Data Base and Data Mining Group of Politecnico di Torino

Data Science and Database Technology

Politecnico di Torino

Problem specifications

As part of the enhancement of the excellence of Made in Italy, the Italian Government wants to analyze the export of Italian wines abroad, to understand, for example, which are the most popular wines, from which areas of Italy they come, and which are the main foreign countries importers.
The Government has collected from Italian wine companies the list of purchase orders. Each purchase order is characterized by date of the order, amount in euros, quantity in liters, type of wine, type of packaging (e.g., bottle, carton, demijohn, etc.), VAT number of the wine company of origin, foreign country of destination. The analysts, starting from the purchase orders, want to create a data warehouse which synthesizes the collected data in order to answer a set of queries.
During the implementation phase, analysts integrate order data with context data in their possession, necessary to carry out the analysis of interest. In particular, each type of wine is associated to the list of quality certifications (DOC, DOP, or DOCG) it owns. For example: the wine type "Barolo" is always associated to the DOCG certification, the wine type "Nebbiolo" is always associated to DOP and DOCG certifications, etc. Each certification can be associated to more than one type of wine. Some types of wine can be without certifications. Each order is also associated to its size: orders up to 100 liters are considered small, medium the ones up to 1000 liters, and large the ones above. Finally, for every winery, it is possible to find out from the VAT number the complete personal data (name, address, province, region, geographical area of the region).

The queries analysts are interested in concern the total quantity of wine exported (in liters), and the average price per liter, according to the variation of:

- month, 2 months period, 3 months period, semester, and year
- province, region and geographical area (North, Center, South) of the winery
- wine type
- quality certifications → TEXT TELLS ME THE CORRISPONDANCE WITH WINE TYPE
- packaging type
- foreign country of destination, and its continent
- order size (small, medium, large)

The data warehouse will track information for the years 2010 through 2013. Following are some specific queries:

a) Considering only wines that have "DOC" certification exported to Asia, select the average price per liter for each year, the percentage of liters exported in the year compared to the total for all years, and the cumulative yearly total of liters exported. Perform the analysis separately for each type of packaging.

→ PAY ATTENTION TO THIS WHEN PARTITIONING

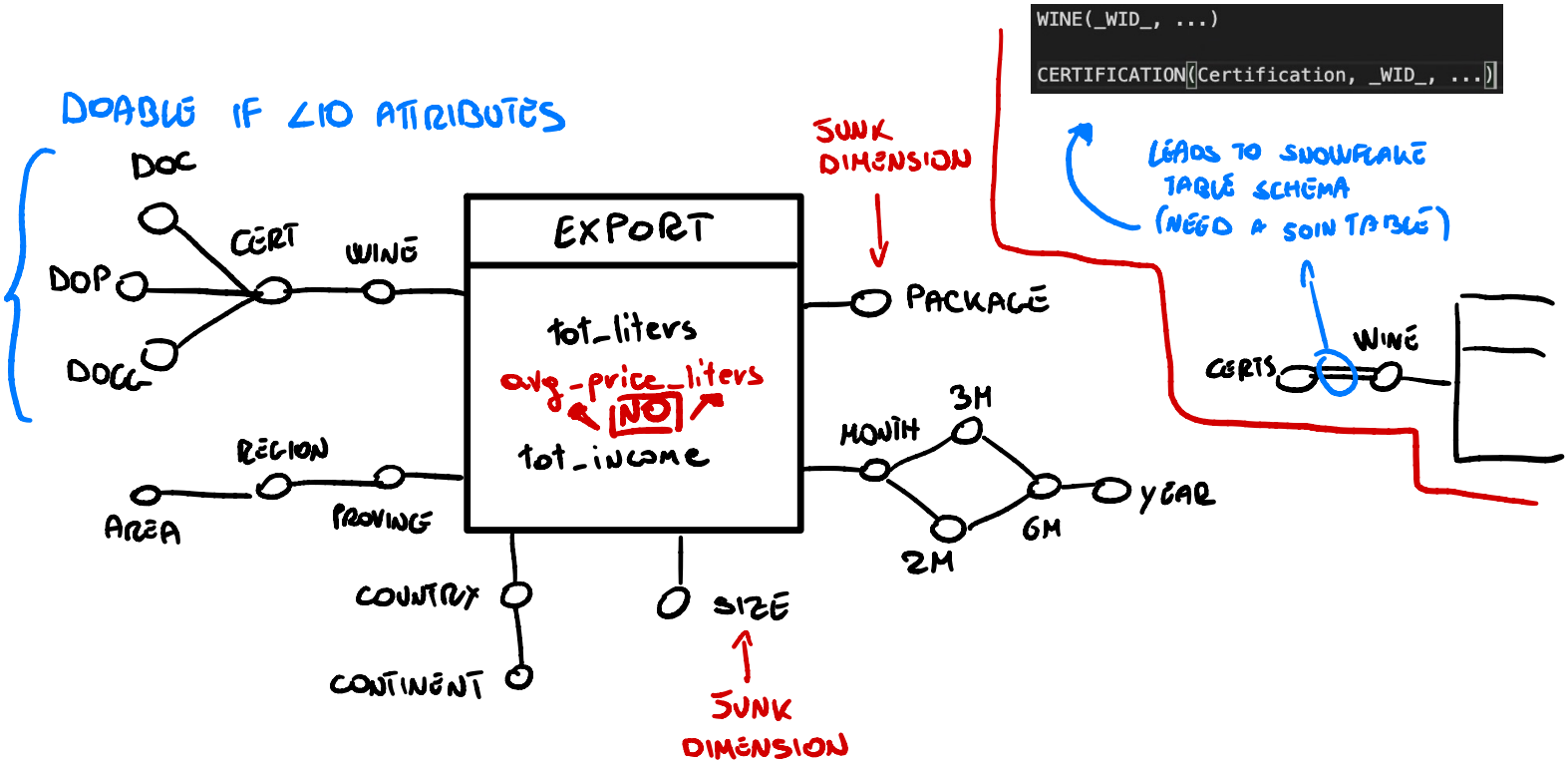
b) For each region, select the average price per liter, the average number of liters of wine exported per province, and the percentage of liters of wine exported from each region with respect to the total of the geographical area to which it belongs (North, Center, South). Perform the analysis separately for each year.

