Some tips and tricks

Big Data Engineer

Lisa Hoving

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)

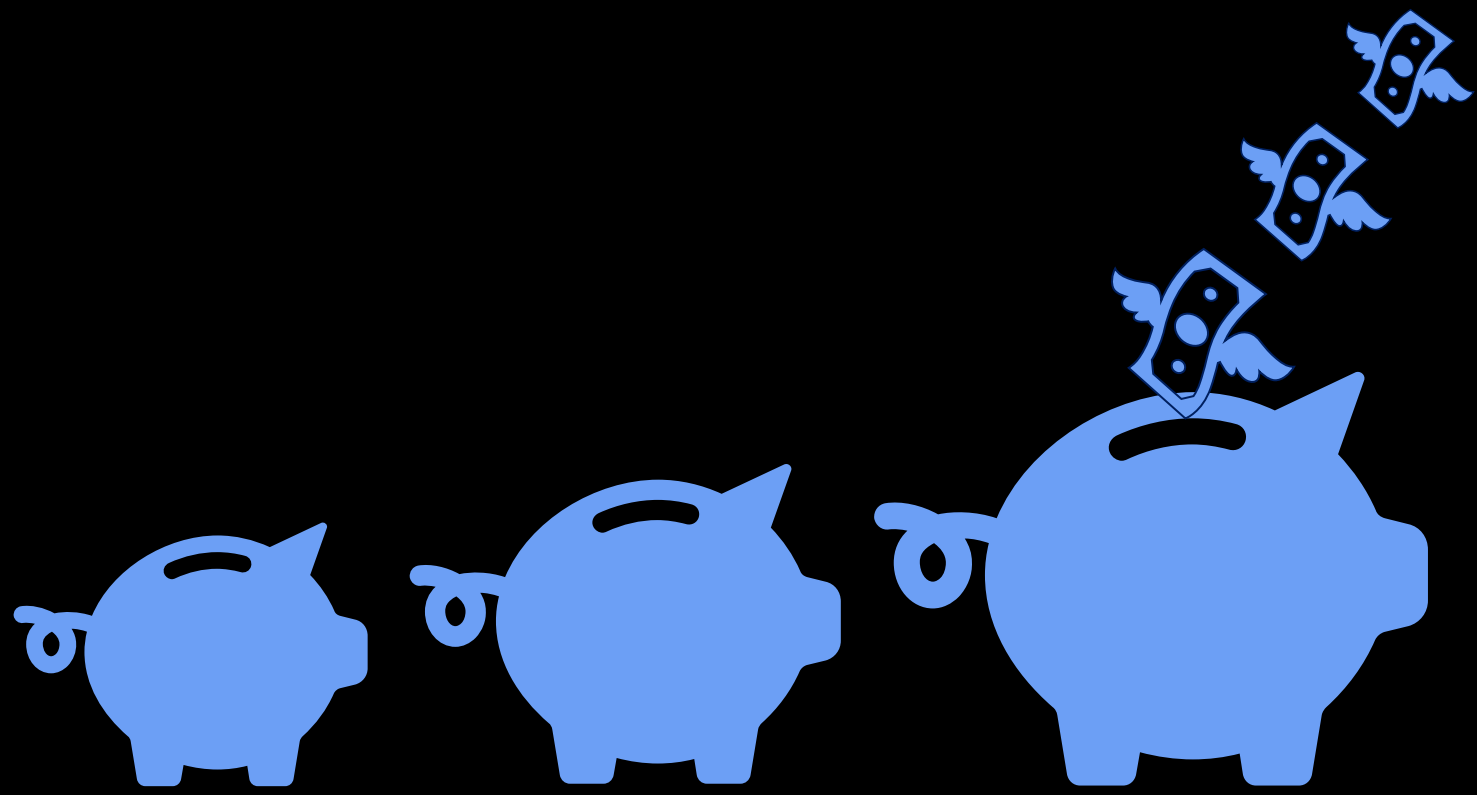
05 Conclusion

