Grafana Labs

LGTM Stack

Or: Philosophy of Observability

Richard “RichiH” Hartmann

Why this matters to you

Or: Who is this person?

My Background

- Director of Community @ Grafana Labs
- Prometheus team member (CNCF graduated project)
  - PromCon lead, Prometheus Dev Summit chair
- CNCF Governing Board
- CNCF Technical Oversight Committee
- CNCF Technical Advisory Group Observability chair
- OpenMetrics founder
- OpenTelemetry member

My Background

- DENOG e.V. founder
- DENOG e.V. board emeritus
- Maintainer snmp_exporter & modbus_exporter
- Built Europe's most modern datacenter, monitored through SNMP and Modbus/TCP only
- Extensive paid and pro bono consulting on how to instrument existing and new applications

My Background

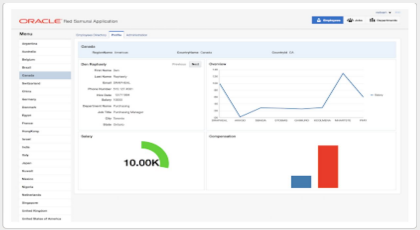
- Ran the backbone of an ISP for eleven years
  - Was the only person on call
  - My life & sanity depended on on-point monitoring & alerting
- Active in RIPE, IETF, DENOG, #networker etc
  - RFC to my name, changed RIPE NCC's IPv4 PI policies, etc.
- Prometheus transition for Germany's oldest ISP, ~5k devices
- Staffed world's largest IRC network for more than a decade
- Run conferences & monitoring from 100s to 18k attendees
  - DENOG, DebConf, FOSDEM, CCC, GrafanaCon, PromCon

I come from the trenches of tech

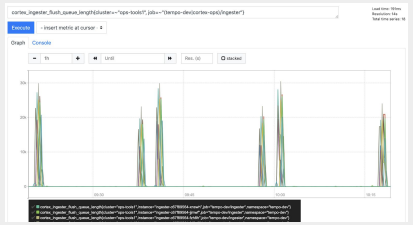
I know the pain, and I want to help

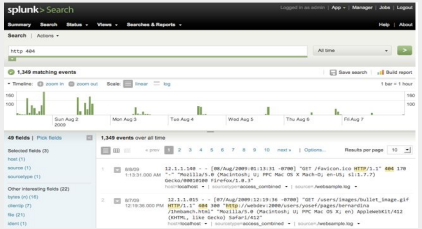
Today's reality:
Disparate systems. Disparate data.

ORACLE



APPDYNAMICS





Grafana

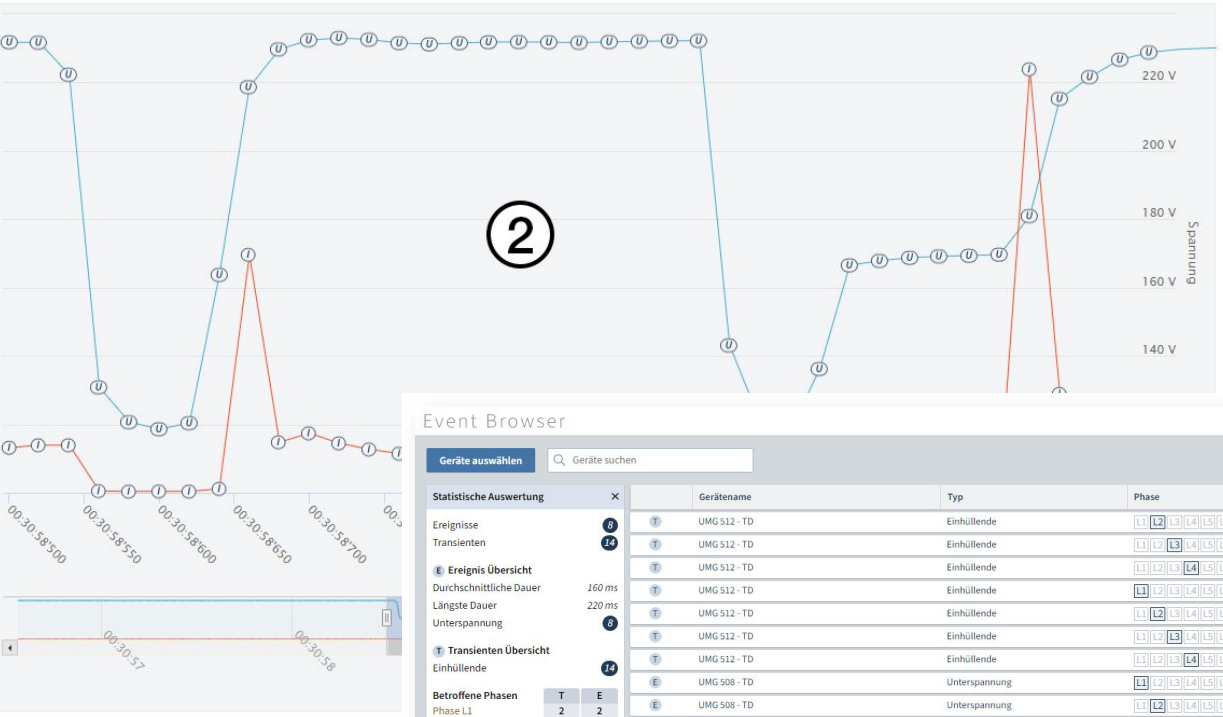
splunk>

elasticsearch

Back to the basics

Let’s rethink this

How humanity deals with data



Event Browser

Geräte auswählen | Geräte suchen | 19.12.2019 - 19.12.2019 00:00 - 23:59

Statistische Auswertung

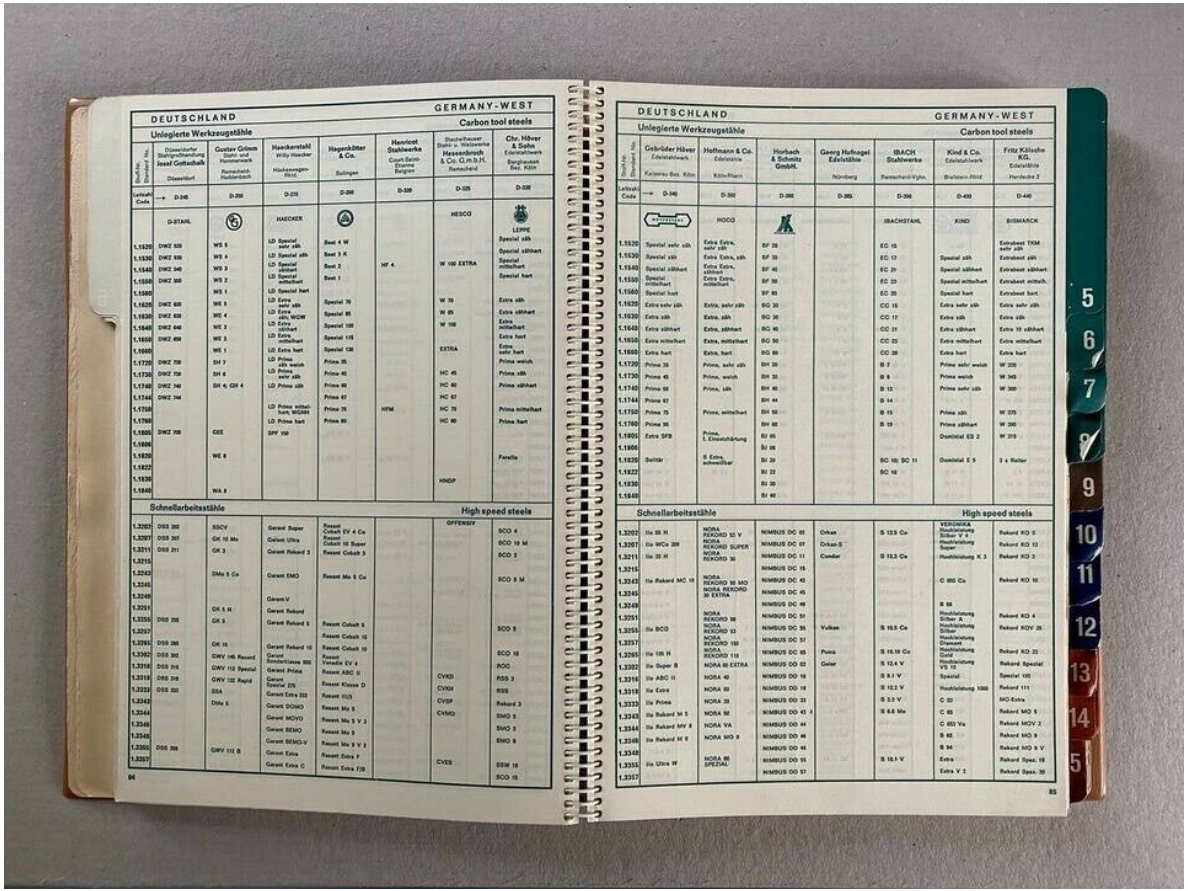
Ereignisse	8
Transienten	14
Ereignis Übersicht	
Durchschnittliche Dauer	160 ms
Längste Dauer	220 ms
Unterspannung	8
Transienten Übersicht	
Einhüllende	14

