

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

●	<u>Описание задачи</u>	3
●	<u>План реализации</u>	4
●	<u>Используемые технологии</u>	5
●	<u>Схема процесса</u>	6
●	<u>ER - диаграмма</u>	7
●	<u>Результаты</u>	8
●	<u>Пример</u>	9



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, sqlalchemy для работы с базой данной и crontab для оркестрации.

PostgreSQL — распространенная система управления базами данных.

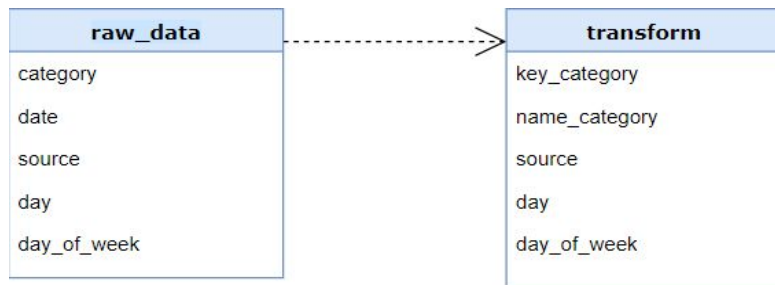
Для сохранения промежуточных и итоговых результатов также используются файлы и таблицы Excel в распространенных форматах csv и xlsx, что позволяет быстро ознакомиться с данными, отправить электронной почтой, построить графики.



Схема процесса

1. Загружаем данные из источников в `start.py`, сохраняем в `raw.csv` и загружаем в БД.
2. Раз в сутки добавляем новые данные `add.py`, сохраняем в `raw.csv` и загружаем в БД.
3. Обрабатываем данные, сохраняем в файл `transform.csv` и загружаем в БД.
4. Проводим расчеты и импортируем витрины данных в БД и результаты 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

key_category
key_category

source-category-count
source
key_category
count

key_category-count-week_day
key_category
count
week_day

key_category-max
key_category
max

name_category
name_category

key_category-count_day
key_category
count_day

source-key_category-count_day
source
key_category
count_day

key_category-avg
key_category
avg

key_category-count
key_category
count

Результаты

Получены витрины данных в Postgres и книга Excel с таблицами, каждая в отдельном листе:

- Суррогатный ключ категории (key_category)
- Название категории (name_category)
- Общее количество новостей из всех источников по данной категории за все время (key_category_count)
- Количество новостей данной категории для каждого из источников за все время (source-category_count)
- Общее количество новостей из всех источников по данной категории за последние сутки (key_category_count_day)
- Количество новостей данной категории для каждого из источников за последние сутки (source-key_category_count_day)
- Среднее количество публикаций по данной категории в сутки (key_category_avg)
- День, в который было сделано максимальное количество публикаций по данной категории (key_category_max)
- Количество публикаций новостей данной категории по дням недели (key_category_count_week_day)

Пример диаграммы в Excel. Новости по категориям за день.

