Group Member

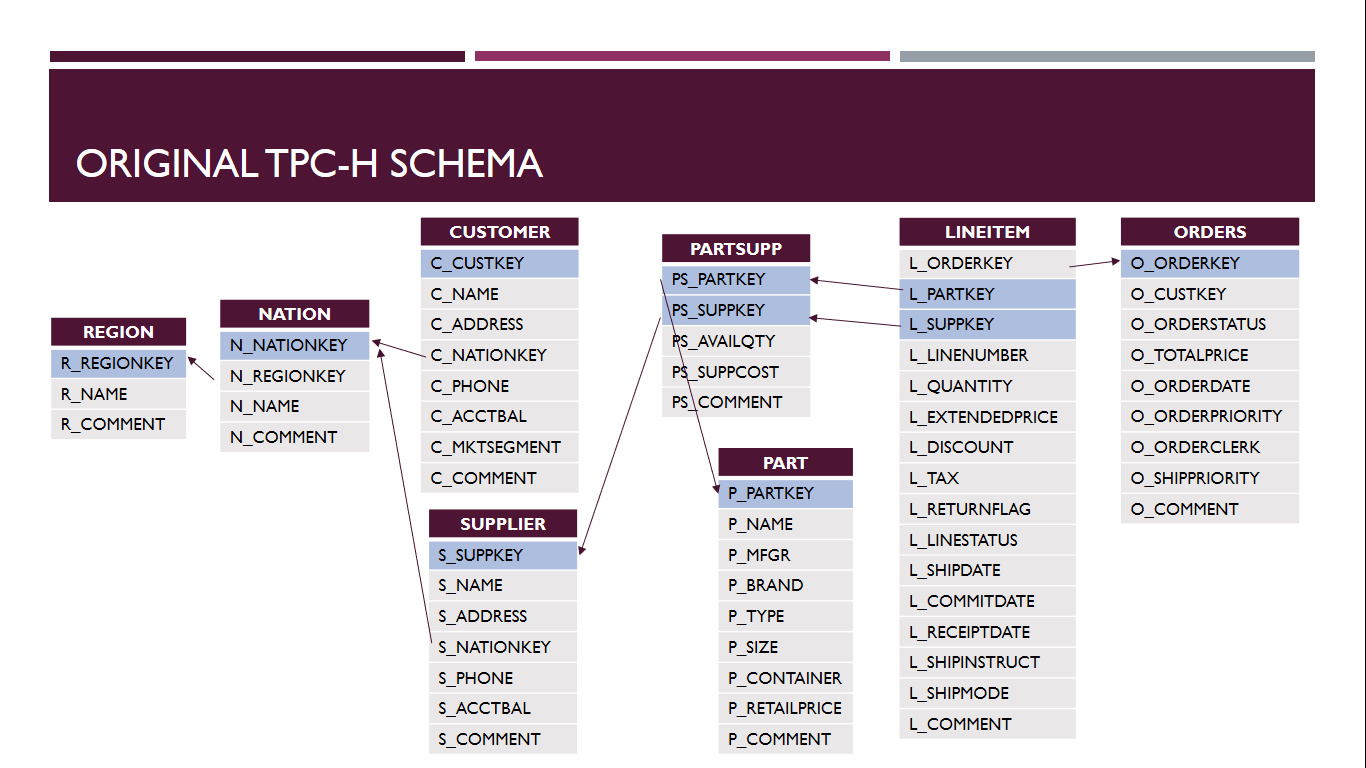
YU Zhen A0160895U  
Lu HE A0160883Y  
YAU Chung Yin A0160936B

