



**Some tips and tricks**



# Lisa

# Hoving

*Big Data Engineer*



Rate my session



Github

# Household announcements

- An overview of options
- Some in preview
  - Need Unity Catalog
- Based on my own experiences

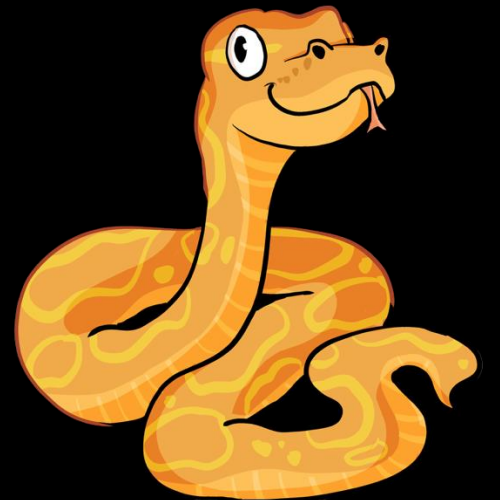


# Agenda

- 01** An expensive Python
- 02** Azure Databricks Pricing
- 03** Monitoring & Alerts
- 04** Solutions (1-5)
- 05** Conclusion



# 1. An expensive python



# An expensive python

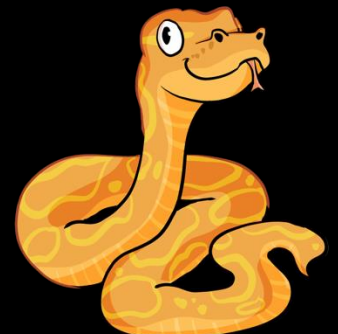
## Azure spend

Microsoft Azure Sponsorship

Subscription Cost: \$920.35

SERVICE NAME	SERVICE RESOURCE	SPEND
Storage	Hot GRS Iterative Read Operations	\$387.44
Azure Databricks	Premium All-Purpose Photon DBU	\$374.16
Virtual Machines	D4ds v5	\$102.32
Storage	P15 LRS Disk	\$22.65
Storage	Hot GRS Write Operations	\$19.56
IoT Hub	S1 Unit	\$4.84

*(14-06-2024 to 16-06-2024)*



## 2. Azure Databricks Pricing





# 2. Azure Databricks Pricing

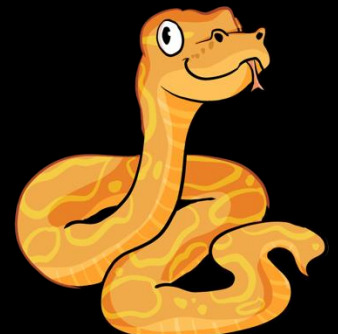
## Azure spend

Microsoft Azure Sponsorship

Subscription Cost: \$920.35

SERVICE NAME	SERVICE RESOURCE	SPEND
Storage	Hot GRS Iterative Read Operations	\$387.44
Azure Databricks	Premium All-Purpose Photon DBU	\$374.16
Virtual Machines	D4ds v5	\$102.32
Storage	P15 LRS Disk	\$22.65
Storage	Hot GRS Write Operations	\$19.56
IoT Hub	S1 Unit	\$4.84

(14-06-2024 to 16-06-2024)

