Lessons from Self-Hosting ClickHouse

Boris Tane

boristane.com
@boristane

About me

Boris Tane

CLOUDFLARE®

↑

baselime

About Baselime

application observability for the next frontier of cloud computing

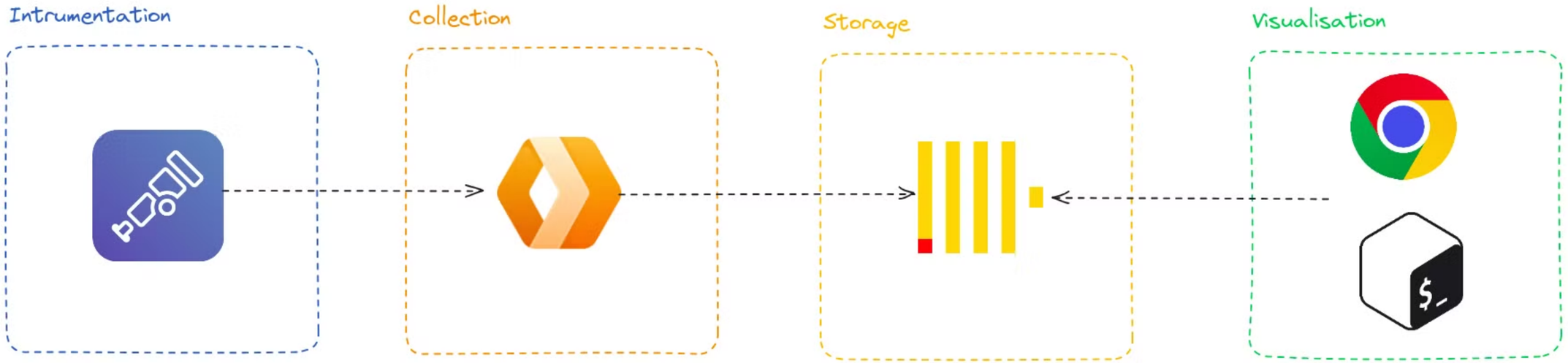
What is observability?

understanding how your software works
by looking at its external outputs

How much data?

~15,000+ events per second

How does it work?



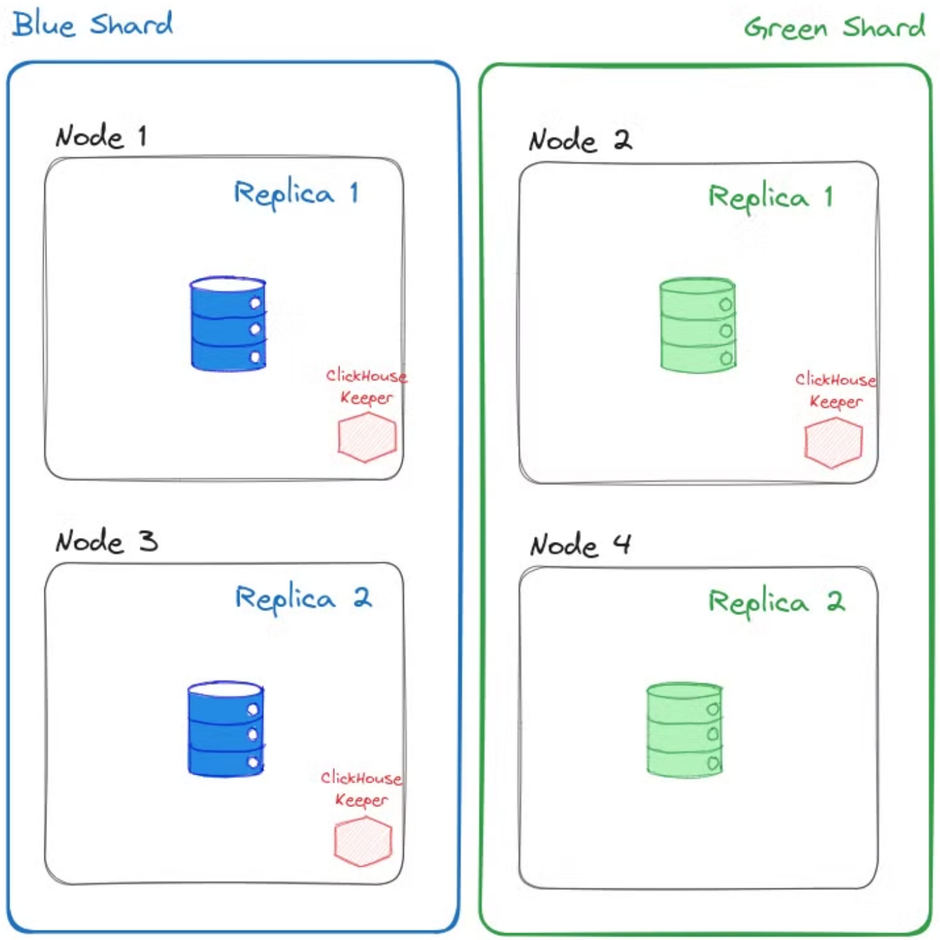
Why ClickHouse?

