Honey, I Shrunk the ~~Kids~~ Bill!

Harjot S. Parmar

Featured

BigQuery's Ridiculous Pricing Model Cost Us $10,000 in Just 22 Seconds!!!

5 min read · Mar 19, 2025

We had a dev query the entire datalake.

The query ran for like 2-3 weeks and ended up costing about 75K.

RedditBeaver42 · 1y ago

Customer did training on Cosmos DB and left instances running. Some timer later they got a call from MS because the bill had reached 45000 USD.

Src: https://blog.det.life/bigquerys-ridiculous-pricing-model-cost-us-10-000-in-just-22-seconds-7d52e3e4ae60

Featured

BigQuery's Ridiculous Pricing Model Cost Us $10,000 in Just 22 Seconds!!!

5 min read · Mar 19, 2025

We had a dev query the entire datalake.

The query ran for like 2-3 weeks and ended up costing about 75K.

RedditBeaver42 · 1y ago

Customer did training on Cosmos DB and left instances running. Some timer later they got a call from MS because the bill had reached 45000 USD.

“Cloud comes with convenience…
and convenience comes at a cost ”

HARJOT SINGH PARMAR

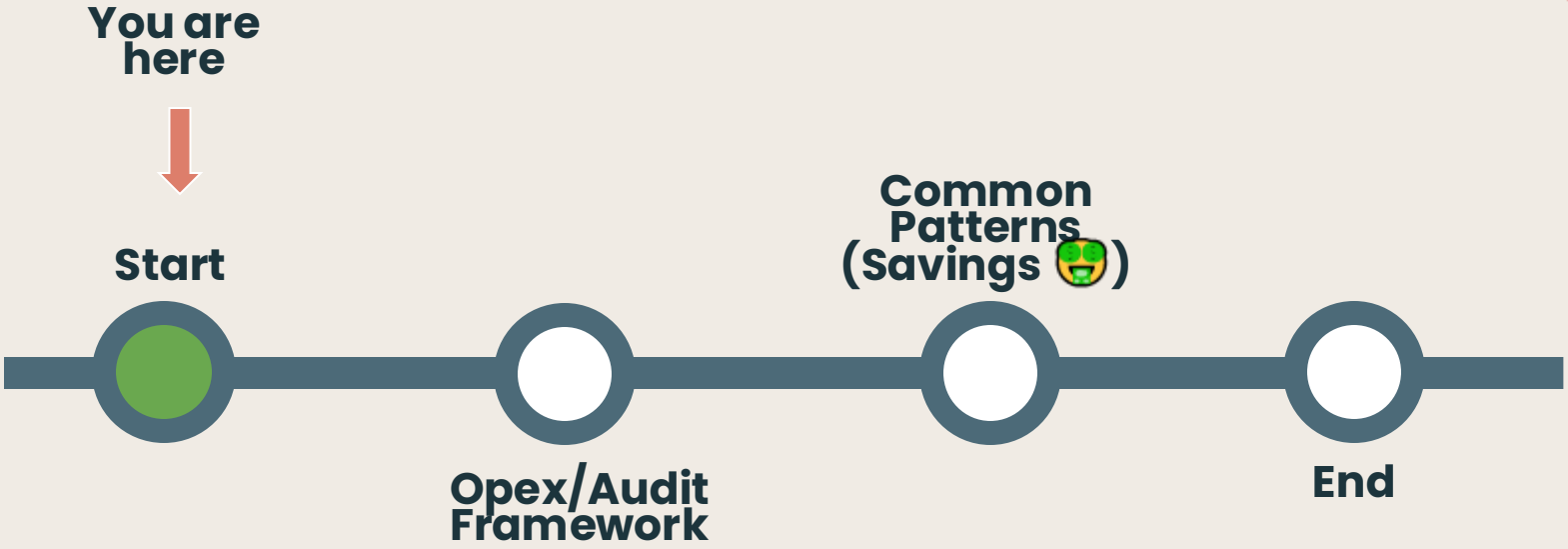
LDX3 2025 EDITION

- **Occupation**
  Staff Machine Learning Engineer @ Intuit
  Data Engineering, Generative AI and Agentic
- **Education**
  University of Waterloo, SYDE
- **Accessories**
  Macbook, Hiking Boots, Soccer Ball, Charging Cords
- **Special Interest**
  Data intensive products, architecture optimization, Trivia nights

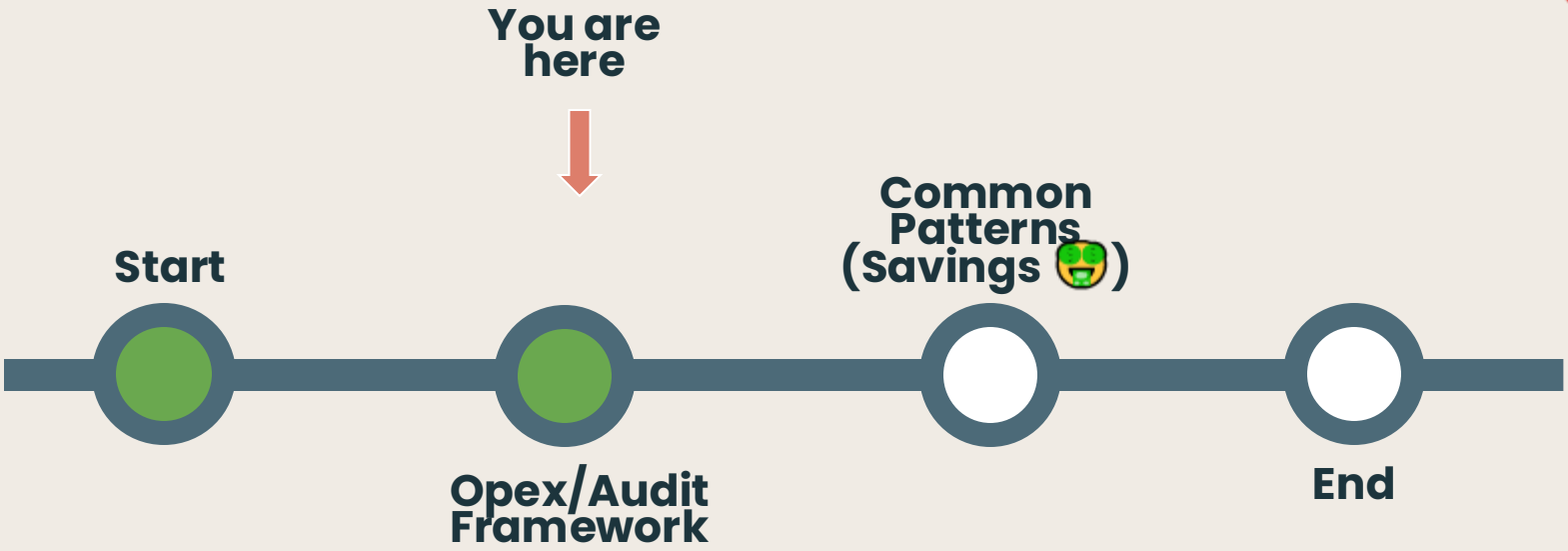
Trains Models, Trek Mountains, Trust the Process

harjotsparmar

1234567895



Travel Time: 25m


You are here
Start
Common Patterns (Savings)
Opex/Audit Framework
End

Minimal Audit Framework

Pre-audit

Laying the groundwork.
Setting up the foundation

- [ ] Technical Setup
- [ ] Cultural Setup

Audit

Put on your detective hats and start digging.

- [ ] Weekly OpEx Reviews
- [ ] Housekeeping Days
- [ ] Bounties and Rewards

Post-Audit

Sustaining the gains

- [ ] Diagramming and Optimization
- [ ] Education and Record Keeping
- [ ] Automate common patterns

Groundwork

Pre-audit

Laying the groundwork.
Setting up the foundation

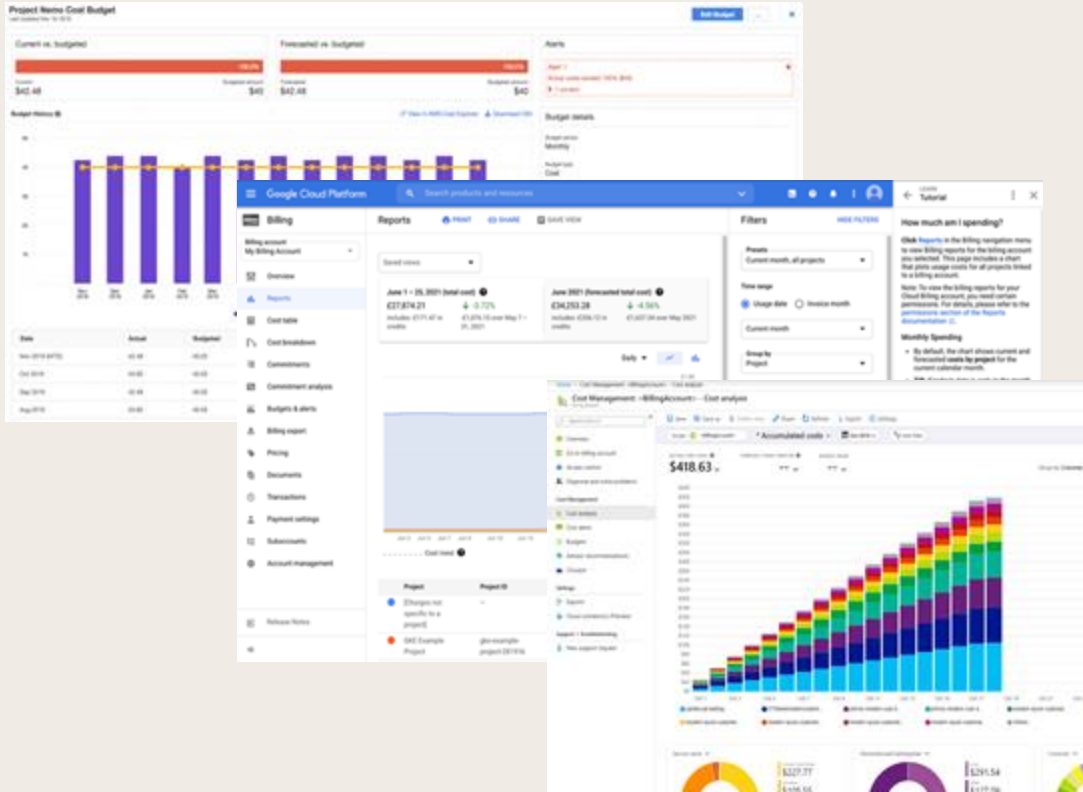
- [ ] Technical Setup
- [ ] Cultural Setup

