CONTENTSQUARE

Experience Matters

ClickHouse at ContentSquare

Summary

What are we doing at ContentSquare?

What have we done so far with ClickHouse?

What do we think about ClickHouse?

Who are we

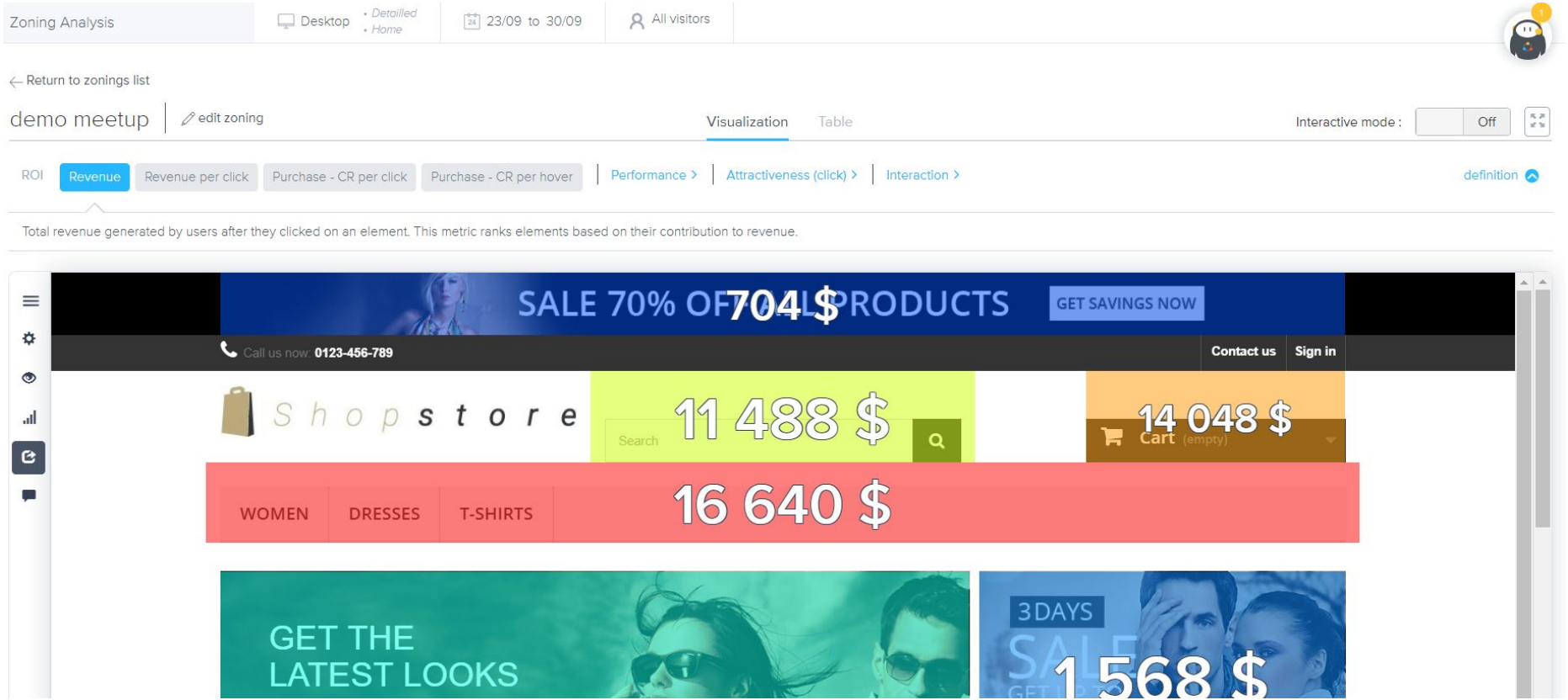
Christophe Kalenzaga
Senior Data Engineer

