Some tips and tricks

Data Engineer
@Rabobank

Lisa Hoving

THANK YOU

Data Community Austria Day 2025

Awesome Partner

Microsoft

Platinum

redgate

ACP IT for innovators.

Gold

b.telligent
smart data. smart decisions.

Lucient

Bronze

Tabular Editor

paiqo
The Platform & AI Company

Power BI Camp
www.linearis.at

Agenda

01 An expensive Python

02 Azure Databricks Pricing

03 Monitoring & Alerts

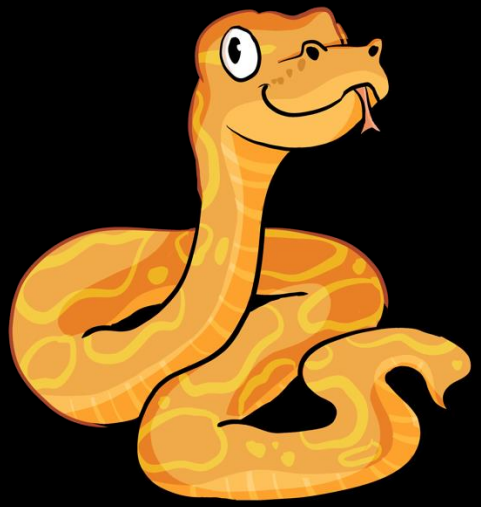
04 Solutions (1-6)

05 Conclusion

Household announcements

- A lot in preview!
  - Needs Unity Catalog
- An overview of options
- Based on my own experiences

1. An expensive python



An expensive python

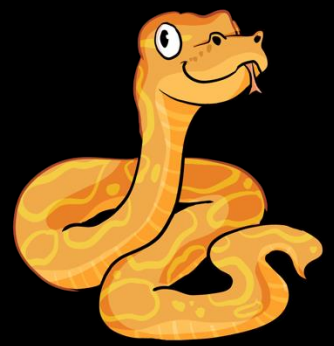
Azure spend

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

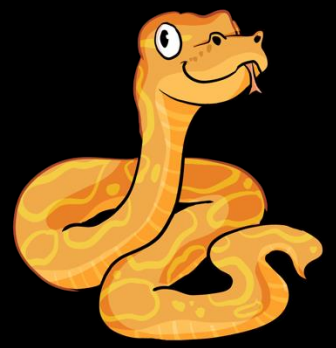
Azure spend

Microsoft Azure Sponsorship

Subscription Cost: $920.35

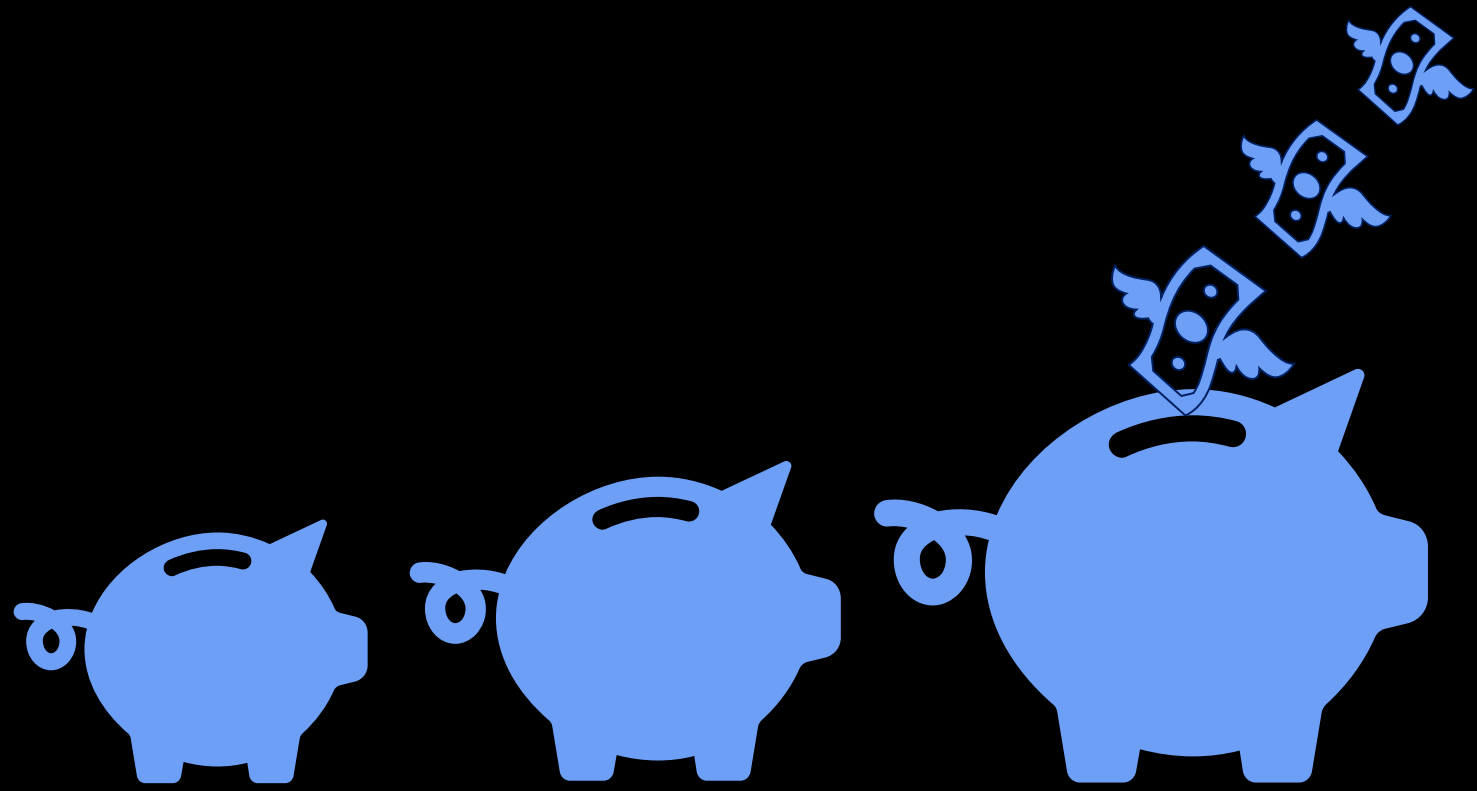
SERVICE NAME	SERVICE RESOURCE	SPEND
Storage	Hot GRS Iterative Read Operations	$387.44
Azure Databricks	Premium All-Purpose Photon DBU	$374.16
Virtual Machines	D4ds v5	$102.32
Storage	P15 LRS Disk	$22.65
Storage	Hot GRS Write Operations	$19.56
IoT Hub	S1 Unit	$4.84

