TWINNING CONTRACT JO/13/ENP/ST/23

# Strengthening the capabilities of the Department of Statistics in Jordan

## *Databases, Data Warehouses and Statistical Dissemination Systems*

*Leonardo Tininini*
*ISTAT*

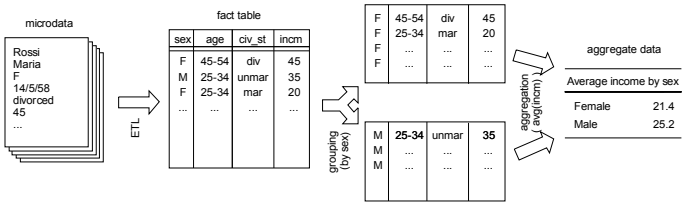
Amman, March 2014

## *The definitions (according to wikipedia)*

- **Database**
  - Organized collection of data. The data are typically organized to model relevant aspects of reality in a way that supports processes requiring this information
- **Data Warehouse**
  - A database used for reporting and data analysis. Integrating data from one or more disparate sources creates a central repository of data, a data warehouse (DW). Data warehouses store current and historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons
- **Statistical Dissemination System / Statistical Data Warehouse ???**

## *Data warehouse basic terminology*

- **aggregate data**
  - obtained by applying aggregations (count, sum, avg, etc.) over elementary data (aka raw data or **microdata**)
- **fact tables** ($D_1$ , $D_2$ , ... , $D_n$; M)
  - dimension codes (used to group data and/or to consider only specific subsets of data)
  - measure(s) (possibly to be aggregated and deriving from microdata quantitative variables)





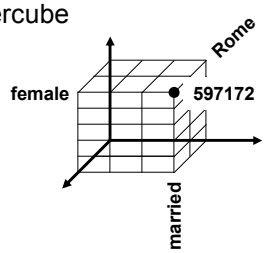

Istat



## *Data warehouse basic terminology (2)*

- **dimensions and dimension levels**
  - dimensions are often articulated in different dimension levels, e.g. a territorial dimension may comprise the levels: national, regional, municipality


- **data cube**
  - the association between dimension code combination and measure is represented by a n-dimensional hypercube

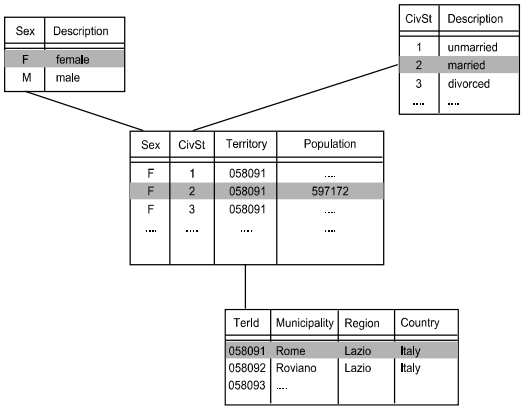





Istat

## *A relational perspective for data cubes (1)*

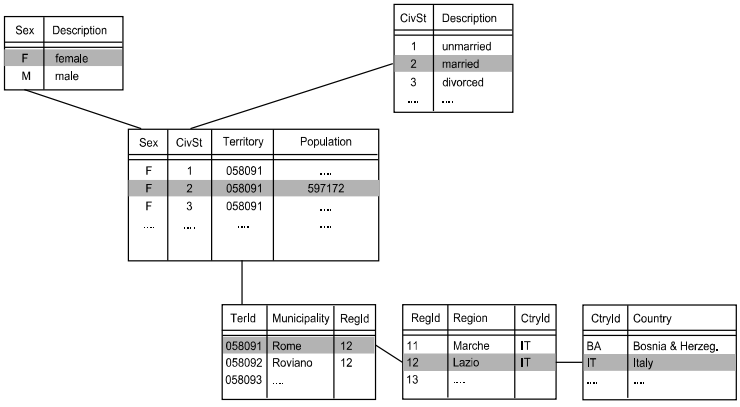
- **star schema**
  - the codes in the fact table/data cube are decoded by (possibly de-normalized) single dimensional tables






## *A relational perspective for data cubes (2)*

- **snowflake schema**
  - the codes in the fact table/data cube are decoded by trees of (normalized) dimensional tables



## *In SDS dimensional cubes are often "sparse"*

- Due to:
  - significance of sample data
  - privacy protection
  - microdata unavailability
  - ...