BT 5110  
Data Management and Warehousing

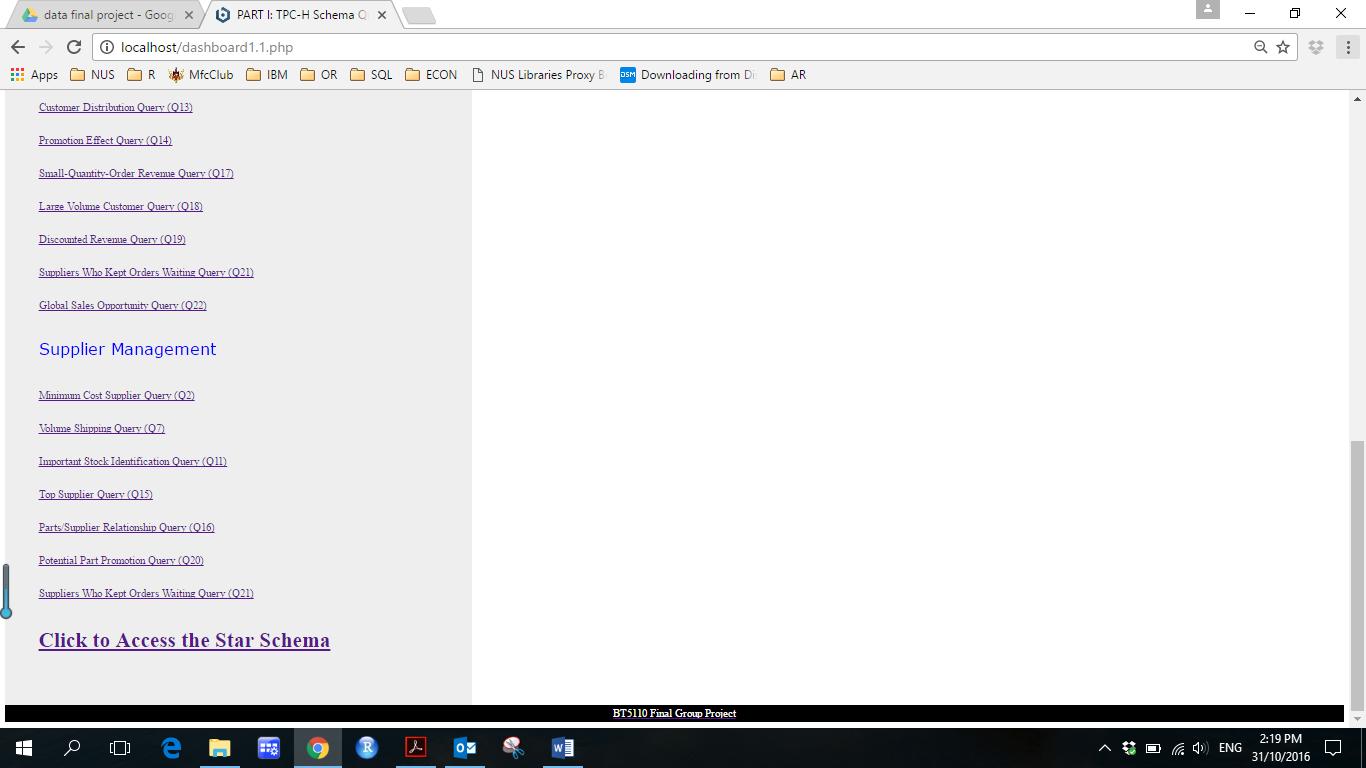
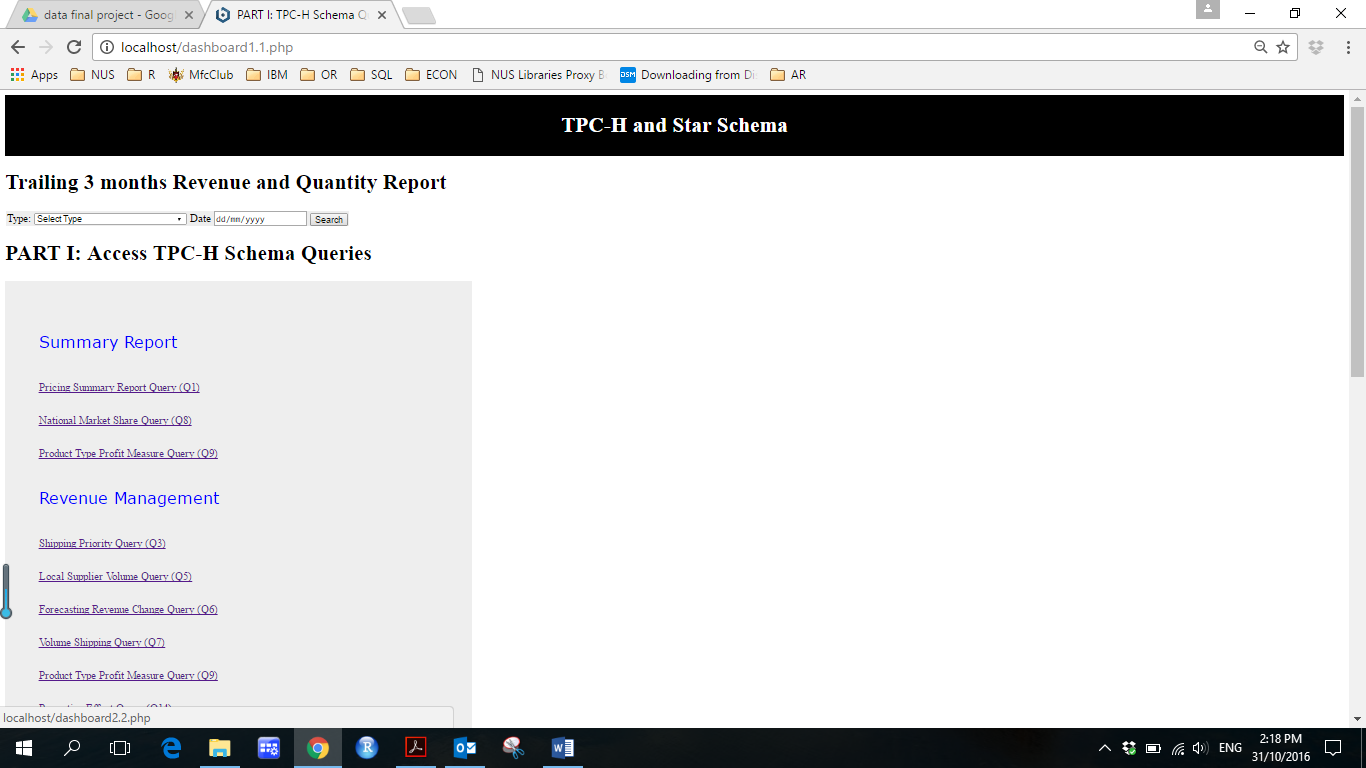
Final Project Group 12

