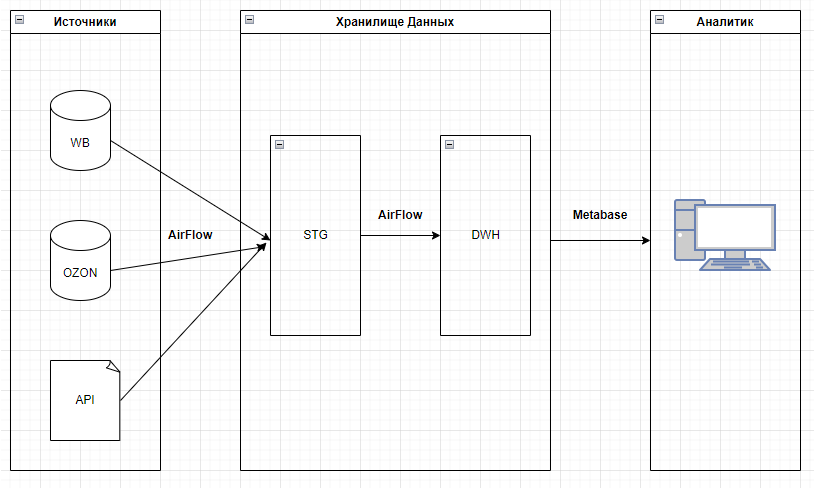
**Архитектура и бизнес-требования к хранилищу данных текстильного магазина.**

Хранилище данных – единая система хранения данных о отгрузках и реализации товара текстильного магазина. ХД необходим для сбора информации с различных систем-источников в единое целевое хранилище с возможностью дальнейшего расширения. Добавлению новых источников и визуализации собранных данных для стратегического анализа.

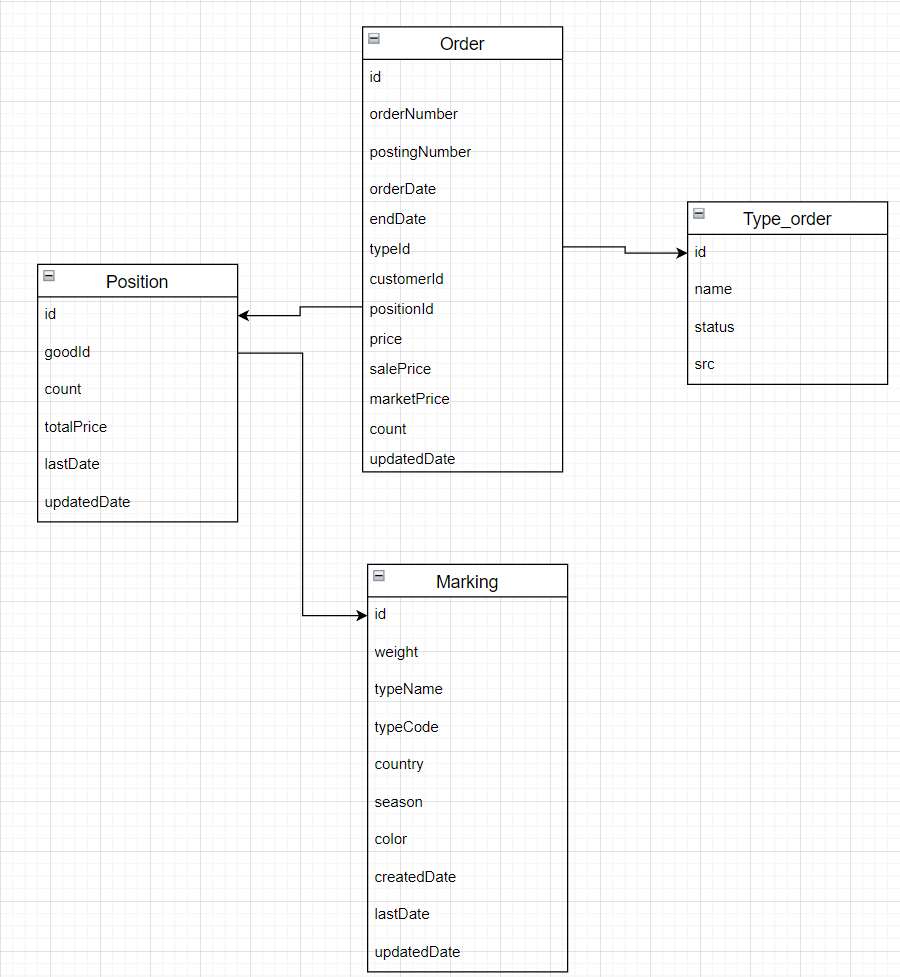
В качестве источников используется 2 БД и API: WB и OZON с данными о реализации товара и данные о поступление товара на реализацию. Хранилище реализовано на БД ClickHouse. Оркестратором служит AirFlow. В нем же реализована логика преобразования и проверки на качество данных. К примеру, ссылочная целостность данных, сумма сделки. BI-решением для визуализации данных выступает Metabase.