2. Azure Databricks Pricing

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

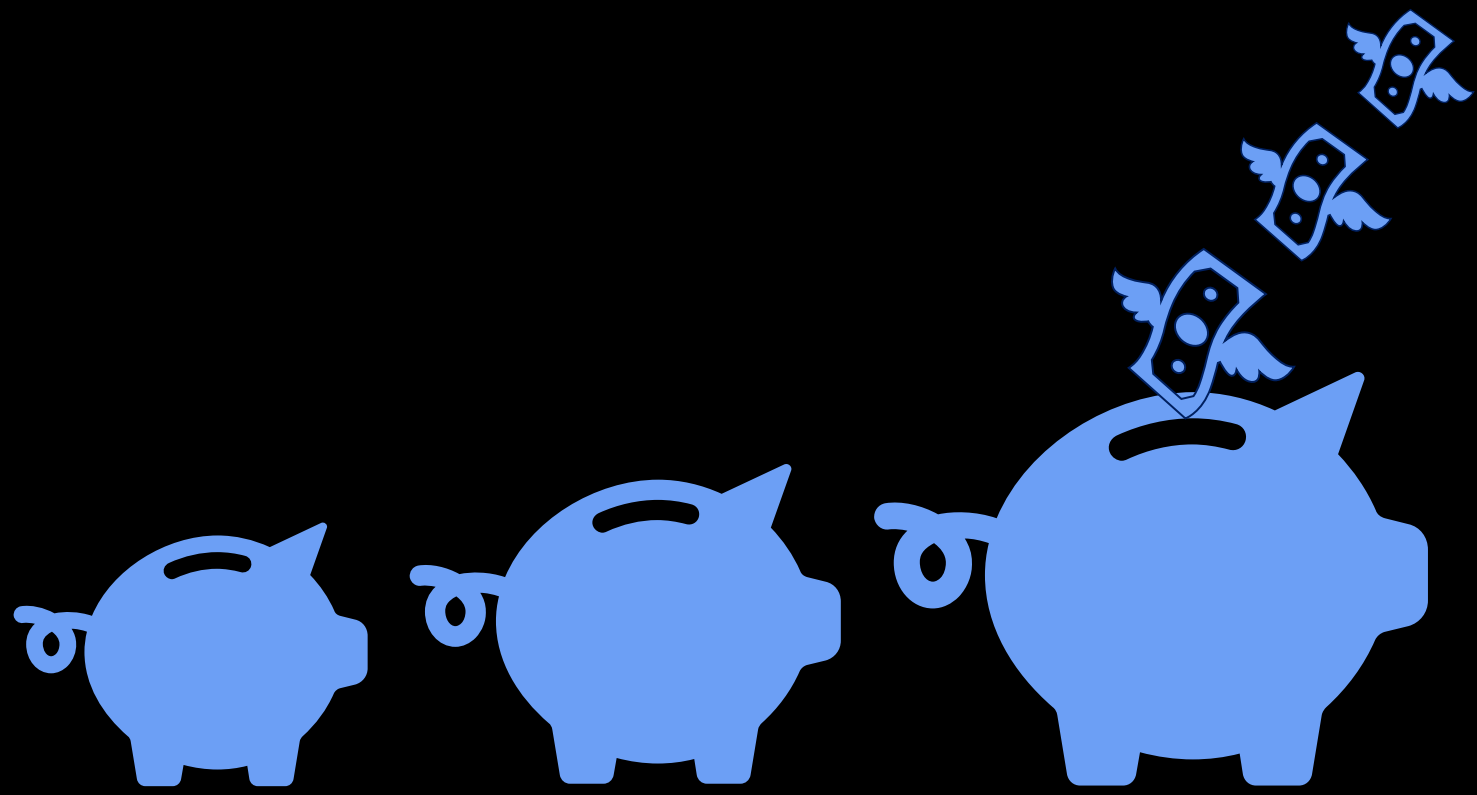
JSON View

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...

