Гибкое хранение данных в ClickHouse
Для чего нужны тома и диски?
Как справляются сейчас?
Что уже есть в ClickHouse?
Секретный слайд
Что появилось нового?
Что собираемся сделать ещё?
Ссылки
Вопросы и ответы

Гибкое хранение данных в ClickHouse на нескольких дисковых томах

Владимир Чеботарёв

Altinity Ltd

vchebotarev@altinity.com

11 декабря 2019 г.

Немного о себе

Гибкое хранение данных в ClickHouse

Для чего нужны тома и диски?

Как справляются сейчас?

Что уже есть в ClickHouse?

Секретный слайд

Что появилось нового?

Что собираемся сделать ещё?

Ссылки

Вопросы и ответы

- последние 3 месяца работаю над ClickHouse в компании Altinity
- работал в Яндекс.Метрике, в Лаборатории Касперского и в Deutsche Bank
- всегда мечтал работать у Лёши в команде

Немного об Altinity

Мы делаем ClickHouse ещё быстрее и ещё удобнее!

Как мы это делаем:

- дорабатываем ClickHouse вместе с Яндексом
  95 пулл реквестов в основной репозиторий в 2018-2019[1]
- разрабатываем экосистемные проекты
  например, C++ и ODBC-драйверы, плагин для Grafana, оператор для k8s
- обеспечиваем 24x7 поддержку ClickHouse-инсталляций
- обучаем и помогаем строить решения на ClickHouse

Содержание

1. Для чего нужны тома и диски?
2. Как справляются сейчас?
3. Что уже есть в ClickHouse?
4. Секретный слайд
5. Что появилось нового?
6. Что собираемся сделать ещё?
7. Ссылки
8. Вопросы и ответы

Для чего нужны тома и диски?

Сколько стоит хранение данных?

Гибкое хранение данных в ClickHouse

Для чего нужны тома и диски?

Как справляются сейчас?

Что уже есть в ClickHouse?

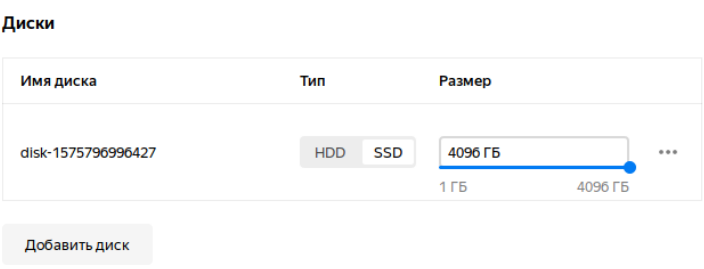
Секретный слайд

Что появилось нового?

Что собираемся сделать ещё?

Ссылки

Вопросы и ответы



Для чего нужны тома и диски?

Сколько стоит хранение данных?

Гибкое хранение данных в ClickHouse

Для чего нужны тома и диски?

Как справляются сейчас?

Что уже есть в ClickHouse?

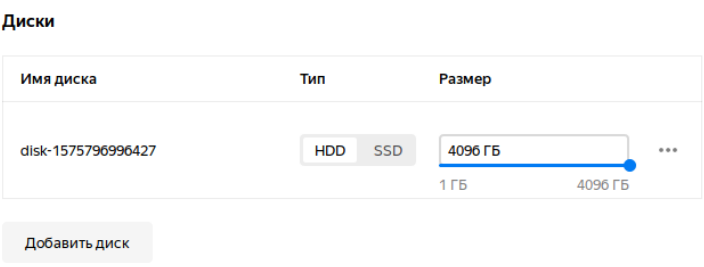
Секретный слайд

Что появилось нового?

Что собираемся сделать ещё?

Ссылки

Вопросы и ответы



Для чего нужны тома и диски?

Сколько стоит хранение данных?

Гибкое хранение данных в ClickHouse

Для чего нужны тома и диски?

Как справляются сейчас?

Что уже есть в ClickHouse?

