

|  |
| --- |
| DWH  WildBerries онлайн-магазин |
| EPBYBI17-WB17 |

| RELATED ARTIFACTS | |
| --- | --- |
| Ref. | Artifact Name |
|  |  |
|  |  |

|  |  |
| --- | --- |
| Abbreviations and Acronyms | |
|  |  |
|  |  |

Оглавление

[WildBerries онлайн-магазин 1](#_Toc498647351)

[1 Описание предметной области 5](#_Toc498647352)

[1.1 Бизнес информация 5](#_Toc498647353)

[1.2 Недостатки существующего управления данными 5](#_Toc498647354)

[1.3 Преимущества внедрения хранилища данных 5](#_Toc498647355)

[2 Бизнес измерения 6](#_Toc498647356)

[2.1 Описание бизнес-процесса 6](#_Toc498647357)

[2.2 Зерно. 6](#_Toc498647358)

[2.3 Измерения. 6](#_Toc498647359)

[2.4 Факты. 8](#_Toc498647360)

[2.5 Источники информации. 9](#_Toc498647361)

[3 Логическая схема 10](#_Toc498647362)

[3.1 Слой доступа к данным DWH - схема Звезда 10](#_Toc498647363)

[3.2 Фундаментальный слой DWH - схема 3 NF 11](#_Toc498647364)

[4 Модель потоков данных 18](#_Toc498647365)

[4.1 Типы измерений 18](#_Toc498647366)

[4.2 Слои данных DWH 18](#_Toc498647367)

[5 Стратегия партиционирования таблицы фактов 20](#_Toc498647368)

[5.1 Партиционирование 20](#_Toc498647369)

[5.2 Бизнес применение 20](#_Toc498647370)

[6 Стратегия пареллельной загрузки 22](#_Toc498647371)

[7 Макеты отчетов 23](#_Toc498647372)

# Описание предметной области

## Бизнес информация

Хранилище данных будет реализовано для нужд белорусского online-магазина Wildberries. Wildberries – это один из крупнейших интернет-магазинов модной одежды, обуви, аксессуаров, косметических средств, существующий уже более 12 лет. Магазин насчитывает более 7 тысяч брендов на сайте и более 100 тысяч заказов в день. Из года в год он развивается, расширяет географию и улучшает качество обслуживания!

## Недостатки существующего управления данными

С каждым днем количество товаров, отделений магазина и клиентов растет в геометрической прогрессии. Это требует пересмотра обработки данных. На данный момент база данных приложения недостаточно производительна, чтобы отрабатывать сложные запросы, поэтому на получение одного простейшего отчета, с которым хорошо построенное DWH справляется за считанные секунды, уходят часы. В такой ситуации многие запросы управляющего сектора не находят отклика, нет возможности быстрого реагирования на изменение рынка, реактивного изменения стратегий продаж, открытия новых. В результате теряется скорость принятия решений, теряются деньги, растет недоверие к ИТ решениям.

## Преимущества внедрения хранилища данных

Внедрение DWH поможет навести порядок и упростит использование данных для анализа. Сложные отчеты будут быстро формироваться, а, следовательно, не будут нагружать приложения, будут чаще использоваться и потому помогут в развитии бизнеса. Увеличится скорость реагирования на изменения рынка, что приведет к принятию стратегически важных решений, улучшится качество обслуживания, увеличится количество клиентов и прибыли соответственно.

Расширение бизнеса приведет к увеличению количества рабочих мест, как в самом онлайн магазине, так и у поставщиков и производителей приобретаемой продукции.

# Бизнес измерения

## Описание бизнес-процесса

Учет продажи товаров различных марок в интернет-магазине Wildberries с учетом различных показателей.

## Зерно

Зерном будет одна продажа конкретного товара одной торговой марки из одной категории одному клиенту в одном пункте самовывоза в пределах одного заказа.

## Измерения

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.

CREATE TABLE "DIM\_Time\_Day"

( "DATE\_ID" NUMBER(8) NOT NULL,

"FULL\_DATE" DATE,

"DAY\_WEEK" NUMBER(1),

"DAY\_WEEK\_DECS" VARCHAR2(50),

"DAY\_MONTH" NUMBER(2),

"DAY\_YEAR" NUMBER(3),

"MONTH\_NUM" NUMBER(2),

"MONTH\_DESC" VARCHAR2(50),

"QUARTER" NUMBER(1),

"HALF\_YEAR" NUMBER(1),

"YEAR" NUMBER(4));

Измерение DIM\_Customers\_SCD – dimension, содержащий информацию о клиентах магазина. Измерение является большим dimension, который включает более 100 000 строк. Это медленно изменяющееся измерение 2 типа (Slowly Changing Dimensions).

