The Observatorium

Using ML & Observability together to reduce Incident Impact

Data Council New York City 2019

alex@digitalocean.com

✓, TOC.

1. `alex@digitalocean:~$ whoami/who_we_are`
2. The Observatorium: ***Foundations*** *and* ***Motivations***
3. Putting the pieces together, 1 event at a time
4. 2020 Vision
5. Questions (and Answers?)

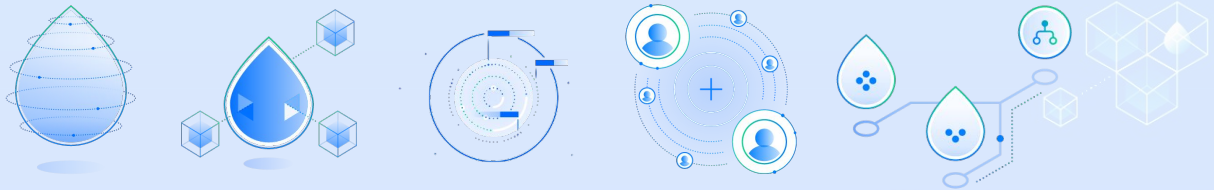
DigitalOcean

alex@digitalocean:~$ whoami/who_we_are

Global Cloud Hosting Provider

12 Data Centers, worldwide

DO builds **products** that help engineering teams build, deploy and scale cloud applications

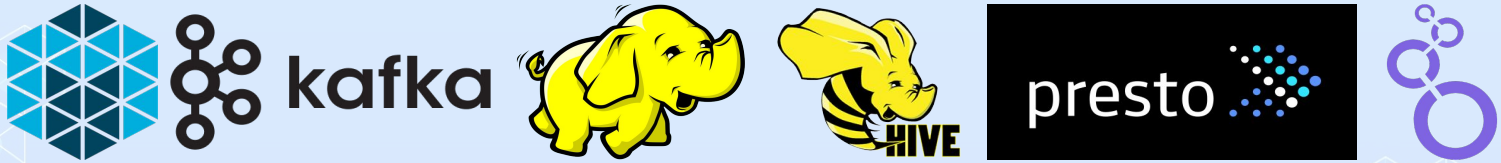


DigitalOcean

alex@digitalocean:~$ whoami/who_we_are

Observability Applications + Infra Analytics

Analytics Infrastructure



DigitalOcean

alex@digitalocean:~$ whoami/who_we_are

Observability Applications + Infra Analytics

What is the OA Mission?

- To **simplify** and **optimize** internal consumption of data from distributed systems
- To **reduce** incident **MTTD/MTTR** through custom applications
- To help **define**, **maintain**, and **broadcast** source-of-truth performance and reliability data to the rest of the organization

DigitalOcean

alex@digitalocean:~$ whoami/who_we_are

Observability Applications + Infra Analytics

What is the IA Mission?

- To **generate insights** through data for the Infrastructure and wider orgs
- To build and oversee a **centralized data platform**
- To help **define**, **maintain**, and **broadcast** source-of-truth ~~performance and reliability~~ data to the rest of the organization

DigitalOcean

alex@digitalocean:~$ whoami/who_we_are

But how can we achieve these things?

- To **simplify** and **optimize** internal consumption of data from distributed systems
- To **reduce** incident **MTTD/MTTR** through custom applications
- To **generate insights** through data for the Infrastructure and wider orgs
- To build and oversee a **centralized data platform**
- To help **define**, **maintain**, and **broadcast** source-of-truth (performance and reliability) data to the rest of the organization

DigitalOcean

alex@digitalocean:~$ whoami/who_we_are

But how can we achieve these things?

The Observatorium

DigitalOcean

The Observatorium

Foundations and Motivations

The Observatorium

The Observatorium: Foundations & **Motivations** (**what**/why)

A centralized application to help **reduce MTTD/MTTR**

i.e. the cost/impact of incidents

DigitalOcean

The Observatorium: Foundations & **Motivations** (what/**why**)

"I want to know the **current health of the cloud**"

DigitalOcean

The Observatorium: Foundations & **Motivations** (what/**why**)

"I want to see the live health and **historical performance** of all services that relate to Droplet Creation."

DigitalOcean

"There's currently an outage. I wonder if any **outages like this one** have occurred before and if so, how they were fixed."

The Observatorium: Foundations & **Motivations** (what/**why**)

"I want to understand the reliability of any/all **customer-facing products over time**."

DigitalOcean

The Observatorium: Foundations & **Motivations** (what/**why**)

"How much of our team's weekly/monthly/annual **error budget** have we depleted as of today?"

DigitalOcean

The Observatorium: Foundations & **Motivations** (what/**why**)

"I want to know if there are **warning signs** around the current performance of my service(s) that will lead to **degradation in the near future**."

DigitalOcean

The Observatorium: **Foundations** & Motivations (**what**/why)

How can we start building to answer these questions?

The Observatorium: **Foundations** & Motivations (**what**/why)

How can we start building to answer these questions?

Foundations:

SLM	Service Catalog	Observability Platforms

DigitalOcean

Service **L**evel **M**anagement

SLAs SLOs SLIs

SLA

an Agreement with consequences

SLO

an Objective, or goal (!= commitment)

SLI

an Indicator, or metric, that reveals whether an SLO is being met

SLA = service consumption (#2)

SLO/SLI = service production (#1)

Q1: Who **owns** the SLOs/SLIs for individual services?

A1: The service owner teams

Q2: Where are these SLOs/SLIs defined?

A2: A “catalog of services”...

Service Catalog

"A Central Authority for Distributed Microservices"

Requirement: a service *must* have a complete SC entry **to be allowed to deploy to production**.

But what is a "complete" entry?

A complete entry:

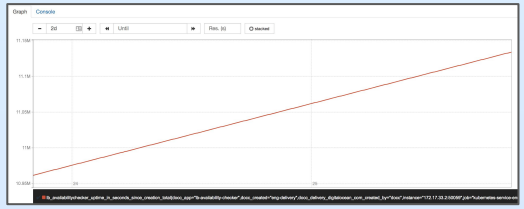
```
contact: TEAM_EMAIL@digitalocean.com
criticality: SEV-1
desc: <text about the Harpoon service ...>
dependencies: [2,5,7,8,13,14]
github: https://link/to/github/repo/README.md
id: 1
jira: HPN
name: harpoon
notes: <more text>
pager_duty: PD_CODE
product: droplet
slack: '#harpoon'
sli: sum(increase(harpoon_server_request_duration_seconds_count{code!="Internal",
code!="Unavailable", docc_app="harpoon-server"}[2m])) /
sum(increase(harpoon_server_request_duration_seconds_count{docc_app="harpoon-server"}[2m]))
slo: .995
team: Harpoon
```

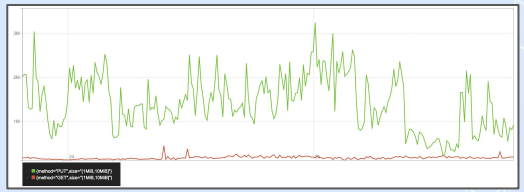
Observability Platforms:

Prometheus / Pandora

Prometheus / Pandora

- Easy to implement and deploy at scale
- Flexible time-series metrics
  - Counters
  - Gauges
  - Recording Rules (SLIs!)