Cloud Provider Setup

Pre-audit

Laying the groundwork.
Setting up the foundation

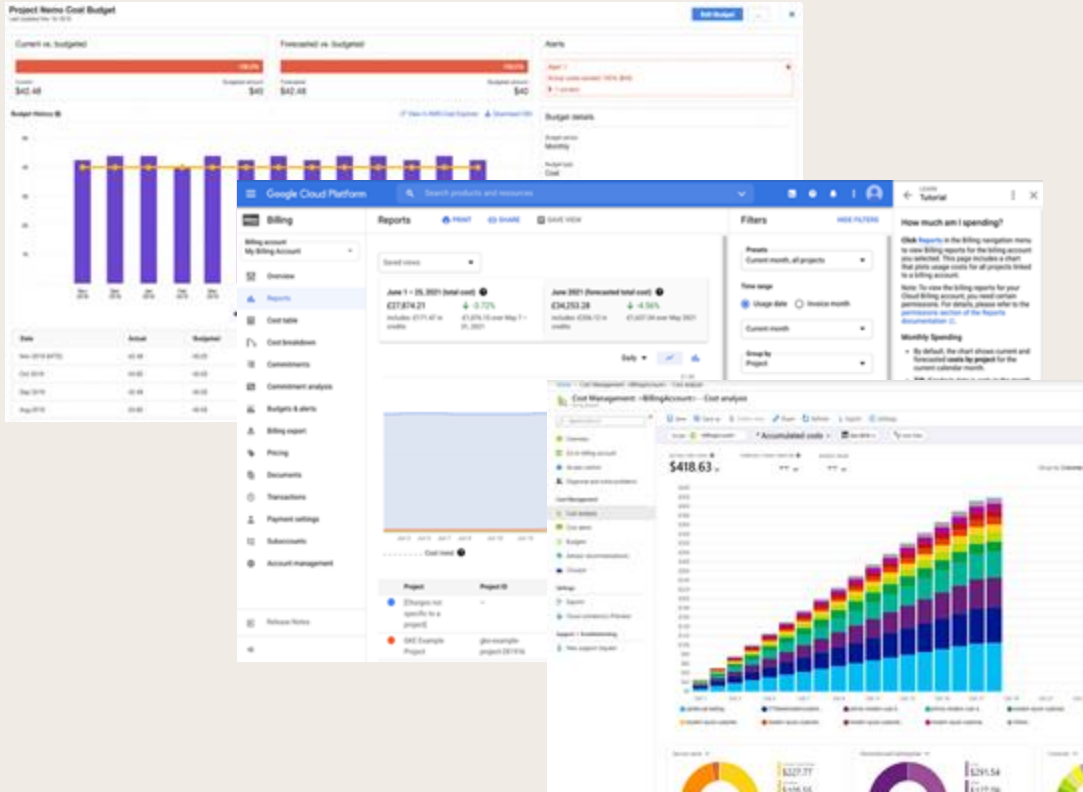
- [x] Technical Setup
- [ ] Cultural Setup



Cloud Provider Setup

Core Features

- Budgeting
- Forecasting
- Anomaly Detection
- Data Exports & Custom Reporting



Microservices Setup

Pre-audit

Laying the groundwork.
Setting up the foundation

- [x] Technical Setup
- [ ] Cultural Setup

Metrics

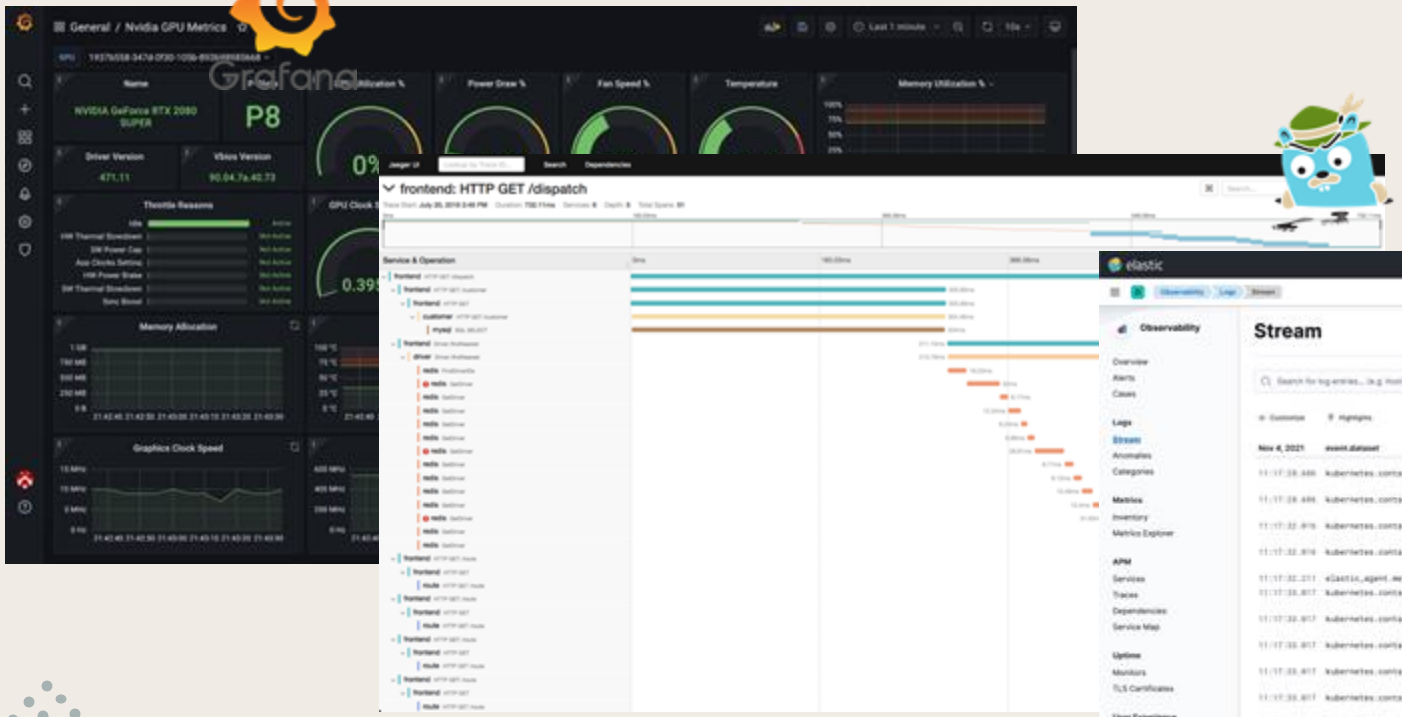
Grafana

Logs

Stack

Tracing

Microservices Setup



https://grafana.com/ | https://www.elastic.co/observability/log-monitoring | https://www.jaegertracing.io/docs/1.49/

