|  |
| --- |
| EPAM Systems, RD Dep. |
| Load and Transformation |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| REVISION HISTORY | | | | | |
| Ver. | Description of Change | Author | Date | Approved | |
| Name | Effective Date |
| 1.0 | Initial status | [Valeryia\_Lupanava](mailto:Kiryl_Bucha@epam.com) | 15-NOV-2017 |  |  |

Contents

[1. Data Modelling Task 3](#_Toc498552724)

[1.1. Detailing diagrams for 3NF and Star/snowflake layers 3](#_Toc498552725)

[1.2. Visual and textual description of layers of data warehouse 3](#_Toc498552726)

[2. Data Vault 3](#_Toc498552727)

# Data Modelling Task

## Detailing diagrams for 3NF and Star/snowflake layers

* There were created seven dimensions. They are described in “Business Description #4” document
* Several dimensions have a SCD2 type. It was chosen due to It is the next dimensions:
  + dim\_customers\_scd;
  + dim\_employees\_scd;
  + dim\_payment\_methods\_scd;
  + dim\_products\_scd;
  + dim\_promotions\_scd;
* dim\_time\_day has calendar type;
* dim\_stores has a SCD1 type, because it is not necessary to have history about closed stores.

## Visual and textual description of layers of data warehouse

In this work was used a two-layer architecture.

### **Source layer**

There were used flat files and html-code as data sources.



### **Data staging layer**

The data stored to sources should be extracted, cleansed to remove inconsistencies and fill gaps, and integrated to merge heterogeneous sources into one common schema.

In this work will be used the following data staging layers step by step.

#### ***Cleansing layer***

This layer was used for data cleansing, filtering wrong data, replace missing values with singletons and performing transformations like code lookups or currency conversions. As the Staging Area, the Cleansing Area contains only data of the last delivery, and data from different sources is not integrated.

#### ***3NF layer***

Here reducing data redundancy and improving data integrity were made. Also, the process of simplifying the design of a database, so that it achieves the optimal structure composed of atomic elements, was achieved on this layer.

#### ***Dimensional model layer***

The Star schema was chosen as a dimensional model for business processes description.

The main reasons:

• Simple structure;

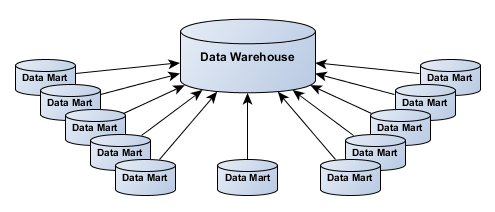
• Absence of a big number of tables to join;

• Denormalized tables are not too large in a specific case of this task;

• Widely support by a large number of business intelligence tools.

### **Data warehouse layer**

Information is stored to one logically centralized single repository: a data warehouse. The data warehouse can be directly accessed, but it can also be used as a source for creating data marts, which partially replicate data warehouse contents and are designed for specific departments. This layer is added after finishing previous steps.



### **Analysis**

In this layer, integrated data is efficiently and flexibly accessed to issue reports, dynamically analyze information, and simulate hypothetical business scenarios. Template of this layer is represented in “Retail Analysis” document.

### **Layers model of data warehouse**



# Data Vault

* Get back to the task 5 from and refactor sample business schema into the Data Vault model.