The Observatorium: **Foundations**

SLM | Service Catalog | **Observability Platforms**

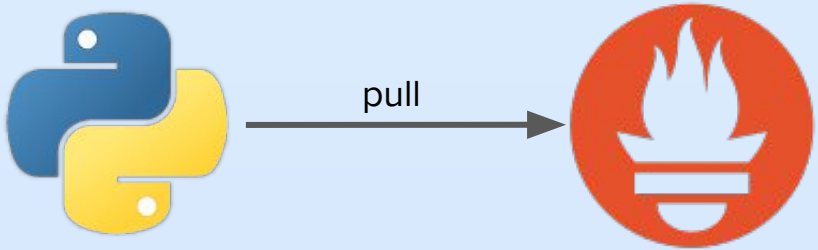
Prometheus / Pandora

```
---
hosts:
  prod-rsyslog-ams2:
    port: 44221
    chef:
      query: fqdn:prod-syslog* AND
region:ams2

relabels:
  -
    regex: |-
      [^\.]+\.([^\.]+)\..*
    replacement: "${1}"
    source_labels:
      - __address__
    target_label: region

scrape_config:
  scrape_interval: 5m
```

DigitalOcean

Prometheus / Pandora

v1:

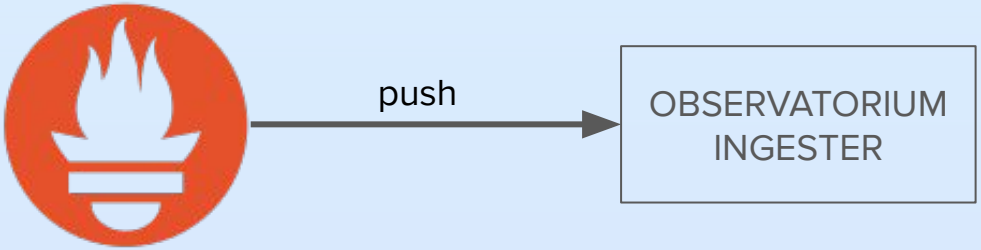


The Observatorium: **Foundations**

SLM | Service Catalog | **Observability Platforms**

Prometheus / Pandora

v2:




```
remote_write:
  - url:
http://observatorium-ingester.internal.digitalocean.com:9190/ingester
    write_relabel_configs:
      - source_labels: [__name__]
        regex: 'sli:.*'
        action: keep
      - source_labels: [observatorium]
        regex: 'sli'
        action: keep
```

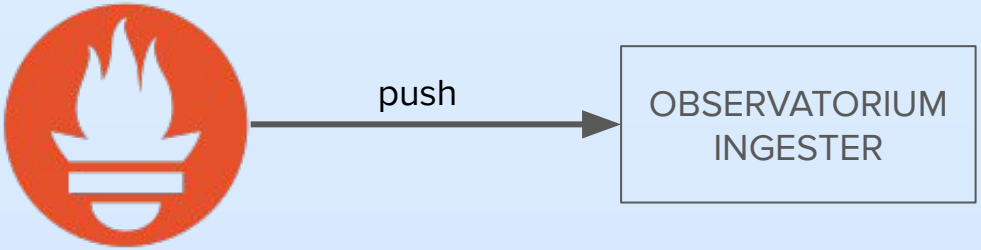
DigitalOcean

The Observatorium: **Foundations**

SLM | Service Catalog | **Observability Platforms**

Prometheus / Pandora

v2:

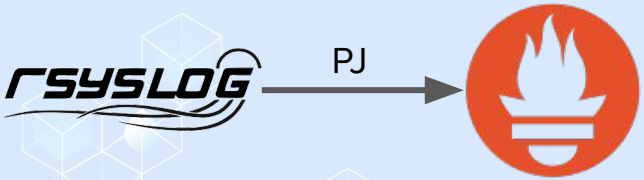



```
remote_write:
  - url:
http://observatorium-ingester.internal.digitalocean.com:9190/ingester
    write_relabel_configs:
      - source_labels: [__name__]
        regex: 'sli:.*'
        action: keep
      - source_labels: [observatorium]
        regex: 'sli'
        action: keep
```

DigitalOcean

Prometheus / Pandora / Polyjuice

```
<190>2019-01-29T19:53:16.450156+00:00 flux-kubernetes03.nyc3.internal.digitalocean.com
polyjuice_flux[1]: @cee: {"response":{"code":201,"time_ms":12}}
```




```
# HELP polyjuice_http_resp_time_ms Polyjuice HTTP response time
(ms)<br>
# TYPE polyjuice_http_resp_time_ms histogram
polyjuice_http_resp_time_ms_bucket{resp_code="201",le="1"} 1
polyjuice_http_resp_time_ms_bucket{resp_code="201",le="4"} 1
polyjuice_http_resp_time_ms_bucket{resp_code="201",le="16"} 1
polyjuice_http_resp_time_ms_bucket{resp_code="201",le="64"} 0
polyjuice_http_resp_time_ms_bucket{resp_code="201",le="256"} 0
polyjuice_http_resp_time_ms_bucket{resp_code="201",le="1024"} 0
polyjuice_http_resp_time_ms_bucket{resp_code="201",le="4096"} 0
polyjuice_http_resp_time_ms_bucket{resp_code="201",le="16384"} 0
polyjuice_http_resp_time_ms_bucket{resp_code="201",le="32768"} 0
polyjuice_http_resp_time_ms_bucket{resp_code="201",le="+Inf"} 0
polyjuice_http_resp_time_ms_sum{resp_code="201"} 12
```

The Observatorium: Motivations

- "I want to know the **current health of the cloud**"
- "I want to see the live health and **historical performance** of all services that relate to Droplet Creation"
- "There's currently an outage. I wonder if any **outages like this one** have occurred before, and if so, how they were fixed."
- "I want to understand the reliability of any/all **customer-facing products over time**"
- "How much of our team's weekly/monthly/annual **error budget** have we depleted as of today?"
- "I want to know if there are **warning signs** around the current performance of my service(s) that will lead to **degradation in the near future**"

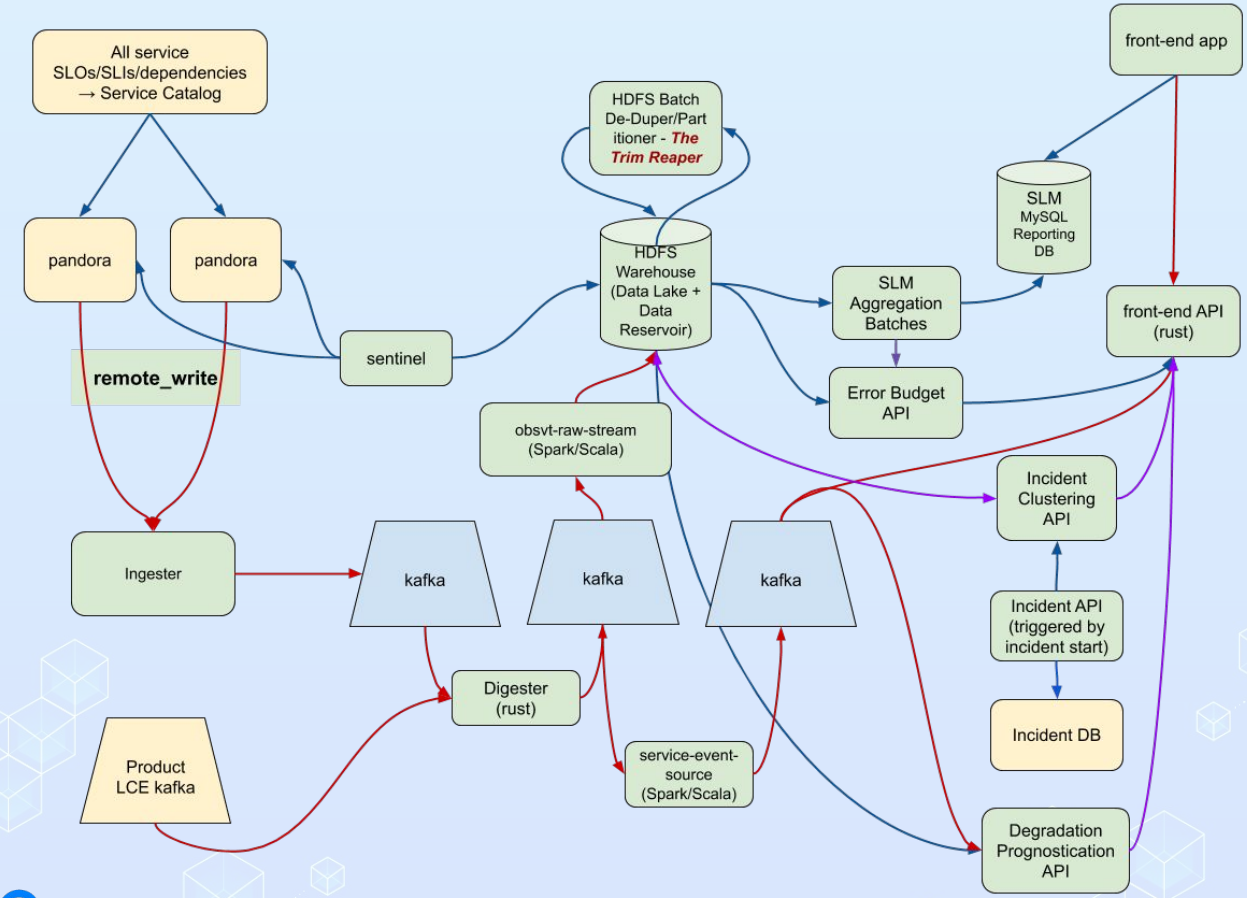
*This is a **data product**, with multiple customer personas*