Betroffene Phasen	T	E
Phase L1	2	2
Phase L2	4	2
Phase L3	4	2
Phase L4	4	2

	Gerätename	Typ	Phase	Start	Ende	Dauer	Wert	
T	UMG 512 - TD	Einhüllende	L2	19.12.2019 01:15:40'663	---	---	---	Analysieren
T	UMG 512 - TD	Einhüllende	L3	19.12.2019 01:15:40'663	---	---	---	Analysieren
T	UMG 512 - TD	Einhüllende	L4	19.12.2019 01:15:40'663	---	---	---	Analysieren
T	UMG 512 - TD	Einhüllende	L1	19.12.2019 01:15:39'345	---	---	---	Analysieren
T	UMG 512 - TD	Einhüllende	L2	19.12.2019 01:15:39'345	---	---	---	Analysieren
T	UMG 512 - TD	Einhüllende	L3	19.12.2019 01:15:39'345	---	---	---	Analysieren
T	UMG 512 - TD	Einhüllende	L4	19.12.2019 01:15:39'345	---	---	---	Analysieren
E	UMG 508 - TD	Unterspannung	L1	19.12.2019 00:30:59'979	19.12.2019 00:30:59'199	220 ms	117.836 V (MIN)	Analysieren
E	UMG 508 - TD	Unterspannung	L2	19.12.2019 00:30:59'979	19.12.2019 00:30:59'199	220 ms	117.806 V (MIN)	Analysieren
E	UMG 508 - TD	Unterspannung	L3	19.12.2019 00:30:59'979	19.12.2019 00:30:59'199	220 ms	117.825 V (MIN)	Analysieren
E	UMG 508 - TD	Unterspannung	L4	19.12.2019 00:30:59'979	19.12.2019 00:30:59'199	220 ms	117.830 V (MIN)	Analysieren
E	UMG 508 - TD	Unterspannung	L1	19.12.2019 00:30:58'559	19.12.2019 00:30:58'659	100 ms	118.640 V (MIN)	Analysieren
E	UMG 508 - TD	Unterspannung	L2	19.12.2019 00:30:58'559	19.12.2019 00:30:58'659	100 ms	118.612 V (MIN)	Analysieren
E	UMG 508 - TD	Unterspannung	L3	19.12.2019 00:30:58'559	19.12.2019 00:30:58'659	100 ms	118.622 V (MIN)	Analysieren
E	UMG 508 - TD	Unterspannung	L4	19.12.2019 00:30:58'559	19.12.2019 00:30:58'659	100 ms	118.631 V (MIN)	Analysieren
T	UMG 512 - TD	Einhüllende	L2	19.12.2019 00:28:40'565	---	---	---	Analysieren
T	UMG 512 - TD	Einhüllende	L3	19.12.2019 00:28:40'565	---	---	---	Analysieren
T	UMG 512 - TD	Einhüllende	L4	19.12.2019 00:28:40'565	---	---	---	Analysieren
T	UMG 512 - TD	Einhüllende	L1	19.12.2019 00:28:40'130	---	---	---	Analysieren
T	UMG 512 - TD	Einhüllende	L2	19.12.2019 00:28:40'130	---	---	---	Analysieren

20 | 1 2

Grafana Labs



DEUTSCHLAND — GERMANY-WEST

Unlegierte Werkzeugstähle — Carbon tool steels

Stahl-Nr. / Standard-Nr.	Düsseldorfer Stahlgroßhandlung Josef Gottschalk, Düsseldorf	Custav Grimm Stahl- und Kunststoffwerk, Remscheid-Hasten-bach	Haeckerstahl Willy Haecker, Hückeswagen-Rhld.	Hagenkötter & Co., Solingen	Henricot Stahlwerke Court-Saint-Etienne, Belgien	Stahlschmiede Stahl u. Walzwerke Hessenbruch & Co. G.m.b.H., Remscheid	Chr. Höver & Sohn Edelstahlwerk, Burgbrohl/Bez. Köln
Leitzahl / Code	→ D-345	D-350	D-370	D-390	D-330	D-335	D-330
	D-STAHL		HAECKER			HESCO	LEPPE
1.1520	DWZ 520	WS 5					Special sähr
1.1530	DWZ 530	WS 4	LD Special sähr	Best 4 W			Special sähhart
1.1540	DWZ 540	WS 3	LD Special sähhart	Best 3 K			Special mittelhart
1.1550	DWZ 550	WS 2	LD Special mittelhart	Best 2	HF 4	W 100 EXTRA	Special hart
1.1560		WS 1	LD Special hart	Best 1			
1.1620	DWZ 620	WE 5	LD Extra sehr sähr	Special 70		W 70	Extra sähr
1.1630	DWZ 630	WE 4	LD Extra sähr, WCW	Special 85		W 85	Extra sähhart
1.1640	DWZ 640	WE 3	LD Extra sähhart	Special 100		W 100	Extra mittelhart
1.1650	DWZ 650	WE 2	LD Extra mittelhart	Special 115			Extra hart
1.1660		WE 1	LD Extra hart	Special 130		EXTRA	Extra sehr hart
1.1720	DWZ 720	SH 7	LD Prima sähr weich	Prima 35			Prima weich
1.1730	DWZ 730	SH 6	LD Prima sehr sähr	Prima 45		HC 45	Prima sähr
1.1740	DWZ 740	SH 4, GH 4	LD Prima sähr	Prima 60		HC 60	Prima sähhart
1.1744	DWZ 744			Prima 67		HC 67	
1.1750			LD Prima mittelhart, WQMH	Prima 75	HFM	HC 75	Prima mittelhart
1.1760			LD Prima hart	Prima 90		HC 85	Prima hart
1.1805	DWZ 705	GEE	SPF 150				
1.1806							
1.1820		WE 6					Ferelle
1.1822							
1.1830						HNDP	
1.1840		WA 8					