Data Engineering & Infra

Pre-audit

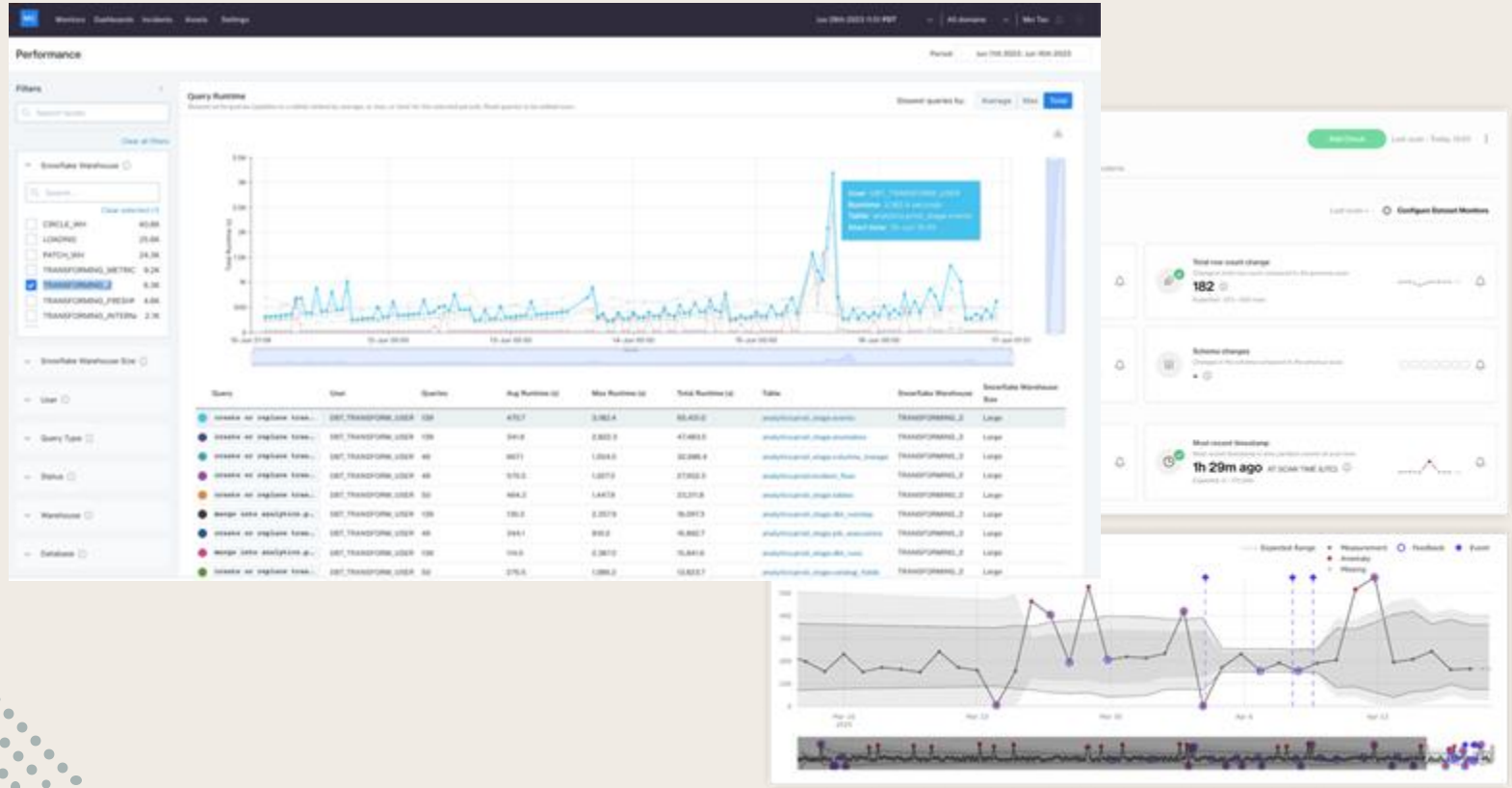
Laying the groundwork.
Setting up the foundation

- [x] Technical Setup
- [ ] Cultural Setup

SODA

MC MONTE CARLO

Data Infra



Cultural Setup

Pre-audit

Laying the groundwork.
Setting up the foundation

- [x] Technical Setup
- [x] Cultural Setup

Enable Engineers

Pre-audit

Laying the groundwork.
Setting up the foundation

- [x] Technical Setup
- [x] Cultural Setup

Enable Engineers

Pre-audit

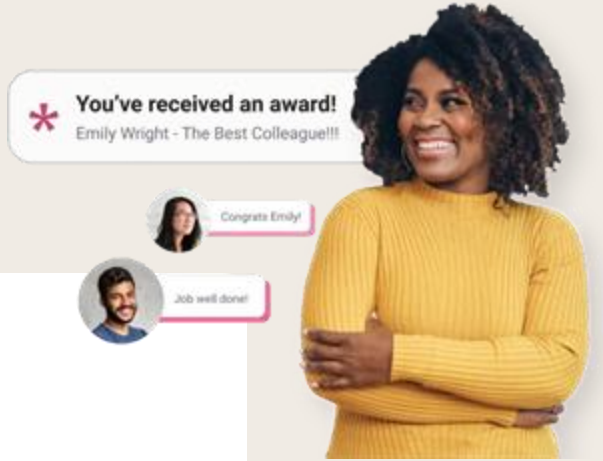
Laying the groundwork.
Setting up the foundation

- [x] Technical Setup
- [x] Cultural Setup

Recognition and Incentives

> Show me the incentive, I'll show you the outcome

Charlie Munger



Pre-audit

Laying the groundwork.
Setting up the foundation

- [x] Technical Setup
- [x] Cultural Setup

ce: https://www.workhuman.com/

Minimal Audit Framework

Pre-audit

Laying the groundwork.
Setting up the foundation

- [x] Technical Setup
- [x] Cultural Setup

Audit

Put on your detective hats and start digging.

- [ ] Weekly OpEx Reviews
- [ ] Housekeeping Days
- [ ] Bounties and Rewards

Weekly OpEx Reviews

Regular meetings focused solely on operational expenditures, including cloud costs

Audit

Put on your detective hats and start digging.

- [x] Weekly OpEx Reviews
- [ ] Housekeeping Days
- [ ] Bounties and Rewards

Monthly Housekeeping Days

Routine chores that prevent cost creep

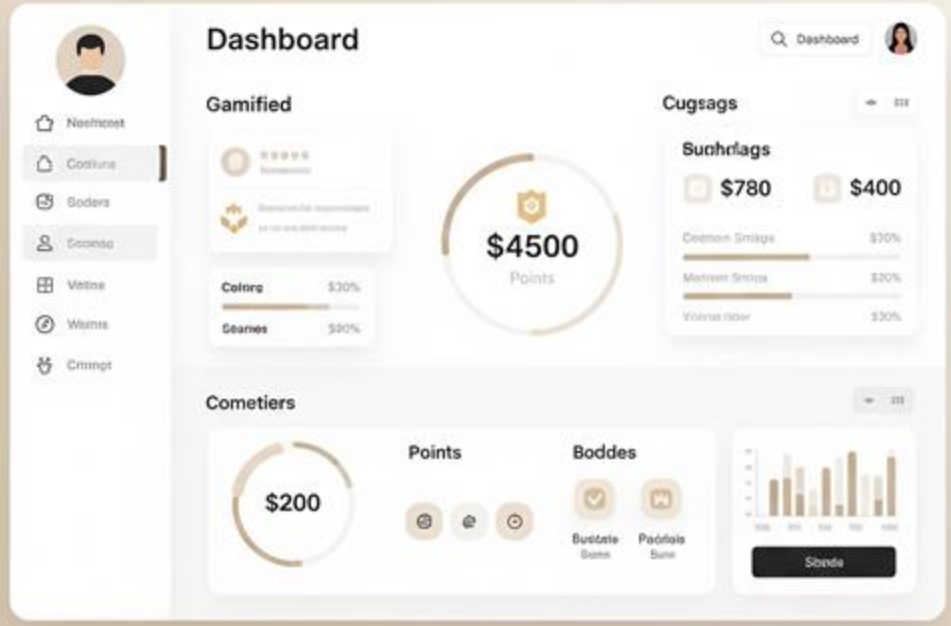
Audit

Put on your detective hats and start digging.

- [x] Weekly OpEx Reviews
- [x] Housekeeping Days

Bounties And Rewards

Actively incentivize engineers to find and suggest cost-saving opportunities



Audit

Put on your detective hats and start digging.

- [x] Weekly OpEx Reviews
- [x] Housekeeping Days
- [x] Bounties and Rewards

Minimal Audit Framework

Pre-audit

Laying the groundwork.
Setting up the foundation

- [x] Technical Setup
- [x] Cultural Setup

Audit

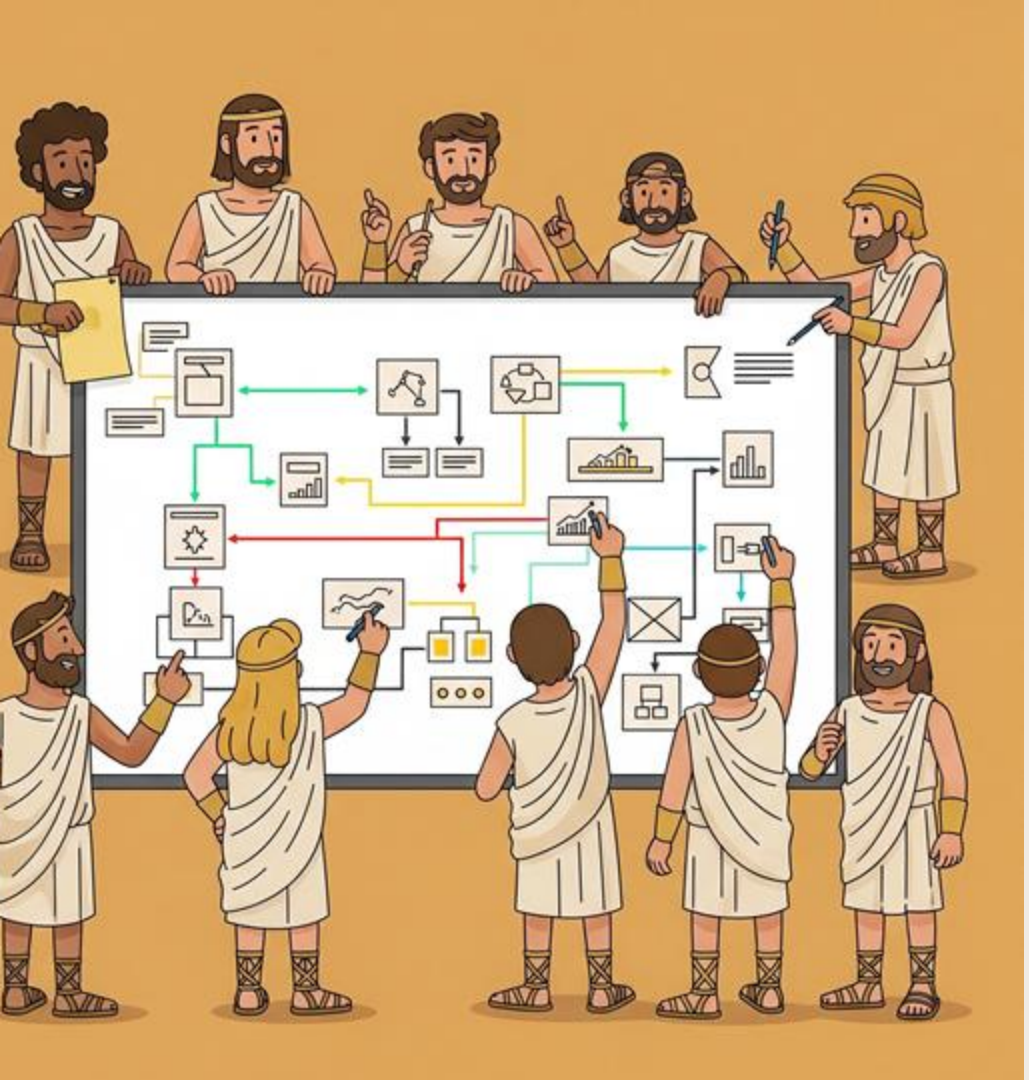
Put on your detective hats and start digging.