CREATE TABLE "DIM\_Customers\_SCD"

( "CUSTOMER\_ID" NUMBER(8) NOT NULL,

"CUSTOMER\_CODE" NUMBER(7),

"CUSTOMER\_LAST\_NAME" VARCHAR2(50),

"CUSTOMER\_FIRST\_NAME" VARCHAR2(50),

"CUSTOMER\_MIDDLE\_NAME" VARCHAR2(50),

"CUSTOMER\_AGE" VARCHAR2(3),

"CUSTOMER\_GENDER" VARCHAR2(1),

"CUSTOMER\_PERSONAL\_DISCOUNT" NUMBER(2),

"CUSTOMER\_LOCATION\_TYPE" VARCHAR2(50),

"CUSTOMER\_LOCATION" VARCHAR2(50),

"CUSTOMER\_DISTRICT" VARCHAR2(50),

"CUSTOMER\_REGION" VARCHAR2(50),

"CUSTOMER\_START\_DT" DATE,

"CUSTOMER\_END\_DT" DATE,

"CUSTOMER\_IS\_ACTIVE" VARCHAR2(1),

"CUSTOMER\_INSERT\_DT" DATE);

Измерение DIM\_Products\_SCD – измерение, которое содержит необходимую информацию о товаре, бренде (производителе) и категории товара. Измерение является иерархическим dimension. Это медленно изменяющееся измерение 2 типа (Slowly Changing Dimensions).

CREATE TABLE "DIM\_Products\_SCD"

( "PRODUCT\_ID" NUMBER(8) NOT NULL,

"PRODUCT\_CODE" NUMBER(9),

"PRODUCT\_DESC" VARCHAR2(50),

"PRODUCT\_BRAND" VARCHAR2(100),

"PRODUCT\_TYPE" VARCHAR2(100),

"PRODUCT\_SUBCATEGORY" VARCHAR2(150),

"PRODUCT\_CATEGORY" VARCHAR2(50),

"PRODUCT\_START\_DT" DATE,

"PRODUCT\_END\_DT" DATE,

"PRODUCT\_IS\_ACTIVE" VARCHAR2(1),

"PRODUCT\_INSERT\_DT" DATE);

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

CREATE TABLE "DIM\_PayDeliveries"

( "PAYDELIVERY\_ID" NUMBER(2) NOT NULL,

"DELIVERY\_DESC" VARCHAR2(50),

"PAYOPTION\_DESC" VARCHAR2(50));

Измерение DIM\_Points – информация о пунктах самовывоза.

CREATE TABLE "DIM\_Points"

( "POINT\_ID" NUMBER(3) NOT NULL,

"POINT\_CODE" VARCHAR2(3),

"POINT\_LOCATION" VARCHAR2(50),

"POINT\_REGION" VARCHAR2(50),

"POINT\_DISTRICT" VARCHAR2(50));

## Факты

Таблица фактов будет содержать количество проданных единиц товара, сумму без скидки, сумму предоставленной скидки и итоговую сумму реализации за месяц в разрезе товара, бренда, категории товара, клиента, способа доставки и оплаты. Содержит дегенеративное измерение – Order\_ID.

CREATE TABLE "FCT\_SalesItems"

( "ORDERDATE\_ID" NUMBER(8) NOT NULL,

"ORDER\_ID" NUMBER(10,2),

"CUSTOMER\_ID" NUMBER(8) NOT NULL,

"PAYDATE\_ID" NUMBER(8) NOT NULL,

"PRODUCT\_ID" NUMBER(8) NOT NULL,

"PAYDELIVERY\_ID" NUMBER(2) NOT NULL,

"POINT\_ID" NUMBER(3),

"QUANTITY" NUMBER(10),

"ITEM\_SUM" NUMBER(35,2),

"DISCOUNT\_SUM" NUMBER(35,2),

"TOTAL\_SUM" NUMBER(35,2));

## Источники информации

Основная информация о товарах, категориях, брендах, способах доставки, оплаты и пунктах самовывоза взята на официальном сайте: <https://www.wildberries.by/>