# **Part I - TPC-H Schema**

## Basic Structure



## Directory Dashboard – TPC-H Schema



* **Revenue Summary report for the C-suites**

The trailing 3 months Revenue and Quantity Report captures the total revenue by types and by nation in the past 3 months. Management can quickly identify revenue contributions of a segment and also identify potential markets.

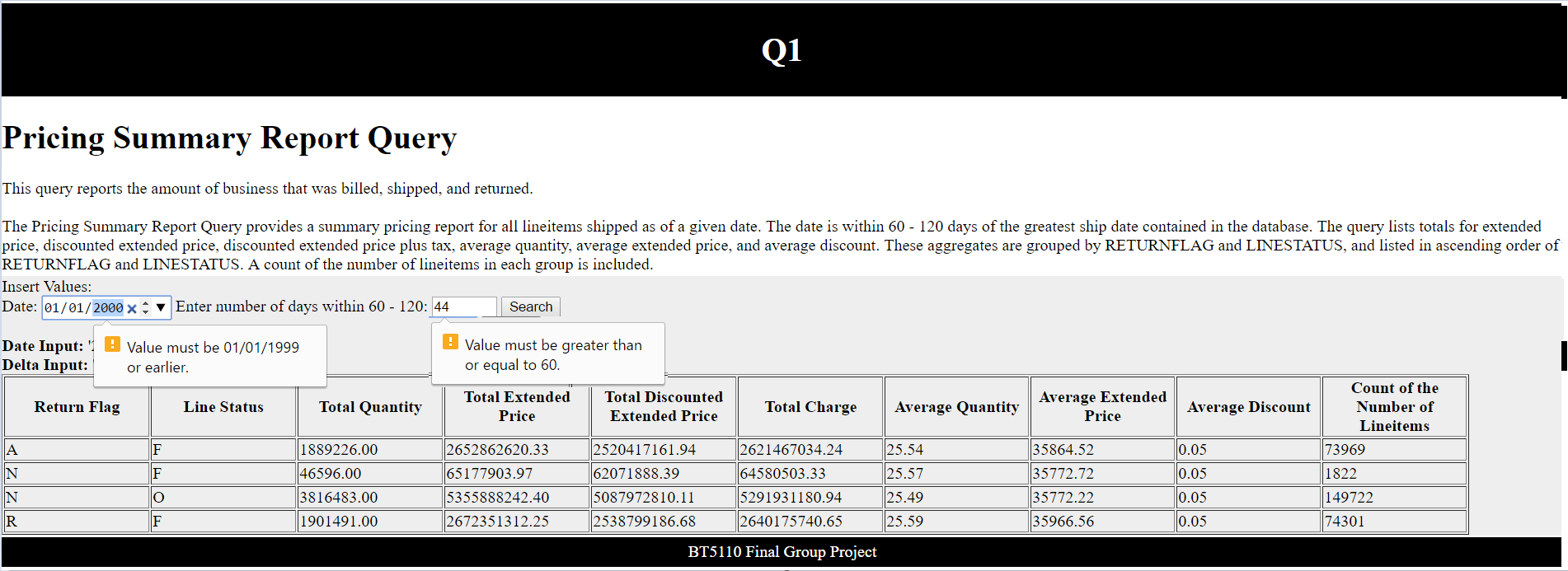
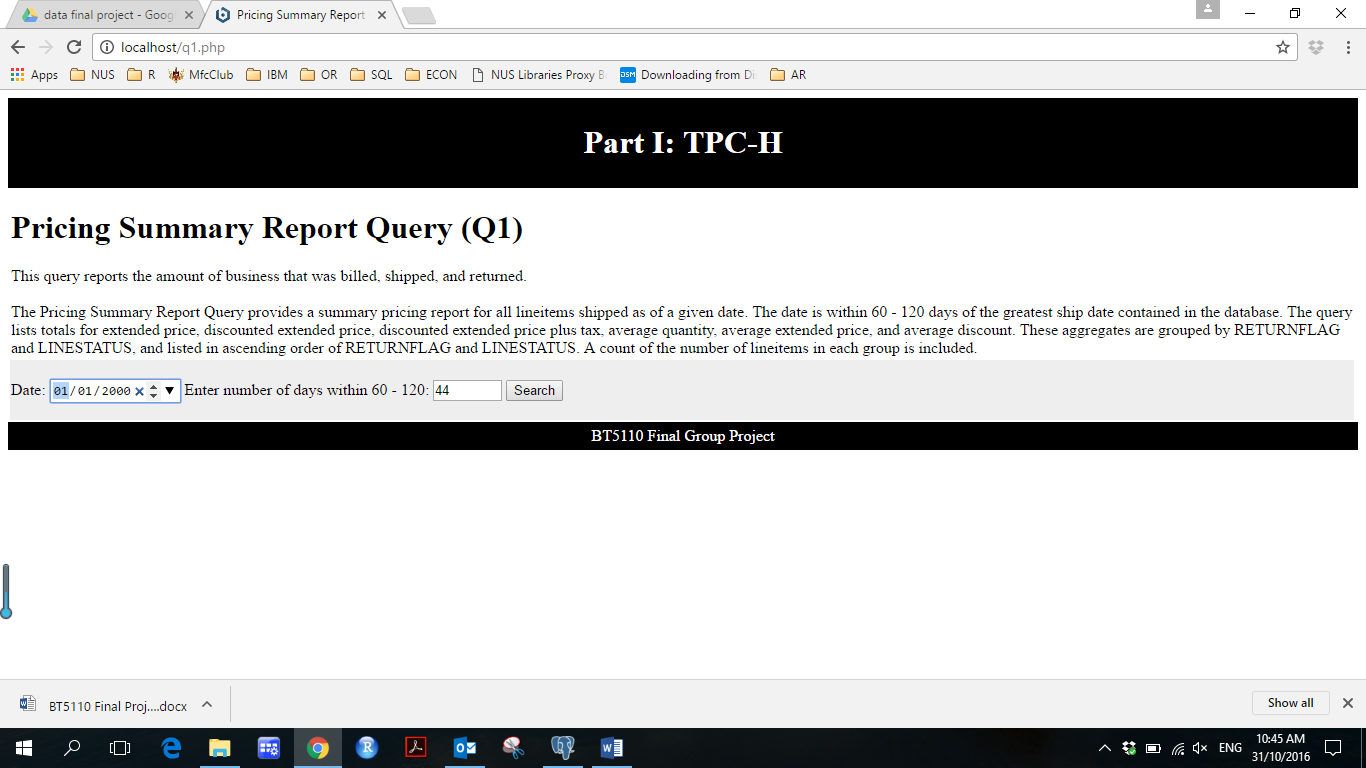
* **Quick directory to all queries and by categories**

Users can access to all queries via directory. Also the reports are categorized to enable quick access for different users in different business domains.