Launch Workspace

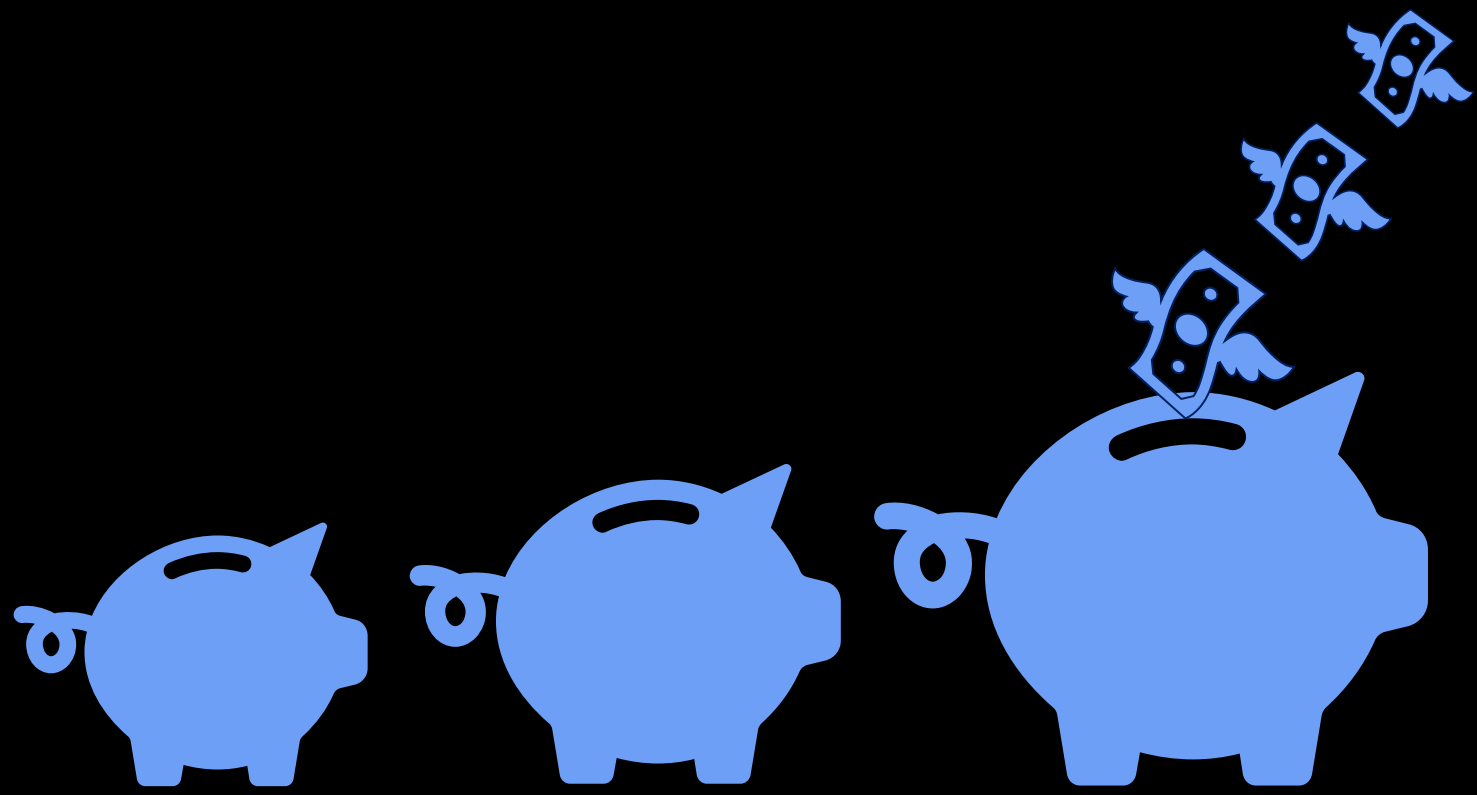
2. Azure Databricks Pricing

- DBU
- Virtual Machines



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

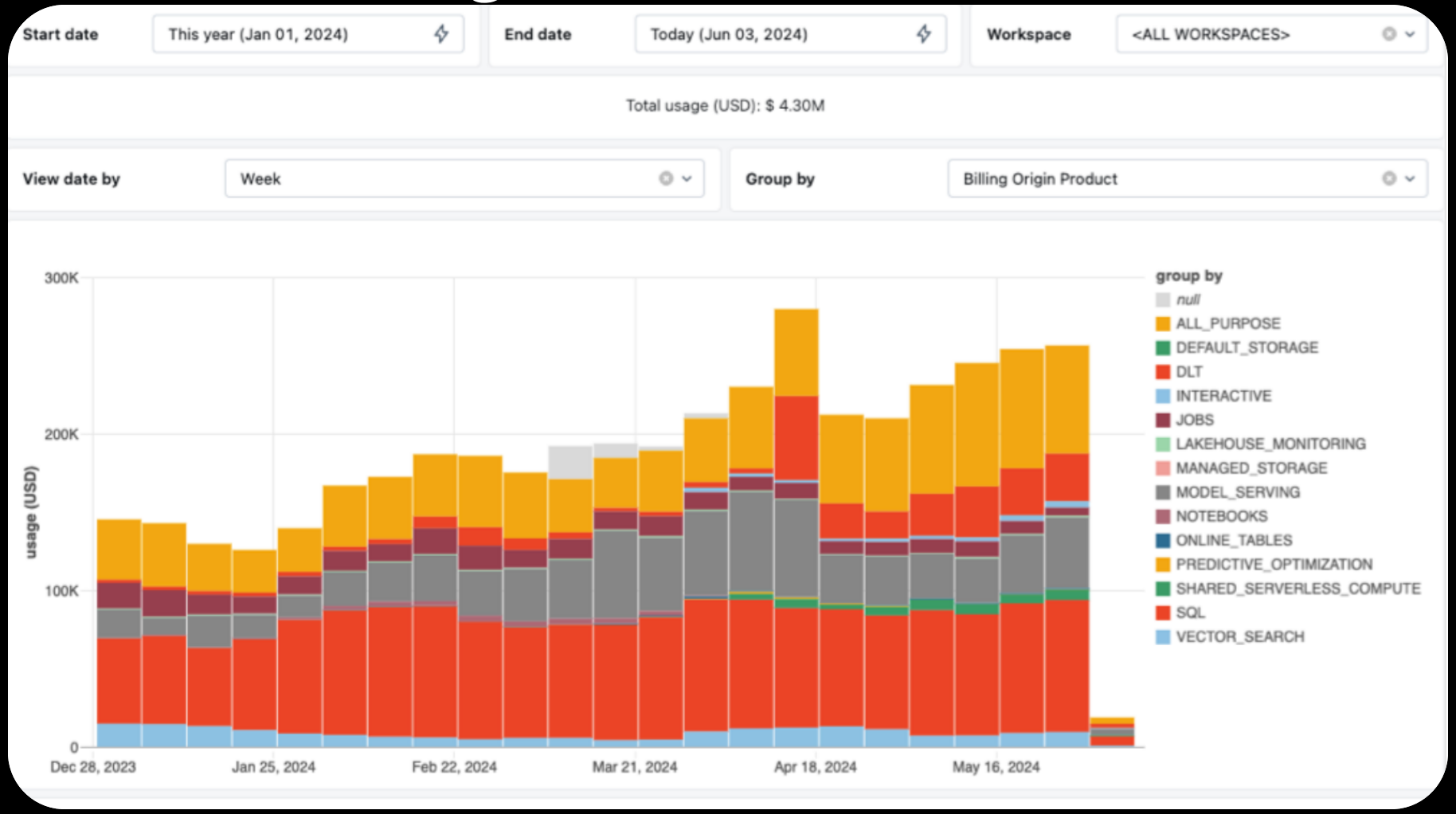
- Azure Databricks (Unity Catalog in preview)
  - Budgets
  - Alerts
  - Serverless
- Azure Portal
  - Budgets
