Описание задачи.

Разработать ETL процесс, получающий ежедневную выгрузку данных (за 3 дня), загружающий ее в хранилище данных и ежедневно строящий отчет.

Выгрузка данных.

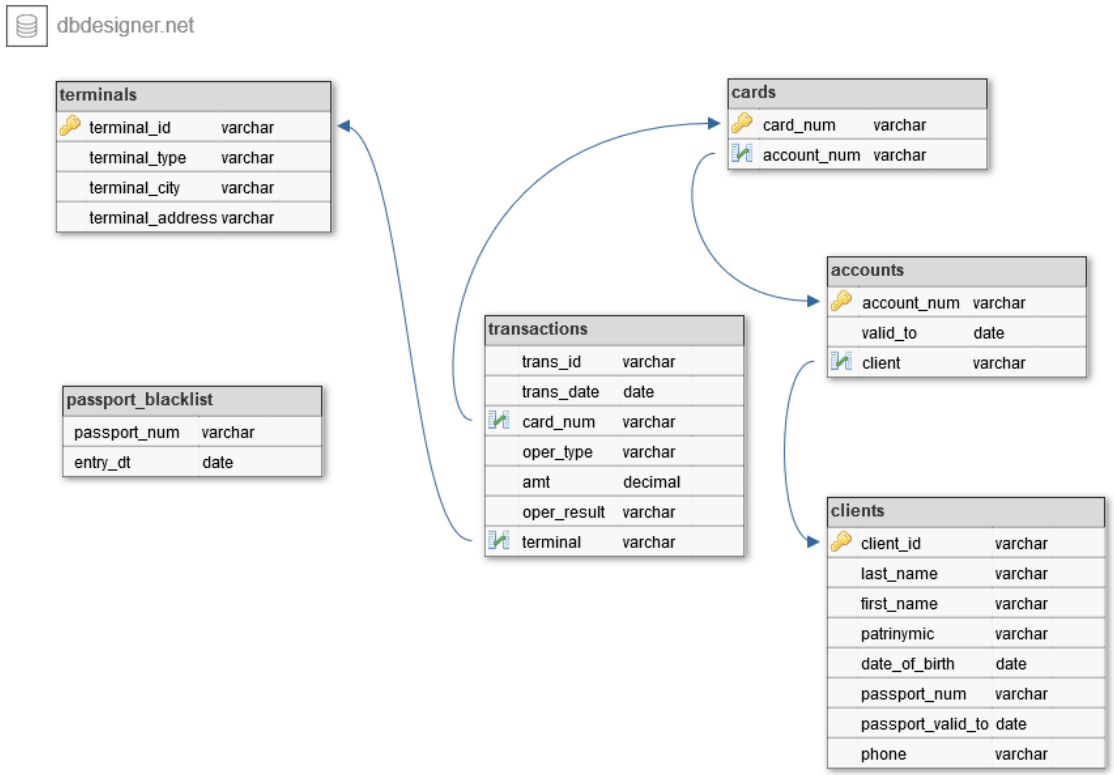
Ежедневно некая информационная система выгружает файл в формате xlsx, который содержит в ненормализованном виде все транзакции, совершенные за предыдущие дни месяца, т.е. с начала месяца происходит накопление. В файле к транзакциям привязаны все дополнительные сведения, типа клиента, номера договора и прочее.

Дополнительно предоставляется файл выгрузки паспортов из «черного списка» для проверки на мошеннические действия, его так же необходимо захватывать в хранилище для построения отчета. Файл содержит только номера паспортов и дату внесения в реестр. Ежедневная выгрузка содержит паспорта, выгруженные только в этот день, без накопления.

Файлы предоставлены в двух видах – с кириллическими символами и транслитерацией. Обратите внимание на критерии оценки, если вы пользуетесь файлами с транслитерацией, то недополучаете баллы. Коллеги, выполняющие индивидуальные проекты – просьба предусмотреть в ваших исходных данных поля с кириллическими символами, поскольку критерий распространяется на всех.

Структура хранилища.

Данные должны быть загружены в следующую нормализованную структуру:



Типы данных в полях можно изменять на однородные если для этого есть необходимость.

Ко всем таблицам SCD1 должны быть добавлены технические поля `create_dt`, `update_dt`; ко всем таблицам SCD2 должны быть добавлены технические поля `effective_from`, `effective_to`, `deleted_flg`.

Построение отчета.

По результатам загрузки ежедневно необходимо строить витрину отчетности по мошенническим операциям. Витрина строится накоплением, каждый новый отчет укладывается

В витрине должны содержаться следующие поля:

`event_dt`	Время наступления события. Если событие наступило по результату нескольких действий – указывается время последнего действия, по которому установлен факт мошенничества.
`passport`	Номер паспорта клиента, совершившего мошенническую операцию.
`fio`	ФИО клиента, совершившего мошенническую операцию.
`phone`	Номер телефона клиента, совершившего мошенническую операцию.
`event_type`	Описание типа мошенничества.
`report_dt`	Время построения отчета.