Vianney Foucault
Senior Platform Engineer

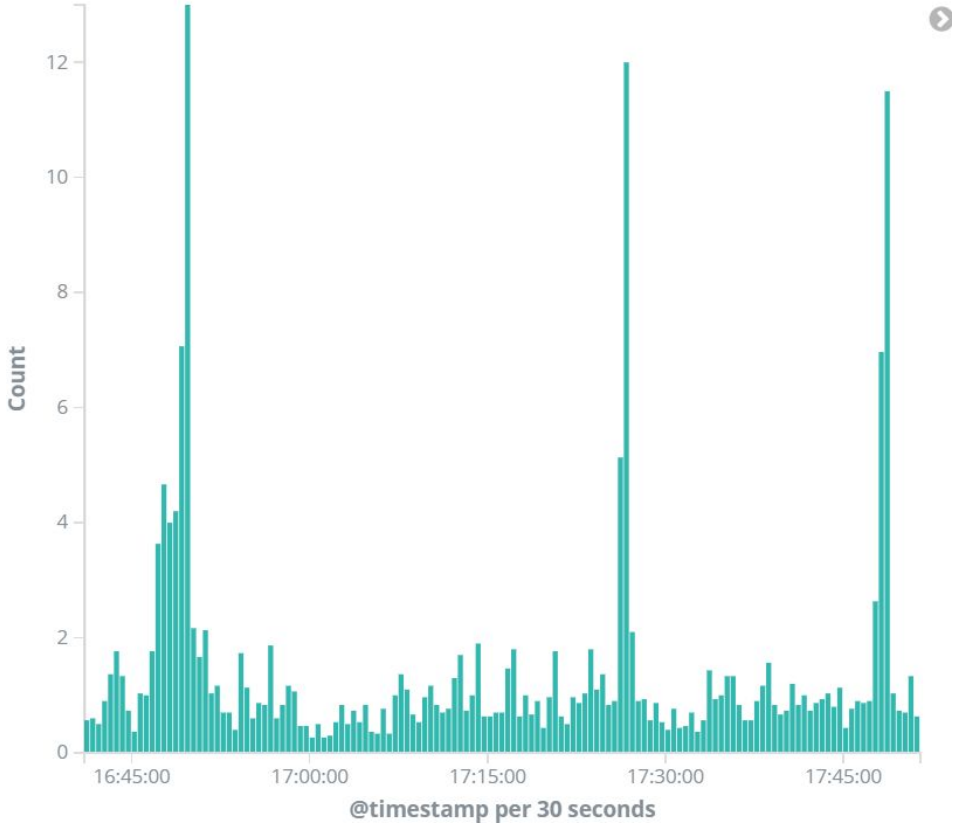
ContentSquare in a few words

- Web analytic Company
  - 600GB collected per day
  - 13 months retention (~230 TBytes)
  - websites and mobile applications
- 100+ clients Worldwide
  - Automotive | BFSI | Retail | Grocery | Luxury
  - B2B | Energy | Travel | Telco | Gaming
- Collected data doubles every year

One of the many challenges of ContentSquare

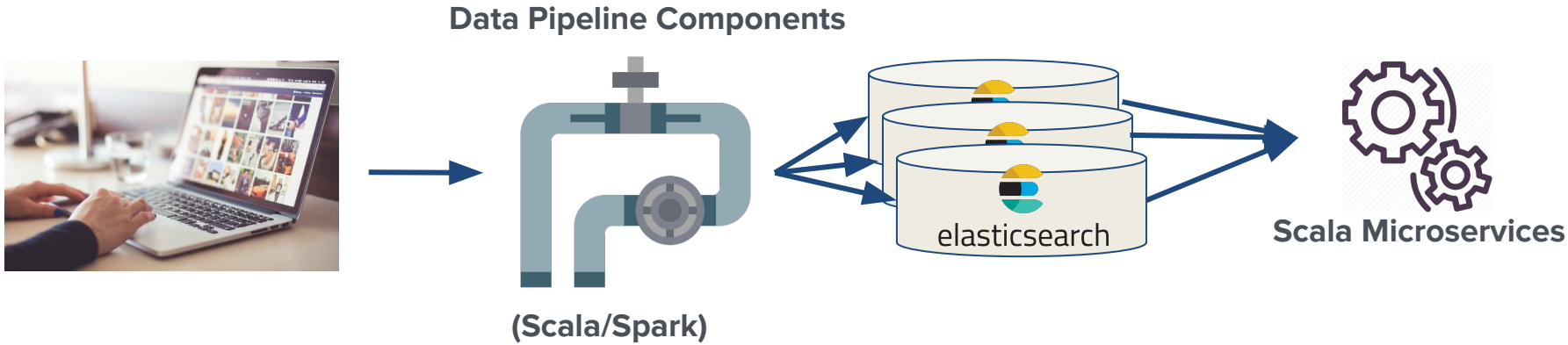


Our workload



- 30k queries per day
- Skewed load
- CPU intensive queries
- 20% of cacheable queries
- Can’t precompute queries

current backend infrastructure



- Many ElasticSearch clusters
- Storage cost
- Compute cost
- Can’t handle clients with too much data

to sum up

what we've done so far

Looking for a new technology

Selection of a few technologies to replace ElasticSearch

Start mobile analytics solution on ClickHouse

Benchmarks to find the right data model and the right configuration

Start migrating the web analytic solution to ClickHouse

insert data from the web analytics solution into ClickHouse check if there are no side effects

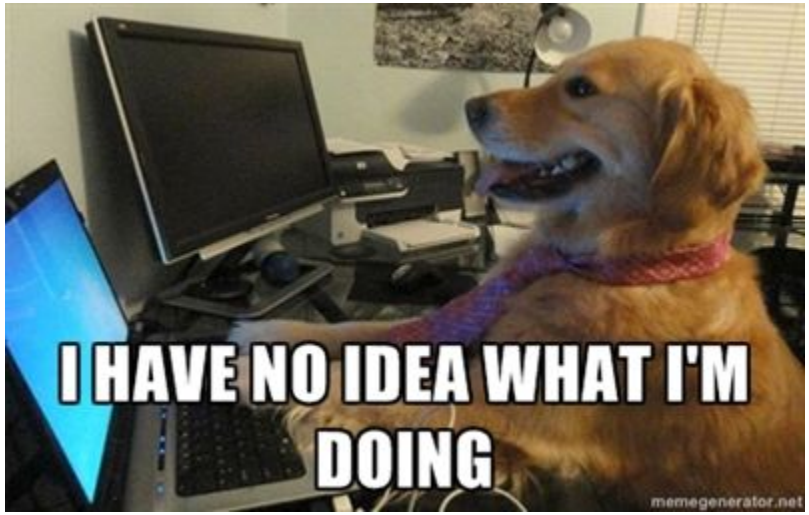
POC of multiple technologies

One-week POC on ClickHouse on a specific feature of the web analytics solution

the mobile analytics solution is released

First release of the mobile analytics solution on clickhouse.

How to benchmark?