  - Alerts

Serverless

Workspace

3. Monitoring & Alerts

Databricks Usage Dashboard



4. Solutions

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

- 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

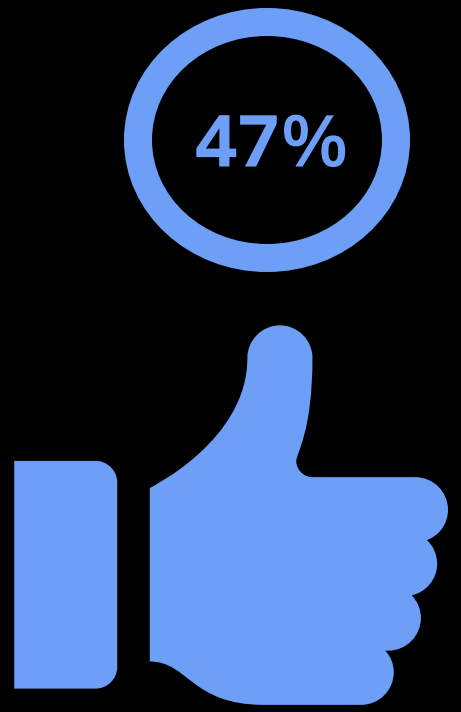
VM	$90.72
DBU	$369.60
Data Source	$407.00
Total	**$867.32**