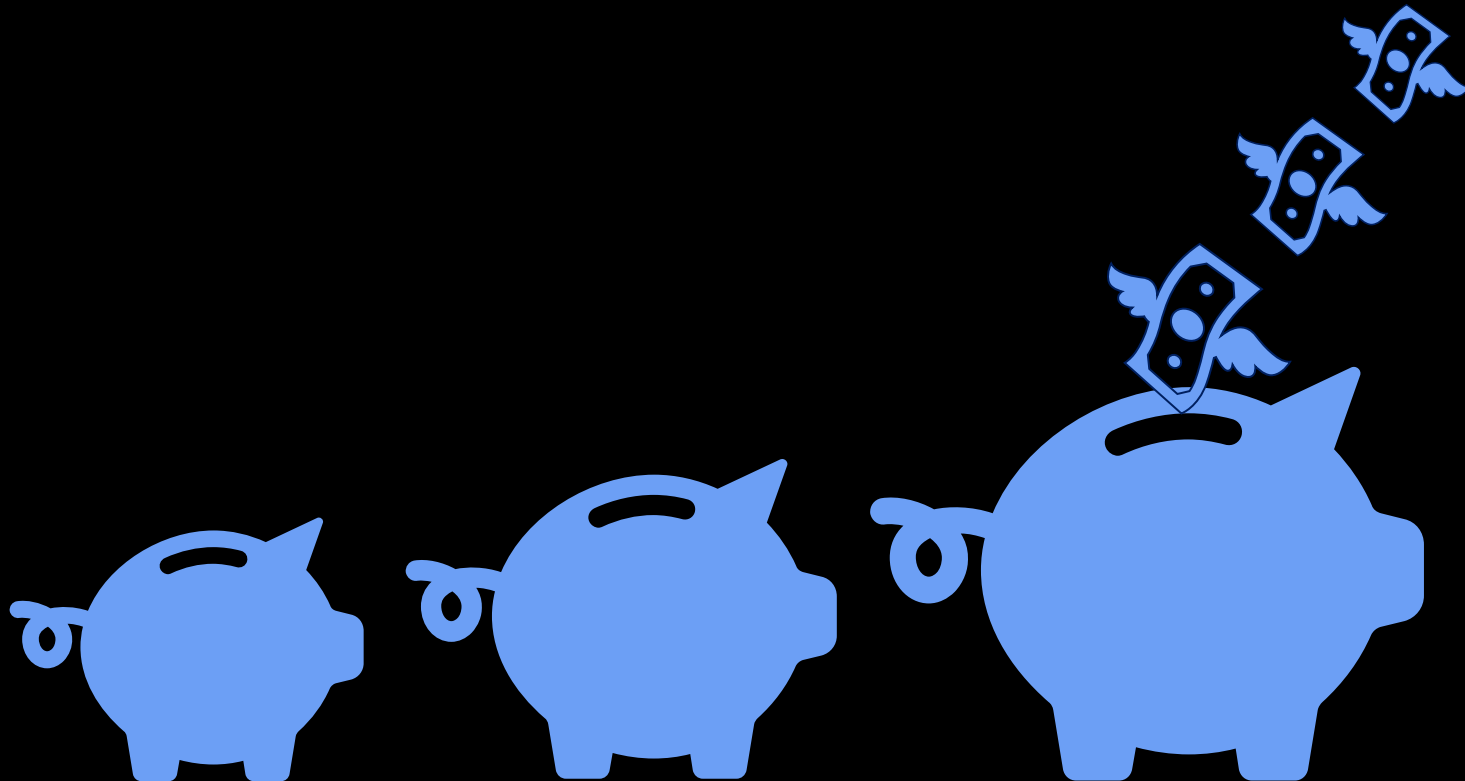




## 2. Azure Databricks Pricing


### Databricks Unit (DBU)




- Normalized unit of processing power
  - Per hour





# 2. Azure Databricks Pricing

Home >


 **DiscountingData**  
Azure Databricks Service


  


Search


 


Overview

 Activity log

 Access control (IAM)

 Tags

 Diagnose and solve problems


 Resource visualizer

> Settings

> Monitoring

> Automation

> Help

 Delete

^ Essentials

Status : Active

Resource group

Location : West Europe

Subscription

Subscription ID


Tags [\(edit\)](#) : workshop : helpmydatabricksistooexpensive

Managed Resource Group

URL

Pricing Tier : [Premium \(+ Role-based access controls\)](#) [\(Click to cha...](#)

