Анализ публикуемых новостей

Голунова Анастасия, DE

Цель

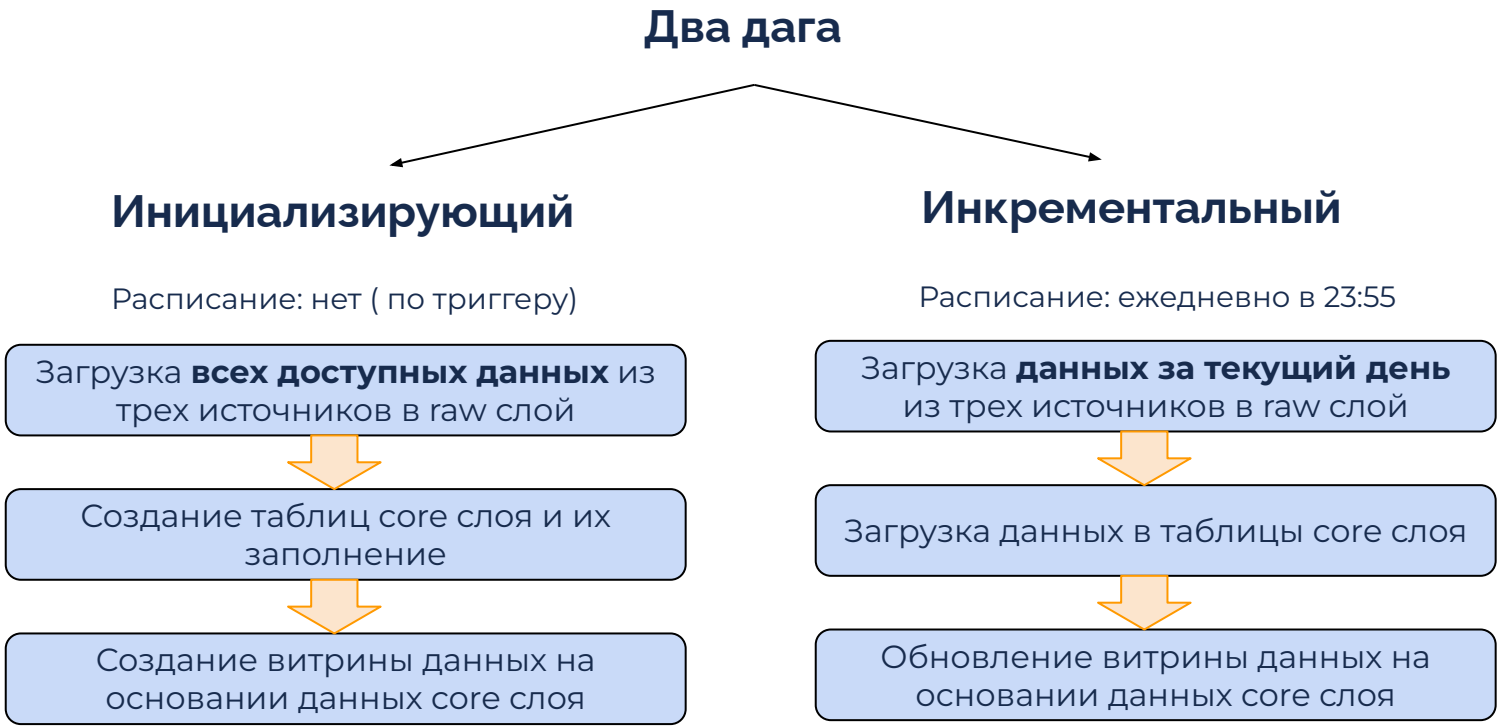
Создать ETL-процесс формирования витрины данных для анализа публикаций новостей из источников:

- https://lenta.ru/rss/
- https://www.vedomosti.ru/rss/news
- https://tass.ru/rss/v2.xml

