Using monitoring to understand Cassandra

THE LAST PICKLE

About TLP and Myself

Hi, I'm Alain.

@arodream
https://www.linkedin.com/in/arodrime

Cassandra operator since 2011 (v 0.8)
Datastax MVP for Apache Cassandra
Montpellier, France

Consultant - The Last Pickle


I am here :-)

About you!

Poll time!

About Cassandra

Cassandra - History

2008

Cassandra - History

2008

2009
-
2010

Cassandra - Today!

310 systems in ranking, November 2016

Rank Nov 2016	Rank Oct 2016	Rank Nov 2015	DBMS	Database Model	Score Nov 2016	Score Oct 2016	Score Nov 2015
1.	1.	1.	Oracle	Relational DBMS	1413.01	-4.09	-67.94
2.	2.	2.	MySQL	Relational DBMS	1373.56	+10.91	+86.71
3.	3.	3.	Microsoft SQL Server	Relational DBMS	1213.80	-0.38	+91.48
4.	↑ 5.	↑ 5.	PostgreSQL	Relational DBMS	325.82	+7.12	+40.13
5.	↓ 4.	↓ 4.	MongoDB	Document store	325.48	+6.67	+20.87
6.	6.	6.	DB2	Relational DBMS	181.46	+0.90	-21.07
7.	7.	↑ 8.	Cassandra	Wide column store	133.97	-1.09	+1.05
8.	8.	↓ 7.	Microsoft Access	Relational DBMS	125.97	+1.30	-14.99
9.	9.	↑ 10.	Redis	Key-value store	115.54	+6.00	+13.13
10.	10.	↓ 9.	SQLite	Relational DBMS	112.00	+3.43	+8.55

Sharing knowledge slow + Growing fast —> **Skills gap**

Source: http://db-engines.com/en/ranking

Cassandra Internals - Main characteristics

A scalable, Highly Available, and eventually consistent Database!

Scalable?	HA?	Eventual consistency?
Shard!	Redundancy!	Anti-entropy systems!
Token range / vnodes Consistent hashing Linear scalability	Replication Factor (RF) No SPOF (Masterless)	Hinted handoff Read Repair Anti-entropy Repair Consistency Level (CL)

Cassandra Internals - Tunable Consistency

- ANY (Writes - uses HH)
- ONE (Two, Three)
- LOCAL_ONE
- QUORUM → floor(RF/2) + 1
- **LOCAL_QUORUM**
- EACH_QUORUM
- ALL