Benchmark methodology

- Define different types of query
- Get the statistical distribution of each type of query in production
- Create a dataset (10 TBytes)
- Create queries with the same statistical distribution as in production
- Be deterministic

Benchmark Constraints for a #Platfomer

Automation

Cost Optimisation

CPU's

Deployment

Deterministic

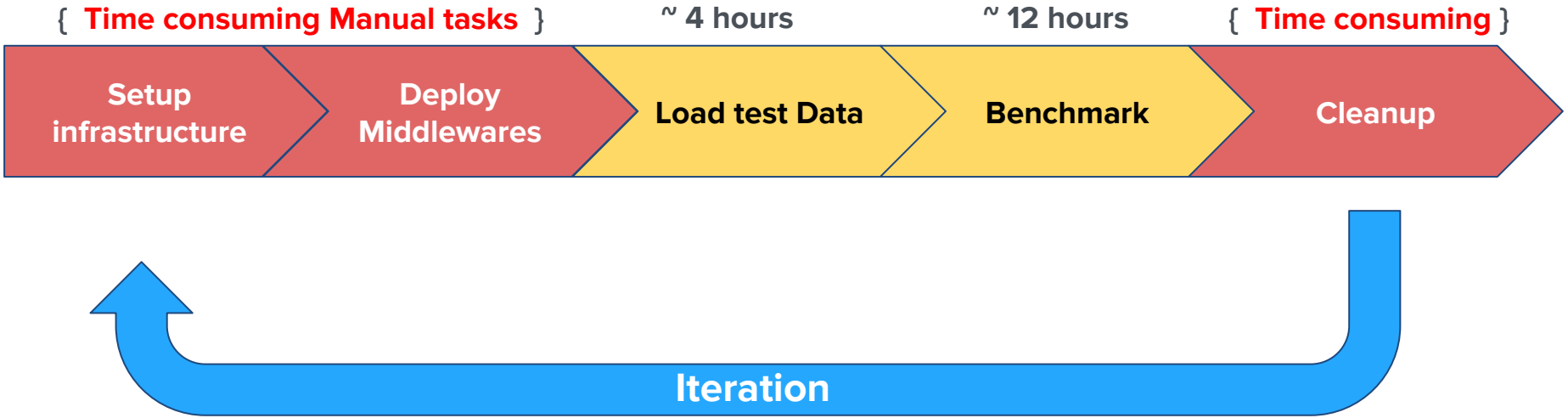
Configuration Management

Data Load

Memory

Monitoring

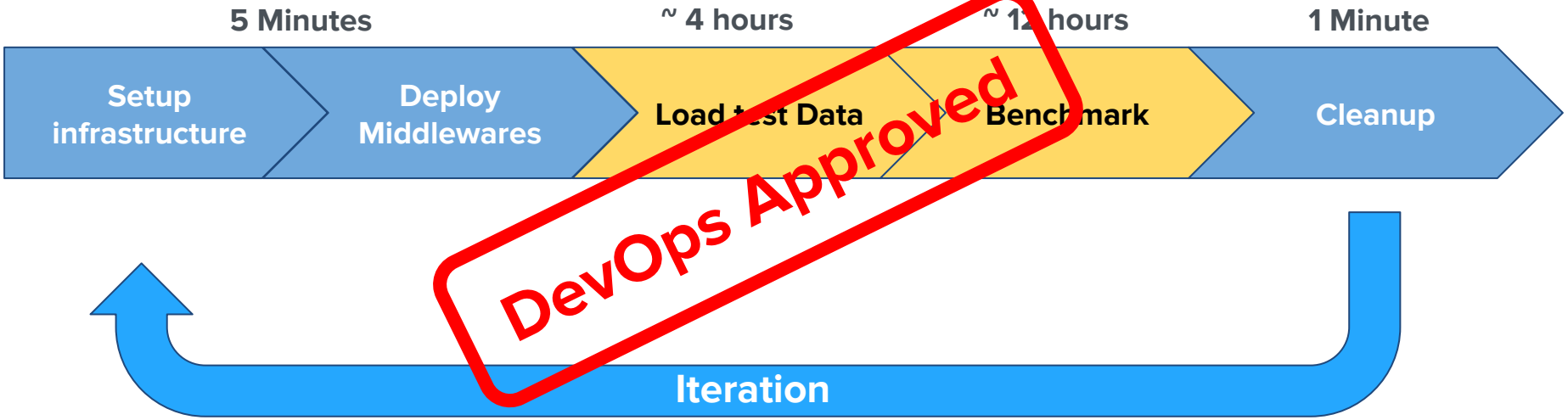
Benchmark timeline



DevOps Motto:

Industrialise As Soon As Possible

Fully Automated Benchmark timeline



HashiCorp Terraform

ANSIBLE

DevOps Leitmotiv:

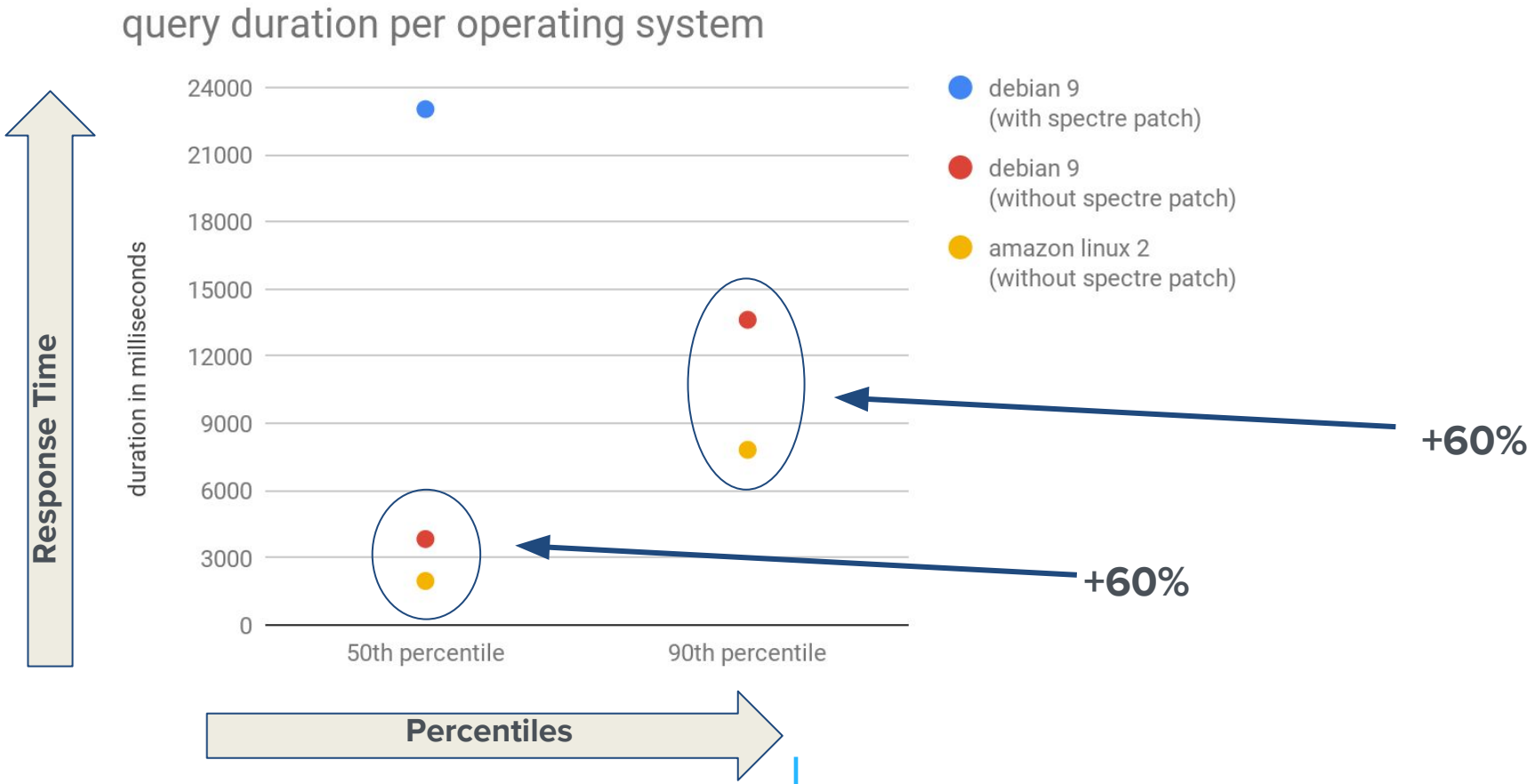
Monitor As Soon As Possible

Benchmark iterations on operating systems

Amazon Linux

http://openbenchmarking.org/result/1704225-TR-AMAZON38819

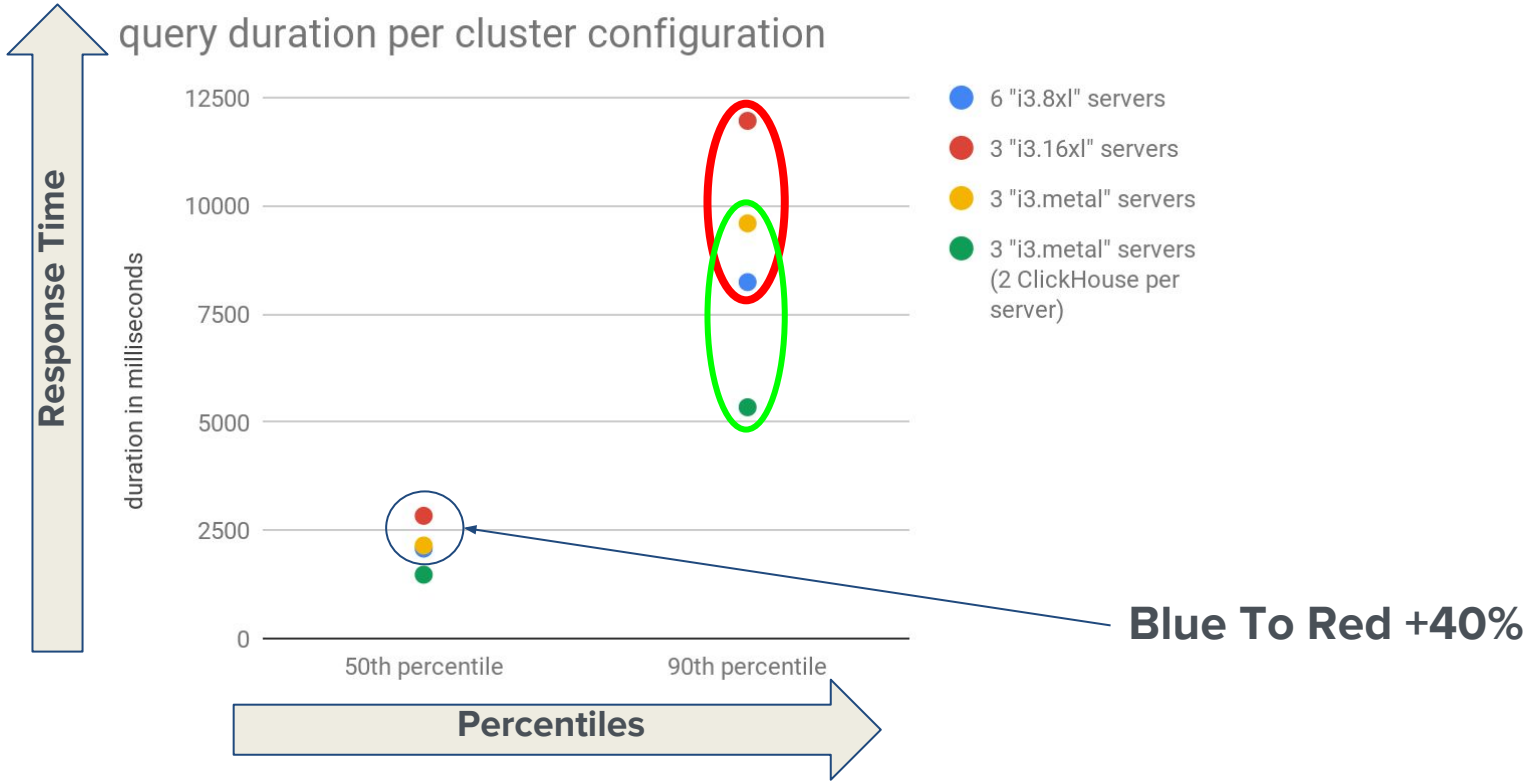
Benchmark iterations on different operating systems



