**Задача по автоматизации репликации данных. Ver. 0.5**

ETL (от англ. Extract, Transform, Load — «извлечение, преобразование, загрузка») — один из основных процессов в управлении хранилищами данных.

Это общий термин для всех процессов миграции данных из одного источника в другой.

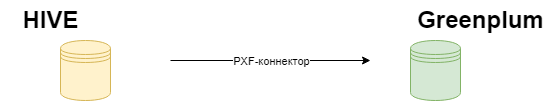
Процесс ETL включает в себя:

1. Извлечение данных из внешних источников.

2. Их трансформация, очистка и обогащение, чтобы они соответствовали потребностям бизнес-модели.

3. Загрузка данных в хранилище данных.

Процесс «перекладки» данных будет проходить с помощью **PXF**-коннектора, для подключения внешней таблицы к целевой базе данных с последующей репликации.



Для простой реализации ETL-процесса с перекладкой данных из источника (Hive) к приемнику на стороне Greenplum необходимо создать несколько служебных схем.

**etl\_service**– Схема, предназначенная для ETL-процессов.

**source\_hive** – схема, предназначенная для внешних таблиц (external table) с целевого источника Hive.

**monitoring\_service** – Схема, предназначенная для мониторинга просмотра загрузок, журнальных таблиц, процесса работы etl-сервиса.

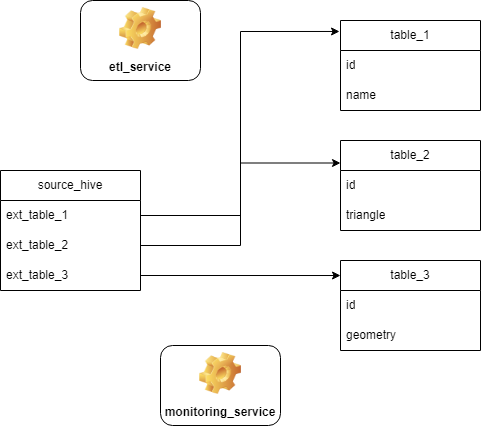
В последствии реализовать еще несколько схем под различные задачи.

**dataquality\_service** – схема предназначенная для реализации поддержки качества данных.

**admin\_service** – схема, предназначенная для администрирования процессов etl.

И тд.

Процесс внутренней перекладки данных регулируется внутренними процедурами, созданные на PLpgSQL.



Процесс выстроен следующим образом:

1. Создается внешняя таблица с помощью PXF в схеме **source\_hive**.
2. С помощью созданных процедур в **etl\_service** начинается процесс репликации данных в таблицу-источник
3. После завершения работы, процесс фиксируется в схему **monitoring\_service.**

Реализация схемы **source\_hive**

* Внутри схемы будут реализованы внешние таблицы, с источником данных на реляционной базе данных Apache Hive.
* Схема не должна хранить в себе никаких физических таблиц.
* Единственная процедура генерации корректных типов данных и создания таблицы.

Для подготовки создания внешнего подключения и создания external-таблицы необходимо придерживаться корректных типов данных. Они представлены в таблице ниже.

|  |  |
| --- | --- |
| **Тип поля в Hive** | **Тип поля в GreenPlum** |
| **Tinyint** | **int2** |
| **Timestamp** | **timestamp** |
| **String** | **varchar** |
| **Smallint** | **int2** |
| **Float** | **float4** |
| **Double** | **float8** |
| **Boolean** | **bool** |
| **Binary** | **byte** |
| **Bigint** | **int8** |
| **Decimal** | **numeric** |

Параметры создания таблицы

CREATE WRITABLE EXTERNAL WEB [TEMPORARY | TEMP] TABLE <table\_name>

    ( <column\_name> <data\_type> [, ...] | LIKE <other\_table> )

    EXECUTE '<command>' [ON ALL]

    FORMAT 'TEXT'

               [( [DELIMITER [AS] '<delimiter>']

               [NULL [AS] '<null string>']

               [ESCAPE [AS] '<escape>' | 'OFF'] )]

          | 'CSV'

               [([QUOTE [AS] '<quote>']

               [DELIMITER [AS] '<delimiter>']

               [NULL [AS] '<null string>']

               [FORCE QUOTE <column> [, ...]] | \* ]

               [ESCAPE [AS] '<escape>'] )]

           | 'CUSTOM' (Formatter=<<formatter specifications>>)

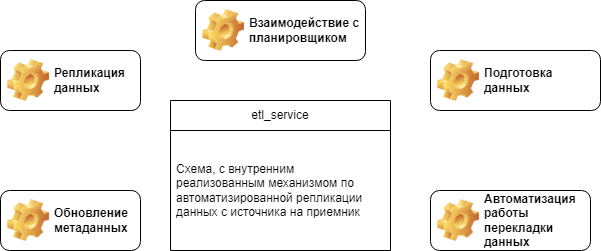
[ OPTIONS ( <key> '<value>' [, ...] ) ]

    [ ENCODING '<write\_encoding>' ]

    [ DISTRIBUTED BY ({<column> [<opclass>]}, [ ... ] ) | DISTRIBUTED RANDOMLY ]

Реализация **etl\_service**

* Схема будет содержать метаданные для необходимого процесса
* Внутри схемы будет располагаться множество процедур, обособленых между собой для соблюдения правила (одна процедура – одно действие)
* Схема будет служить основных механизмом взаимодействия с «шедулером» на внешней стороне
* Процедуры будут написаны согласно СПИРу, который будет реализован позднее.
* По работе процедур и их метаданных информация будет после согласования
* Процедуры будут реализованы на языке PlpgSQL.



Процесс работы etl\_service:

* С помощью подготовленных процессов на стороне планировщика задач, с помощью метаданных, подготовленных на стороне **etl\_service** запускается процесс репликации с внешних таблиц в целевые таблицы.
* Подробнее о процессах будет описано позднее

Реализация **monitoring\_service**

* В схеме будут созданы необходимые таблицы логов (журнальные таблицы) для корректного мониторинга процессов
* Также, будут созданы представления для сжатой информации обо всех процессах
* Процедуры для записи работы ETL-процессов а также процедуры отправки сообщений на почту
* Процедуры будут реализованы PLpgSQL, а также на Python/Java

Для корректной реализации поставленной задачи по репликации и автоматизации данных необходимы еще некоторые дополнительные компоненты.

* CRUD права для взаимодействия с базой данных, а также права на execute.
* Доступ с теми же правами на контуры разработки и тестирования
* Созданные подключения к Apache Hive (PXF)
* Возможность работать с Apache Airflow или же целевым решением.
* Созданная страница на проекта на GITLAB с реализацией к последующему развертыванию с процессом CI/CD (Если есть возможность, я могу сам сделать всю структуру процессов)