+ Strong Availability

- Poor Consistency

+ Strong Consistency

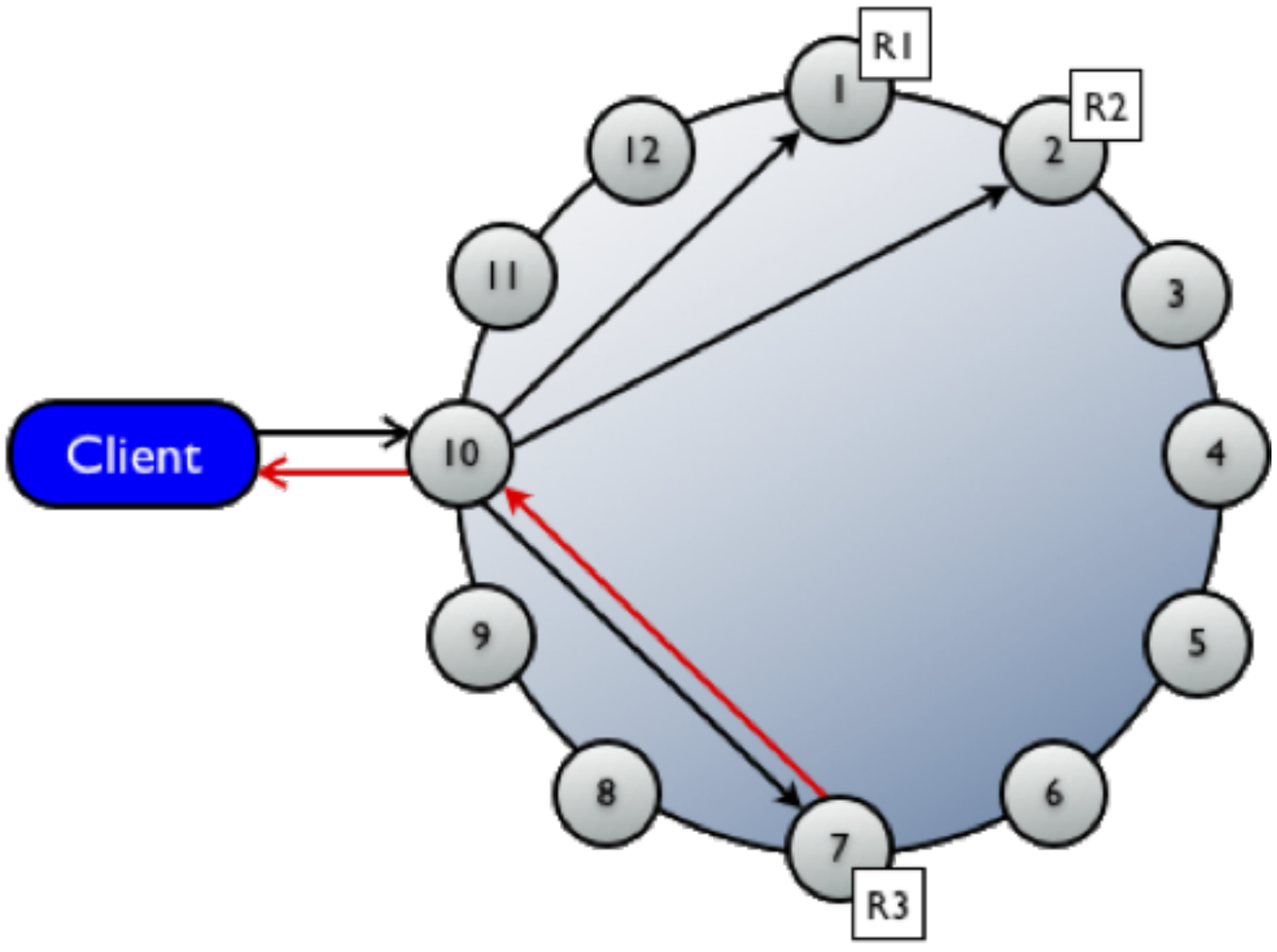
- Poor Availability

Cassandra Internals - Writes

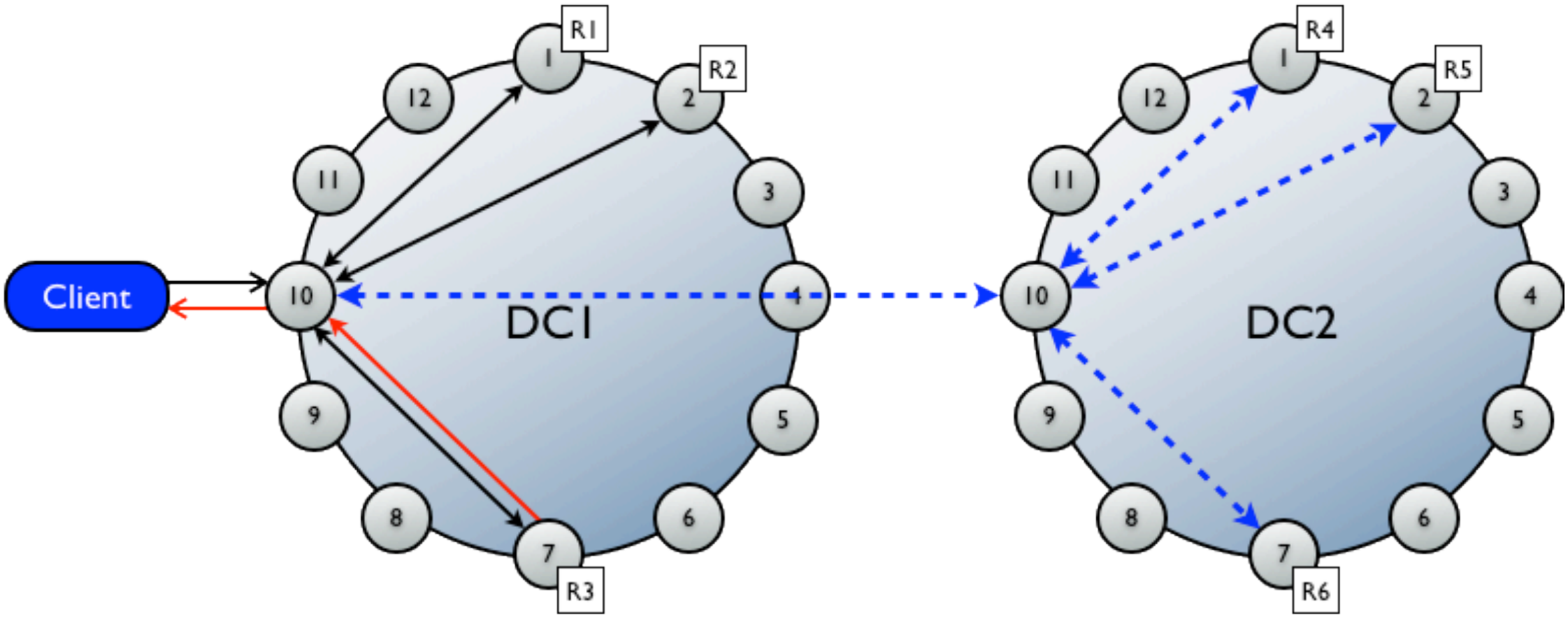
Write with

CL: ONE

RF = 3



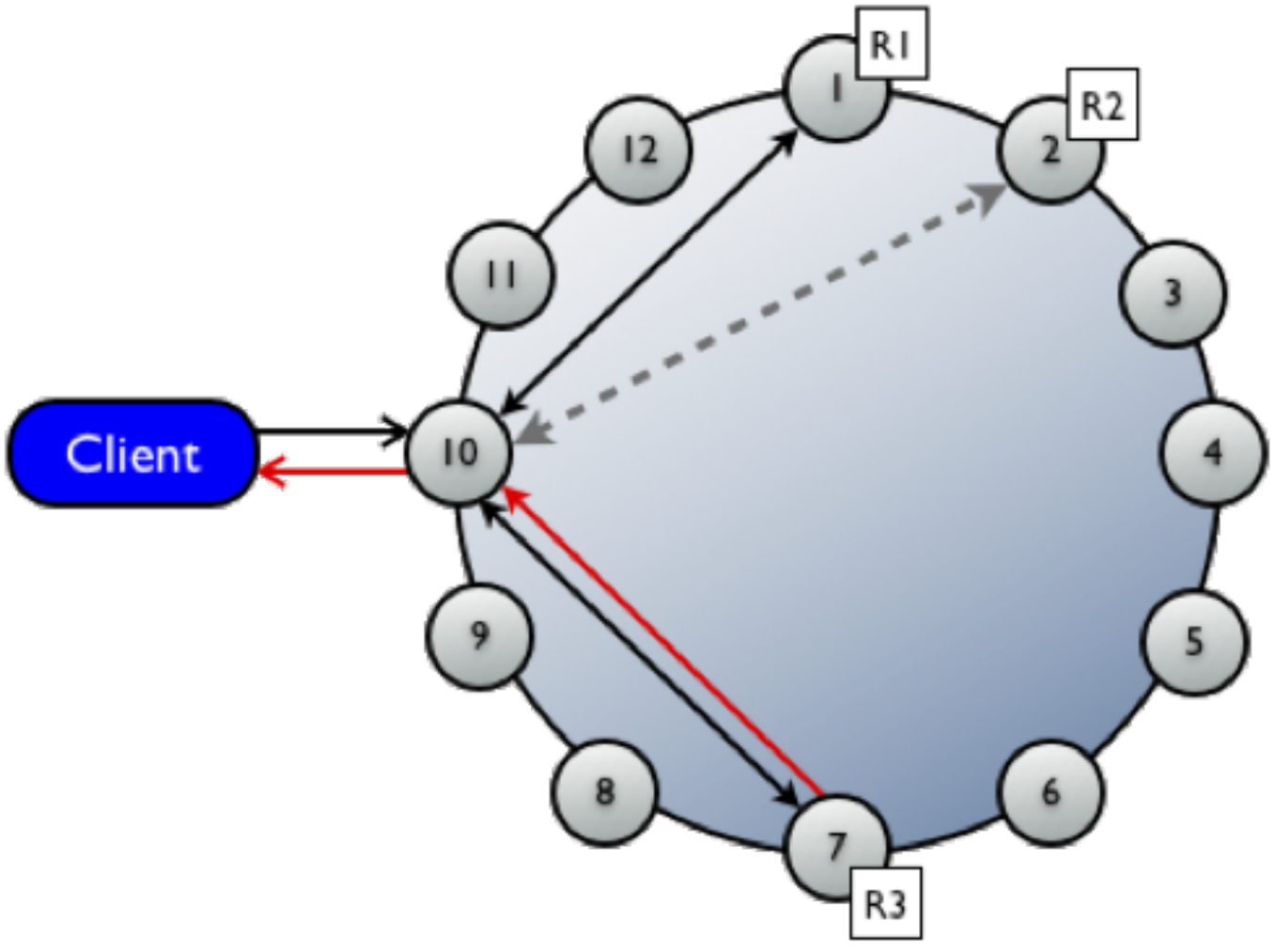
Cassandra Internals - R&W Operations



Multiple Data Center - Write requests

RF {DC1: 3, DC2: 3} and CL: ONE

Cassandra Internals - R&W Operations



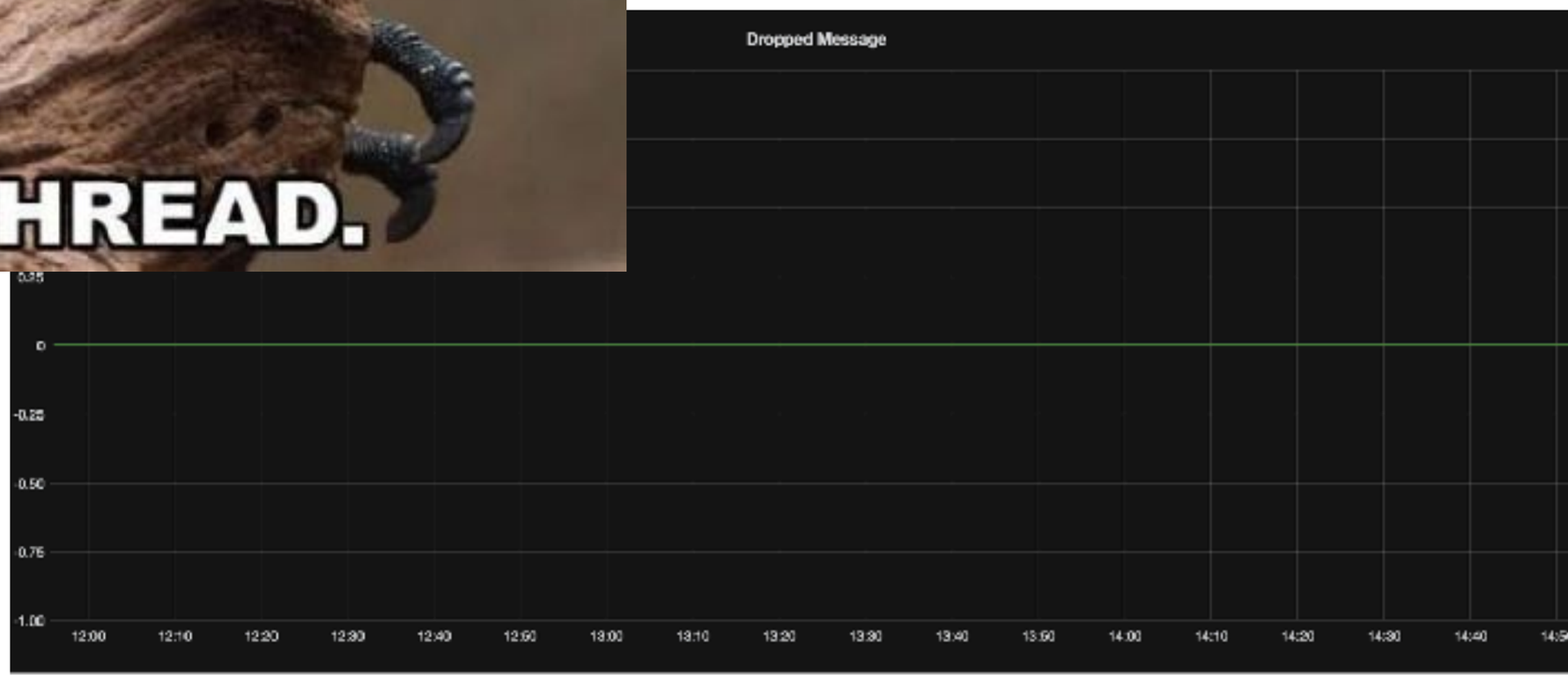
Read with

CL: Quorum

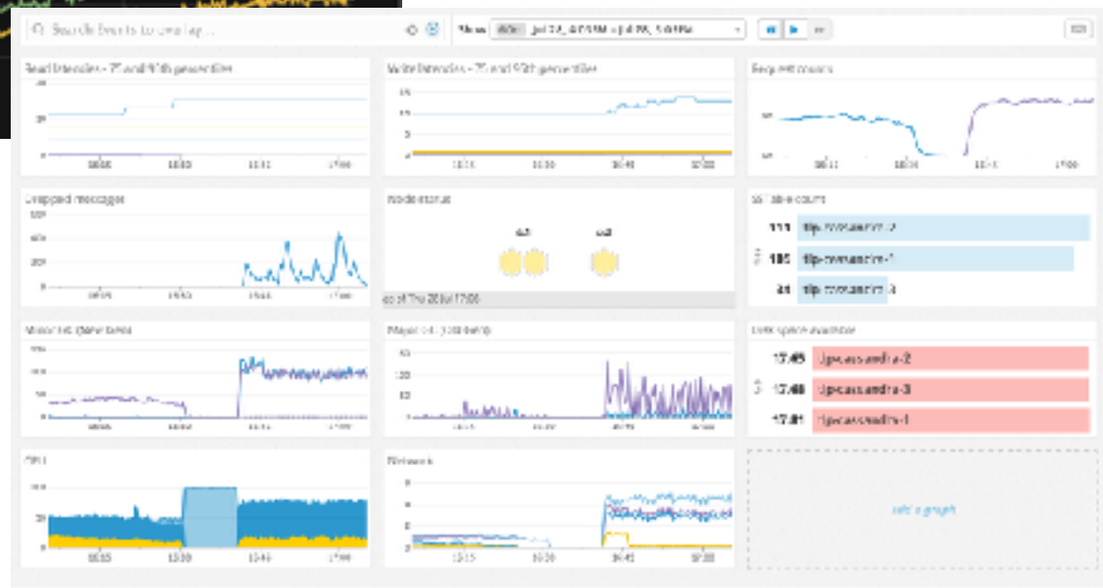
RF = 3

About Monitoring

Monitoring is **important**!



I meant: **Efficient** monitoring is **important**!

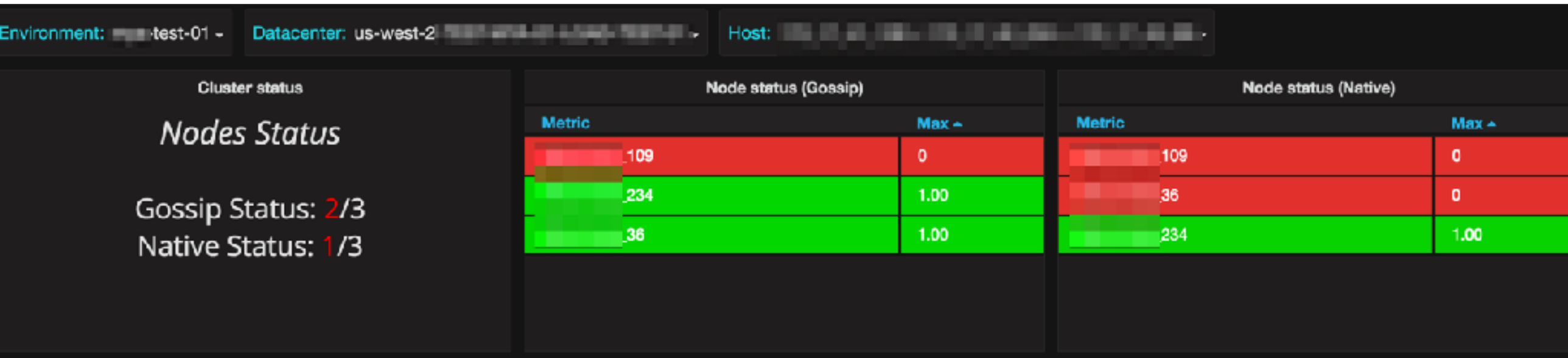


I have been advocating this way

Cassandra Summit 2016: https://www.youtube.com/watch?v=Q9AAR4UQzMk

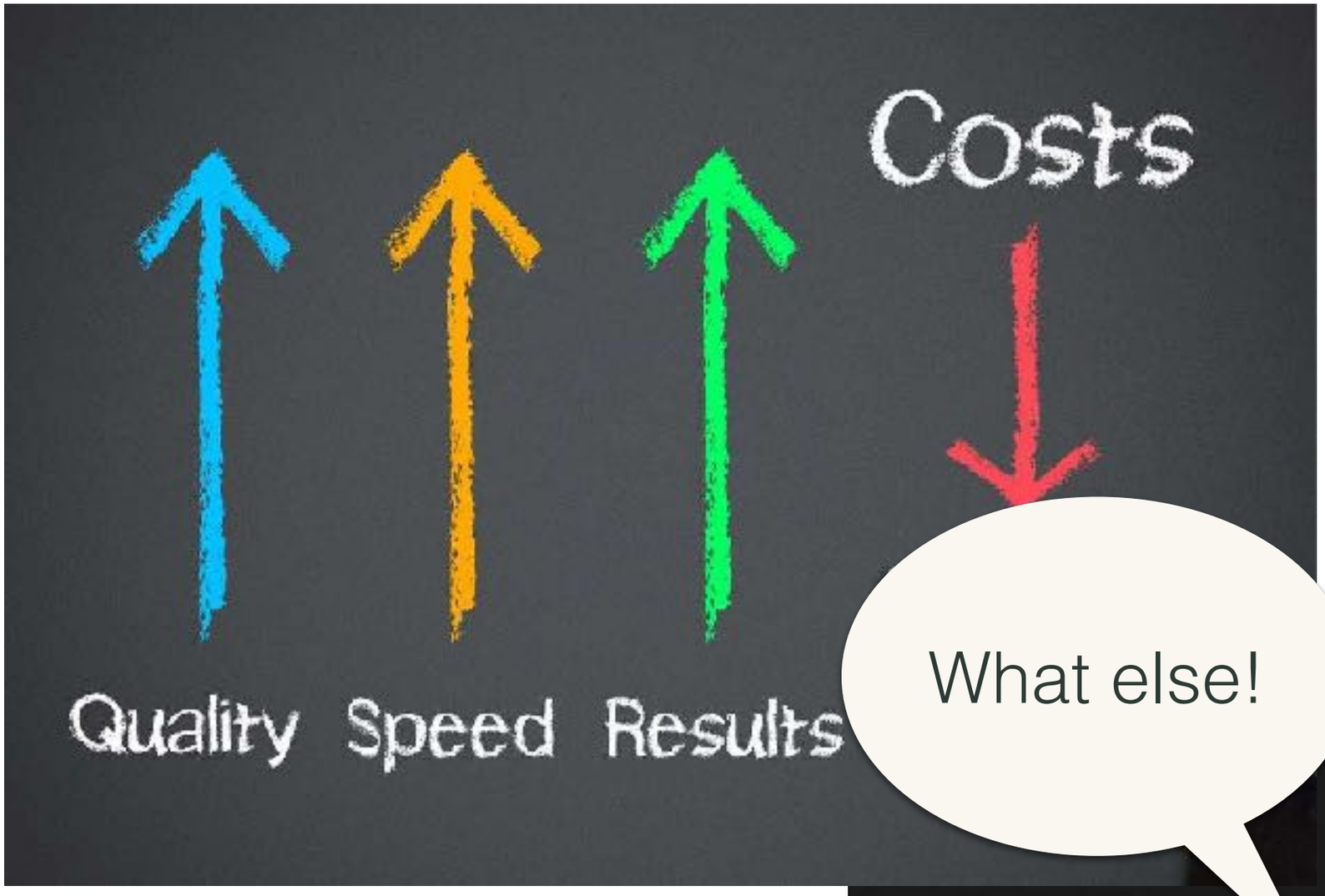
We want to prevent this!

Or at least be aware that it is happening!



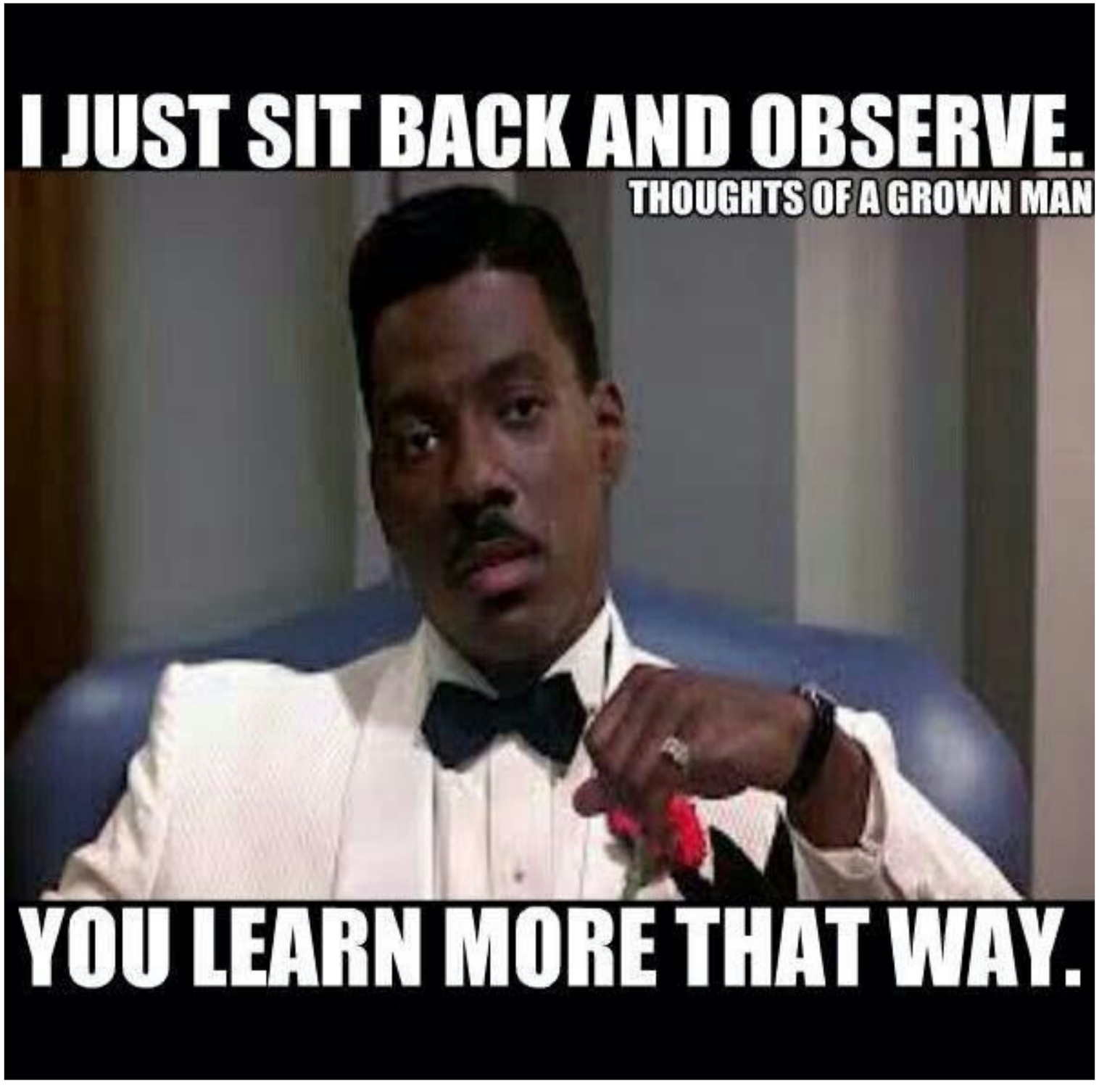
An accurate feel for the system

Troubleshooting and optimization



What else!

