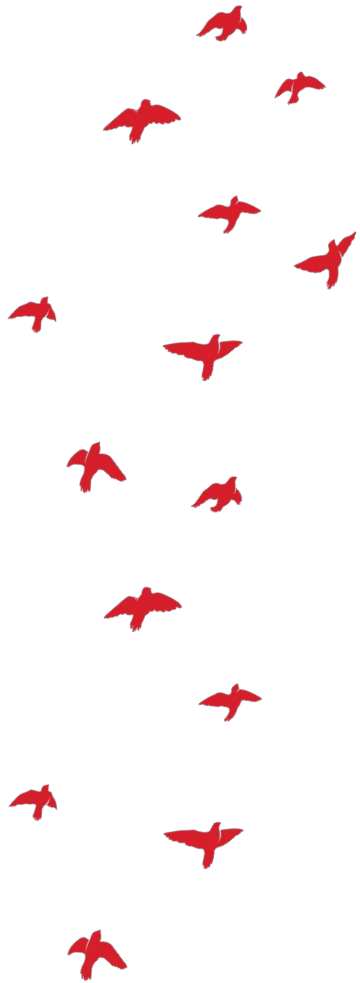
Show Me The Money!

METRICS EDITION!

Tracking your runtime cost using observability!



red canary
a zscaler company

Who am I?

Brian Davis

PRINCIPAL SOFTWARE ARCHITECT

RED CANARY, a Zscaler Company

@brianthedavis

- Member of the
  Red Canary Architecture team
- I've been building complex distributed systems for more than two decades
- Avid hiker,
  photographer,
  home automator,
  and beekeeper based in Denver

Your new job....

You're the new lead architect at a media analysis company...

Chonkify Industries: We Chonkify Your Media Into Insights!

Product	**Document Extract**
Billing Unit	Pages processed
Cost	**$**$$$: text extraction, sentiment, keyword tagging

Your new job....

You're the new lead architect at...

Chonkify Industries: We Chonkify Your Media Into Insights!

Product	**Document Extract**	**Image Intelligence**
Billing Unit	Pages processed	Images processed
Cost	**$**$$$: text extraction, sentiment, keyword tagging	**$$**$$: resize, classify, OCR, brand recognition

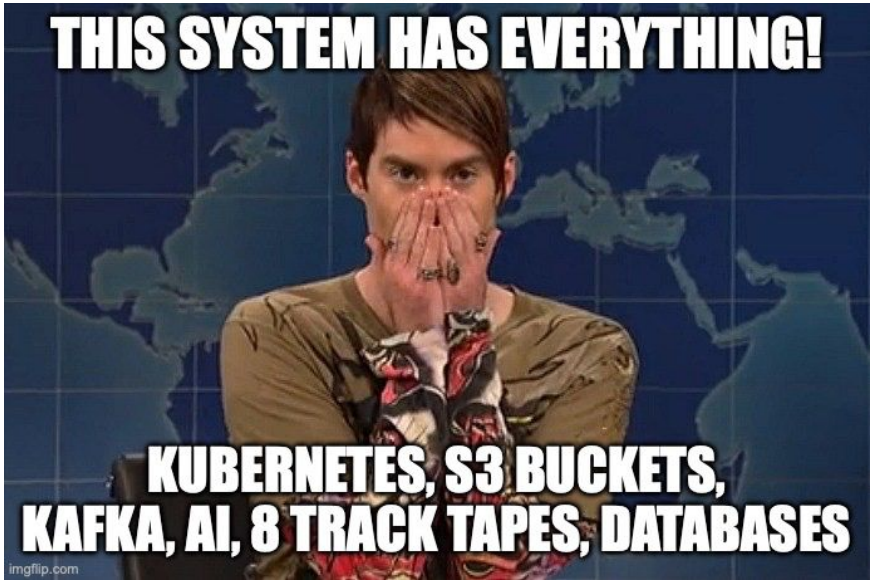
Your new job....

You're the new lead architect at...

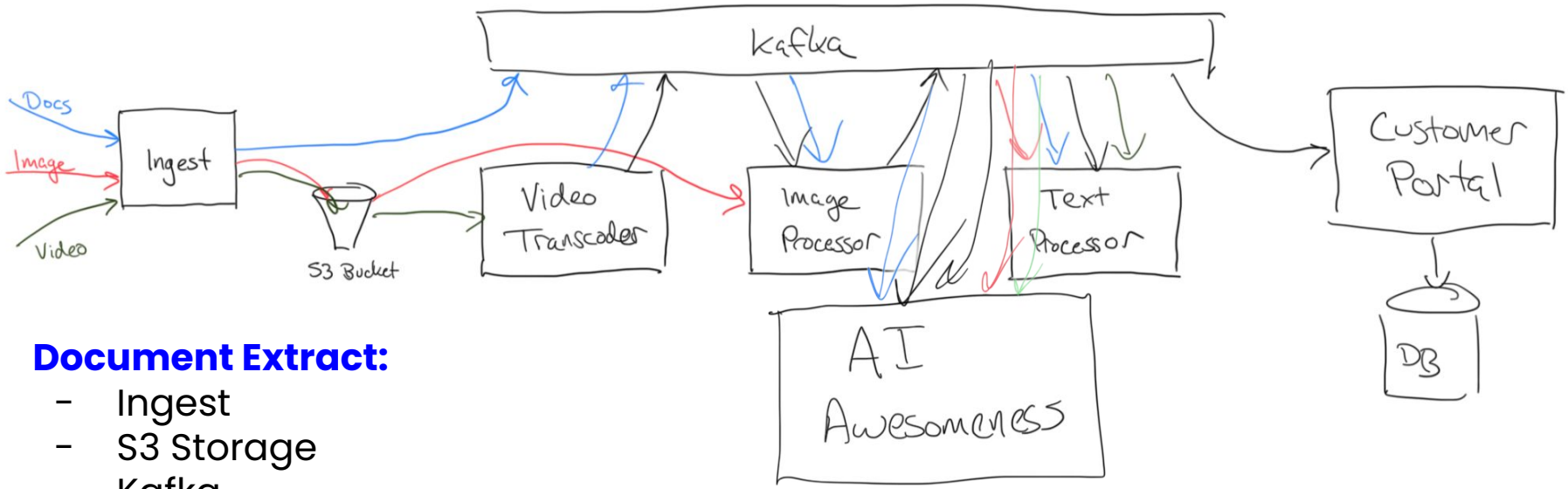
Chonkify Industries: We Chonkify Your Media Into Insights!

