Data Warehousing and Data Mining

Presentation for Project Evaluation

FREIE UNIVERSITÄT BOZEN
LIBERA UNIVERSITÀ DI BOLZANO
FREE UNIVERSITY OF BOZEN · BOLZANO

Ondrej Platek
Peteris Nikiforovs

January 19, 2012

Domain

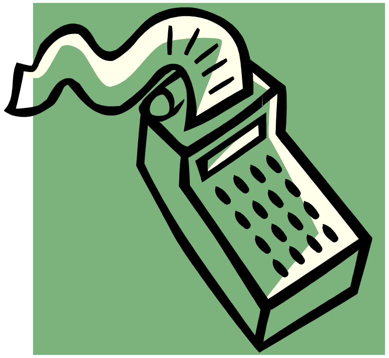
Daily newspaper

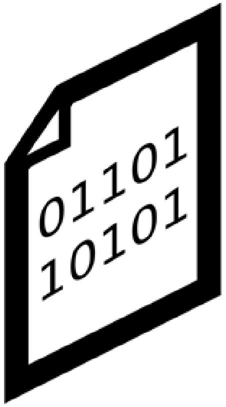
The Wall Street Journal

printed

on-line

Data Warehouse Objectives





Multiple data sources

Data Warehouse Objectives

- Largest newspaper in the US by circulation
- 400k online subscribers

Business Processes

- Selling subscriptions
- Advertising
- Online content analysis

Subscriptions

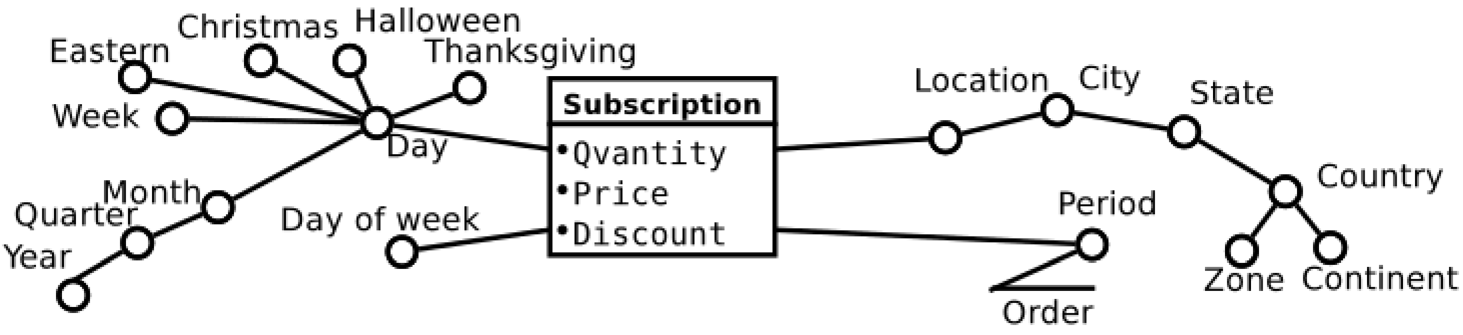
Dimensions

- International newspaper: location dimension by city, country, region
- Date dimension including holidays
- Subscription interval (month, quarter, year)