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



2. Azure Databricks Pricing

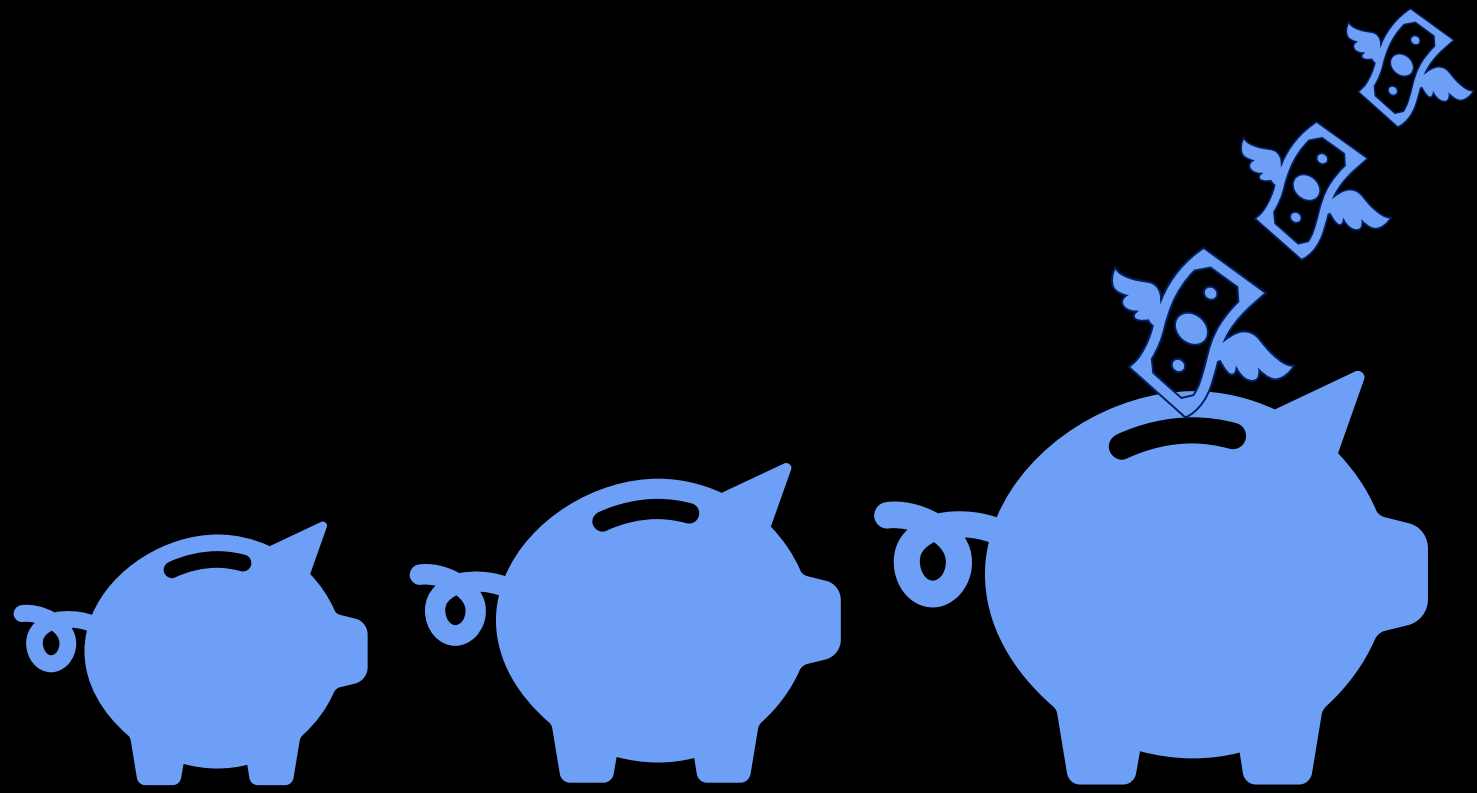
- DBU
- Virtual Machines
- Data Sources
- Other Resources



2. Azure Databricks Pricing

Databricks Unit (DBU)