Product	**Document Extract**	**Image Intelligence**	**Video Insights**
Billing Unit	Pages processed	Images processed	Minutes of video processed
Cost	**$**$$$: text extraction, sentiment, keyword tagging	**$$**$$: resize, classify, OCR, brand recognition	**$$$$**: transcoding, frame analysis, ML inference

The Chonkifier 9000™



The Chonkifier 9000™



Document Extract:

- Ingest
- S3 Storage
- Kafka
- Text Processor Kubernetes Component
- Customer Portal Compute
- Customer Portal Database

The Chonkifier 9000™

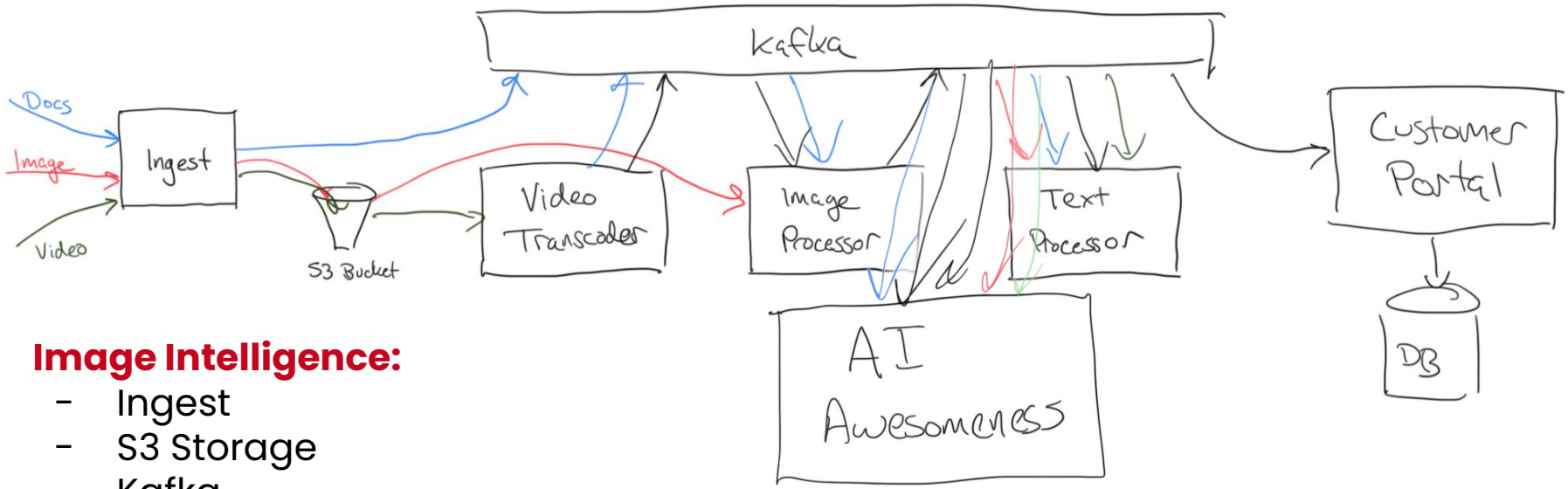
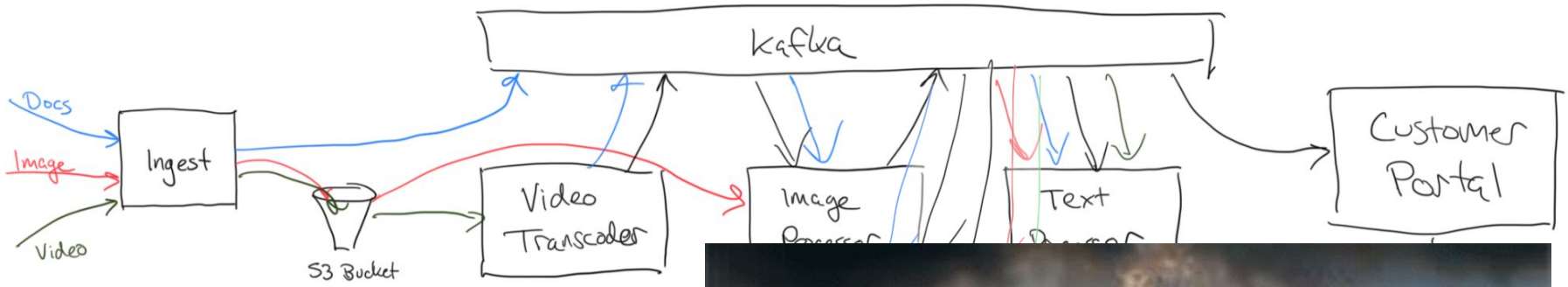


Image Intelligence:

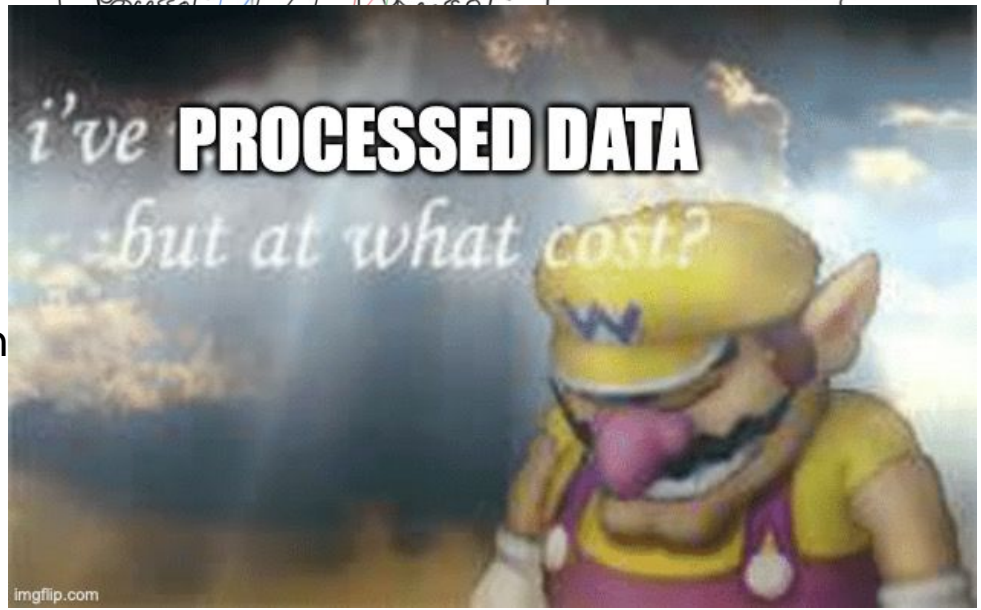
- Ingest
- S3 Storage
- Kafka
- Text Processor Kubernetes Component
- Customer Portal Compute
- Customer Portal Database
- **Image Processor**

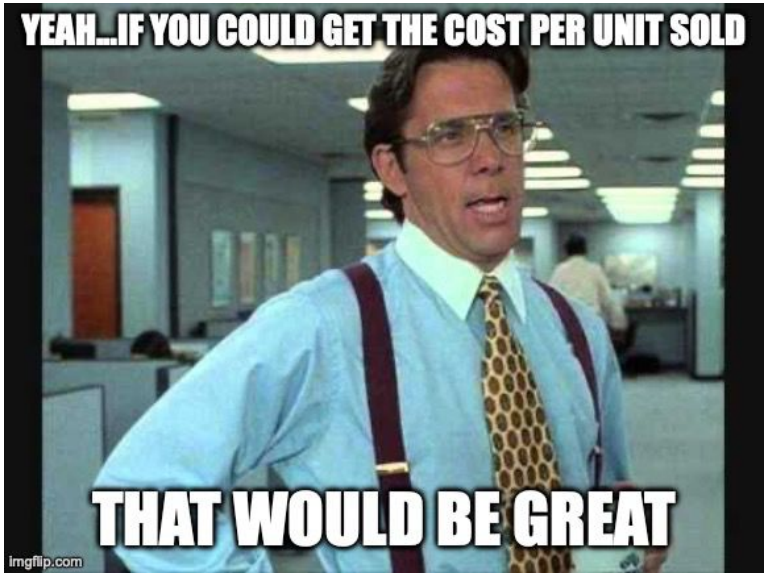
The Chonkifier 9000™



Video Insights:

- Ingest
- S3 Storage
- Kafka
- Text Processor Kubernetes Com
- Customer Portal Compute
- Customer Portal Database
- **Image Processor**
- **Video Transcoder**