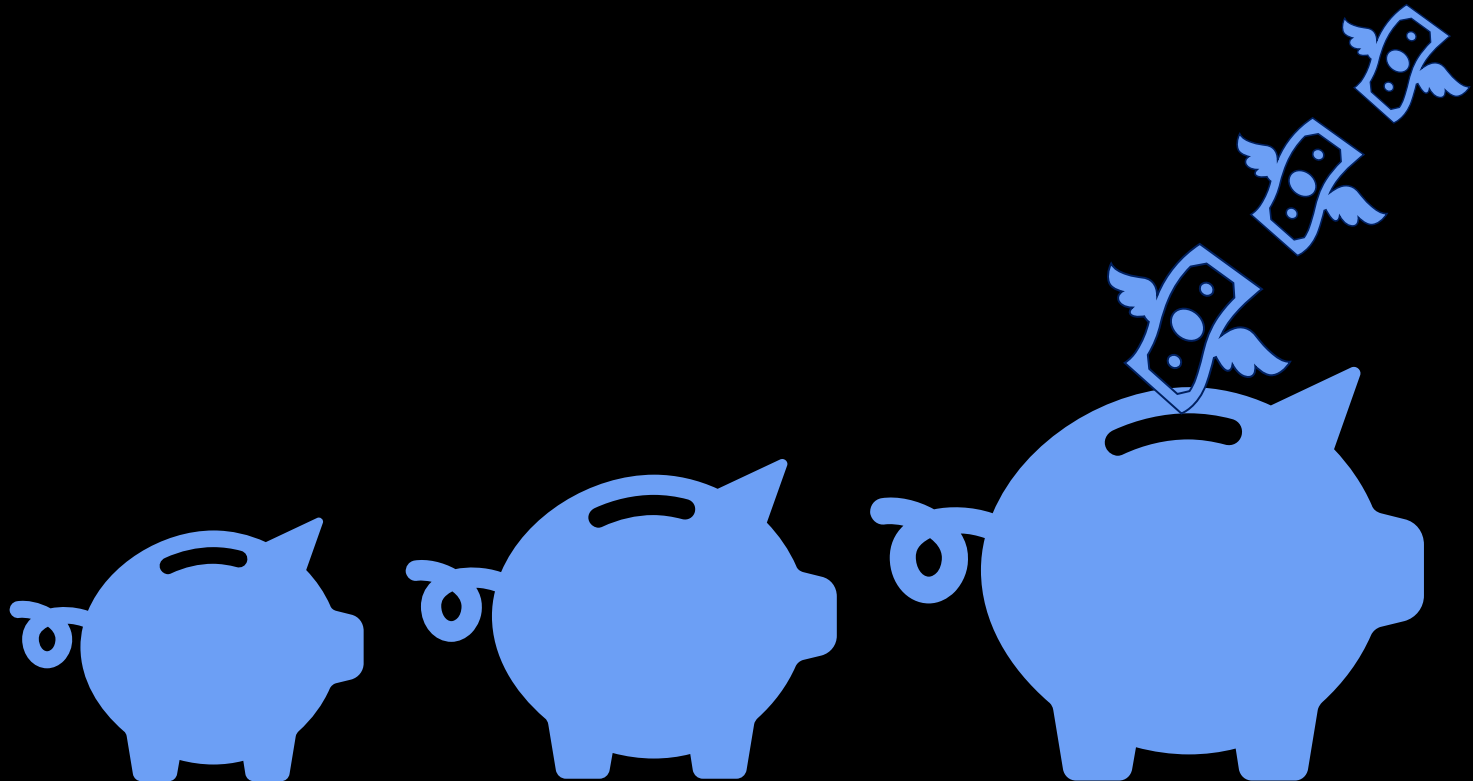
JSON View



Launch Workspace

## 2. Azure Databricks Pricing

- DBU
- Virtual Machines



# 2. Azure Databricks Pricing

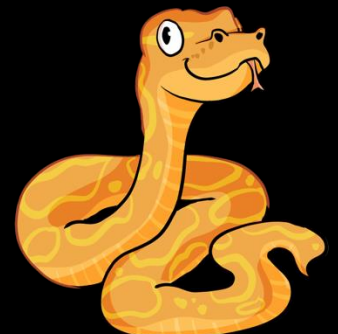
## Azure spend

Microsoft Azure Sponsorship

Subscription Cost: \$920.35

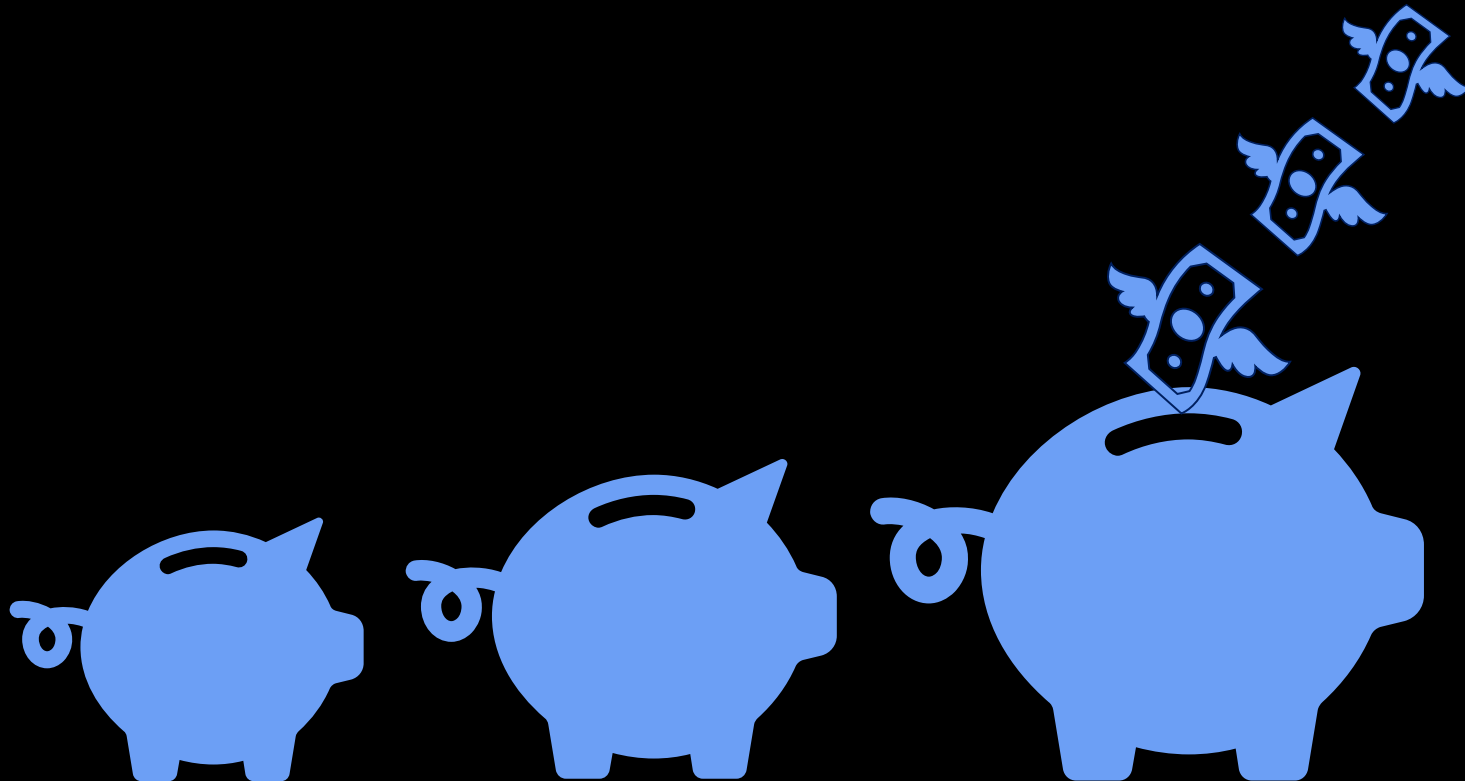
SERVICE NAME	SERVICE RESOURCE	SPEND
Storage	Hot GRS Iterative Read Operations	\$387.44
Azure Databricks	Premium All-Purpose Photon DBU	\$374.16
Virtual Machines	D4ds v5	\$102.32
Storage	P15 LRS Disk	\$22.65
Storage	Hot GRS Write Operations	\$19.56
IoT Hub	S1 Unit	\$4.84

(14-06-2024 to 16-06-2024)



## 2. Azure Databricks Pricing

- DBU
- Virtual Machines
  - Disks
  - IP Address
- Other resources
  - Storage Account
  - Key Vault
  - Log Analytics
  - Data sources



**This is why you  
should monitor!**





## **3. Monitoring & Alerts**



# 3. Monitoring & Alerts

## 2 Options

- Azure Databricks (Unity Catalog in preview)
  - Budgets
  - Serverless
- Azure Portal
  - Budgets
  - Alerts