Schnellarbeitsstähle — High speed steels

1.3202	DSS 202	SSCV	Garant Super	Resant Cobalt EV 4 Co		OFFENSIV	SCO 4
1.3207	DSS 207	GK 10 Mo	Garant Ultra	Resant Cobalt 10 Super			SCO 10 M
1.3211	DSS 211	GK 3	Garant Rekord 3	Resant Cobalt 3			SCO 3
1.3215							
1.3243		DMo 5 Co	Garant EMO	Resant Mo 5 Co			SCO 5 M
1.3245							
1.3249			Garant V				
1.3251		GK 5 Ni	Garant Rekord				
1.3255	DSS 255	GK 5	Garant Rekord 5	Resant Cobalt 5			SCO 5
1.3257				Resant Cobalt 15			
1.3265	DSS 265	GK 10	Garant Rekord 10	Resant Cobalt 10			SCO 10
1.3302	DSS 302	GWV 145 Record	Garant Sonderklasse 500	Resant Vanadia EV 4			ROC
1.3310	DSS 310	GWV 112 Special	Garant Prima	Resant ABC II		CVKD	RSS 2
1.3318	DSS 318	GWV 132 Rapid	Garant Special 215	Resant Klasse D		CVDH	RSS
1.3333	DSS 333	SSA	Garant Extra 333	Resant 1123		CVSP	Rekord 3
1.3343		DMo 5	Garant DOMO	Resant Mo 5		CVMO	SMO 5
1.3344			Garant MOVO	Resant Mo 5 V 2			SMO 3
1.3348			Garant REMO	Resant Mo 9			SMO 8
1.3349			Garant REMO-V	Resant Mo 9 V 2			
1.3355	DSS 355	GWV 112 S	Garant Extra	Resant Extra F		CVES	SSW 18
1.3357			Garant Extra C	Resant Extra 735			SCO 15

64

DEUTSCHLAND — GERMANY-WEST

Unlegierte Werkzeugstähle — Carbon tool steels

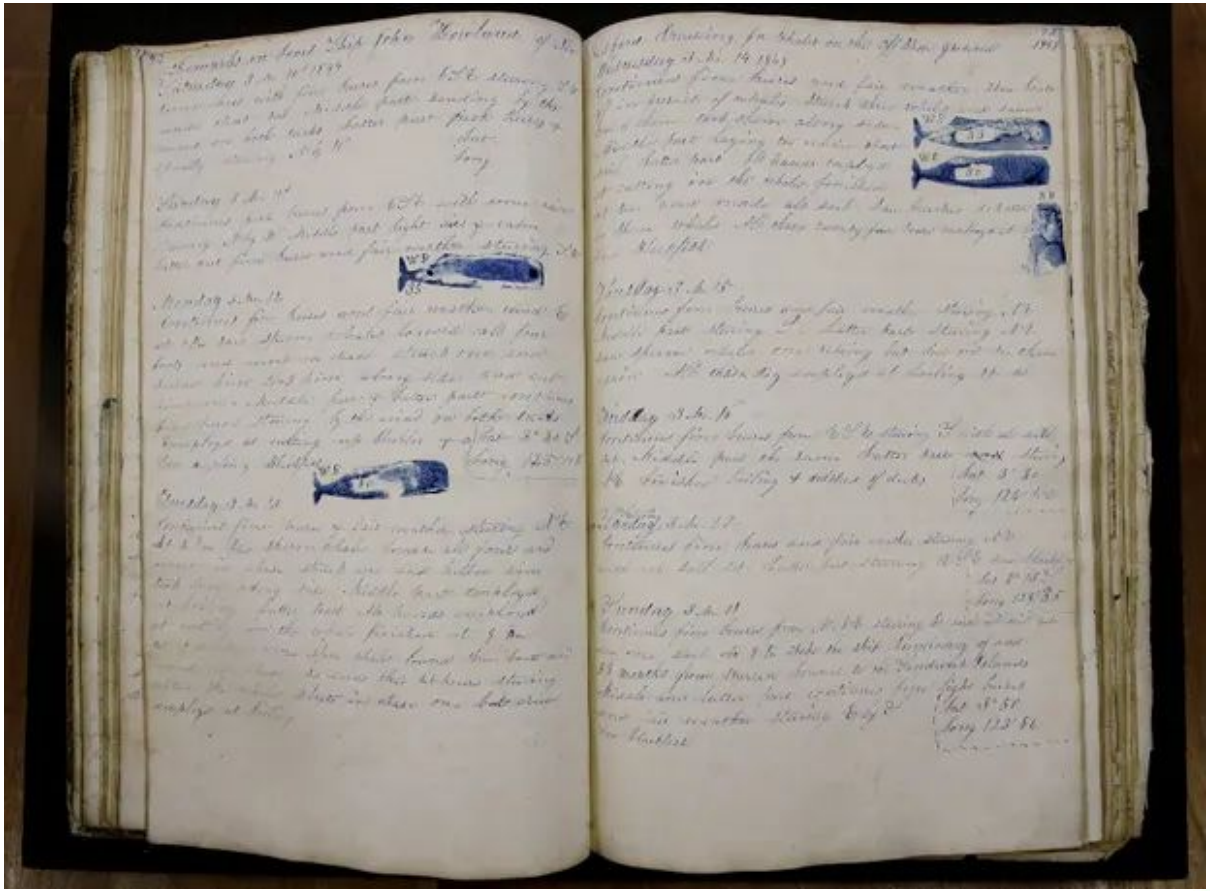
Stahl-Nr. / Standard-Nr.	Gebrüder Höver Edelstahlwerk, Kaiserau Bez. Köln	Hoffmann & Co. Edelstahle, Köln-Rhein	Horbach & Schmitz GmbH,	Georg Hofnagel Edelstähle, Nürnberg	IBACH Stahlwerke, Pamschard/Vgbs.	Kind & Co. Edelstahlwerk, Brebach-Rhld.	Fritz Kälsche KG. Edelstähle Hardecke 2
Leitzahl / Code	→ D-340	D-362	D-366	D-365	D-360	D-400	D-440
		HOCO			IBACHSTAHL	KIND	BISMARCK
1.1520	Special sehr sähr	Extra Extra, sehr sähr	BF 35		EC 15		Extrabest TKM sehr sähr
1.1530	Special sähr	Extra Extra, sähr	BF 30		EC 17	Special sähr	Extrabest sähr
1.1540	Special sähhart	Extra Extra, sähhart	BF 40		EC 21	Special sähhart	Extrabest sähhart
1.1550	Special mittelhart	Extra Extra, mittelhart	BF 50		EC 23	Special mittelhart	Extrabest mittelh.
1.1560	Special hart		BF 60		EC 25	Special hart	Extrabest hart
1.1620	Extra sehr sähr	Extra, sehr sähr	BQ 35		CC 15	Extra sehr sähr	Extra sehr sähr
1.1630	Extra sähr	Extra, sähr	BQ 30		CC 17	Extra sähr	Extra sähr
1.1640	Extra sähhart	Extra, sähhart	BQ 40		CC 21	Extra sähhart	Extra 10 sähhart
1.1650	Extra mittelhart	Extra, mittelhart	BQ 50		CC 23	Extra mittelhart	Extra mittelhart
1.1660	Extra hart	Extra, hart	BQ 60		CC 25	Extra hart	Extra hart
1.1720	Prima 35	Prima, sehr sähr	BH 35		B 7	Prima sehr weich	W 335
1.1730	Prima 45	Prima, weich	BH 30		B 9	Prima weich	W 345
1.1740	Prima 60	Prima, sähr	BH 40		B 13	Prima sehr sähr	W 360
1.1744	Prima 67		BH 44		B 14		
1.1750	Prima 75	Prima, mittelhart	BH 50		B 15	Prima sähr	W 370
1.1760	Prima 90		BH 60		B 19	Prima sähhart	W 390
1.1805	Extra SFB	Prima, f. Einsatzhärtung	BJ 05			Dominiat ES 2	W 315
1.1806			BJ 06				
1.1820	Solidur	S Extra, schweißbar	BJ 20		SC 10; SC 11	Dominiat E 9	3 x Reiter
1.1822			BJ 22		SC 16		
1.1830			BJ 30				
1.1840			BJ 40				

