

Clickhouse для аналитиков

со знанием SQL

Как пришли к Clickhouse

Возникла потребность в хранении логов Метрики (Metrika Logs API).

Был Hive. Залили hits. 2 ТБ данных.

запустили `select count(*) from hits..` 17 минут.

Долго мучались.

Поставили Clickhouse, залили hits.. 250 Гб.

`select count from hits(),` 3 секунды. (8GB RAM)

Мучаться перестали.

Что получили:

Всего 2 таблицы для ежедневного обновления (visits & hits). - не нужно просить программистов настроить сбор данных в определенном срезе - всё на месте.

Удобный и красивый графический интерфейс, доступный из браузера (tabix).

http интерфейс - автоматическое обновление отчетов где угодно, например в Excel и PowerBI.

Целый набор функций специфичных для данных из логов Метрики. (работа с массивами, обработка урлов, параметров визитов итд.)

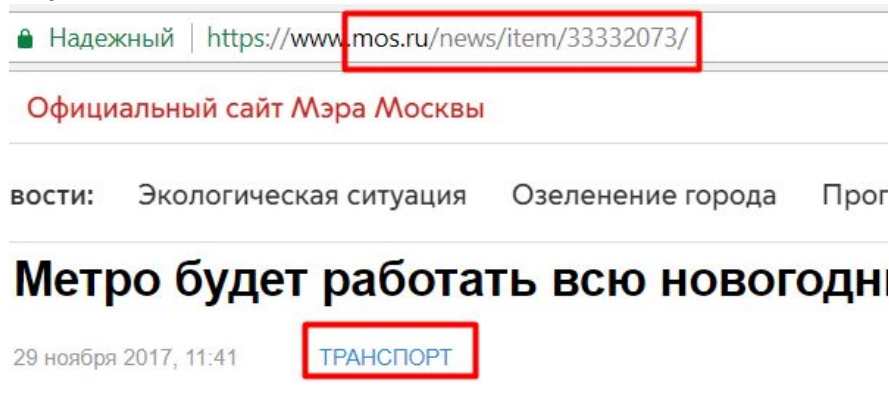
Время на анализ, а не на сбор и обработку данных.

Зачем аналитикам сырые данные (пример 1)

Настройка целей задним числом.

Задача - посмотреть динамику посещаемости новостей из разных сфер.

Определить сферу ни по URL ни по заголовку нельзя - только настройка целей.



Надежный | <https://www.mos.ru/news/item/33332073/>

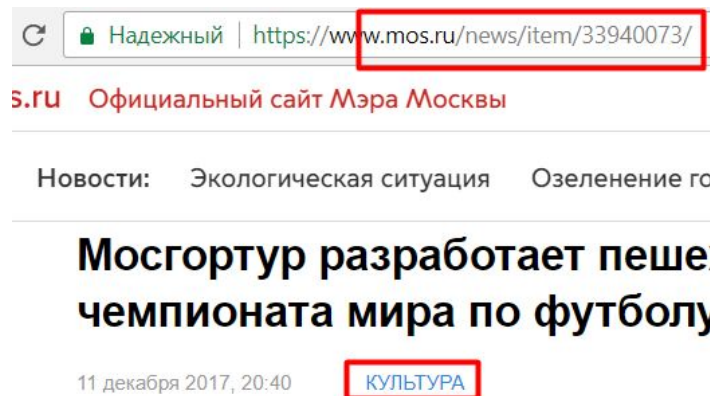
Официальный сайт Мэра Москвы

вости: Экологическая ситуация Озеленение города Прог

Метро будет работать всю новогодн

29 ноября 2017, 11:41

TRANСПОРТ



Надежный | <https://www.mos.ru/news/item/33940073/>

s.ru Официальный сайт Мэра Москвы

Новости: Экологическая ситуация Озеленение го

Мосгортур разработает пеше: чемпионата мира по футболу

11 декабря 2017, 20:40

КУЛЬТУРА

Зачем аналитикам сырые данные (пример 1)

Если к логам добавить
табличку с соответствием URL
и сферы - задача решается за
28 сек при 1 млрд. строк на
настройках “из коробки”

```
1 select toStartOfMonth(date) as month, cat2, uniq(client_id)
   users
2 from hits
3 any inner join dict_url using url
4 where 1=1
5     and date between [redacted]
6     and cat1 = 'Новости'
7     and artificial = 0
8     and cat1 = 'Новости'
9 group by month, cat2
```

RUN ALL ⬆ + ⚙ + 📄 RUN CURRENT ⚙ + 📄 🗄 📁

USE default ▼ 28.61 sec. | 947,171,430 rows. | 39 GB

TABLE 🗃 DRAW 📊 📌

	month	cat2
1	2017-09-01	Юбилей Москвы
2	2017-08-01	День города
3	2017-09-01	Moscow housing renc
4	2017-08-01	My Street
5	2017-11-01	Science and innovat

Зачем аналитикам сырые данные (пример 2)

Нестандартные метрики

Имея сырые данные с лёгким доступом можно выйти за рамки стандартных метрик, не выходя за рамки SQL.





Например, можно посчитать “прочитаемость” статей.

```
1- select cat2, countIf(spent_on_page >= 30) reads, count() pv, reads / pv as read_rate
2- from (
3-   with if(runningDifference(client_id) = 0, 1, 0) as same_client_id,
4-        - runningDifference(date_time) * same_client_id as spent_on_page
5-   select client_id,
6-          date_time,
7-          cat2,
8-          spent_on_page
9-   from (
10-     select client_id,
11-            date_time,
12-            url, cat2
13-     from hits
14-     any inner join dict_url using url
15-     where date between [redacted]
16-            and cat1 = 'Новости'
17-            and client_id <> 0
18-            and artificial = 0
19-     order by client_id, date_time desc
20-   )
21-   where spent_on_page <= 30*60
22- ) group by cat2
23 limit 100
```

RUN ALL ⚙ + ⌨ + ⌨ RUN CURRENT ⚙ + ⌨ :: 📄 USE default ▾

11.78 sec. | 633,876,468 rows. | 27 GB

TABLE 📊 DRAW 📈

	cat2	reads	pv	
1		321	1125	
2		35	123	
3		9	31	
4		1362	5241	

Спасибо за внимание!

Вопросы.