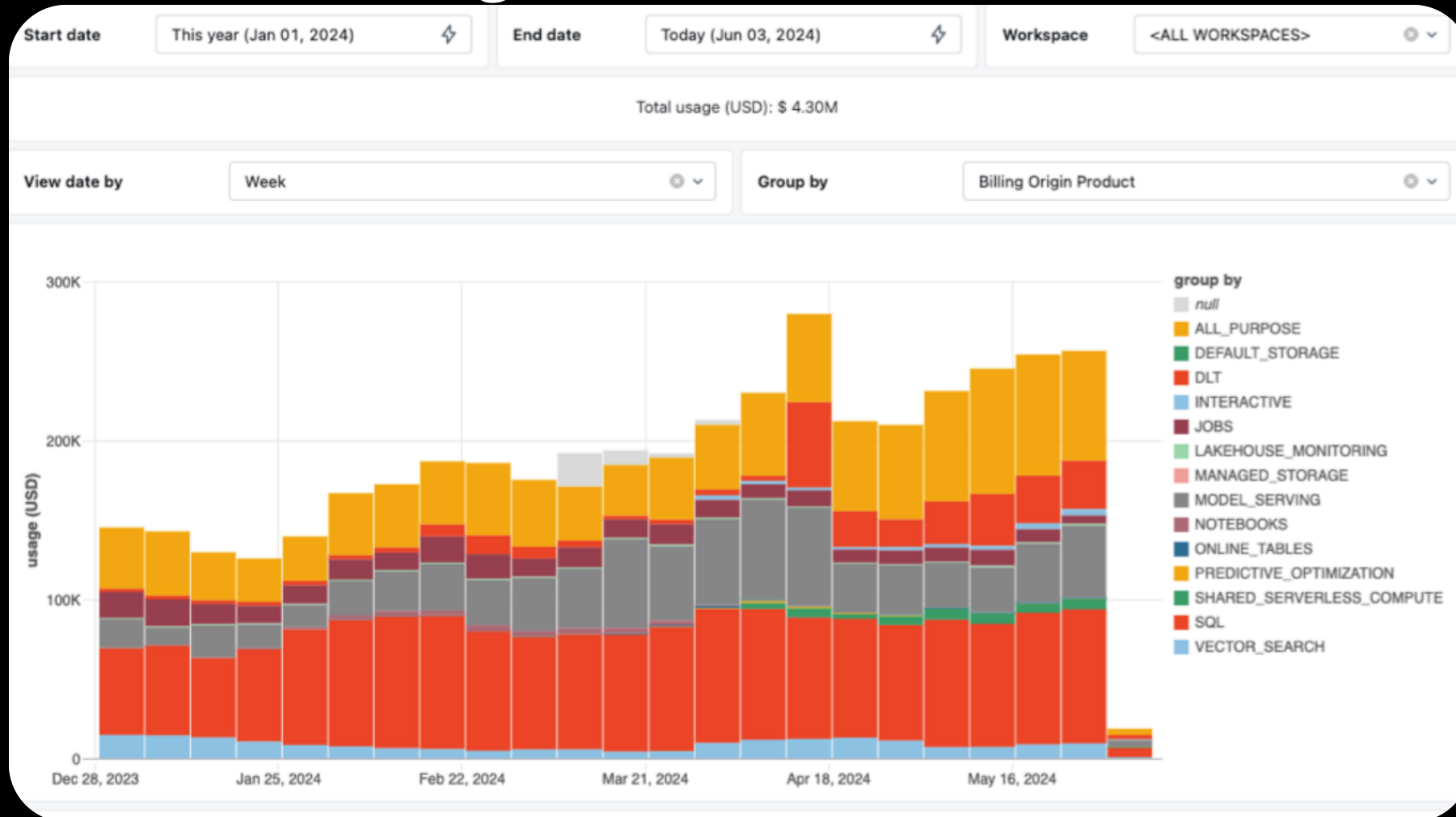
Serverless



Workspace

# 3. Monitoring & Alerts

## Databricks Usage Dashboard



# 4. Solutions

A blue lightbulb icon is positioned above the end of the text "Solutions". The lightbulb is drawn with a simple blue outline, featuring a rounded top and a base with three horizontal lines and a small semi-circle at the bottom.

# Solutions

- 01** Optimize Data Source
- 02** Cluster Settings
- 03** Optimize Code
- 04** Make it a Job!
- 05** Stream or Micro Batch?



# 4. Solutions

*(1) Optimize Data Source*



# 4. Solutions

## Optimize Data Source

### Microsoft Azure Sponsorship

Subscription Cost: \$920.35

SERVICE NAME	SERVICE RESOURCE	SPEND
Storage	Hot GRS Iterative Read Operations	\$387.44
Azure Databricks	Premium All-Purpose Photon DBU	\$374.16
Virtual Machines	D4ds v5	\$102.32
Storage	P15 LRS Disk	\$22.65
Storage	Hot GRS Write Operations	\$19.56
IoT Hub	S1 Unit	\$4.84

(14-06-2024 to 16-06-2024)

# 4. Solutions

## *Optimize Data Source*

- What do your queries cost?
- What techniques are used?
- Can they be more efficient?





---

# Azure cost calculator

West Europe

VM: D4ds\_v5

0.27 VM/hour

West Europe

All purpose compute (Photon)

Premium Workspace

0.55 DBU/hour

DBU

2/VM

Number of VM's

(workers + driver)

7

Hours

48

---

## Total Cost

VM

\$90.72

DBU

\$369.60

Data Source

\$407.00

**Total**

**\$867.32**

---

# Azure cost calculator

West Europe

VM: D4ds\_v5

0.27 VM/hour

West Europe

All purpose compute (Photon)

Premium Workspace

0.55 DBU/hour

DBU

2/VM

Number of VM's  
(workers + driver)