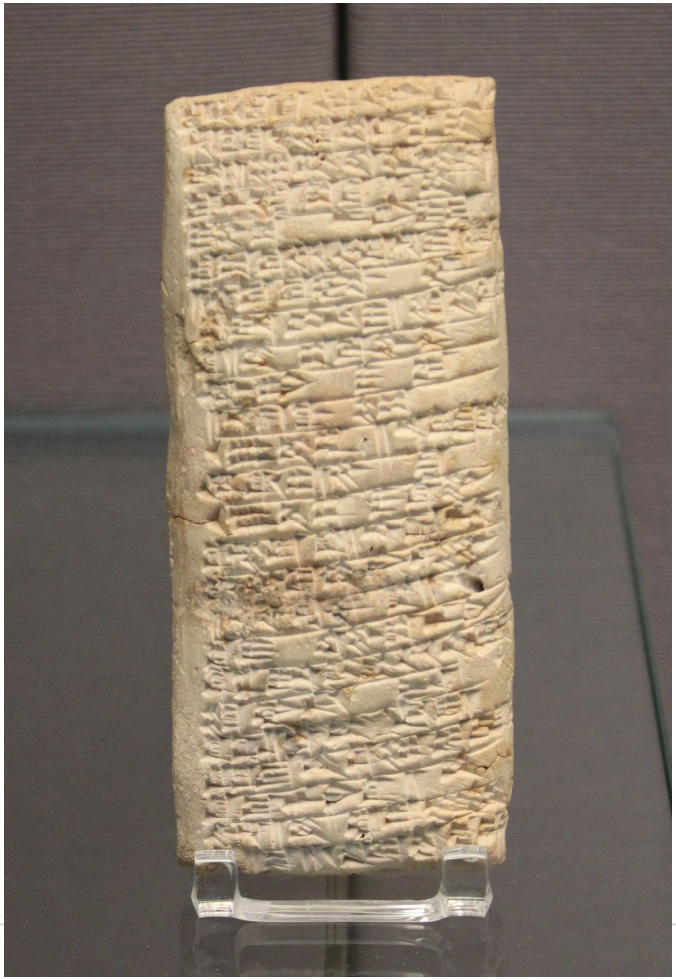
Schnellarbeitsstähle — High speed steels

1.3202	Ila 55 H	NORA REKORD 53 V	NIMBUS DC 55	Orkan	S 12,5 Co	VERONIKA Hochleistung Silber V 5	Rekord KO 5
1.3207	Ila WCo 209	NORA REKORD SUPER	NIMBUS DC 57	Orkan-S		Hochleistung Super	Rekord KO 12
1.3211	Ila 35 H	NORA REKORD 30	NIMBUS DC 11	Condor	S 12,3 Co	Hochleistung K 3	Rekord KO 3
1.3215			NIMBUS DC 15				
1.3243	Ila Rekord MC 10	NORA REKORD 55 MO	NIMBUS DC 45			C 850 Co	Rekord KO 10
1.3245		NORA REKORD 35 EXTRA	NIMBUS DC 45				
1.3249			NIMBUS DC 49			B 55	
1.3251		NORA REKORD 58	NIMBUS DC 51			Hochleistung Silber V 4	Rekord KO 4
1.3255	Ila SCO	NORA REKORD 53	NIMBUS DC 55	Vulkan	S 10,5 Co	Hochleistung Silber	Rekord KOV 25
1.3257		NORA REKORD 150	NIMBUS DC 57			Hochleistung Diamant	
1.3265	Ila 155 H	NORA REKORD 110	NIMBUS DC 65	Puma	S 10,10 Co	Hochleistung Gold	Rekord KO 22
1.3302	Ila Super B	NORA 66 EXTRA	NIMBUS DD 62	Geier	S 13,4 V	Hochleistung VS 10	Rekord Special
1.3310	Ila ABC II	NORA 40	NIMBUS DD 10		S 9,1 V	Special	Special 100
1.3318	Ila Extra	NORA 60	NIMBUS DD 18		S 12,2 V	Hochleistung 1000	Rekord 111
1.3333	Ila Prima	NORA 33	NIMBUS DD 33		S 3,3 V	C 33	MO-Extra
1.3343	Ila Rekord M 5	NORA 50	NIMBUS DD 43		S 6,6 Mo	C 65	Rekord MO 5
1.3344	Ila Rekord MV 8	NORA VA	NIMBUS DD 44			C 653 Va	Rekord MOV 2
1.3348	Ila Rekord M 9	NORA MO 9	NIMBUS DD 48			B 65	Rekord MO 9
1.3349			NIMBUS DD 49			B 94	Rekord MO 9 V
1.3355	Ila Ultra W	NORA 88 SPEZIAL	NIMBUS DD 55		S 18,1 V	Extra	Rekord Spez. 18
1.3357			NIMBUS DD 57			Extra V 2	Rekord Spez. 35

5 6 7 8 9 10 11 12 13 14 15

83





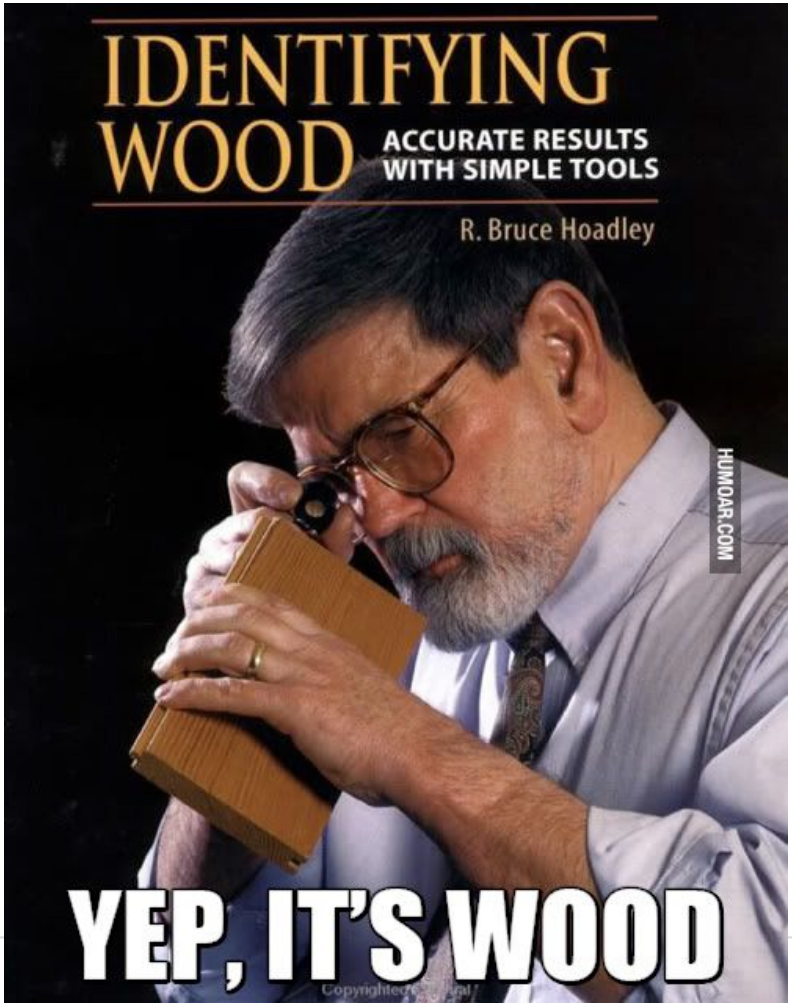
Humanity has optimized detailed accounts into key events into numbers for millenia, again and again

Observability & SRE

Or: Buzzwords, and their useful parts

Observability, the buzzword

- Cool new term, almost meaningless by now, what does it mean?
  - Pitfall alert: Cargo culting
  - It's about changing the behaviour, not about changing the name
- "Monitoring" has taken on a meaning of collecting, not using data
  - One extreme: Full text indexing
  - Other extreme: Data lake
- "Observability" is about enabling humans to understand complex systems
  - Ask new questions on the fly
  - Ask **why** it's not working instead of just knowing that it's not


IDENTIFYING WOOD
ACCURATE RESULTS WITH SIMPLE TOOLS
R. Bruce Hoadley
HUMOAR.COM
YEP, IT'S WOOD

[...] observations are [...] approximations to the truth [...] this can be accomplished in no other way than by a suitable combination of more observations than the number absolutely requisite for the determination of the unknown quantities

Carl Friedrich Gauß, 1809

Observability is a measure of how well internal states of a system can be inferred from knowledge of its external outputs.

