26
ORDERS	1,500,000	104	149
LINEITEM[3]	6,001,215	112	641
NATION[1]	25	128	< 1
REGION[1]	5	124	< 1
Total	8,661,245		956

[1] Fixed cardinality: does not scale with SF.

[2] Typical lengths and sizes given here are examples, not requirements, of what could result from an implementation (sizes do not include storage/access overheads).

[3] The cardinality of the LINEITEM table is not a strict multiple of SF since the number of lineitems in an order is chosen at random with an average of four (see Clause 4.2.5.2).

TPC-H – Benchmark download

- **Metrics for supported scale factor**

Table 4: LINEITEM Cardinality shows the cardinality of the LINEITEM table at all authorized scale factors.

Table 4: LINEITEM Cardinality

Scale Factor (SF)	Cardinality of LINEITEM Table
1	6001215
10	59986052
30	179998372
100	600037902
300	1799989091
1000	5999989709
3000	18000048306
10000	59999994267
30000	179999978268
100000	599999969200

TPC-H – Benchmark download

- **Download the dbgen package.**
- **Implement the DB in the DBMS PostgreSQL following the specification (structure and datatypes)**
- **Populate the DB using data generated by dbgen (csv files)**

A SF = 10 is suggested

- Collect some table statistics
- Number of rows
- Table size
- Number of distinct values for each attribute
- MinValue and MaxValue for each meaningful attribute.

TPC-H queries (1)

- **Local revenue value.**
- **Aggregation of the revenue of lineitems which are locally sold. Locally means that the customer and the supplier are in the same nation (→ region). The revenue is obtained by** l_extendedprice * (1 - l_discount) of the considered lineitems.
- **The aggregations should be performed with the following roll-up**
- **Month → Quarter → Year**
- **Type**
- **Nation → Region**

TPC-H queries (2)

- **Export/import revenue value.**
- **Aggregation of the export/import of revenue of lineitems between two nations (E,I) where E is the nation of the lineitem supplier and I the nations of the lineitem customer. The revenue is obtained by** l_extendedprice * (1 - l_discount) of the considered lineitems
- **The aggregations should be performed with the following roll-up**
- **Month → Quarter → Year**
- **Type**
- **Nation → Region**

TPC-H queries (3)

- **Late delivery.**
- Number of orders where at least one lineitem has been received later than the committed date.
- The aggregations should be performed with the following **roll-up**
- Month → Year
- Nation → Region (Customer)

The same query can be issued with the following **slicing**

A specific Month

A specific Type of lineitem

TPC-H queries (4)

- **Returned item loss.**

- The query gives the revenue loss for customers who might be having problems with the parts that are shipped to them. Revenue lost is defined as sum(l_extendedprice*(1-l_discount)) for all qualifying lineitems.

- **The aggregations should be performed with the following roll-up**
- **Month → Quarter → Year**
- **Customer**

The query can be issued with the following slicing (combined)

- **Name of a customer**
- **A specific quarter**