|  |
| --- |
| , RD Dep.  **introduction to dwh and etl** |
|  |



# Initial Guidelines of Star Aggregation Layer Modeling

The core objectives of building and implementing naming standard in an enterprise are:

* Business as well as technical users should be able to describe any data entity or data element just by looking at its name. Users can be internal as well as external (vendors) to the organization.
* The name decided by more than one professional for an entity or a data element should be same if they are exposed to same business and technical descriptions of the data asset.

Describing and naming data correctly is critical. If it is done right, it can help an enterprise:

* Minimize misunderstandings among business functions, which can reduce the amount of total effort needed in a BI/DW project.
* Facilitate operational efficiency and strategic use of the data.
* Reduce time to deploy new business challenges.
* Build Mapping business Dictionary to technical realizations.

The main goal of a DB objects Name convention chapter is to standardize all naming technics overall development process. Next naming standards will be described below:

* Table Naming Conventions
* Dimensions
* Slowly Changing Dimensions
* Facts Tables
* Aggregates
* Commenting

Contents

[Table Naming Conventions 4](#_Toc151045985)

[column Naming Conventions 5](#_Toc151045986)

[Rules for SCD TYPE 1 7](#_Toc151045987)

[rules for SCD TYPE 2 9](#_Toc151045990)

[Hierarchy on DM Layer 11](#_Toc151045993)

[Rules for Facts table 12](#_Toc151045994)

## Table Naming Conventions

The common Data warehouse objects for BL\_DM and BL\_3NF layers have to be named as shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| N | **Objects** | **Layer** | **Naming templates** |
| **1** | **Business Layer Cleansing Schema** | CL | **BL\_CL** |
| 2 | **Metadata tables** | CL | **MTA\_<Name in Plural>**  *MTA\_CUSTOMERS* |
| 3 | **Work tables** | CL | **WRK\_<Name in Plural>**  *WRK\_ BOOKS* |
| 4 | **Lookup tables** | CL | **LKP\_<Name in Plural>**  *LKP \_BOOK\_TYPES* |
| **5** | **Name of the source system schemas** | SA | **SA\_<SOURCE\_SYSTEM\_1\_NAME>**  separate schema for separate division of your datasets  *SA\_ASIA\_SALES* |
| **6** | External tables | SA | **EXT\_<FILE\_NAME>**  for both source systems  *EXT\_ASIA\_SALES* |
| **7** | **Source tables** | SA | SRC\_<FILE\_NAME>  for both source systems  *SRC\_ASIA\_SALES* |
| 8 | **Business Layer 3NF**  **schema name** | 3NF | **BL\_3NF** |
| 9 | **Core entity tables** | 3NF | **CE\_<Name in Plural>**  all tables should be named with prefix CE\_  *CE\_BOOKS* |
| 10 | **Entity tables SCD Type 2** | 3NF | **CE\_<Name in Plural>\_SCD**  all tables with SCD2 logic (historical data)  *CE\_PRODUCTS\_SCD* |
| 11 | **Business Layer Dimensional Model**  **schema name** | DM | **BL\_DM** |
| 12 | **Dimensions SCD Type 1 or Type 0** | DM | **DIM\_<Name in Plural>**  *DIM\_BOOKS* |
| 13 | **Dimensions SCD Type 2** | DM | **DIM\_<Name in Plural>\_SCD**  all tables with SCD 2 Type  *DIM\_PRODUCTS\_SCD* |
| 14 | **Dimensions: Calendar** | DM | **DIM\_TIME\_<Level>**  only 1 time dimension should be, based on data granularity  *DIM\_TIME\_DAY*  *DIM\_TIME\_MM* |
| 15 | **Fact tables** | DM | **FCT\_<Name in Plural>\_<Level>**  Level should be detected based on data granularity (daily/ manthly)  *FCT\_BOOKS\_DD* |

