LAB 1: Data preparation to make Star Schema

Data Warehouse:

A data warehouse is a centralized repository that stores large volumes of structured, semi-structured, or unstructured data from various sources within an organization. It is designed to support business decision-making processes by enabling the analysis and reporting of data from different systems into a single, unified format.

Key characteristics of a data warehouse:

- Consolidation of Data
- Subject-oriented
- Time-variant
- Non-volatile

SnowFlake:

SnowFlake is a data platform that would harness the immense power of the cloud. The founders of SnowFlake engineered Snowflake to power the Data Cloud, where thousands of organizations have seamless access to explore, share, and unlock the true value of their data.

Understanding the database:

Tables of Snowflake_Sample_Data / TPCH_SF10:

a. CUSTOMER
b. LINEITEM
c. NATION
d. ORDERS
e. PART
f. PARTSUPP
g. REGION
h. SUPPLIER

Columns of above each tables:

a. CUSTOMER

- C_ACCTBAL
- C_ADDRESS
- C_COMMENT
- C_CUSTKEY
- C_MKTSEGMENT
- C_NAME
- C_NATIONKEY
- C_PHONE

b. LINEITEM

- L_COMMENT
- L_COMMITDATE
- L_DISCOUNT
- L_EXTENDEDPRICE
- L_LINENUMBER
- L_LINESTATUS
- L_ORDERKEY
- L_PARTKEY
- L_QUANTITY
- L_RECEIPTDATE
- L_RETURNFLAG
- L_SHIPDATE
- L_SHIPINSTRUCT
- L_SHIPMODE
- L_SUPPKEY
- L_TAX

c. NATION

- N_COMMENT
- N_NAME
- N_NATIONKEY
- N_REGIONKEY

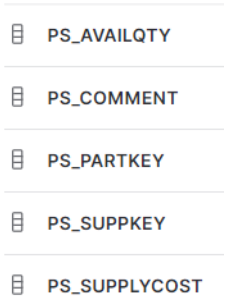
d. ORDERS

- O_CLERK
- O_COMMENT
- O_CUSTKEY
- O_ORDERDATE
- O_ORDERKEY
- O_ORDERPRIORITY
- O_ORDERSTATUS
- O_SHIPPRIORITY
- O_TOTALPRICE

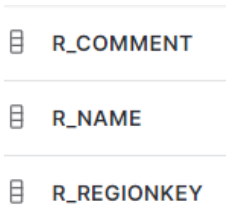
e. PART

- P_BRAND
- P_COMMENT
- P_CONTAINER
- P_MFGR
- P_NAME
- P_PARTKEY
- P_RETAILPRICE
- P_SIZE
- P_TYPE

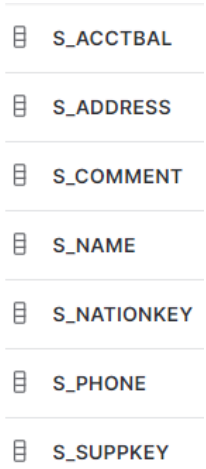
f. PARTSUPP



g. REGION



h. SUPPLIER



For the blueprint of preparation for star schema, the fact table is orders table and the dimensions for this fact table is time and location.

The following SQL commands are used to understand the star schema

a. Count of orders when date is '1995-3-06'

SELECT count(O_ORDERKEY) FROM ORDERS WHERE O_ORDERDATE = '1995-3-06'

	...	COUNT(O_ORDERKEY)
1		6293

b. Count of orders when year is 1995 and month is 3 order by date

SELECT count(O_ORDERKEY)

FROM ORDERS

WHERE year(o_orderdate) = 1995 and month(o_orderdate) = 03
GROUP BY o_orderdate

	...	COUNT(O_ORDERKEY)
1		6353
2		6268
3		6369
4		6264
5		6296
6		6236
7		6223
8		6440
9		6261
10		6152
11		6324
12		6361
13		6107
14		6251
15		6255
16		6215
17		6152
18		6228
19		6210
20		6304
21		6132
22		6185
23		6225
24		6214
25		6142
26		6293
27		6387
28		6300
29		6296
30		6153
31		6123

c. Count of orders of each month in 1995 year

SELECT count(O_ORDERKEY), MONTH(o_orderdate)

FROM ORDERS

WHERE year(o_orderdate) = 1995 GROUP BY MONTH(o_orderdate)

	··· COUNT(O_ORDERKEY)	MONTH(O_ORDERDATE)
1	193057	10
2	193116	5
3	193902	1
4	193607	7
5	187016	4
6	187281	9
7	193207	12
8	186946	6
9	193223	8
10	193719	3
11	173991	2
12	186510	11

d. Count of orders of years from 1992 to 1998

SELECT count(O_ORDERKEY), year(o_orderdate)

FROM ORDERS

GROUP BY year(o_orderdate)

	··· COUNT(O_ORDERKEY)	YEAR(O_ORDERDATE)
1	2275511	1997
2	1333214	1998
3	2275919	1994
4	2275575	1995
5	2281205	1992
6	2276638	1993
7	2281938	1996

e. Count of orders joining customer and nation by nation name where nation name is India

select count(n.N_NAME)

```
from orders as o
left join customer as c on (o.O_CUSTKEY= c.C_CUSTKEY)
left join nation as n on (n.N_NATIONKEY = c.C_NATIONKEY)
where n.N_NAME ='INDIA';
group by (n.N_NAME)
```

	COUNT(N.N_NAME)
1	600735