Monitoring to learn about internals!



Monitor to learn about internals!

The idea?

We can use monitoring to learn

But we need efficient monitoring tools!

The problem?

How to have a relevant set of dashboards?

How could a new user know what dashboards to build?

A strong set of dashboards

- Is long, painful and expensive to build
- Requires an expert knowledge to be built

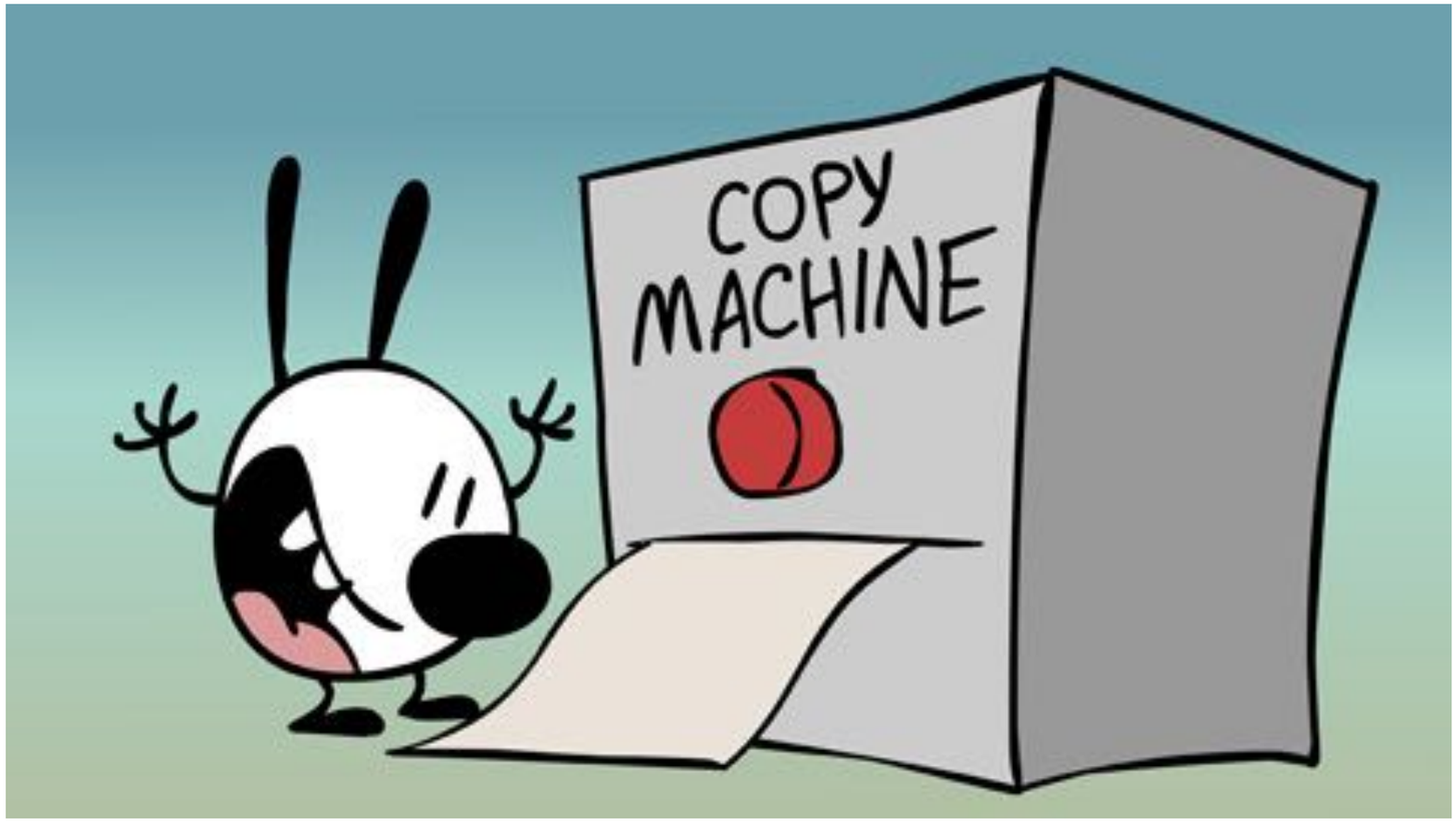
An utopic world ?

So is this idea a fantasy?

Yet another unrealistic recommendation?

A strong set of dashboards

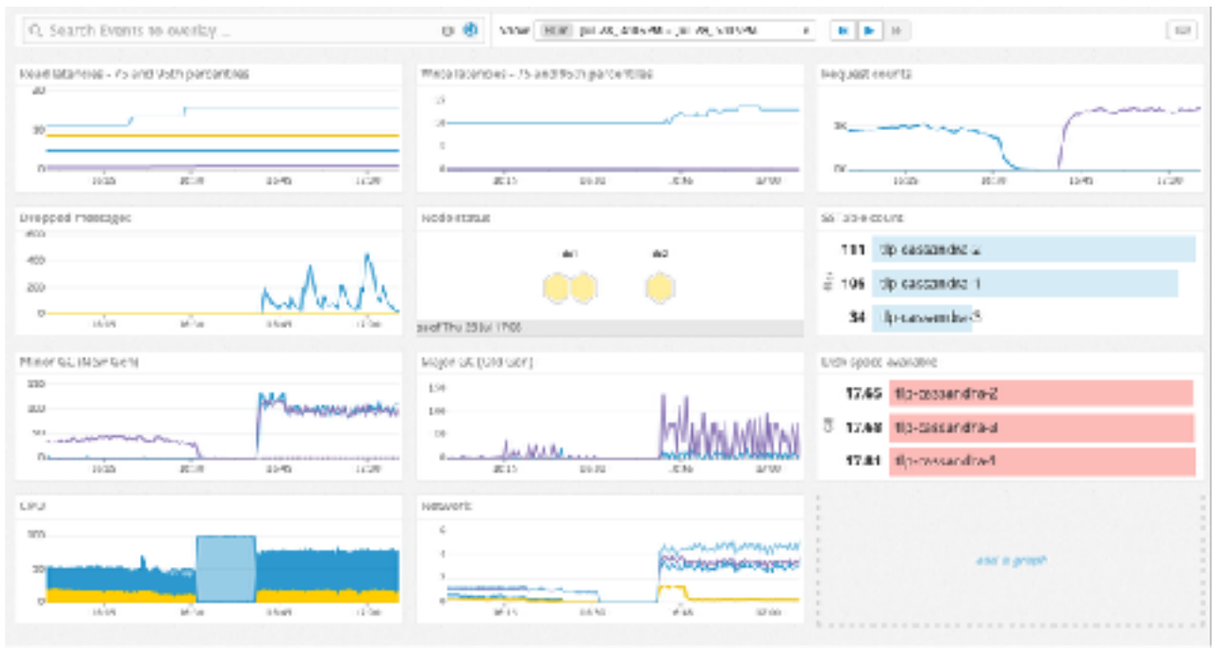
+ Is very useful to any operator from day 1
+ A template is easy to use and immediately available



+ A strong set of dashboards is actually expert knowledge easily sharable!

Out of the box dashboards?

An ideal world as an operator would probably look like this



Building dashboards: Templates idea origin

(From my talk at Cassandra Summit 2016)

- TLP = Consulting company = many Cassandra clusters

Building dashboards: Templates idea origin

(From my talk at Cassandra Summit 2016)

- TLP = Consulting company = many Cassandra clusters
- Need to build dashboards for many clients

Building dashboards: Templates idea origin

(From my talk at Cassandra Summit 2016)

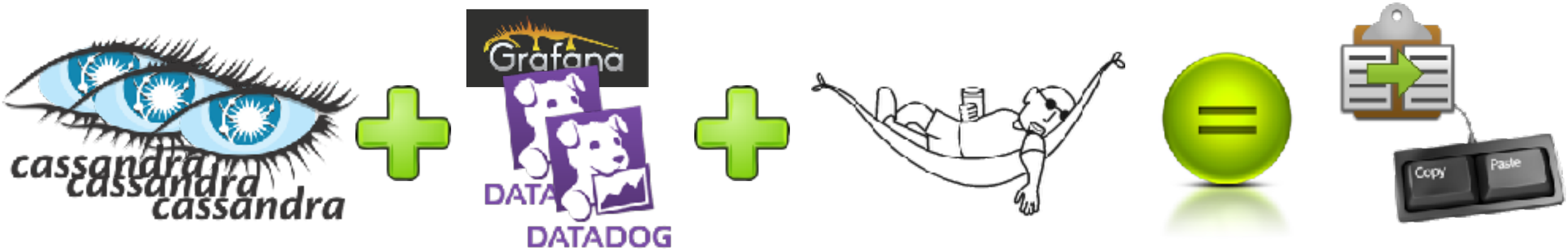
- TLP = Consulting company = many Cassandra clusters
- Need to build dashboards for many clients
- Laziness, I do not like repeating tasks

Building dashboards: Templates idea origin

(From my talk at Cassandra Summit 2016)

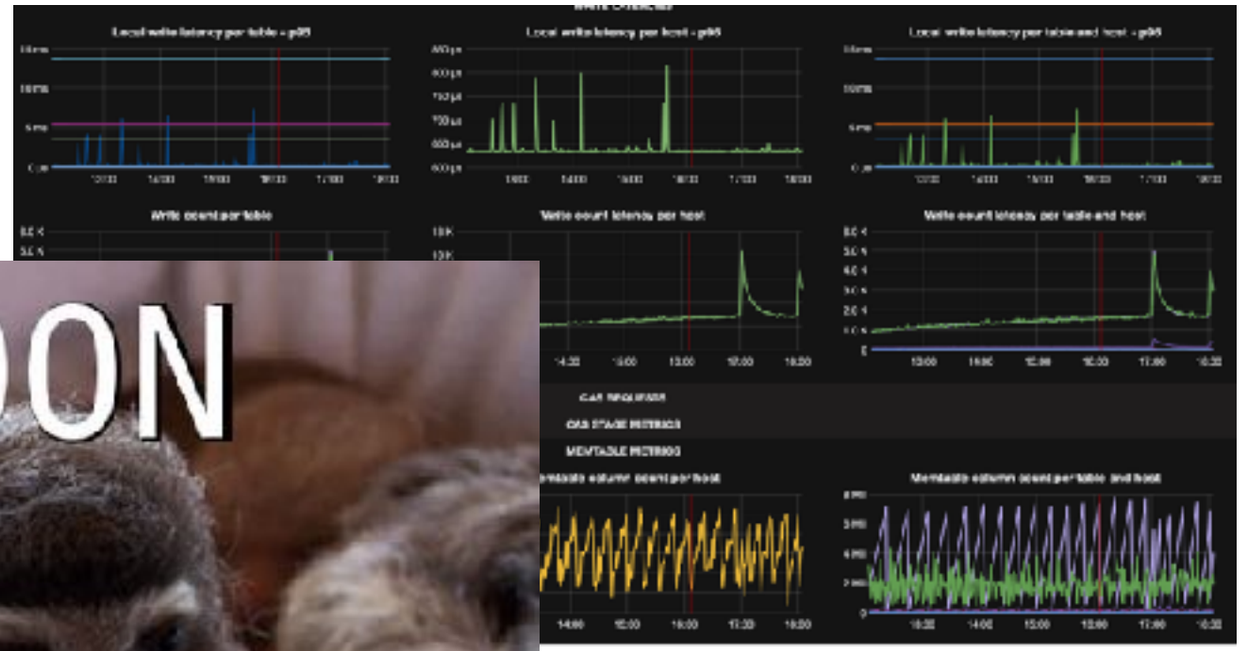
- TLP = Consulting company = many Cassandra clusters
- Need to build dashboards for many clients
- Laziness, I do not like repeating tasks

We decided to build templates and share with the community



Out of the box TLP-dashboards!

Ongoing work to be shared with the C* community “soon”

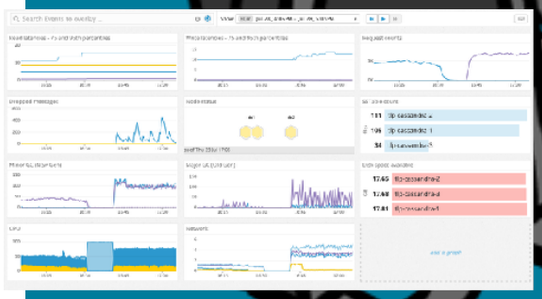


more awesome pictures at THEMETAPICTURE.COM

Back to the idea…

We can use monitoring to learn

With the right tools!



