Задание №1

В некотором банке внедрили новую frontend-систему для работы с клиентами, а так же обновили и саму базу данных. Большую часть данных успешно были перенесены из старых БД в одну новую централизованную БД. Но в момент переключения со старой системы на новую возникли непредвиденные проблемы в ETL-процессе, небольшой период (конец 2017 начало 2018 года) так и остался в старой базе. Старую базу отключили, а не выгруженные данные сохранили в csv-файлы. Недавно банку потребовалось построить отчёт по 101 форме. Те данные что остались в csv-файлах тоже нужны. Загрузить их в новую БД не получиться из-за архитектурных и управленческих сложностей, нужно рассчитать витрину отдельно. Но для этого сначала нужно загрузить исходные данные из csv-файлов в детальный слой (DS) хранилища в СУБД Oracle / PostgreSQL.

**Задача 1.1**

Разработать ETL-процесс для загрузки «банковских» данных из csv-файлов в соответствующие таблицы СУБД Oracle или PostgreSQL. Покрыть данный процесс логированием этапов работы и всевозможной дополнительной статистикой (на усмотрение вашей фантазии). В исходных файлах могут быть ошибки в виде некорректных форматах значений. Но глядя на эти значения вам будет понятно, какие значения имеются в виду.

Исходные данные:

* Данные из 6 таблиц в виде excel-файлов:
  + md\_ledger\_account\_s – справочник балансовых счётов;
  + md\_account\_d – информация о счетах клиентов;
  + ft\_balance\_f – остатки средств на счетах;
  + ft\_posting\_f – проводки (движения средств) по счетам;
  + md\_currency\_d – справочник валют;
  + md\_exchange\_rate\_d – курсы валют.
* Файл «Структура таблиц.docx» – поможет создать таблицы в детальном слое DS.

Требования к реализации задачи:

* В своей БД создать пользователя / схему «DS».  
  Примеры команд:  
  <https://oracle-dba.ru/docs/architecture/schemas/basics/>

<https://postgrespro.ru/docs/postgresql/9.6/sql-createschema>

* Создать в DS-схеме таблицы под загрузку данных из csv-файлов.
* Начало и окончание работы процесса загрузки данных должно логироваться в специальную логовую таблицу. Эту таблицу нужно придумать самостоятельно;
* После логирования о начале загрузки добавить таймер (паузу) на 5 секунд, чтобы чётко видеть разницу во времени между началом и окончанием загрузки. Из-за небольшого учебного объёма данных – процесс загрузки быстрый;
* Для хранения логов нужно в БД создать отдельного пользователя / схему «LOGS» и создать в этой схеме таблицу для логов;
* (В случае реализации процесса в Talend) В зависимости от мощностей рабочей станции – сделать загрузку из всех файлов одним потоком в параллели или отдельными потоками (может не хватить оперативной памяти для Java-heap);
* Для корректного обновления данных в таблицах детального слоя DS нужно выбрать правильную Update strategy и использовать следующие первичные ключи для таблиц фактов, измерений и справочников (должно быть однозначное уникальное значение, идентифицирующее каждую запись таблицы):

|  |  |
| --- | --- |
| **Таблица** | **Первичный ключ** |
| DS.FT\_BALANCE\_F | ON\_DATE, ACCOUNT\_RK |
| DS.FT\_POSTING\_F | OPER\_DATE, CREDIT\_ACCOUNT\_RK, DEBET\_ACCOUNT\_RK |
| DS.MD\_ACCOUNT\_D | DATA\_ACTUAL\_DATE, ACCOUNT\_RK |
| DS.MD\_CURRENCY\_D | CURRENCY\_RK, DATA\_ACTUAL\_DATE |
| DS.MD\_EXCHANGE\_RATE\_D | DATA\_ACTUAL\_DATE, CURRENCY\_RK |
| DS.MD\_LEDGER\_ACCOUNT\_S | LEDGER\_ACCOUNT, START\_DATE |

Технологические требования

ETL-процесс по загрузке файлов вы можете сделать с помощью различных технологий, которые вам будут удобней. Возможные варианты технологий:

* Talend – бесплатная (для учебных целей) ETL-платформа;