Rudolf Emil Kálmán, 1960

Complexity

- Fake complexity, a.k.a. bad design
  - Can be reduced
- Real, system-inherent complexity
  - Can be moved (monolith vs client-server vs microservices)
  - Must be compartmentalized (service boundaries)
  - Should be distilled meaningfully (observability…)

Services

- What's a service?
  - Compartmentalized complexity, with an interface
  - Different owners/teams
  - Contracts define interfaces
- Why "contract": Shared agreement which MUST NOT be broken
  - Internal and external customers rely on what you build and maintain
- Other common term: layer
  - The Internet would not exist without network layering
  - Enables innovation, parallelizes human engineering
- Other examples: CPUs, hard drive, compute nodes, your lunch

SRE, an instantiation of DevOps

- At its core: Align incentives across the org
  - Error budgets allow devs, ops, PMs, etc. to optimize for shared benefits
- Measure it!
  - SLI: Service Level Indicator: What you measure
  - SLO: Service Level Objective: What you need to hit
  - SLA: Service Level Agreement: When you need to pay

Shared understanding

- Everyone uses the same tools & dashboards
  - Shared incentive to invest into tooling
  - Pooling of institutional system knowledge
  - Shared language & understanding of services

Alerting

- Customers care about services being up, not about individual components
- Discern between different SLIs
  - Primary: service-relevant, for alerting
  - Secondary: informational, debugging, might be underlying's primary

Anything currently or imminently impacting customer service must be alerted upon

But nothing(!) else

Prometheus

Prometheus 101

- Inspired by Google's Borgmon
- Time series database
- Rich ecosystem, 10,000s of instrumentation & exporters
- Cloud-native default

Time series

- Time series are recorded values which change over time
- Individual events are usually merged into counters and/or histograms
- Changing values are recorded as gauges
- Typical examples
  - Requests to a webserver (counter)
  - Temperatures in a datacenter (gauge)
  - Service latency (histograms)

Cloud-native default

- Kubernetes =~ Borg
- Prometheus =~ Borgmon
- Google couldn't have run Borg without Borgmon (plus Omega and Monarch)
- Kubernetes & Prometheus are designed and written with each other in mind

Prometheus 101

- Black-box monitoring: Looking at a service from the outside (Does the server answer to HTTP requests?)
- White-box monitoring: Instrumenting code from the inside (How much time does this subroutine take?)
- Every service should have its own metrics endpoint
- Hard API commitments within major versions
- New release candidate every six weeks

Main selling points

- Highly dynamic, built-in service discovery
- No hierarchical model, n-dimensional label set
- PromQL: for processing, graphing, alerting, and export
- Simple operation
- Highly efficient

Super easy to emit, parse & read

```
http_requests_total{env="prod",method="post",code="200"} 1027
http_requests_total{env="prod",method="post",code="400"} 3
http_requests_total{env="prod",method="post",code="500"} 12
http_requests_total{env="prod",method="get",code="200"} 20
http_requests_total{env="test",method="post",code="200"} 372
http_requests_total{env="test",method="post",code="400"} 75
```

PromQL

All partitions in my entire infrastructure with more than 100GB capacity that are not mounted on root?

```
node_filesystem_bytes_total{mountpoint!="/"} / 1e9 > 100
```

```
{device="sda1", mountpoint="/home", instance="10.0.0.1"} 118.8
{device="sda1", mountpoint="/home", instance="10.0.0.2"} 118.8
{device="sdb1", mountpoint="/data", instance="10.0.0.2"} 451.2
{device="xdvc", mountpoint="/mnt", instance="10.0.0.3"} 320.0
```

PromQL

What's the ratio of request errors across all service instances?

```
sum by(path) (rate(http_requests_total{status="500"}[5m])) /
 sum by(path) (rate(http_requests_total[5m]))

 {path="/status"} 0.0039
 {path="/"} 0.0011
 {path="/api/v1/topics/:topic"} 0.087
 {path="/api/v1/topics} 0.0342
```

New features

- Remote Write Receiver (v2.25 (feature flag) v2.33 (stable))
- Trigonometric functions (v2.31)
- Agent mode (v2.32)
- Long term support versions (v2.27)
- Out of order ingestion (v2.39)

Next highlight feature: Native histograms

Prometheus scale

- 1,000,000+ samples/second no problem on current hardware
- ~200,000 samples/second/core
- 16 bytes/sample compressed to 1.36 bytes/sample
- Reliable into the tens of millions of active series

OpenMetrics

OpenMetrics

Text format MUST be supported:

```
vendor_interface_in_bytes_total{interface="TenGigE0/0/2/0"} 12345678
vendor_interface_out_bytes_total{interface="TenGigE0/0/2/0"} 34567890
vendor_psu_status{module="psu1"} 1
vendor_temperature_celsius{module="psu1",location="inlet"} 35
```

Equivalent protobuf MAY be used

OpenMetrics

- Standardizing Prometheus exposition format
- Submitted to IETF OPSAWG, presented at IETF 110
- Substantial interest from WG
  - Document working 1.0 first
  - Open floor for careful evolution afterwards
  - High-resolution histograms one of the 2.x main features
- Feel free to slip the ID / RFC into your tender documents ;)

Mimir

Mimir

- For **M**etrics
- Prometheus → Cortex → Grafana Enterprise Metrics → Mimir
- Scales to more than 1,000,000,000 Active Series
- Blazingly fast query performance
- Hard multi-tenancy, access control, and three-way replication
- Can ingest native OpenTelemetry, DataDog, Graphite, and Influx

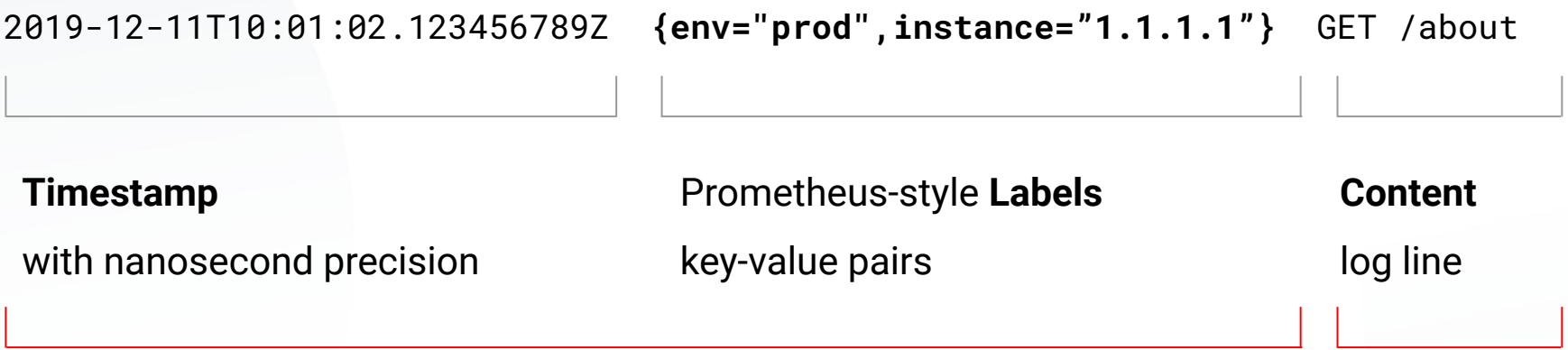
Mimir @ Grafana

- 1,000,000,000 Active Series - in one cluster
- 1,500 machines
- 7,000 CPU cores
- 30 TiB RAM