DigitalOcean

The Observatorium

Putting the pieces together

Putting the pieces together

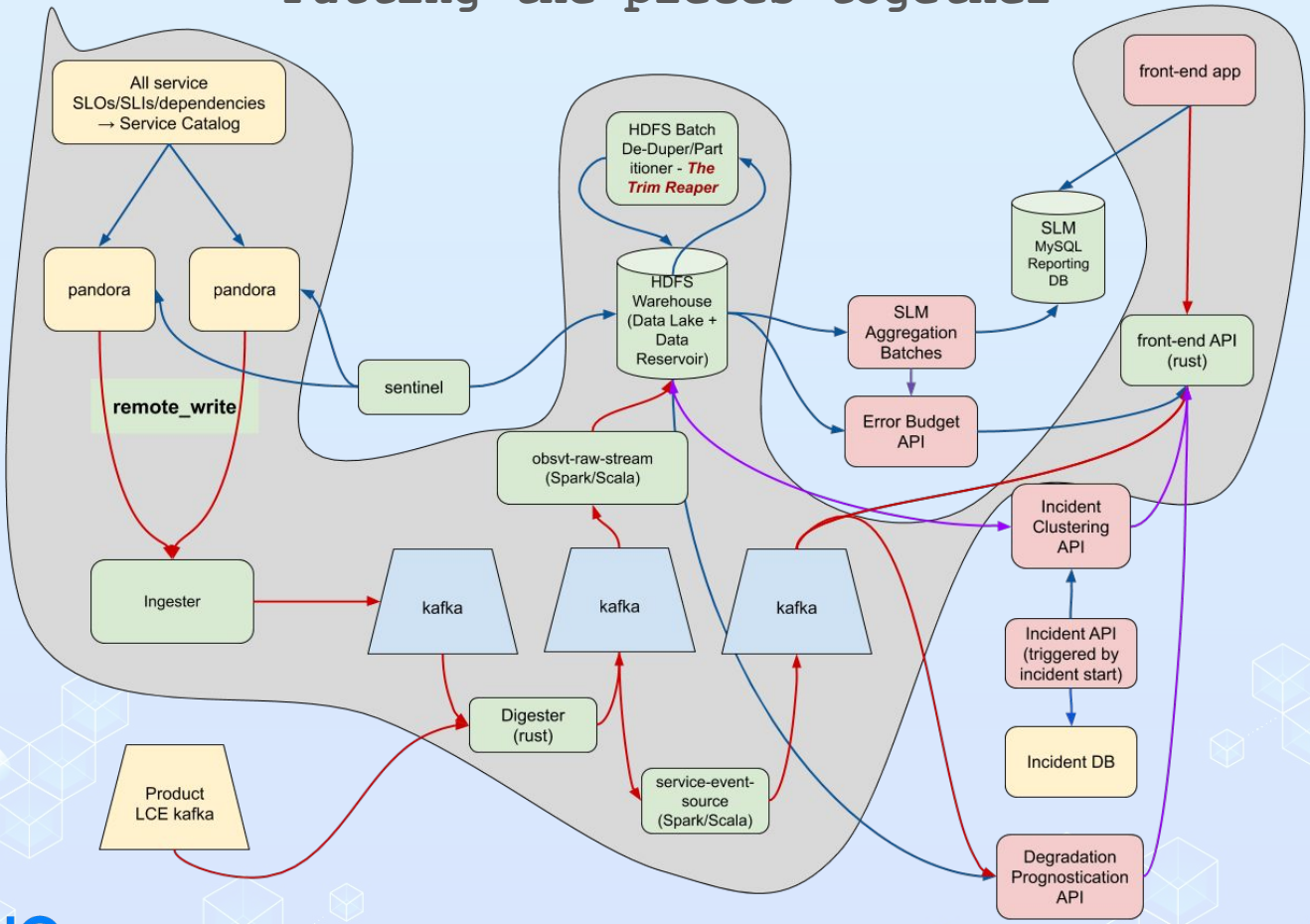


DigitalOcean

Putting the pieces together

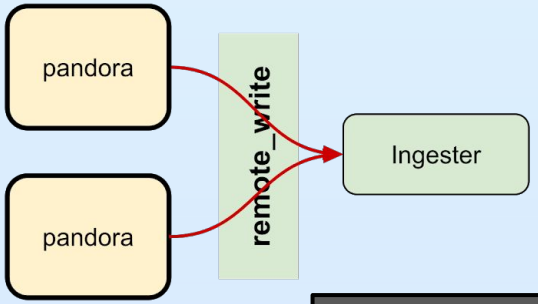
(record scratch sound)

Putting the pieces together



DigitalOcean

Putting the pieces together



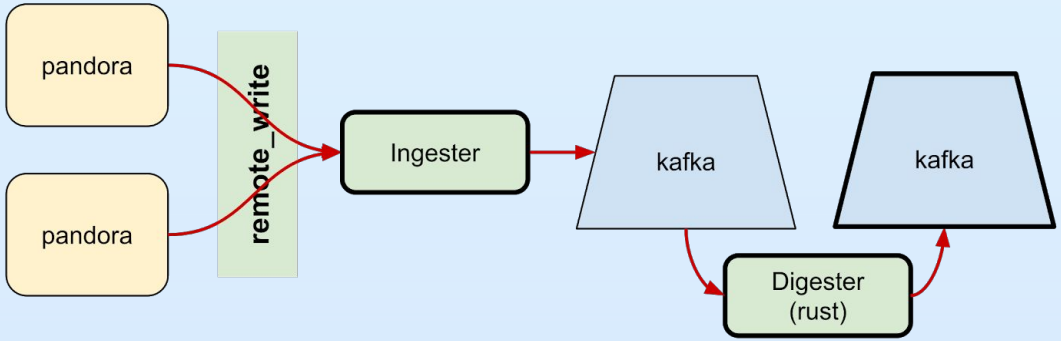

```
recording_rules:
  - record: sli:alpha_write_latency:p99
    expr: |-

histogram_quantile(0.99,sum(rate(mysql_info_schema_write_query_response_time_seconds_bucket{cluster="al
pha"}[5m]))
      by (le))
    labels:
      observatorium: sli
```


```
{"status":"success","data":{"resultType":"vector","result":[{"metric":{"__name__":"sli:alpha_write_latency:p
99","observatorium":"sli"},"value":[1572182521.252,"0.020096308724832153"]}]}}
```


Element	Value
sli:alpha_write_latency:p99{observatorium="sli"}	0.020096308724832153