Measures

$$\sum(price\ * (1 - discount) * quantity)$$

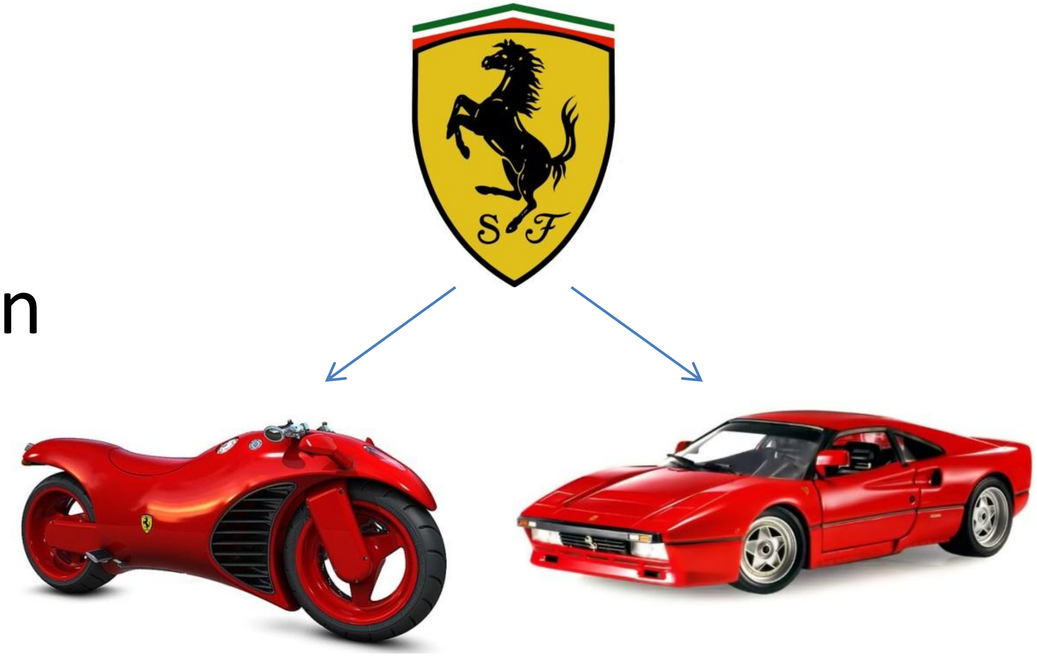
Subscriptions: Fact



Subscriptions: Business Queries

- Revenue from subscriptions by year and country
- Top 10 least profitable cities taking into account subscription sales & population
- How much would we earn without applying discounts on subscriptions, by period type and by year?
- Compare sales on various holidays in different countries

Advertising



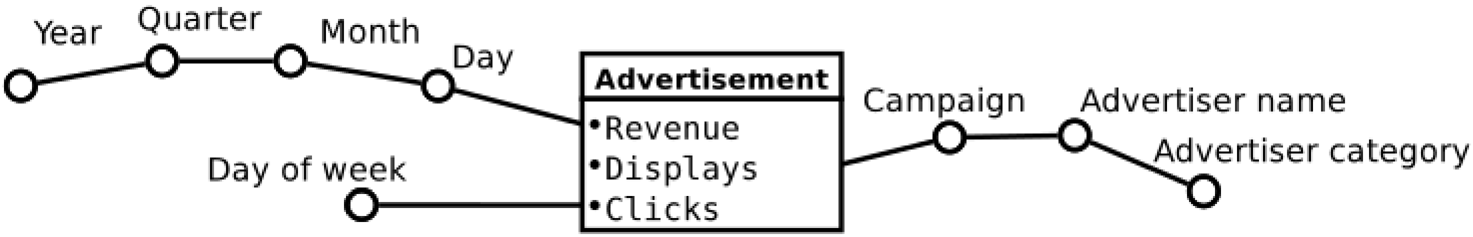
Dimensions

- Date dimension
- Campaign dimension

Measures

- $\sum revenue$
- CPM = $\frac{\sum clicks}{\sum displays}$

Advertising: Fact



Advertising: Business Queries

- Revenue by advertisers from the ”Middle Fish” category who have greater revenue than the average of the ”Big Fish” bias=0.5 together with average of ”Big Fish” advertisers?

- CPM (clicks divided by displays) for the top 10 advertisers by revenue together with the average CPM for advertisers category for the advertisers that have at least 15 campaigns?
- The campaigns that lasted more than 5 months with revenue bigger than 140k at least in once over the past 5 month, all in year 2011?

Content Analysis

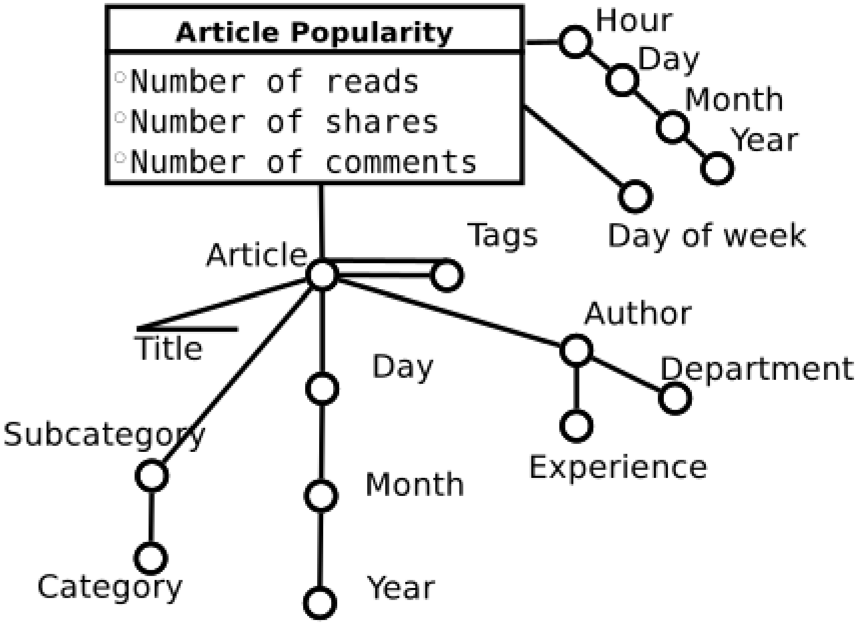
Dimensions

- Date dimension
- Article dimension
  - Publication dates
  - Categories
  - Authors
  - Tags

Mesasures

- $\sum reads, \sum comments, \sum shares$

Content Analysis: Fact



Content Analysis: Business Queries

- Top 10 read articles and their authors for every month in year 2011?
- Compare the number of reads/shares/comments of articles tagged with tags `*positive*` and `*negative*` for each year?
- Hours of the day when most articles are read grouped by category in year 2010 together with the average number of articles read during this hour in the same year?

Logical & Physical Design

- Star schema
- Oracle