- Normalized unit of processing power
  - Per hour



2. Azure Databricks Pricing

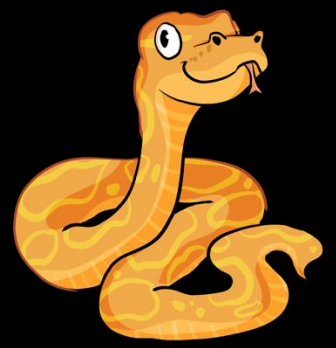
Azure spend

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

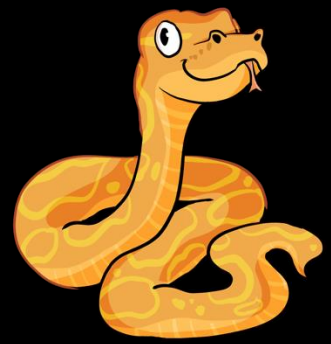
Azure spend

Microsoft Azure Sponsorship

Subscription Cost: $920.35

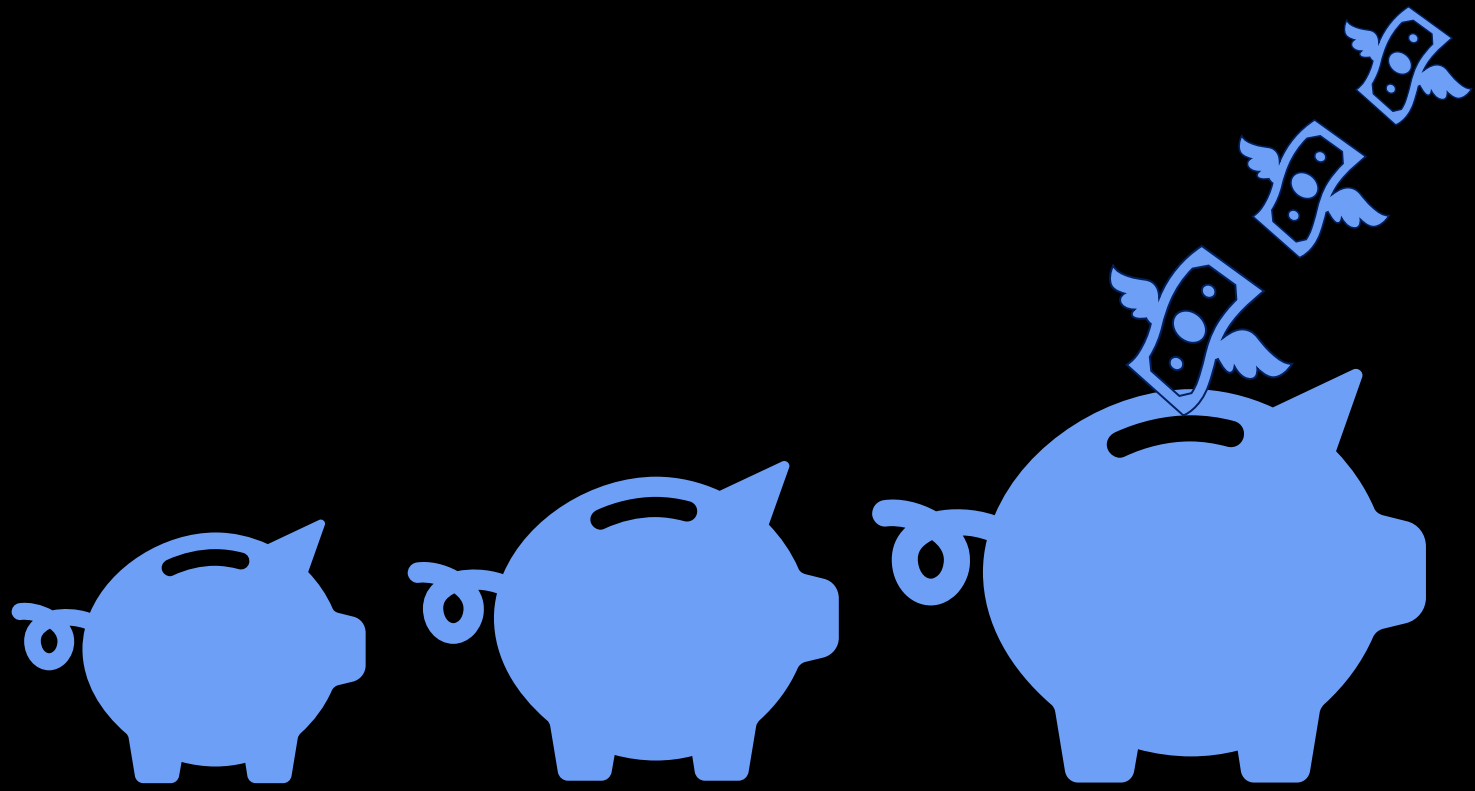
SERVICE NAME	SERVICE RESOURCE	SPEND
Storage	Hot GRS Iterative Read Operations	$387.44
Azure Databricks	Premium All-Purpose Photon DBU	$374.16
Virtual Machines	D4ds v5	$102.32
Storage	P15 LRS Disk	$22.65
Storage	Hot GRS Write Operations	$19.56
IoT Hub	S1 Unit	$4.84