DigitalOcean

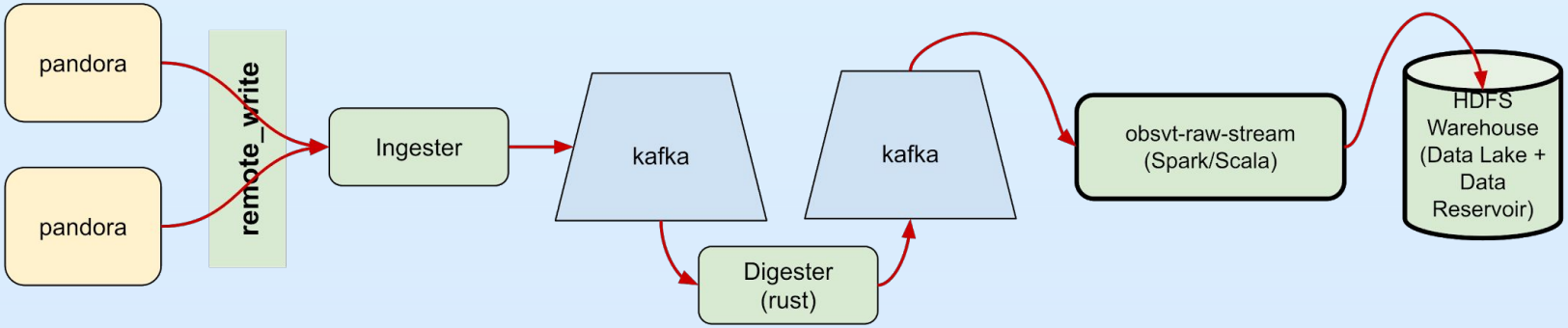
Putting the pieces together




```
labels {key: "observatorium" value: "sli"}
labels {key: "replica" value: "general-2d3a637.fra1"} labels {key: "__name__" value:
"sli:alpha_write_latency:p99"} samples {key: 1572182521.252 value: 0.020096308724832153}
```

DigitalOcean

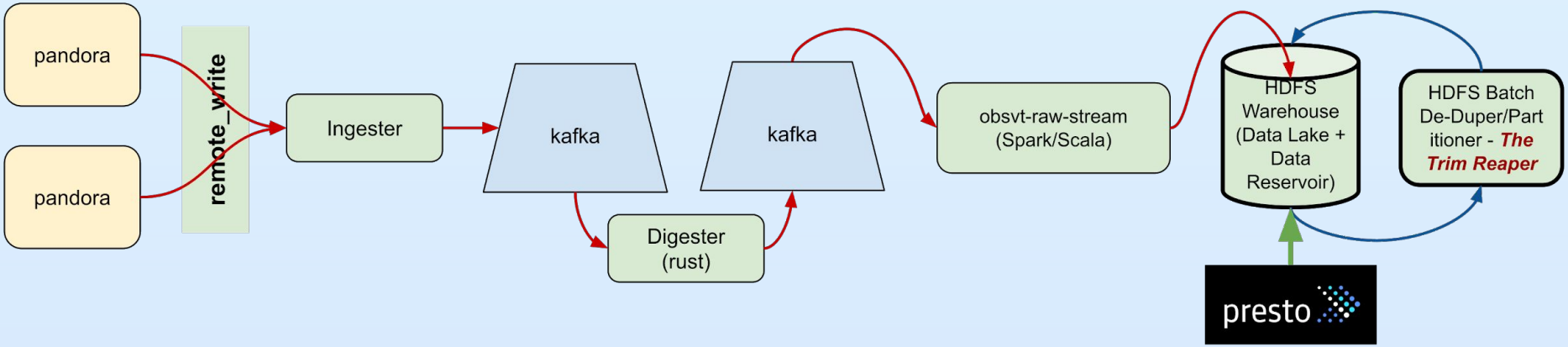
Putting the pieces together




```
Row(
    metric_name='sli:alpha_write_latency:p99',
    time=datetime.datetime(2019, 10, 27, 13, 22, 1, 379000),
    value=0.020096308724832153,
    labels={'replica': 'general-49ae403.nyc3', '__name__': 'sli:alpha_write_latency:p99', 'observatorium':
'sli'},
    meta={}
)
```

DigitalOcean

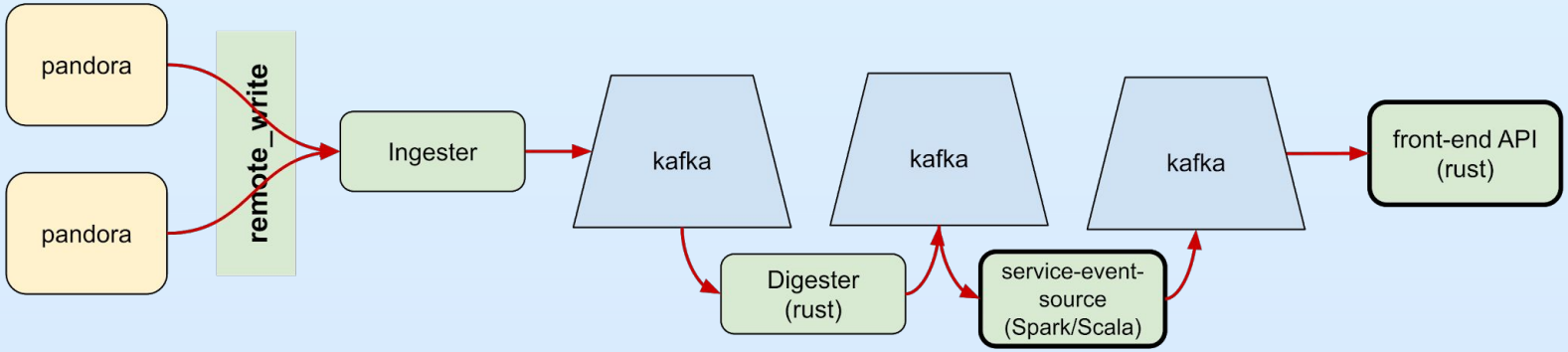
Putting the pieces together




```
select * from hive.observatorium.metrics_data where metric_name = 'sli:alpha_write_latency:p99' limit 1\G

-[ RECORD 1 ]-----------------------------------------------------------------------------------------
time        | 2019-10-27 13:22:01.379
value       | 0.020096308724832153
labels      | {replica=general-2d3a637.fra1, __name__=sli:alpha_write_latency:p99, observatorium=sli}
meta        | {}
metric_name | sli:alpha_write_latency:p99
year        | 2019
month       | 10
day         | 27
hour        | 13
```