Loki

Loki 101

- For **L**ogs
- Following the same label-based system as Prometheus
  - Only index what you need often, query the rest
  - “Index the labels, query the data”
- Work with logs at scale, without the massive cost
  - Scalable low latency write path
  - Flexible schema on read
- Access logs with the same label sets as metrics
  - Turn logs into metrics, to make it easier & cheaper to work with them


2019-12-11T10:01:02.123456789Z
{env="prod",instance="1.1.1.1"}
GET /about
Timestamp
with nanosecond precision
Prometheus-style Labels
key-value pairs
Content
log line
indexed
unindexed

Loki @ Grafana Labs

- Largest user cluster (as of 2022-09): 180 TiB per day
- Queries regularly see 80 GiB/s
- Query terabytes of data in under a minute
  - Including complex processing of result sets

Tempo

Tempo

- For **T**races
- Historic problem: Traces require extremely rich metadata for analysis
  - Expensive, slow, and mandates sampling
- Exemplars: Leverage the extracted logs & metrics
  - Exemplars work at Google scale, with the ease of Grafana
  - Native to Prometheus, Cortex, Thanos, and Loki
- Index and search by labelsets available for those who need it
- 100% compatible with OpenTelemetry Tracing, Zipkin, Jaeger

Tempo @ Grafana Labs (2022-09)

- 2,200,000 samples per second @ 350 MiB/s
  - 5,000,000 samples second peak
- 14-day retention @ 3 copies stored
- Latencies:
  - p99 - 2.5s
  - p90 - 2.3s
  - p50 - 1.6s

Phlare

Phlare

- For **P**rofiling
- Pronounced “Flare”
- Profiles
  - “How much CPU & RAM am I spending in what areas of the code?”
  - “...and how does this change over time?”
- Go: pprof
- Java: https://github.com/grafana/JPProf

Data (and cost) savings

Logs to metrics

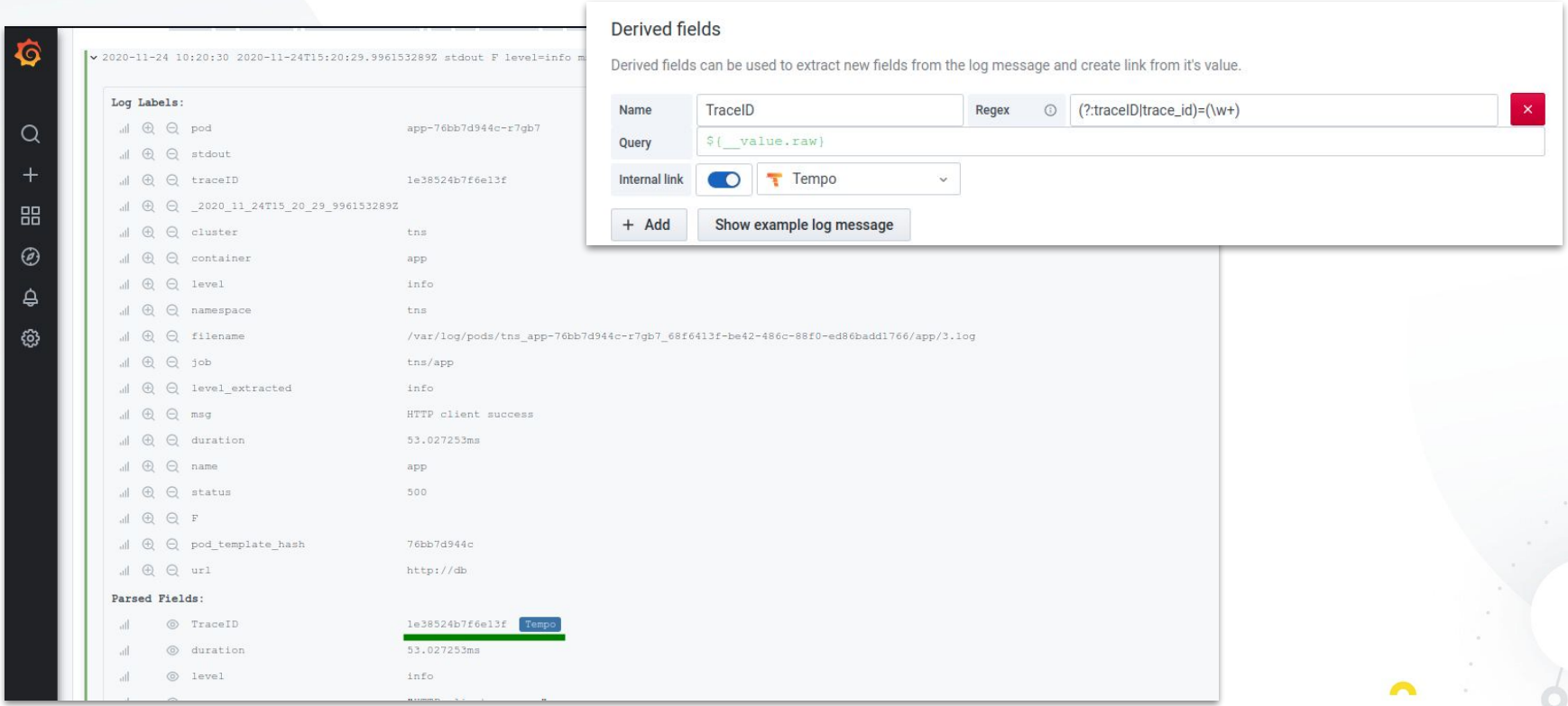
- Full text indexing: 10 TiB logs → ~20 TiB index
- Loki: 10 TiB logs → ~200 MiB index

- Logs @ Grafana ~600 Byte average per line
- Metrics ~1.36 Byte per metric sample