- [x] Weekly OpEx Reviews
- [x] Housekeeping Days
- [x] Bounties and Rewards

Post-Audit

Sustaining the gains

- [ ] Diagramming and Optimization
- [ ] Education and Record Keeping
- [ ] Automate common patterns



Post-Audit: Diagramming and Optimization

Post-Audit

Sustaining the gains

- [x] Diagramming and Optimization
- [ ] Education and Record Keeping
- [ ] Automate common patterns

Post-Audit: Education and Record Keeping

Post-Audit

Sustaining the gains

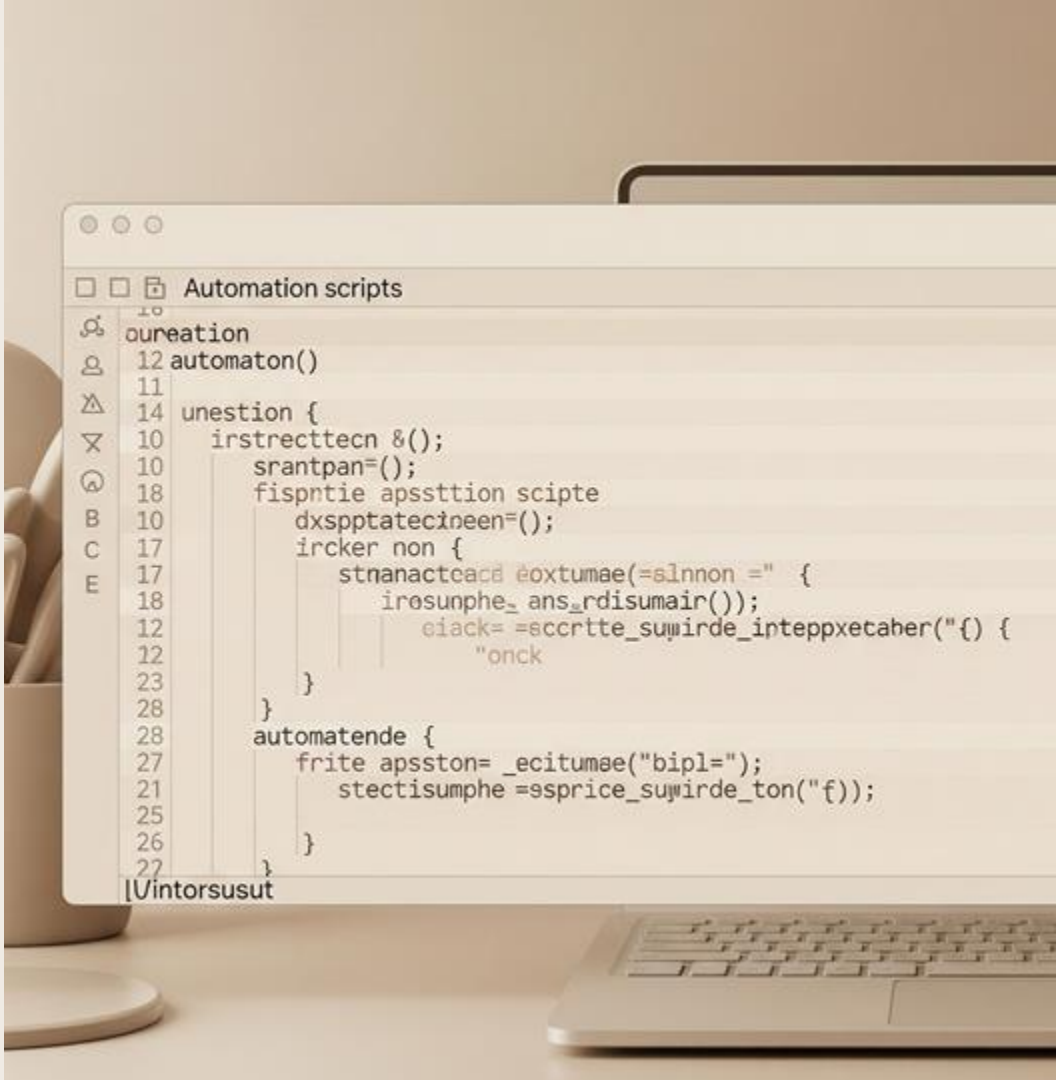
- [x] Diagramming and Optimization
- [x] Education and Record Keeping
- [ ] Automate common patterns

Post-Audit: Automate Common Patterns

Post-Audit

Sustaining the gains

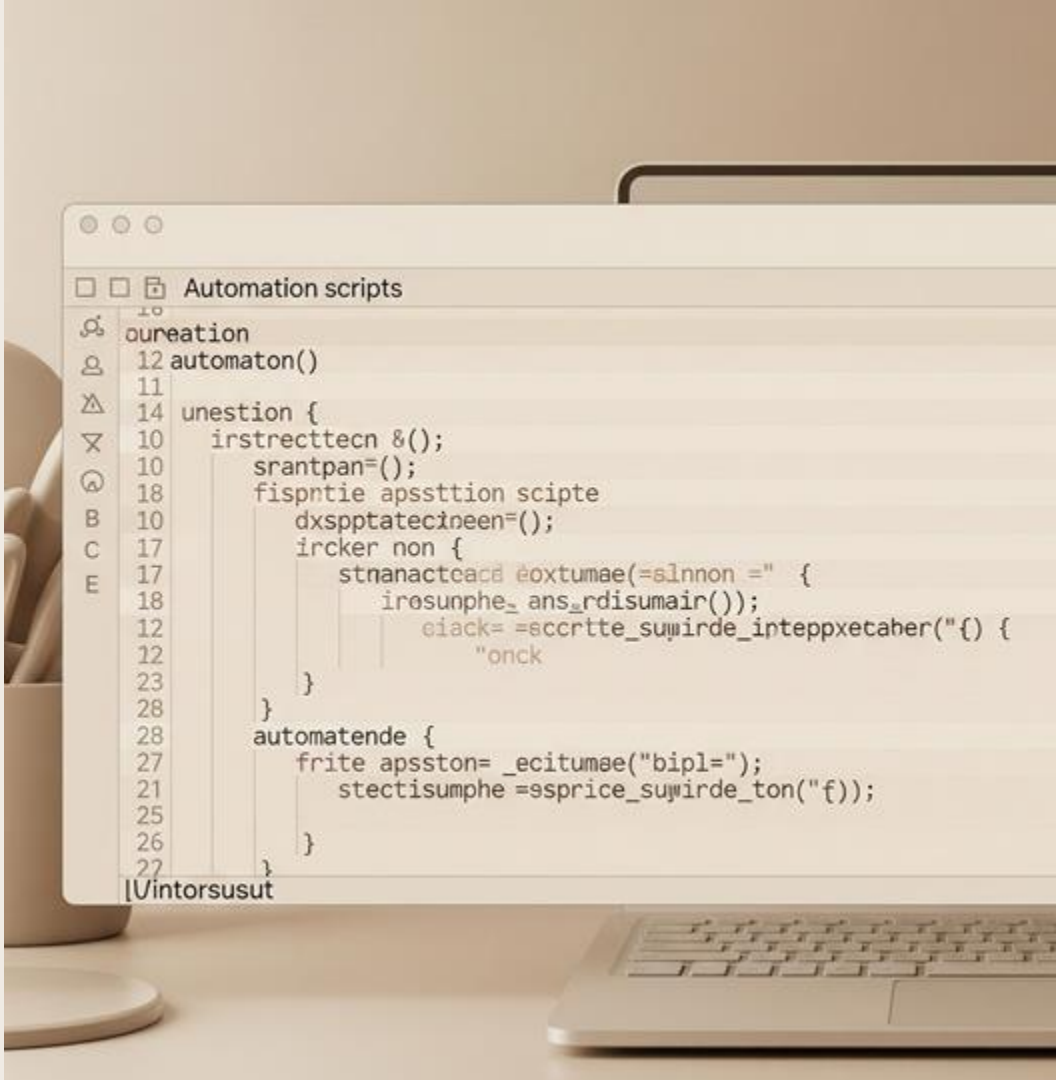
- [x] Diagramming and Optimization
- [x] Education and Record Keeping
- [x] Automate common patterns



Post-Audit: Automate Common Patterns

Eg:

- Automated shutdown
- Script for the cleanup
- lifecycle policy



Minimal Audit Framework

Pre-audit

Laying the groundwork.
Setting up the foundation

- [x] Technical Setup
- [x] Cultural Setup

Audit

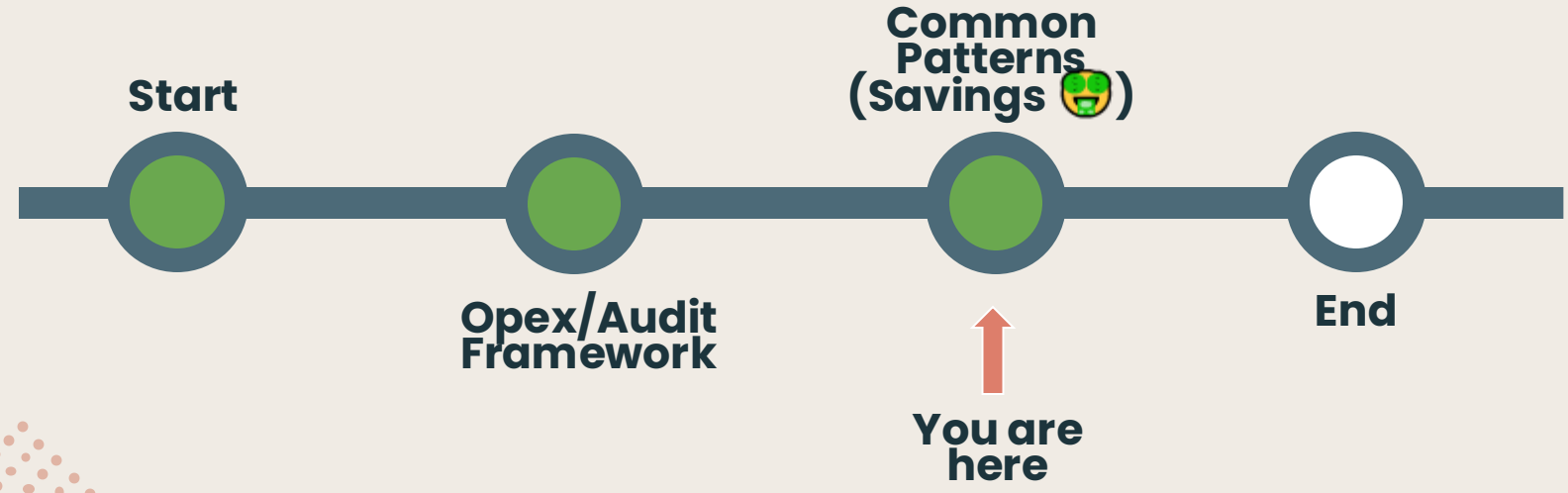
Put on your detective hats and start digging.

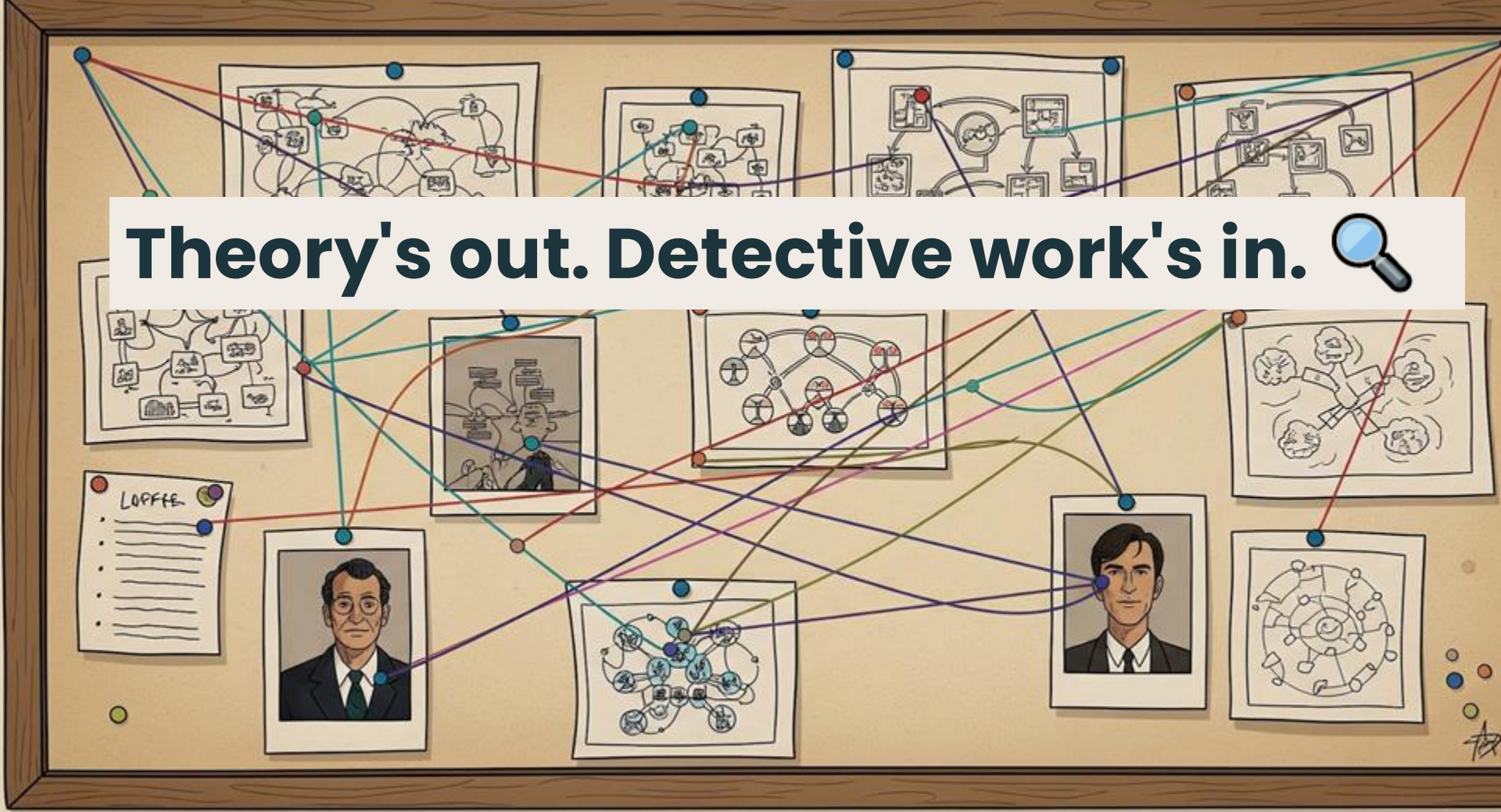
- [x] Weekly OpEx Reviews
- [x] Housekeeping Days
- [x] Bounties and Rewards

Post-Audit

Sustaining the gains

- [x] Diagramming and Optimization
- [x] Education and Record Keeping
- [x] Automate common patterns