7

Hours

48

## Total Cost

VM

\$90.72

DBU

\$369.60

Data Source

\$3.11

**Total**

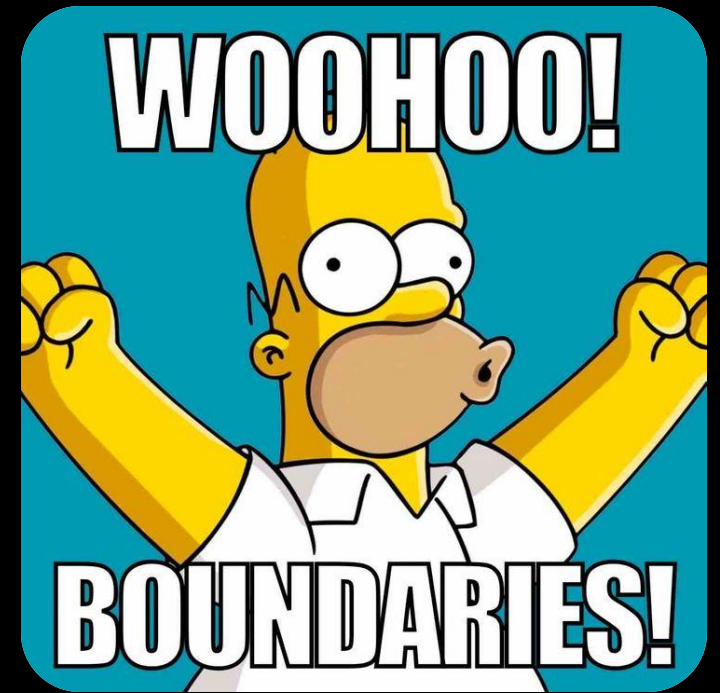
**\$463.43**

47%



# 4. Solutions

## *(2) Cluster Settings*



# 4. Solutions

## Cluster Settings

### Microsoft Azure Sponsorship

Subscription Cost: \$920.35

SERVICE NAME	SERVICE RESOURCE	SPEND
Storage	Hot GRS Iterative Read Operations	\$387.44
Azure Databricks	Premium All-Purpose Photon DBU	\$374.16
Virtual Machines	D4ds v5	\$102.32
Storage	P15 LRS Disk	\$22.65
Storage	Hot GRS Write Operations	\$19.56
IoT Hub	S1 Unit	\$4.84

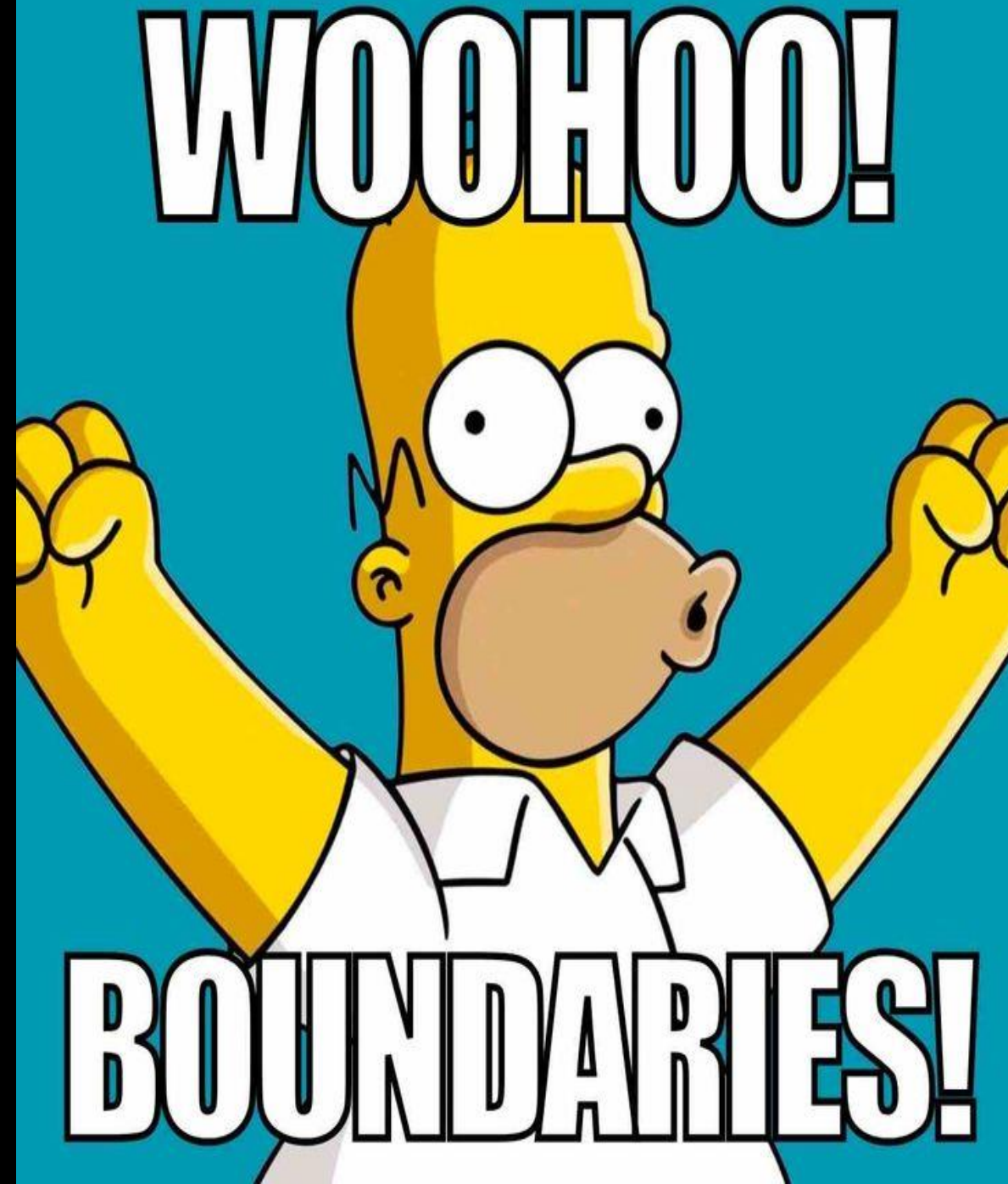
(14-06-2024 to 16-06-2024)