DigitalOcean

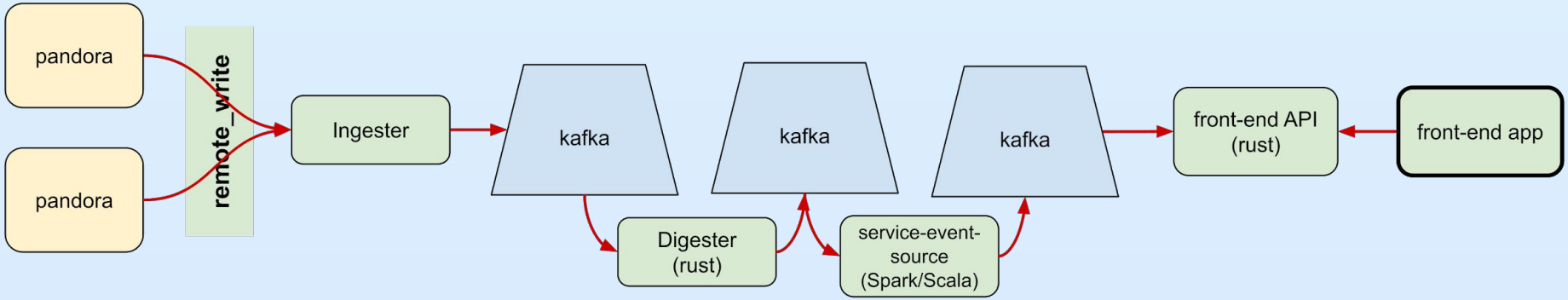
Putting the pieces together

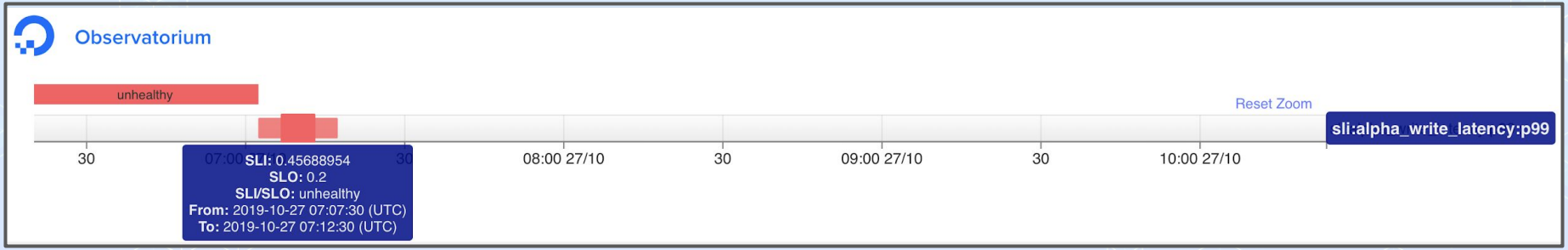



```
|name                       |start              |end                |aggregator|aggregatorLabel|objective|value     |observations|
+---------------------------+-------------------+-------------------+----------+---------------+---------+----------+------------+
|sli:alpha_write_latency:p99|2019-10-27 09:45:00|2019-10-27 09:55:00|null      |null           |0.2      |0.02772143|20          |
```

DigitalOcean

Putting the pieces together





DigitalOcean

Putting the pieces together
Stepping Back

"I want to know the current health of the cloud"

"I want to understand the reliability of any/all customer-facing products over time"

*"I want to see the live health and historical performance of all services that relate to **Droplet Creation**"*

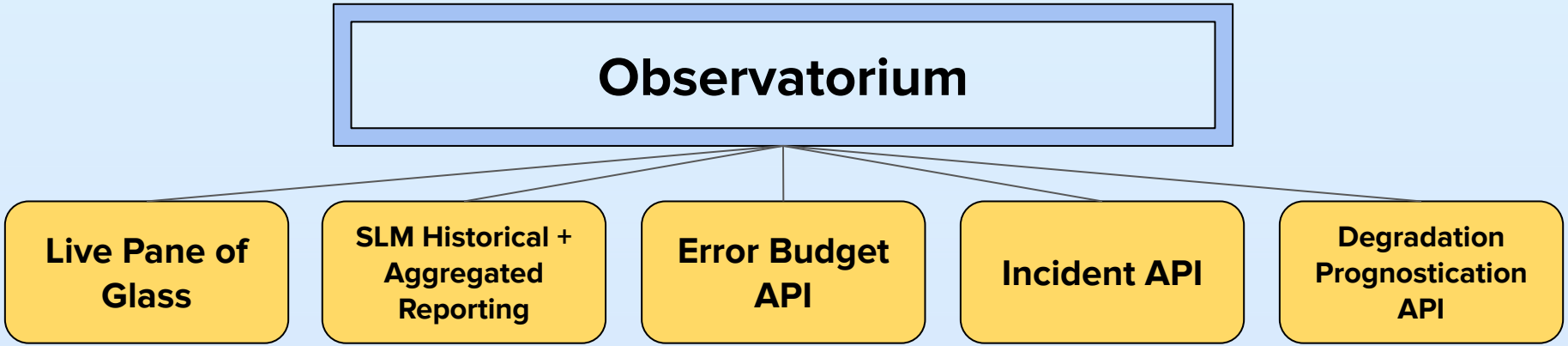
"How much of our team's weekly/monthly/annual error budget have we depleted as of today?"

"There's currently an outage. I wonder if any outages like this one have occurred before, and if so, how they were fixed."

"I want to know if there are warning signs around the current performance of my service(s) that will lead to degradation in the near future.

