|  |
| --- |
|  |
| DWH Project |



Contents

[1. Business Description 3](#_Toc498274127)

[1.1. Business background 3](#_Toc498274128)

[1.2. Problems because of poor data management 3](#_Toc498274129)

[1.3. Benefits from implementing a Data Warehouse 3](#_Toc498274130)

[2. Dimensions of a Business 4](#_Toc498274131)

[2.1. Бизнес-процесс. 4](#_Toc498274132)

[2.2. Зерно. 4](#_Toc498274133)

[2.3. Измерения. 4](#_Toc498274134)

[2.4. Факты. 5](#_Toc498274135)

[2.5. Источники информации. 6](#_Toc498274136)

[2.6. Слой доступа к данным DWH - схема Звезда 6](#_Toc498274137)

[2.7. Фундаментальный слой DWH - схема 3 NF 7](#_Toc498274138)

[3. Logical Scheme 10](#_Toc498274139)

[4. Data Flow 10](#_Toc498274140)

[5. Fact Table Partitioning Strategy 10](#_Toc498274141)

[6. Strategy of Parallel Load 10](#_Toc498274142)

[7. Report Layouts 10](#_Toc498274143)

# Business Description

## Business background

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!

## Problems because of poor data management

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.

## Benefits from implementing a Data Warehouse

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.

# Dimensions of a Business

## Бизнес-процесс.

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 6 dimensions:

Измерение DatesDim – 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 "DatesDim"

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

"FULL\_DATE" DATE,

"DAY\_WEEK" NUMBER(1),

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

"DAY\_MONTH" NUMBER(2),

"DAY\_YEAR" NUMBER(3),

"MONTH\_NUM" NUMBER(2),

"MONTH\_NAME" VARCHAR2(50),

"QUARTER" NUMBER(1),

"HALF\_YEAR" NUMBER(1),

"YEAR" NUMBER(4));

Измерение CustomersDim – dimension, содержащий информацию о клиентах магазина. Это большой dimension, который включает более 100 000 строк.

CREATE TABLE "CustomersDim"

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

"CUSTOMER\_CODE" NUMBER(7),

"LAST\_NAME" VARCHAR2(50),

"FIRST\_NAME" VARCHAR2(50),

"MIDDLE\_NAME" VARCHAR2(50),

"AGE" VARCHAR2(3),

"GENDER" VARCHAR2(50),

"PERSONAL\_DISCOUNT" NUMBER(2),

"CITY" VARCHAR2(50),

"REGION" VARCHAR2(50));

Измерение ProductsDim – измерение, которое содержит необходимую информацию о товаре, бренде (производителе) и категории товара. Является иерархическим dimension.

CREATE TABLE "ProductsDim"

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

"PRODUCT\_VENDOR\_CODE" NUMBER(9),

"PRODUCT\_TYPE" VARCHAR2(50),

"PRODUCT\_CATEGORY" VARCHAR2(50));

Измерение PayDeliveriesDim – описание способа оплаты и способа доставки. Идет как Junk dimension.

CREATE TABLE "PayDeliveriesDim"

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

"DELIVERY\_NAME" VARCHAR2(50),

"PAYOPTION\_NAME" VARCHAR2(50));

Измерение PickupPointsDim – информация о пунктах самовывоза.

CREATE TABLE "PickupPointsDim"

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

"PP\_CODE" VARCHAR2(3),

"PP\_ADDRESS" VARCHAR2(50),

"PP\_CITY" VARCHAR2(50),

"PP\_REGION" VARCHAR2(50));

Измерение BrandsDim – информация о брендах товаров.

CREATE TABLE "BrandsDim"

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

"BRAND\_NAME" VARCHAR2(50));

## Факты.

Таблица фактов будет содержать количество проданных единиц товара и сумму реализации за месяц в разрезе товара, бренда, категории товара, клиента, способа доставки и оплаты.

CREATE TABLE "SalesItemsFacts"

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

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

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

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

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

"PICKUPPOINT\_ID" NUMBER(3),

"BRAND\_ID" NUMBER(8),

"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>

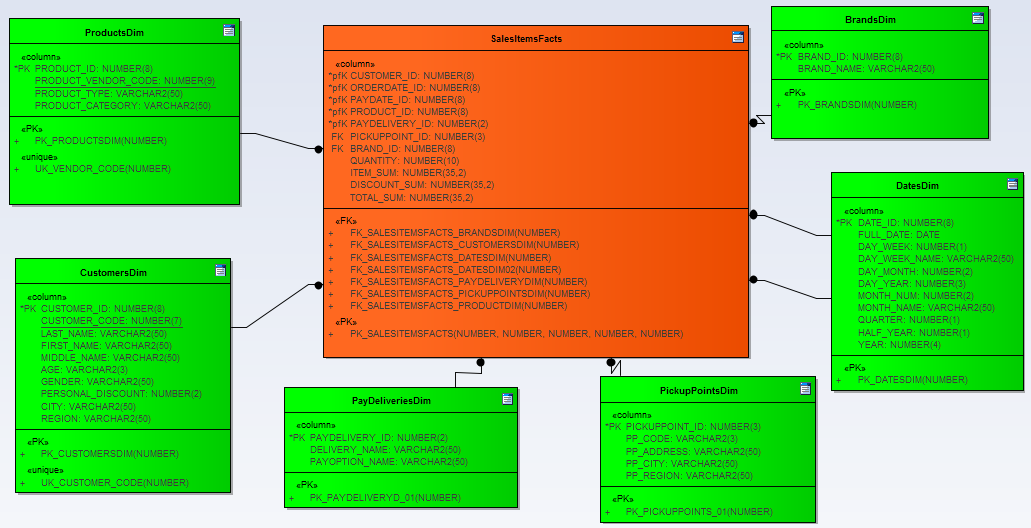
<https://www.wildberries.by/catalog/0/brand.aspx>