## 4. Solutions

### *Cluster Settings*

#### Change DBU

- Photon
- Number of workers (VM's)
- Worker/driver Type

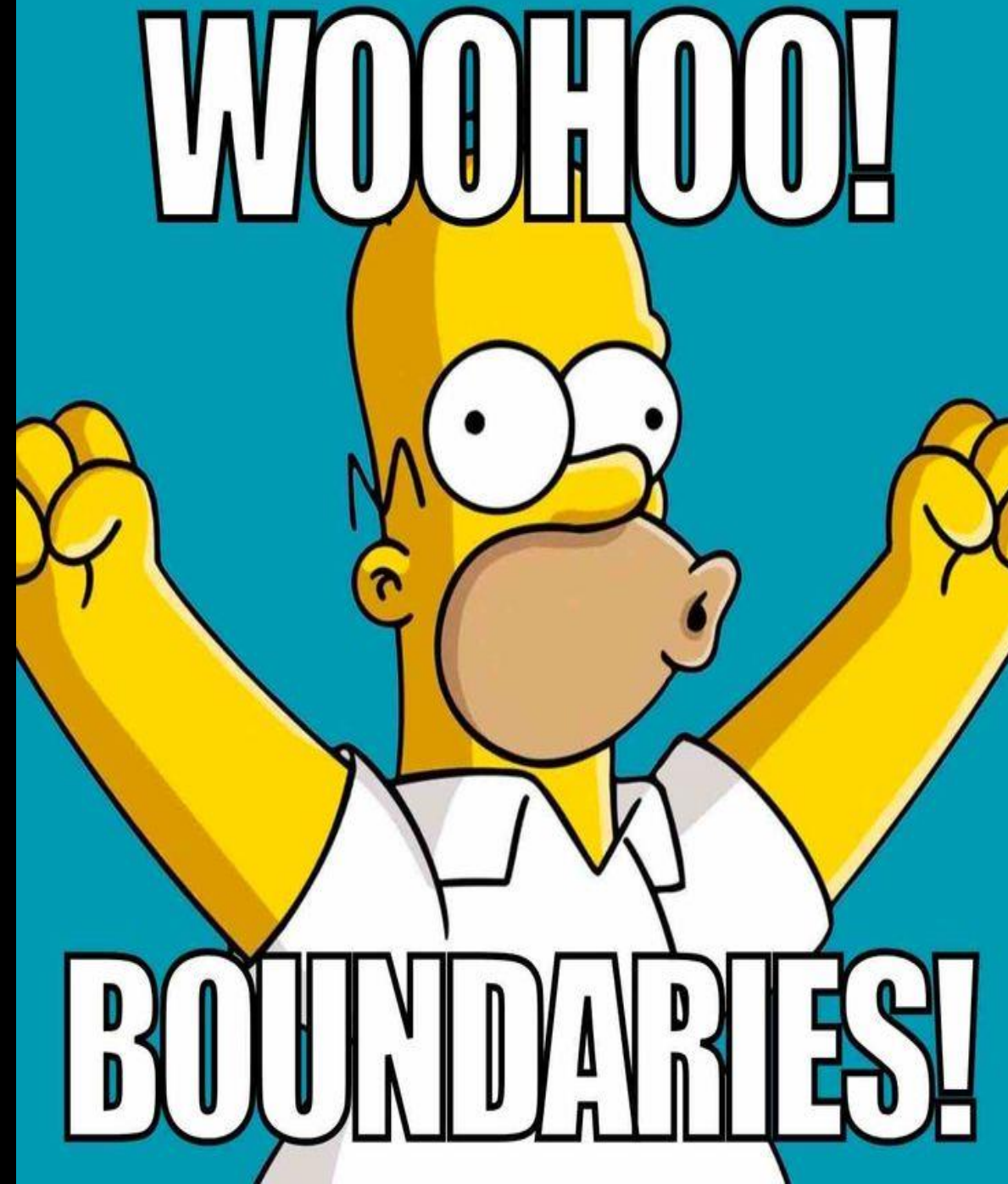


## 4. Solutions

### *Cluster Settings*

#### Decrease cluster time

- Auto terminate
- Spark version



# 4. Solutions

## *Cluster Settings*

### Spot instances

- Might decrease price
- Might make workloads
  - Unstable
  - Run longer
- Not for driver nodes





# Azure cost calculator

West Europe

VM: D4ds\_v5

0.27 VM/hour

West Europe

All purpose compute (Photon)

Premium Workspace

0.55 DBU/hour

DBU

2/VM

Number of VM's

(workers + driver)

7

Hours

48

## Total Cost

VM

\$90.72

DBU

\$369.60

Data Source

\$3.11

**Total**

**\$463.43**



# Azure cost calculator

West Europe

VM: D4s\_v3

0.24 VM/hour

West Europe

All purpose compute (no photon)

Premium Workspace

0.55 DBU/hour

DBU

0.75/VM

Number of VM's

(workers + driver)

2 - 4

Hours

48

## Total Cost

VM

\$23.04 - \$46.08

DBU

\$39.60 – \$79.20

Data Source

\$3.11

**Total**

**\$65.75 – \$128.39**

72%



# 4. Solutions

## *(3) Code Optimization*



# 4. Solutions

## *Code Optimization*

- The most expensive resource? It's-a me!
  - *At some point, you should stop*



# 4. Solutions

## *Code Optimization*

### When?

- Upgrade Apache Spark
- Change UDF to Apache Spark native



## 4. Solutions

*(4) Make it a job!*



# 4. Solutions

*Make it a job!*

- DBU price differs per workload type
- Jobs compute < All-purpose compute
  - $\$0.30 < \$0.55$  per DBU/hour





