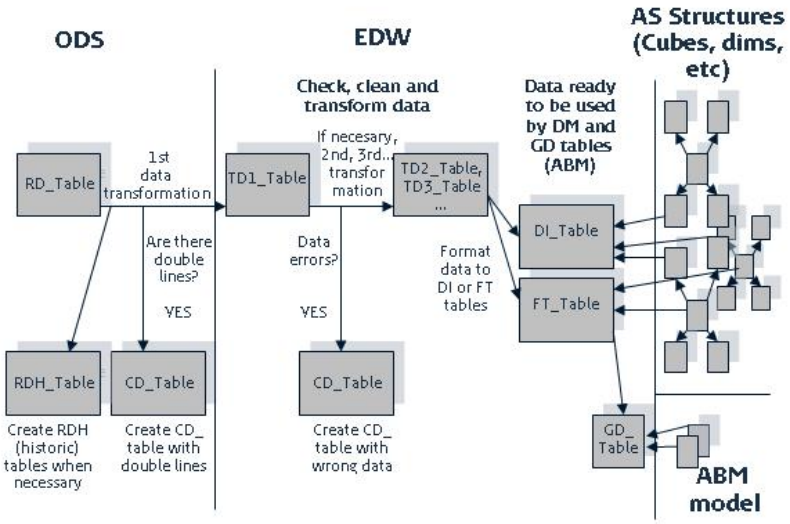
The EDW-schema

Dokumentbrugere:	Dokumentansvarlig:	Redaktør:	Dokumentnummer:	Godkendt af:	Niveau:
KONC-POEM-MT	**D-PØMQansvar**	Version: **1.1**	**12. 6. 1.07**	**PAMA2610** **08.03.2010**	

1) Introduction

This schema is called Enterprise DW because it contains information from different areas in DSB. The purpose of the EDW schema is to keep transformed data. Some data have to be transformed several times before it gets the format and content that suits the desired structure. The purpose of the EDW data tables, after being transformed, is to serve the DM database structures and the ABM model.

The following diagram shows the transformations performed to the data in the EDW schema, and its different stages. It is worth to mention, that not all the data pass through every step of the transformations, it depends on the source systems data quality and the DM structures and data needs.



The data in the EDW-schema are:

- Organised by subject (fx. Togpersonale, Litrakm, costobjekt, etc.)
- Integrated, because it contains data from the whole company or systems considered in the DW. This feature gives the possibility of crossing data from different systems and platforms across the company.
- Non-volatile because when EDW data are refreshed (with new data from the ODS schema), the historic or previous EDW data remind.
- Contain different time periods.

Data logged in this schema:

The most relevant operation to the EDW data tables, is data refresh/update. The data coming from the ODS scheme are checked, cleaned and transformed before they refresh the EDW tables. The information about EDW data refresh is logged in the table etl.Edw_TabRefLog (Notice that this table is defined in the ETL schema).

Every time a table is refreshed/updated in the EDW schema, the etl.Edw_TabRefLog table loges data about the refreshed/updated EDW table. For a full description of the table etl.Edw_TabRefLog, refer to the ETL-schema document under the title - Etl.Edw_TabRefLog.

As described before, there might be several transformations of data in this scheme, for this purpose, temporal tables are defined in this schema with the prefix TD1_, TD2, etc. There is no limit in the amount of transformations performed to the data and therefore no limit in the amount of TD tables defined. The transformation process stops only when the data is completely cleaned, checked, transformed and kept in a format that serves the DM structures and ABM model.

Processes of the EDW-schema

- The amount of historic data storage in EDW tables is defined by business process, according to their data needs and the source data availability.
- The refreshment strategy of the EDW tables is defined by table/business process, according to the availability of the data and the particularity of each system. For example, when accumulated data is refreshed in EDW tables, it is worth to consider loading the whole accumulated period and probably use partitions to speed the data refreshment. On the other hand, for slowly changing dimensions (which are described below in this

document), the new data values should be appended and the dimension values should not be deleted.

Fact Tables data refresh strategy

Data in fact tables can never be modified and the strategy to use is
- Load by period (Insert or Append data).

In the case of changes to existing data, the changes must be handled as corrections. Se document 12.6.4.

Dimension Tables data refresh strategy

The dimension tables refresh strategy depends on the treatment to the changes in dimension values and data related, especially hierarchies changes and hierarchy definition.

The simpliest dimensions without hierarchies can be refreshed using truncate/insert or insert (recommended).

The dimensions whose parent hierarchy changes must be registered (and probably FT data reaggregated), will need to be trated according to the needs, these dimension are called Slowly Changing Dimensions (SCD). Two type of SCD have been defined for this system:

- Type 1: Overwrite the parent, the description and reaggregate the data
- Type 2: Add a new row for the new dimension value (even when the dimension value name is the same), update parent for both dimension values, the old (which is not going to be used any longer) and the new one (which is the current parent). For this method to work properly, surrogate keys must be used in the dimension definition.

2) Prefixes

To identify which tables are used to keep temporal transformations and which tables are ready to serve the DM structures and ABM model, the following prefixes will be used in the tables of this schema:

Prefixes used in EDW transformations:

Prefix	Description
CD_	Table containing checking data
TD1, TD2 ...	Table containing temporal data from a first, second, third, etc. transformation
GD_	Table containing ground data in the format required by the ABM model
FT_	Table containing Fact table structure
DI_	Table containing Dimension table structure
DIH_	Table containing a dimension structure from a previous period
MD_	Table containing manual data
RT_	Table containing customized output from fact-tables
VW_	Virtual table containing af customized view of exsting data

Revisioner

Version	Godkendt	Revisions information
1	13.02.2009	
1.1	08.03.2010	