(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

3 options

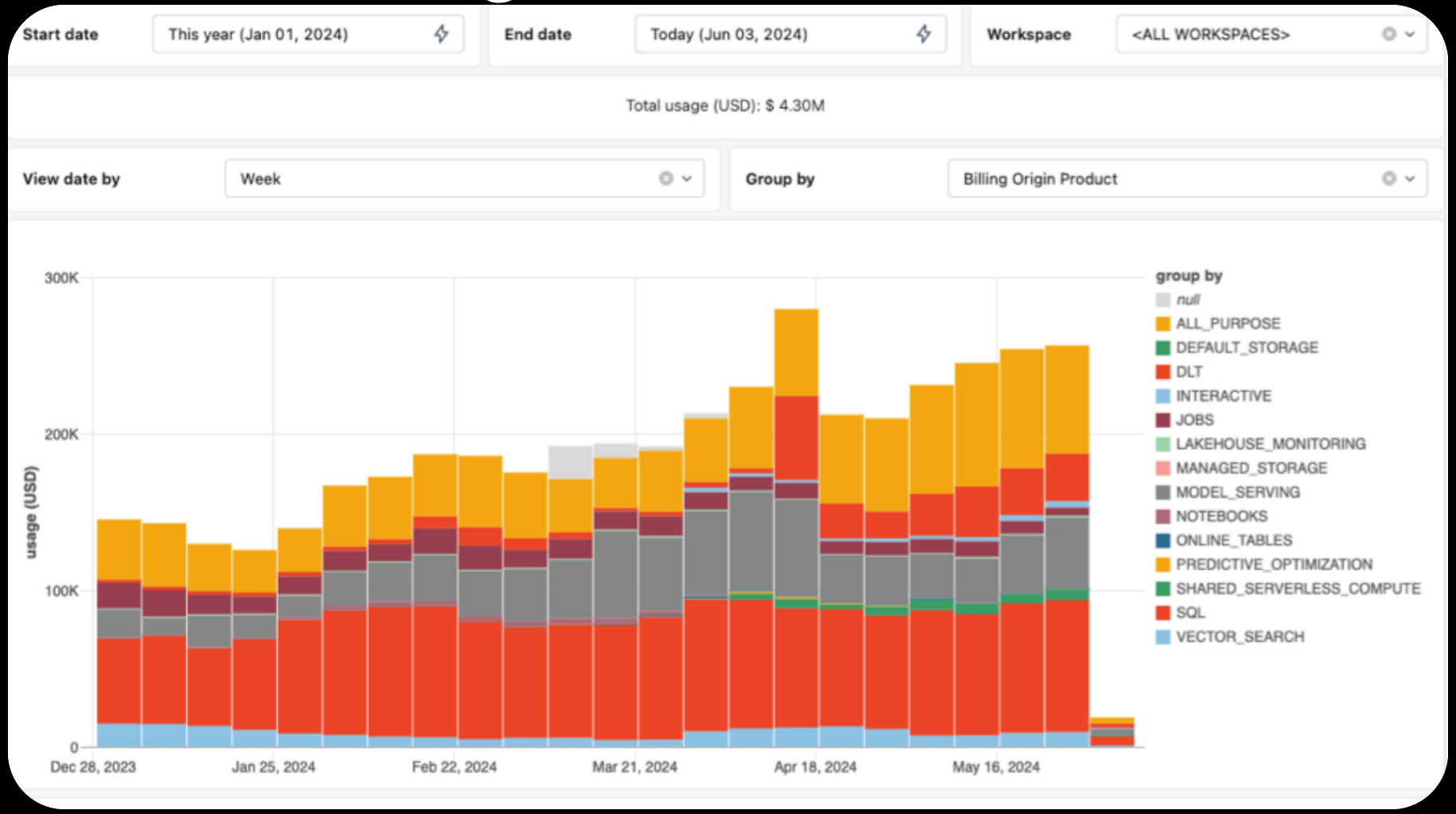
- Azure Portal
  - Budgets
  - Alerts
- Azure Databricks (Unity Catalog)
  - Budgets
  - Alerts
  - Serverless

Serverless

Workspace

3. Monitoring & Alerts

Databricks Usage Dashboard



4. Solutions

Solutions

01 Optimize Data Source

02 Optimize Code

03 Cluster Settings

04 Make it a Job!

05 Stream or Micro Batch?

06 Prepay

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

47%

Total Cost

VM	$90.72
DBU	$369.60
Data Source	$3.11
Total	**$463.43**

4. Solutions

(2) Code Optimization

4. Solutions

Code Optimization

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

4. Solutions

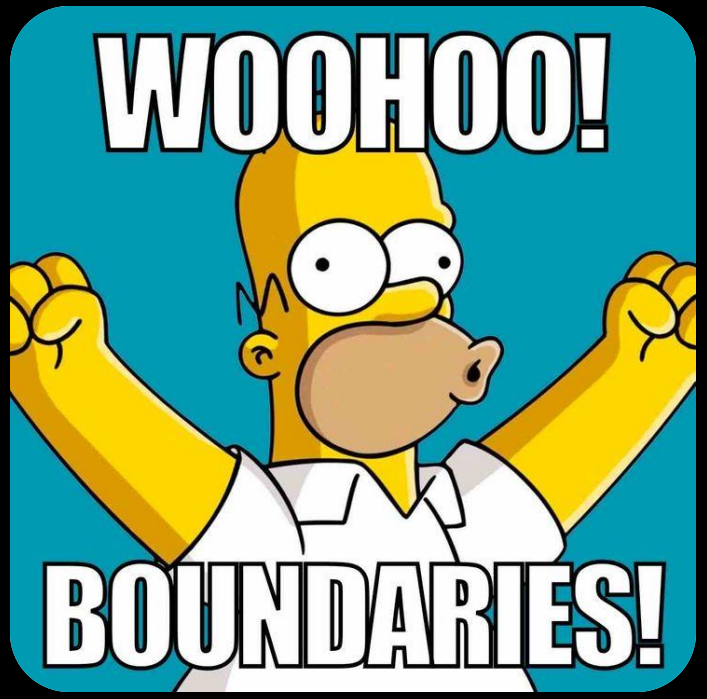
Code Optimization

When?