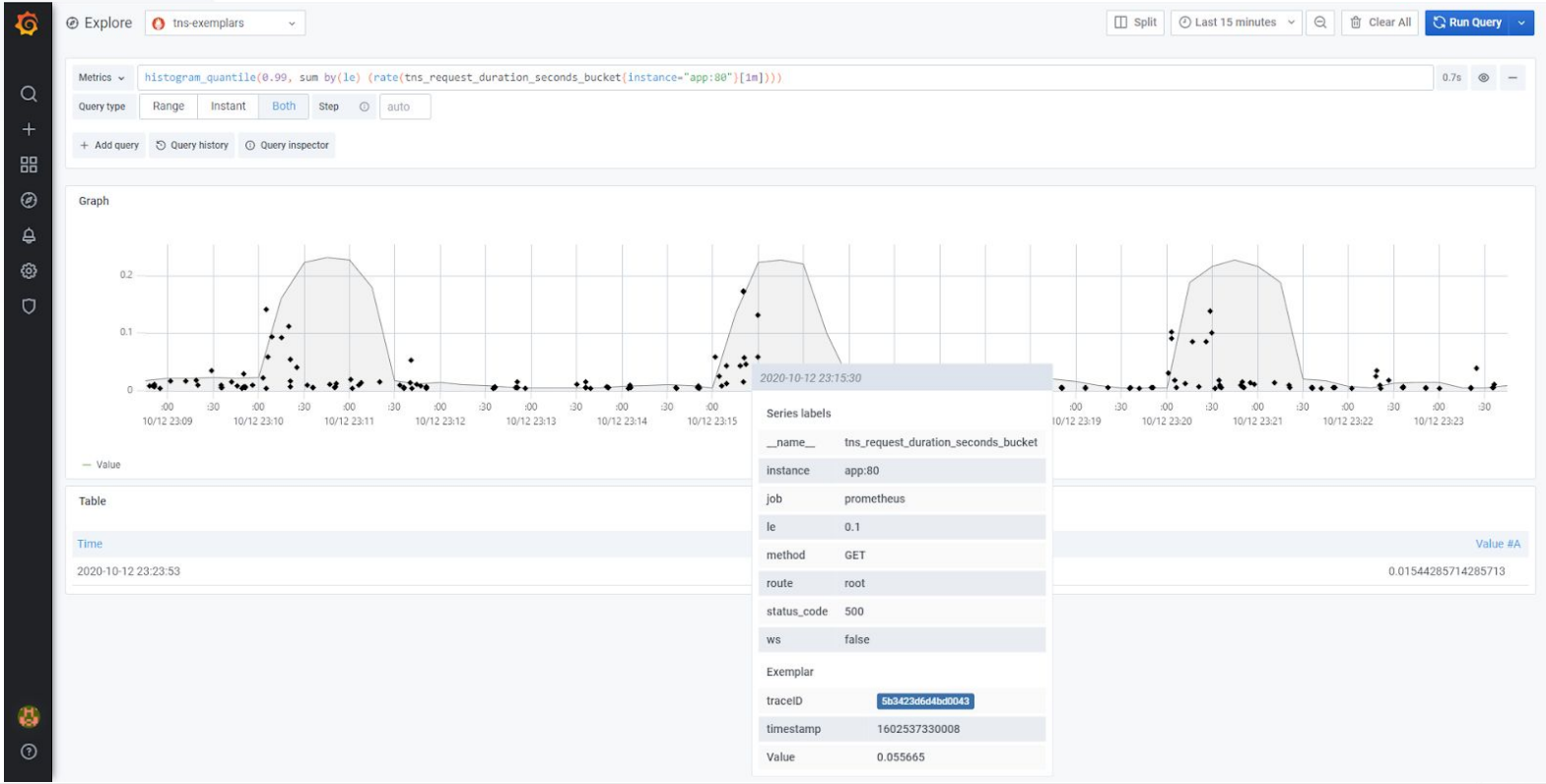
→ 99.8% reduction in storage size for first log line
~100% for every follow-up log line