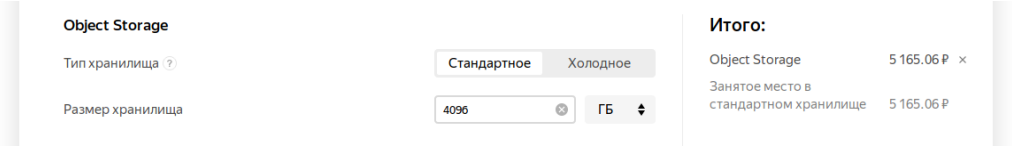
Секретный слайд

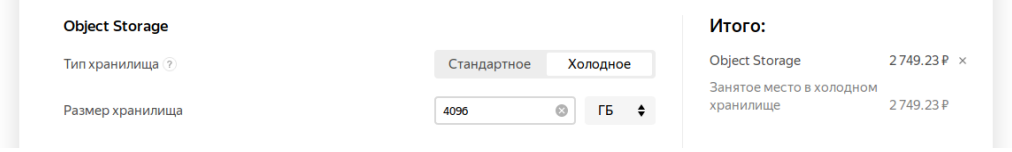
Что появилось нового?

Что собираемся сделать ещё?

Ссылки

Вопросы и ответы





Для чего нужны тома и диски?

Для чего же нужны тома и диски?

Гибкое хранение данных в ClickHouse

Для чего нужны тома и диски?

Как справляются сейчас?

Что уже есть в ClickHouse?

Секретный слайд

Что появилось нового?

Что собираемся сделать ещё?

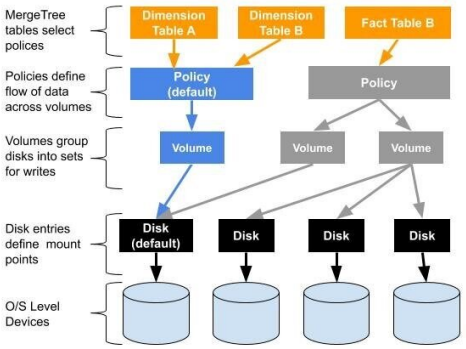
Ссылки

Вопросы и ответы

- стоимость
- максимальный размер
- распределение нагрузки
- скорость

Для чего нужны тома и диски?
Что хотят люди?

Гибкое хранение данных в ClickHouse
Для чего нужны тома и диски?
Как справляются сейчас?
Что уже есть в ClickHouse?
Секретный слайд
Что появилось нового?
Что собираемся сделать ещё?
Ссылки
Вопросы и ответы



[3]

- создавать гетерогенные хранилища данных, даже на одной реплике

Гибкое хранение данных в ClickHouse
Для чего нужны тома и диски?
Как справляются сейчас?
Что уже есть в ClickHouse?
Секретный слайд
Что появилось нового?
Что собираемся сделать ещё?
Ссылки
Вопросы и ответы

Как справляются сейчас?

- различные техники кэширования
  - HDFS + Alluxio[5], 2019
  - HDFS + GridGain[6], 2019
  - и многие другие
- нормальные решения
  - Amazon UltraWarm[7], 2019
  - HDFS[8], 2016
  - Vertica
  - Oracle
- никому не известные решения
  - HPC + Data Jockey[9], 2019
  - OpenEdge
  - OctopusFS[10], 2017
  - hStorage-DB[11], 2012
  - и другие

Что уже есть в ClickHouse?
Терминология

Гибкое хранение данных в ClickHouse
Для чего нужны тома и диски?
Как справляются сейчас?
Что уже есть в ClickHouse?
Секретный слайд
Что появилось нового?
Что собираемся сделать ещё?
Ссылки
Вопросы и ответы

- диск
- диск по умолчанию
- том
- политика хранения

Что уже есть в ClickHouse?
Что умеет ClickHouse?

Гибкое хранение данных в ClickHouse

Для чего нужны тома и диски?

Как справляются сейчас?

Что уже есть в ClickHouse?

Секретный слайд

Что появилось нового?

Что собираемся сделать ещё?

