СБЕР УНИВЕРСИТЕТ

Дипломный проект на тему: «Разработка ETL-процесса: построение витрины отчетности по мошенническим операциям»

Слушатель:
Харламов Филипп Александрович

Актуальность темы и ее проблематика

Данная тема является актуальной, т.к. согласно статистике:

1. В 2021 году Россияне потеряли 3,15 млрд рублей из-за мошенничества с фейковыми платёжными системами (tadviser.ru)
2. Пандемия спровоцировала увеличение активности мошенников. Рост объема и количества несанкционированных операций связан с активным переходом граждан на дистанционный формат потребления продуктов и услуг, в том числе финансовых (rbc.ru)
3. Объем рынка продаж краденых данных банковских карт приблизился к $2 млрд

Ежедневное отслеживание мошеннических операций поможет в кратчайшие сроки выявить обман и отменить выполненные транзакции

Разработка ETL-процесса: построение витрины отчетности по мошенническим операциям

Для решения данной задачи использовались следующие технологии:

1) Python (библиотеки pandas, os, re, jaydepeapi)
2) Oracle SQL (строковые функции, оконные функции, базовый синтаксис)

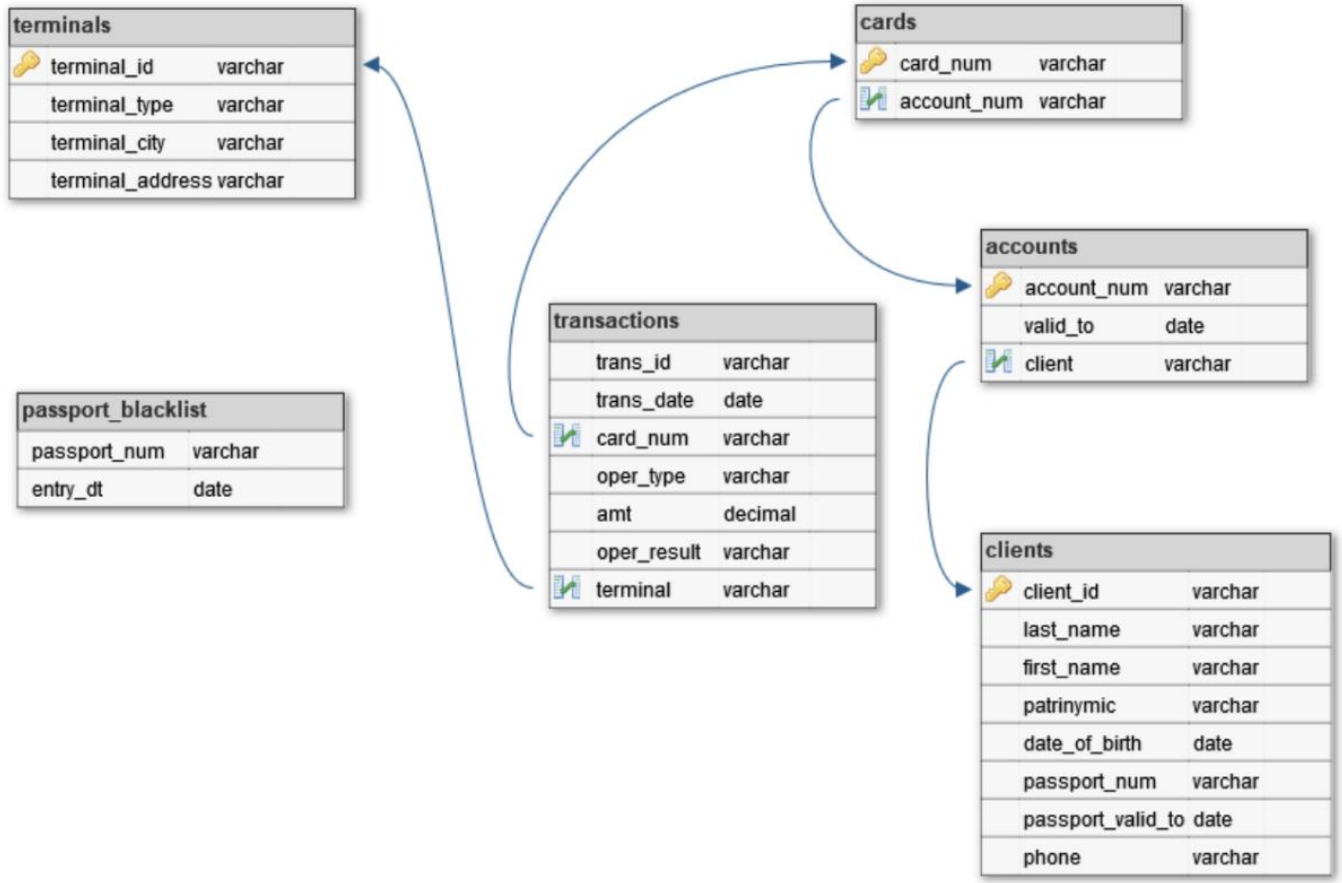
Входные данные:

1) MS EXCEL (xlsx)
2) csv-файлы
3) Существующие данные в таблицах БД

Слушатель: Харламов Ф.А.

Разработка ETL-процесса: построение витрины отчетности по мошенническим операциям

Схема данных в БД



Слушатель: Харламов Ф.А.