Benchmark iterations on Instances

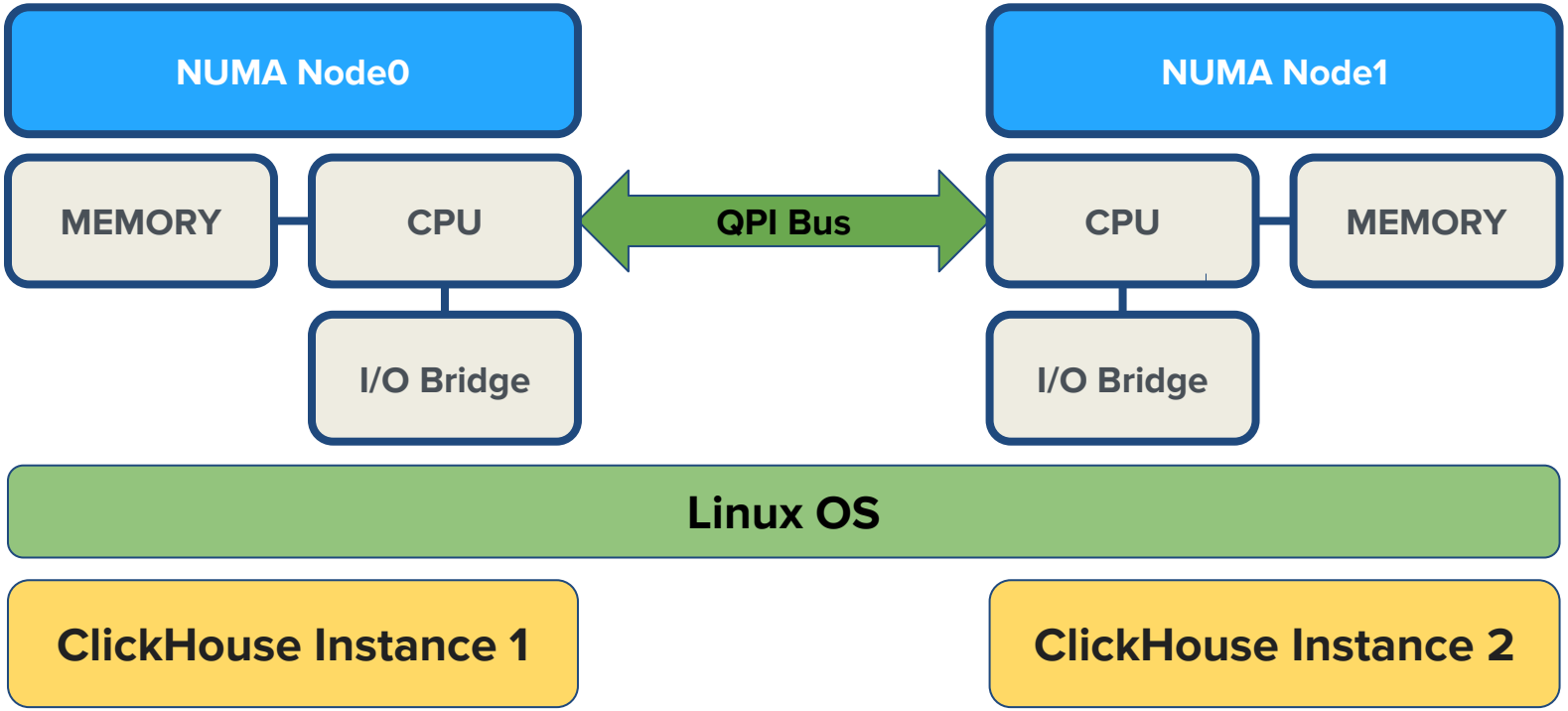
Instance Type	vCPU	Mem (GiB)	Networking Perf.	$ Per Hour	Perf Index
i3.large	2	15,25	Up to 10 Gigabit	0.172	1
i3.xlarge	4	30,5	Up to 10 Gigabit	0.344	1
i3.2xlarge	8	61	Up to 10 Gigabit	0.688	1
i3.4xlarge	16	122	Up to 10 Gigabit	1.376	1
i3.8xlarge	32	244	10 Gigabit	2.752	1
i3.16xlarge	64	488	25 Gigabit	5.504	1
i3.metal	72	512	25 Gigabit	5.504	1.3