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

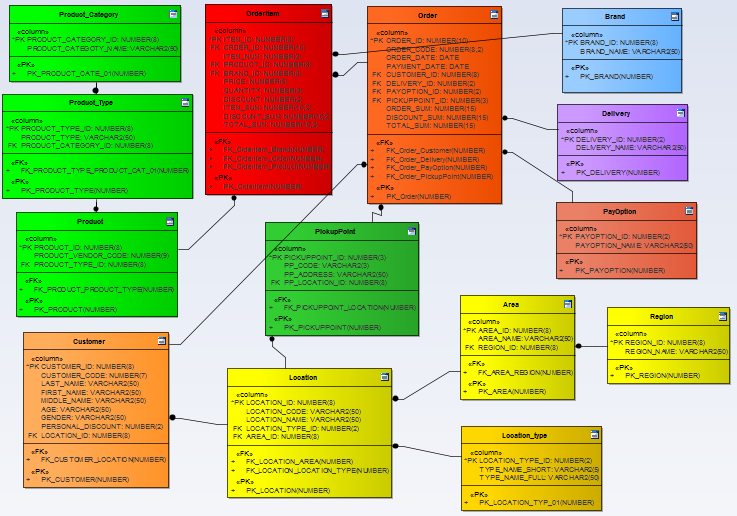
Какая-то часть данных будет взята с сайтов, какая-то сгенерирована c помощью других интернет сайтов, например http://www.fakenamegenerator.com, также будет применена генерация с помощью PLSQL и вставка данных вручную.

## Слой доступа к данным DWH - схема Звезда

Все связи на Access layer между dimensions и таблицей фактов реализованы как один ко многим.

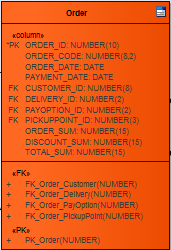


## Фундаментальный слой DWH - схема 3 NF



Одна из основных таблиц Order 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\_Id в один календарный день Order\_Date. Он может быть оплачен с помощью одного из вариантов PayOption\_id и может быть доставлен курьером, почтой или получен в пункте самовывоза Delivery\_Id , в третьем случае указывается какой именно пункт PickupPoint\_Id. Фиксируются две даты – дата заказа Order\_Date и дата оплаты Payment\_Date.



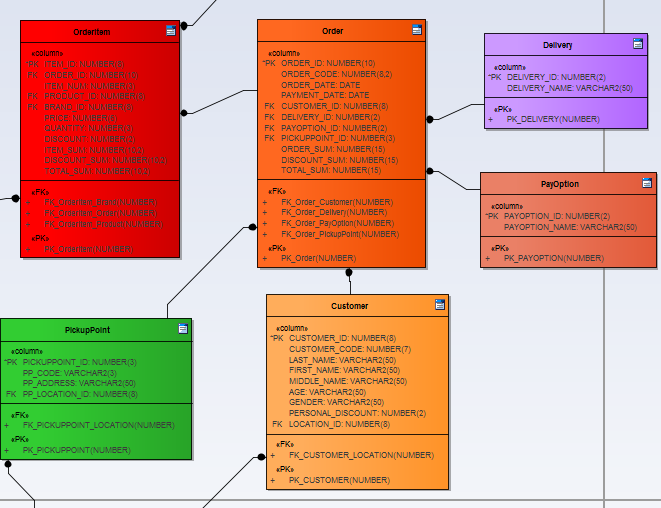
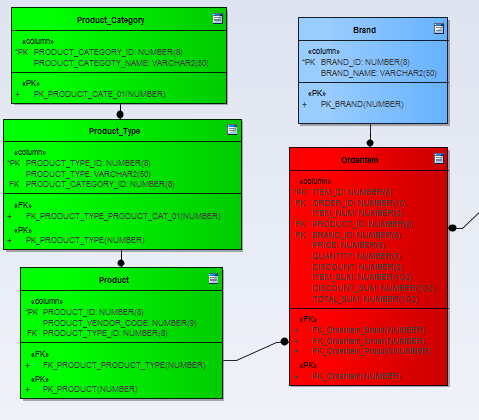


Таблица OrderItem содержит в себе информацию о пунктах заказа, идентификатор заказа, идентификатор товара, идентификатор бренда товара, а также цену товара, его количество, процент скидки (суммируются скидка клиента и разовые скидки), сумму за указанный товара, сумму скидки и итоговую сумму к оплате.

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



Для понимания приведу пример данных для ветки товаров: *Product* Платье Bracegirdle, артикул: 2120550 🡪 *Product\_Type* [Платья-миди](https://www.wildberries.by/catalog/zhenshchinam/odezhda/platya-midi) 🡪 *Product\_Category* Платья



Пункты самовывоза и клиенты имеют location\_id, который связывает их с таблицей location. Поскольку доставка производится по Беларуси, то это отражается в формировании ветки локаций. Пример для понимания:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Location\_Code | Location\_Name | Area\_Name | Region\_Name | Type\_name\_short |
| 6240833091 | Нарочь | Мядельский | Минская область | аг. |

# Logical Scheme

# Data Flow

# Fact Table Partitioning Strategy

# Strategy of Parallel Load

# Report Layouts