## column Naming Conventions

The columns have to match rules below:

|  |  |  |  |
| --- | --- | --- | --- |
| N | **Object category** | **Layer** | **Rules** |
| 1 | **Date Columns** | DM/3NF | Postfix:**<Column Name>\_DT**  *EVENT\_DT* |
| 2 | **All Columns**  Except KPI (Metrics) | DM/3NF | Column Attribute: **ARE NOT NULL**  NOT NULL constraint should be set for all columns, except KPI  *BOOK\_ID BIGINT NOT NULL* |
| 3 | **Time period** | DM/3NF | Use 2 columns to describe period:  these columns should be in SCD Type 2 dimension  *from* **START\_DT** *till* **END\_DT** |
| 4 | **Natural Keys**  ID’s from source systems | SA/3NF/DM | Source/3NF **<Column Name>\_ID** should be loaded accordingly as 3NF/DM **<ColumnName>\_SRC\_ID** |
| 5 | **Technical Attributes** | DM/3NF | Prefix: **TA\_<Column Name>**  Technical dates like date when data was inserted or updated  *TA\_INSERT\_DT* |
| 6 | **ID’s columns** | DM/3NF | **Use only SEQUENCES to generate IDs.**  **Do not use the SERIAL type.** |

Whitelist object abbreviations which could be used within project:

|  |  |  |  |
| --- | --- | --- | --- |
| N | **Abbreviation** | **Category** | **Description** |
| 1 | **ACT** | Postfix | **Actual Data** |
| 2 | **AVG** | Postfix | **Average KPI (Metrics)** |
| 3 | **BUS** | Prefix | **Business information** |
| 4 | **CNT** | Postfix | **Count KPI (Metrics)** |
| 5 | **DT** | Postfix | **Data & Time** |
| 6 | **EOP** | Postfix | **End of Period** |
| 7 | **GRP** | Prefix | **Group data** |
| 8 | **NO**  **NUM** | Postfix | **Number or Numeration** |
| 9 | **OPT** | Prefix | **Options** |
| 10 | **SRC** | Postfix | **Source system data** |
| 11 | **SOP** | Postfix | **Start of Period** |
| 12 | **TOT** | Postfix | **Total** |

## 

## Rules for SCD TYPE 1

The common script design for SCD Type 1:

|  |  |
| --- | --- |
| For 3NF layer | For DM layer |
| CREATE TABLE **CE\_<Name in Plural>**(  **<Name>\_ID** BIGINT NOT NULL,  **<Name>\_SRС\_ID** VARCHAR(<n>) NOT NULL,  **<Name>\_NAME** VARCHAR(<m>) NOT NULL,  **<Name>\_DESC** VARCHAR(<m>) NOT NULL,  ...  **INSERT\_DT** DATE NOT NULL,  **UPDATE\_DT** DATE NOT NULL**,**  **SOURCE\_SYSTEM** VARCHAR(<n>) NOT NULL,  **SOURCE\_ENTITY** VARCHAR(<n>) NOT NULL ); | CREATE TABLE **DIM\_<Name in Plural>**(  **<Name>\_SURR\_ID** BIGINT NOT NULL,  **<Name>\_SRC\_ID** VARCHAR(<n>) NOT NULL,  **<Name>\_NAME** VARCHAR(<m>) NOT NULL,  **<Name>\_DESC** VARCHAR(<m>) NOT NULL,  **...**  <Name>\_**<HIER>\_ID** BIGINT NOT NULL,  <Name>\_**<HIER>\_DESC** VARCHAR(<m>) NOT NULL,  **...**  **INSERT\_DT** DATE NOT NULL,  **UPDATE\_DT** DATE NOT NULL,  **SOURCE\_SYSTEM** VARCHAR(<n>) NOT NULL,  **SOURCE\_ENTITY** VARCHAR(<n>) NOT NULL ); |

Next rules are obligatory for SCD Type 1 on both 3NF and DM layers:

|  |  |  |
| --- | --- | --- |
|  | **Example** | **Description** |
| 1 | **Source\_system**  **Source\_entity**  **<Name>\_SRС\_ID** | The tables have obligatory source triplet columns to identify the whole path of current row from Sources to Target table   * Source\_system – describes the system (source) this row is loaded from. * Source\_entity (could be source\_table as well) – describes the entity this row is loaded from. Usually, entity is a table. It could also be ‘MANUAL’ if the values were inserted manually, like default row. * Source\_ID (<Name>\_SRС\_ID) – natural (business) key. Describe how the row can be identified in source entity. |
| 2 | **INSERT\_DT** DATE  **UPDATE\_DT** DATE | The tables have obligatory columns last time changed information columns:  **INSERT\_DT –** Date of row creation  **UPDATE\_DT –** Date of last time row modification |
| 3 | **-1 “n.a.” Not Applicable** | Every table has obligatory default rows:  -1 (or -98 depends on the project). A default value for the case when the data a generally sensible, but not available at the current case.  Please check “Default Rows” video to understand this point deeply. |
| 4 | <Name>**\_ID** in 3NF  <Name>**\_SURR\_ID** in DM | Should be generated by **SEQUENCES** only**.** New sequence for each table. |
| 5 | Table CE\_<Name in Plural> <Name>**\_SRC\_ID**  <Name>**\_ID**  *Example:*  *Table* ***CE\_BOOKS***  *Columns:* ***BOOK\_SRC\_ID***  ***BOOK\_ID*** | Main columns names have corresponded to table name in the singular. On both 3NF and DM layers. |