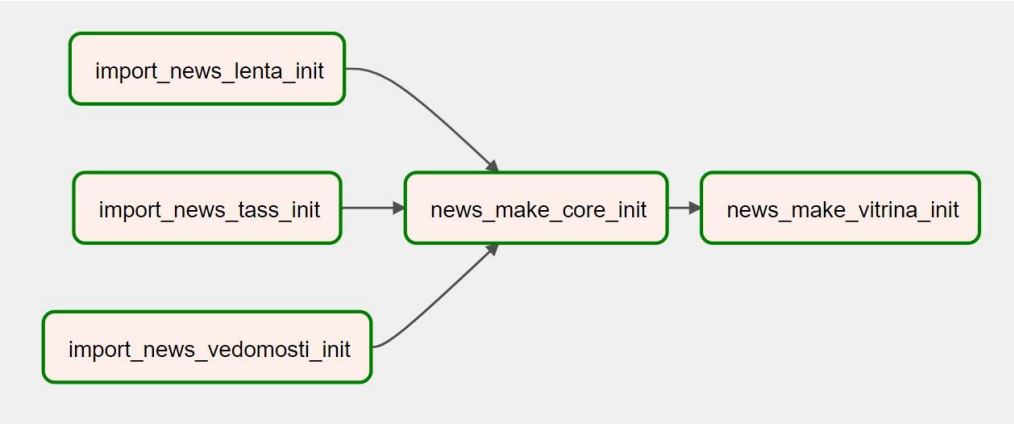
Инструменты

1. Для хранения и обработки данных по новостям выбрана СУБД Postgres, т.к. данные не большие (не более 1000 новостей в день со всех источников)
2. В качестве оркестратора ETL процесса выбран Apache Airflow
3. Таски дагов написаны при помощи Python оператора

