Описание задачи

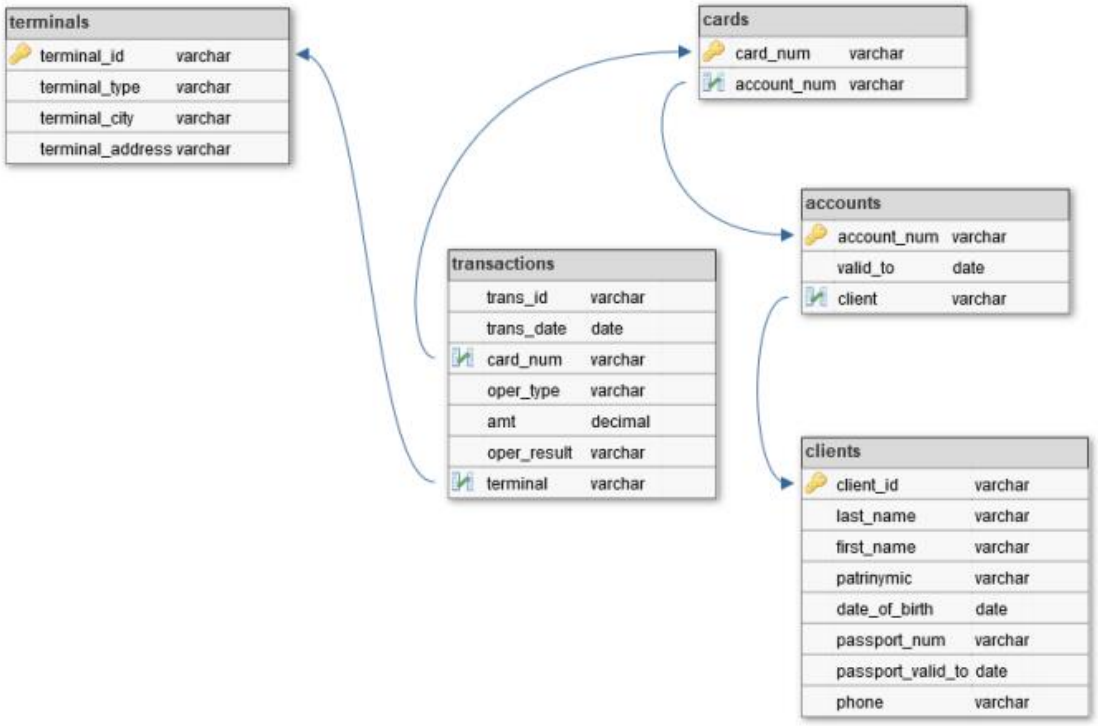
Ежедневно некоторая информационная система выгружает файл в формате xlsx, в котором в ненормализованном виде содержатся накопительно все транзакции (за 3 дня), совершенные за предыдущие дни месяца (накопление происходит с начала месяца). В файле к транзакциям привязаны сведения о клиенте, номера договора, карт и пр.

В файле «transactions_01052020.xlsx» - содержатся транзакции за 01.05.2020. В файле «transactions_02052020.xlsx» - содержатся транзакции за 01.05.2020 и 02.05.2020. В файле «transactions_03052020.xlsx» - содержатся транзакции за 01.05.2020, 02.05.2020 и 03.05.2020.

Необходимо разработать ETL-процесс, который ежедневно получает выгрузку, загружает ее в хранилище данных согласно структуре хранилища и ежедневно строит отчет.

Структура хранилища

Данные должны быть загружены в следующую нормализованную структуру:



При загрузке данных должна быть обеспечена версионность данных 1 или 2 типа. Ко всем таблицам SCD1 должны быть добавлены технические поля create_dt (дата создания версии), update_dt (дата изменения версии). Ко всем таблицам SCD2 должны быть добавлены технические поля start_dt и end_dt – начало и конец периода существования версии.

Правила наименования таблиц

Необходимо придерживаться следующих правил наименования таблиц:

- Таблицы для промежуточного выделения инкремента, а также любые временные таблицы:

STG_<TABLE_NAME>

- Таблицы фактов, загруженные в хранилище. В качестве фактов выступают сами транзакции:

FACT_TRANSACTIONS

- Таблицы измерений (terminals, accounts, clients, cards) в SCD1:

DIM_<TABLE_NAME>

- Таблицы измерений (terminals, accounts, clients, cards) в SCD2:

DIM_<TABLE_NAME>_HIST

- Таблица с отчетом:

REPORT

- Таблица для хранения метаданных:

META_<TABLE_NAME>

Построение отчета

По результатам загрузки ежедневно необходимо строить витрину данных для финансового отдела по подозрению на проведение мошеннических операций. Витрина должна строиться накопительно.

В витрине должны содержаться следующие поля:

- FRAUD_DT – Время наступления предполагаемого мошенничества. Если событие наступило по результату нескольких действий, указывается время последнего действия.

- PASSPORT – Номер паспорта клиента

- FIO – ФИО клиента

- PHONE – Номер телефона клиента

- FRAUD_TYPE – Описание типа предполагаемого мошенничества

- REPORT_DT – Время построения отчета.

Предполагаемыми мошенническими действиями являются:

1) Совершение операции при просроченном паспорте.
2) Совершение операции при недействующем договоре.
3) Совершение операции в разных городах в течение 1 часа.
4) Попытка подбора сумм. В течение 20 минут проходит более 3х операций со следующим шаблоном – каждая последующая меньше предыдущей, при этом отклонены все кроме последней. Последняя операция (успешная) в такой цепочке считается мошеннической.

Оценка результатов

На проверку должны быть отправлены скрипты DDL, ETL и Report. При создании DDL в начале должна идти команда DROP TABLE всех таблиц, после этого их создание.

Критерии оценки

1. Структурированность кода: отступы, комментирование, разделение на логические блоки.
2. Качество обработки инкремента: Инкремент должен выделяться правильно, эффективно и без лишних операций.

3. Общая сложность процесса обработки данных: Необоснованное ухудшение процесса обработки данных снижает балл. Дополнительные баллы будут начислены за создание constraints, использование метаданных, хранение всех измерений в SCD2 формате.
4. Качество получаемого результата: необходимо найти все предполагаемые мошеннические операции.
5. Загрузка оригинальных Excel-файлов в таблицу Oracle возможна ручным процессом. Дополнительные баллы будут начислены за автоматизацию процесса загрузки данных, например, средствами python с использованием модуля cx_Oracle или типовой процедурой загрузки внешних файлов. Допускается любая обоснованная технологичная реализация обработки данных.