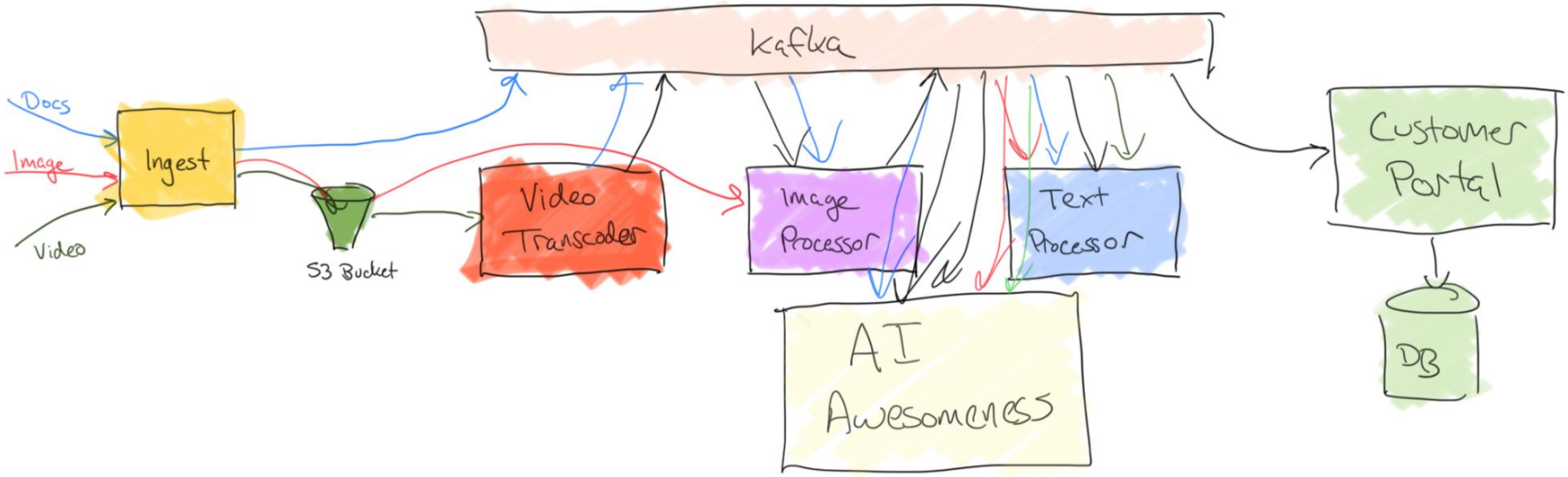




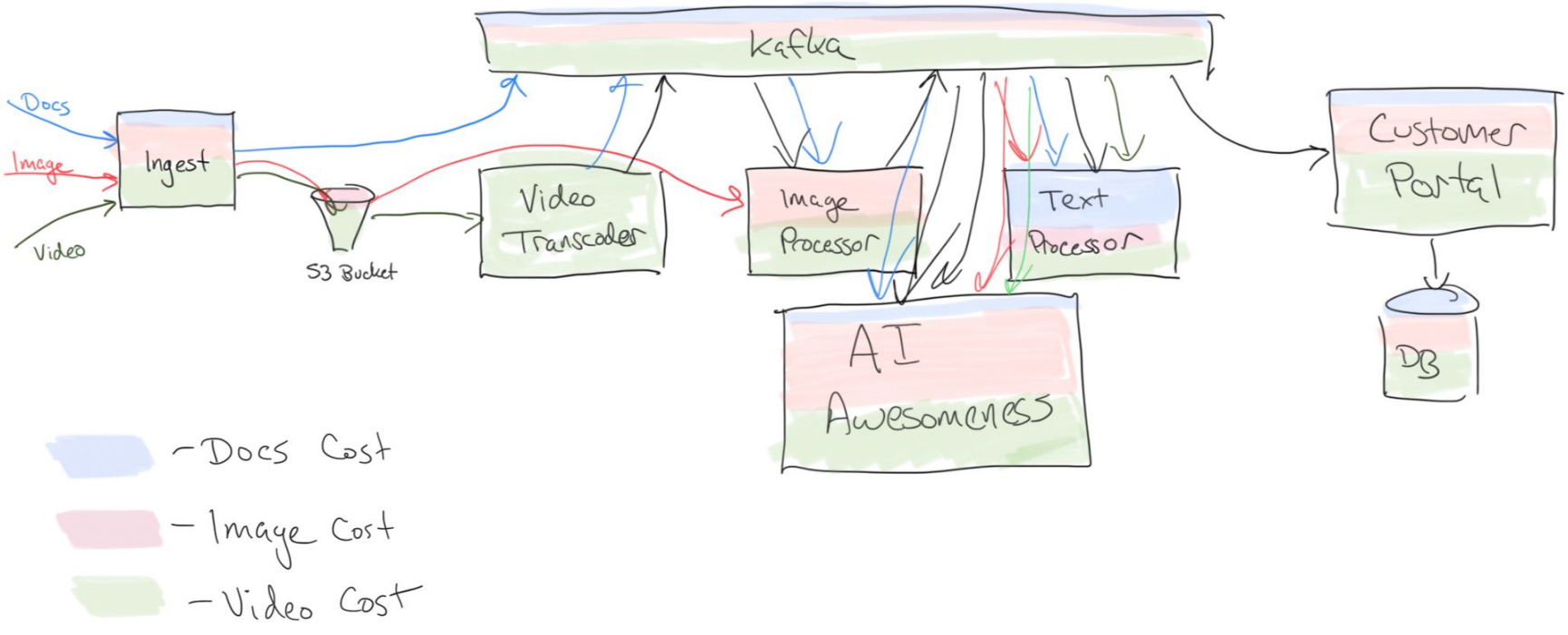
Your finance team wants to get:

- Cost per minute of video processed
- Cost per image processed
- Cost per page processed

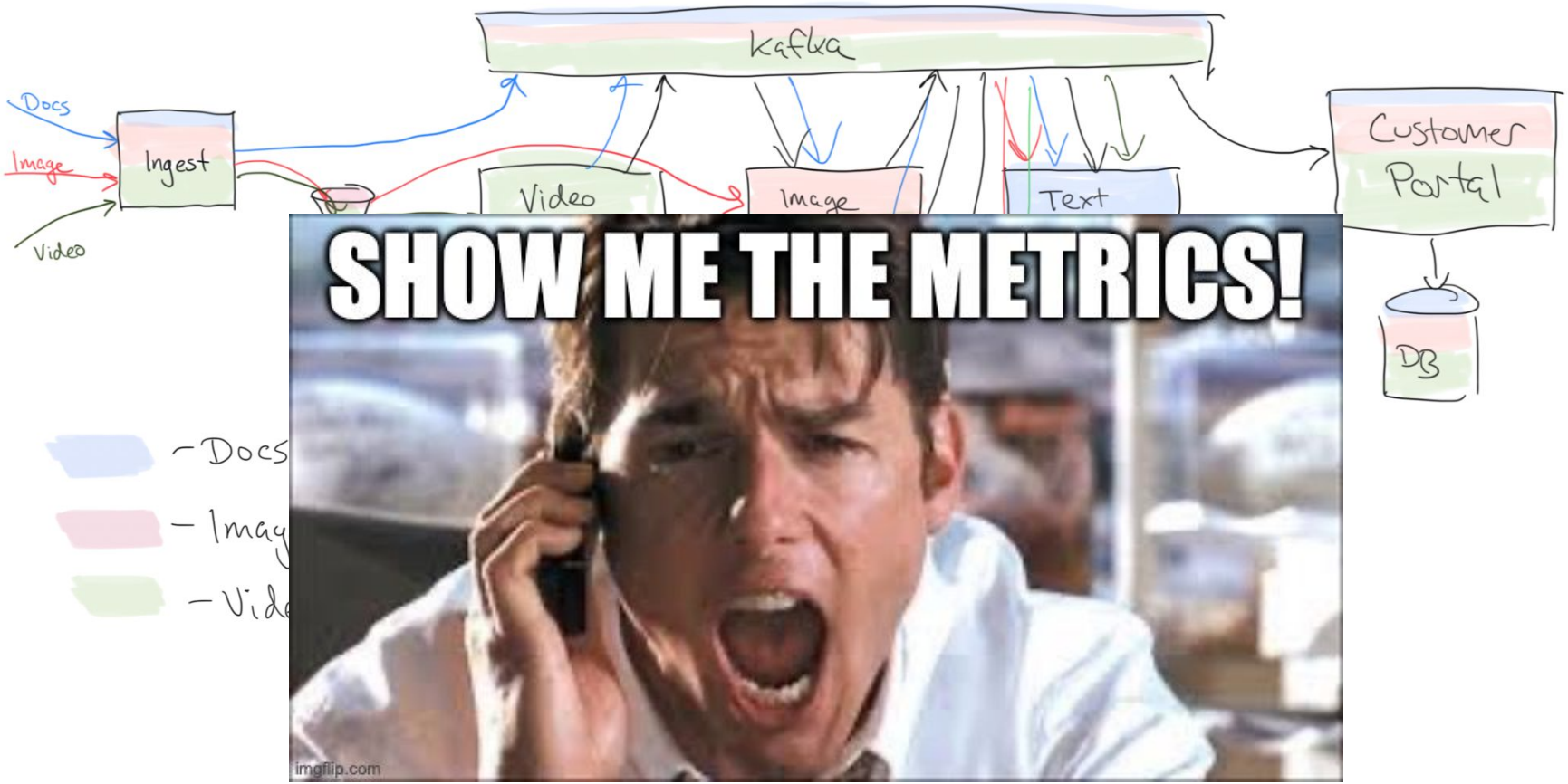
The problem...