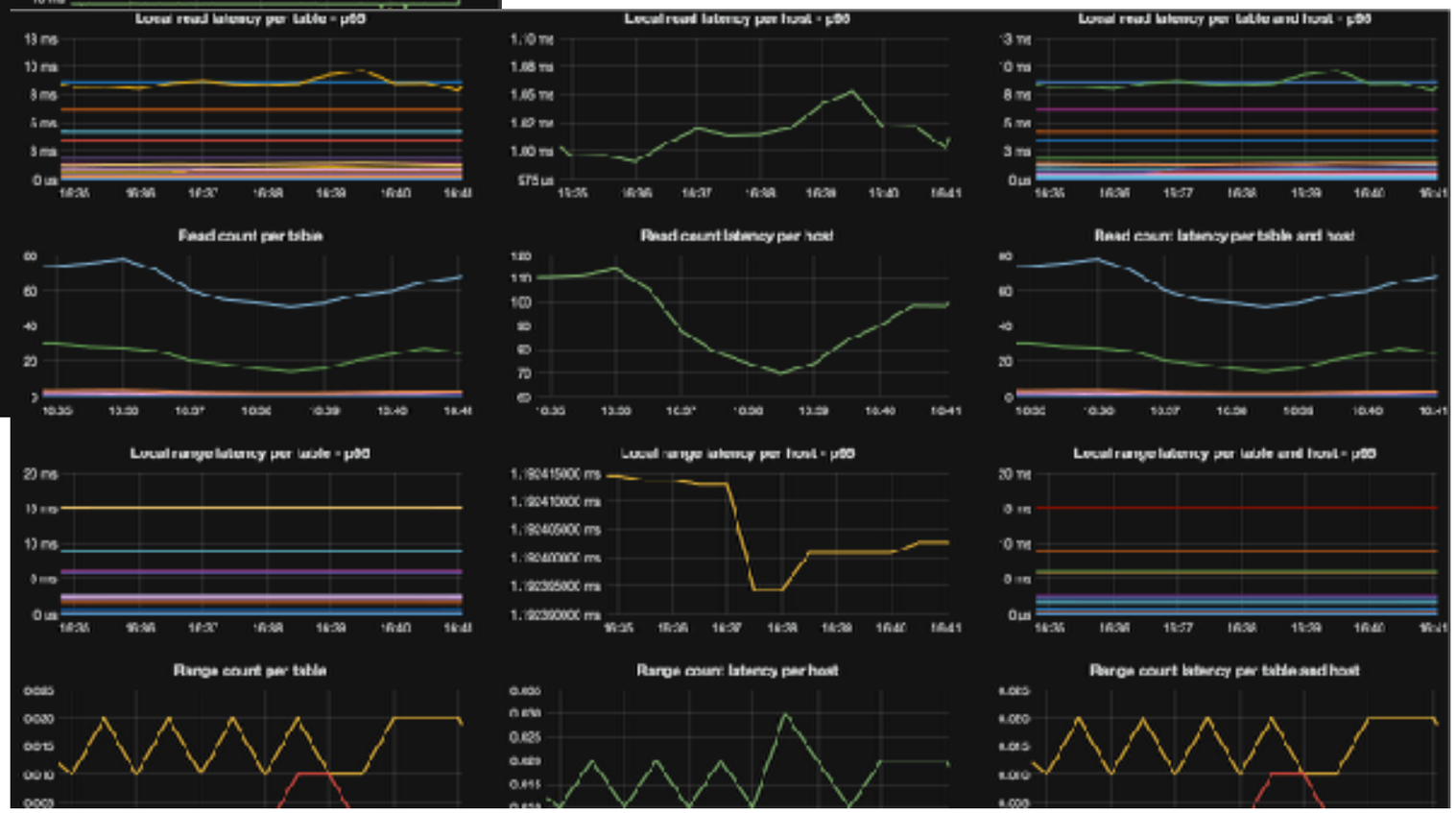
The right tools

1 - Overview Dashboards - Detect

The right tools

1 - Overview Dashboards - Detect

2 - Themed Dashboards - Troubleshoot

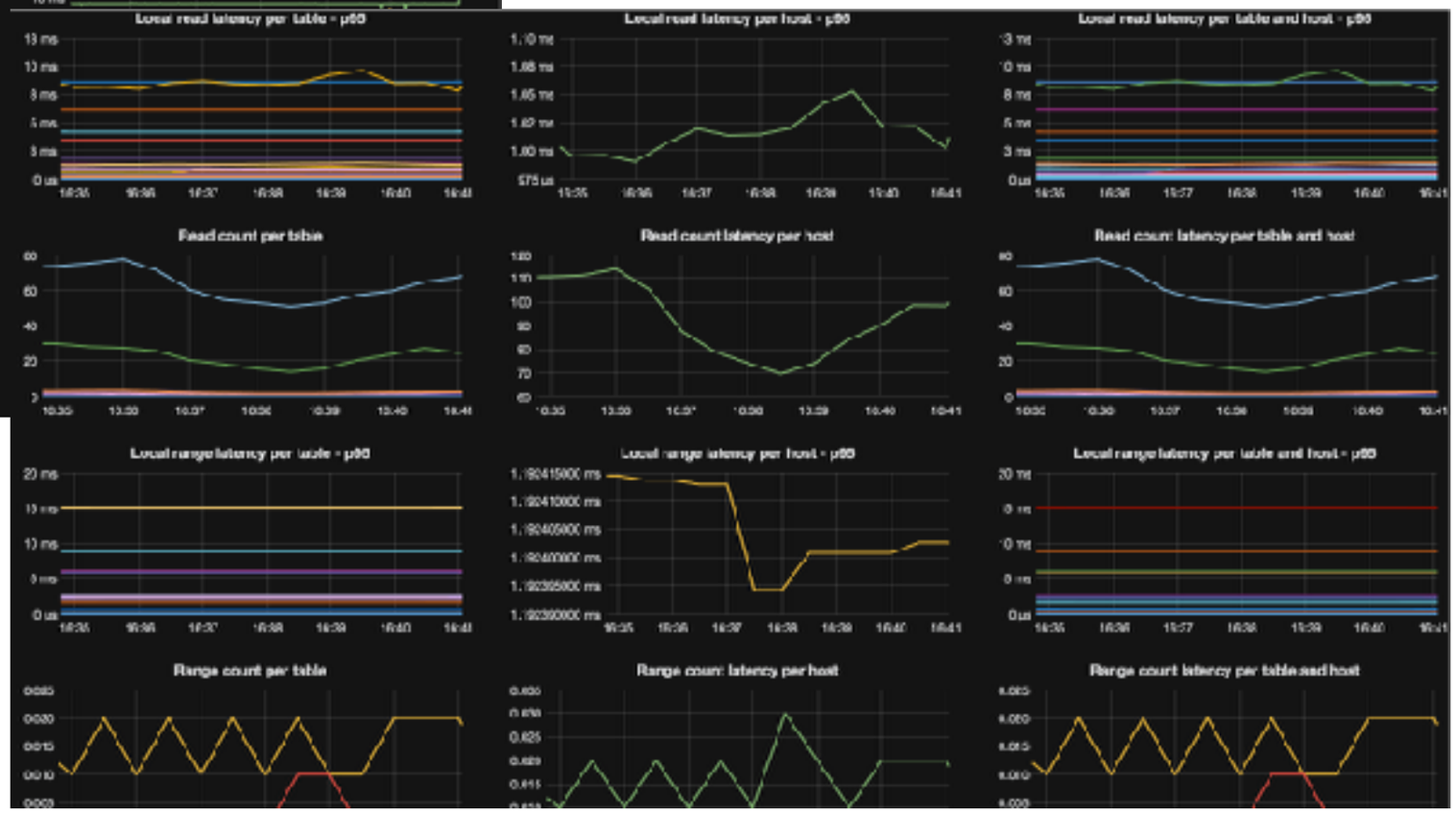


The right tools

1 - Overview Dashboards - Detect

2 - Themed Dashboards - Troubleshoot

3 - Themed Dashboards - Optimize



Applying to Cassandra

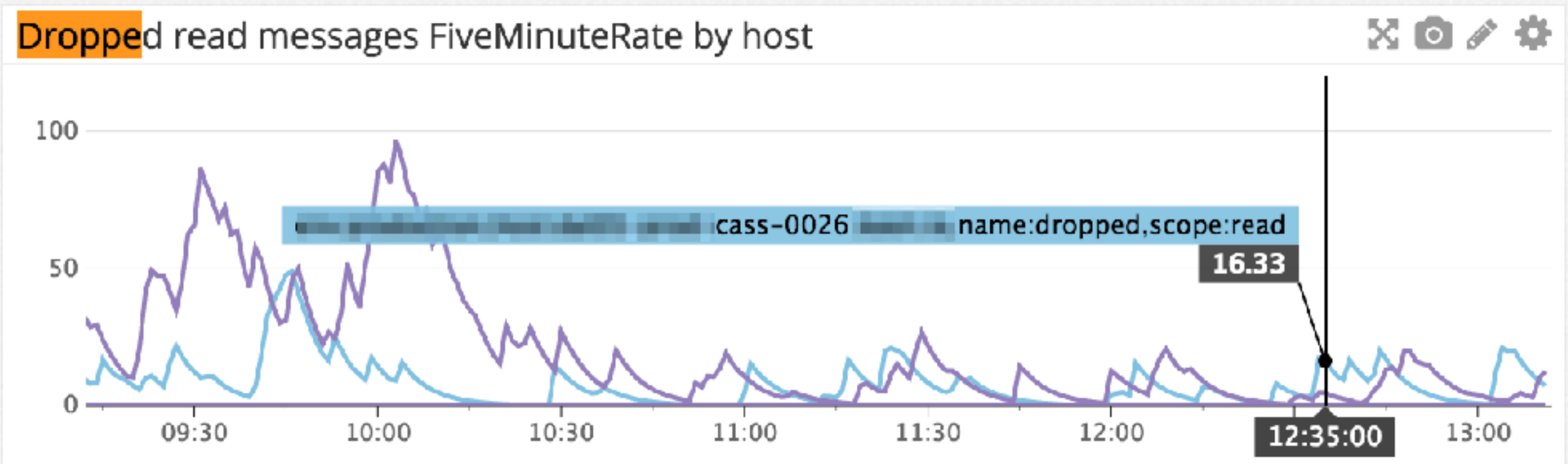
How to learn Cassandra relying on monitoring?