---

# Azure cost calculator

West Europe

VM: D4s\_v3

0.24 VM/hour

West Europe

All purpose compute (no photon)

Premium Workspace

0.55 DBU/hour

DBU

0.75/VM

Number of VM's

(workers + driver)

2 - 4

Hours

48

---

## Total Cost

VM

\$23.04 - \$46.08

DBU

\$39.60 – \$79.20

Data Source

\$3.11

**Total**

**\$65.75 – \$128.39**

---

# Azure cost calculator

West Europe

VM: D4s\_v3

0.24 VM/hour

West Europe

Job Compute

Premium Workspace

0.30 DBU/hour

DBU

0.75/VM

Number of VM's  
(workers + driver)

2 - 4

Hours

48

## Total Cost

VM

\$23.04 - \$46.08

DBU

\$21.60 – \$43.20

Data Source

\$3.11

**Total**

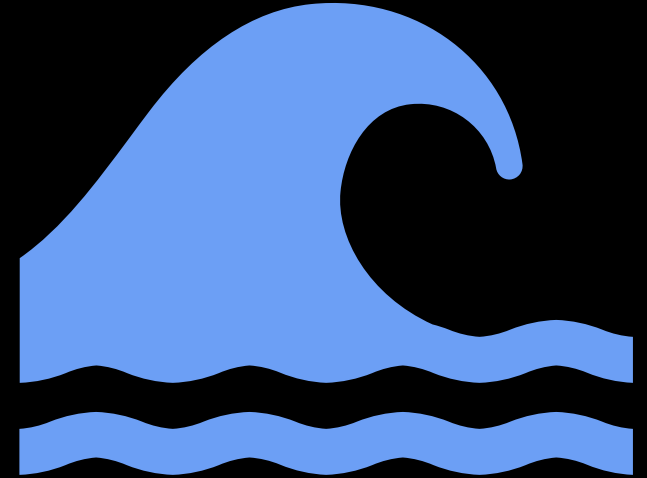
**\$47.75 – \$92.39**

28%



## 4. Solutions

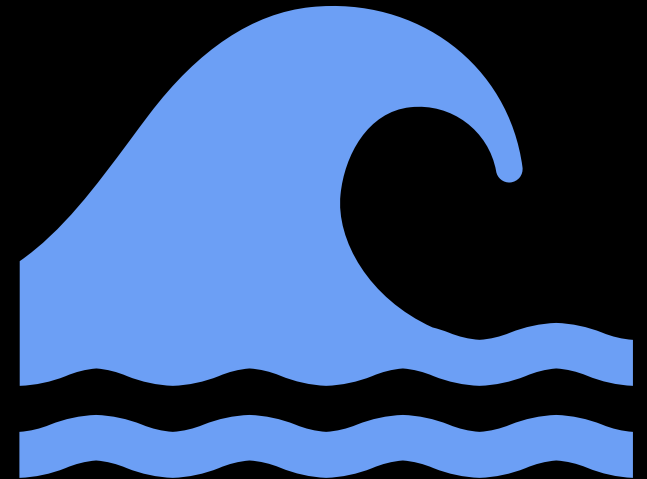
*(5) Stream or Micro Batch?*



# 4. Solutions

*Stream or Micro Batch?*

**How real time do you need it to be?**



---

# Azure cost calculator

West Europe

VM: D4s\_v3

0.24 VM/hour

West Europe

Job Compute

Premium Workspace

0.30 DBU/hour

DBU

0.75/VM

Number of VM's

(workers + driver)

2 - 4

Hours

48

---

## Total Cost

VM

\$23.04 - \$46.08

DBU

\$21.60 – \$43.20

Data Source

\$3.11

**Total**

**\$47.75 – \$92.39**

---

# Azure cost calculator

West Europe

VM: D4s\_v3

0.24 VM/hour

West Europe

Job Compute

Premium Workspace

0.30 DBU/hour

DBU

0.75/VM

Number of VM's  
(workers + driver)