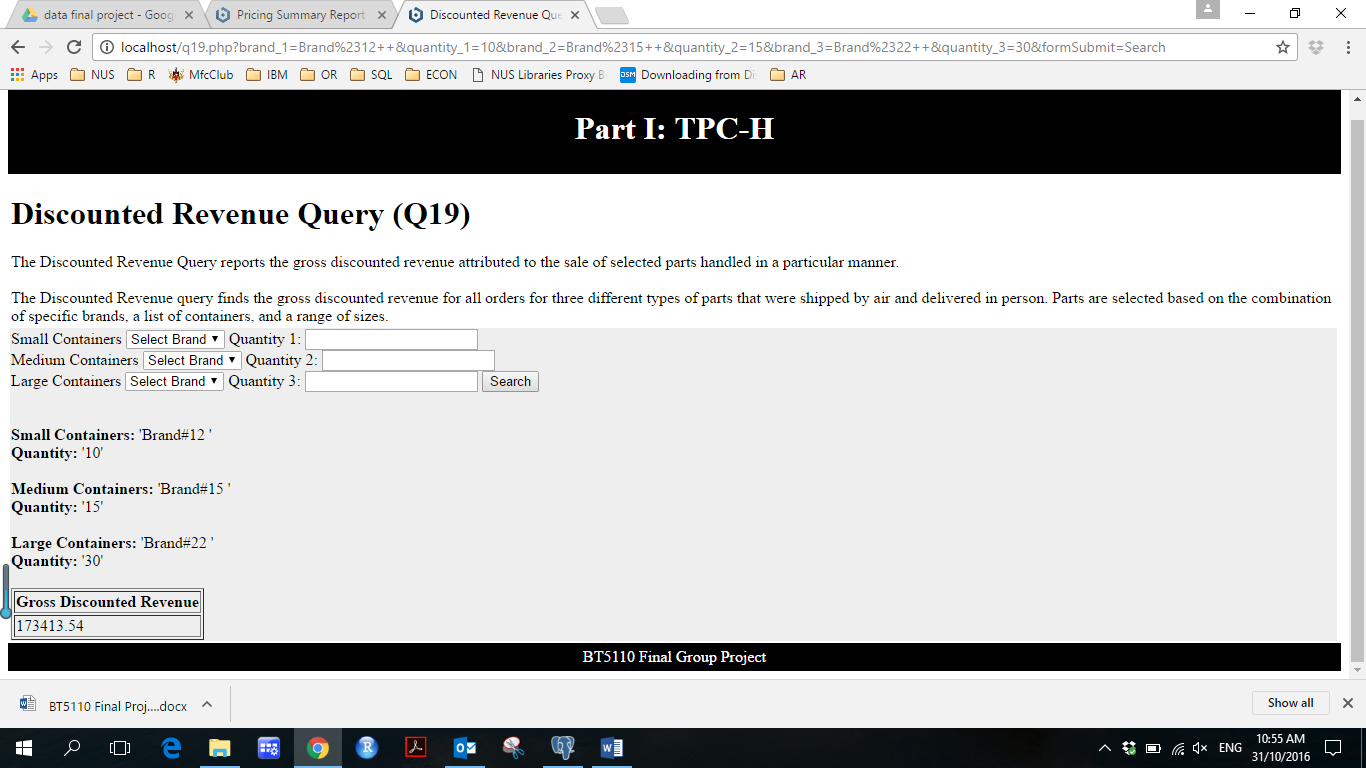
* **Easy switch between TPC-H and Star Schema**

Staff from IT departments will find it useful to evaluate the query performance between TPC-H schema and Star Schema.

## Sample Query Dashboard (Q1):



## Sample Query Dashboard (Q19):



* **Date and value constraints**

Only meaningful values are allowed as inputs. This is to minimize the likelihood of generating irrelevant outputs

* **Calendar selection of dates**

A user-friendly feature to select date value with the date constraints

* **Drop down selection of values**

Users can choose predefined values only to avoid unrecognized text input

* **Input Records**

Users are able to use their input as reference to trace the error if unexpected table is generated