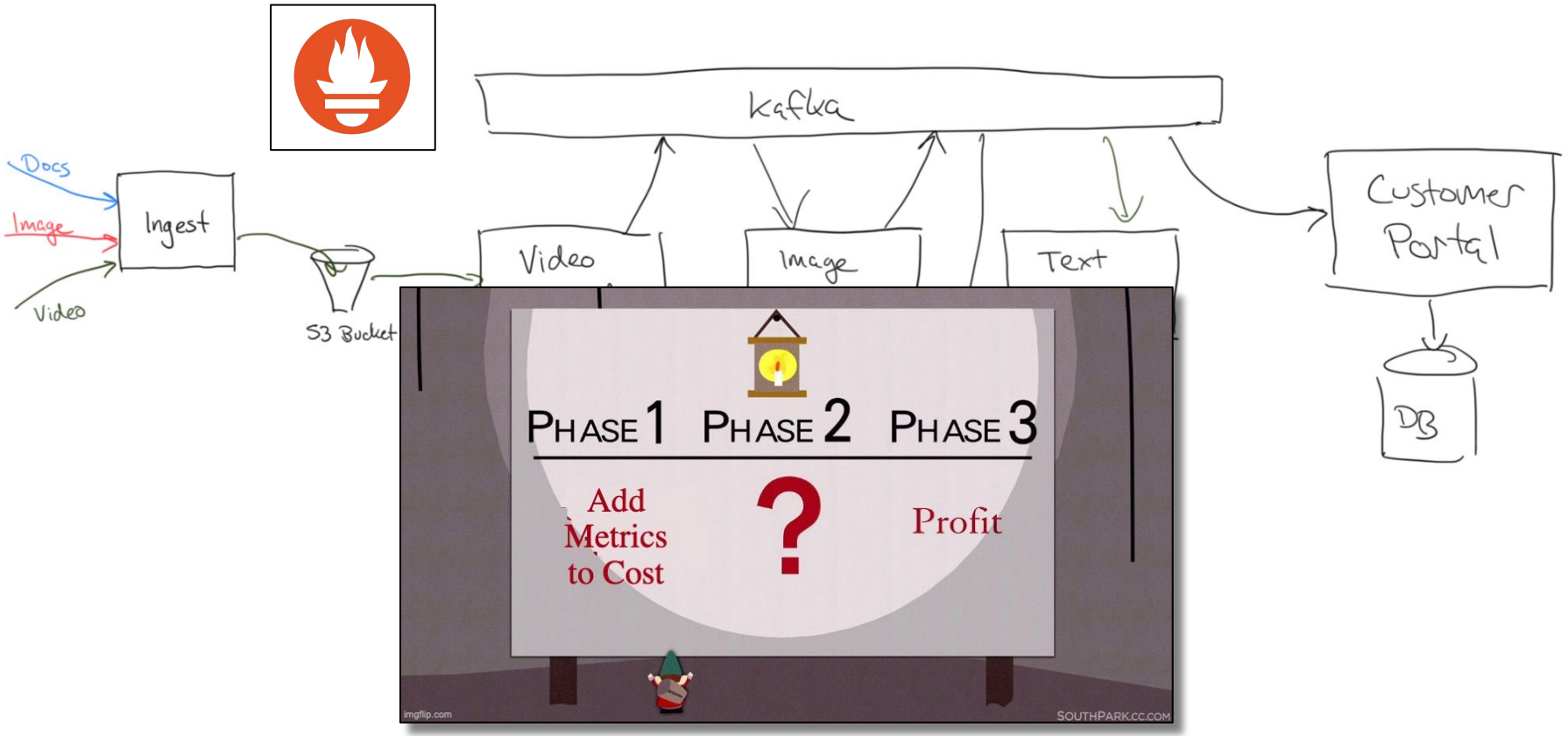

kafka
Docs
Image
Video
Ingest
S3 Bucket
Video Transcoder
Image Processor
Text Processor
AI Awesomeness
Customer Portal
DB