Let’s try to understand a couple production issues together

Example 1: Read messages dropped

A dropped read issue - Monitoring

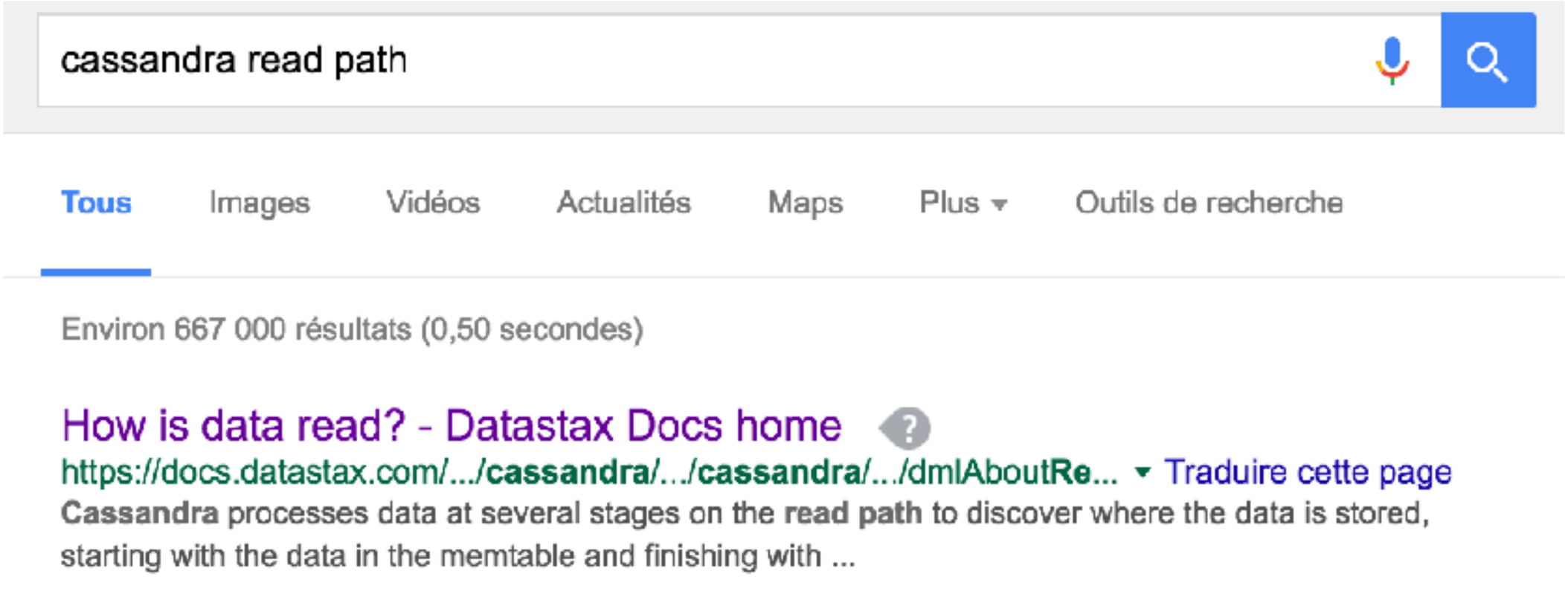
From overview dashboard



Read path issue

A dropped read issue - Internals

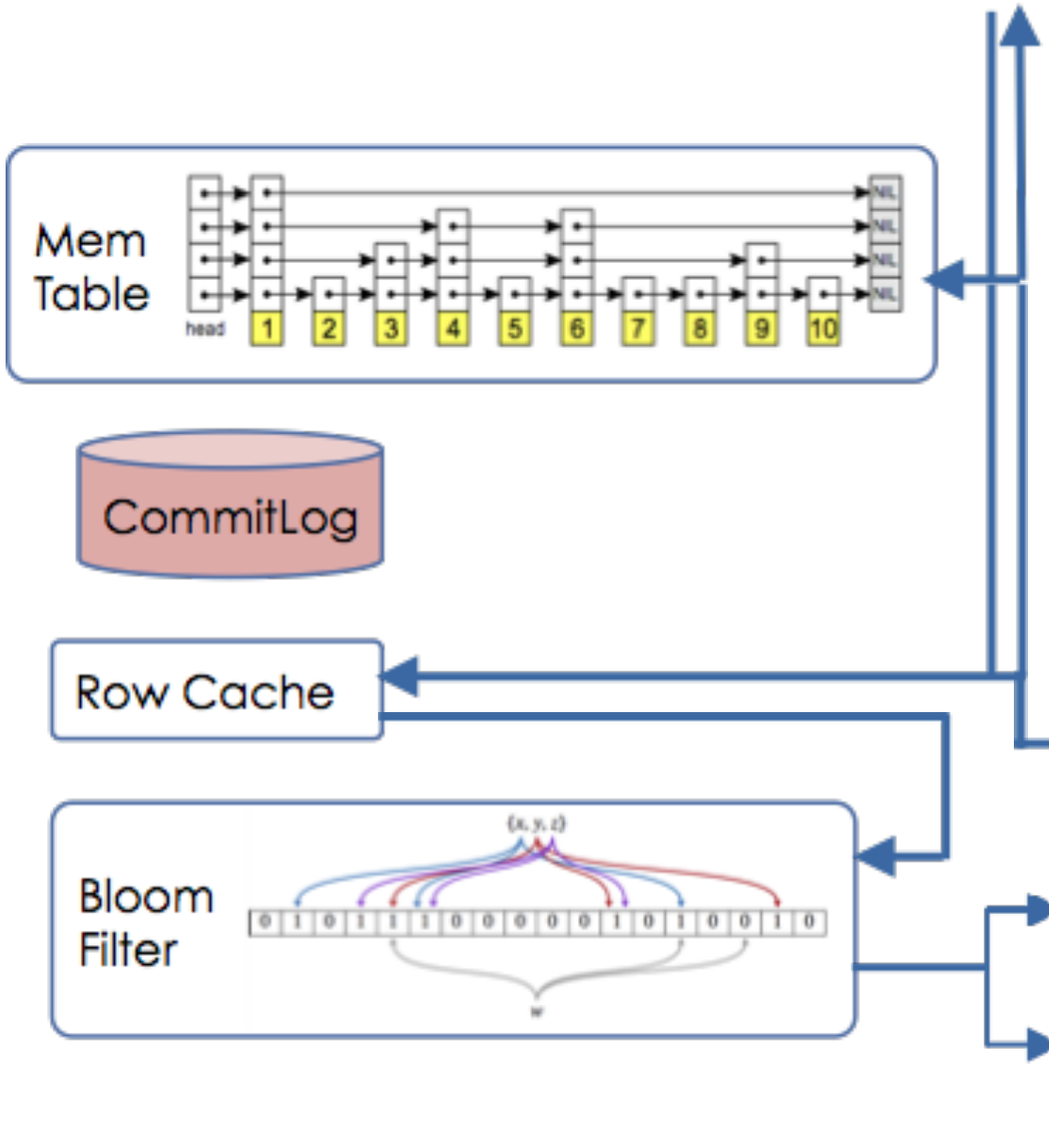
How does Cassandra read data?



Let’s read and learn about it

A dropped read issue - Internals

How does Cassandra read data?



Get values from Memtable

Get values from Row Cache if present

Check BF - find possible SSTables

Check Key Cache for fast search in SSTables

Get values from SSTables

Repopulate the Row Cache?

Read issue —> Read Path Dashboard!

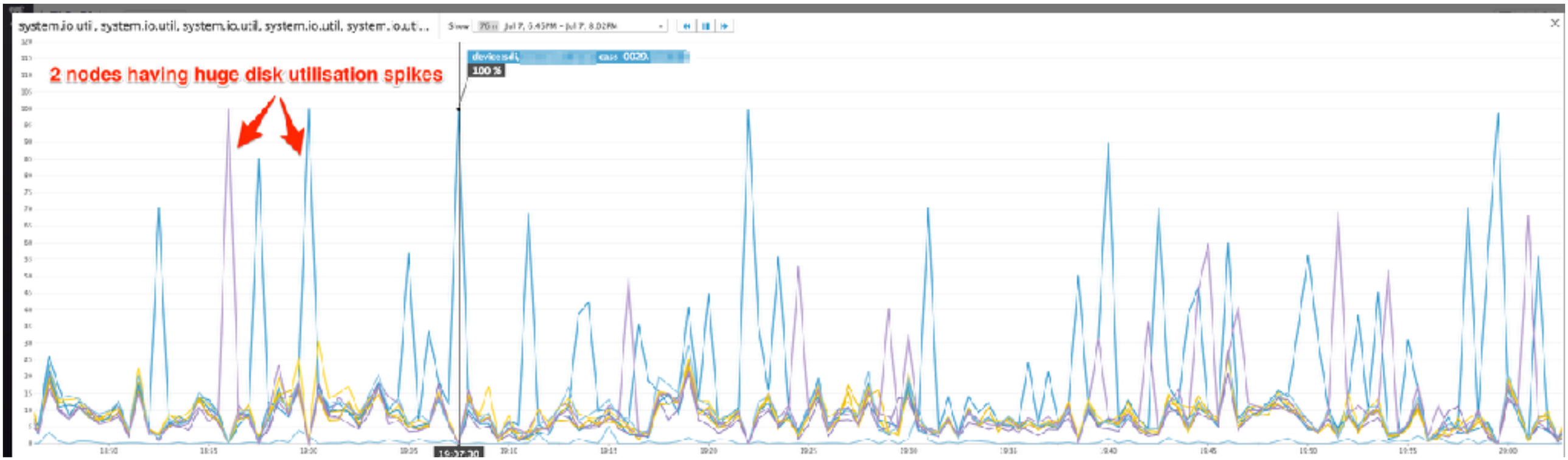
A dropped read issue - Internals

So what can be causing this issue?

- Under-sized cluster
- Inefficient bloomfilters
- Inefficient key caching
- Tombstones read
- Too many SSTables
- Wide rows
- Poorly designed queries
- JVM / GC issues
- Hardware issue
- …

A dropped read issue - Monitoring

From read path dashboard



Spikes only appear on disks charts

-

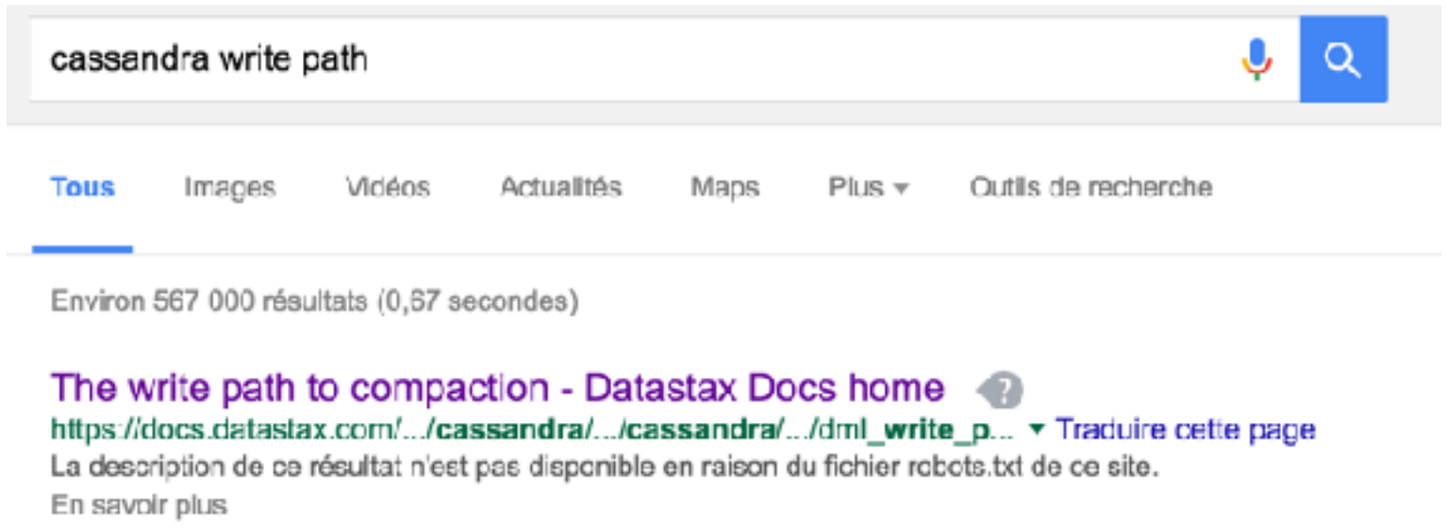
On 2 nodes

A dropped read issue - Thinking

Looks like a hardware issue…

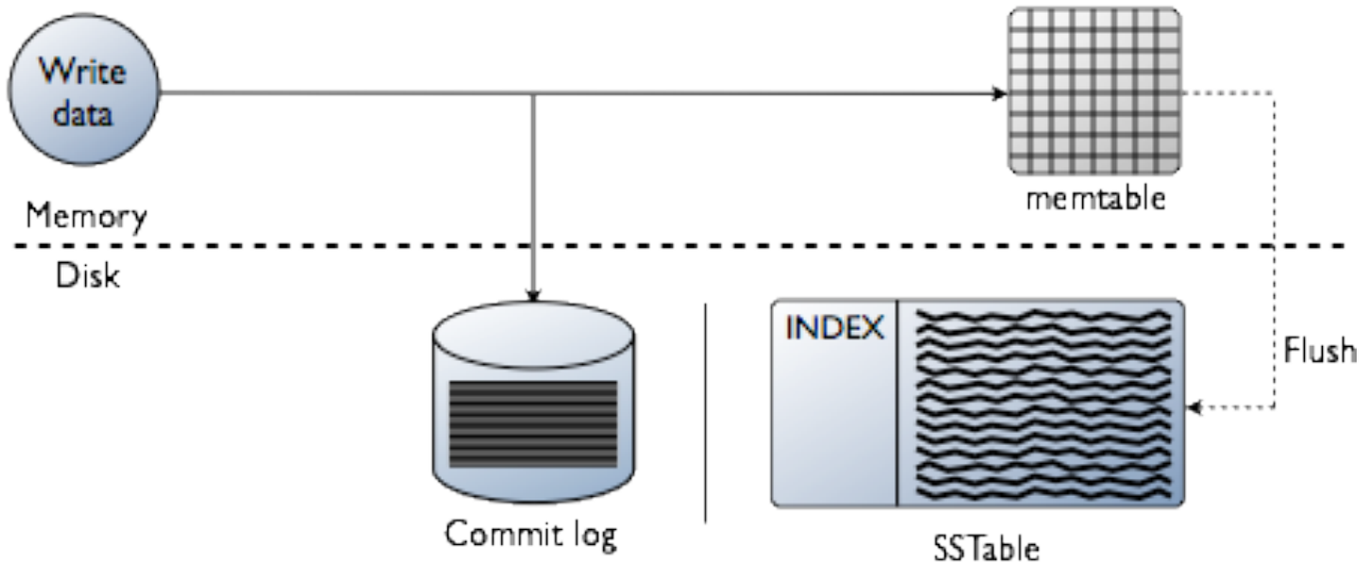
… but then, why aren’t writes dropped as well?

A dropped read issue - Internals



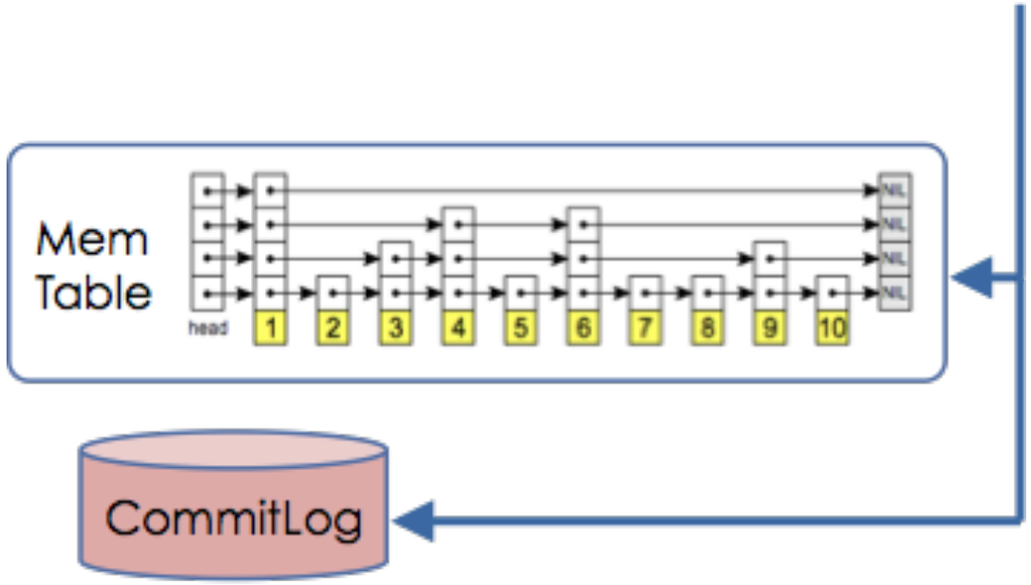
Searching to understand internals

Cassandra write path



A dropped read issue - Internals

Cassandra write path



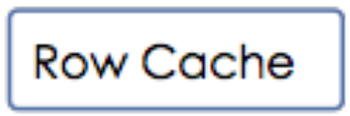
Insert into Memtable

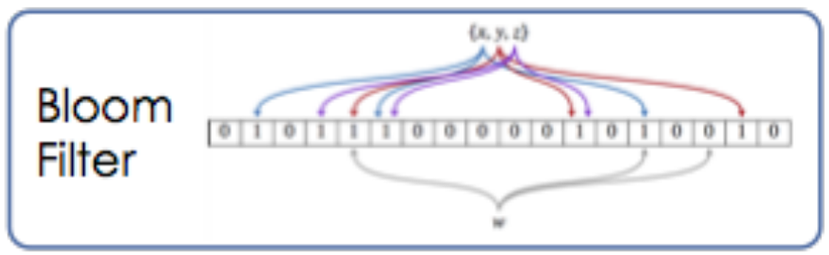
Append to commitlog

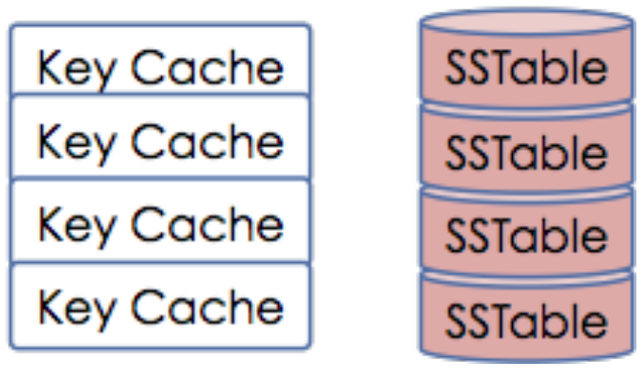
→ No read, no data disk hit

→ Very fast

→ Bottleneck CPU before I/O !







Written data is **not** hitting the data disk in Cassandra!