* **Column in layman terms**

Non-technical users will also be able to interpret the tables

* **Values with max 2 decimal places**

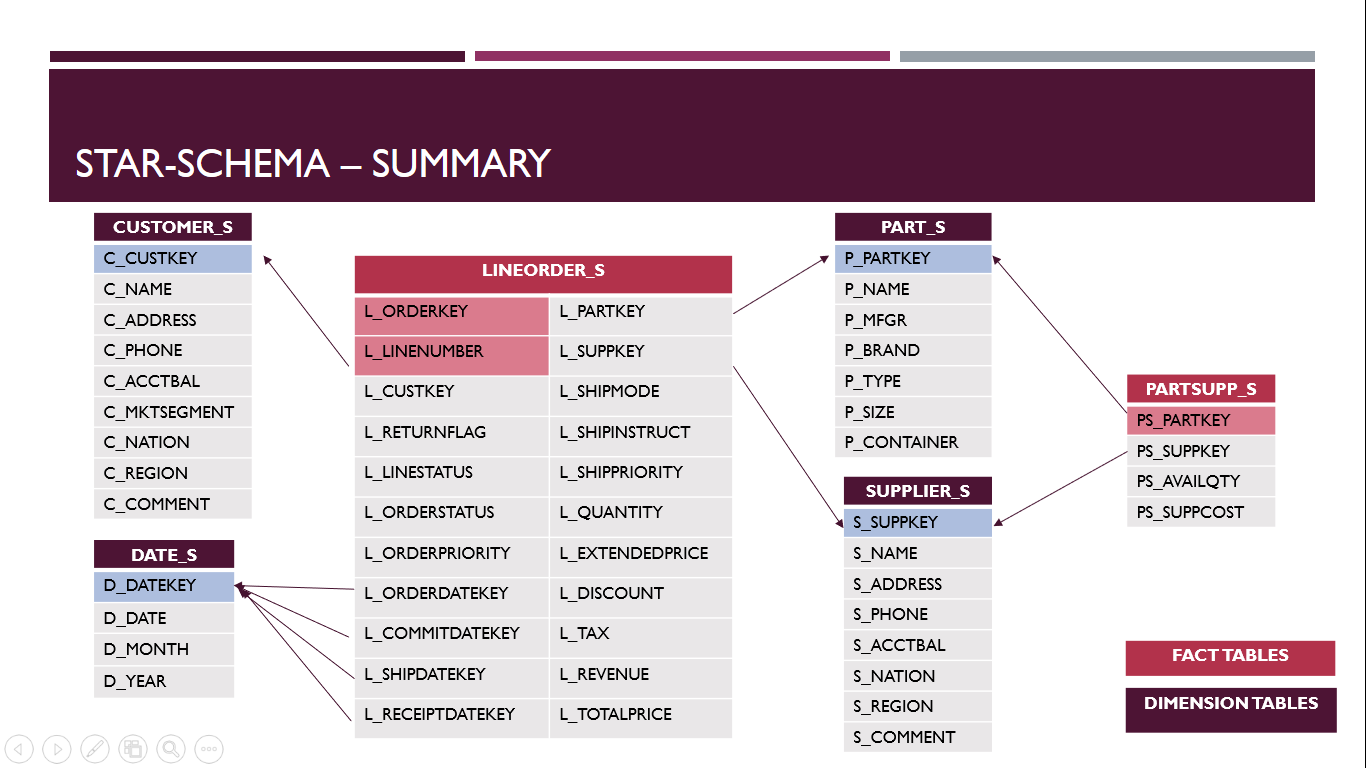
Make tables easier to read and more organized