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

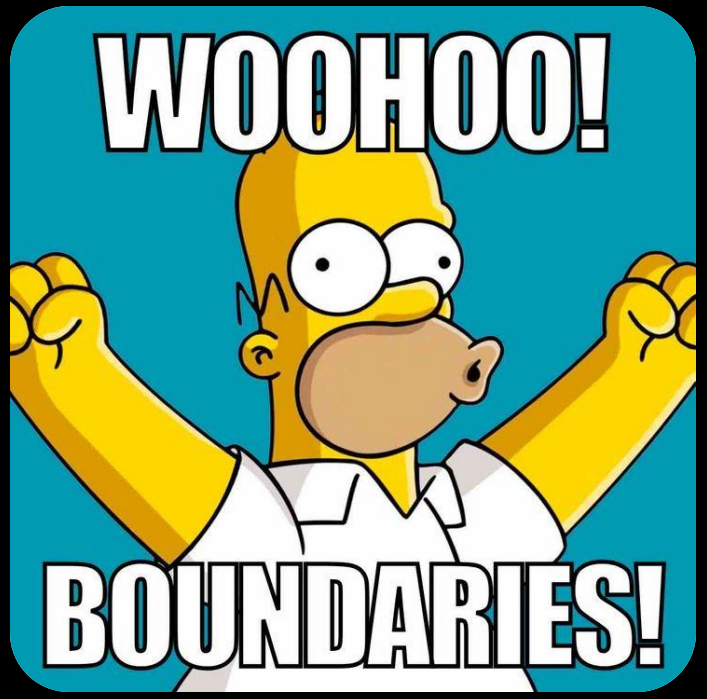
Total Cost

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



4. Solutions

(2) Cluster Settings

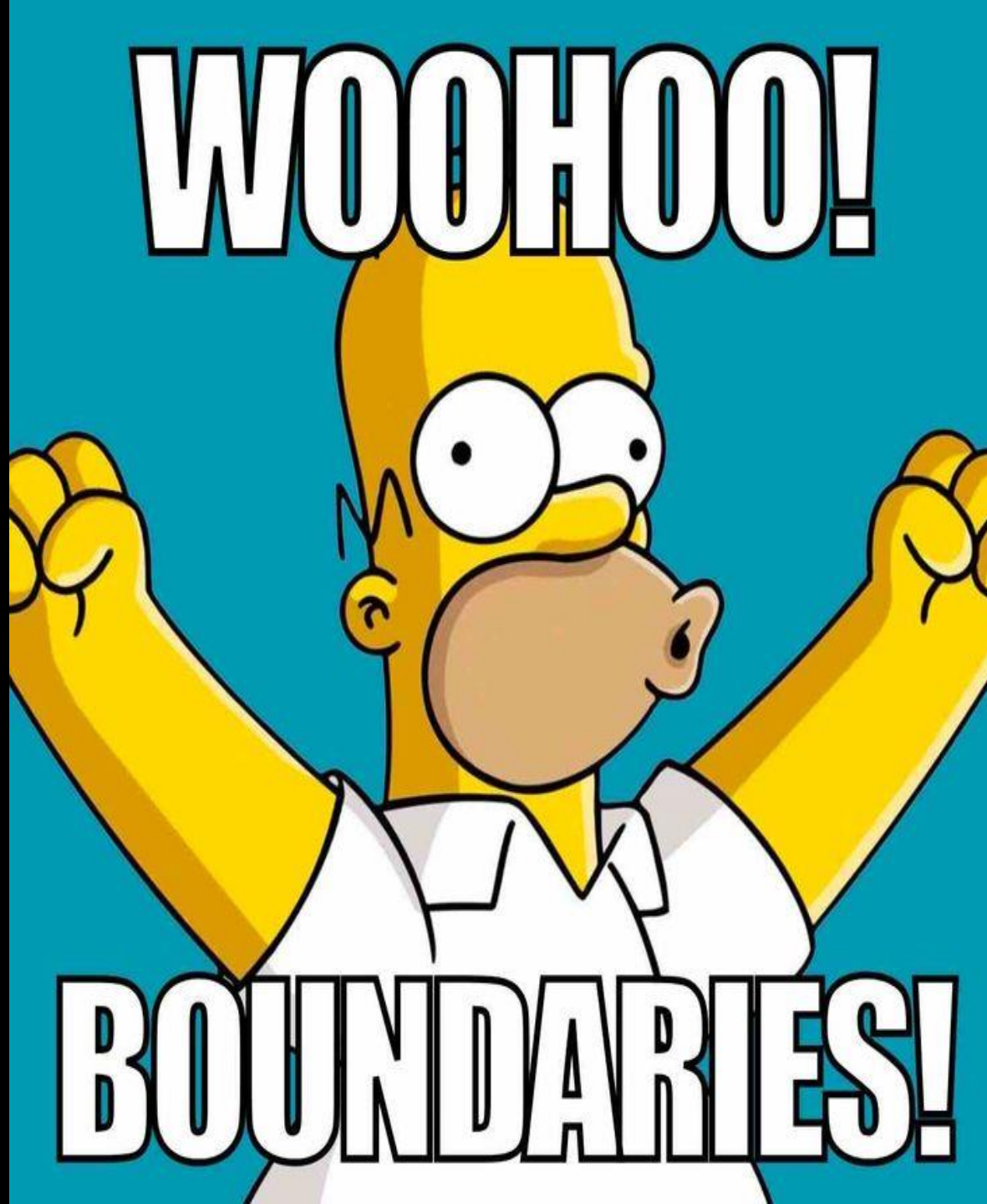


4. Solutions

Cluster Settings

Change DBU

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