## rules for SCD TYPE 2

The common script design for SCD Type 2:

|  |  |
| --- | --- |
| For 3NF layer | For DM layer |
| CREATE TABLE **CE\_<Name in Plural>\_SCD**(  **<Name>\_ID** BIGINT NOT NULL,  **<Name>\_SRС\_ID** VARCHAR(<n>) NOT NULL,  <Name>\_**NAME** VARCHAR(<n>) NOT NULL,  <Name>\_**DESC** VARCHAR(<m>) NOT NULL,  ...  **START\_DT** DATE NOT NULL,  **END\_DT** DATE NOT NULL,  **IS\_ACTIVE** VARCHAR(1) NOT NULL,  **INSERT\_DT** DATE NOT NULL  **SOURCE\_SYSTEM** VARCHAR(<n>) NOT NULL,  **SOURCE\_ENTITY** VARCHAR(<n>) NOT NULL  ); | CREATE TABLE **DIM\_<Name in Plural>\_SCD** (  <Name>\_**SURR\_ID** BIGINT NOT NULL,  <Name>\_**SRC**\_**ID** VARCHAR(<n>) NOT NULL,  <Name>\_**NAME** VARCHAR(<n>) NOT NULL,  <Name>\_**DESC** VARCHAR(<m>) NOT NULL,  ...  <Name>\_**<HIER>\_ID** BIGINT NOT NULL,  <Name>\_**<HIER>\_DESC** VARCHAR(<m>) NOT NULL,   ...  **START\_DT** DATE NOT NULL,  **END\_DT** DATE NOT NULL,  **IS\_ACTIVE** VARCHAR(1) NOT NULL,  **INSERT\_DT** DATE NOT NULL,  **SOURCE\_SYSTEM** VARCHAR(<n>) NOT NULL,  **SOURCE\_ENTITY** VARCHAR(<n>) NOT NULL ); |

Next rules are obligatory for both 3NF and DM layers:

|  |  |  |
| --- | --- | --- |
|  | **Example** | **Description** |
| 1 | **Source\_system**  **Source\_entity**  **<Name>\_SRC\_ID** | The tables have obligatory source triplet columns to identify the whole path of current row from Sources to Target table.   * Source\_system – describes the system (source) this row is loaded from. * Source\_entity (could be source\_table as well) – describes the entity this row is loaded from. Usually, entity is a table. It could also be ‘MANUAL’ if the values were inserted manually, like default row. * Source\_ID (<Name>\_SRС\_ID) – natural (business) key. Describe how the row can be identified in source entity. |
|  | **START\_DT** DATE **END\_DT**  DATE **IS\_ACTIVE** VARCHAR(1) **INSERT\_DT** DATE | The tables have obligatory columns for application of the slow changing (SCD2) functionality:   * **START\_DT –** Start Date of period in which row values was in Active status. Default value: 01/01/1990, “DD/MM/YYYY” * **END\_DT –** End Date of period in which row values was in Active status. Default value: 31/12/9999, “DD/MM/YYYY” * **IS\_ACTIVE** –Indicator of row values status at the current “storage” day. Default values: Y / N * **INSERT\_DT –**  Date of row creation (Initial date) |
| 3 | **-1 “n.a.” Not Applicable** | Every table has obligatory default rows:  -1 (or -98 depends on the project). A default value for the case when the data a generally sensible, but not available at the current case.  Please check “Default Rows” video to understand this point deeply. |
| 4 | <Name>**\_ID** in 3NF  <Name>**\_SURR\_ID** in DM | Should be generated by **SEQUENCES** only**.** New sequence for each table. |
| 5 | Primary key **on 3NF layer** | should be COMPOSITE (<NAME>\_ID + START\_DT)  FK to this table in such case should be logical only (no need to create FK constraint in fact table **for SCD2** dimensions!). |
| 6 | Table CE\_<Name in Plural> <Name>**\_SКС\_ID**  <Name>**\_ID**  *Example:*  *Table* ***CE\_PRODUCTS\_SCD***  *Columns:* ***PRODUCT\_SRC\_ID***  ***PRODUCT\_ID*** | Main columns names have corresponded to table name in the singular. |