- Schemaless
- High Cardinality and Dimensionality
- Fast

Why Self-Host?

- More fun :)
- ClickHouse Cloud wasn't around yet
- Unit Economics

What's the architecture?



How's it going?

How's it going?

10 Lessons from operating ClickHouse

Lesson 1: Sending data in small batches is a very bad idea

Lesson 1: Sending data in small batches is a very bad idea

TOO_MANY_PARTS

Lesson 1: Sending data in small batches is a very bad idea

TOO_MANY_PARTS

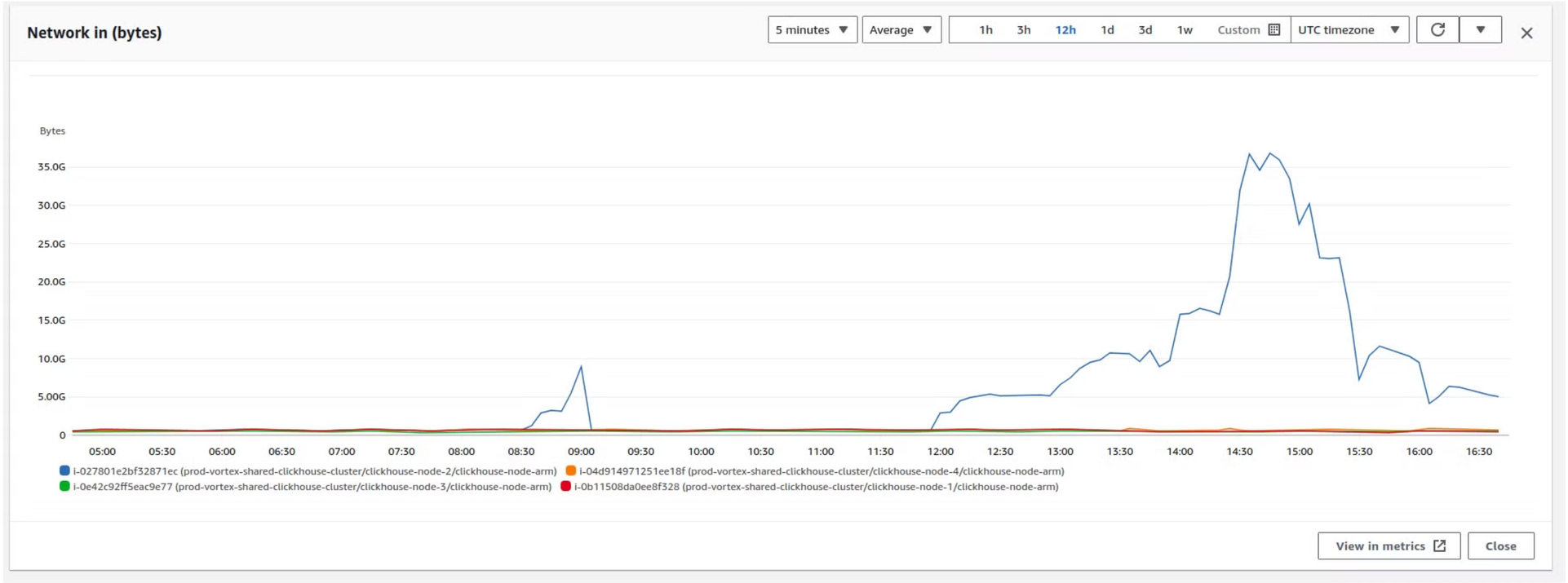
ClickHouse is struggling to merge all the parts as data is coming through