Benchmark iterations on different Amazon servers

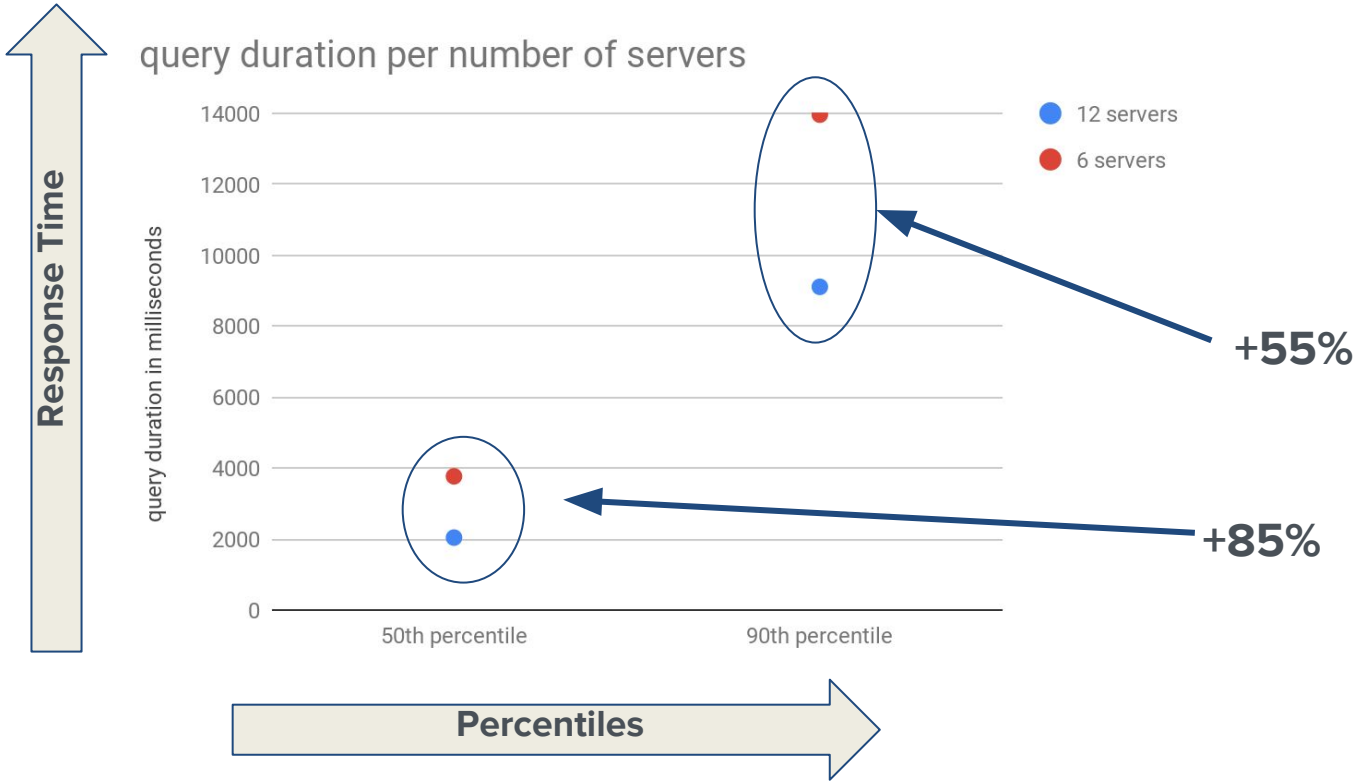


All the cluster configurations have the same price!