<https://www.wildberries.by/services/sposoby-oplaty>

<https://www.wildberries.by/services/besplatnaya-dostavka#pickupPoints>

<https://www.wildberries.by/catalog/0/brand.aspx>

Населенные пункты РБ - <https://www.bygeo.ru/materialy/naselenie-belarusi/>

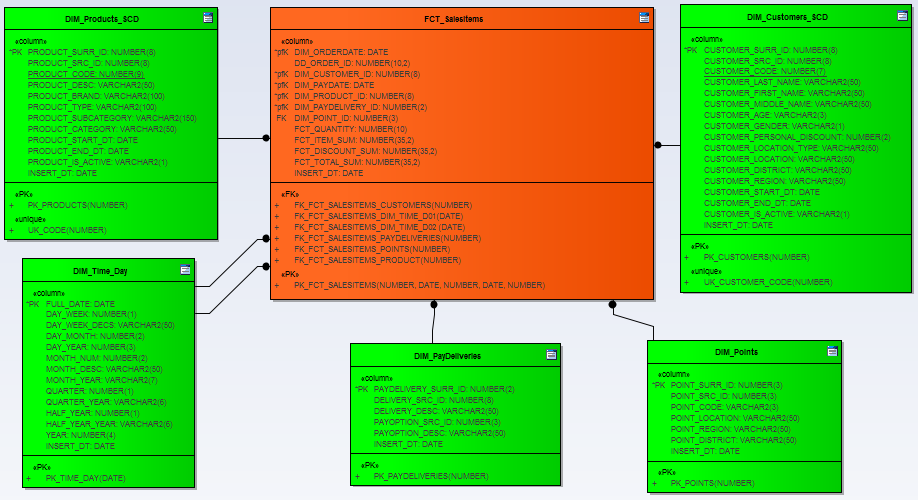
Названия улиц - <https://ato.by/streets>.

Какая-то часть данных будет взята с сайтов, какая-то сгенерирована c помощью других интернет сайтов, например http://www.fakenamegenerator.com, также будет применена генерация с помощью PLSQL и поиск, копирование и вставка данных вручную.

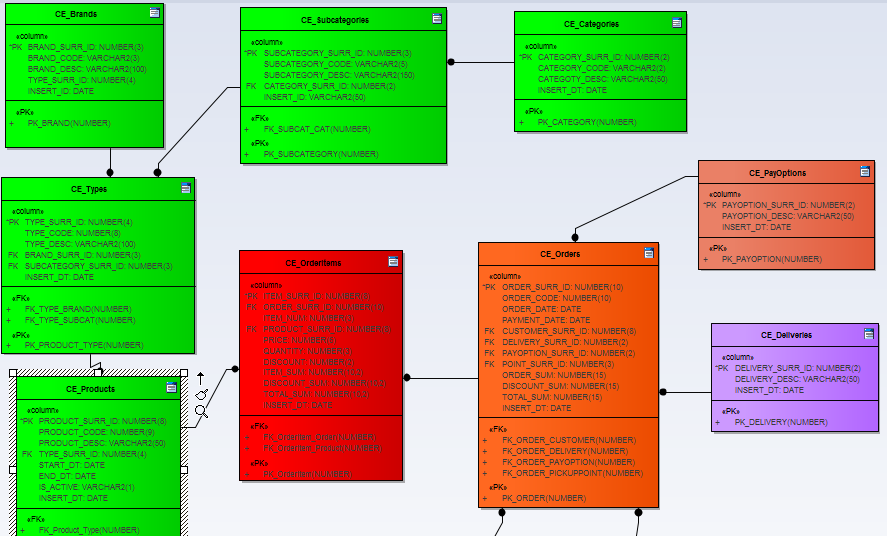
# Логическая схема

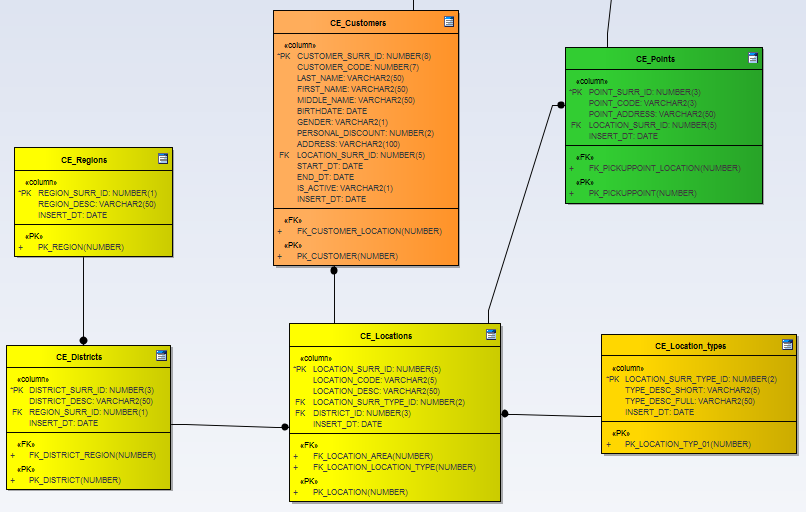
## Слой доступа к данным DWH - схема Звезда