- Upgrade Apache Spark
- File format != Parquet or Delta Lake
- Change UDF to Apache Spark native
- Data is inefficiently partitioned
    - High shuffle
    - Disk spill

4. Solutions

(3) 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

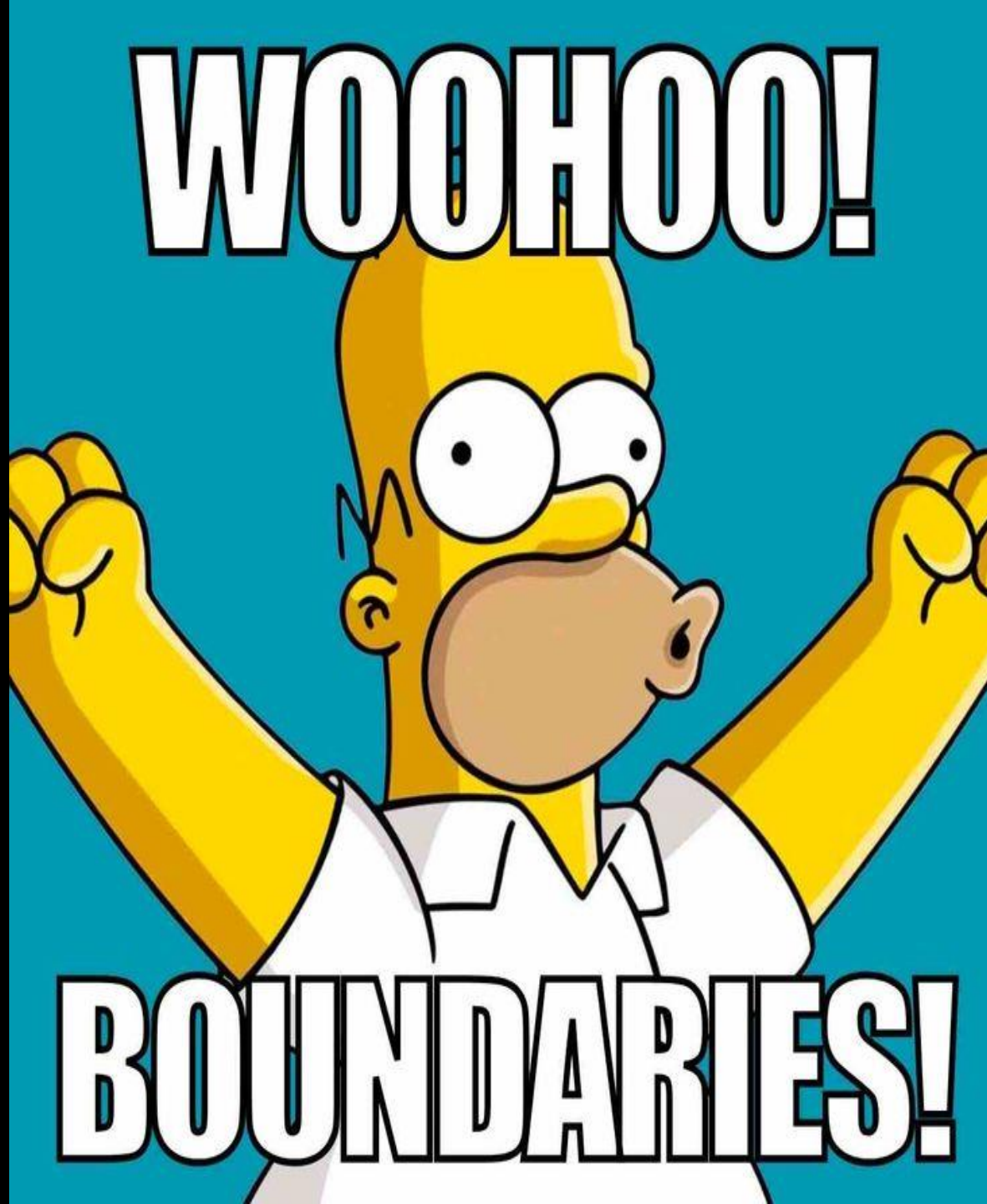
(14-06-2024 to 16-06-2024)