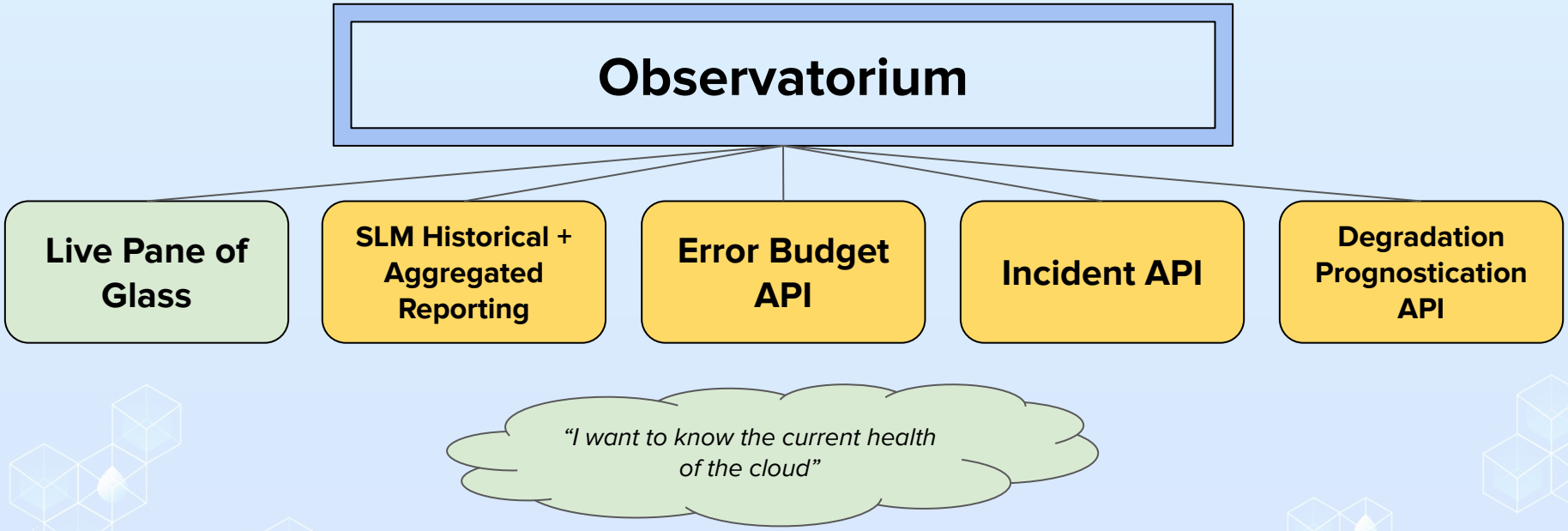
DigitalOcean

Putting the pieces together



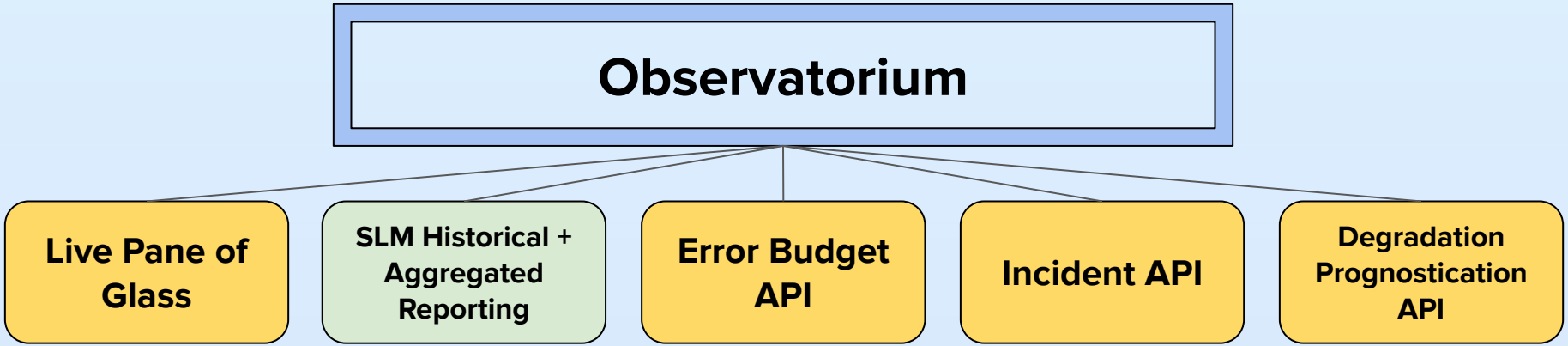
DigitalOcean

Putting the pieces together



DigitalOcean

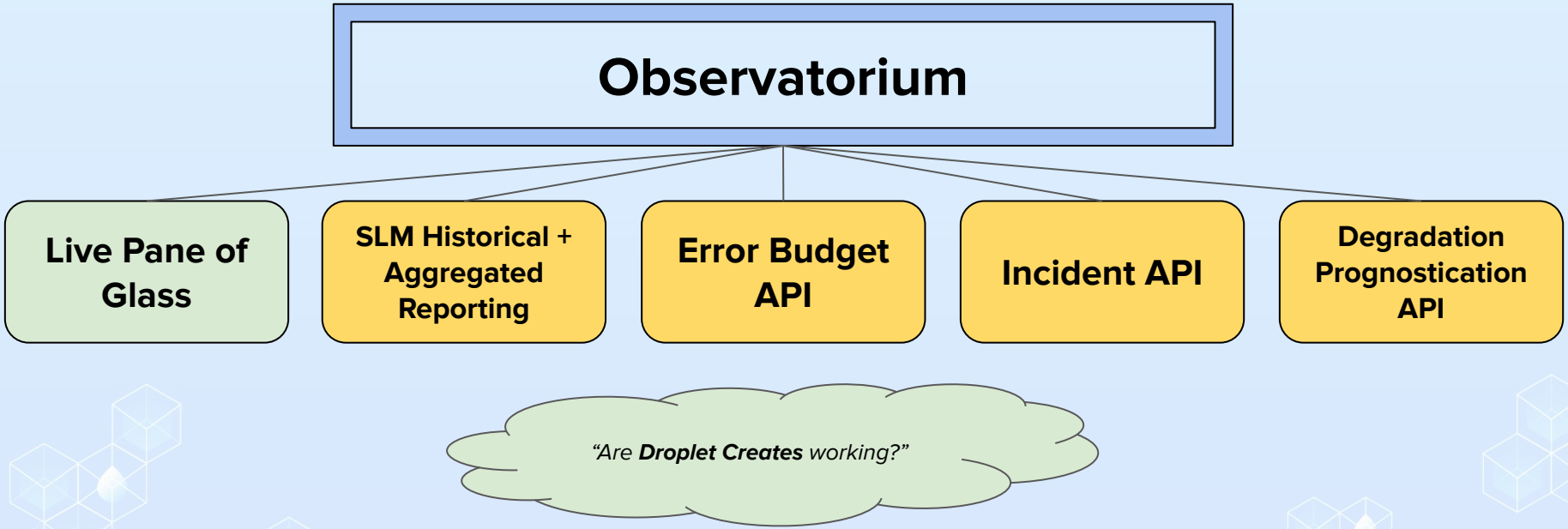
Putting the pieces together



"I want to understand the reliability of any/all customer-facing products over time"

