Итоговый проект

1. Описание задачи

a. **Предметная область** для предоставленных данных связана с коммерцией и электронной коммерцией (e-commerce), конкретно сосредоточенной на магазине велосипедов под названием "99Bikers" в Австралии. Данные включают информацию о продуктах (product_id, brand, product_line и др.), клиентах (first_name, last_name, gender и др.), транзакциях (transaction_date, online_order, order_status и др.) и адресах клиентов (address, postcode, state, country и др.). Такой тип данных обычно встречается в магазинах и интернет-магазинах, продающих товары оффлайн и онлайн, например, в магазине велосипедов.

b. **Применимость кейса**. На основе магазина велосипедов "99Bikers" в Австралии можно выделить несколько потенциальных вариантов использования или сценариев для бизнеса. Созданный ETL пайплайн позволяет приступить к решению следующих задач:

1. Сегментация клиентов и Таргетированный маркетинг:

Мы можем использовать демографические данные клиентов, их предыдущие покупки и интересы в области велосипедов для разделения клиентов на разные группы (например, случайные райдеры, энтузиасты) и создания таргетированных маркетинговых кампаний, которые будут соответствовать их предпочтениям и потребностям.

2. Управление запасами на складах и повторные заказы:

Анализируя данные о продажах товаров для прогнозирования спроса и обеспечения оптимальных уровней запасов мы можем определить когда следует повторно заказать популярные товары, как управлять хранящимся на складах и избавляться от избыточного товара.

3. Анализ корзины покупателя:

Изучая данные о транзакциях клиентов, мы можем выявить частые совместно покупаемые товары. Это может помочь создавать комплекты товаров или рекомендации, увеличивающие возможности кросс-продаж.

4. Прогнозирование оттока и удержание клиентов (Churn):

Используя историю клиента, покупок и других атрибутов мы можем спрогнозировать возможный оттока клиентов и перспективы внедрения стратегий их удержания.

5. Оптимизация цен:

Проанализировав список цен, себестоимость и истории покупок мы можем максимизировать доходы при сохранении конкурентоспособности.

6. Стратегия запуска продукта:

Проанализировав исторические данные о продажах и предпочтениях клиентов мы можем выработать стратегию успешного запуска новых продуктов, например лучшее время и методы для введения новых товаров на рынок.

7. Улучшение обслуживания клиентов и user-experience:

Используя данные о предыдущих покупках и взаимодействиях мы можем улучшить качество обслуживания клиентов и выявить общие проблемы или запросы и принять меры по их предотвращению.

c. Структура хранилища

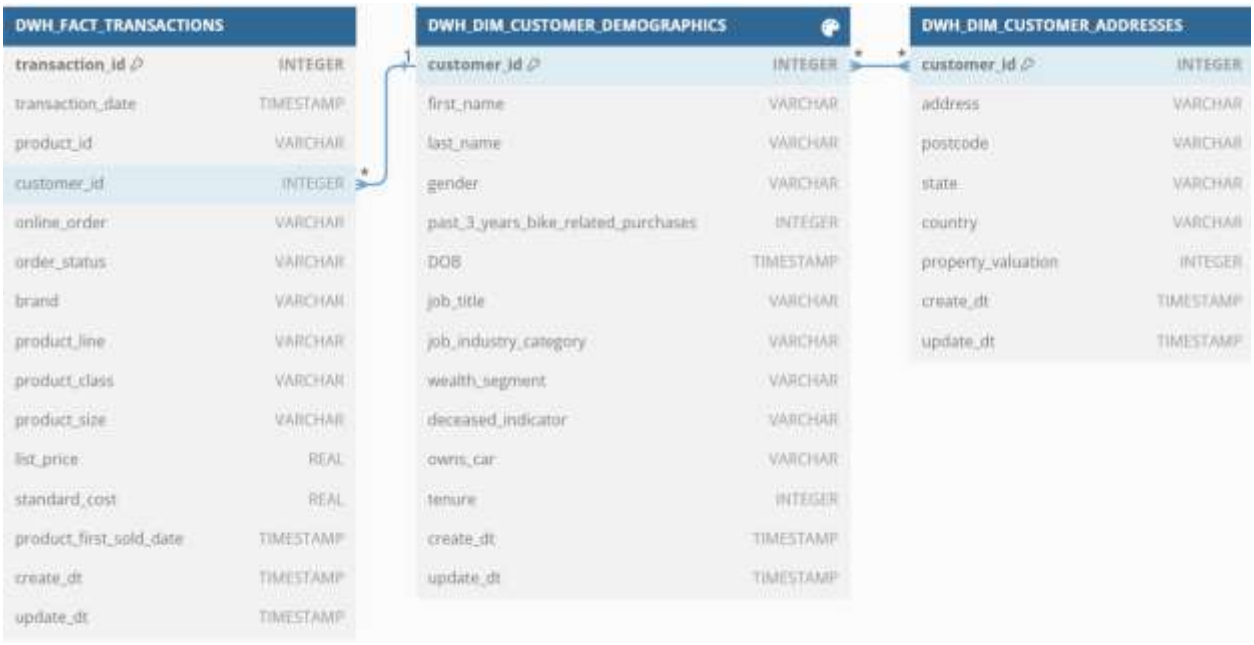
Данные загружены в хранилище со следующей структурой:

- Стейджинговая таблица STG_TRANSACTIONS
- Стейджинговая таблица STG_CUSTOMER_DEMPGRAPHICS
- Стейджинговая таблица STG_NEW_CUSTOMERS_LIST
- Стейджинговая таблица STG_CUSTOMER_ADRESSES

- Основная таблица DWH_FACT_TRANSACTIONS
- Основная табл. DWH_DIM_CUSTOMER_DEMOGRAPHIC
- Основная таблица DWH_DIM_CUSTOMER_ADRESSES

Основные таблицы имеют триггеры создания и дополнения данных, обновление осуществляется согласно SCD1.

2. ER-Диаграмма



3. Бизнес-процесс

Исходные данные представляют собой excel файл с четырьмя листами.

1. Лист с транзакциями операций. Включающий в себя 20 000 записей и 12 столбцов – айди продукта, клиента, дата продажи, онлайн\оффлайн заказ, статус заказа, бренд, линейка товаров, класс товара, размер, рекомендуемая цена продажи, ожидаемая цена продажи, дата первой продажи продукта.
2. Лист с информацией о клиентах. Включает в себя 4000 записей и 12 столбов – имя, фамилия, пол, покупки велосипедных товаров за последние три года, дата рождения, должность, сфера деятельности, класс обеспеченности, бинарный индикатор жив\мертв клиент, владеет ли авто, срок владения. Один параметр “default” содержит мусорные данные.
3. Лист с информацией о новых клиентах. Включает в себя 1000 записей и 23 столбца по большей части состоящий из двух листов – о клиентах и их адресах. Имя, фамилия, пол, покупки велосипедных товаров за последние три года, дата рождения, должность, сфера деятельности, класс