4. Solutions

Cluster Settings

Change DBU

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

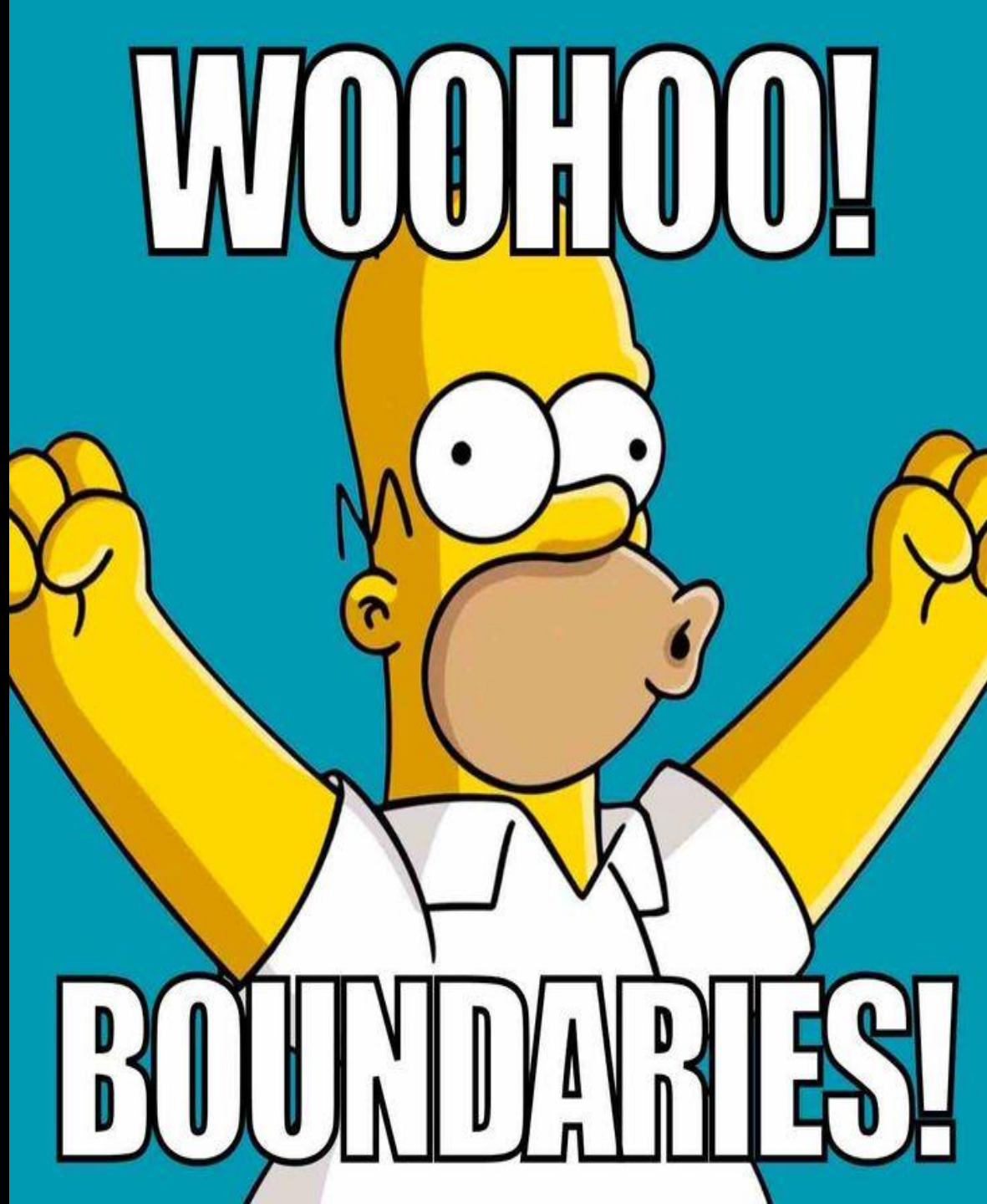


4. Solutions

Cluster Settings

Decrease cluster time

- Auto Terminate
- Spark version



4. Solutions

Cluster Settings

Spot instances

- Might decrease price
- 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: D4ds_v5	0.27 VM/hour
West Europe All purpose compute (no photon) Premium Workspace	0.55 DBU/hour
DBU	1/VM
Number of VM's (workers + driver)	2 - 4
Hours	48

65%

Total Cost

VM	$25.92 - $51.84
DBU	$52.80 – $105.60
Data Source	$3.11
Total	**$81.83 – $160.55**

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: D4ds_v5	0.27 VM/hour
West Europe All purpose compute (no photon) Premium Workspace	0.55 DBU/hour
DBU	1/VM
Number of VM's (workers + driver)	2 - 4
Hours	48

Total Cost

VM	$25.92 - $51.84
DBU	$52.80 – $105.60
Data Source	$3.11
Total	**$81.83 – $160.55**

Azure cost calculator

West Europe VM: D4ds_v5	0.27 VM/hour
West Europe Job Compute Premium Workspace	0.30 DBU/hour
DBU	1/VM
Number of VM's (workers + driver)	2 - 4
Hours	48

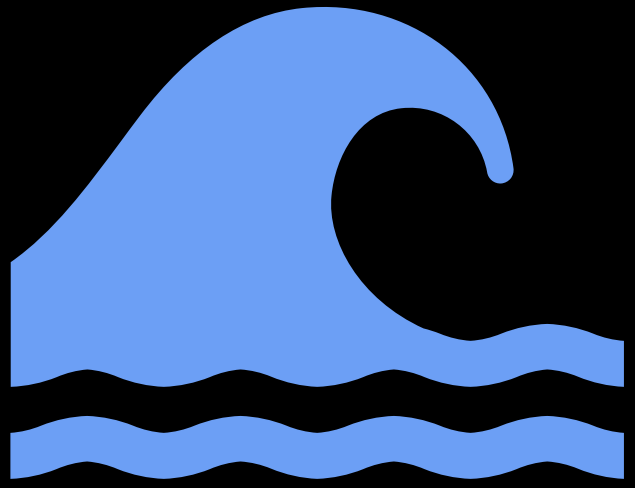
30%

Total Cost

VM	$25.92 - $51.84
DBU	$28.80 – $57.60
Data Source	$3.11
Total	**$57.83 – $112.55**

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: D4ds_v5	0.27 VM/hour
West Europe Job Compute Premium Workspace	0.30 DBU/hour
DBU	1/VM
Number of VM's (workers + driver)	2 - 4
Hours	48

Total Cost

VM	$25.92 - $51.84
DBU	$28.80 – $57.60
Data Source	$3.11
Total	**$57.83 – $112.55**

Azure cost calculator

West Europe VM: D4ds_v5	0.27 VM/hour
West Europe Job Compute Premium Workspace	0.30 DBU/hour
DBU	1/VM
Number of VM's (workers + driver)	2 - 4
Hours	21