Ссылки

Вопросы и ответы

- автоматически перемещать куски данных при заполнении диска на один из томов, следующих за томом текущего диска в данной политике
- вручную перемещать партиции или куски на заданный диск или том

Гибкое хранение данных в ClickHouse
Для чего нужны тома и диски?
Как справляются сейчас?
Что уже есть в ClickHouse?
Секретный слайд
Что появилось нового?
Что собираемся сделать ещё?
Ссылки
Вопросы и ответы

Конфигурация

```
<yandex>
    <storage_configuration>
        <disks>
            <ram32>
                <path>/var/lib/clickhouse/ram32/</path>
            </ram32>
        </disks>

        <policies>
            <default_with_ram32>
                <volumes>
                    <ram32>
                        <disk>ram32</disk>
                    </ram32>
                    <main>
                        <disk>default</disk>
                    </main>
                </volumes>
            </default_with_ram32>
        </policies>
    </storage_configuration>
<yandex>
```

Что уже есть в ClickHouse?

Как выглядит работа с дисками?

Информация о дисках

```
SELECT *
FROM system.disks
```

name	path	free_space	total_space	keep_free_space
default	/var/lib/clickhouse/	74803624960	502468107264	1024
external	/var/lib/clickhouse/external/	74803625984	502468108288	0
jbod1	/var/lib/clickhouse/jbod1/	74803625984	502468108288	0
jbod2	/var/lib/clickhouse/jbod2/	10485760	10485760	0
ram32	/var/lib/clickhouse/ram32/	26017378304	34359738368	0

```
5 rows in set. Elapsed: 0.004 sec.
```

Гибкое хранение данных в ClickHouse
Для чего нужны тома и диски?
Как справляются сейчас?
Что уже есть в ClickHouse?
Секретный слайд
Что появилось нового?
Что собираемся сделать ещё?
Ссылки
Вопросы и ответы

Информация о политиках хранения

```
SELECT *
FROM system.storage_policies

┌─policy_name─────────────┬─volume_name─┬─volume_priority─┬─disks─────────────┬─max_data_part_size─┬─move_factor─┐
│ default                 │ default     │               1 │ ['default']       │                  0 │           0 │
│ default_with_ram32      │ ram32       │               1 │ ['ram32']         │                  0 │         0.1 │
│ default_with_ram32      │ main        │               2 │ ['default']       │                  0 │         0.1 │
│ jbods_with_external     │ main        │               1 │ ['jbod1','jbod2'] │           10485760 │         0.1 │
│ jbods_with_external     │ external    │               2 │ ['external']      │                  0 │         0.1 │
│ small_jbod_with_external │ main        │               1 │ ['jbod1']         │                  0 │         0.1 │
│ small_jbod_with_external │ external    │               2 │ ['external']      │                  0 │         0.1 │
└─────────────────────────┴─────────────┴─────────────────┴───────────────────┴────────────────────┴─────────────┘

7 rows in set. Elapsed: 0.003 sec.
```

Гибкое хранение данных в ClickHouse
Для чего нужны тома и диски?
Как справляются сейчас?
Что уже есть в ClickHouse?
Секретный слайд
Что появилось нового?
Что собираемся сделать ещё?
Ссылки
Вопросы и ответы

Создание таблицы

```
CREATE TABLE reddit_ram
(
    `author` String,
    `body` String,
    `created_utc` DateTime,
    `downs` Int64,
    `id` String,
    `link_id` String,
    `parent_id` String,
    `score` Int64,
    `subreddit` String,
    `subreddit_id` String,
    `ups` Int64
)
ENGINE = MergeTree()
PARTITION BY toDate(created_utc)
ORDER BY link_id
SETTINGS index_granularity = 8192, storage_policy = 'default_with_ram32'

Ok.

0 rows in set. Elapsed: 0.011 sec.
```

Что уже есть в ClickHouse?

Как выглядит работа с дисками?

Перемещение куска

```
ALTER TABLE reddit_ram
    MOVE PART '20150101_42_42_0' TO DISK 'default'


Ok.

0 rows in set. Elapsed: 0.177 sec.
```

Перемещение партиции

```
ALTER TABLE reddit_ram
    MOVE PARTITION '2015-01-02' TO VOLUME 'main'


Ok.

0 rows in set. Elapsed: 1.249 sec.
```

Что уже есть в ClickHouse?

Как выглядит работа с дисками?

