**Проект создания ETL-процесса для анализа публикуемых новостей.**

Проект создает ETL-процесс формирования витрины данных для анализа публикаций новостей. В качестве источников данных используются rss-ленты новостей следующих ресурсов: <https://lenta.ru/rss>, <https://www.vedomosti.ru/rss/news>, <https://tass.ru/rss/v2.xml>

Итоговая витрина данных содержит следующие данные:

* 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 новости
* Количество публикаций новостей данной категории по дням недели

Т.к. в разных источниках категории новостей отличаться, был проведен их анализ, в результате выделены обобщенные категории по всем источникам.

**Стек технологий.**

Т.к. источники данных предоставляют структурированный набор данных в rss-формате небольшого размера, в качестве базы данных используется PostgreSql.

Для обработки данных применяется язык python, со следующими библиотеками:

* psycopg2 – используется для работы с PostgreSql
* requests – используется для работы с HTTP-запросами
* beautifulsoup4, lxml – используется для парсинга xml (rss-формата)
* python-dotenv – используется для обработки .env файла

Для оркестрации процесса обработки данных используется - Apache Airflow

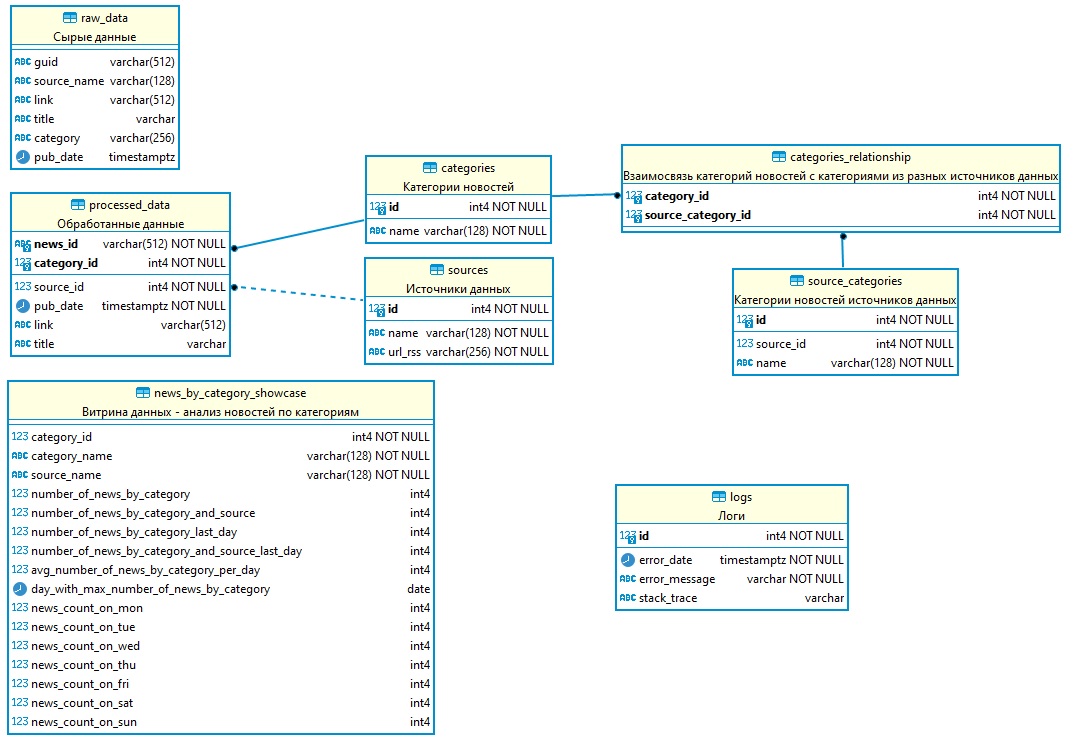
Для автоматизации создания, управления и развертывания приложения используется – Docker

**Реализация.**

Структуру хранения данных состоит из трех слоев:

* Сырой слой данных
* Промежуточный слой
* Слой витрин

ER-диаграмма БД:



1. Сырые данные из источников сохраняются в таблицу raw\_data без изменений.
2. Перед каждой загрузкой, таблица raw\_data очищается.
3. Перед загрузкой в сырой слой, все данные валидируются на корректность, все некорректные данные помещаются в таблицу logs.
4. Данные обрабатываются на основе таблицы categories\_relationship, в которой лежат данные взаимосвязи итоговых категорий новостей, c категориями из разных источников данных. Все обработанные данные сохраняются в таблицу processed\_data. Если таблица processed\_data еще пустая в нее добавляются все данные из raw\_data, иначе только новые данные, которые ранее не были добавлены.
5. На основе данных из processed\_data, создается итоговая витрина данных, которая сохраняется в таблицу news\_by\_category\_showcase
6. Apache Airflow последовательно запускает все этапы один раз в сутки

Все источники данных хранятся в таблицы sources, в нее можно добавлять дополнительные источники.

**Результаты разработки, выводы:**

В результате разработки сформирован ETL-процесс формирования витрины данных для анализа публикаций новостей. Полученные данные помогут проанализировать наиболее востребованные категории и источники новостей.

**Для запуска системы:**

1. Перед запуском в корне проекта создать файл .env с параметрами, описанными в файле .env.example.
2. Проинициализировать airflow командой: docker compose up airflow-init
3. Запускаем контейнеры: docker compose up –d
4. Для входа в airflow используем url: http://localhost:8080