Lesson 2: If in doubt, check that you have enough disk space

Lesson 2: If in doubt, check that you have enough disk space

```
$ df -h
Filesystem       Size  Used Avail Use% Mounted on
/dev/root         88G  7.3G   80G   9% /
devtmpfs         7.7G     0  7.7G   0% /dev
tmpfs            7.7G     0  7.7G   0% /dev/shm
tmpfs            1.6G  972K  1.6G   1% /run
tmpfs            5.0M     0  5.0M   0% /run/lock
tmpfs            7.7G     0  7.7G   0% /sys/fs/cgroup
/dev/loop0        22M   22M     0 100% /snap/amazon-ssm-agent/7629
/dev/loop1        50M   50M     0 100% /snap/core18/2794
/dev/nvme0n1p15   98M  6.3M   92M   7% /boot/efi
/dev/loop2        36M   36M     0 100% /snap/snapd/20298
/dev/loop3        92M   92M     0 100% /snap/lxd/24065
/dev/nvme1n1     1.4T  1.4T   11G 100% /clickhouse/data
/dev/loop4        60M   60M     0 100% /snap/core20/2019
$
```

Lesson 2: If in doubt, check that you have enough disk space



Lesson 3: Beware of the disks you select

Lesson 3: Beware of the disks you select

16TB Limit for General Purpose SSDs On AWS

Lesson 4: Mutating data is a really bad idea

Lesson 4: Mutating data is a really bad idea

```
ALTER TABLE <table> ON CLUSTER <cluster>
MODIFY COLUMN <column> Float64;
```

Lesson 4: Mutating data is a really bad idea

ClickHouse creates a new version of the data and then merges it with the existing data

Lesson 4: Mutating data is a really bad idea

- Resource-intensive
- Significantly performance impact
- Potentially returns stale data to users

Lesson 5: ClickHouse TTLs mutate data

Lesson 5: ClickHouse TTLs mutate data

- Resource-intensive
- Performance impact

Lesson 5: ClickHouse TTLs mutate data

Instead, delete partitions individually

```
PARTITION BY toDate(timestamp)
```

Lesson 5: ClickHouse TTLs mutate data

Instead, delete partitions individually

```
ALTER TABLE <table> ON CLUSTER <cluster> DROP
PARTITION "2024-05-21"
```

Lesson 6: ClickHouse default values are pretty conservative, don't hesitate to increase them

Lesson 6: ClickHouse default values are pretty conservative, don't hesitate to increase them

