**Описание бизнес кейса:**

Я хочу для онлайн магазина Hoff сделать таргетированную рекомендацию продуктов в режиме онлайн клиентам, которые используют его.

**Представим единичный случай:**

Степан Иванов заказывает садовые качели. Степан - вероятнее всего, мужчина. Он потратил 50 минут на просмотр 20 разных качелей и аксессуаров к ним перед окончательным выбором. Можно сделать вывод, что у него не было четких требований к товару. Как он попал на страницу компании? Если добавить сопутствующую информацию, выяснится, что он ввел поисковый запрос в Google и перешёл на сайт компании. К нашему анализу можно добавить полную историю его покупок онлайн, и сделать вывод, что Степан купил много товаров для ремонта дома, а за последние две недели количество таких покупок резко увеличилось. Для приобретения товаров Степан пользуется поисковиком Google, а не заходит на конкретный сайт магазина, поэтому можно сказать, что у него нет приоритетных брендов для покупок. При оплате товара он указал адрес доставки, на основании которого можно определить некоторые демографические данные. Это коттеджный поселок “Анютины глазки” на западе Подмосковья. В этом районе очень дорогая недвижимость, что говорит о том, что в основном, там живут очень состоятельные люди. Допустим, у нас есть доступ к единой базе недвижимости, тогда мы еще узнаем, что этот дом был продан полтора месяца назад, а на его территории есть баня и бассейн. Значит, Степан новосёл и сейчас занимается вопросами ремонта. Это вторичное жильё, а значит, не стоит предлагать Степану материалы для шумоизоляции или выравнивания стен. Также он нажал на кнопку “Позвать друга” и тем самым принял условия пользовательского соглашения с VK/Instagram, что открыло его социальную сеть. Мы собрали много сырых данных, которые помогут нам нарисовать портрет конкретного покупателя. Теперь мы знаем, что в ближайший месяц стоит предлагать более дорогие товары для ремонта и уюта: обои(если по предыдущим покупкам понятно, что Степан хочет сделать именно ремонт, а не обновить часть интерьера с помощью декора), коврики, шторы, мб шкафы для ванной и тд. А чуть позже живые цветы для украшения дома.

Имеет смысл провести такой анализ для других покупателей магазина, автоматизируя процесс хотя бы частично.

**Построение аналитического хранилища данных.**

Источниками данных будем считать всё, что как-то может рассказать нам о клиенте, то есть мы выгружаем данные из API Hoff и других магазинов для дома и ремонта, API VK, API Instagram, Единая база данных недвижимости и тд. Теперь нужно собрать данные из нескольких источников в нескольких форматах и поместить их в одно хранилище данных.

Извлечение, преобразование и загрузка представляет собой конвейер, в рамках которого данные собираются из различных источников, преобразовываются (объединение, очистка, дедупликация и проверка данных на качество) в соответствии с бизнес-правилами и загружаются в целевое хранилище данных.

