DATA HACK

5 — 7
АВГУСТА 2022

КЕЙС #3

Создание прототипа ETL Движка
из Postgres, Oracle, ClickHouse
в HDFS на Spark

КЕЙС 3

Прототип ETL/ELT движка. Между системами источниками (Postgres) в HDFS на SPARK.

Требования к реализации ETL движка

Справа изображена концептуальная архитектура решения, представляющая собой высокоуровневую модель, определяющую компоненты и общую структуру решения.

Источники данных:

- РСУБД
- Текстовые файлы (csv, json, xml)
- Сервисы API

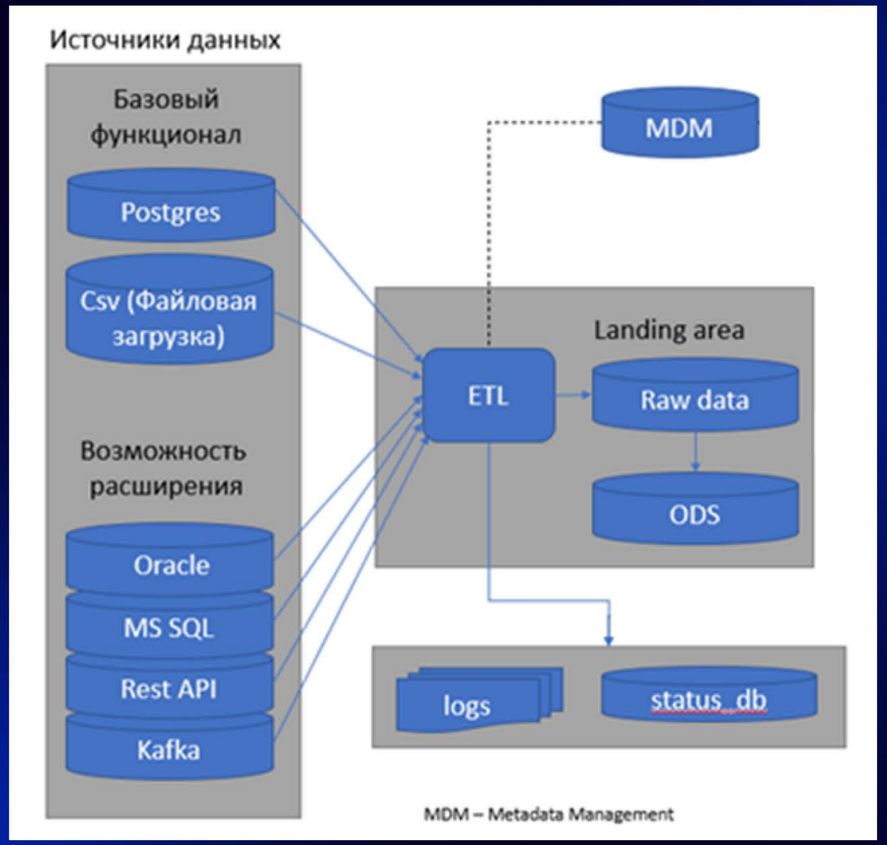

Рисунок 1. Общая архитектура решения.

КЕЙС 3

Прототип ETL/ELT движка. Между системами источниками (Postgres) в HDFS на SPARK.

Основные функциональные требования:

- В базовом функционале необходимо реализовать загрузку с системы Postgresql и организовать загрузку текстовых (csv) файлов, заложить функционал для подключения к любой РСУБД

- Параметризация выгрузки - источник, таргет, параметры подключения, конфиг SPARK-сесии, конфиг загрузки, количество выдаваемых файлов на выходе, выходной формат файлов, разные режимы записи, разложение по партициям. включенная компрессия файлов;

- Разные способы выгрузки:

• Overwrite (перезапись конкретных партиций либо таблицы целиком)

• Append (запись либо в таблицу либо в партицию)

- Учесть возможность выделения инкремента на стороне источника:

Способы выделения инкремента:

• По дате изменения (по техническому полю обновления);

• По ключу или составному ключу

КЕЙС 3

Прототип ETL/ELT движка. Между системами источниками (Postgres) в HDFS на SPARK.

Основные функциональные требования:

Добавить возможность выгрузки порциями данных:

• Если таблица большая, то автоматическое разбиение на мелкие интервалы для забора таблицы частями (например данные нужно забрать с 2010 по 2020 год при выставлении интервала в год, данные будут разбиты на 10 разных частей с 2010 по 2011, с 2011 по 2012 и так далее)

- Параметризация выгрузки должна быть реализована либо в виде файлов конфигов, либо в виде мета информации сохраняемой в PG

- Статус выгрузки должен сохраняться в Meta раздел

- ETL инструмент должен уметь работать с несколькими слоями данных, которые будут располагаться на HDFS;

КЕЙС 3

Прототип ETL/ELT движка. Между системами источниками (Postgres) в HDFS на SPARK.