A dropped read issue - takeaway

Changing Machines / disks solved the issue!

We learned about read and write path at the node level

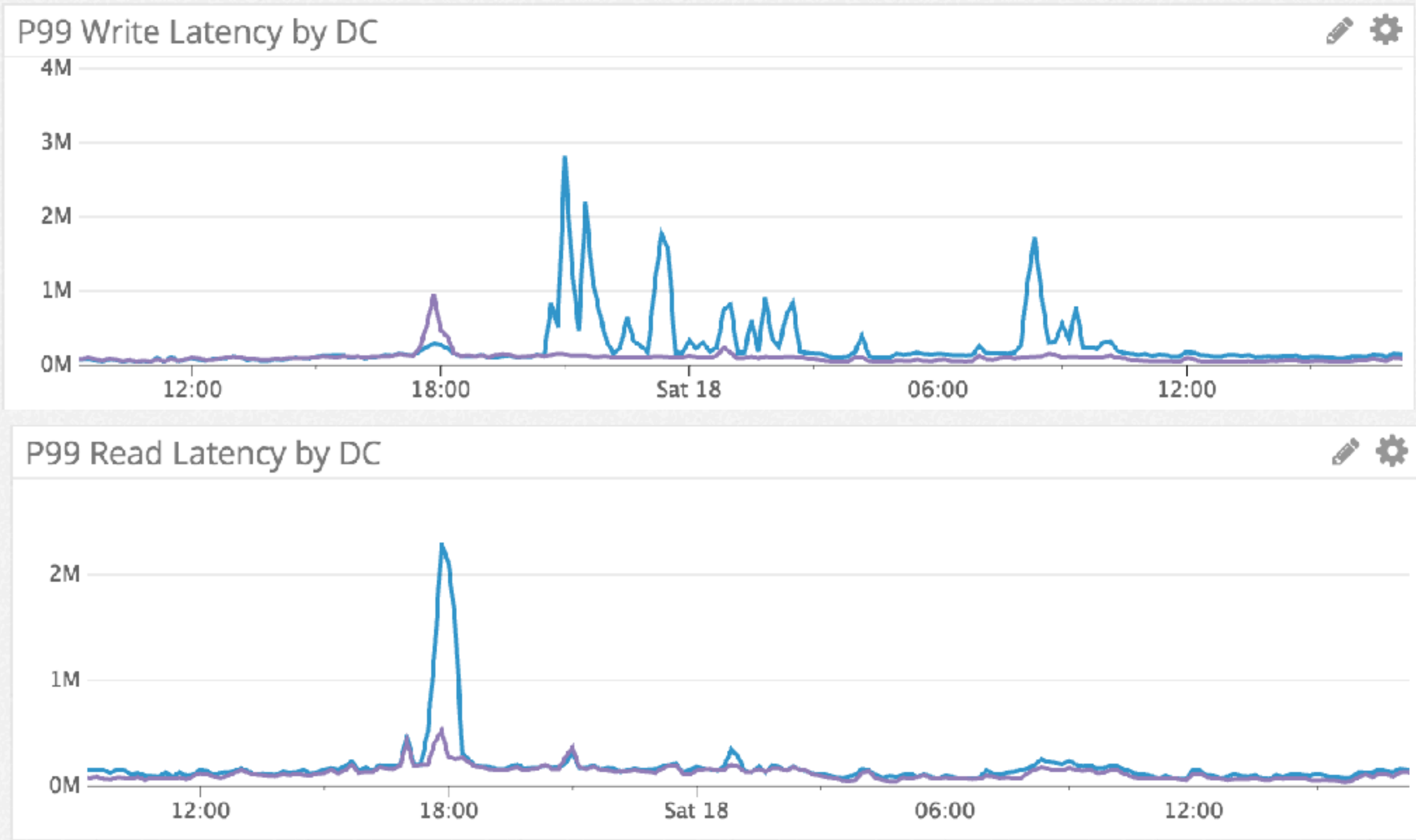
How to learn Cassandra relying on monitoring?

Let’s try to understand a few production issues together

Example 2: Latency issue

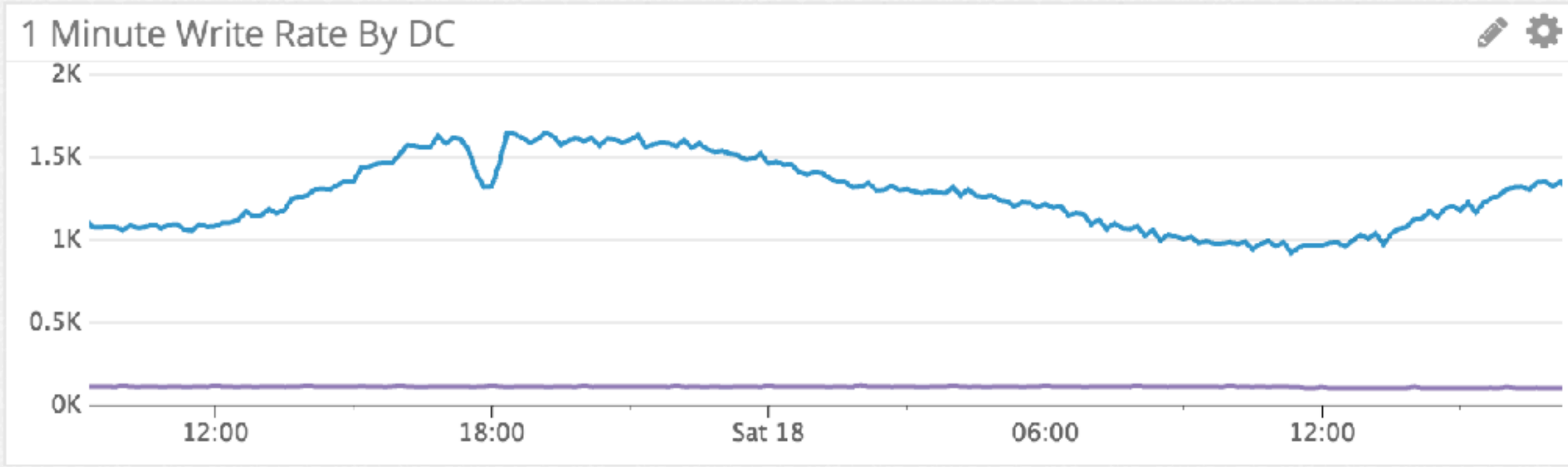
A latency issue - Monitoring

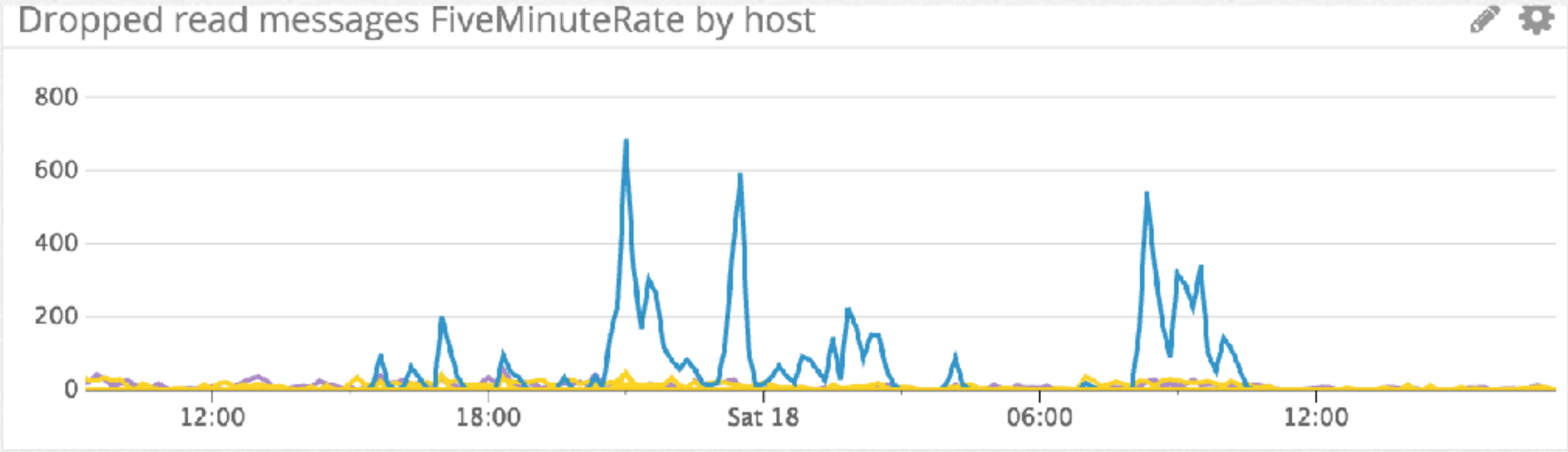
From overview dashboard



A latency issue - Monitoring

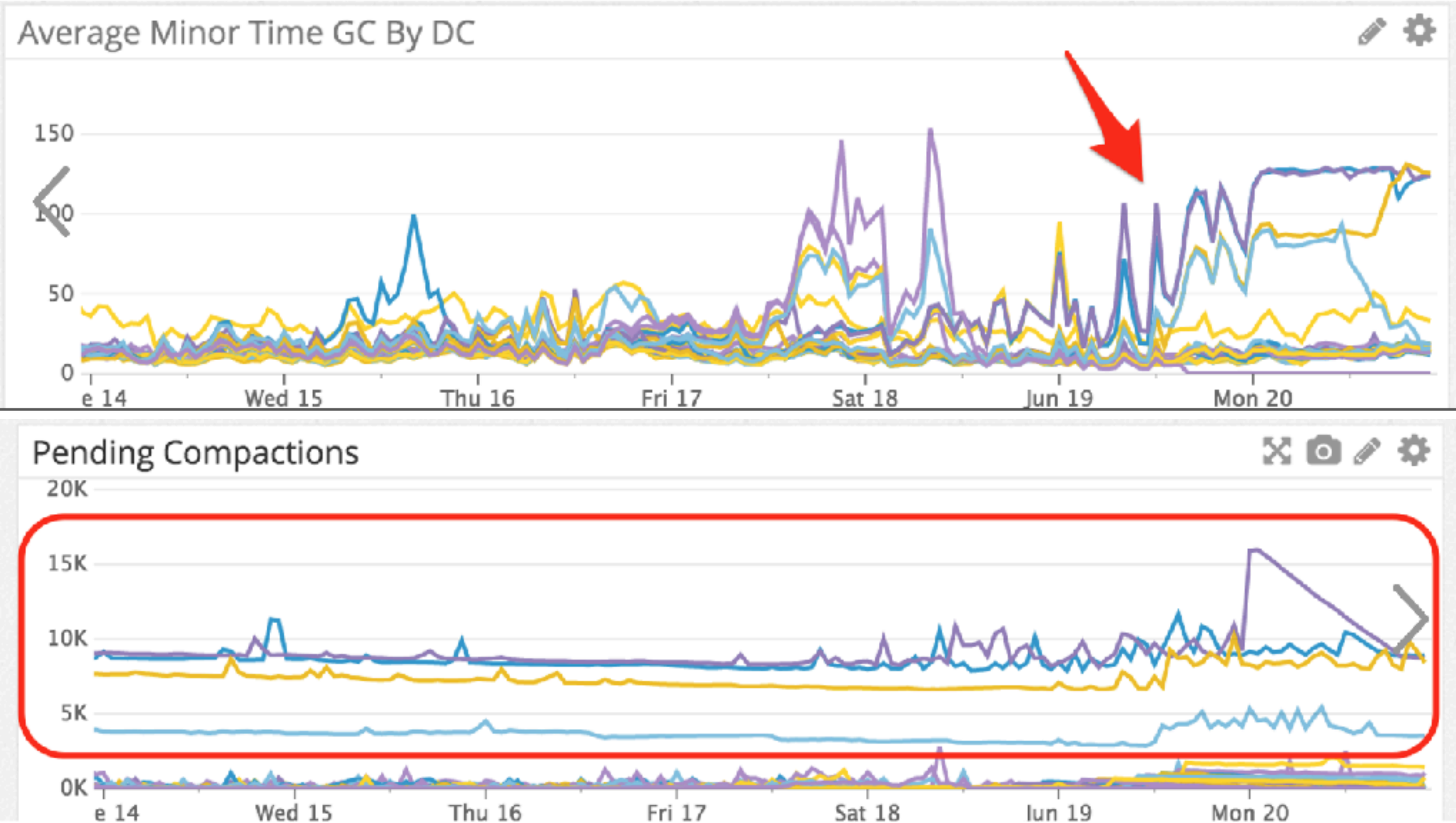
From overview dashboard





A latency issue - Troubleshooting

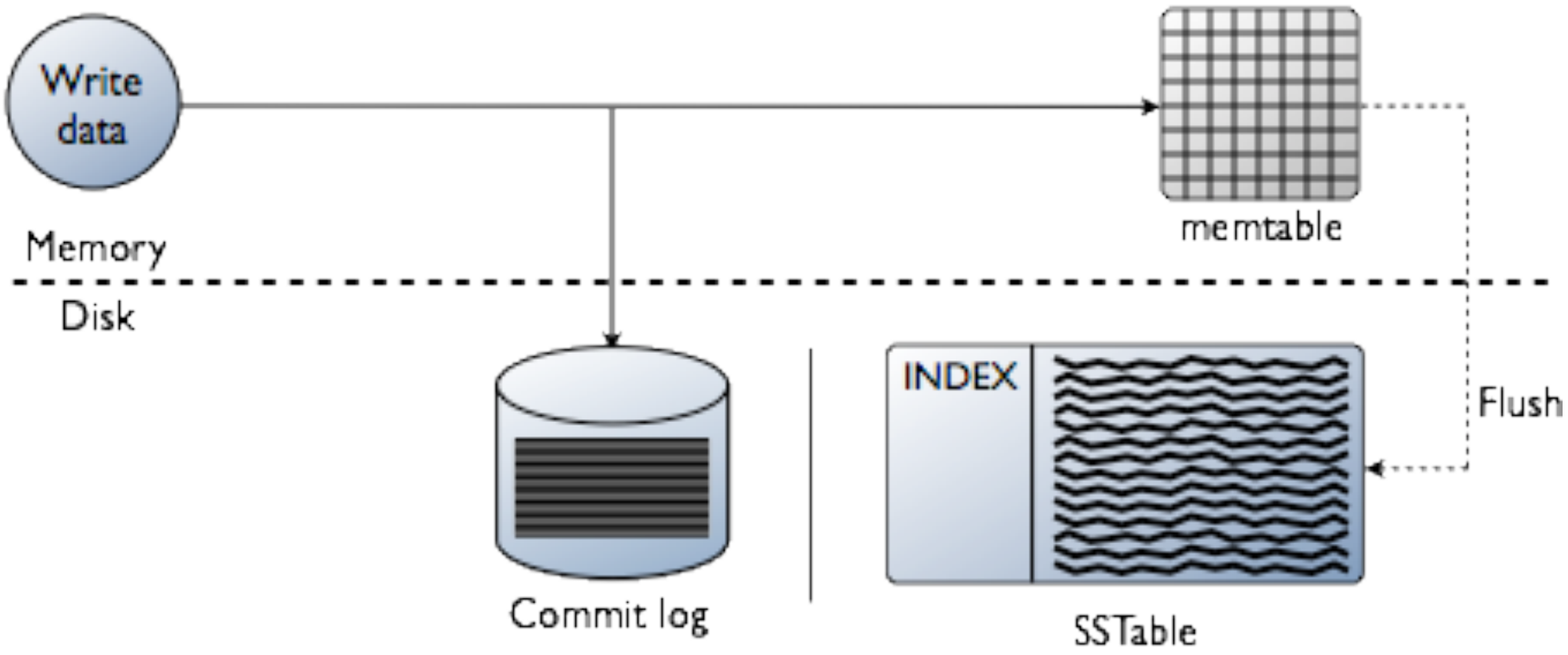
From the read path dashboard



A latency issue - Internals

Wait a minute, what are compactions in Cassandra?