Как узнать, где лежит кусок?

```
SELECT
    partition_id,
    name,
    disk_name,
    data_compressed_bytes,
    path
FROM system.parts
WHERE (table = 'reddit_ram') AND (active = 1) AND (partition_id = '20150101')
```

partition_id	name	disk_name	data_compressed_bytes	path
20150101	20150101_5_35_1	ram32	98834419	/var/lib/clickhouse/ram32/data/default/reddit_ram/20150101_5_35_1/
20150101	20150101_42_42_0	default	16648331	/var/lib/clickhouse/data/default/reddit_ram/20150101_42_42_0/
20150101	20150101_47_47_0	ram32	14094509	/var/lib/clickhouse/ram32/data/default/reddit_ram/20150101_47_47_0/
20150101	20150101_49_49_0	ram32	22941919	/var/lib/clickhouse/ram32/data/default/reddit_ram/20150101_49_49_0/
20150101	20150101_56_56_0	ram32	10957197	/var/lib/clickhouse/ram32/data/default/reddit_ram/20150101_56_56_0/

```
5 rows in set. Elapsed: 0.008 sec.
```

Гибкое хранение данных в ClickHouse

Для чего нужны тома и диски?

Как справляются сейчас?

Что уже есть в ClickHouse?

Секретный слайд

Что появилось нового?

Что собираемся сделать ещё?

Ссылки

Вопросы и ответы

- управлять механизмом автоматического перемещения кусков, в зависимости от данных

Торжественное вливание нового функционала

Гибкое хранение данных в ClickHouse

Для чего нужны тома и диски?

Как справляются сейчас?

Что уже есть в ClickHouse?

Секретный слайд

Что появилось нового?

Что собираемся сделать ещё?

Ссылки

Вопросы и ответы

Нужен доброволец!

Что появилось нового?

Что теперь умеет ClickHouse (если всё прошло успешно)?

Гибкое хранение данных в ClickHouse

Для чего нужны тома и диски?

Как справляются сейчас?

Что уже есть в ClickHouse?

Секретный слайд

Что появилось нового?

Что собираемся сделать ещё?

Ссылки

Вопросы и ответы

- автоматически перемещать куски данных на нужный том или диск, в зависимости от их “возраста”*

Что появилось нового?

Синтаксис

Создание таблицы

```
CREATE TABLE {table-name}
(
    {columns-list}
)
ENGINE = MergeTree()
PARTITION BY {partition-by-expr}
ORDER BY {order-by-expr}
TTL {time-or-datetime-expr} [DELETE|TO DISK 'disk-name'|TO VOLUME 'volume-name'], ...
SETTINGS {settings-list}
```

Изменение структуры таблицы

```
ALTER {table-name}
    MODIFY TTL {time-or-datetime-expr} [DELETE|TO DISK 'disk-name'|TO VOLUME 'volume-name'], ...
```

Что появилось нового?

Как настраивать автоматическое перемещение?

Создание таблицы

```
CREATE TABLE reddit_ram
(
    `author` String,
    `body` String,
    `created_utc` DateTime,
    `downs` Int64,
    `id` String,
    `link_id` String,
    `parent_id` String,
    `score` Int64,
    `subreddit` String,
    `subreddit_id` String,
    `ups` Int64
)
ENGINE = MergeTree()
PARTITION BY toDate(created_utc)
ORDER BY link_id
TTL created_utc + toIntervalDay(7) TO VOLUME 'main'
SETTINGS index_granularity = 8192, storage_policy = 'default_with_ram32'

Ok.

0 rows in set. Elapsed: 0.011 sec.
```

Что появилось нового?

Какая логика изменилась?

Гибкое хранение данных в ClickHouse

Для чего нужны тома и диски?

Как справляются сейчас?

Что уже есть в ClickHouse?

Секретный слайд

Что появилось нового?

Что собираемся сделать ещё?

Ссылки

Вопросы и ответы

- INSERT
- слияния
- фоновые перемещения

Гибкое хранение данных в ClickHouse
Для чего нужны тома и диски?
Как справляются сейчас?
Что уже есть в ClickHouse?
Секретный слайд
Что появилось нового?
Что собираемся сделать ещё?
Ссылки
Вопросы и ответы

Как хранится информация о TTL в кусках?

```
# cat /var/lib/clickhouse/data/default/reddit_ram/20150101_1_1_0/ttl.txt
ttl format version: 1
{"moves":[{"expression":"plus(created_utc, toIntervalDay(7))","min":1420678855,"max":1420715278}]}
```

Что собираемся сделать ещё?