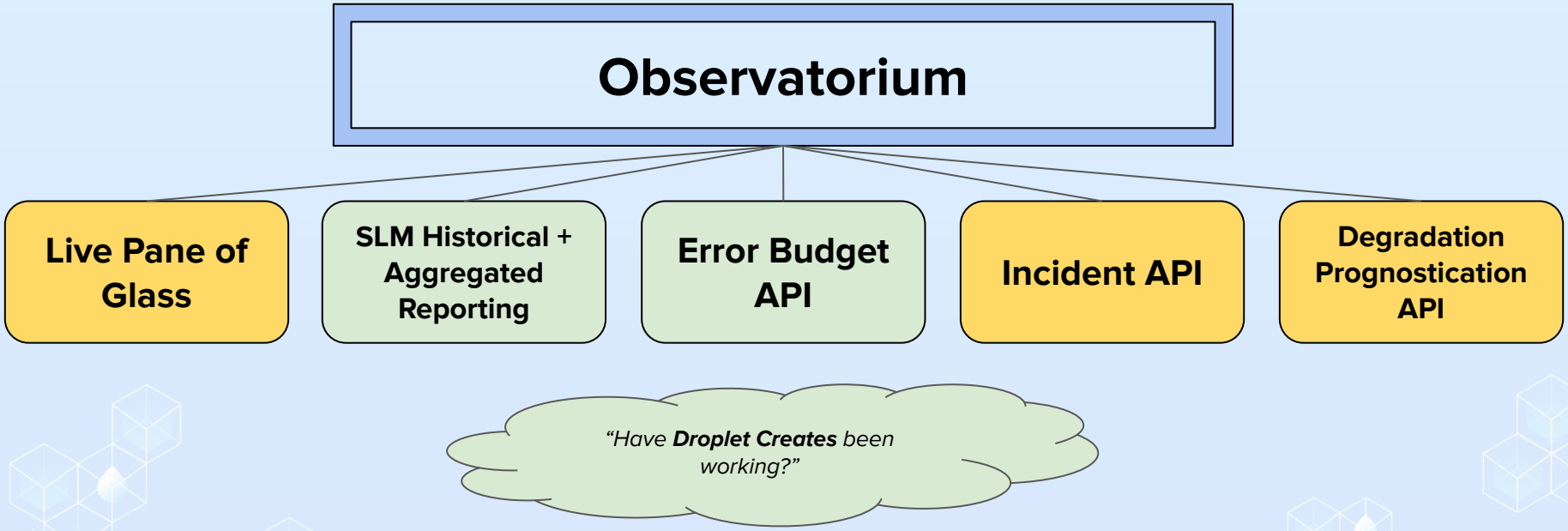
DigitalOcean

Putting the pieces together



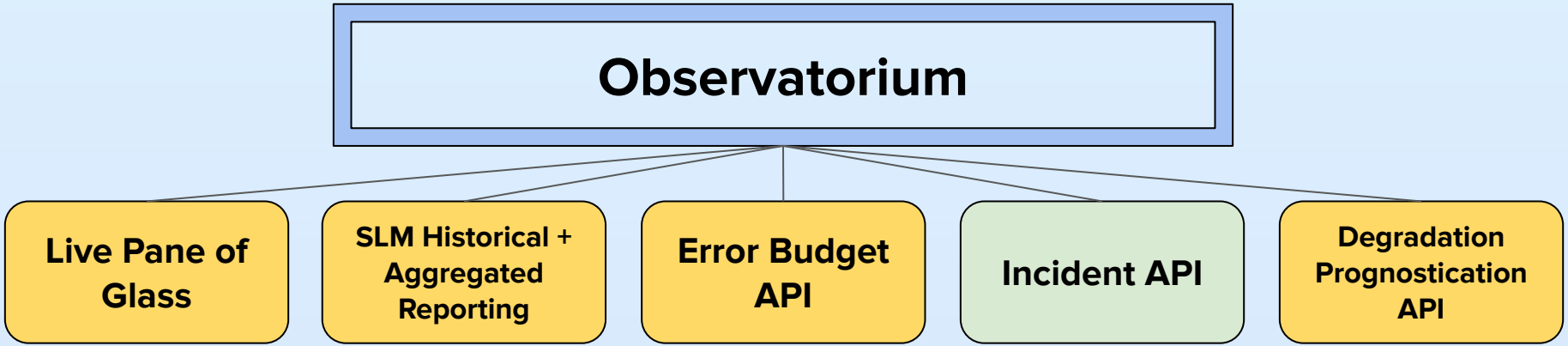
DigitalOcean

Putting the pieces together



DigitalOcean

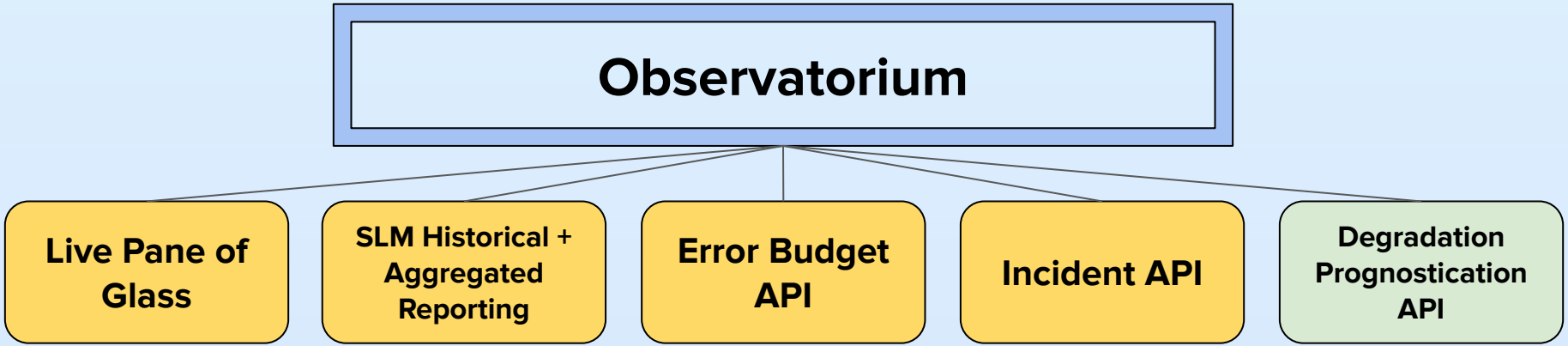
Putting the pieces together



"There's currently an outage. I wonder if any outages like this one have occurred before, and if so, how they were fixed."

DigitalOcean

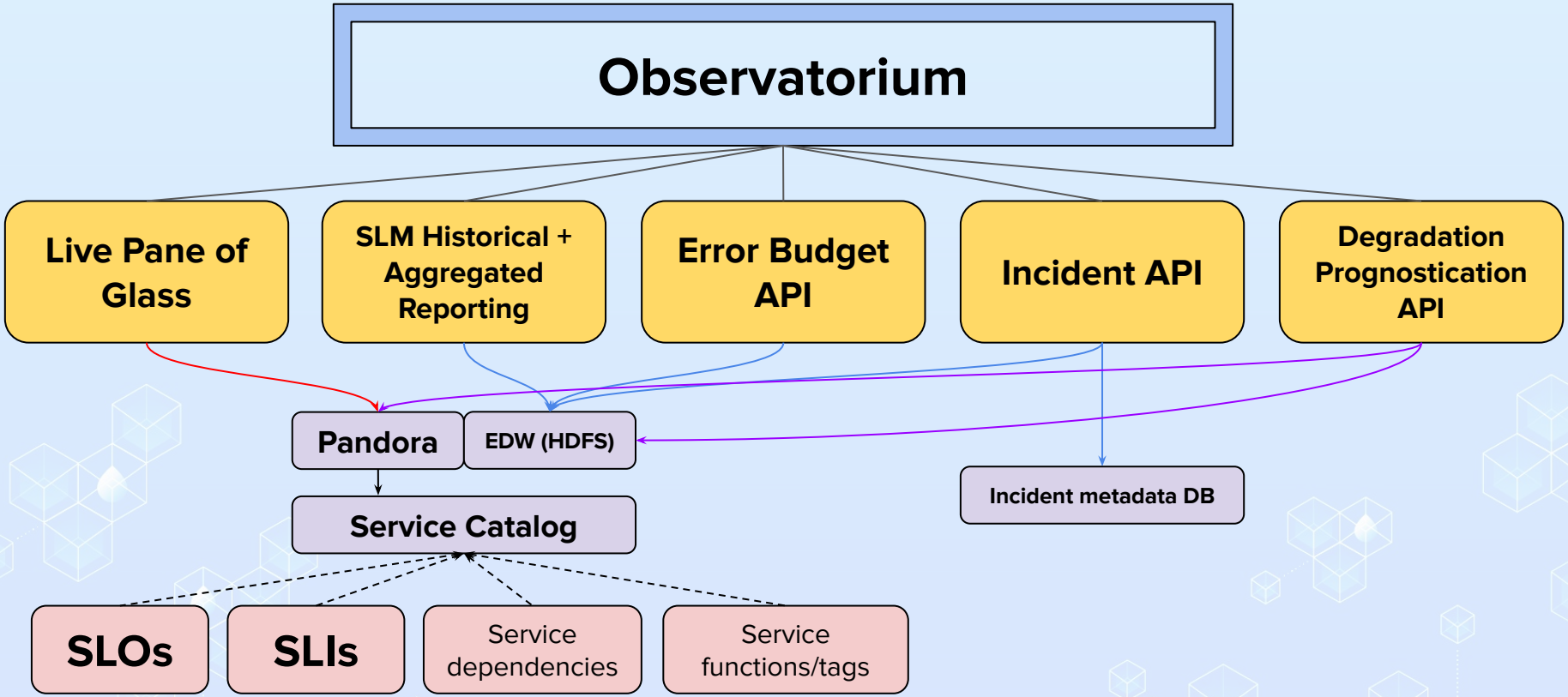
Putting the pieces together



"I want to know if there are warning signs around the current performance of my service(s) that will lead to degradation in the near future.

DigitalOcean

Putting the pieces together



DigitalOcean

Putting the pieces together
UI/API components

Live Pane of Glass

SLM Historical + Aggregated Reporting

product	slo_name	slo_type	region	slo_target	current_month	delta
droplet	live migration	pct	sfo1	0.99	0.99435	0.00435
droplet	create latency	latency	sfo1	55	131.992582	76.992582
droplet	resize duration	latency	sfo1	55	256.306905	201.306905
droplet	uptime (node)	pct	sfo1	0.9999	1	0.0001
droplet	create success rate	pct	sfo1	0.99	0.980314	-0.009686
droplet	resize success rate	pct	sfo1	0.99	0.992327	0.002327