Все связи на слое доступа к данным между измерениями и таблицей фактов реализованы как один ко многим.



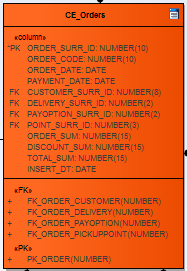
## Фундаментальный слой DWH - схема 3 NF





Одна из основных таблиц CE\_Orders содержит в себе информацию о заказе, его номер, дата заказа, дата оплаты, идентификатор клиента, совершившего данный заказ, идентификатор типа оплаты, идентификатор типа доставки и идентификатор пункта самовывоза, а также сумму заказа, сумму предоставленной скидки и общая сумма, оплаченная клиентом.

Один заказ Order\_Code может быть сделан одним клиентом Customer\_Surr\_Id в один календарный день Order\_Date. Он может быть оплачен с помощью одного из вариантов PayOption\_Surr\_Id и может быть доставлен курьером, почтой или получен в пункте самовывоза Delivery\_Surr\_Id , в третьем случае указывается какой именно пункт Point\_Surr\_Id. Фиксируются две даты – дата заказа Order\_Date и дата оплаты Payment\_Date.



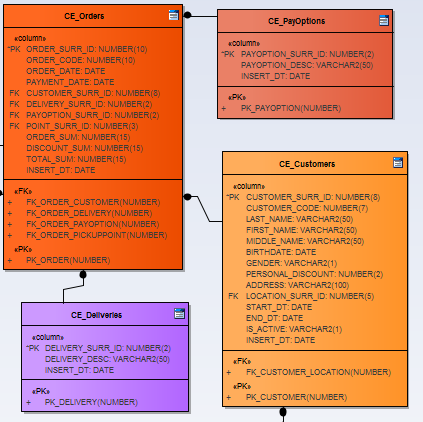


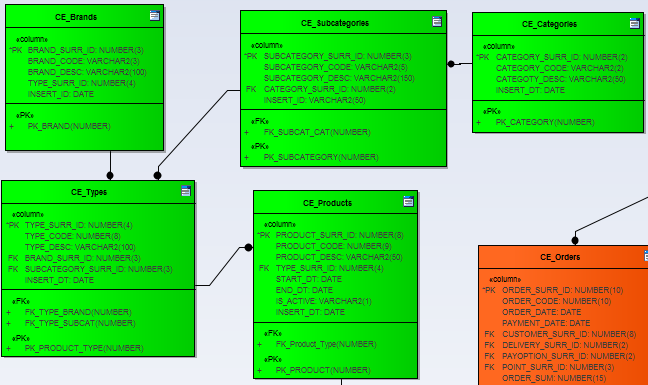
Таблица CE\_OrderItems содержит в себе информацию о пунктах заказа, идентификатор заказа, идентификатор товара, идентификатор бренда товара, а также цену товара, его количество, процент скидки (суммируются скидка клиента и разовые скидки), сумму за указанный товара, сумму скидки и итоговую сумму к оплате.

Один пункт заказа Item\_Num относится только к одному заказу Order\_Surr\_Id. Один заказ может содержать N пунктов. Пункт относится исключительно к одному товару Product\_Surr\_Id и соответственно только к 1 бренду.

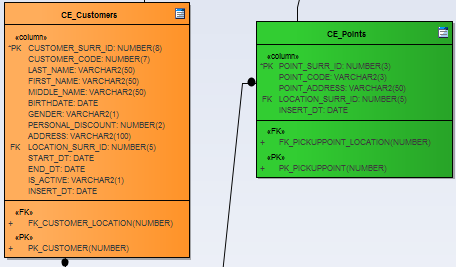


Пример данных для ветки товаров:

* Таблица CE\_Product - Product Desc: Платье Bracegirdle, Product Code: 2120550 🡪
* Таблица CE\_Type – Type Desc: [Платья-миди](https://www.wildberries.by/catalog/zhenshchinam/odezhda/platya-midi) 🡪
* Таблица CE\_Brand – Brand Desc: ONLY 🡪
* Таблица CE\_Subcategory: Subcategory Desc - Платья 🡪
* Таблица CE\_Category: Category Desc - Женская одежда



Пункты самовывоза и клиенты имеют location\_Surr\_Id, который связывает их с таблицей CE\_locations.



