Анализ публикуемых новостей

Итоговая аттестация, курс «Инженер данных» (2022 год)

Создание простого ETL- процесса формирования витрин данных

На основе источников:
https://lenta.ru/rss/
https://www.vedomosti.ru/rss/news
https://tass.ru/rss/v2.xml

Оглавление

Описание задачи

1. Разработать скрипты загрузки данных в 2-х режимах: инициализирующий – загрузка полного слепка данных источника, инкрементальный – загрузка дельты данных за прошедшие сутки.
2. Организовать правильную структуру хранения данных: сырой слой данных, промежуточный слой, слой витрин.
3. Написать скрипт, который формирует витрину данных.

План реализации

1. Изучение бизнес-задачи и требований.
2. Разработка архитектуры проекта.
3. Выбор стека технологий, инструментов решения.
4. Написание кода.
5. Тестирование.
6. Оформление результатов работы.

Используемые технологии

Нашей целью является создание доступного решения для повседневного использования конечным пользователем.

В качестве языка программирования для реализации использован Python как один из популярных языков программирования с необходимыми библиотеками: pandas для работы с данными, feedparser для парсинга новостных лент и crontab для оркестрации.

Для сохранения промежуточных и итоговых результатов выбраны файлы и таблицы Excel в распространенных форматах csv и xlsx, что позволяет легко ознакомиться с данными, проводить дальнейшую обработку и анализ данных, в том числе импорт в базы данных.

Схема процесса

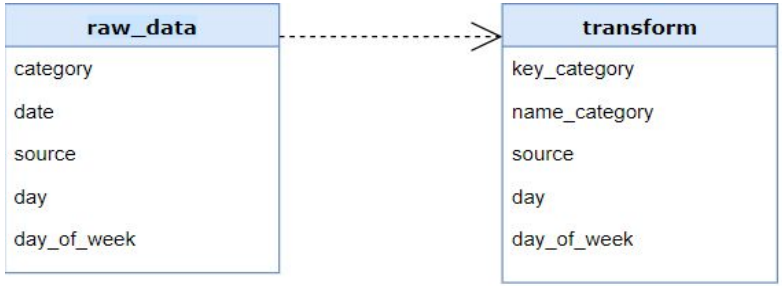
1. Загружаем данные из источников в start.py и сохраняем в raw.csv

2. Раз в сутки добавляем новые данные add.py и сохраняем в raw.csv

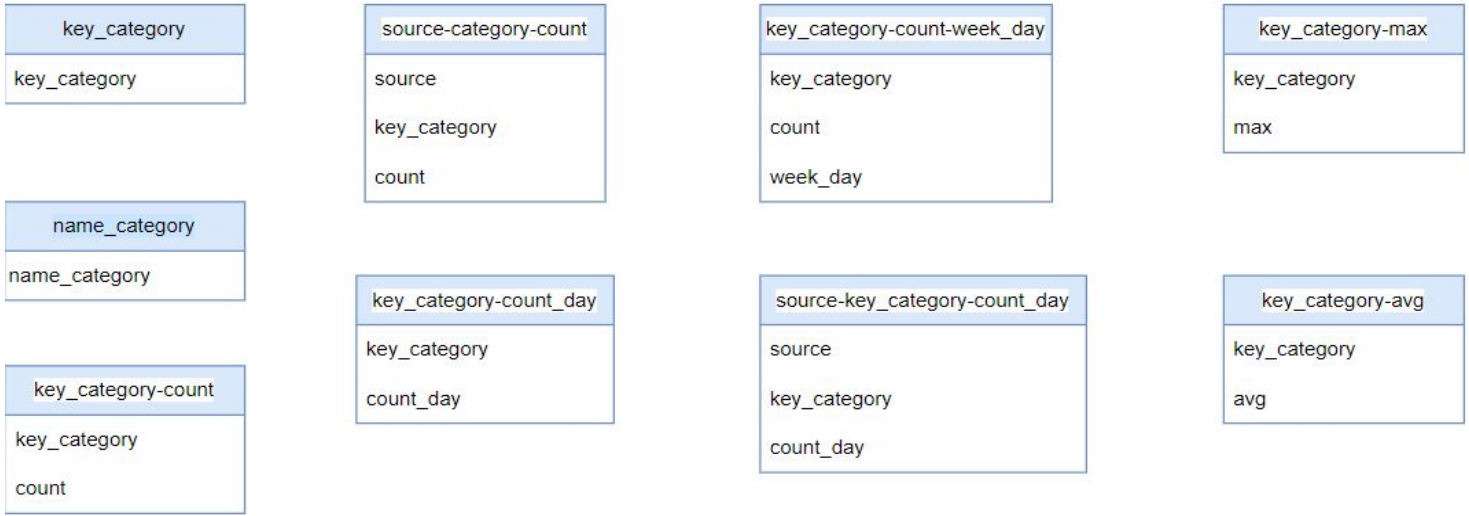
3. Обрабатываем данные и загружаем в файл transform.csv

4. Проводим расчеты и выгружаем витрины данных в лист книги Excel.

Структура выбрана исходя из требований задачи и согласно рекомендациям. Скрипты и файлы с данными в проекте расположены в разных папках.



ER - диаграмма



Результаты

Получилась книга Excel с таблицами, каждая в отдельном листе:

- Суррогатный ключ категории
- Название категории
- Общее количество новостей из всех источников по данной категории за все время
- Количество новостей данной категории для каждого из источников за все время
- Общее количество новостей из всех источников по данной категории за последние сутки
- Количество новостей данной категории для каждого из источников за последние сутки
- Среднее количество публикаций по данной категории в сутки
- День, в который было сделано максимальное количество публикаций по данной категории
- Количество публикаций новостей данной категории по дням недели

Пример диаграммы в Excel. Новости по категориям за день.