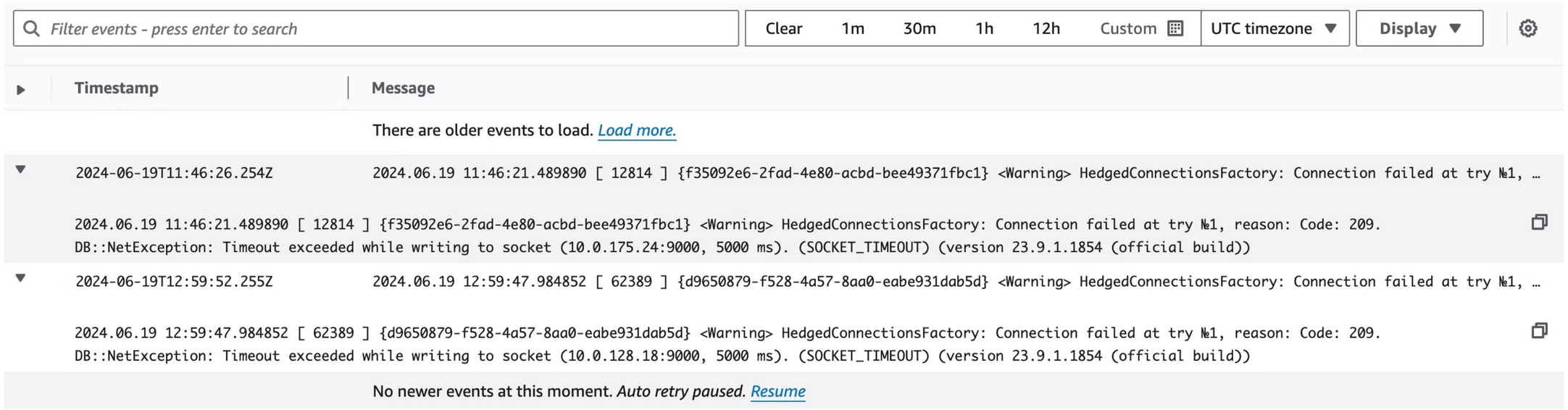
max_partition_size_to_drop: 50GB

max_table_size_to_drop: 50GB

keep_alive_timeout: 10s

max_concurrent_queries: now unlimited (was 100)

Lesson 7: Ship your ClickHouse Logs somewhere - don't rely on tail



Lesson 8: Ship your ClickHouse Metrics somewhere

```
"metrics_collected": {
      "disk": {
        "measurement": ["used_percent"],
        "metrics_collection_interval": 60,
        "resources": ["*"]
      },
      "mem": {
        "measurement": ["mem_used_percent"],
        "metrics_collection_interval": 60
      },
      "diskio": {
        "measurement": ["read_bytes", "write_bytes", "io_time"],
        "metrics_collection_interval": 60,
        "resources": ["/dev/root", "/dev/nvme1n1"]
      },
      "cpu": {
        "measurement": ["usage_idle", "usage_system", "usage_user"],
        "metrics_collection_interval": 60,
        "totalcpu": true
      },
      "net": {
        "measurement": ["bytes_sent", "bytes_recv", "packets_sent", "packets_recv", "drop_in", "drop_out"],
        "metrics_collection_interval": 60,
        "interface": "eth0"
      }
}
```

Lesson 8: Ship your ClickHouse Metrics somewhere

```
<clickhouse>
    <prometheus>
        <endpoint>/metrics</endpoint>
        <port>9363</port>
        <metrics>true</metrics>
        <events>true</events>
        <asynchronous_metrics>true</asynchronous_metrics>
        <errors>true</errors>
    </prometheus>
</clickhouse>

```

Lesson 8: Ship your ClickHouse Metrics somewhere

818 metrics

Lesson 8: Ship your ClickHouse Metrics somewhere

Total time spent merging parts

MergesTimeMilliseconds

Lesson 8: Ship your ClickHouse Metrics somewhere

Bytes sent over the network, for reads, writes and replications

NetworkSendBytes

Lesson 8: Ship your ClickHouse Metrics somewhere

Bytes received from the network, for reads, writes and replications

NetworkReceiveBytes

Lesson 8: Ship your ClickHouse Metrics somewhere

Slow writes

InsertQueryTimeMicroseconds

Lesson 8: Ship your ClickHouse Metrics somewhere

Slow reads

SelectQueryTimeMicroseconds

Lesson 9: Maintain a directory of useful ClickHouse queries

Lesson 9: Maintain a directory of useful ClickHouse queries

You'll need them when crap hits the fan at 2am

Lesson 9: Maintain a directory of useful ClickHouse queries

Recent ClickHouse Errors

```
select * from system.errors order by last_error_time desc
```

Lesson 9: Maintain a directory of useful ClickHouse queries

Replication queue - view queues with errors

```
select * from system.replication_queue limit 1 format Vertical;
```

Lesson 9: Maintain a directory of useful ClickHouse queries

Partitions still in sync - potential issues with merges

```
select * from system.merges where elapsed > 10 and progress <
0.99 limit 20 format Vertical;
```

Lesson 9: Maintain a directory of useful ClickHouse queries

List partitions - view potentially stale partitions

```
select partition, count(), sum(bytes_on_disk) as size from
system.parts group by partition order by size desc
```

Lesson 10: If something looks odd, even just slightly, something is wrong - check it before it gets worse

Thanks for having me here!

Boris Tane

boristane.com
@boristane

Pitch

Want to make a presentation like this one?

Start with a fully customizable template, create a beautiful deck in minutes, then easily share it with anyone.

Create a presentation (It's free)