## Results of 22 Queries (Same Results for using TPC-H and Star Schema)

|  |  |  |
| --- | --- | --- |
| **Query** | **Input** | **Sample Output** |
| **Q1** | Date Input: '1995-02-02' Delta Input: '60' | |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Return Flag | Line Status | Total Quantity | Total Extended Price | Total Discounted Extended Price | Total Charge | Average Quantity | Average Extended Price | Average Discount | Count of the Number of Line items | | A | F | 1602408.00 | 2250594238.00 | 2138388502.00 | 2224050855.63 | 25.54 | 35868.90 | 0.05 | 62745 | | R | F | 1613845.00 | 2267757527.30 | 2154253624.45 | 2240134838.31 | 25.58 | 35947.65 | 0.05 | 63085 | |
| **Q2** | Region Input: ‘AFRICA ‘ Type Input: ‘ECONOMY ANODIZED NICKEL’ Size Input: ‘4’ | |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | Supplier’s Account Balance | Supplier’s Name | Supplier’s Nation | Part’s Number | Part’s Manufacturer | Supplier’s Address | Supplier’s Phone Number | Comment Information | | 2291.35 | Supplier#000000228 | KENYA | 3831 | Manufacturer#3 | pyTY uocaSasIUlrHUbBwM,r, | 24-920-216-6514 | s above the final deposits detect along the furiously idle packages. B | |
| **Q3** | Segment Input: ‘BUILDING ‘ Region Input: ‘1994-07-07’ | |  |  |  |  | | --- | --- | --- | --- | | Order Number | Revenue | Order Date | Shipping Priority | | 290851 | 339179.84 | 1994-06-15 | 0 | | 221060 | 313037.57 | 1994-07-03 | 0 |   …… |
| **Q4** | Date Input: ‘1995-07-07’ | |  |  | | --- | --- | | Order Priority | Number of Orders | | 1-URGENT | 497 | | 2-HIGH | 532 |   …… |
| **Q5** | Region Input: ‘EUROPE ‘ Date Input: ‘1992-06-06’ | |  |  | | --- | --- | | Nation | Revenue | | UNITED KINGDOM | 3629750.37 | | GERMANY | 3582108.27 |   …… |
| **Q6** | Date Input: ‘1994-04-04’ Discount Input: ‘0.01’ Quantity Input: ‘2’ | |  | | --- | | Revenue | | 3874.8785 | |
| **Q7** | Supplier’s Nation: ‘ALGERIA ‘ Customer’s Nation: ‘BRAZIL ‘ | |  |  |  |  | | --- | --- | --- | --- | | Supplier Nation | Customer Nation | Shipping Year | Revenue | | ALGERIA | BRAZIL | 1995 | 3012801.74 | | ALGERIA | BRAZIL | 1996 | 2460837.66 |   …… |
| **Q8** | Nation Input: ‘CHINA ‘ Region Input: ‘AMERICA ‘ Type Input: ‘ECONOMY BRUSHED BRASS’ | |  |  | | --- | --- | | Year | Market Share | | 1995 | 0.16 | | 1996 | 0.07 | |
| **Q9** | Keyword Input: ‘blue’ | |  |  |  | | --- | --- | --- | | Nation | Year | Sum of Profit | | ALGERIA | 1998 | 1154411.93 | | ALGERIA | 1997 | 1726304.02 |   …… |
| **Q10** | Date Input: ‘1993-02-02’ | |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | Customer Key | Customer Name | Revenue | Account Balance | Nation | Address | Phone | Comment | | 2821 | Customer#000002821 | 528360.05 | 5378.90 | GERMANY | 3WSNaGzMWQ sej PQGRPygZN4YjzWAH2c | 17-964-888-4096 | c, express pains are carefully about the furiously even pinto beans…… | | 1771 | Customer#000001771 | 477738.80 | 3151.21 | IRAQ | evmQypmt DbfynZ4bXvm0KUNtyvynyDp3zjcXX | 21-345-763-5234 | tes wake carefully according to the unusual accounts…… |   …… |