Back to the write path!



A latency issue - Internals

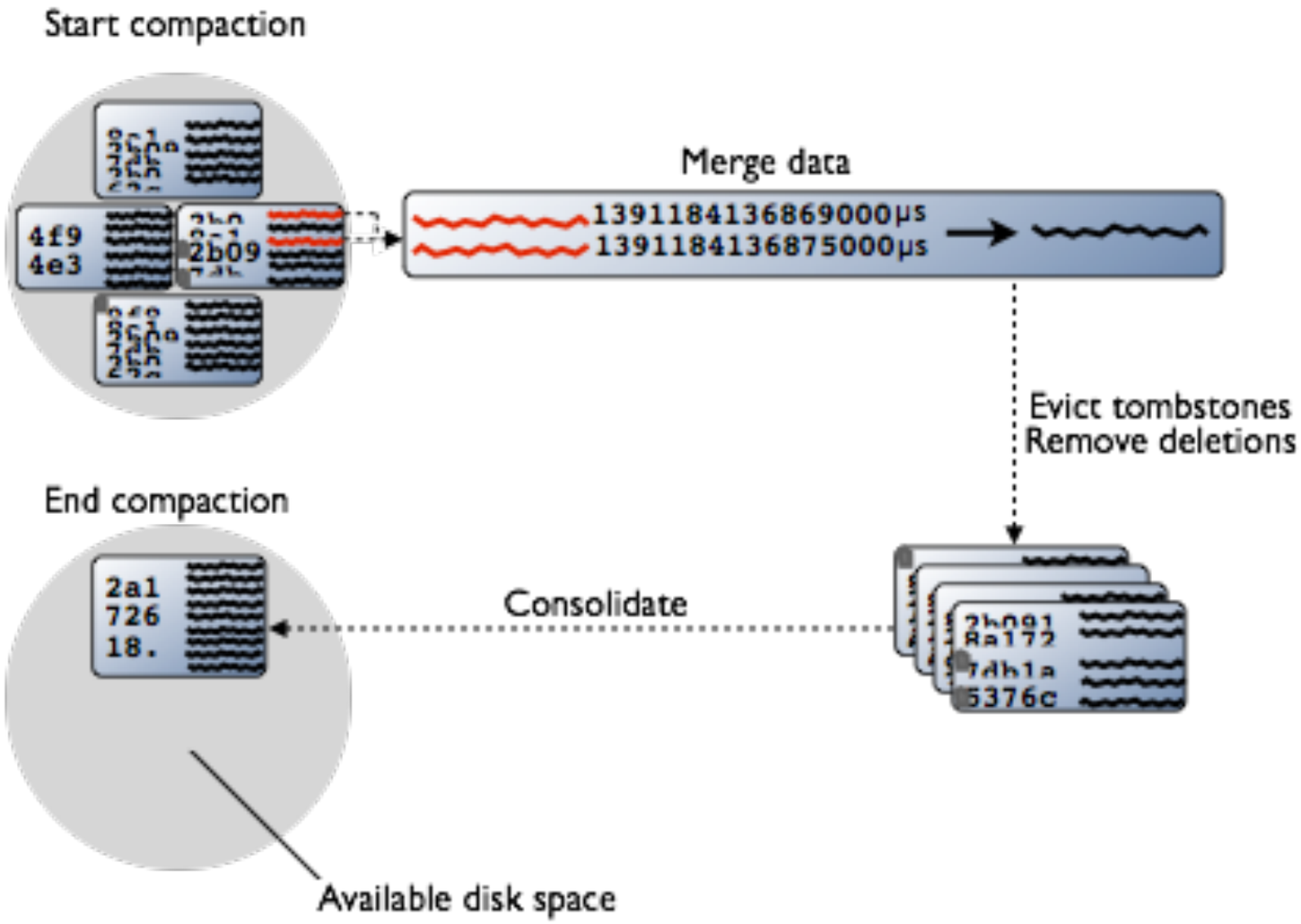
Wait a minute, what are compactions in Cassandra?

Inserts → Size up / Spread rows

Updates → Inserts

Deletes → Insert tombstone

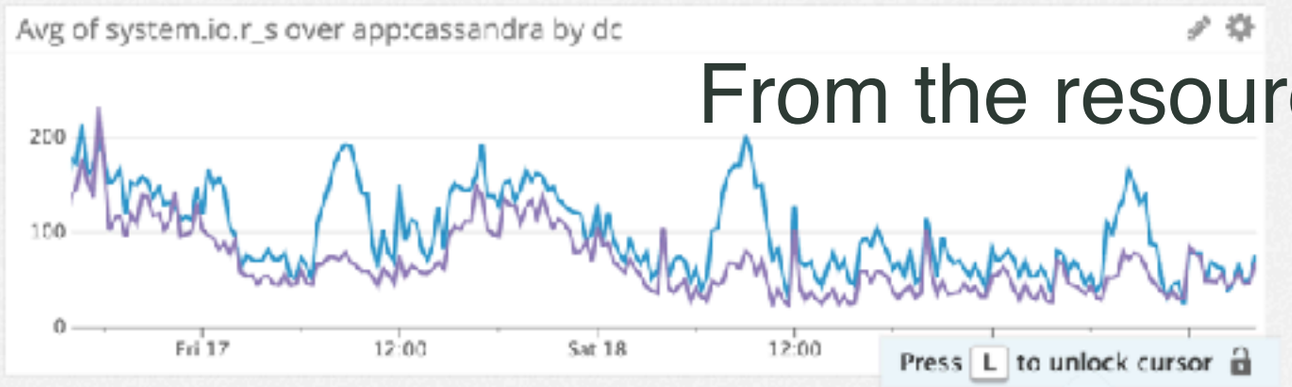
SSTables are immutable !

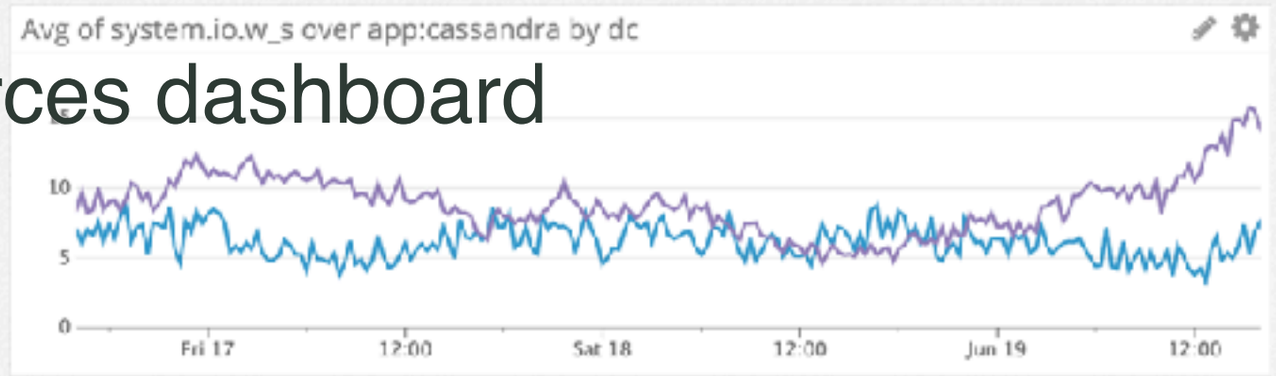


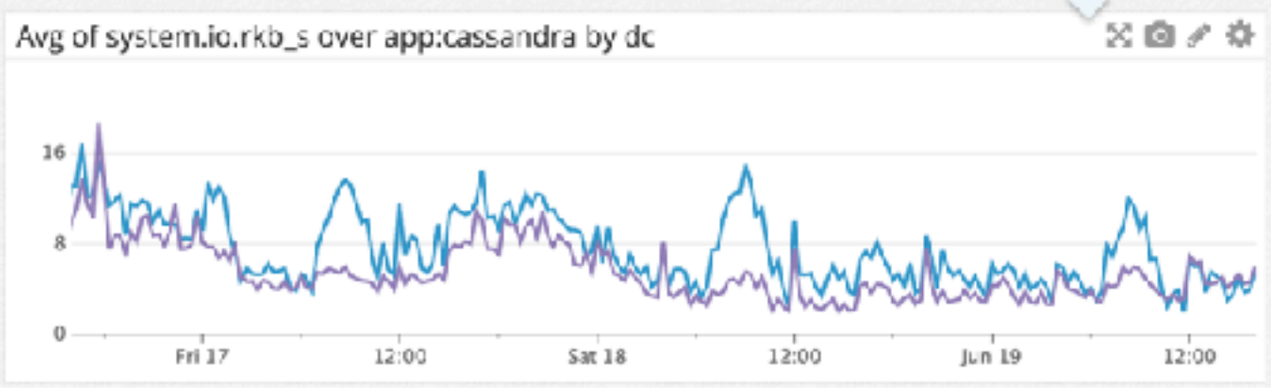
So, why are compactions not keeping up?