Архитектурная диаграмма хранилища:



ХД имеет два слоя данных: STG и DWH. Данные из источников переливаются на stg слой без изменений с добавлением поля вставки данных. Каждой таблице на источнике должна соответствовать одной таблице на STG слое. Из STG слоя данные прогружаются в DWH слой с заложенной бизнес-логикой.

Схема слоя DWH:



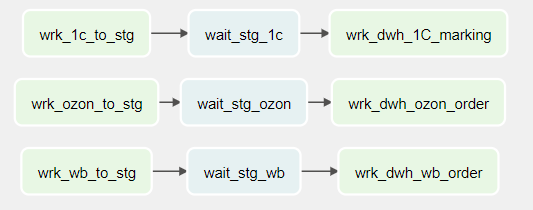
Справочники заполняются в двух режимах: автоматически и вручную. И имеют дефолтное значение - “Неопределенно”, которым заменяется NULL.

Основная таблица продаж(order) должна содержать отдельный строки с продажей/отгрузкой и отменой заказа в разбивке по товарам. В дальнейшем на основе этих данных можно рассчитать агрегированные данные. S2T между слоями STG и DWH для таблицы order, для 5 таблиц-источников.

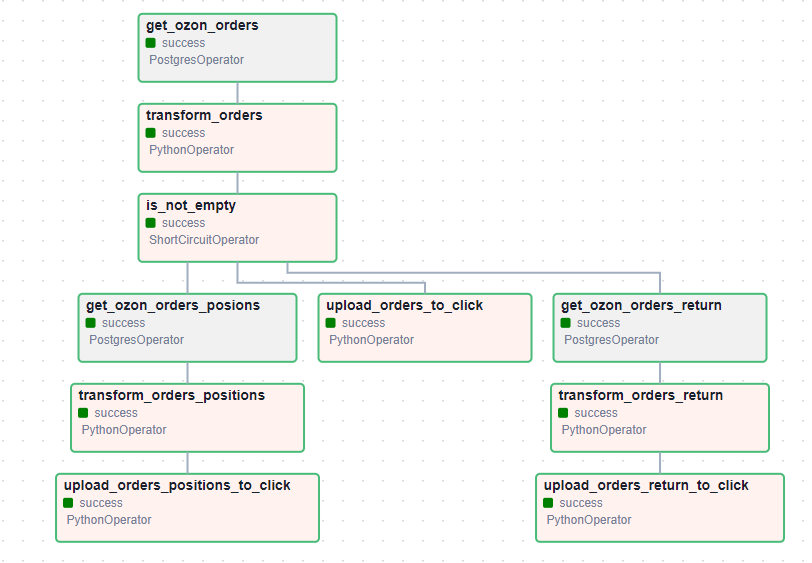
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SRC | sourse\_table | sourse \_column | target\_table | target\_column | PK | Описание |
| - | - | - | order | id | Y | Уникальный хэш(orderNumber, postingNumber, goodId) |
| OZON | ozon\_order | orderNumber | order | orderNumber | N | Номер заказа |
| OZON | ozon\_order | postingNumber | order | postingNumber | N | Номер позиции |
| OZON | ozon\_order | fromD | order | orderDate | N | Дата заказа |
| OZON | ozon\_order | toD | order | endDate | N | Дата завершения заказа |
| OZON | ozon\_order | ozonStatus | order | typeId | N | Статус заказа.FK для type\_order |
| OZON | ozon\_order | customerId | order | customerId | N | Клиент |
| OZON | ozon\_order\_position | goodId | order | positionId | N | Товар. FK для position |
| OZON | ozon\_order\_position | totalPrice | order | price | N | Цена магазина |
| OZON | ozon\_order\_position | salePrice | order | salePrice | N | Размер скидки |
| OZON | ozon\_order\_position | marketPrice | order | marketPrice | N | Итоговая цена с учетом скидки площадки |
| OZON | ozon\_order\_position | count | order | count | N | Кол-во товара |
| - | - | - | order | updateDate | N | Дата изменения |
| - | - | - | order | id | Y | Уникальный хэш(orderNumber, postingNumber, goodId, returnid) |
| OZON | ozon\_order | orderNumber | order | orderNumber | N | Номер заказа |
| OZON | ozon\_order /ozon\_order\_return | orderNumber/ returnId | order | postingNumber | N | Номер позиции |