| **Q11** | Nation Input: ‘CANADA ‘ Percentage Input: ‘0.001’ | |  |  | | --- | --- | | Part Key | Value | | 8860 | 12156678.15 | | 6019 | 11678899.58 |   …… |
| **Q12** | First Ship Mode Input: ‘MAIL ‘ Second Ship Mode Input: ‘SHIP ‘ Date Input: ‘1992-08-01’ | |  |  |  | | --- | --- | --- | | Ship Mode | High Line Count | Low Line Count | | MAIL | 331 | 484 | | SHIP | 308 | 477 | |
| **Q13** | First Special Category Input: ‘special’ Second Special Category Input: ‘requests’ | |  |  | | --- | --- | | Number of Orders | Count of Customers | | 0 | 2608 | | 9 | 322 |   …… |
| **Q14** | Date Input: '1995-09-09' | |  | | --- | | Percentage of Revenue | | 15.98 | |
| **Q15** | Date Input: '1998-11-11' | |  |  |  |  |  | | --- | --- | --- | --- | --- | | Supplier Key | Name | Address | Phone | Total Revenue | | 74 | Supplier#000000074 | uM3yV5NOc6b5wNdpxF69CW 8QvDxqvKubRJtA | 30-166-486-1559 | 11874 | |
| **Q16** | Brand Excluded: 'Brand#11 '  Type Excluded: 'STANDARD POLISHED TIN'  Size 1: '1' Size 2: '5' Size 3: '10'  Size 4: '15' Size 5: '20' Size 6: '25'  Size 7: '30' Size 8: '35' | |  |  |  |  | | --- | --- | --- | --- | | Brand | Type | Size | Count | | Brand#12 | ECONOMY POLISHED COPPER | 25 | 8 | | Brand#12 | MEDIUM PLATED STEEL | 20 | 8 |   …… |
| **Q17** | Brand: 'Brand#25 '  Container: ‘LG BOX’ | |  | | --- | | Average Yearly Revenue | | 16310.85 | |
| **Q18** | Quantity: ‘300’ | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Customer Name | Customer ID | Order ID | Order Date | Total Price | Sum of Quantity | | Customer#000003328 | 3328 | 29158 | 1995-10-21 | 403769.47 | 305 | | Customer#000000889 | 889 | 6882 | 1997-04-09 | 361395.73 | 303 | | |  | |
| **Q19** | Small Containers: 'Brand#11 ' Quantity: '10'  Medium Containers: 'Brand#22 ' Quantity: '20'  Large Containers: 'Brand#33 ' Quantity: '30' | |  | | --- | | Gross Discounted Revenue | | 192650.31 | |
| **Q20** | Part Name: ‘blush%’  Date: ‘2014-01-01’  Nation: ’FRANCE’ | |  |  | | --- | --- | | Supplier Name | Supplier Address | | Supplier#000000070 | "INWNH2w,OOWgNDq0BRCcBwOMQc6PdFDc4" | | Supplier#000000090 | "bPE6Uhz1f2m3gwSGMrnRt,g,3gq37r5kxgphqss1" |   …… |
| **Q21** | Nation: 'FRANCE' | |  |  | | --- | --- | | Supplier Name | Number Wait | | Supplier#000000468 | 17 | | Supplier#000000104 | 15 |   …… |
| **Q22** | Country Code 1: '10'  Country Code 2: '12'  Country Code 3: '14'  Country Code 4: '16'  Country Code 5: '18'  Country Code 6: '20'  Country Code 7: '24' | |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  |  | | --- | --- | --- | | Country Code | Number of Customer | Total Account Balance | | 10 | 44 | 342467.86 | | 12 | 38 | 275479.28 |   …… | |  | |