c) Assign a rank to each foreign country according to the total quantity of wine imported. Perform the analysis separately for each type of wine. Consider only large orders.

Design

1. Design the data warehouse needed to manage the needs of the appliance company to meet the requirements described in the problem specification. The designed data warehouse must also be able to efficiently respond to **all** of the frequent queries proposed in the problem specification.

2. Write frequent queries (a) and (b) using the extended SQL language.



TABLES

```
--DIMENSIONS
TIMEDIM (_TID_, Month, 2M, 3M, 6M, year)
WINERY (_WID_, Province, Region, Geo_area)
DESTINATION(_DID_, Country, Continent)
WINE_TYPE(_WTID_, Wine, DOC, DOP, DOCG)

--JUNK DIMENSIONS (TO BE PUSHED DOWN)
-- PACKAGE(_PID_, Package)
-- SIZE(_SID_, Size)

--FACT TABLE
EXPORT(_TID_, _WID_, _DID_, _WTID_, _Package_, _Size_, tot_income, tot_liters)
```

JUNK DIMS ARE BETTER PUSHED DOWN TO THE FACT TABLE

QUERIES

#1

```
--Query #1
SELECT T.year, E.Package,
    SUM(E.tot_income)/SUM(e.tot_liters),
    100*SUM(E.tot_liters)/SUM(SUM(E.tot_liters)) OVER (PARTITION BY E.Package),
    SUM(SUM(E.tot_liters)) OVER (PARTITION BY E.Package,
                                 ORDER BY T.year
                                 ROWS UNBOUNDED PRECEDING),
FROM EXPORT E, DESTINATION D, TIMEDIM T, WINETYPE W
WHERE E.DID = D.DID AND E.TID = T.TID AND E.WTID = W.WTID
    AND D.Continent = "Asia" AND W.DOC = True
GROUP BY T.year, E.Package;
```

```
--Query #2 (1)
SELECT T.year, DISTINCT W.Region,
    SUM(SUM(E.tot_income)) OVER (PARTITION BY W.Region, T.year)/
    SUM(SUM(E.tot_liters)) OVER (PARTITION BY W.Region, T.year),
    AVG(SUM(E.tot_liters)) OVER (PARTITION BY W.Region, T.year),
    --didn't quite get the one below, TODO: asks prof
    100 * SUM(SUM(E.tot_liters)) OVER (PARTITION BY W.Region, T.year)/
    SUM(SUM(E.tot_liters)) OVER (PARTITION BY W.Geo_area, T.year)
FROM EXPORT E, TIMEDIM T, WINERY W
WHERE E.WID = W.WID AND E.TID = T.TID
GROUP BY T.year, W.Region, W.Geo_area, W.Province;
```

```
--Query #2 (2)
SELECT T.year, W.Region,
    SUM(E.tot_income) / SUM(E.tot_liters),
    SUM(E.tot_liters) / COUNT(DISTINCT W.Province),
    100 * SUM(E.tot_liters) /
    SUM(SUM(E.tot_liters)) OVER (PARTITION BY W.Geo_area, T.year)
FROM EXPORT E, TIMEDIM T, WINERY W
WHERE E.WID = W.WID AND E.TID = T.TID
GROUP BY T.year, W.Region, W.Geo_area;
```

#2