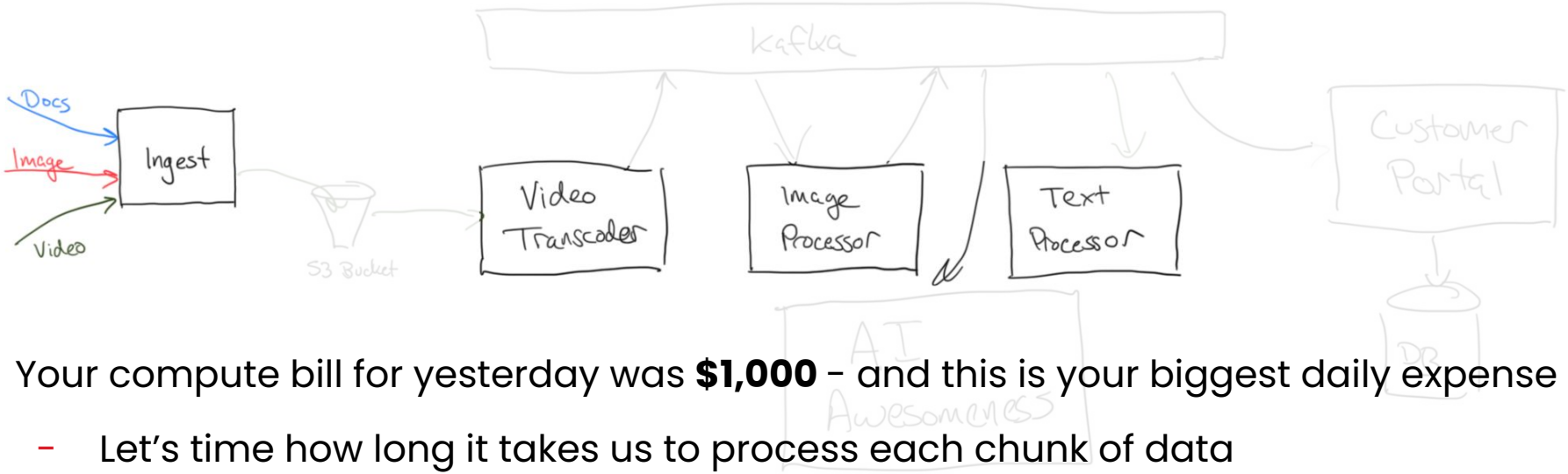
What you REALLY want...



What you REALLY want...



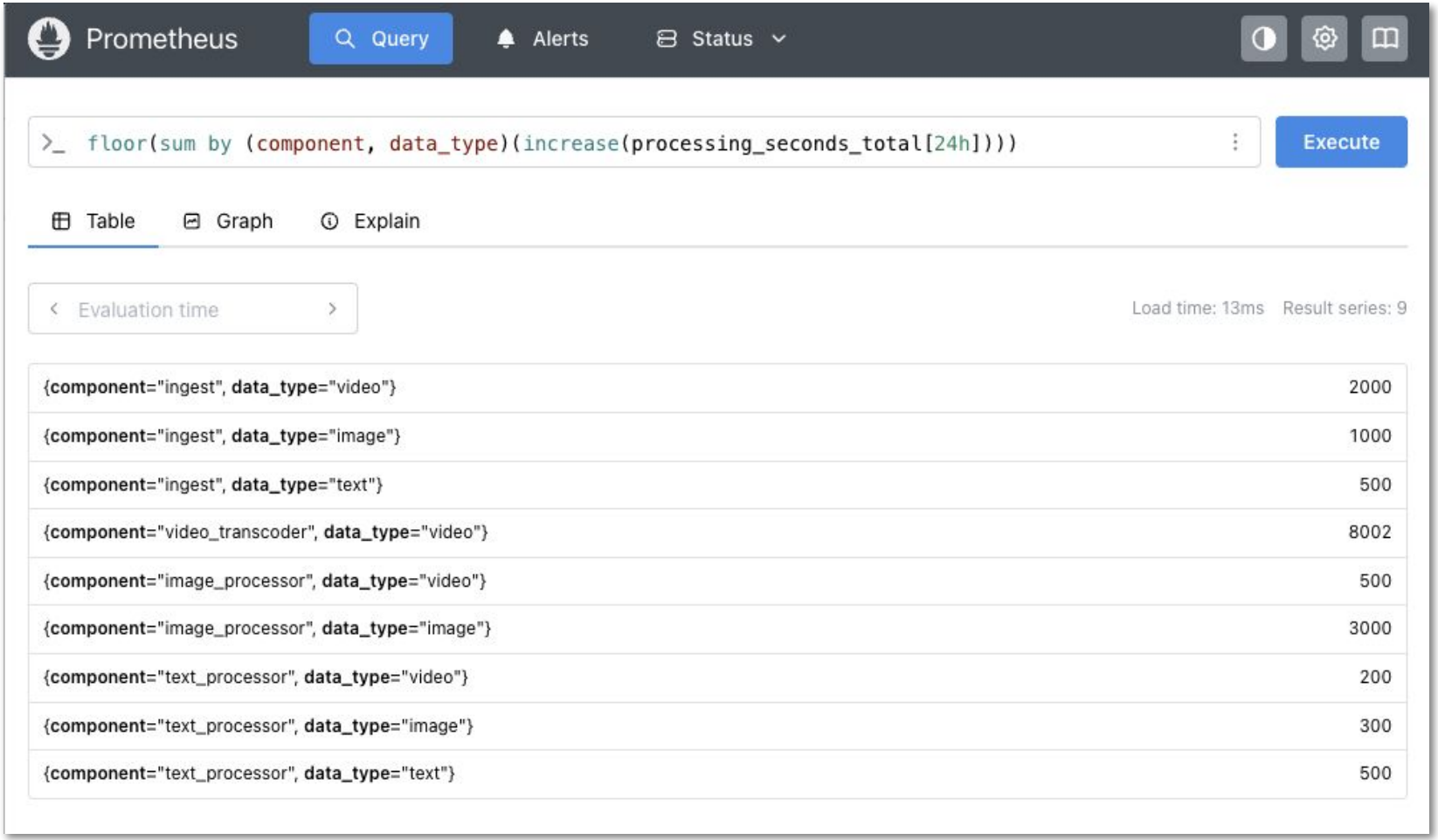

Docs
Image
Video
Ingest
S3 Bucket
kafka
Video
Image
Text
Customer Portal
DB
PHASE 1
PHASE 2
PHASE 3
Add Metrics to Cost
?
Profit
imgflip.com
SOUTHPARK.CC.COM