Поскольку доставка производится по Беларуси, то это отражается в формировании ветки локаций. Пример для понимания:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Location\_Code | Location\_Desc | District\_Desc | Region\_Desc | Type\_desc\_short |
| 6240833091 | Нарочь | Мядельский | Минская область | аг. |

# Модель потоков данных

## Типы измерений

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.

## Слои данных DWH

Наше хранилище состоит из 4 слоев:

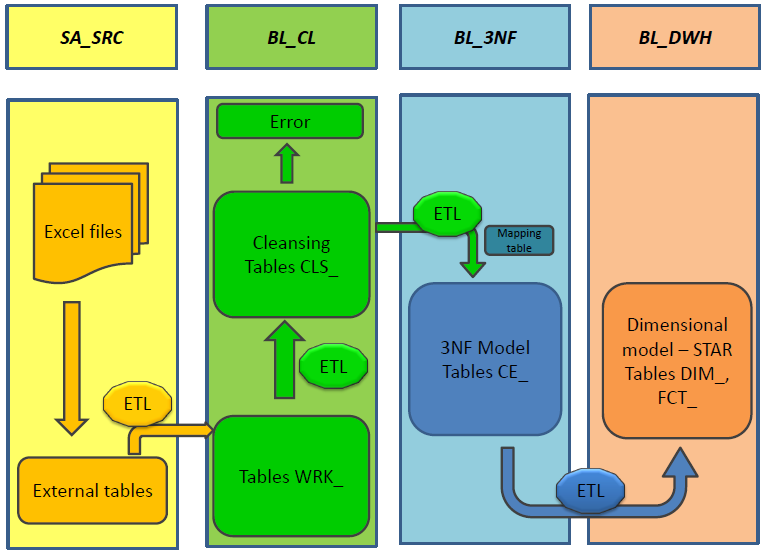
sa\_src – source layer. Это слой на котором находятся внешние источники информации в виду Excel файлов и external tables.

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.



# Стратегия партиционирования таблицы фактов

## Партиционирование

Для ускорения доступа к данным и скорости выполнения запросов используется Composite Partitioning для таблицы фактов FCT\_SalesItems. Вначале таблица разделяется методом Range на партиции по годам продаж. Затем каждая партиция разбивается на List субпартиции по категориям и подкатегориям продуктов. Удобно для отслеживания продаж конкретных категорий товаров по годам. Так, например, первая партиция будет содержать все продажи за 2010 год и будет разделена на 3 субпартиции – женская одежда, мужская одежда и детская одежда, вторая – за 2011 год и разделена на такие же партиции, и т.д.

## Бизнес применение

Я буду анализировать ежемесячные и годовые продажи товаров по подкатегориям и категориям, в разрезе клиентов или пунктов самовывоза, по регионам и другим доступным критериям. Ниже приведены примеры возможных вариантов отчетов.

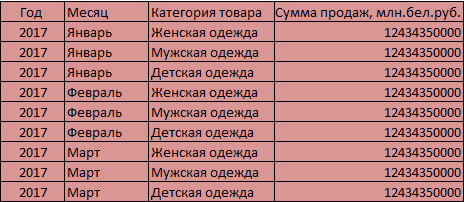
Отчет о ежемесячных продажах товаров Платья из категории Женская одежда за 2017 год.



Отчет о продажах мужской обуви за январь и февраль 2017 года в разрезе областей.



Отчет о совокупных продажах товаров по основным категориям в разрезе месяцев за 1 квартал 2017 года.



# Стратегия пареллельной загрузки

# Макеты отчетов

1. Example of 1st lvl appendix

This is Appendix A.

* 1. Example of 2nd lvl appendix

This is Appendix A.1

| REVISION HISTORY | | | | | |
| --- | --- | --- | --- | --- | --- |
| Ver. | Description of Change | Author | Date | Approved | |
| Name | Effective Date |
| 1.1. | Начальная версия 2.0 | Марина Глазунва | 17-ноя-2017 |  |  |
|  |  |  |  |  |  |