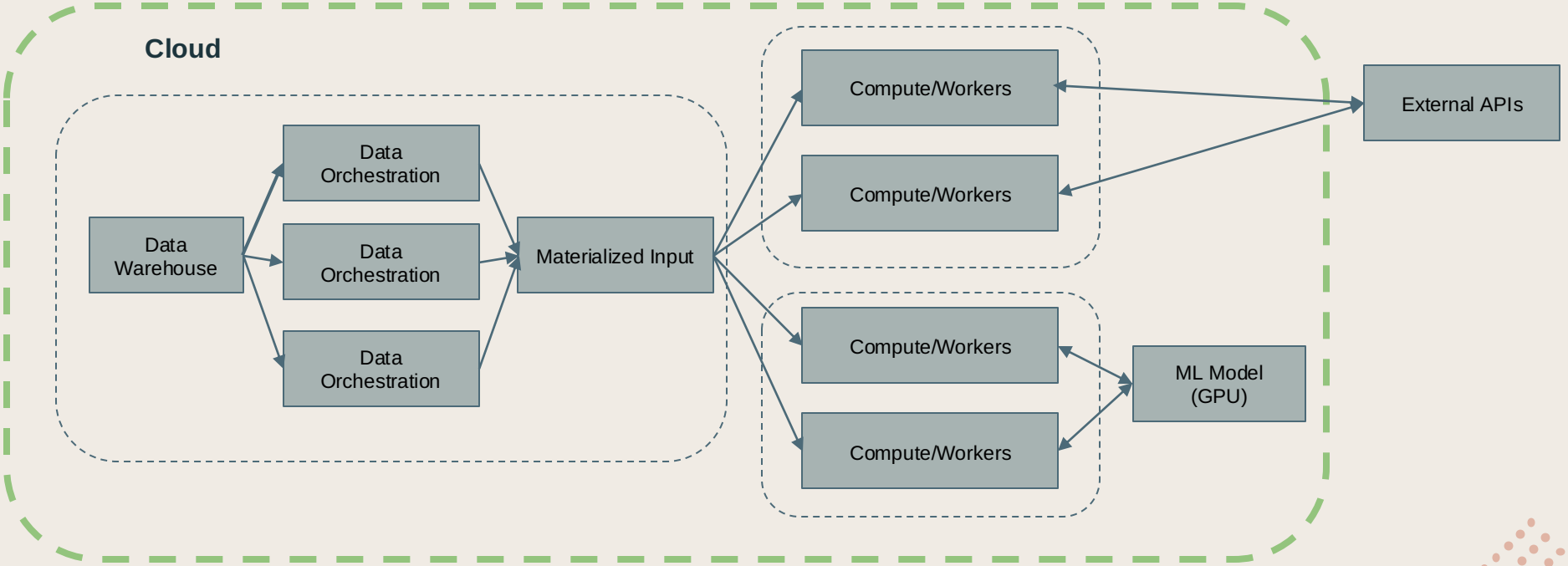

Start
Opex/Audit Framework
Common Patterns (Savings)
You are here
End


Theory's out. Detective work's in. 🔍

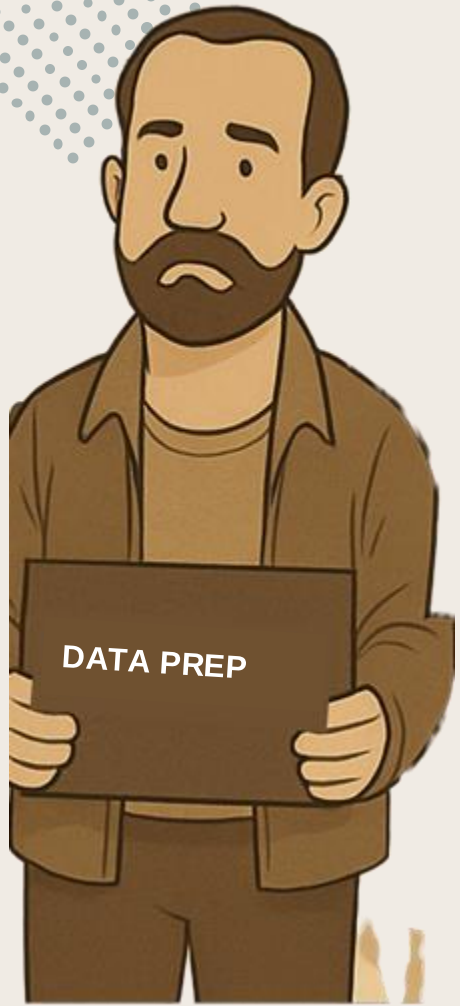
Optimizing for Hidden Cost

Identifying Common patterns for wastage

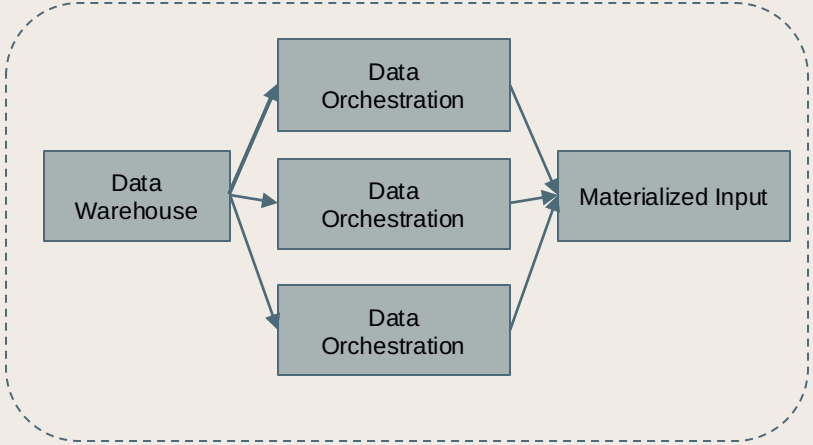
ML Pipeline


Cloud
Data Warehouse
Data Orchestration
Data Orchestration
Data Orchestration
Materialized Input
Compute/Workers
Compute/Workers
Compute/Workers
Compute/Workers
External APIs
ML Model (GPU)


6'0"
6'6"
6'0"
5'6"
DATA PREP
COMPUTE
GPU
External APIs
LOGS
NETWORKING



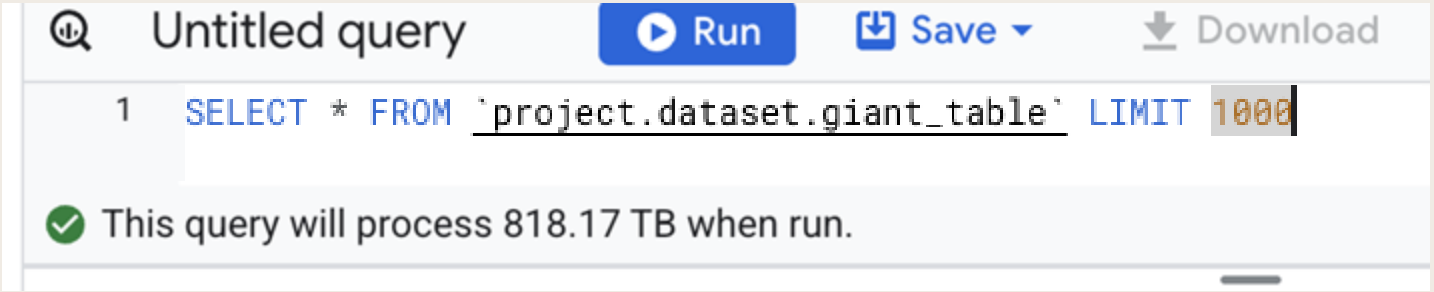
Suspect #1
Data Preparation



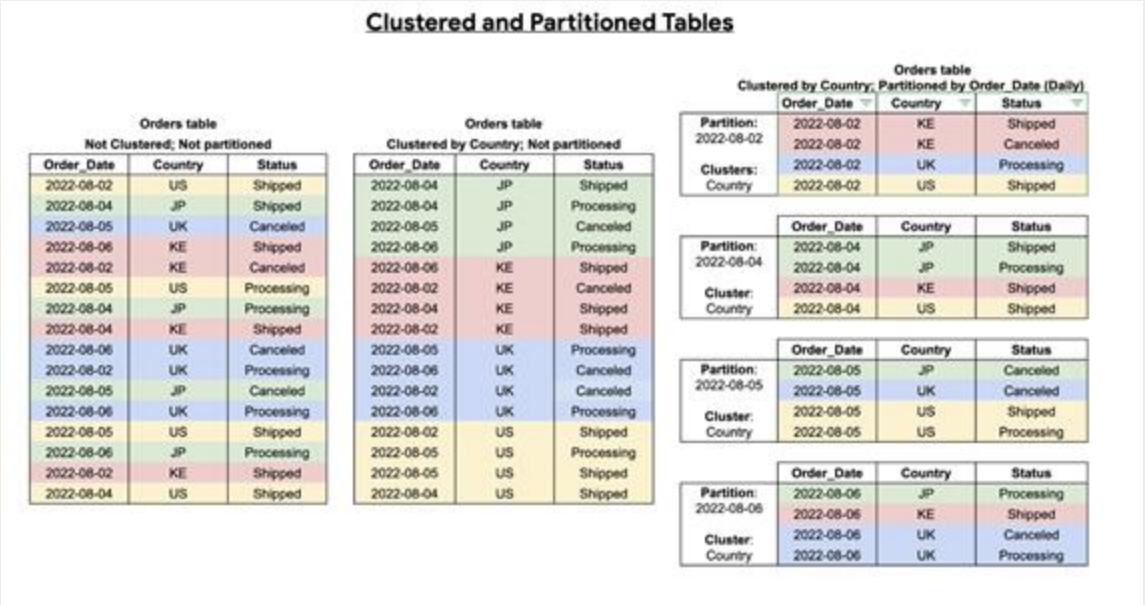
Excessive data scans

$4,000 per run

- Select * from _____



Bigquery



Clustered and Partitioned Tables

Orders table
Not Clustered; Not partitioned

Order_Date	Country	Status
2022-08-02	US	Shipped
2022-08-04	JP	Shipped
2022-08-05	UK	Canceled
2022-08-06	KE	Shipped
2022-08-02	KE	Canceled
2022-08-05	US	Processing
2022-08-04	JP	Processing
2022-08-04	KE	Shipped
2022-08-06	UK	Canceled
2022-08-02	UK	Processing
2022-08-05	JP	Canceled
2022-08-06	UK	Processing
2022-08-05	US	Shipped
2022-08-06	JP	Processing
2022-08-02	KE	Shipped
2022-08-04	US	Shipped

Orders table
Clustered by Country; Not partitioned

Order_Date	Country	Status
2022-08-04	JP	Shipped
2022-08-04	JP	Processing
2022-08-05	JP	Canceled
2022-08-06	JP	Processing
2022-08-06	KE	Shipped
2022-08-02	KE	Canceled
2022-08-04	KE	Shipped
2022-08-02	KE	Shipped
2022-08-05	UK	Processing
2022-08-06	UK	Canceled
2022-08-02	UK	Canceled
2022-08-06	UK	Processing
2022-08-02	US	Shipped
2022-08-05	US	Processing
2022-08-05	US	Shipped
2022-08-04	US	Shipped

Orders table
Clustered by Country; Partitioned by Order_Date (Daily)

	Order_Date	Country	Status
Partition: 2022-08-02	2022-08-02	KE	Shipped
	2022-08-02	KE	Canceled
Clusters: Country	2022-08-02	UK	Processing
	2022-08-02	US	Shipped

	Order_Date	Country	Status
Partition: 2022-08-04	2022-08-04	JP	Shipped
	2022-08-04	JP	Processing
Cluster: Country	2022-08-04	KE	Shipped
	2022-08-04	US	Shipped

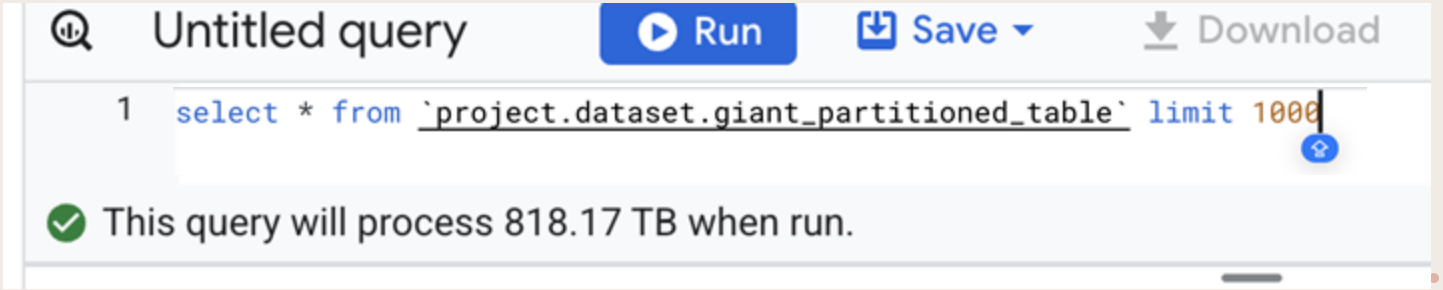
	Order_Date	Country	Status
Partition: 2022-08-05	2022-08-05	JP	Canceled
	2022-08-05	UK	Canceled
Cluster: Country	2022-08-05	US	Shipped
	2022-08-05	US	Processing