Your compute bill for yesterday was **$1,000** - and this is your biggest daily expense

- Let's time how long it takes us to process each chunk of data
- Measure and emit a metric (in every component) for each piece of work done:

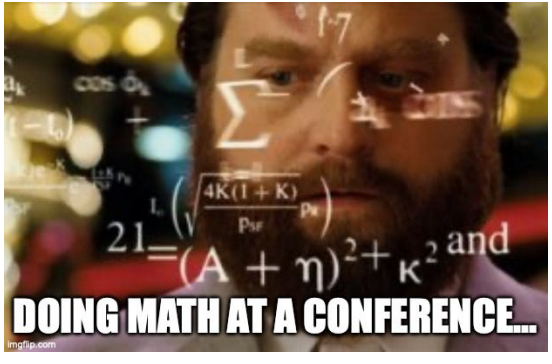
```
processing_seconds_total{
    component=<component_name>,
    data_type=<type>,
    tenant=<tenant_id>
}
```


Prometheus
Query
Alerts
Status
floor(sum by (component, data_type)(increase(processing_seconds_total[24h])))
Execute
Table
Graph
Explain
Evaluation time
Load time: 13ms
Result series: 9
{component="ingest", data_type="video"} 2000
{component="ingest", data_type="image"} 1000
{component="ingest", data_type="text"} 500
{component="video_transcoder", data_type="video"} 8002
{component="image_processor", data_type="video"} 500
{component="image_processor", data_type="image"} 3000
{component="text_processor", data_type="video"} 200
{component="text_processor", data_type="image"} 300
{component="text_processor", data_type="text"} 500

Compute Costs

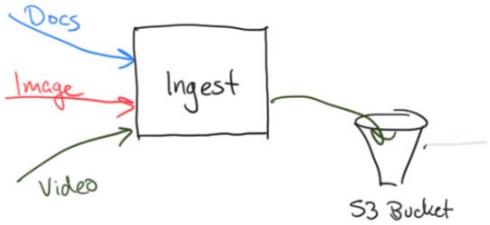
Total compute bill: $1,000

Data Type	Seconds	% of Total Sec	Allocated Cost
Video	10,700s	66.88%	**$668.75**
Image	4,300s	26.87%	**$268.75**
Text	1,000s	6.25%	**$62.50**



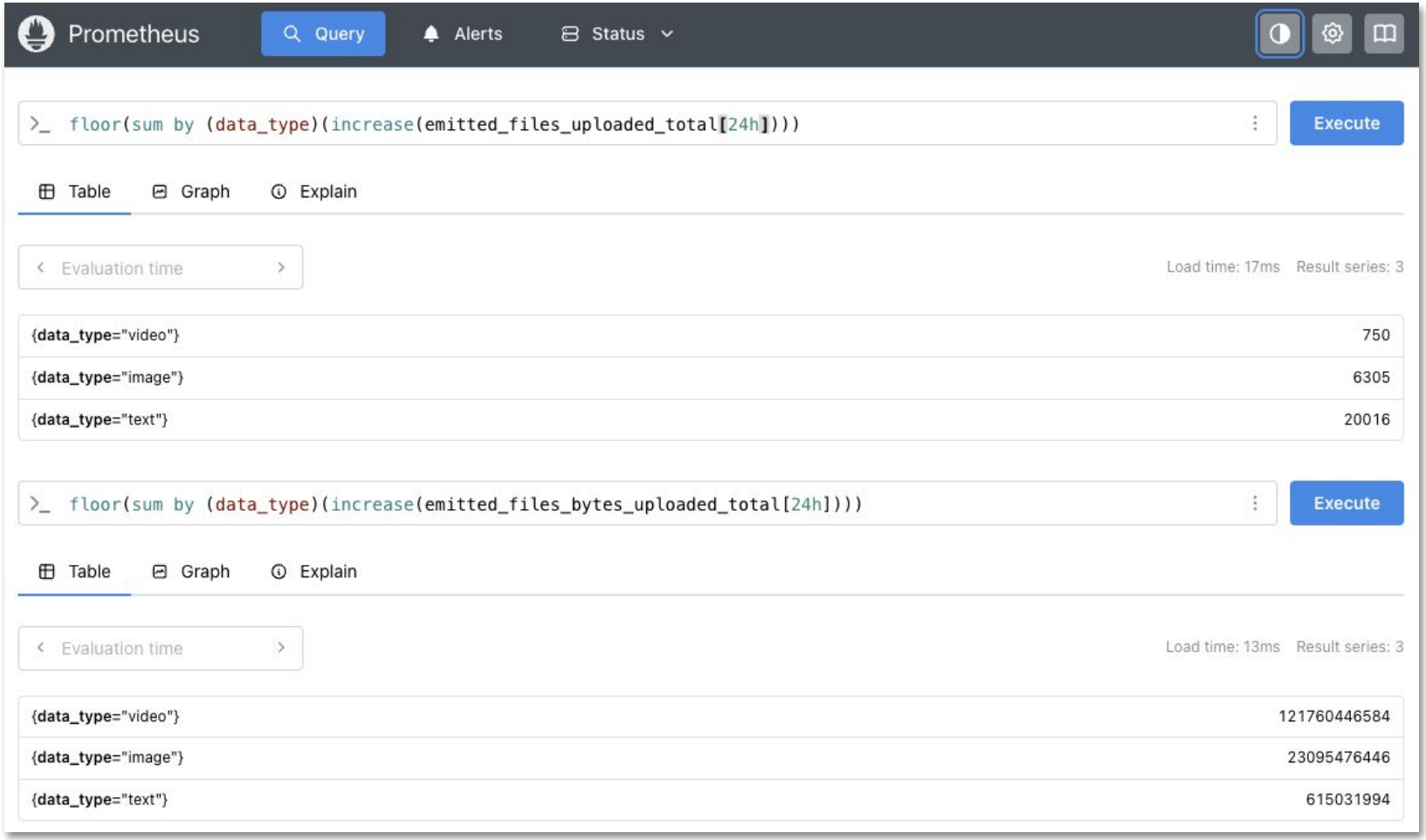
Your S3 bill yesterday for this bucket was **$80**