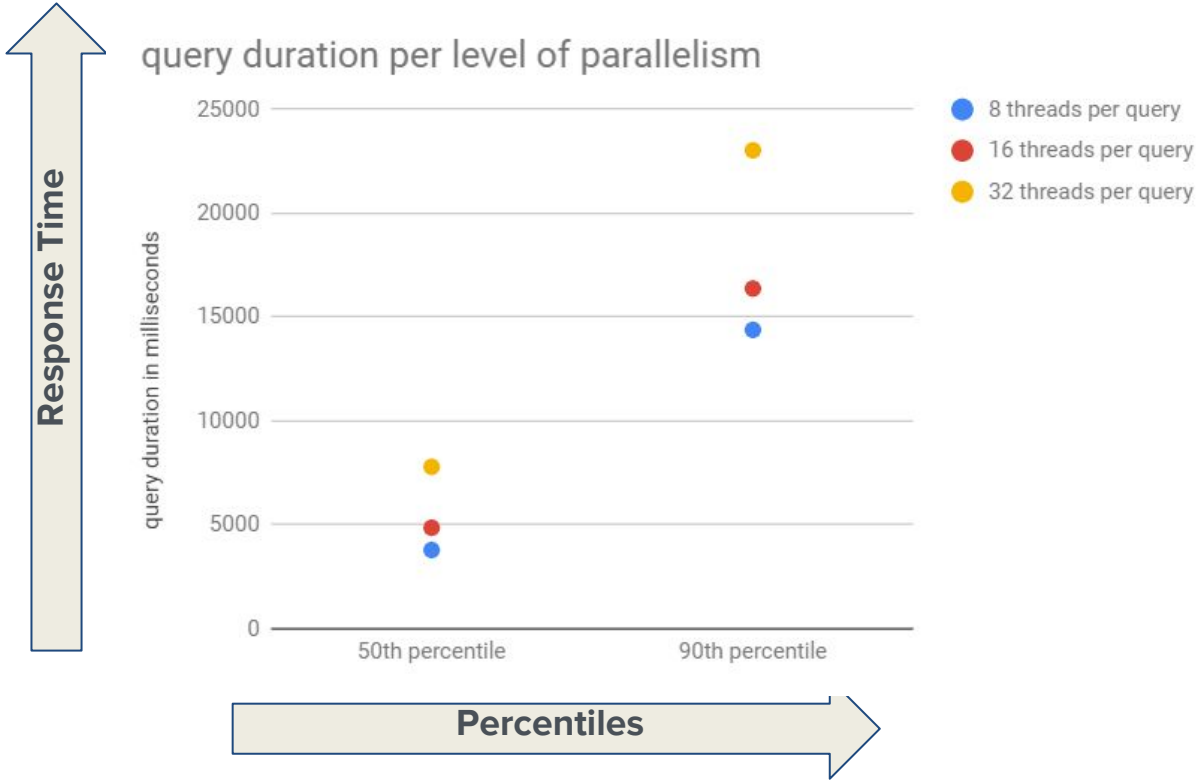
Numa Architecture



Benchmark iterations on scalability

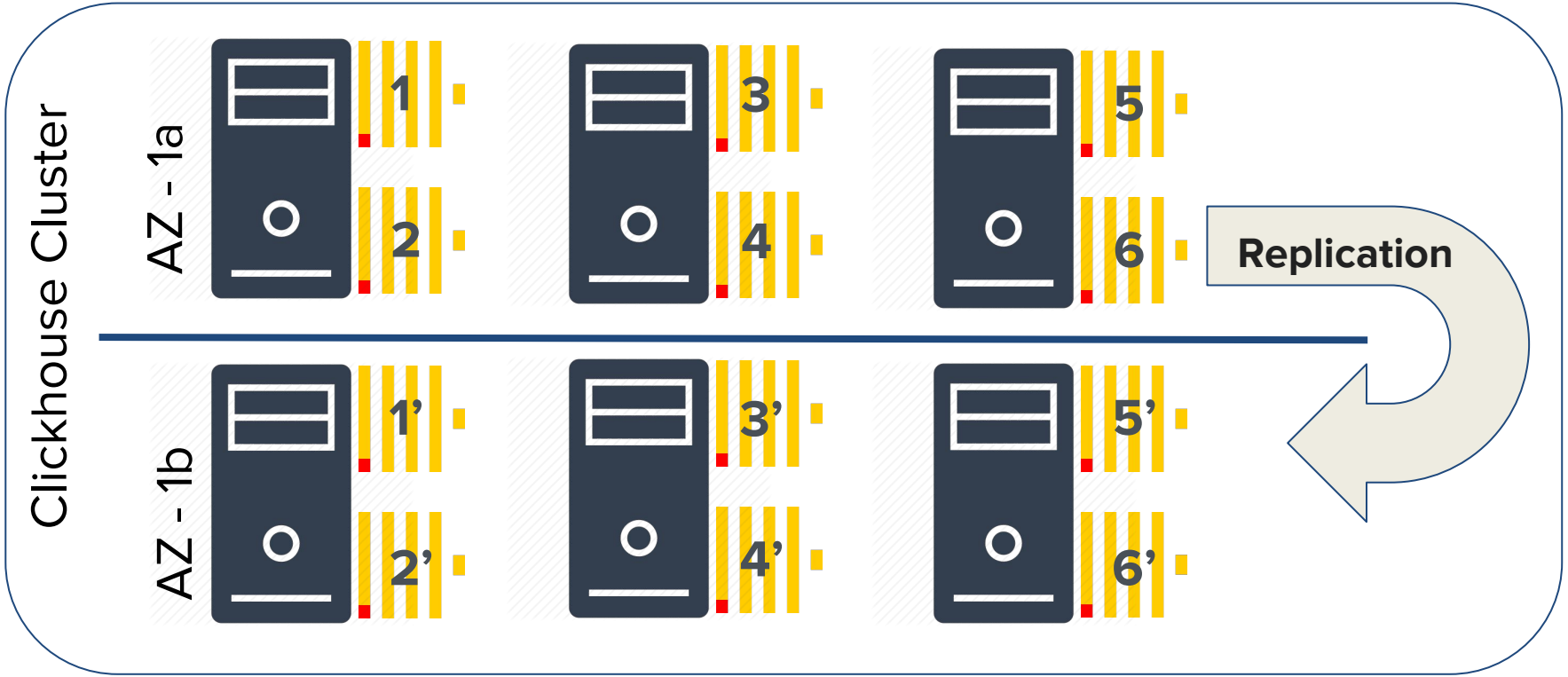


Benchmark iterations on parallelism

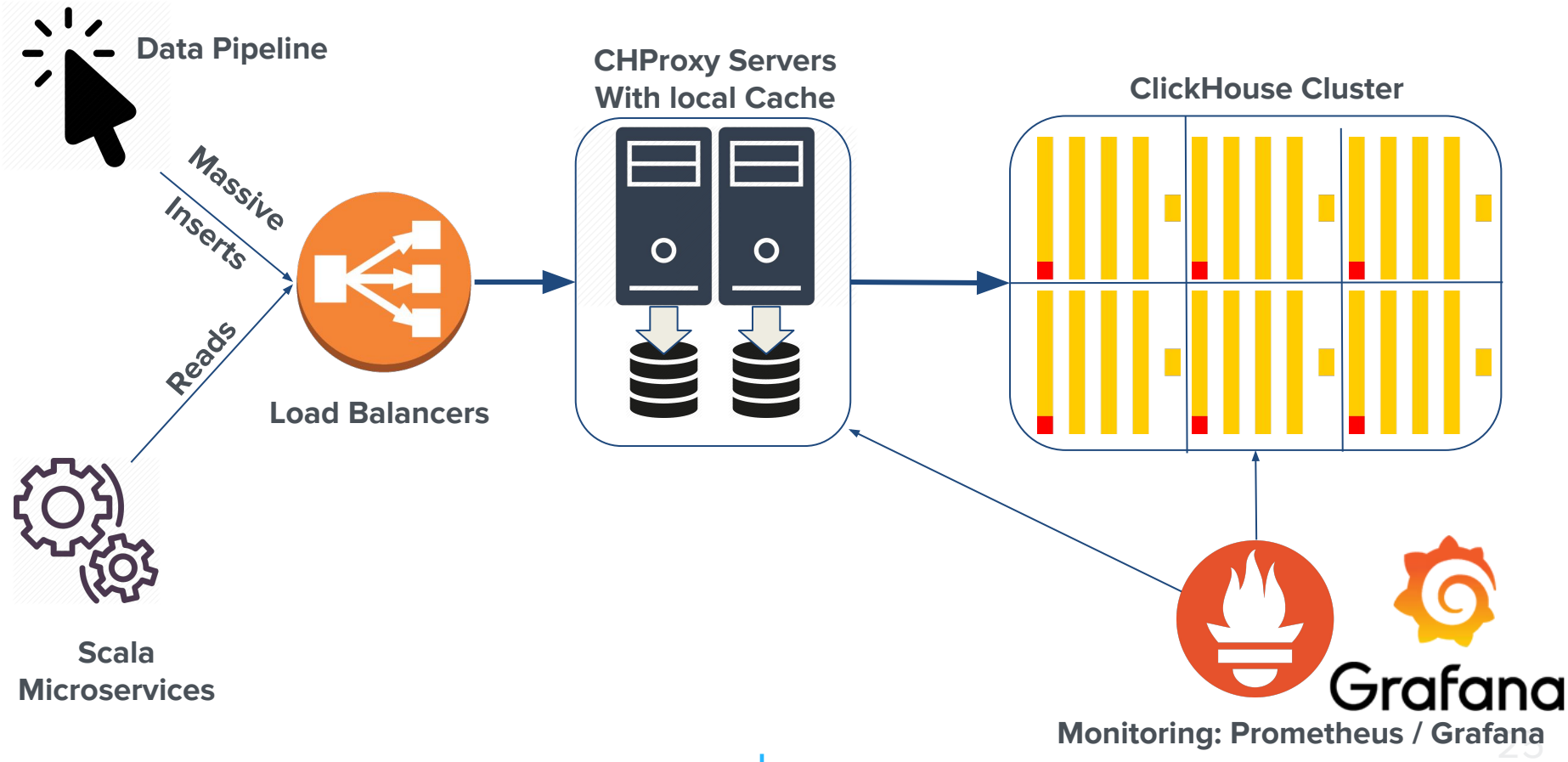


New Backend Infrastructure (Zoom in)

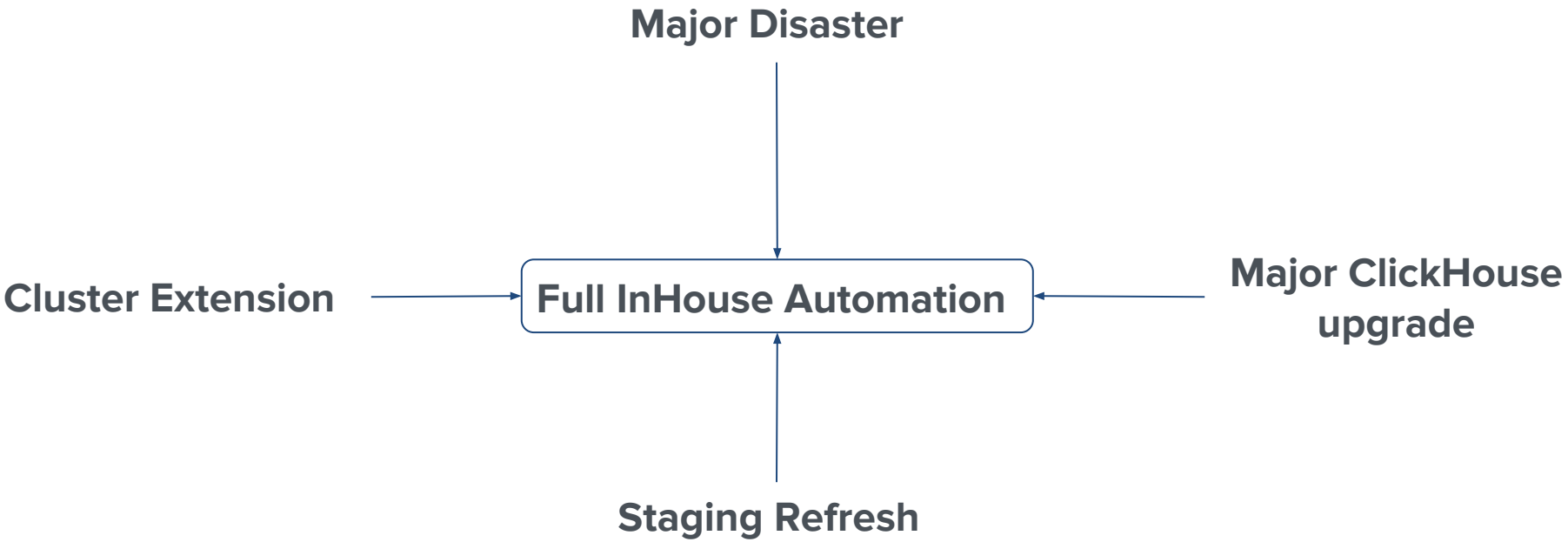
1 EC2 (i3.metal) instance hosts 2 clickhouse servers



New Backend Infrastructure (Big picture)



Next Steps


Major Disaster
Cluster Extension
Full InHouse Automation
Major ClickHouse upgrade
Staging Refresh

What do we like with ClickHouse

Native sampling

Stability

Easy to understand

Fast

Active developments

What parts of ClickHouse still need to be improved

Difficult to master

Lack of tooling

Random stability of a new version

No query optimizer

Do we recommend ClickHouse?

CONTENTSQUARE
Experience Matters

Thank you

Any question?

We’re hiring!