СБЕР УНИВЕРСИТЕТ

Дипломный проект на тему:

«Хранилище данных (DWH) банковских транзакций с отчётностью по мошенническим операциям»

Слушатели:
Киргизов Роман Анатольевич

Актуальность темы и ее проблематика

Для банковской организации, как для коммерческого предприятия, вопрос хранения и обработки оперативной информации критически важен в условиях резко возросшего её потока. Правильный подход позволяет выявить проблемные места в различных процессах, увеличивать скорость и безопасность их проведения, принимать верное решение по развитию бизнеса

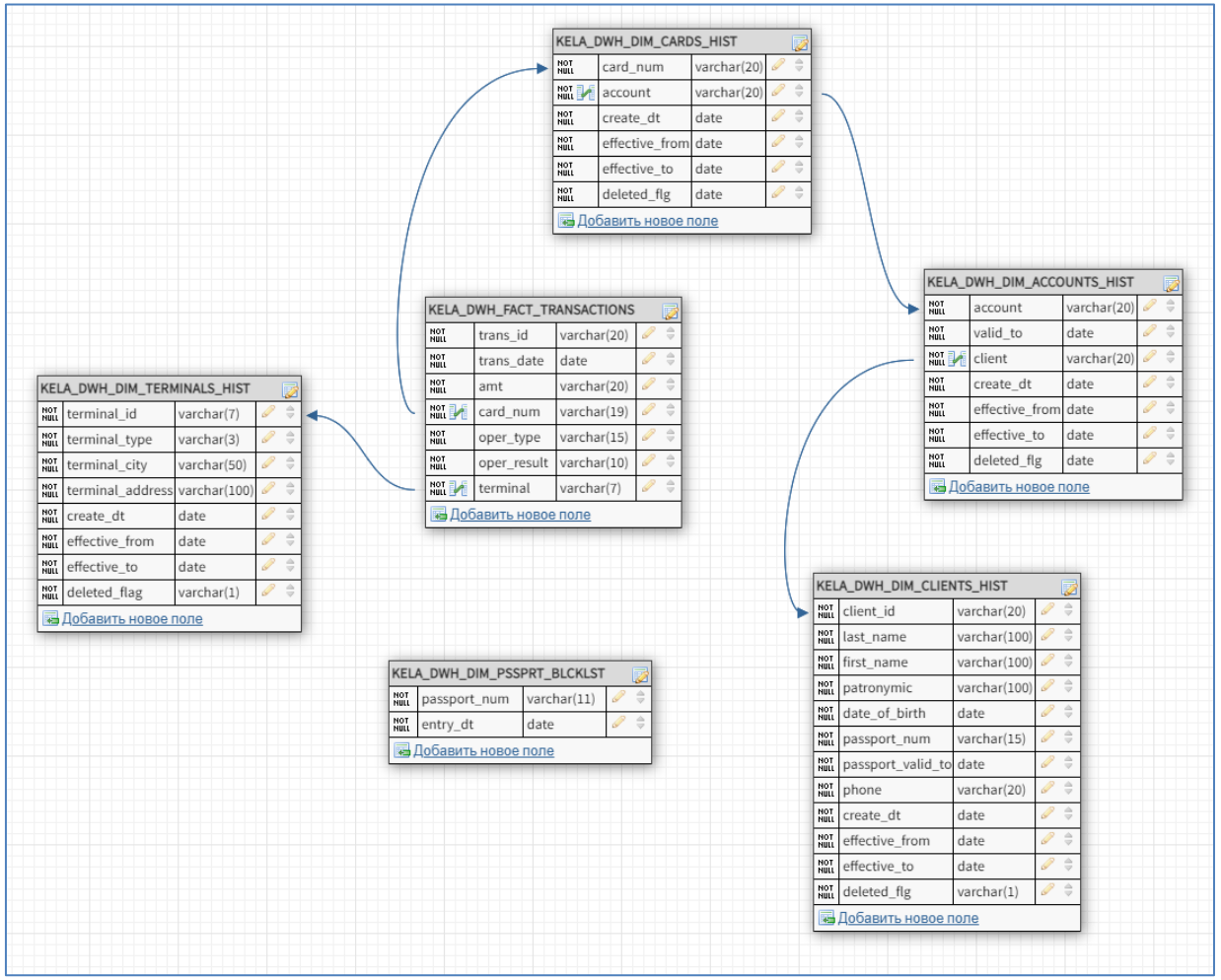
Используемые в оперативной деятельности OLTP-системы (Online Transaction Processing), не подходят для обработки и анализа данных, в связи с чем возникла потребность в создании системы, использующих информацию из OLTP-систем, не вмешиваясь в их структуру и процессы. Такие системы получили название хранилищ данных (Data Warehouse).

Data Warehouse – спроектированная специальным образом информационная база данных, предназначенная для подготовки отчётов и бизнес-анализа с целью поддержки принятия решений в организации.

Для дипломного проекта было выбрано построение DWH на основе данных банковских транзакций по карточным счетам физических лиц с анализом их на предмет выявления мошеннических операций.

Схема данных

Для построения хранилища была выбрана схема «Снежинка» с частичной нормализацией. Наиболее важные данные находятся в родительских таблицах измерений, но дополнительные сведения, которые могут не понадобиться или не понадобиться в каждом отчете, вынесены в дочерние таблицы.