План реализации

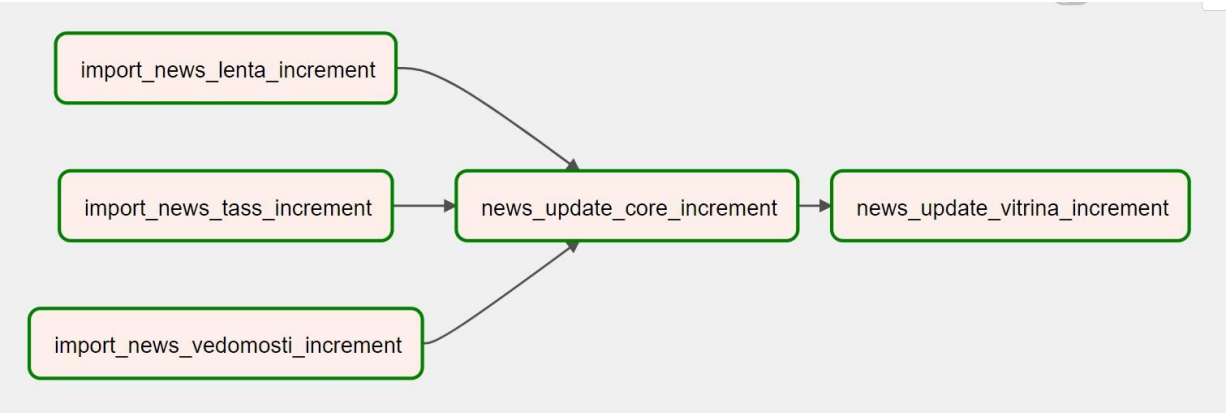


Инициализирующий даг



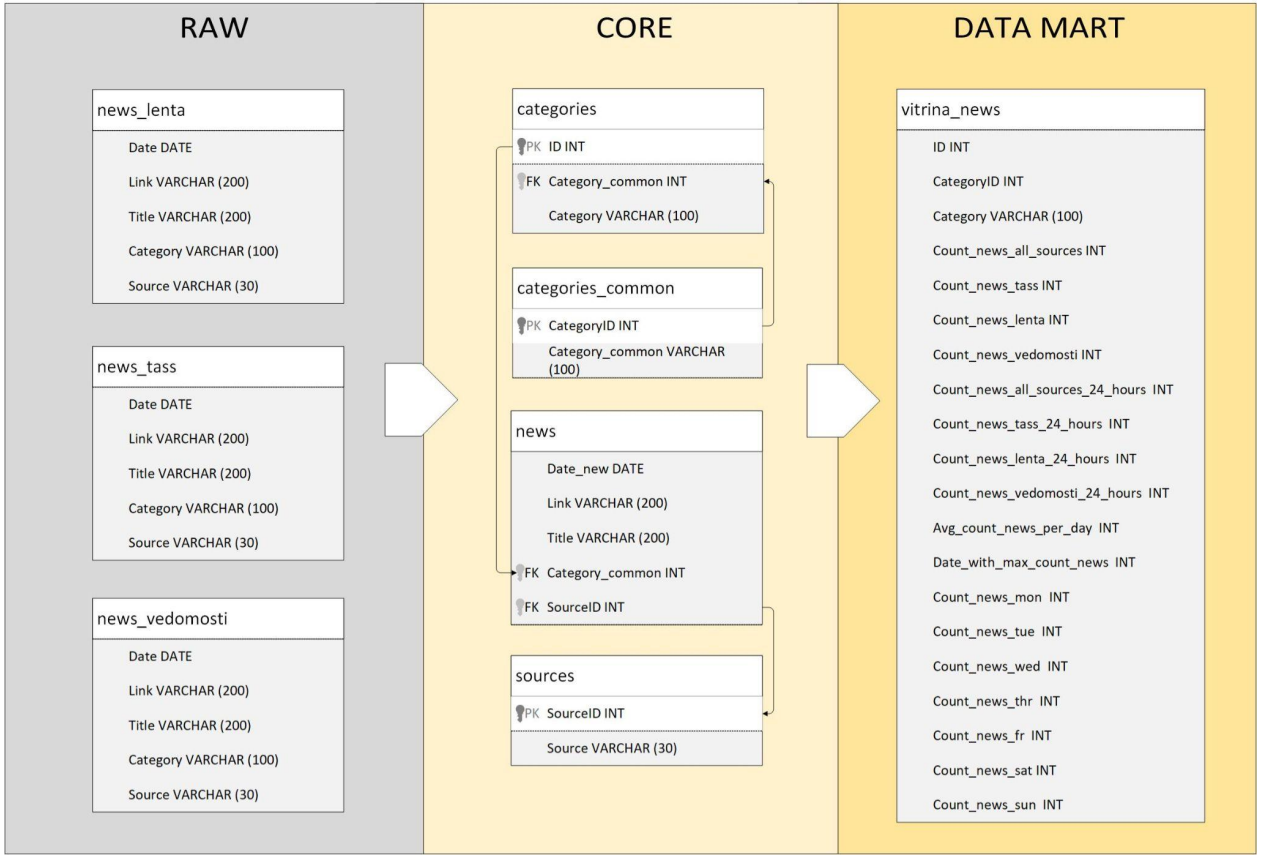
1. import_news_lenta_init, import_news_tass_init, import_news_vedomosti_init - загружают все доступные данные из трех источников, формируют датафреймы и в виде датафреймов отправляют в Postgres с методом replace.
2. news_make_core_init - создает таблицы core слоя и заполняет их данными с raw слоя
3. news_make_vitrina_init - создает витрину данных на основе данных core слоя

Инкрементальный даг



1. import_news_lenta_inrement, import_news_tass_inrement, import_news_vedomosti_inrement - загружают данные за текущий день из трех источников, формируют датафреймы и в виде датафреймов отправляют в Postgres с методом append. Время для дага выбрано, так как источник Тасс передает только данные за текущий день.
2. news_update_core_inrement - апдейтит таблицы core слоя данными с raw слоя
3. news_update_vitrina_inrementt - апдейтит витрину данных на основе данных core слоя

ER схема


RAW
CORE
DATA MART
news_lenta
Date DATE
Link VARCHAR (200)
Title VARCHAR (200)
Category VARCHAR (100)
Source VARCHAR (30)
news_tass
Date DATE
Link VARCHAR (200)
Title VARCHAR (200)
Category VARCHAR (100)
Source VARCHAR (30)
news_vedomosti
Date DATE
Link VARCHAR (200)
Title VARCHAR (200)
Category VARCHAR (100)
Source VARCHAR (30)
categories
PK ID INT
FK Category_common INT
Category VARCHAR (100)
categories_common
PK CategoryID INT
Category_common VARCHAR (100)
news
Date_new DATE
Link VARCHAR (200)
Title VARCHAR (200)
FK Category_common INT
FK SourceID INT
sources
PK SourceID INT
Source VARCHAR (30)
vitrina_news
ID INT
CategoryID INT
Category VARCHAR (100)
Count_news_all_sources INT
Count_news_tass INT
Count_news_lenta INT
Count_news_vedomosti INT
Count_news_all_sources_24_hours INT
Count_news_tass_24_hours INT
Count_news_lenta_24_hours INT
Count_news_vedomosti_24_hours INT
Avg_count_news_per_day INT
Date_with_max_count_news INT
Count_news_mon INT
Count_news_tue INT
Count_news_wed INT
Count_news_thr INT
Count_news_fr INT
Count_news_sat INT
Count_news_sun INT