	Order_Date	Country	Status
Partition: 2022-08-06	2022-08-06	JP	Processing
	2022-08-06	KE	Shipped
Cluster: Country	2022-08-06	UK	Canceled
	2022-08-06	UK	Processing

Bigquery

```sql
CREATE TABLE
  project.dataset.giant_table_partitioned (transaction_id INT64, transaction_date DATE)
PARTITION BY
  transaction_date
AS (
  SELECT
    transaction_id, transaction_date
  FROM
    project.dataset.giant_table

);
```



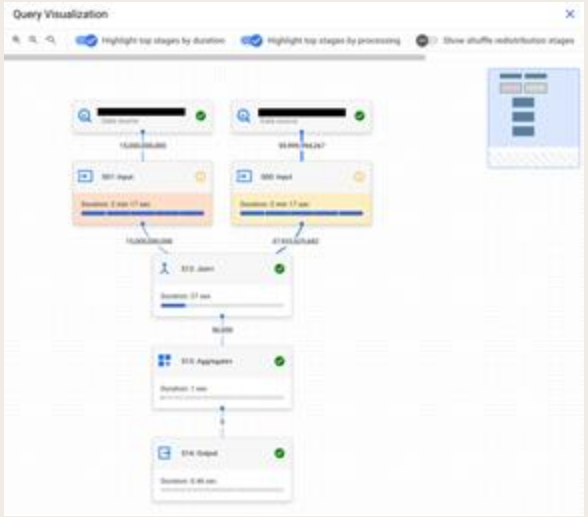
Bigquery

```
CREATE TABLE
  `project.dataset.giant_partitioned_table_with_required_filter`
(transaction_id INT64, transaction_date DATE)
PARTITION BY
  transaction_date
  OPTIONS (
    require_partition_filter = TRUE) as
  select * from `project.dataset.giant_table`
```

```
select transaction_id, transaction_date from
`project.dataset.giant_partitioned_table_with_required_filter`
where transaction_date = "2025-06-16"
```

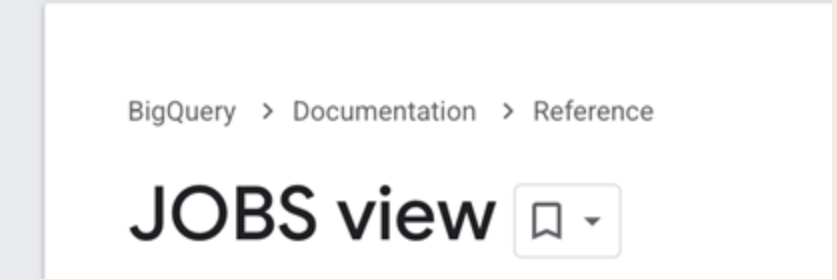
This query will process 64.23 MB when run.

Bigquery



Dig in for more insights

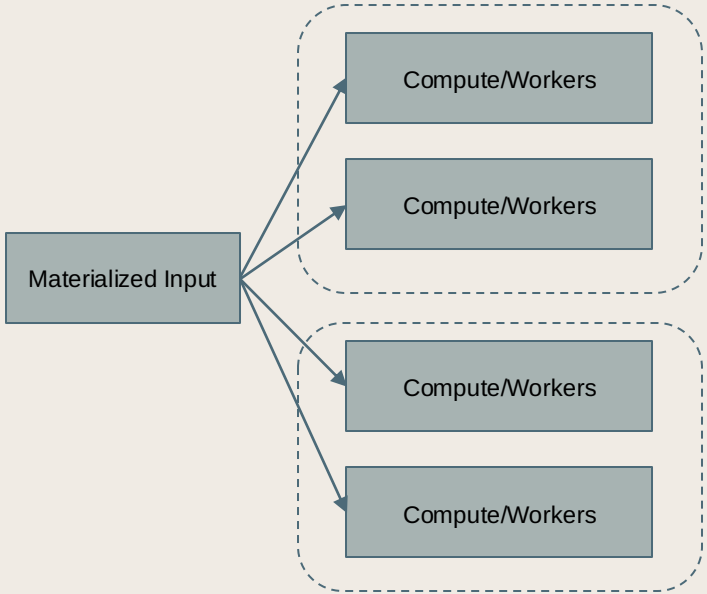
- Execution graph
- Jobs view



Remedies

- **Reduce data before joins:** Apply `WHERE` clauses as early as possible in queries to minimize data shuffling during joins.
- **Choose appropriate pricing model:** Select on-demand pricing for unpredictable workloads and capacity pricing for consistent, high-volume workloads.
- **Avoid frequent table overwrites:** Use incremental data loads instead of overwrites to prevent hidden storage costs from time travel retention.
- **Use query quotas:** Set custom daily query quotas at the project or user level to limit the amount of data processed.
    - Utilize `Maximum bytes billed` per query: Define a maximum number of bytes a query can process to avoid unexpected expenses.

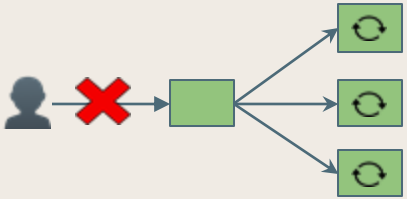
Suspect #2
Distributed Compute



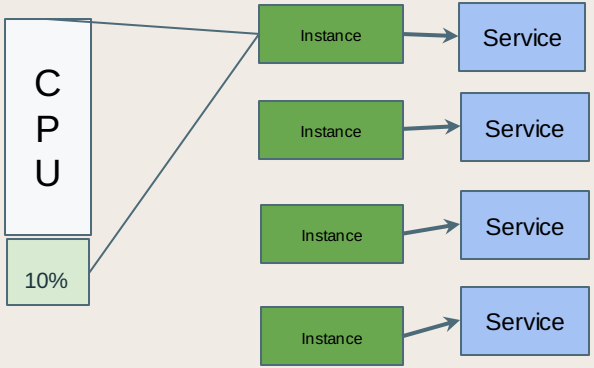
Compute Wastage

Straggler Tasks

Partial/Failed Executions



Orphan Computations



Mixed I/O and CPU bound executions

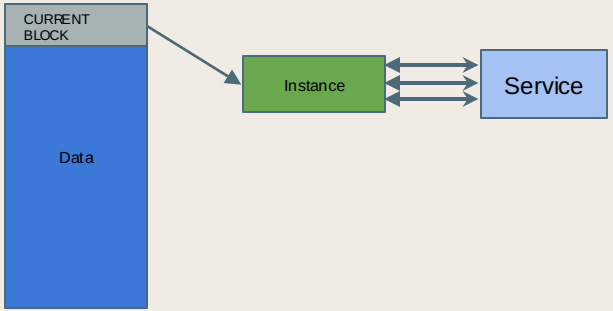
- Underutilized compute
- Excessive Idle time
- Repeated tasks
- Overprovisioning
- Dangling temporary artifacts

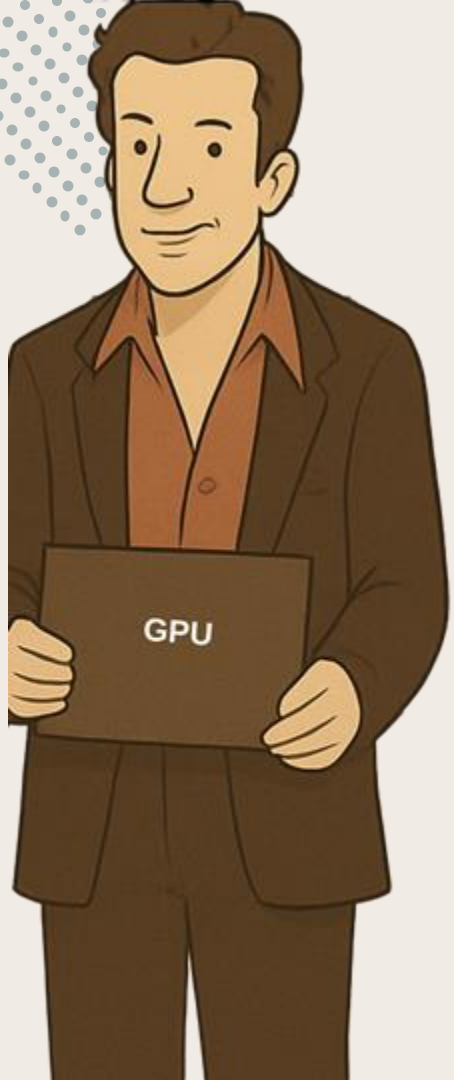
Remedies

- Implement custom checkpoints behaviour in batch pipelines
- Delta tracking queries as intermediate steps

Complete input - p[illegible]ial result = Δ

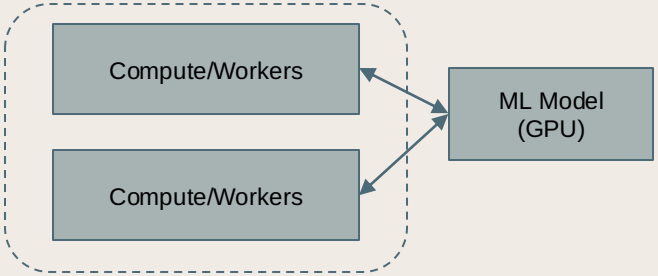
- Implement custom data loaders for paged reading from tables
- Investing in building async API executors on the single nodes