Bringing it together

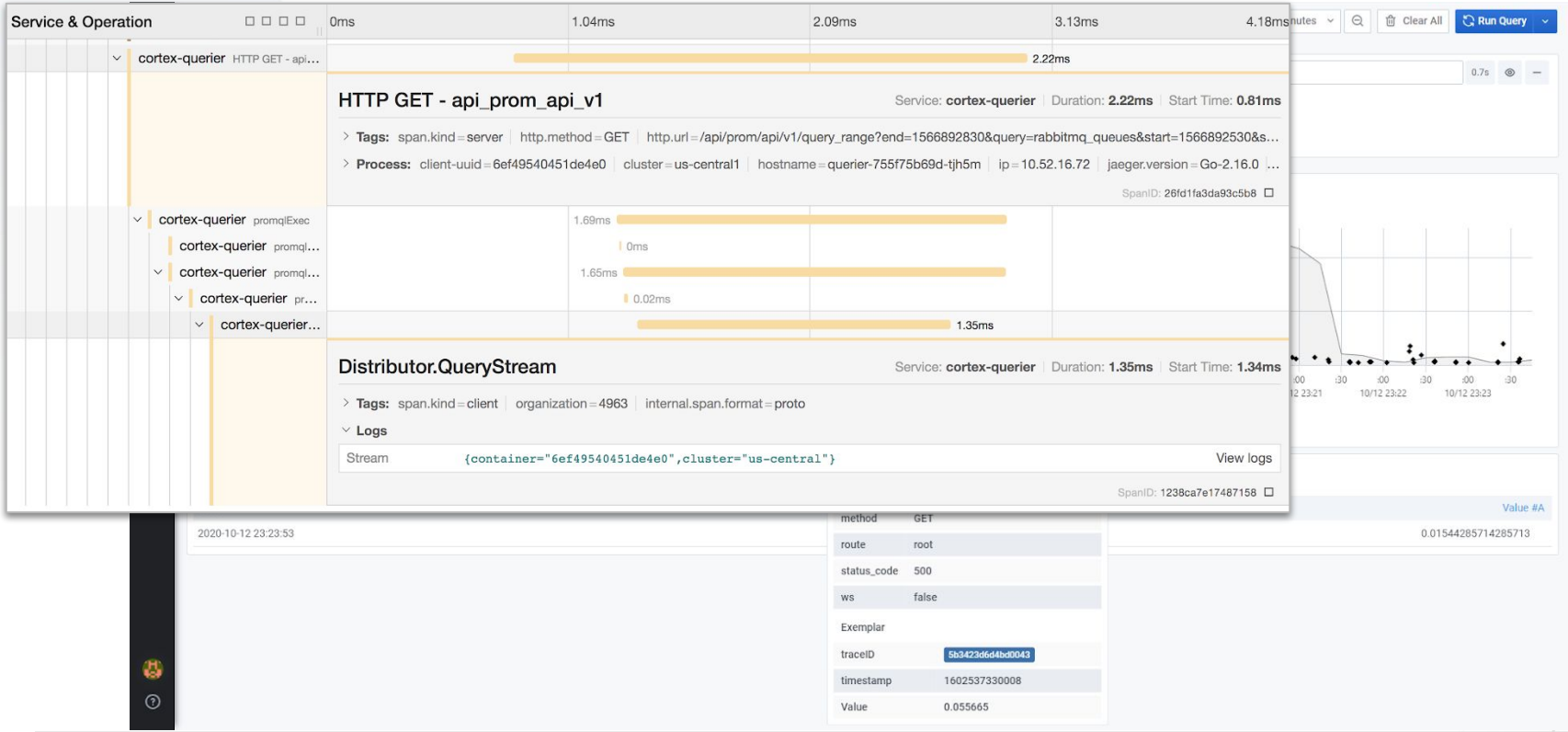
From logs to traces



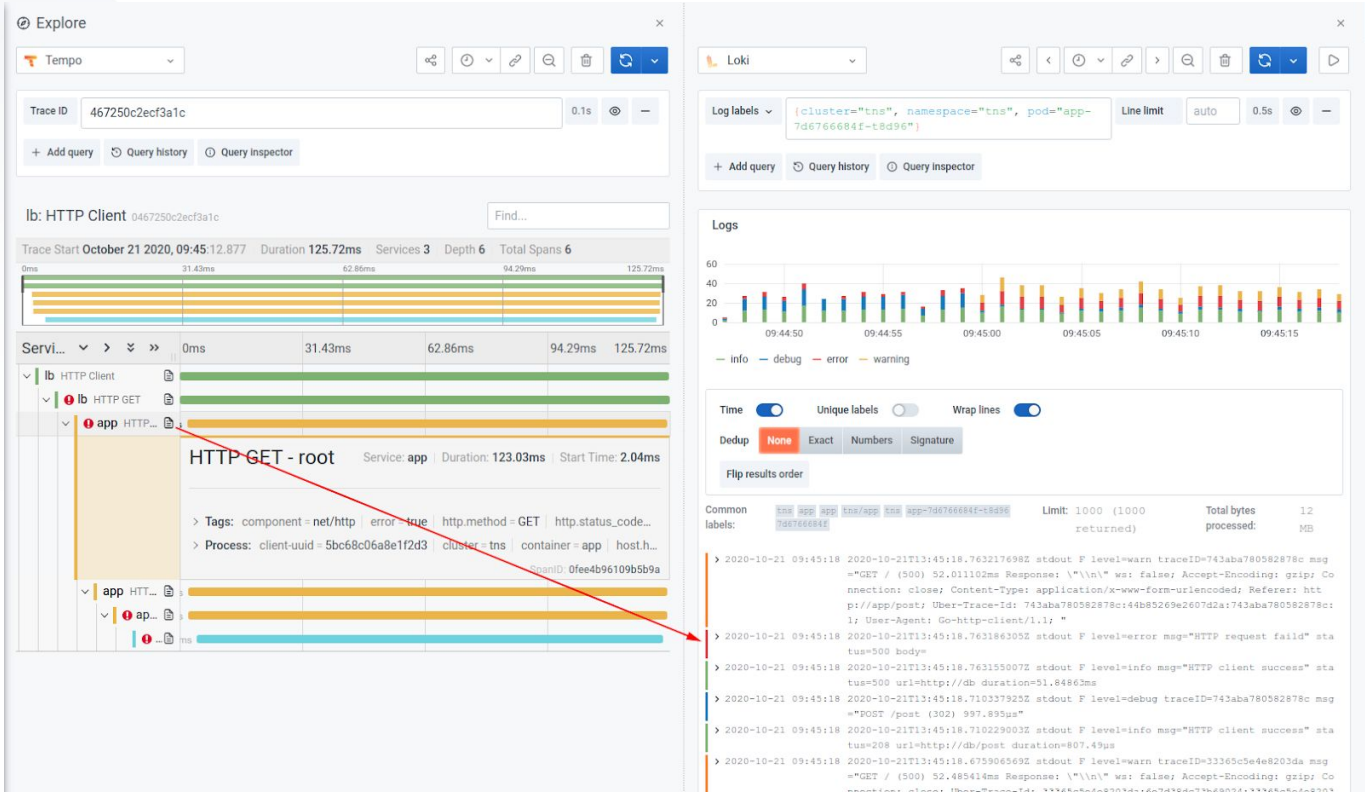
From metrics to traces



From metrics to traces

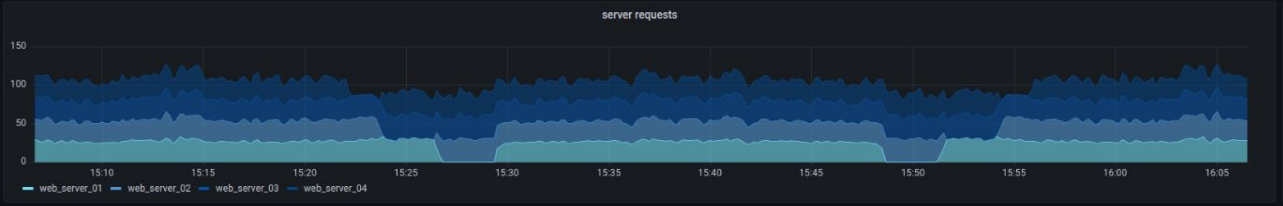


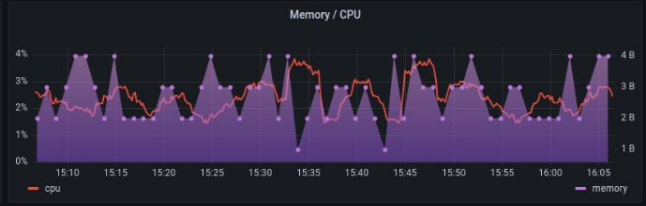
...and from traces to logs

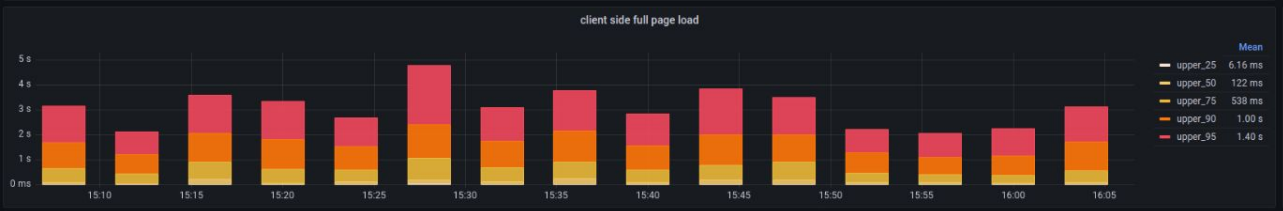


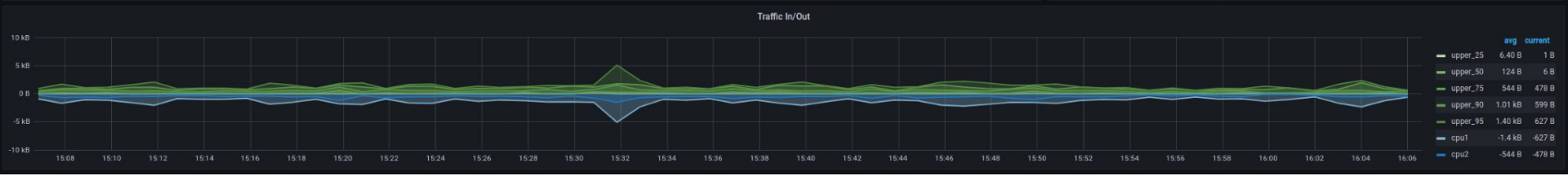
All of this is Open Source and you can run it yourself

(But we will also sell it to you happily)

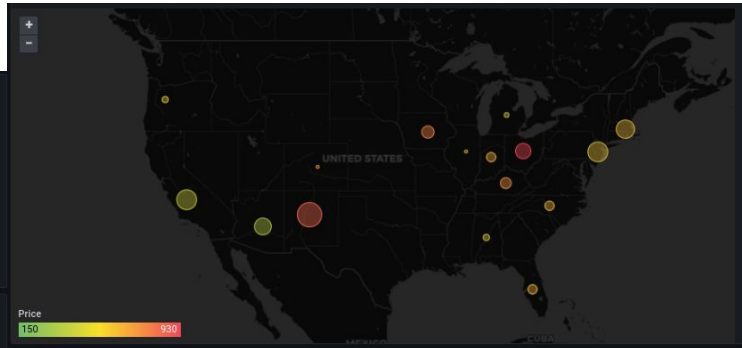


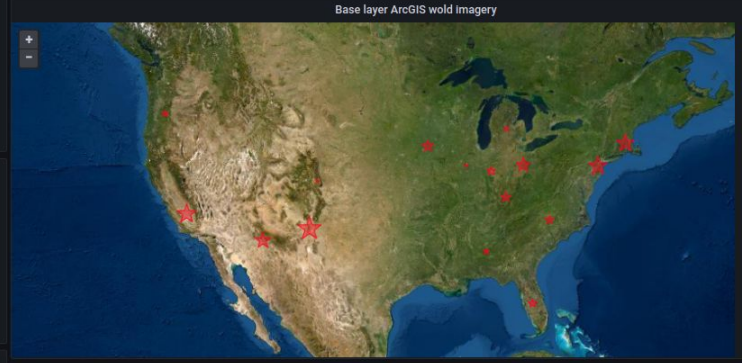


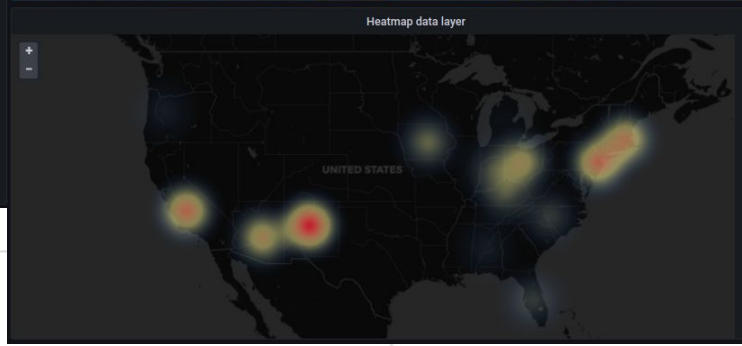




Grafana Labs







Grafana Labs

For a deeper dive into
Open Source Observability, go to:

grafana.com/about/events/observabilitycon/2022/

Thank you!

richih@richih.org
chaos.social/@RichiH
github.com/RichiH/talks