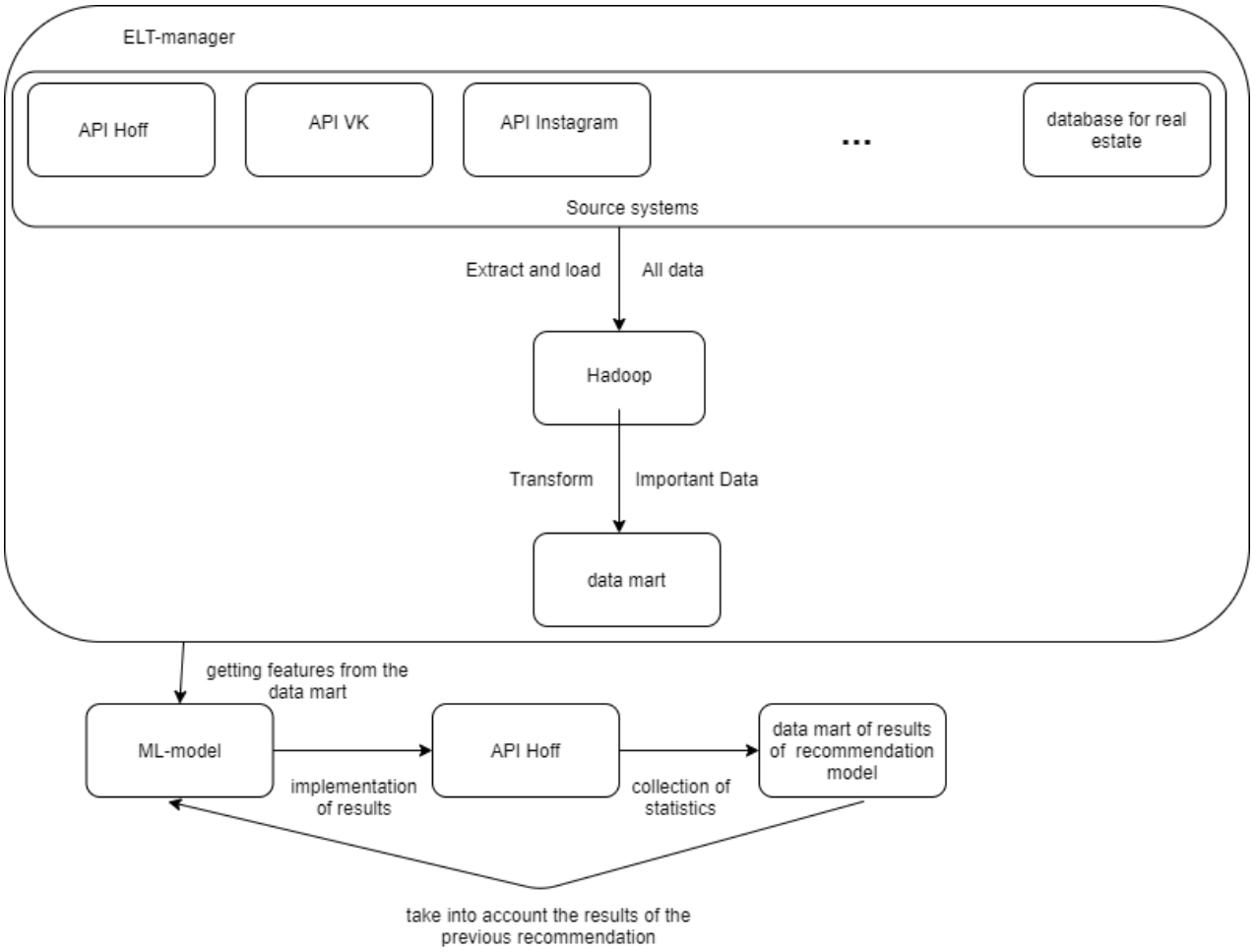
Я хочу использовать ELT конвейер для обработки больших объемов данных. В нем преобразование происходит в целевом хранилище данных. Это упрощает архитектуру за счет удаления механизма преобразования из конвейера. Также большим преимуществом является то, что масштабирование целевого хранилища данных также улучшает производительность конвейера ELT.

Мы извлекаем все исходные данные в неструктурированные файлы в масштабируемое хранилище (например, распределенную файловую систему Hadoop (HDFS)). Затем для выполнения запроса исходных данных будем использовать Spark(достаточно mini-batch streaming). Ключевой особенностью ELT является то, что хранилище данных, используемое для выполнения преобразования, — это то же хранилище, в котором данные в конечном счете потребляются. Это хранилище данных считывает данные непосредственно из масштабируемого хранилища, вместо того чтобы загружать их в собственное защищённое хранилище. Этот подход пропускает этап копирования (присутствующий в ETL), который может занимать много времени при обработке больших наборов данных, а мы хотим проводить все операции достаточно быстро.

И так, будем считать, что целевое хранилище данных — это хранилище данных, использующее кластер Hadoop (с использованием Hive или Spark). Схема накладывается на данные неструктурированных файлов во время выполнения запроса и сохраняется в виде таблиц(внешнее масштабируемое хранилище), позволяя запрашивать данные таким же образом, как и любую другую таблицу в хранилище данных. Здесь же собираем витрины(1) данных, необходимые для ML-модели.

Следующий этап, это разработка ML-модели: написание рекомендательной системы для отбора продуктов для последующей рекламы клиенту на основании признаков из витрины(1).

На основание этой модели рекомендуем определенные товары покупателю на сайте. Также ведем статистику для каждого клиента сайта: какие товары показали, на что клиент обратил внимание, что в итоге купил. Непрерывно совершенствуем рекомендательную модель с учетом этой статистики(то есть мы считаем, что клиент не купит определенный товар только тогда, когда мы его уже показали, а клиент не отреагировал. Иначе возможна ситуация, когда клиент не купил не потому, что ему не понравился товар, а потому что он не видел). Добавляем новые данные в витрину(1).


ELT-manager
API Hoff
API VK
API Instagram
...
database for real estate
Source systems
Extract and load
All data
Hadoop
Transform
Important Data
data mart


getting features from the data mart
ML-model
API Hoff
data mart of results of recommendation model
implementation of results
collection of statistics
take into account the results of the previous recommendation