Базовые слои необходимые для загрузки данных:

- **Raw data** – слой, который содержит порции данных, выгруженных из интерфейсов систем источников. Порции данных содержат инкременты или полные срезы данных в зависимости от режима загрузки;
- **ODS** – слой оперативных данных, содержит данные в модели, близкой к моделям систем источников, обновляется на основании порций данных, которые поступают в слой raw data;
- **MDM** – система хранения и управления метаданными.

Процессы трансформации данных использует подход, называемый ELT, согласно которому ETL-платформа выполняет функцию загрузки данных в область Stage и дальнейший контроль последовательности запуска скриптов в разрабатываемом инструменте.

КЕЙС 3

Прототип ETL/ELT движка. Между системами источниками (Postgres) в HDFS на SPARK.

Система логирования и мониторинга выполнения загрузок:

Метаданные, включая информацию об системах источниках необходимо хранить в отдельном файле – конфигурации (JSON)– либо базе данных (например Postgres). В метаданных можно предусмотреть справочники маппингом типов данных между системой источником и теми типами данных, как мы их будем обрабатывать в хранилище.

- **Логи** (в текстовых файлах) либо в Meta разделе, в которых будет собираться информация по выполняемым шагам и результатам их работы.
- **Статусная информация**, по которой можно понять какие задачи выполняются ETL инструментов в настоящий момент времени, уровень загрузки системы, результат работы системы, разработать статусную модель состояния ETL задачи (например: запущена, ошибка, в процессе, выполнена, приостановлен; можно расширить данную статусную модель и мониторить разные элементы системы).

Перед загрузкой данных выполнять проверку на соответствие метаданных (что метаданные (структура, наименование полей и их типов на источнике) не изменились и препятствий к их обработке нет);
Расширенная возможность: предусмотреть возможность изменения метаданных источников, так что бы ETL инструмент мог это обрабатывать, хранить историю изменений, и при этом хранить состояние всех данных (до изменений и после них).

КЕЙС 3

Прототип ETL/ELT движка. Между системами источниками (Postgres) в HDFS на SPARK.

Область Raw Data:

- Продумать Naming Convention для новых таблиц (чтобы имя таблицы включало в себя схему, имя таблицы источника и по неймингу можно было понять к какому источнику она относится)
- Все данные грузим AS IS (как есть)
- Для всех загружаемых данных необходимо добавить технические поля выгрузки с источника и даты загрузки в raw

При подготовке данных к помещению в область ODS необходимо числить инкремент, при помощи окна загрузки, (либо провести действия, которые будут определены типом загрузки)

КЕЙС 3

Прототип ETL/ELT движка. Между системами источниками (Postgres) в HDFS на SPARK.

Область ODS:

Данные обогащаются PROCESSED_DTTM – дата и время операции загрузки данных;

- **EXTRACT_ID** – уникальный номер порции данных в рамках всех загрузок.
- **CLOSED_EXTRACT_ID** – номер порции данных – которым закрыта актуальная версия.
- **VALID_FROM_DTTM** – техническая версионность записи. Дата и время, с которого данная запись является действительной (SCD2)
- **VALID_TO_DTTM** – техническая версионность записи. Дата и время, по которое данная запись является действительной (SCD2)
- **DELETED_FLG** – флаг удаления записи.

В этой области для хранения накапливаются данные, которые загружаются в хранилище. У каждой записи есть период действия, который отражает в каком состоянии находилась запись в определенный период времени (назовем это версионностью данных, либо историей объектов)

КЕЙС 3

Прототип ETL/ELT движка. Между системами источниками (Postgres) в HDFS на SPARK.

Общая архитектура ETL:

Для реализации использовать:

- **Python 3.7+;**
- **PySpark 2+;**
- **Hadoop 2+;**
- **Postgres 10+;**

Процессы, которые реализуются при помощи ETL продукта должны быть:

- **Оптимальными**. Оптимизировать пересылку данных по сети, предварительные запросы выполнять на стороне источника;
- **Потенциально изменяемы переменные и настройки** должны быть вынесены в параметры, либо в базе метаданных; («Har code» должен быть минимизирован, поскольку мы делаем прототипом, хотелось бы видеть возможность этой оптимизации в бедующем)
- **Идемпотентность**. Многократное выполнение процесса приводит к одному результату;
- **Как плюс** предложение соглашение о наименовании модулей, процедур, функций, объектов, переменных;

КЕЙС 3

Прототип ETL/ELT движка. Между системами источниками (Postgres) в HDFS на SPARK.

Исходный код и стенды проекта:

В процессе реализации проекта необходимо файлы конфигураций и исходные коды программ разместить в git репозиторий.

Среды, для разработки развернуть в Docker контейнерах, каждое в отдельном контейнере. На каждый контейнер предоставить файл Docker образа, в котором будет поднято окружение, перенесен и запущен проект.

Должно быть представлено описание реализованного функционала, и возможностей его расширений.

Проверка и тестирование:

Команда должна подготовить тестовые данные для проведения демонстрации загрузок и демонстрации функционала инструмента.