Что осталось за кадром?

Гибкое хранение данных в ClickHouse

Для чего нужны тома и диски?

Как справляются сейчас?

Что уже есть в ClickHouse?

Секретный слайд

Что появилось нового?

Что собираемся сделать ещё?

Ссылки

Вопросы и ответы

- поддержка изменения конфигурации налету
- более хорошая интеграция с ALTER
- throttling операций перемещения
- системная таблица с информацией о перемещениях
- более хорошая интеграция с CollapsingMergeTree и другими разновидностями
- более умные согласованные перемещения

Что собираемся сделать ещё?
Что будет дальше?

Гибкое хранение данных в ClickHouse

Для чего нужны тома и диски?

Как справляются сейчас?

Что уже есть в ClickHouse?

Секретный слайд

Что появилось нового?

Что собираемся сделать ещё?

Ссылки

Вопросы и ответы

- Object Storage как один из уровней хранения данных

Ссылки I

Гибкое хранение данных в ClickHouse
Для чего нужны тома и диски?
Как справляются сейчас?
Что уже есть в ClickHouse?
Секретный слайд
Что появилось нового?
Что собираемся сделать ещё?
Ссылки
Вопросы и ответы


Pull Requests by Altinity to ClickHouse/ClickHouse (2018-2019)
https://github.com/ClickHouse/ClickHouse/pulls?q=is:pr+label:altinity


Mikhail Filimonov (2019)
Do-It-Yourself Multi-Volume Storage in ClickHouse
https://www.altinity.com/blog/2019/3/5/do-it-yourself-multi-volume-storage-in-clickhouse


Mikhail Filimonov (2019)
Amplifying ClickHouse Capacity with Multi-Volume Storage (Part 1)
https://www.altinity.com/blog/2019/11/27/amplifying-clickhouse-capacity-with-multi-volume-storage-part-1


Mikhail Filimonov (2019)
Amplifying ClickHouse Capacity with Multi-Volume Storage (Part 2)
https://www.altinity.com/blog/2019/11/29/amplifying-clickhouse-capacity-with-multi-volume-storage-part-2

Ссылки II

Гибкое хранение данных в ClickHouse
Для чего нужны тома и диски?
Как справляются сейчас?
Что уже есть в ClickHouse?
Секретный слайд
Что появилось нового?
Что собираемся сделать ещё?
Ссылки
Вопросы и ответы


Alluxio: In Memory Distributed Storage (2019)
http://www.alluxio.org/


GridGain In-Memory Computing Platform (2019)
http://www.gridgain.com/


Steve Roberts (2019)
Announcing UltraWarm (Preview) for Amazon Elasticsearch Service
https://aws.amazon.com/blogs/aws/announcing-ultrawarm-preview-for-amazon-elasticsearch-service/


A. Agarwal (2016)
Enable Support for Heterogeneous Storages in HDFS
https://issues.apache.org/jira/browse/HDFS-2832

Ссылки III

Гибкое хранение данных в ClickHouse
Для чего нужны тома и диски?
Как справляются сейчас?
Что уже есть в ClickHouse?
Секретный слайд
Что появилось нового?
Что собираемся сделать ещё?
Ссылки
Вопросы и ответы


W. Shin and C. D. Brumgard and B. Xie and S. S. Vazhkudai and D. Ghoshal and S. Oral and L. Ramakrishnan (2019)
Data Jockey: Automatic Data Management for HPC Multi-tiered Storage Systems
2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)


Kakoulli, Elena and Herodotou, Herodotos (2017)
OctopusFS: A Distributed File System with Tiered Storage Management
2017 ACM International Conference


T. Luo, R. Lee, M. Mesnier, F. Chen, and X. Zhang (2012)
hStorage-DB: heterogeneity-aware data management to exploit the full capability of hybrid storage systems
VLDB Endowment, vol. 5, no. 10, pp. 1076-1087

Гибкое хранение данных в ClickHouse

Для чего нужны тома и диски?

Как справляются сейчас?

Что уже есть в ClickHouse?

Секретный слайд

Что появилось нового?

Что собираемся сделать ещё?

Ссылки

Вопросы и ответы

Вопросы и ответы

info@altinity.com
https://www.altinity.com
https://www.altinity.com/blog