Признаки мошеннических операций.

1. Совершение операции при просроченном или заблокированном паспорте.

2. Совершение операции при недействующем договоре.

3. Совершение операций в разных городах в течение одного часа.

4. Попытка подбора суммы. В течение 20 минут проходит более 3х операций со следующим шаблоном – каждая последующая меньше предыдущей, при это отклонены все кроме последней. Последняя операция (успешная) в такой цепочке считается мошеннической.

Правила именования таблиц.

Необходимо придерживаться следующих правил именования (для автоматизации проверки):

`DE5.<CODE>_STG_<TABLE_NAME>`	Таблицы для размещения стейджинговых таблиц (первоначальная загрузка), промежуточное выделение инкремента если требуется. Временные таблицы, если такие потребуются в расчете, можно

	также складывать с таким именованием. Имя таблиц можете выбирать произвольное, но смысловое.
DE5.<CODE>_DWH_FACT_<TABLE_NAME>	Таблицы фактов, загруженных в хранилище. В качестве фактов выступают сами транзакции и «черный список» паспортов. Имя таблиц – как в ER диаграмме.
DE5.<CODE>_DWH_DIM_<TABLE_NAME>	Таблицы измерений, в данном случае все справочники. Имя таблиц – как в ER диаграмме.
DE5.<CODE>_DWH_DIM_<TABLE_NAME>_HIST	Таблицы измерений, хранящиеся в SCD2 формате (только для тех, кто выполняет усложненное задание). Имя таблиц – как в ER диаграмме.
DE5.<CODE>_REP_FRAUD	Таблица с отчетом.
DE5.<CODE>_META_<TABLE_NAME>	Таблицы для хранения метаданных. Имя таблиц можете выбирать произвольное, но смысловое.

<CODE> - 4 буквы вашего персонального кода.

Если результирующее имя не удовлетворяет ограничениям Oracle – необходимо из имени таблицы удалить гласные буквы:

```
PASSPORT_BLACKLIST = PSSPRT_BLCKLST
```

Проверка результата.

На проверку должен быть отправлен на почту запакованный в zip набор файлов (скрипты python, скрипты SQL, скрипты DDL). Обязательным является один главный файл python с именем `main.py`, а также файл с DDL. При создании DDL в начале файла должны идти команда `DROP TABLE`, после этого их создание. Остальные файлы являются дополнительными и при правильном использовании повышают балл за структурированность кода.

Тема письма – строго «Индивидуальный проект Фамилия». Будет настроен отдельный фильтр для оперативного реагирования на эти письма.

Данные в таблицах будут проверены автоматически исходя из правил наименования. Будьте внимательны, если имя таблицы не соответствует выставленным требованиям – проверка не происходит.

Обработка файла инкремента

В каталоге с файлом `main.py` ищутся следующие файлы:

`transactions_DDMMYYYY.xlsx`

`passport_blacklist_DDMMYYYY.xlsx`

Предполагается что в один день приходит по одному такому файлу. После загрузки соответствующего файла он должен быть переименован в файл с расширением .backup чтобы при следующем запуске файл не искался:

`transactions_DDMMYYYY.xlsx.backup`

`passport_blacklist_DDMMYYYY.xlsx.backup`

Желающие могут придумать, обосновать и реализовать более технологичные способы обработки.

Критерии оценки.

Оценка выставляется по нескольким критериям:

1. Структурированность кода – восприятие кода (отступы, табуляции), комментирование, разделение на отдельные файлы логических блоков. **До 15%.**

2. Качество обработки инкремента. Инкремент должен выделяться правильно, максимально эффективно и без лишних операций, контроль проводится в том числе автоматически по нескольким операциям. **До 15%. Дополнительные 5%** добавляются за использование оригинальных файлов, содержащих кириллицу. Если вы пользовались файлами с транслитерацией – 5% добавлены не будут.

3. Общая сложность процесса обработки данных. При выполнении задания необходимо придерживаться стандартов, изученных в курсе. Необоснованное ухудшение процесса обработки будет снижать балл. Приветствуется создание constraints, изученные алгоритмы выделения инкремента, использование метаданных. **До 30%.**

4. Качество получаемого результата. Необходимо найти все предусмотренные мошеннические операции. Всего их 7, по 5% за каждую найденную операцию. Итого **до 35%.**

5. Дополнительные баллы за сложность. Проверяющий оставляет за собой право добавлять **до 25%** дополнительных баллов за дополнительное полезное улучшение (и усложнение) проекта. Первый кандидат на такое улучшение – хранение всех измерений в SCD2 формате.