Проект "Анализ публикуемых новостей"

Комарова А.В.

Цель и задачи:

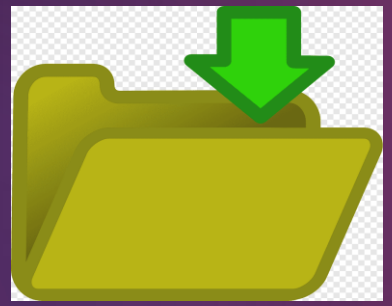
- Цель: создать ETL-процесс формирования витрин данных для анализа публикаций новостей

- Задачи:
  - Разработать скрипты загрузки данных в 2-х режимах:
  - o Инициализирующий – загрузка полного слепка данных источника
  - o Инкрементальный – загрузка дельты данных за прошедшие сутки

  - Организовать правильную структуру хранения данных
  - o Сырой слой данных
  - o Промежуточный слой
  - o Слой витрин

Структура хранения данных и используемые технологии:

Инициализирующий режим:

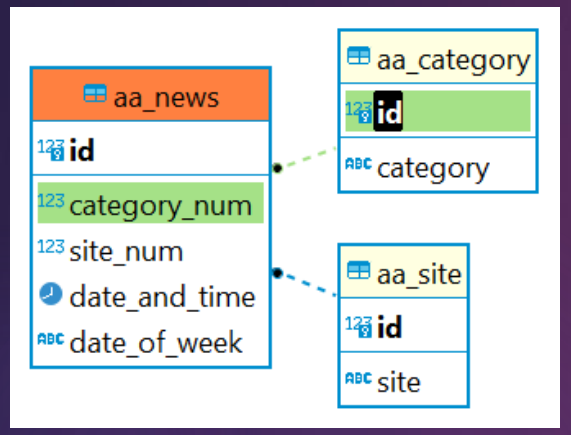


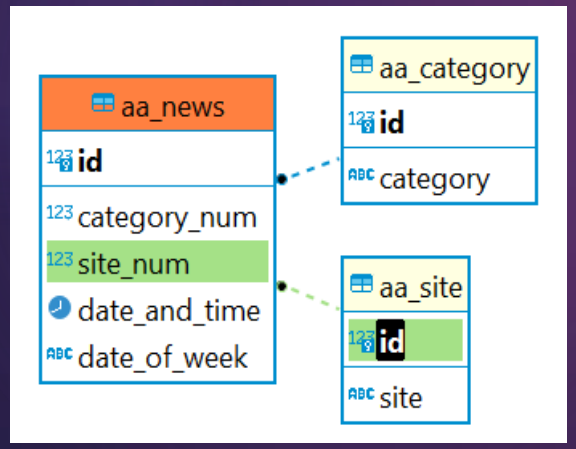
СЫРОЙ СЛОЙ данных храним в json

Инкрементальный режим:

ER-диаграмма и таблицы

Связи один ко многим.





Промежуточный слой данных держим в PostgreSQL

1	0	ВЕДОМОСТИ
2	1	TASS
3	2	LENTA.RU
4	3	Фонтонка.ру

Таблица / Текст

	id	category
1	0	Наука и техника
2	1	Всякое другое
3	2	Политика
4	3	Экономика и биз
5	4	Мир
6	5	Спорт
7	6	Происшествия
8	7	Культура
9	8	Армия и ОПК
10	9	Россия
11	10	Здоровье
12	11	Интернет и СМИ
13	12	Путешествия

Таблица / Текст / Запись

	id	category_num	site_num	date_and_time	date_of_w
1	1	0	0	2022-12-29 22:41:42.000	Thu
2	2	1	0	2022-12-29 22:19:00.000	Thu
3	3	1	0	2022-12-29 21:57:07.000	Thu
4	4	2	0	2022-12-29 21:32:44.000	Thu
5	5	3	0	2022-12-29 21:12:16.000	Thu
6	6	2	0	2022-12-29 20:52:17.000	Thu
7	7	3	0	2022-12-29 20:44:34.000	Thu
8	8	2	0	2022-12-29 20:22:18.000	Thu
9	9	1	0	2022-12-29 19:58:09.000	Thu
10	10	2	0	2022-12-29 19:34:50.000	Thu
11	11	1	0	2022-12-29 19:18:13.000	Thu
12	12	2	0	2022-12-29 19:13:57.000	Thu
13	13	2	0	2022-12-29 19:06:35.000	Thu
14	14	3	0	2022-12-29 18:49:34.000	Thu
15	15	2	0	2022-12-29 18:40:23.000	Thu
16	16	0	0	2022-12-29 18:37:56.000	Thu
17	17	1	0	2022-12-29 18:27:03.000	Thu
18	18	1	0	2022-12-29 18:18:45.000	Thu
19	19	3	0	2022-12-29 18:04:57.000	Thu
20	20	1	0	2022-12-29 17:54:31.000	Thu
21	21	1	0	2022-12-29 17:47:28.000	Thu
22	22	1	0	2022-12-29 17:39:04.000	Thu