Разработка ETL-процесса: построение витрины отчетности по мошенническим операциям

ETL (от англ. *Extract, Transform, Load* — дословно «извлечение, преобразование, загрузка») — один из основных процессов в управлении хранилищами данных, который включает в себя:
1) извлечение данных из внешних источников;
2) их трансформация и очистка, чтобы они соответствовали потребностям бизнес-модели;
3) и загрузка их в хранилище данных.

С точки зрения процесса ETL, архитектуру хранилища данных можно представить в виде трёх компонентов:
1) источник данных: содержит структурированные данные в виде таблиц, совокупности таблиц или просто файла (данные в котором разделены символами-разделителями);
2) промежуточная область: содержит вспомогательные таблицы, создаваемые временно и исключительно для организации процесса выгрузки.
3) получатель данных: хранилище данных или база данных, в которую должны быть помещены извлечённые данные.

Слушатель: Харламов Ф.А.

Разработка ETL-процесса: построение витрины отчетности по мошенническим операциям

Схема ETL процесса

Слушатель: Харламов Ф.А.

Разработка ETL-процесса: построение витрины отчетности по мошенническим операциям

Ввод данных и STG слой

Ввод данных осуществляется при помощи файлов, полученных из других информационных систем. Файлы могут быть получены как все вместе, так и по отдельности. На случай, если файлы не будут переставать генерироваться, а система продолжительное время будет проявлять неработоспособность (по разным причинам), то в случае запуска нашей программы она начнёт обработку с самого раннего файла (и с задержкой в 5 сек). После каждого прохождения процесса, таблицы STG слоя автоматически удаляются.

Ввод данных и формирование STG слоя происходит в одном методе. Если файлов нет, то программа завершает работу, иначе продолжается, пока в папке не закончатся нужные файлы. В целях экономии, в случае отсутствия файла, программа не создаёт для него временную таблицу.

Так же STG слой формируется для таблиц (cards, clients…), уже предусмотренных в БД.

Слушатель: Харламов Ф.А.

Разработка ETL-процесса: построение витрины отчетности по мошенническим операциям

Слой FACT

Данный слой формируется из очищенных STG таблиц. В нашем случае это таблицы transactions и passport_blacklist. На текущий момент им не нужна очистка, поэтому таблицы слоя Fact были загружены как есть. В случае отсутствия таблицы Fact она создаётся и тут же заполняется, по окончанию процесса таблицы удаляются.

Слой DIM

Это слой измерений. Я выбрал путь SCD2, он предполагает историчность нормативно-справочной информации (НСИ) на уровне строк. Данный слой формируется посредством инкрементальной загрузки таблиц cards, clients, accounts и terminals. Каждая таблица имеет период актуальности (дата начала, дата окончания) и признак пометки удаления.

Слушатель: Харламов Ф.А.

Разработка ETL-процесса: построение витрины отчетности по мошенническим операциям

Отчет REP_FRAUD

Схема полей отчета:

1) **event_dt** – дата наступления события. Это дата операции (транакции), которая позволяет узнать, когда произошла мошеннической операция
2) **passport** – номер паспорта физического лица, совершившего операцию
3) **fio** – фамилия, имя, отчество физического лица
4) **phone** – номер телефона физического лица
5) **event_type** – тип мошеннической операции. Данные типы не предопределены в задаче, поэтому с ними можно ознакомиться в **Приложении 1**
6) **report_dt** – время построения отчета

Слушатель: Харламов Ф.А.

Разработка ETL-процесса: построение витрины отчетности по мошенническим операциям

Отчет REP_FRAUD

Условия задачи:

К общим условиям можно отнести то, что каждая мошенническая операция должна быть выполнена (SUCCESS)

Признаки мошеннических операций.

- Совершение операции при просроченном или заблокированном паспорте.
- Совершение операции при недействующем договоре.
- Совершение операций в разных городах в течение одного часа.
- Попытка подбора суммы. 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.

Слушатель: Харламов Ф.А.

Разработка ETL-процесса: построение витрины отчетности по мошенническим операциям

Отчет REP_FRAUD

Условия 1 и 2 выполняются посредством 3-х пакетов. Здесь используются исключительно базовые принципы SQL.

Условие 3 выполнено благодаря соединению “self join” и работе с датами

Условие 4 – самое сложное. Здесь понадобилось использовать оконные функции, для вычисления аналитики по 2-м последующим транзакциям внутри действий каждого физического лица

Слушатель: Харламов Ф.А.

Выводы

В ходе работы над дипломным проектом разработан ETL-процесс построения витрины отчетности по мошенническим операциям:

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 сумм

Слушатель: Харламов Ф.А.

Список использованных источников

1. Грофф, Джеймс . SQL : Энциклопедия : пер. с англ. / Д. Р. Грофф, П. Н. Вайнберг. — 3-е изд. — СПб. : Питер.
2. Уэс Маккинли Python и анализ данных [Электронный ресурс]/ Уэс Маккинли— Электрон. текстовые данные.— Саратов: Профобразование, 2017.— 482 c.
3. Кузнецов, Сергей Дмитриевич. Основы баз данных : курс лекций : учебное пособие / С. Д. Кузнецов.
4. https://pythontutor.ru/
5. www.sql-ex.ru
6. https://ru.wikipedia.org/wiki/ETL

Слушатель: Харламов Ф.А.

Приложение 1.
Типы мошеннических операций

1) **Not valid or blocked passport** – Недействительный или заблокированный пасспорт
1) **Not valid contract** – Просроченный договор
2) **Different cities into 1 hour** – Разные города в течении 1-го часа
3) **Selection amount** – Подбор сумм

Слушатель: Харламов Ф.А.