ER схема. Комментарии

1. Схема core слоя - звезда
2. В разных источниках категории новостей отличаются, поэтому создана доп. таблица с унифицированными категориями. НО так как результаты витрины зависят от выбранных общих категорий - лучше их уточнять у аналитиков.

```
SET Category_common =
    CASE
        WHEN Category LIKE ANY (ARRAY['%кономика%', '%изнес']) THEN 'Экономика/Бизнес'
        WHEN Category LIKE ANY (ARRAY['%нвестиции%', '%инансы%']) THEN 'Финансы/ Инвестиции'
        WHEN Category LIKE ANY (ARRAY['%реда обитания%', '%бщество%', '%жизни%', '%енности%',  '%абота%']) THEN 'Общество'
        WHEN Category LIKE ANY (ARRAY['%стран%', '%оссия%', '%9 паралл%']) THEN 'Россия'
        WHEN Category LIKE ANY (ARRAY['%осква%', '%еверо-Запад%']) THEN 'Москва'
        WHEN Category LIKE ANY (ARRAY['%олити%', '%иловые%', '%рмия%']) THEN 'Политика'
        WHEN Category LIKE ANY (ARRAY['%ехнолог%', '%аука%']) THEN 'Технологии'
        WHEN Category LIKE ANY (ARRAY['%еждунар%', 'Мир%']) THEN 'Мир'
        WHEN Category like '%ультур%' THEN 'Культура'
        WHEN Category LIKE ANY (ARRAY['%нтернет%', '%овости партнер%']) THEN 'Интернет/ СМИ'
        WHEN Category like '%утешествия%' THEN 'Путешествия'
        WHEN Category like '%роисшествия%' THEN 'Происшествия'
        WHEN Category like '%едвижимост%' THEN 'Недвижимость'
        WHEN Category like '%порт%' THEN 'Спорт'
        ELSE 'Другое'
    END
```

Результат

Разработан ETL-процесс формирования витрины данных для анализа публикаций новостей из трех источников.

Скрин

	category	count_	cou	cc	cour	count	cou	cour	cou	avg_count_n	date_wi	cc	count_	count_	cou	cou	cou	c
1	Экономика/Бизнес	80	18	30	32	68	18	30	20	1,6326530612	2023-09-19	8	62	0	0	0	8	0
2	Финансы/ Инвестиции	15	[NULL]	[NULL]	15	14	[NULL]	[NULL]	14	1	2023-09-18	12	3	0	0	0	0	0
3	Общество	154	34	32	88	108	34	32	42	1,6923076923	2023-09-19	28	80	0	0	0	26	0
4	Россия	30	[NULL]	30	[NULL]	30	[NULL]	30	[NULL]	1,0344827586	2023-09-19	0	30	0	0	0	0	0
5	Москва	8	8	[NULL]	[NULL]	8	8	[NULL]	[NULL]	1	2023-09-19	0	8	0	0	0	0	0
6	Политика	259	35	13	211	131	35	13	83	1,8633093525	2023-09-19	66	71	0	0	0	52	0
7	Технологии	20	[NULL]	11	9	17	[NULL]	11	6	1	2023-09-19	3	15	0	0	0	1	0
8	Мир	64	22	42	[NULL]	64	22	42	[NULL]	1,0158730159	2023-09-19	0	64	0	0	0	0	0
9	Культура	18	4	14	[NULL]	18	4	14	[NULL]	2	2023-09-19	0	18	0	0	0	0	0
10	Путешествия	8	[NULL]	8	[NULL]	8	[NULL]	8	[NULL]	1	2023-09-19	0	8	0	0	0	0	0
12	Спорт	22	8	14	[NULL]	22	8	14	[NULL]	2	2023-09-19	0	22	0	0	0	0	0
13	Другое	36	1	25	10	34	1	25	8	1	2023-09-19	7	28	0	0	0	0	0
14	Интернет/ СМИ	10	[NULL]	10	[NULL]	10	[NULL]	10	[NULL]	1	2023-09-19	0	10	0	0	0	0	0
15	Происшествия	7	7	[NULL]	[NULL]	7	7	[NULL]	[NULL]	1	2023-09-19	0	7	0	0	0	0	0