## Hierarchy on DM Layer

Imagine you have the following hierarchy on BL\_3NF layer:

* + CATEGORY
    - SUBCATEGORY
      * TYPE
        + ITEMS

All of them will be represented as separated CE\_ tables:

* CE\_ITEMS (Columns: item\_id, item\_name, item\_type\_id, ……)
* CE\_ITEM\_TYPES (Columns: item\_type\_id, item\_type\_name, item\_subcategory\_id, ……)
* CE\_ITEM\_SUBCATEGORIES (Columns: itemz\_subcategory\_id, item\_subcategory\_name, item\_category\_id, ……)
* CE\_ITEM\_CATEGORIES (Columns: item\_category\_id, item\_category\_name, ……)

CE\_ITEMS

CE\_ITEM\_TYPES

CE\_ITEM\_SUBCATEGORIES

CE\_ITEM\_CATEGORIES

In each table you will have at least 2 main columns: ID + Name.

When you will load data into DM layer you should load all descriptive attributes **AND ID COLUMN**.

For mentioned example you should get 1 DIM table: **DIM\_ITEMS**

Here is a list of columns which should be in **DIM\_ITEM** dimension:

CREATE TABLE **DIM\_ITEMS** (  
 ITEM**\_SURR\_ID** BIGINT NOT NULL, (gen. by sequence)   
 ITEM**\_NAME** VARCHAR(100) NOT NULL,

...  
 ITEM\_**TYPE\_ID** BIGINT NOT NULL, (from ce\_item\_types)  
 ITEM\_**TYPE\_NAME** VARCHAR(100) NOT NULL,

...  
 ITEM\_**SUBCATEGORY\_ID** BIGINT NOT NULL, (from ce..subcategories)  
 ITEM\_**SUBCATEGORY\_NAME** VARCHAR(100) NOT NULL,

...  
 ITEM\_**CATEGORY\_ID** BIGINT NOT NULL, (from ce..categories)  
 ITEM\_**CATEGORY\_NAME** VARCHAR(100) NOT NULL,   
 ...  
 INSERT\_DT DATE NOT NULL,  
 UPDATE\_DT DATE NOT NULL,

**ITEM \_SRC\_ID** VARCHAR(10) NOT NULL, (from ce\_items)

**SOURCE\_SYSTEM** VARCHAR(100) NOT NULL,

**SOURCE\_ENTITY** VARCHAR(100) NOT NULL  
  
);

## Rules for Facts table

The common script design for fact table on DM layer has look like figure below:

CREATE TABLE FCT\_SAMPLE\_MM

(

**EVENT\_DT** DATE NOT NULL,

<DIM NAME>**\_SURR\_ID** BIGINT NOT NULL,

...

**FCT\_**<NAME>\_<UNIT> BIGINT,

...

**INSERT\_DT** DATE NOT NULL,

**UPDATE\_DT** DATE NOT NULL

);

**Fact table on 3NF** should be created and named as usual CE\_ table.

Next rules are obligatory:

|  |  |  |
| --- | --- | --- |
|  | **Example** | **Description** |
| 1 | **EVENT\_DT** DATE NOT NULL | Each fact table (type: timing fact aggregation) has an obligatory column **EVENT\_DT**. This column stores the main time key – date of event.  FCT\_SAMPLE\_MM -**> MM** ->**Monthly granularity.**  **EVENT\_DT** store Months. |
| 2 | FK for SCD2 | FK between SCD2 table and fact table should be created on DM layer only! As PK for SCD2 table on DM – ID only.  Creation of DB constraint references for SCD2 tables on 3NF layer are strongly **PROHIBITED .**  *ADD CONSTRAINT fk\_fct\_books2dim\_book FOREIGN KEY (book\_id)*  *REFERENCES CE\_BOOK\_SCD*  *t\_lng\_scopes (book\_surr\_id, start\_dt);*  **NOT ALLOWED!** |
| 3 | **INSERT\_DT** DATE  **UPDATE\_DT** DATE | The tables have obligatory columns last time changed information columns:  **INSERT\_DT –** Date of fact row creation  **UPDATE\_DT –**Date of fact last time row modification |