- …let's track how much we write to S3!




```
emitted_files_uploaded_total{
     component=<component_name>,
     data_type=<type>,
     tenant=<tenant_id>,
     bucket=<bucket_name>
}

emitted_files_bytes_uploaded_total{
     component=<component_name>,
     data_type=<type>,
     tenant=<tenant_id>,
     bucket=<bucket_name>
}
```


Prometheus
Query
Alerts
Status
floor(sum by (data_type)(increase(emitted_files_uploaded_total[24h])))
Execute
Table
Graph
Explain
Evaluation time
Load time: 17ms
Result series: 3
{data_type="video"} 750
{data_type="image"} 6305
{data_type="text"} 20016
floor(sum by (data_type)(increase(emitted_files_bytes_uploaded_total[24h])))
Execute
Table
Graph
Explain
Evaluation time
Load time: 13ms
Result series: 3
{data_type="video"} 121760446584
{data_type="image"} 23095476446
{data_type="text"} 615031994

Storage Costs

red canary
a zscaler company

Total storage bill: $80

S3 Cost	Daily Cost	Attribute Metric
Storage	$60	Bytes uploaded
API requests	$20	Files uploaded

Data Type	Bytes	Share	x $60
Video	113.4 GB	83.7%	$50.22
Image	21.5 GB	15.9%	$9.54
Text	573 MB	0.4%	$0.24
Total	**135.5 GB**		

Data Type	Files	Share	x $20
Video	750	2.8%	$0.55
Image	6,305	23.3%	$4.66
Text	20,016	73.9%	$14.79
Total	**27,071**		

Allocated Costs…so far

red canary
a zscaler company

Data Type	Compute	Storage	Total
Video	$668.75	$50.77	**$719.52**
Image	$268.75	$14.20	**$282.95**
Text	$62.50	$15.03	**$77.53**

But what about the database? And AI? And other stuff?

- You can add metrics to track usage of each of these OR…

Extrapolate some more…

Other costs: **$560**

Data Type	Compute	Storage	Total	% of Known Cost	x $560	Total
Video	$668.75	$50.77	**$719.52**	66.6%	$373.08	**$1,092.60**
Image	$268.75	$14.20	**$282.95**	26.2%	$146.71	**$429.66**
Text	$62.50	$15.03	**$77.53**	7.2%	$40.20	**$117.73**

You can use cost as a "common denominator" across all your products sharing components

Breaking down the numbers...

red canary
a zscaler company

So, there does that leave our unit economics?

Data Type	Daily Total	Units	Cost per Unit
Video	$1,092.60	1,200 min	$0.91 / minute
Image	$429.66	2,000 images	$0.21 / image
Text	$117.73	10,000 pages	$0.01 / page

The addition of a few key metrics allows us to break down cost per unit across the product

So what can YOU track?

- emitted_files_uploaded_total
- emitted_files_bytes_uploaded_total
- triggered_files_downloaded_total
- triggered_bytes_downloaded_total
- processing_seconds_by_type_total
- dynamo_db_wcu_estimate_by_type_total
- dynamo_db_rows_read_by_type_total
- redis_bytes_written_by_type_total
- inference_input_tokens_total
- inference_output_tokens_total
- kafka_bytes_written_total

If it costs a lot of money, output a metric tracking its use!

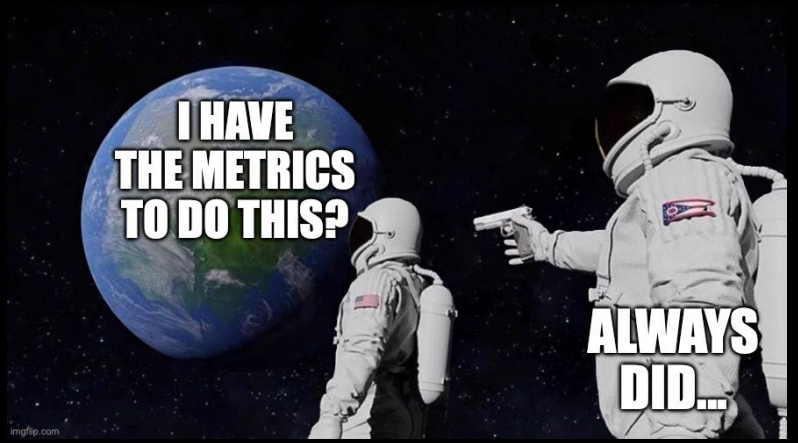
Tying it all together

We have several cronjobs running to orchestrate our cost metrics:

- Hourly cronjob **scrapes Prometheus metrics** and stores in a Postgres DB
- 5 minutely cronjob **scrapes Kubernetes metrics** and stores in Postgres
- Daily cronjob to mash it all together:
  - **download the AWS Cost and Usage Report** (CUR)
  - Merge prometheus and kubernetes metrics
  - Write merged data to postgres DB
- Daily cronjob to send prometheus metrics to a third-party cloud cost service

Your observability tooling can help you track runtime costs!

Questions?



@brianthedavis

red canary
a zscaler company