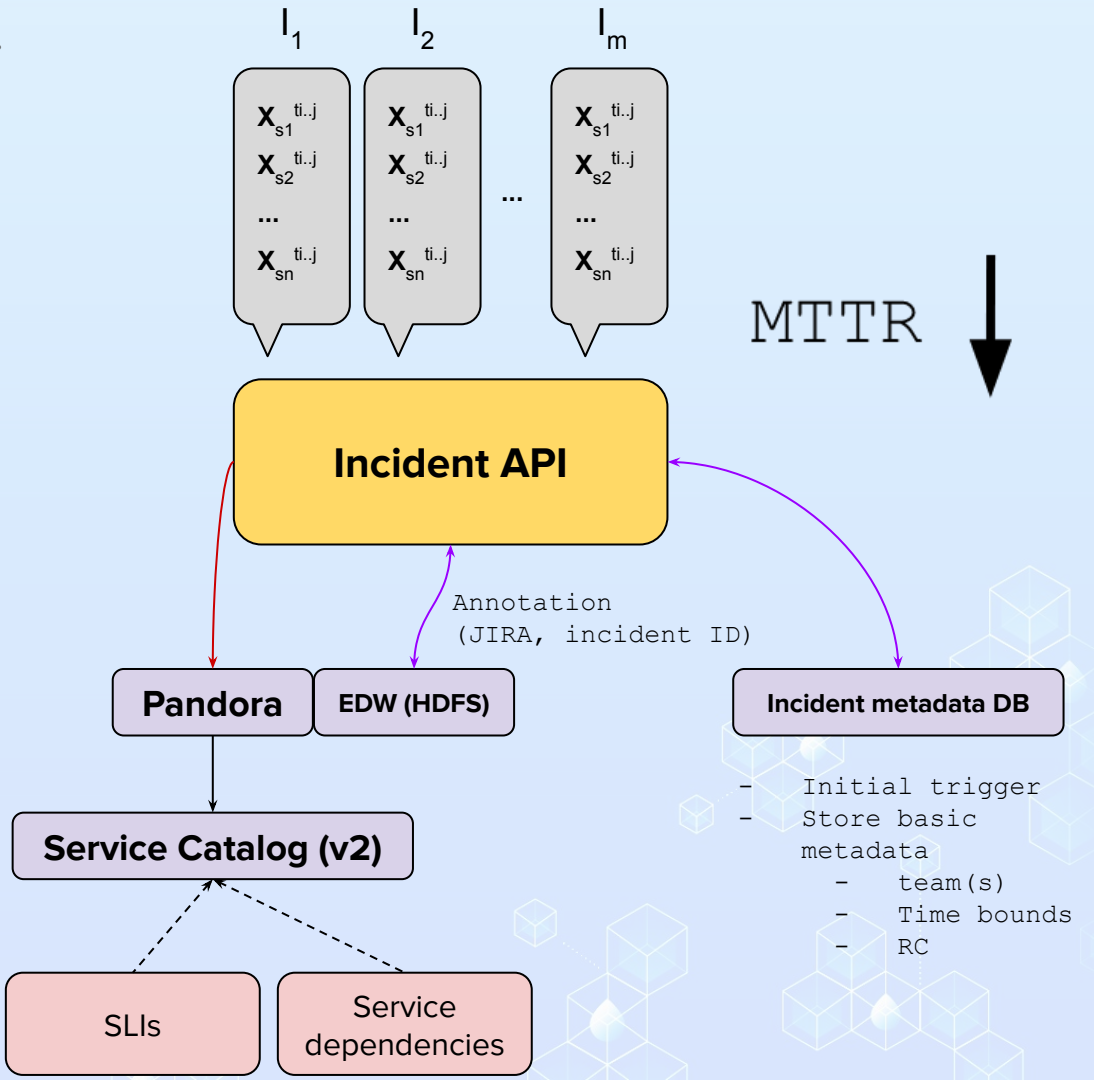
Error Budget API

```
SLO: 99.9% uptime
Monthly allowance: 43.2 minutes
MTD: <n> minutes missed
```

DigitalOcean

Putting the pieces together
Clustering Incidents

1) Incident triggered
2) Annotation begins against all services ➔ EDW
3) Historical records of previous incidents are surfaced
4) Matrices of Service performance vectors are pulled from EDW and compared/clustered
5) Clustering algorithms generate best matching incident(s) given live test data
6) Suggestions surfaced to end user, including metadata
7) After Incident concludes, post-mortem metadata written back to DB

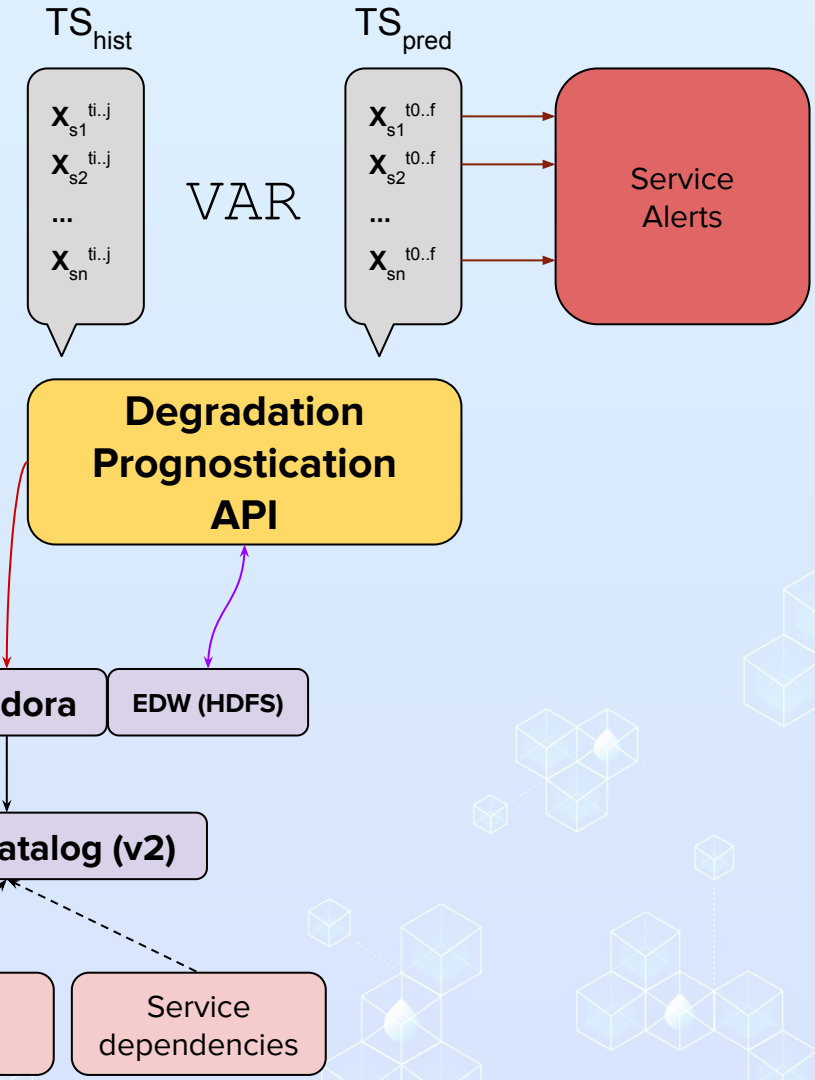


DigitalOcean

Putting the pieces together Forecasting Failures

1) Historical performance/reliability metrics already exist/are warehoused for services and their dependencies
2) Vector AutoRegressive models batched/refreshed regularly
3) Forecasts predicting degradation with enough significance enter the Alerting Protocol
4) Warnings/Messaging arrive to the owner teams before service drops too low

Overall Incident Count



DigitalOcean

2020 Vision

Adoption | Expansion | Impact

DigitalOcean

2020 Vision

Adoption | Expansion | Impact

- Service Catalog as Gatekeeper:
  - “If you don’t comply, you can’t deploy”™
- Bringing ML into the broader data product toolkit/lexicon across the org
- New product SLAs to be predicated on official SLM data
- Telemetry to reveal who uses the product and how often
- Reliability measured in staging/pre-prod environments before deploying to production

DigitalOcean

2020 Vision

Adoption | **Expansion** | Impact

- All services have SLOs and SLIs no matter their proximity to customers
- Error budgets available ad hoc for any historical time period
- Source metric format expands to include non-Pandora data
  - Kafka streams
  - RDBMS
  - NoSQL
- Integration with production/staging Deployment Tracking

DigitalOcean

2020 Vision
Adoption | Expansion | **Impact**

- Fewer customer tickets/complaints about reliability
- Teams iterate on their SLOs and work to reduce outage counts/overall time running degraded services
- More mature pattern recognition among microservices leads to better cross-team developmental collaboration and more cohesive architecture
- Significant reduction of MTTR

DigitalOcean

Q/A

DigitalOcean

Thank you!