# Вопросы

1. Где содержится информация, что Система 1 (справочник товар – категория) обновилась?
2. Можно ли зацепиться за ETL поток, который обновляет Система 1, чтобы потом по триггеру запускать следующий поток?
3. Естественный ключ Системы 1 (справочник товар – категория), это номер товара? Или же это номер товара + название?
4. Не совсем понятно, что значит высказывание в отношении возможностей окна: «категоризация должна быть в актуальном состоянии»? Требуются уточнения!!! А так как их нет, придется принимать ответственность за то, как считаем корректным проводить аналитику, после сбиться с аналитиками, как реально надо.
5. Для чего дата в источнике Система 1 – таблица фактических продаж относится к части естественного ключа таблицы? Какой смысл несет данное нагромождение ключа.

# Допущения

1. Принимаем, что ключ в Системе 1 (Справочник товар – категория) это «**Номер товара**». \**Возможный вариант*\*: «**Номер товара**» + «**Название товара**», пример:

|  |  |  |  |
| --- | --- | --- | --- |
| Номер товара | Название товара | Номер категории | Название категории |
| 1001 | яблоки 1 кг | 1 | Фрукты |
| 1001 | яблоки 3 кг | 1 | Фрукты |

1. Принимаем тот факт, что в Системе 1 (Справочник товар – категория) атрибут «**Название категории**» не может быть частью естественного ключа, а является частью измерения. Так принято, в соответствии с описанием источника.
2. Принимаем, что в разрезе дат аналитика строится так: товар в заданную дату, соответствует той категории, которая у него была на заданную дату. То есть, если у него [2024.01.01-2024.06.30] была категория «Спорт», а после этого и по настоящее время является категория «Обувь», тогда для аналитика до 2024.06.30 будет рассматриваться этот товар как категория «Спорт», а далее строиться как «Обувь». То-есть для бизнеса ставим цель, что важно сохранить исторический контекст — не пересчитывать, а фиксировать изменения с определённой даты. \*Возможный вариант\*: рассматривать только актуальную категорию по товару для аналитики за весь 2024-ый год, возможно такая задача реальна, как ad-hoc (Возможно именно это и имелось в виду в разделе «Будущая визуализация» в задании), также если важна точность для отчетности и прогнозирования на новой категории, при изменении аналитики в отношении метрики, то следует учитывать актуальную дату и пересчитать всё.
3. Система 1 (Таблица фактических продаж) содержит информацию по всем продажам компании, однако в описании источника и ТЗ нигде не говорится о том, где хранится информация о возвратах и нужно ли ее учитывать при построении дашборда, что было бы логично. В соответствии с имеющейся информацией принимаем допущение, что товары невозможно вернуть обратно. \*Возможные вариант – 1) В таблице фактов будет строка с другой датой временем, но тем же номером заказа и номером товара, где сумма будет уже отрицательной, в таком случае, необходима будет дополнительная агрегация при реализации построения витрины. 2) Будет дополнительная таблица по возвратам, где будет содержаться информация по возвратам. В соответствии с этой таблицей тогда придется корректировать данные, которые будут поступать на дашборд.
4. Принимаем следующее допущение, что атрибут «**Дата**» в Системе 1 – таблица фактических продаж, является частью естественного ключа таблица, лишь из-за того, что это ключ партиции, в которой содержится запись по «**Номер заказа**» + «**Номер продукта**». \*Возможный вариант\* существует еще метрики / атрибуты в источнике по фактическим продажам, которые зависят от даты, поэтому необходимо сохранять полный естественный ключ.
5. Что данные в Системе 2 «Таблица с рекламными вложениями продавцов» добавляет на ежедневной основе, только те данные в источник, которые были обновлены (если стоимость рекламы не была изменена, тогда эта запись не будет добавлена, так как есть более ранняя запись, которая зафиксировала цену).
6. Также принимаем тот факт что везде данные не имеют дублей и все источники существуют с наложенными на них ограничениями в ddl схеме при создании таблиц.

# Ход работы

## Описание таблиц в системах

Описание источников в системе 1:

* Источник «Справочник товар – категория», содержит только актуальную на сегодняшнюю дату i в себе. То-есть полностью обновляется и содержит только актуальную информацию. Поэтому, помимо того, чтобы завязаться на обновление этого источника, необходимо разработать схему SCD (именно SCD, а не RCD), так как данные являются информацией о категории товара, а не о транзакциях и продажах. Справочники имеют тип SCD, поэтому необходимо подобрать тип SCD и разработать витрину, которая может сохранять историчность данных и поможет строить аналитику в разрезе дат. Принимаем, что ключом является атрибут «Номер товара» по соответствующим допущениям.
* Источник «Таблица фактических продаж» содержит в себе факт об операциях, связанных с продажами. Обновляется инкрементально и данные добавляются в источник, а не полностью его перезаписывают. А что насчет возвратов? Примем допущения, что такое событие мы не учитываем или принимаем его невозможным, так как отсутствует точное описание, как такое событие обрабатывать. Также в таблице отсутствует состояние заказа / продажи, так что исходя из данного факта, принимаем, что все записи имеет завершённый статус заказа. То есть одна запись является реальной завершенной продажей продукта. Естественным ключом в таблице с рекламными вложениями продавцов будет связь «**Номер заказа**» + «**Номер товара**», дата же в данном случае выступает как ключ партиции в которой находится запись, для ускорения поиска по данным. Иначе не совсем понятно почему естественный ключ состоит из трех полей.

Описание источника из системы 2:

* Источник «Таблица с рекламными вложениями продавцов» содержит информацию по всем рекламным вложениям продавцов компании. Ежедневно таблица пополняется данными за вчерашний день. Естественным ключом является «**Дата**» + «**Номер товара**», то есть данные за 1 день не могут дублироваться по заданным полям. Источник уже имеет четкую версионность кортежей данных в зависимости от даты. SCD следует применить в таком, формате каждая строчка соответствует началу действия предложения на заданный продукт. Поэтому необходимо будет использовать диапазоны дат start\_dt + end\_dt для учета версии рекламного вложения продавцов.

## Построение Data Lineage

Следующим шагом