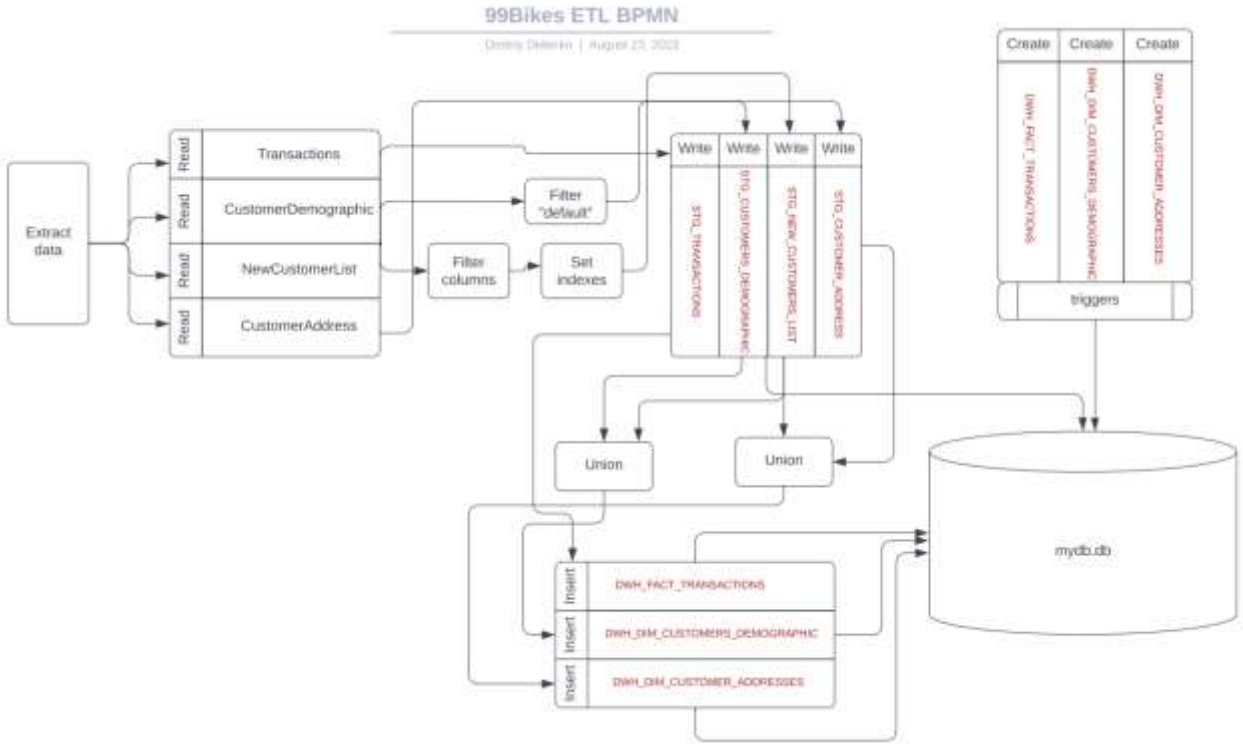
обеспеченности, бинарный индикатор жив\мертв клиент, владеет ли авто, срок владения. Пять скрытых столбцов не имеют названия. Есть столбцы ранг и значение.

4. Лист с информацией об адресах клиентов. 3999 значений с 6 пропусками в индексах, 5 столбцов – адрес, почтовый код, штат, страна, стоимость недвижимости.

Блок трансформации представляет собой операции осуществляемые над данными перед и после подачи в стендинговый слой с помощью библиотек pandas и sqlite3. Данные приводятся к единым индексам, осуществляется компоновка и объединение данных.

Финальные данные представляют три таблицы – таблица транзакций, таблица клиентов и таблица адресов. Стендинговая таблица новых клиентов была преобразована так, чтоб дополнить таблицы клиентов и адресов.

BPMN



4. Архитектура

a. База данных

- SQLite

b. Компоненты

- Python
- SQL

Архитектура пайплайна представляет собой последовательную обработку исходных файлов формата (excel) с последующей записью данных в базу (SQLite). Весь код обработки данных написан на языке python и SQL. Раз в день запускается скрипт и в базу добавляются новые данные из нового файла. Сама база данных представляет собой двухуровневое хранение данных (стейдинговый слой и основной). Отношения сущностей в базе представлено в 1НФ.

5. Выбор СУБД

При анализе исходных данных были сделаны выводы, что табличную структуру хранить выгоднее в реляционной базе. Так же, при недостатке ресурсов можно воспользоваться нормализаций и привести отношение сущностей к 4НФ.

6. Выбор СХД

Stage - это промежуточные данные, ждущие обработки, хранящиеся в виде файлов. Поэтому, нам удобнее всего использовать файловую СХД (например s3). Дальнейшее преобразование и запись данных будет осуществляться в реляционной базе SQLight.

7. Алгоритмы и методов анализа и обработки

Для обработки исходных данных используется язык python и библиотека pandas. Дальнейшая трансформация и обновление данных происходит на стороне базы данных с помощью запросов SQL. Взаимодействие соуществляется с помощью sqlite3.

8. Описание модели угроз

Перечень угроз:

- Несанкционированный доступ в базу (расположена во внутреннем контуре с ограниченным доступом)
- Неверно заданные права доступа и подбор пароля
- Права доступа к хранимой информации (ограниченный круг лиц, кто может взаимодействовать с этой информацией)
- Защита от внешних атак (SQL injection)
- Отсутствие резервного копирования и регулярных бекапов.