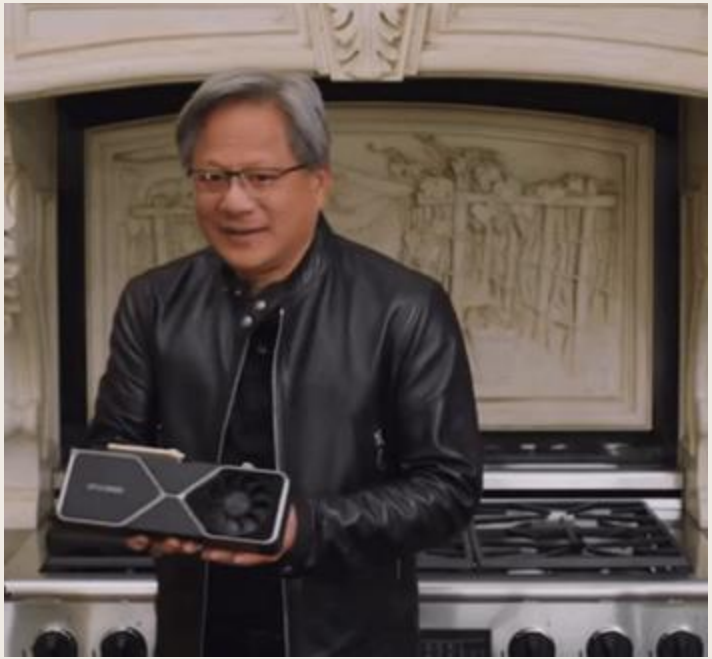


Suspect #3

Graphical Processing Unit (GPU)



GPU Wastage



Beautiful beast, very powerful for ML inferences, but a few understand them well…and that leads to underutilized instances.

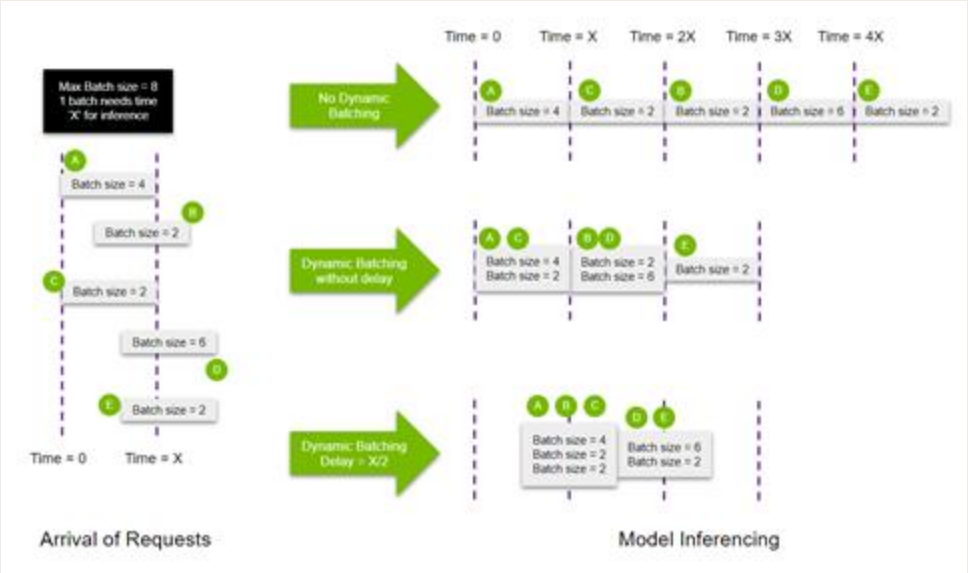
CPU	GPU (usage 0%)
CPU	GPU (usage 0%)
CPU	GPU (usage 0%)

Misconfigured executions

- VRAM not utilized
- Incompatible Driver and Toolkit versions on host and image
- FastAPI service wrappers for inference

Remedies

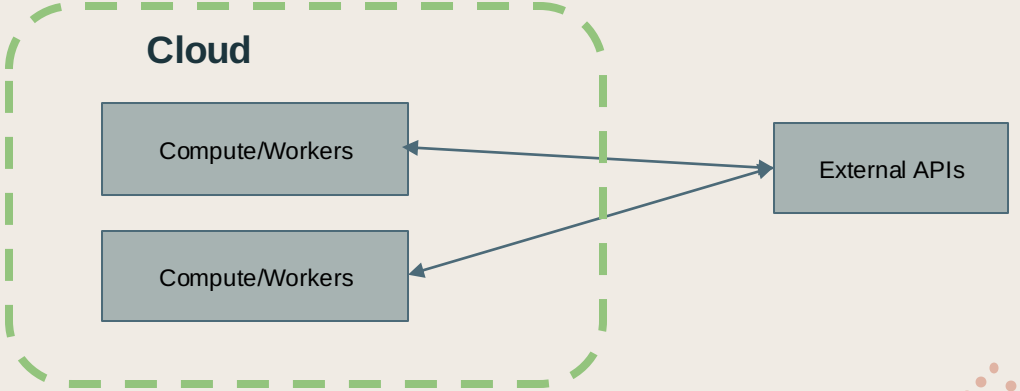
- Automate the shutdowns
- Monitor for GPU memory utilization with custom dashboards
- Optimize GPU inference with proprietary inference servers with features:
    - Multi instance deployment
    - server side batching for inference



https://docs.nvidia.com/deeplearning/triton-inference-server/user-guide/docs/tutorials/Conceptual_Guide/Part_2-improving_resource_utilization/README.html

Suspect #4-5

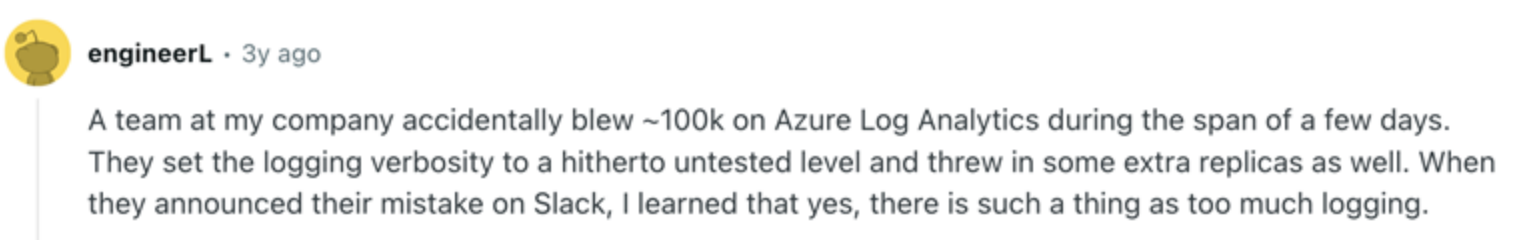
External API
Logs
Egress



API and Networking kerfuffle

- Chatty Services generating lot of logs, incorrect log levels
- External API calls/Data Processing and Sharing over public internet
- Cross-Region and Cross-Availability Zone Transfers and Networking

No such thing as too much logging?

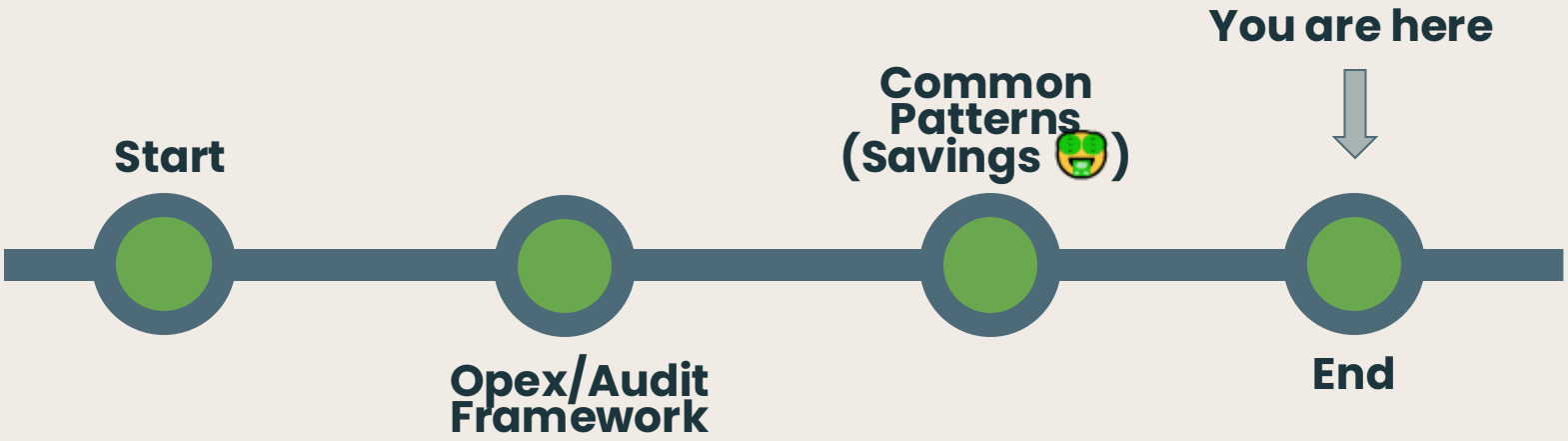


Src: https://www.reddit.com/r/devops/comments/udgohy/there_is_no_such_thing_as_too_much_logging_or_is/

Remedies

- Understand logging service pricing model
- Ensure proper log levels and verbosity, proper Structured Logs
- Implement log segregation and offload to cold storage if needed
- Avoid cross region service calls
- Use VPC peering for cross cloud service communication

https://www.harness.io/blog/cloud-logging-optimization


You are here
Common Patterns (Savings)
Start
Opex/Audit Framework
End