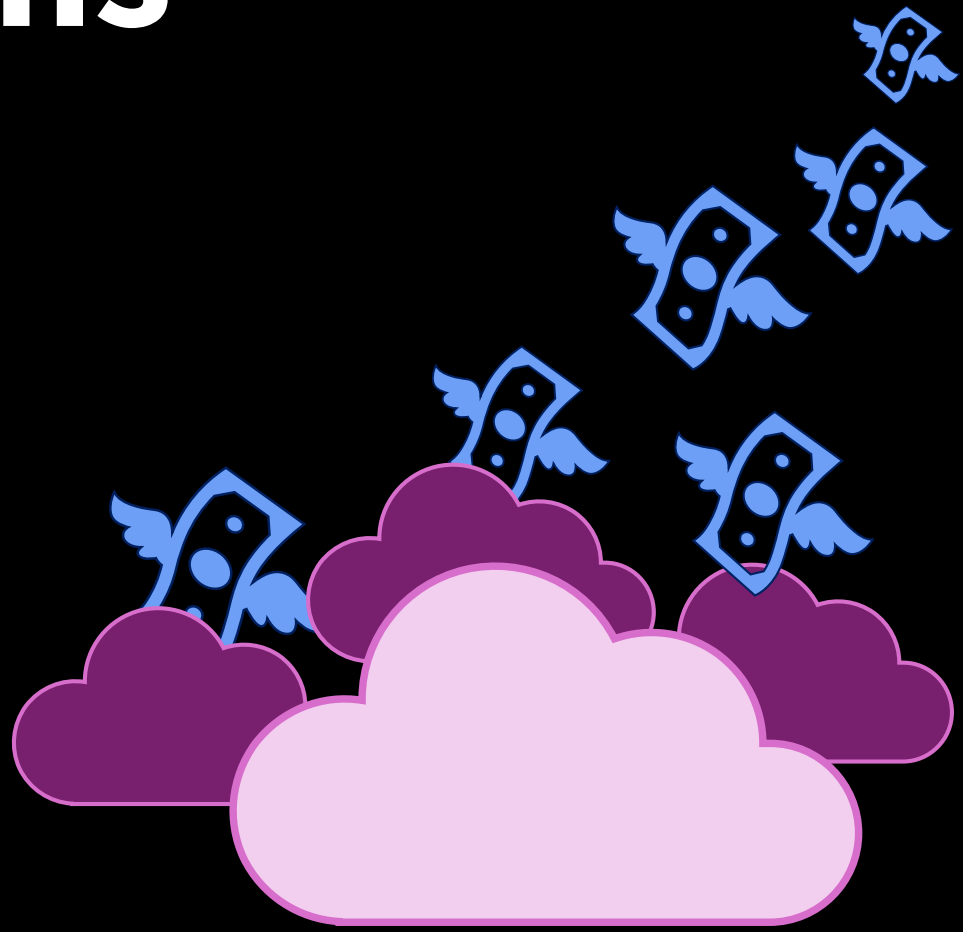
55%

Total Cost

VM	$11.34 - $22.68
DBU	$12.60 – $25.20
Data Source	$3.11
Total	**$27.05 – $50.99**

4. Solutions

(6) Prepay



4. Solutions

Prepay

VM’s

- Reserved instances
- Savings plan

I don’t have clusters running 24/7

4. Solutions

Prepay

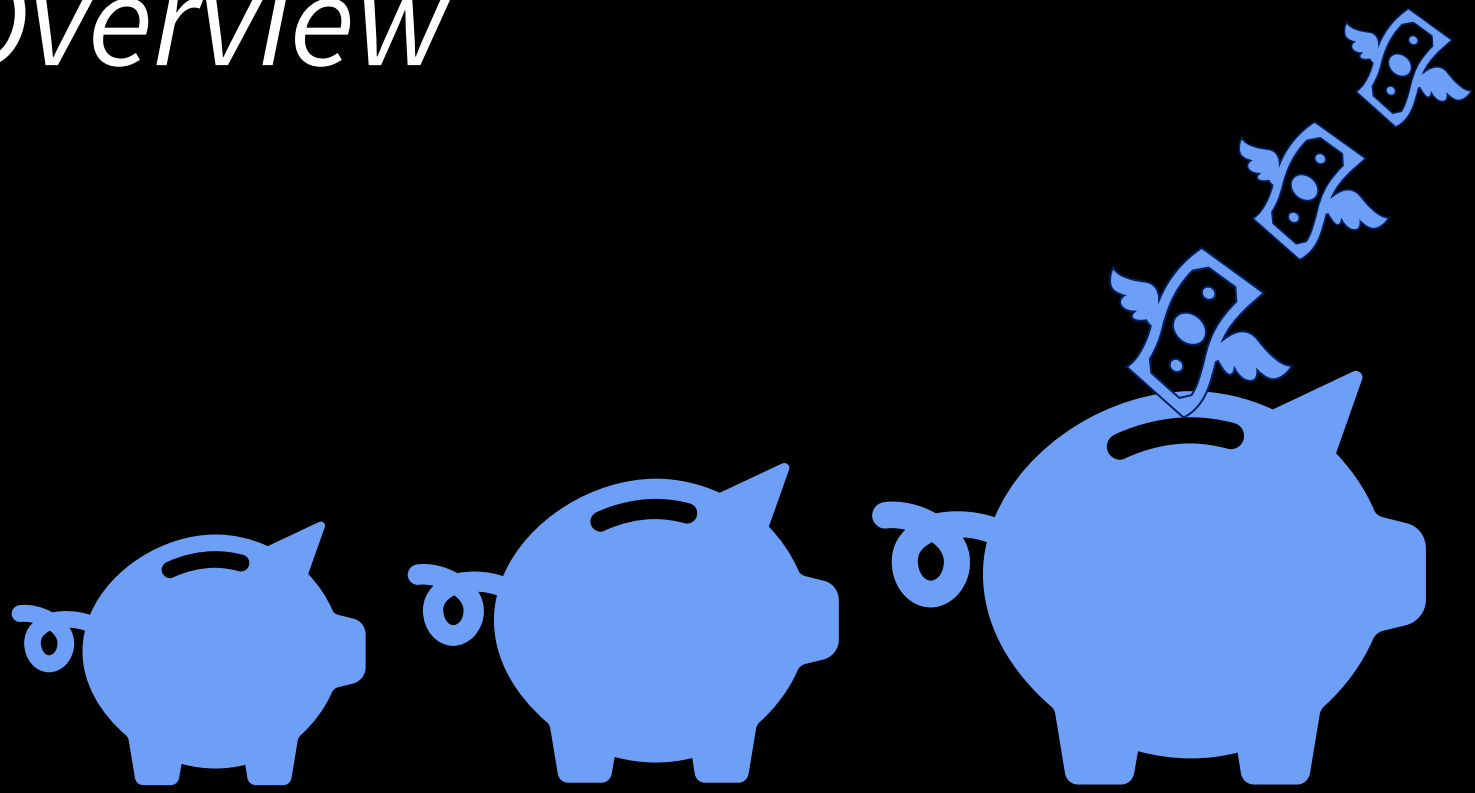
DBCU

- Savings depends DBCU
  - Starts at 12,500 DCBU
  - 1 Year: 4 - 33 %
  - 3 Years: 6 - 37%