2 - 4

Hours

24

## Total Cost

VM

\$11.52 - \$23.04

DBU

\$10.80 – \$21.60

Data Source

\$3.11

**Total**

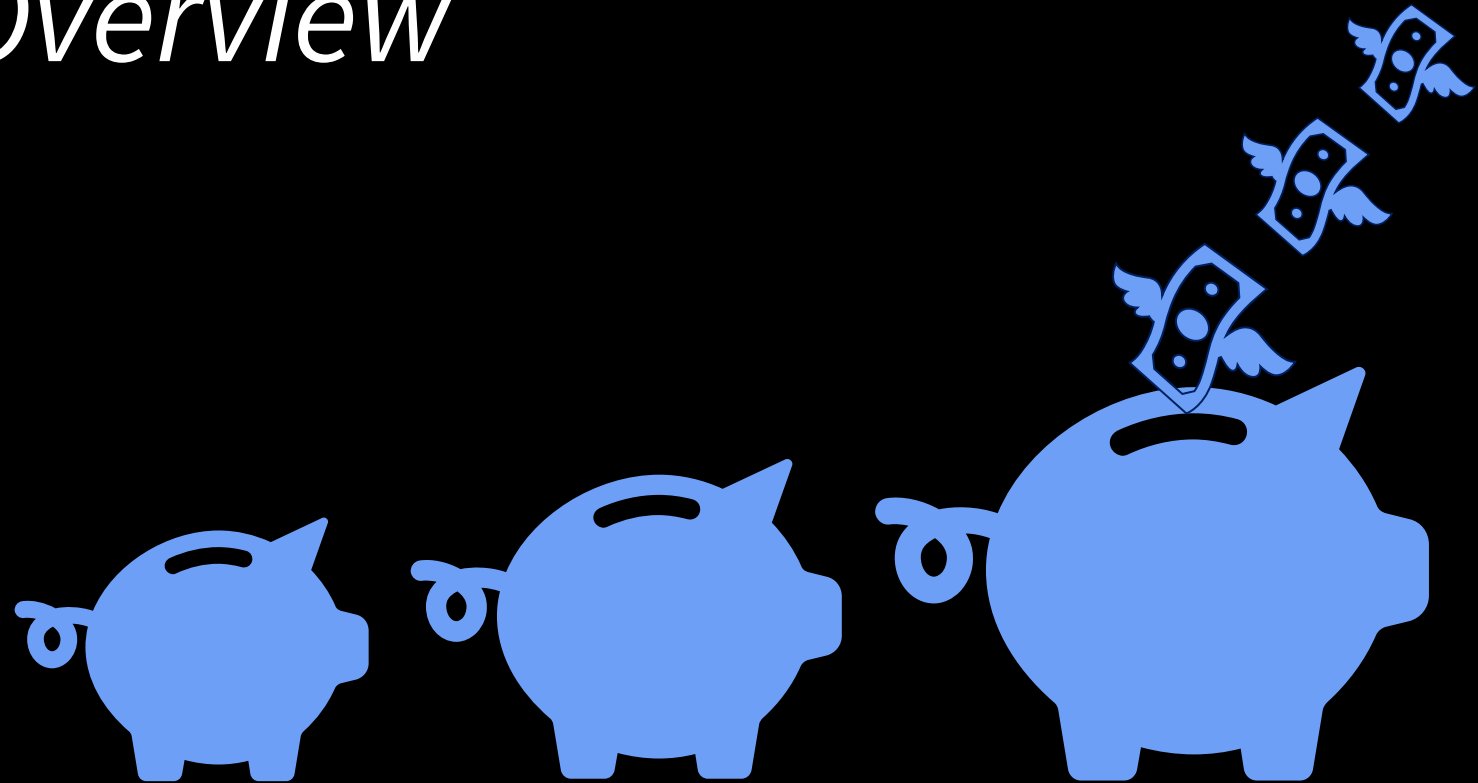
**\$25.43 – \$47.75**

48%



# 4. Solutions

## *Overview*



# 4. Solutions

Overview of cost savings



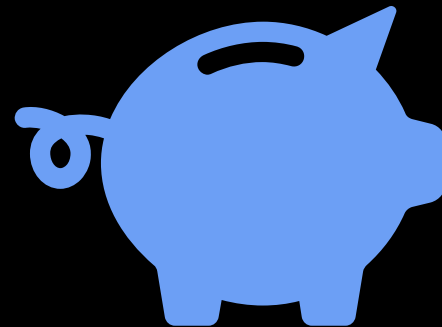
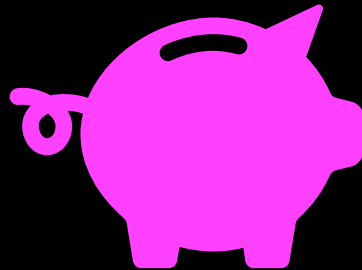
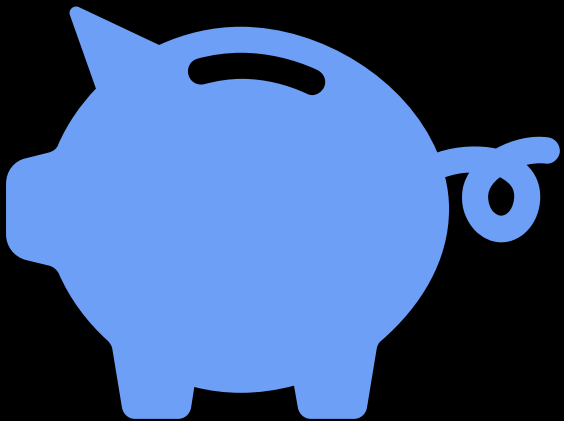


## 4. Solution

Overview of c

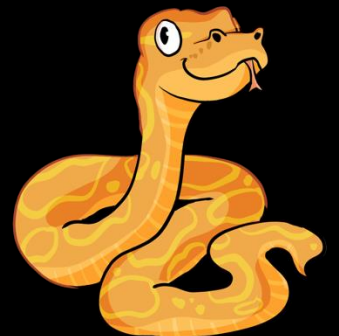


**I could have  
saved >\$800**



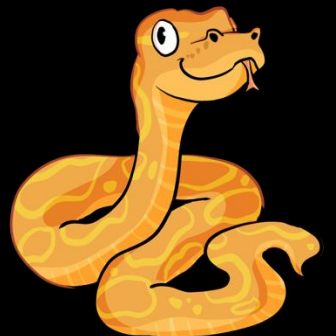
# 5. Conclusion

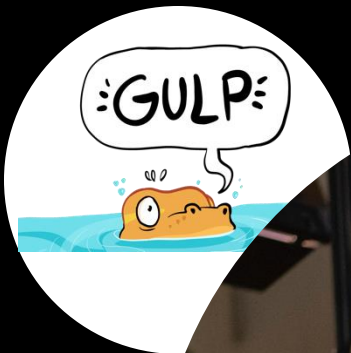
- Total Azure spend
  - DBU
  - VM
  - Data Sources
  - Other resources
- Monitoring
  - Alerts: At least in Azure Portal
  - Usage Dashboard: for optimizing workloads



# 5. Conclusion

- Don't forget your Data Sources
- Optimize your cluster settings
- Avoid unnecessary gold-plating of code
- Job clusters are cheaper than General Purpose Compute
- Streaming is expensive





# Lisa

# Hoving

*Big Data Engineer*



Rate my session



Github