Витрины данных

categories(+) 1

```
with num as( select category_n
```

Введите SQL выражение чтобы

Таблица | Текст | Запись

	id	category	date_of_week	sum_pub
1	0	Наука и техника	Wed	744 329 865
2	0	Наука и техника	Tue	400 986 039
3	0	Наука и техника	Thu	1 291 270 321
4	0	Наука и техника	Fri	640 948 959
5	1	Всякое другое	Wed	1 178 660 045
6	1	Всякое другое	Tue	834 181 629
7	1	Всякое другое	Thu	2 204 166 943
8	1	Всякое другое	Fri	1 126 266 225
9	2	Политика	Wed	1 412 890 490
10	2	Политика	Tue	746 789 316
11	2	Политика	Thu	1 882 222 131
12	2	Политика	Fri	496 698 810
13	3	Экономика и биз	Wed	1 346 945 898
14	3	Экономика и биз	Tue	717 690 506
15	3	Экономика и биз	Thu	2 092 167 732
16	3	Экономика и биз	Fri	649 780 918
17	4	Интернет и СМИ	Thu	1 724 779 226
18	4	Интернет и СМИ	Fri	1 069 830 708
19	5	Мир	Thu	2 175 424 017
20	5	Мир	Fri	1 015 475 918
21	6	Армия и ОПК	Thu	1 009 754 494
22	6	Армия и ОПК	Fri	785 994 193
23	7	Происшествия	Thu	786 639 297

Строятся в PostgreSQL, оркестрируются с помощью Apache AirFlow.

categories 1

```
with nnum_4 as ( with num_4 a
```

Введите SQL выражение чтобы отфильтровать результаты

Таблица | Текст

	id	category	sum_last_day	sum_all	avg_all	avg_per_last_day
1	0	Наука и техника	1 873 517 607	2 984 281 555	167 525	201 489
2	1	Всякое другое	3 290 809 441	5 259 953 861	124 682	139 629
3	2	Политика	2 320 980 016	4 448 140 174	117 436	190 076
4	3	Экономика и бизнес	2 726 699 612	4 784 222 426	118 821	111 390
5	4	Интернет и СМИ	2 767 649 720	2 767 649 720	132 709	147 290
6	5	Мир	3 168 521 051	3 168 521 051	111 678	131 336
7	6	Армия и ОПК	1 783 533 573	1 783 533 573	141 764	141 617
8	7	Происшествия	1 398 903 520	1 398 903 520	165 943	179 710
9	8	Культура	1 191 943 534	1 191 943 534	109 867	125 831
10	9	Спорт	1 678 955 414	1 678 955 414	127 668	168 012
11	10	Россия	1 128 881 824	1 128 881 824	104 808	144 420
12	11	Здоровье	787 909 715	787 909 715	141 710	185 176
13	12	Путешествия	653 109 883	653 109 883	147 997	215 166

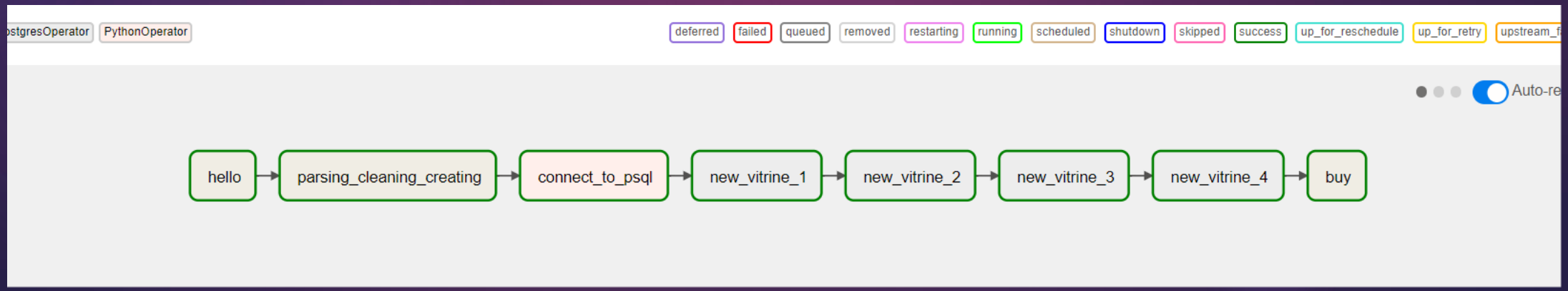
Результаты:

Инициализирующий режим:

- парсинг сайтов новостей,
- сохранение сырых данных в json,
- очистка данных,
- формирование датафреймов,
- подключение к БД,
- создание таблиц в БД,
- загрузка всех таблиц в БД

Инкрементальный режим:

- парсинг сайтов новостей, начиная с последней сохраненной новости
- сохранение дельты сырых данных в json,
- очистка данных,
- формирование датафреймов,
- подключение к БД,
- загрузка таблицы с новостями в БД,
- создание витрин



ВЫВОДЫ:

Профессия Дата Инженер ох какая не простая,
но очень интересная!

Большое спасибо за курс!