I don’t have enough DBU usage

4. Solutions

Overview

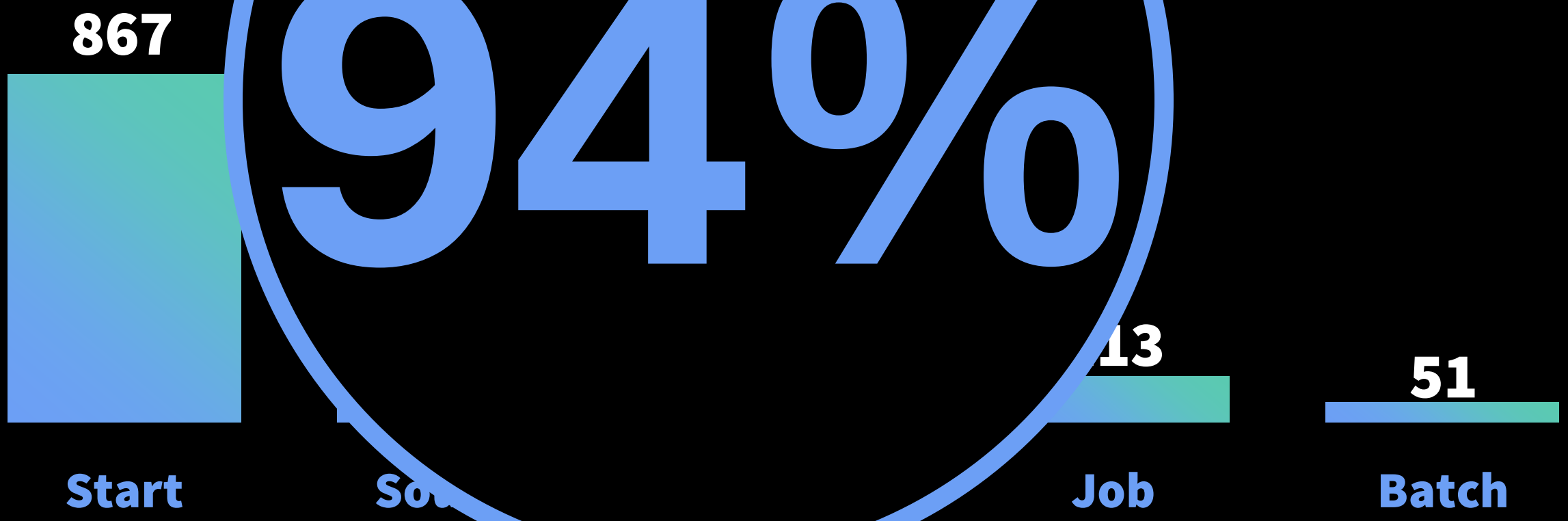


4. Solutions

Overview of cost savings

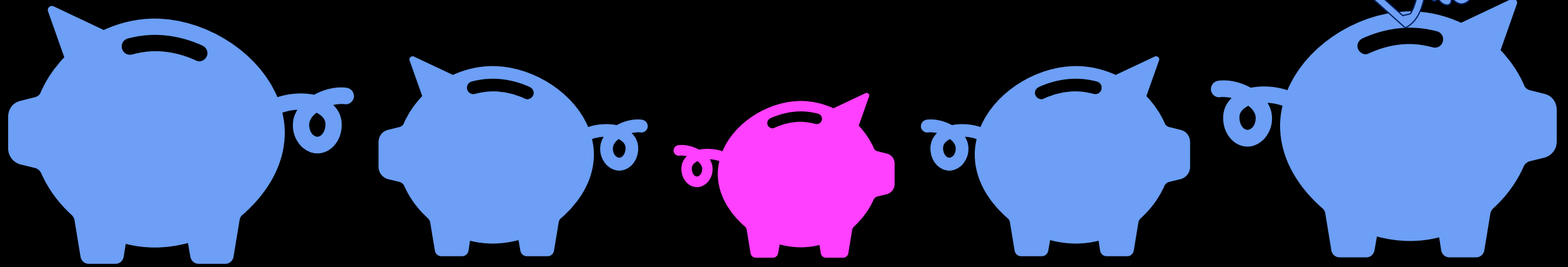
4. Solutio

Overview of c



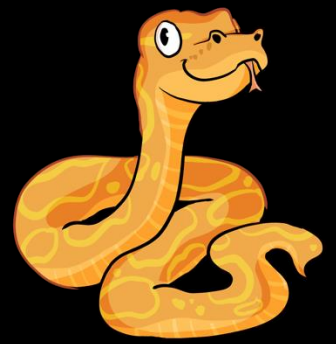
94%

I could have saved >$800



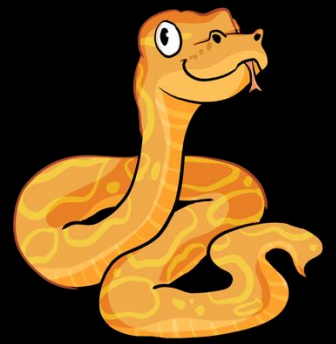
5. Conclusion

- 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
- Prepay if you have a higher usage
- Streaming is expensive



Lisa Hoving

Data Engineer
@Rabobank

Review

Documentation
&
Slides