In Statistical Dissemination Systems many "points" of the multidimensional cube may correspond to data that can not be disseminated



## *DWH vs SDS*

- Dimensional combinations in a data warehouse…

| | C1 | C2 | C3 | C4 | C5 | C6 | C7 |
|---|---|---|---|---|---|---|---|
| M | X | X | X | X | X | X | X |

- Dimensional combinations in a statistical dissemination system…

| | C1 | C2 | C3 | C4 | C5 | C6 | C7 |
|---|---|---|---|---|---|---|---|
| M | X | | X | X | | | |
| M | X | X | | | | | |
| M | | X | X | X | | | |
| M | | X | | | X | X | |
| M | | | | X | | | X |
| M | X | | | | | X | |

## *Dimensional combinations in a DWH*

| | C1 | C2 | C3 | C4 | C5 | C6 | C7 |
|---|---|---|---|---|---|---|---|
| M | X | X | X | X | X | X | X |

•Dimensions can be combined in a completely arbitrary way:

- C1 with C2, C1 with C3, … , C6 with C7
- C1 with C2 and C3, C1 with C2 and C4, …
- …
- C1 with C2 and C3 and C4 and C5 and C6 and C7

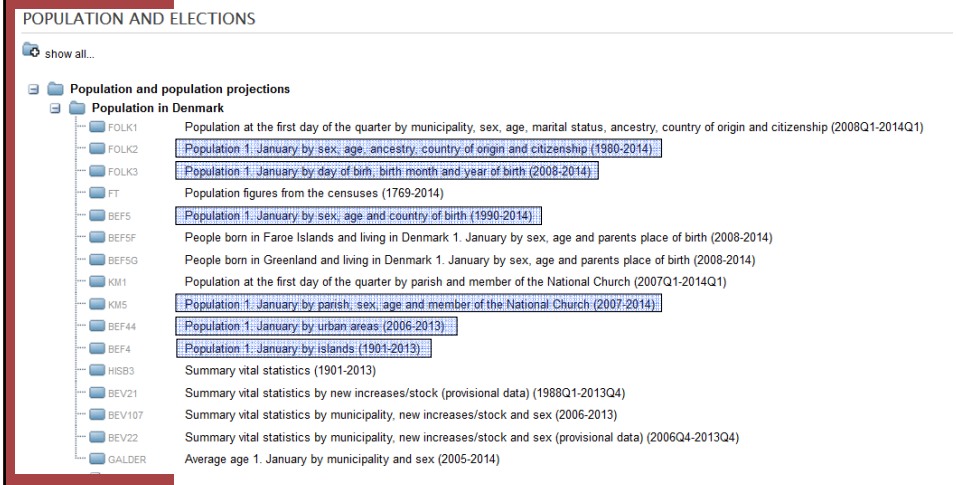


## *Dimensional combinations in a SDS*

| | C1 | C2 | C3 | C4 | C5 | C6 | C7 |
|---|---|---|---|---|---|---|---|
| M | X | | X | X | | | |
| M | X | X | | | | | |
| M | | X | X | X | | | |
| M | | X | | | X | X | |
| M | | | | X | | | X |
| M | X | | | | | X | |

- Possible combinations of dimensions are limited:
  - C1 with C4 and C4 with C7 are OK but…
  - C1 with C7 is not

## *An example (from http://www.statbank.dk/FT)*






## *The agreed common terminology*

- Internal Data Warehouse (IDWH)
  - Data are best modelled using star/snowflake schemas
  - Fact tables contain microdata (1 record = 1 unit of analysis)
  - Grouping and aggregations are performed on the fly
- Statistical Dissemination Database (SDDB)
  - Data are best modelled using specifically designed data models (e.g. the Nordic Data Model)
  - Multidimensional (data-warehouse-like) navigation based on data cubes, dimensions, slice&dice, etc.
  - Fact tables (typically) contain already aggregated data, to minimize the dissemination system's response times (1 record = 1 "cell" of a dissemination table)
  - Minimal amount of aggregations performed on the fly

## *The proposed scenario*