DWH – Task 10

Contents

[1. Data Modeling Task 2](#_Toc498636286)

[1.1. Detail your diagrams for 3NF and Star/snowflake layers. 2](#_Toc498636287)

[1.2. Create visual and textual description of layers of your data warehouse. 2](#_Toc498636288)

# Data Modeling Task

## Detail your diagrams for 3NF and Star/snowflake layers.

В данном проекте будет реализовано 5 dimensions:

Измерение DIM\_Time\_Day – содержит необходимые данные для дальнейшего анализа в разрезе времени – дата, день недели, название дня недели, день месяца, день года, номер месяца, названия месяца, квартал, полугодие, год. Будет применяться как измерение для дат заказов и дат оплат товаров.

Измерение DIM\_Customers\_SCD – dimension, содержащий информацию о клиентах магазина. Измерение является большим и включает более 100 000 строк.

Измерение DIM\_Products\_SCD – измерение, которое содержит необходимую информацию о товаре, бренде (производителе) и категории товара. Измерение является иерархическим dimension.

Измерение DIM\_PayDeliveries – описание способа оплаты и способа доставки. Измерение является Junk dimension.

Измерение DIM\_Points – информация о пунктах самовывоза.

DIM\_Customers\_SCD и DIM\_Products\_SCD – это медленно меняющиеся измерения 2 типа (Slowly Changing Dimensions). DIM\_PayDeliveries и DIM\_Points - измерения 1 типа.

## Create visual and textual description of layers of your data warehouse.

Наше хранилище состоит из 4 слоев:

* sa\_src – source layer. Это слой на котором находятся внешние источники информации в виду Excel файлов и external tables.
* bl\_cl – это слой временных данных, который состоит из двух основных частей – область Work и область Cleansing. В таблицы с префиксом wrk\_ данные заливаются напрямую из external tables без применения каких-либо преобразований. Далее данные очищаются, валидируются, фильтруются. Перед каждым новой загрузкой старые данные из этих таблиц будут удаляться, а загрузка будет производиться в очищенные таблицы.
* bl\_3nf – это слой постоянного хранения данных, куда загружаются данные из таблиц cls\_ в третью нормальную форму. Здесь же производится маппинг данных согласно специальной таблице. При этом устраняется аномальность, избыточность и оптимизируется производительность доступа к данным.
* bl\_dwh – это слой постоянного хранения данных, куда загружаются данные из bl\_3nf после применения агрегирования. В таблицах dim\_ и fct\_ хранятся денормализованные данные, которые позволяют значительно ускорить доступ к данным. В качестве dimensional модели была выбрана схема Star. Причиной выбора была ее простота и малое количество таблиц в сравнении со схемой Snowflake.

Все этапы загрузок и изменения данных производятся с помощью ETL. В нашем случае это будут процедуры PL/SQL.