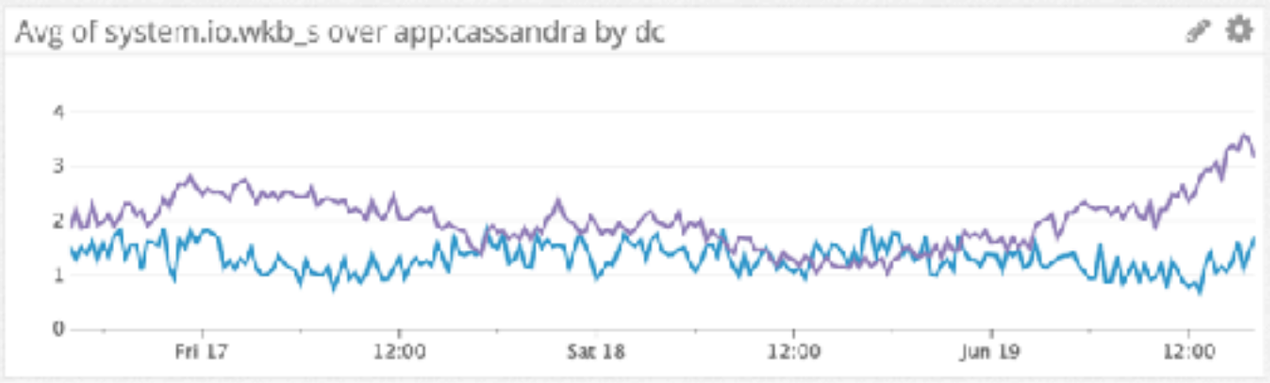
A latency issue - Troubleshooting

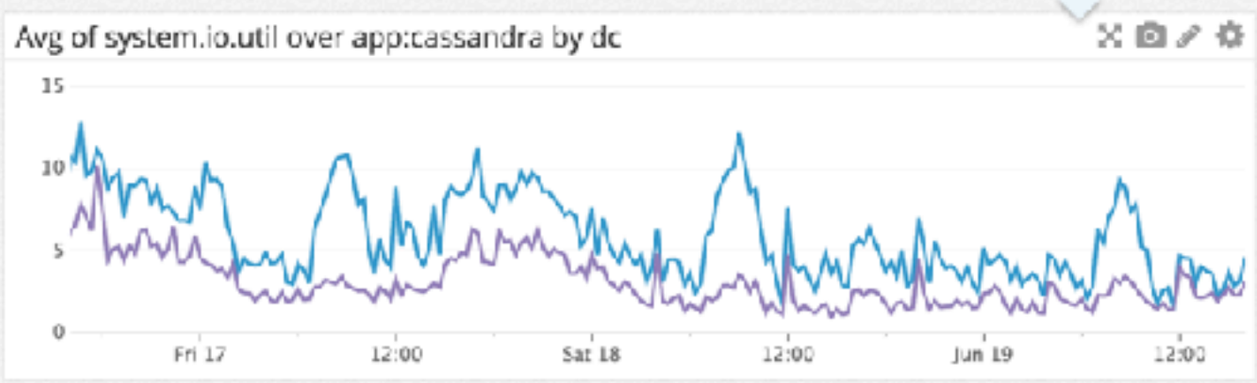
From the resources dashboard

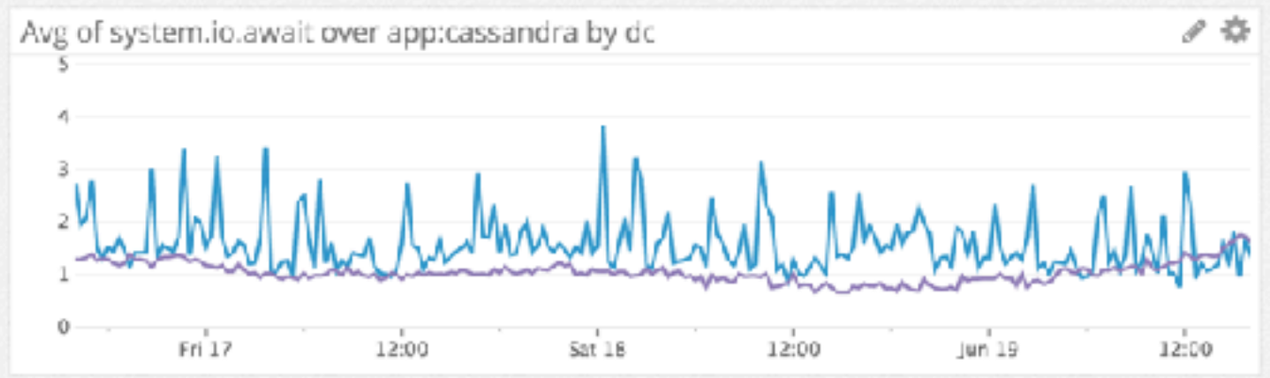


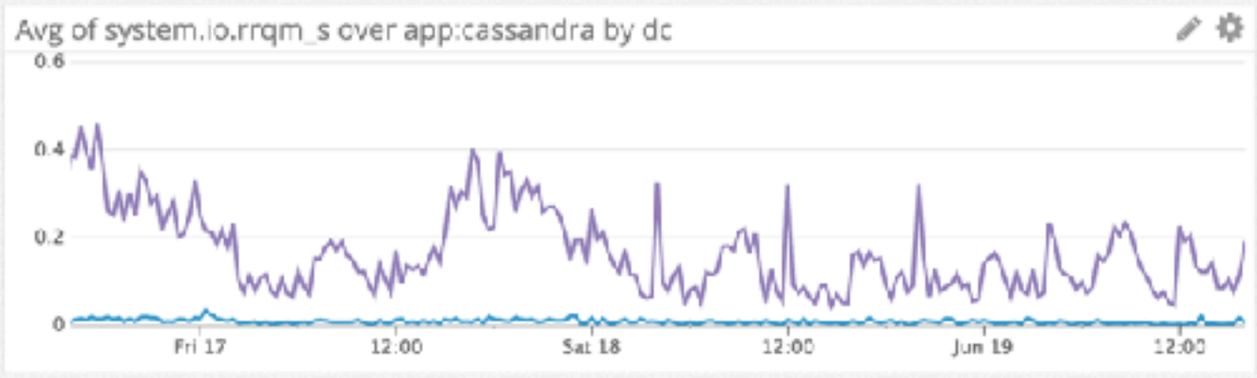


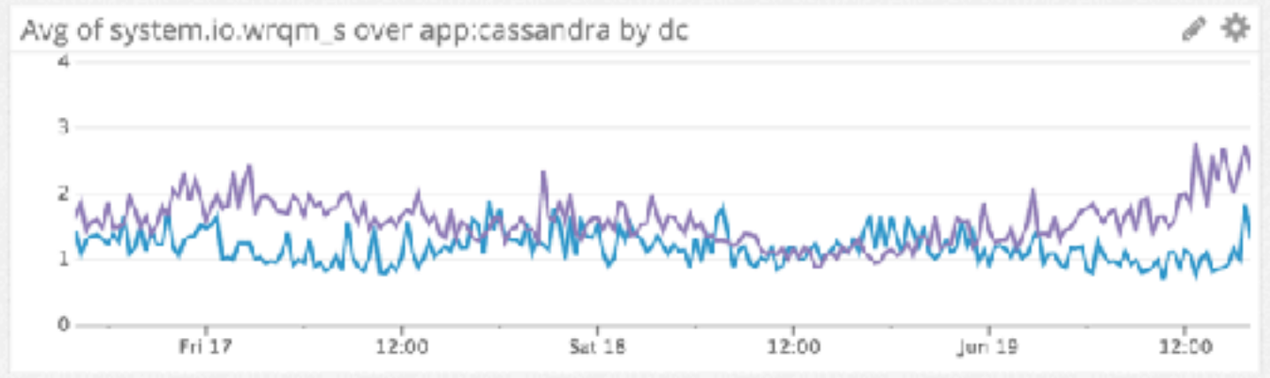












A latency issue - Troubleshooting

Summing things up

From Monitoring

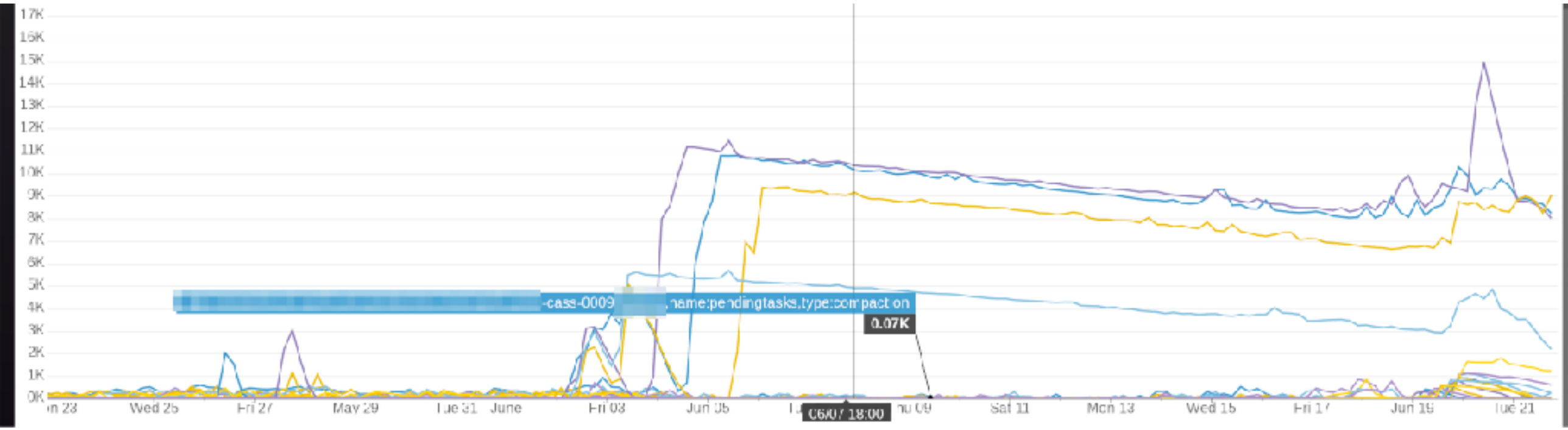
- Latency issue
- Dropped messages
- **High # of compactions pending**
- Frequent and long GC activity
- Read and Write paths affected
- Looks like a global outage
- Resources doing OK though

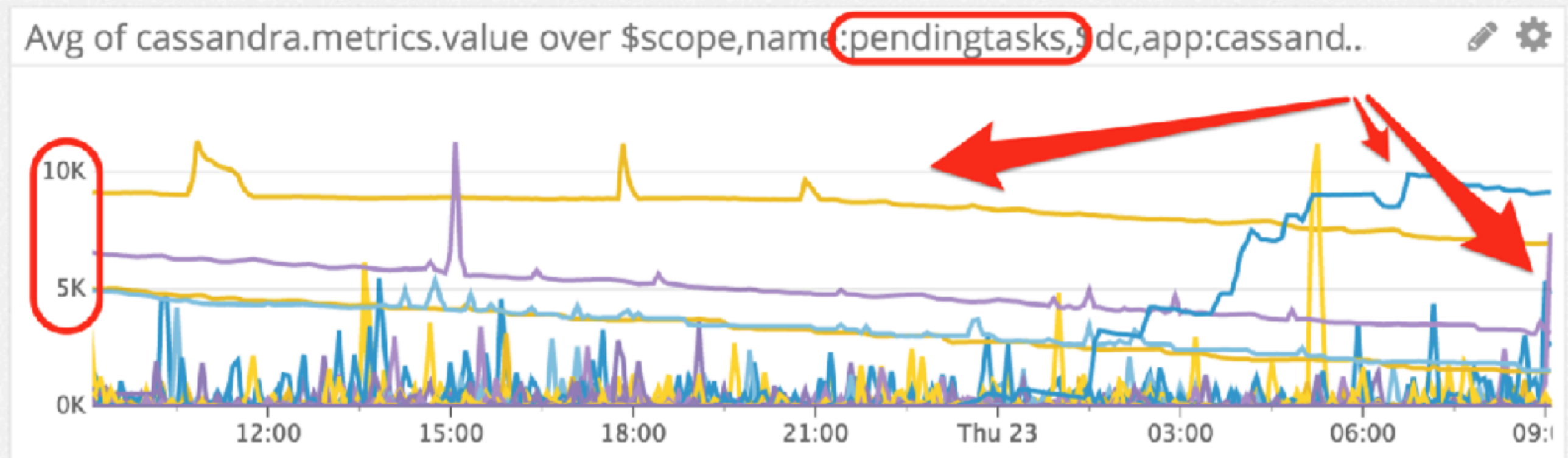
From context

- **Old Cassandra version**
- **Huge dataset (4 TB / node)**

A latency issue - Acting

Take 1 - Improve compactions





A latency issue - Internals

Take 1 - Improve compactions

Facing CASSANDRA-9662: No-op compactions triggering

A latency issue - Think

- Compaction max speed is not enough
- Disks are not a bottleneck, nor is CPU
- Not a resource issue, rather a Cassandra one
- Adding node is always possible solution
- Solving CASSANDRA-9662 is an other solution

A latency issue - Acting

Take 2 - Add nodes + Upgrade Cassandra

We added some nodes as a “quick” fix

Then planned an upgrade of the cluster

Problem solved

A dropped read issue - takeaway

We learned that about compactions in Cassandra

We also learned about a Cassandra issue

Conclusion

Do not undervalue Monitoring!

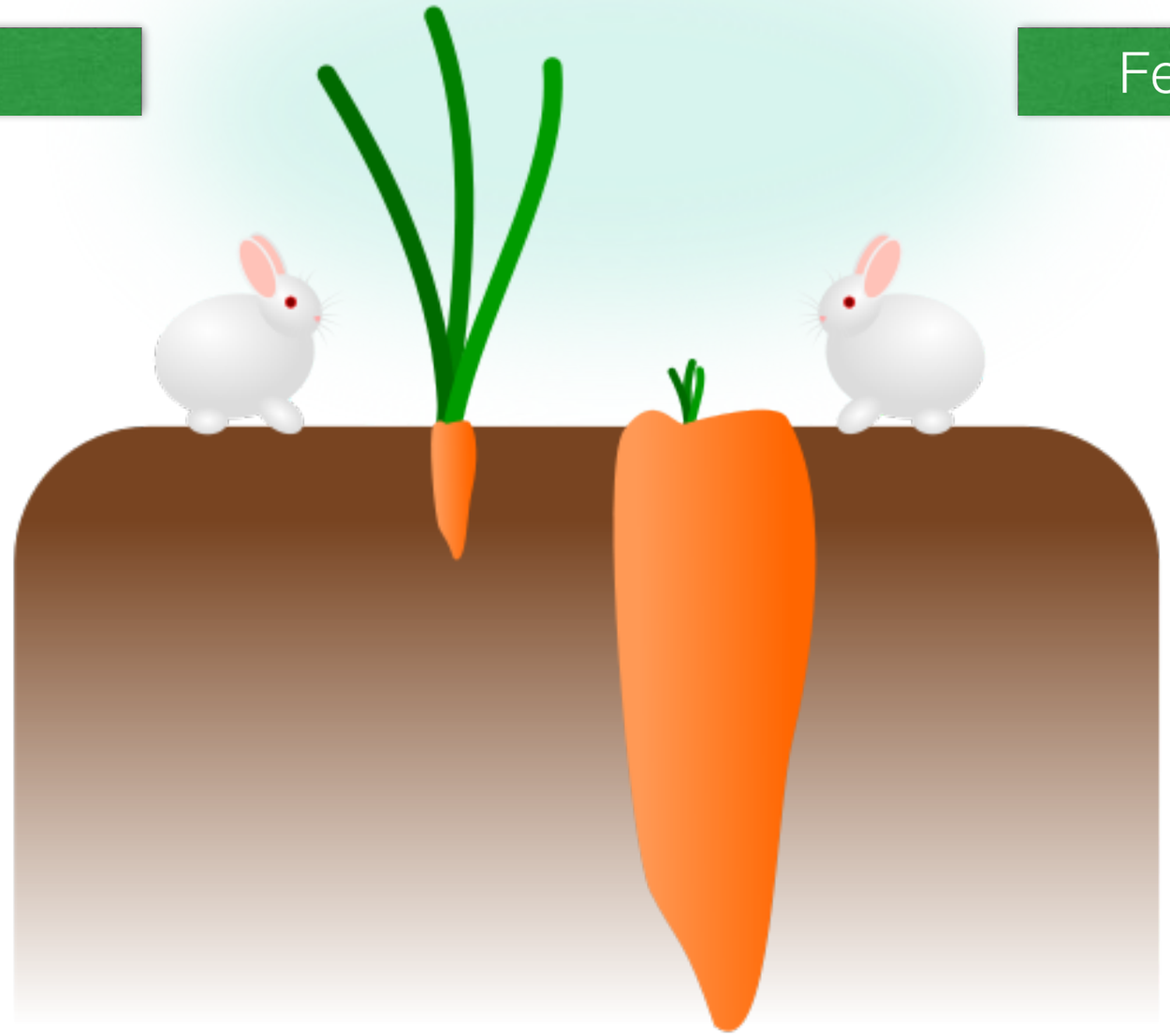
Anomaly detection

Knowledge

Troubleshooting

Optimization

Feel for the cluster



Learning

Use Monitoring to learn as you go!

Happy trip, I hope you'll chose the right way ;-)

Thanks!

@arodrime

I hope this was “enlightening.”

Questions?

THE LAST PICKLE