# **Part II - Star Schema**

## Basic Structure



## Rationale

We follow the 4-Step processes in designing the Star-Schema as suggested in Kimball Chapter2.

|  |  |  |
| --- | --- | --- |
| Step 1: Business Process Involved | **Sales/Orders** | **Purchase/Inventory** |
| Step 2: Grain of the Business Process | A line item on a customer’s order | a snapshot of the inventory level of a specific part from a specific supplier |
| Step 3: Dimensions | Customer, Supplier, Part, Date | Supplier, Part |
| Step 4: Numeric Facts | Addictive fact: Quantity, Extended Price, Revenue (extended price after discount), Total Price.  Non-addictive fact: Discount, Tax. | Facts:  Cost  Available quantity |
| Other attributes | Order Details, Shipment Details | N/A |

## Improvements

* **Separate the information of Sales/Order with Purchase/Inventory**

The original *LINEITEM* and *Orders* gives transactional-level sales records over 7 years, while *PARTSUPP* table gives periodic snapshot (assuming current date) of available quantity and cost. The difference in granularity makes it necessary to have two different fact tables in the Star schema.

* **Combine *LINEITEM* and *ORDERS* to have the lowest possible grain of each dimension**

Combining the *LINEITEM* table and *ORDERS* table is a necessary de-normalization to capture the lowest possible grain while avoiding too many unnecessary joins.

* **De-normalised *NATION* and *REGION* tables**

Snowflaking (dimension normalisation) is necessary in most data mart designs. In this case, however, the geographic information on nation pertaining to a specific customer/supplier will rarely change. Also, it is rare that one nation would belong to a different region. Considering the costs and benefits in this case, we adopt the approach to de-normalised *NATION* and *Region* tables and store such information in the *CUSTOMER\_S, SUPPLIER\_S* tables instead, to avoid too many joints and to improve query performance.

* **Drop unnecessary columns such as *P\_RETAILPRICE, P\_COMMENT, O\_COMMENT, PS\_COMMENT, O\_ORDERCLERK, L\_COMMENT*.**

Such columns are not relevant to the business queries while taking up substantial storage space.

* **Add *DATE* dimension with use of surrogate key**

As mentioned in Kimball Chapter 2, the date dimension is almost guaranteed in every data mart. To make use of the advantage of surrogate key, we updated the date value of *L\_COMMITDATE, L\_ORDERDATE, L\_SHIPDATE, L\_RECEIPTDATE* to date key and set foreign constraints by referencing to the *DATE* table. For illustration purpose, we only included “month” and “year” as attributes of the *DATE* table. Depending on the business need, more attributes can be stored in the *DATE* table, e.g. fiscal year, holiday seasons, or other seasonal indicators that is relevant to the business.

* **Add frequently-used fact to improve query performance: L\_REVENUE**

We noticed that L\_EXTENDEDPRICE \* (1 - L\_DISCOUNT), which represented the revenue (after discount and before tax charge), is one of the most demanded business information and used in 11 queries out of 22, namely Q1, Q3, Q5, Q6, Q7, Q8, Q9, Q10, Q14, Q15 and Q19. By saving this value into each row of the fact table *LINEORDER* as L\_REVENUE, the query performance will be improved by reducing the additional computation time.

## Directory Dashboard – Star Schema

## 

## 

## Results of 22 Queries by using Star Schema

Please refer to Part I for the same results.

## Comparison of Query Performance related to L\_REVENUE

We tested certain Queries (requires information of L\_REVENUE = L\_EXTENDEDPRICE \* (1 - L\_DISCOUNT)) by comparing the average processing time.

|  |  |  |
| --- | --- | --- |
| **Selected Queries involve Revenue** | **Average Processing Time**  **(using L\_REVENUE)** | **Average Processing Time**  **(using L\_EXTENDEDPRICE\*(1-L\_DISCOUNT))** |
| Q3 | 255msec | 272msec |
| Q5 | 234msec | 235msec |
| Q7 | 218msec | 222msec |
| Q10 | 190msec | 202msec |
| Q15 | 348msec | 358msec |
| Q19 | 177msec | 183msec |