Extract - Transform - Loading

ИСТОЧНИКИ ДАННЫХ

Источниками данных выступают:

1. Таблицы базы данных Oracle (ACCOUNTS, CARDS, CLIENTS)
2. Плоские файлы форматов csv, xlsx:
   - passport_blacklist_DDMMYYYY.xlsx
   - terminals_DDMMYYYY.xlsx
   - transactions_DDMMYYYY.txt

ПОСТРОЕНИЕ ОТЧЁТА

Отчёт по мошенническим операциям формируется также в среде Python 3.4.10 на учебном сервере накопительным способом после завершения ETL-процесса.

ETL-ПРОЦЕСС

ETL-процесс построен в среде разработки Python 3.4.10 на учебном сервере и включает в себя все необходимые шаги:

1. Извлечение данных из источников в таблицы временного хранения: `kela_stg_%`. Данные из источников могут храниться там установленное в файле config.ini время.
2. Трансформация и добавление служебных полей.
3. Загрузка чистых данных в слой постоянного хранения с соблюдением версионности SCD2, за исключением таблиц транзакций `kela_dwh_fact_transactions` и недействительных паспортов `kela_stg_pssprt_blcklst`, которые используются как фактовые таблицы.
4. Сохранение даты последнего обновления по каждой таблице в DWH для контроля загрузки новых данных.

Особенности разработки

1. Реализованы возможность как ручного запуска процессов в терминале, так и автоматического – посредством функционала CRON, с периодическим запуском ETL-процесса ежедневно в 05:00.
2. При использовании терминала реализовано меню, которое запускается из файла main.py и позволяет инициализировать хранилище данных, запустить ETL-процесс с возможностью выбора даты, сформировать отчётность.
3. Для реализации проекта были использованы как встроенные модули Python, так и разработанные непосредственно мной:
    1. Utils.py – используется для хранения вынесенных из основных модулей функций.
    2. Msql.py – содержит класс `KelaSQL`, который позволяет конструировать SQL-запросы посредством шаблонов и словарей.
    3. Logger.py – содержит класс `KelaLogger`, который позволяет вести раздельное логгирование происходящих процессов как только в файл, так и одновременно в файл и терминал.

```
   Kela Project
   1. Загрузка данных (etl)
   2. Формирование отчёта (report)
   7. Инициализация хранилища (init)
   0. Выход (exit)
   Введите команду или цифру: 7
   Инициализация хранилища данных проекта.
   Старые данные будут уничтожены!
   Вы уверены? Да-Yes-Enter/...:
   ИНИЦИАЛИЗАЦИЯ ХРАНИЛИЩА...
   1. Удаляем существующие таблицы...
   2. Создаём таблицы...
   3. Редактируем таблицы...
   4. Заполняем метаданные...
   Процесс завершён успешно. Лог записан в kela.log
[de3at@data-engineering kela]$ python main.py
```

```
   Kela Project
   1. Загрузка данных (etl)
   2. Формирование отчёта (report)
   7. Инициализация хранилища (init)
   0. Выход (exit)
   Введите команду или цифру: 1
   Введите дату загрузки:
   Будут загружены все данные по датам. Начать загрузку? Да-Yes-Enter/...
   ЗАГРУЗКА ВСЕХ ДАННЫХ...
   1. Обрабатываем данные из базы Oracle
   Очистка временных хранилищ...
   Загрузка и трансформация данных в DWH...
   2. Обрабатываем файлы с данными /home/de3at/kela/data ...
   Очистка временных хранилищ...
   Данные за 2021-03-01
   Загрузка и трансформация данных в DWH...
   Очистка временных хранилищ...
   Данные за 2021-03-02
   Загрузка и трансформация данных в DWH...
   Очистка временных хранилищ...
   Данные за 2021-03-03
   Загрузка и трансформация данных в DWH...
   Файлы данных не найдены. Err 003: пустая папка: /home/de3at/kela/data/
   ФОРМИРОВАНИЕ ОТЧЁТОВ...
   Попытка подбора суммы в течение 20 минут...
   Операции при недействующем договоре...
   Операции в разных городах в течение одного часа...
   Операции при просроченном или заблокированном паспорте...
   Процесс завершён успешно. Лог записан в kela.log
[de3at@data-engineering kela]$
```

Выводы

Проект реализован в соответствии с проектным заданием и выполняет все необходимые функции.

Сформированный отчёт по мошенническим операциям за 01, 02, 03 марта 2021 года выявил в общей сложности 2302 операции. Из них:

1. 1641 операция, 2 клиента – неверный или просроченный паспорт.
2. 643 операции, 1 клиент – недействующий договор.
3. 16 операций, 3 клиента – операции в разных городах за 1 час.
4. 2 операции, 2 клиента – попытка подбора суммы операции.

В ходе работы над проектом были успешно использованы полученные в ходе обучения знания и навыки.

Список использованных источников

1. Кузнецов, Сергей Дмитриевич. Основы баз данных : курс лекций : учебное пособие / С. Д. Кузнецов.
2. Моисеенко Сергей. SQL задачи и решения.
3. AMA-DMBOK. Свод знаний по управлению данными, 2-ое издание, 2020.
4. Oracle Help Center
5. Oracle PL/SQL учебник — Oracle PL/S
6. Python.org
7. JayDeBeApi: Documentation | Openbase
8. https://stackoverflow.com/
9. Data Engineering: ETL, ELT, Data Pipeline, Data Warehouse
10. What is a Data Warehouse? | IBM