| OZON | ozon\_order\_return | Date(updateDate) | order | orderDate | N | Дата заказа |
| OZON |  | NULL | order | endDate | N | Дата завершения заказа |
| OZON |  |  | order | typeId | N | Статус заказа.FK для type\_order |
| OZON | ozon\_order | customerId | order | customerId | N | Клиент |
| OZON | ozon\_order\_return | goodId | order | positionId | N | Товар. FK для position |
| OZON | ozon\_order\_position | totalPrice | order | price | N | Цена магазина |
| OZON | ozon\_order\_position | salePrice | order | salePrice | N | Размер скидки |
| OZON | ozon\_order\_return | marketPrice | order | marketPrice | N | Итоговая цена с учетом скидки площадки |
| OZON | ozon\_order\_return | count | order | count | N | Кол-во товара |
| - | - | - | order | updateDate | N | Дата изменения |
| - | - | - | order | id | Y | Уникальный хэш(orderNumber, postingNumber, goodId) |
| WB | wb\_order | orderNumber | order | orderNumber | N | Номер заказа |
| WB | wb\_order | postingNumber | order | postingNumber | N | Номер позиции |
| WB | wb\_order | fromD | order | orderDate | N | Дата заказа |
| WB | wb\_order | toD | order | endDate | N | Дата завершения заказа |
| WB | wb\_order | wbStatus | order | typeId | N | Статус заказа.FK для type\_order |
| WB | wb\_order | customerId | order | customerId | N | Клиент |
| WB | wb\_order\_position | goodId | order | positionId | N | Товар. FK для position |
| WB | wb\_order\_position | totalPrice | order | price | N | Цена магазина |
| WB | wb\_order\_position | salePrice | order | salePrice | N | Размер скидки |
| WB | wb\_order\_position | marketPrice | order | marketPrice | N | Итоговая цена с учетом скидки площадки |
| WB | wb\_order\_position | count | order | count | N | Кол-во товара |
| - | - | - | order | updateDate | N | Дата изменения |

Каждый DAG должен брать данные, которые изменялись или добавлялись после последнего успешно отработанного DAG-а. После прогрузки основной таблицы от источника должна производиться прогрузка связанных таблиц. Данные, вставляемые в базу, не должны дублироваться. Если новые данные отсутствуют, то процесс можно считать успешно завершенным и не запускать прогрузку в таблицу и связанные задачи.

Зависимость DAG-ов:



Зависимость задач для wrk\_ozon\_to\_stg:



В данной конфигурации хранится последняя версия пришедших данных. В дальнейшем можно расширить до хранения полной истории.