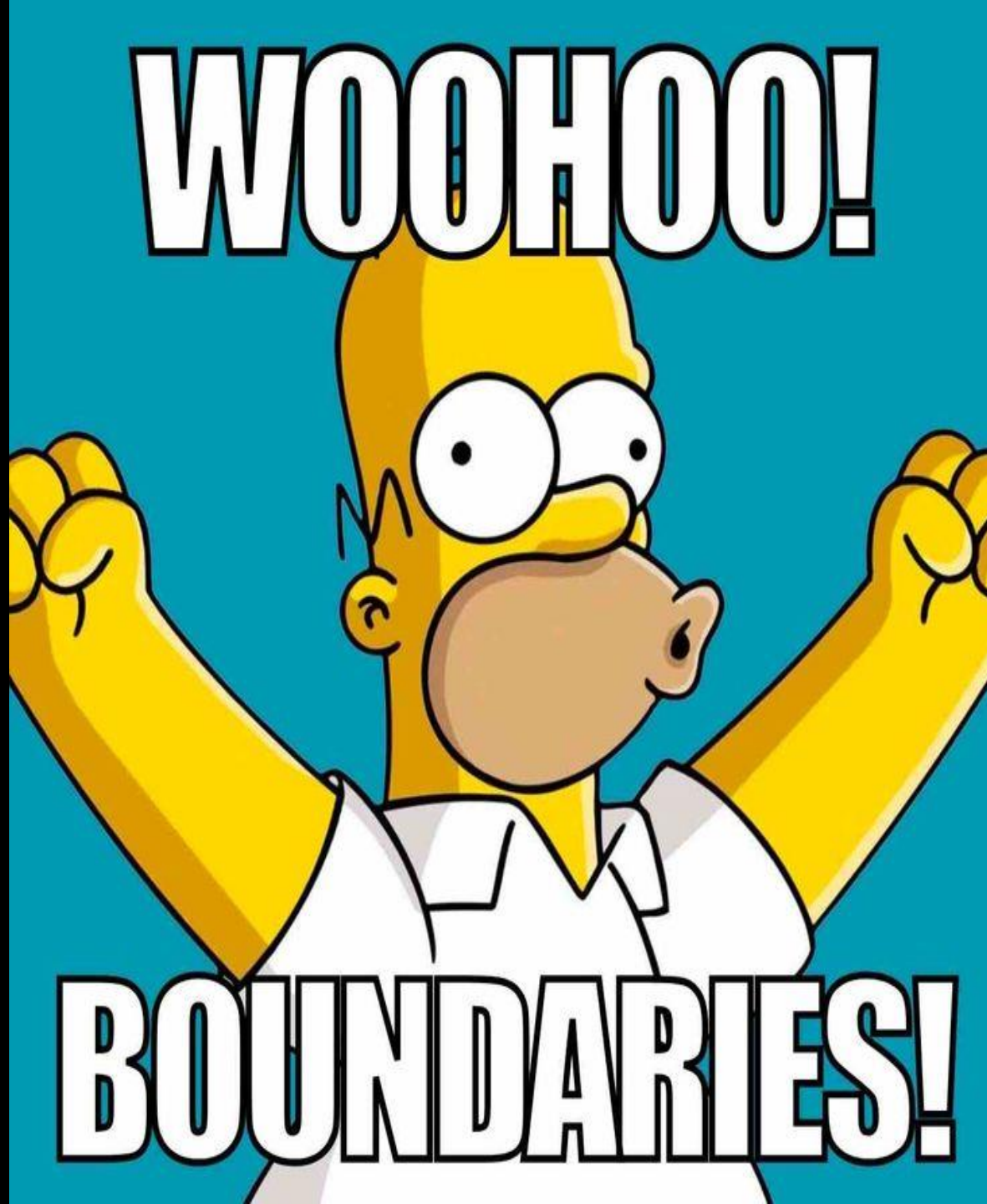


4. Solutions

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: 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**

65%

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: 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

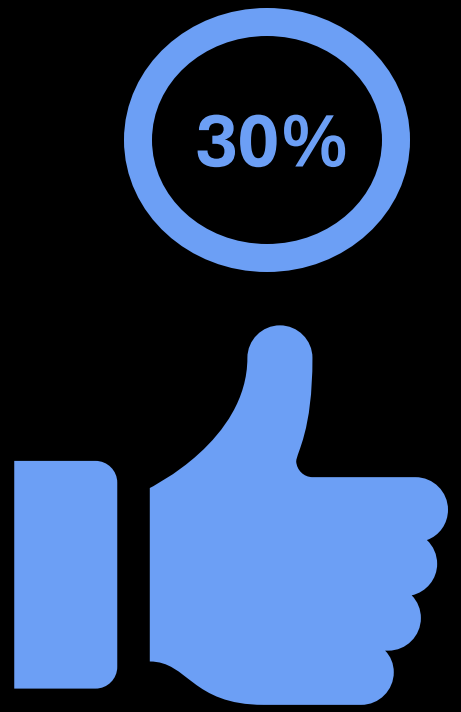
VM	$25.92 - $51.84
DBU	$52.80 – $105.60
Data Source	$3.11
Total	**$81.83 – $160.55**

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

Total Cost

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



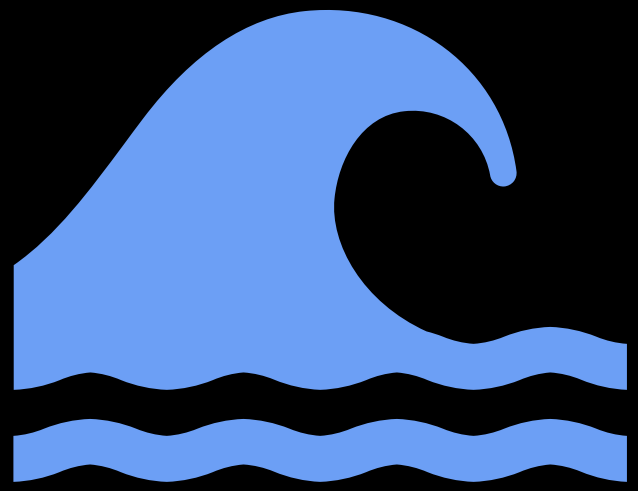
4. Solutions

(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

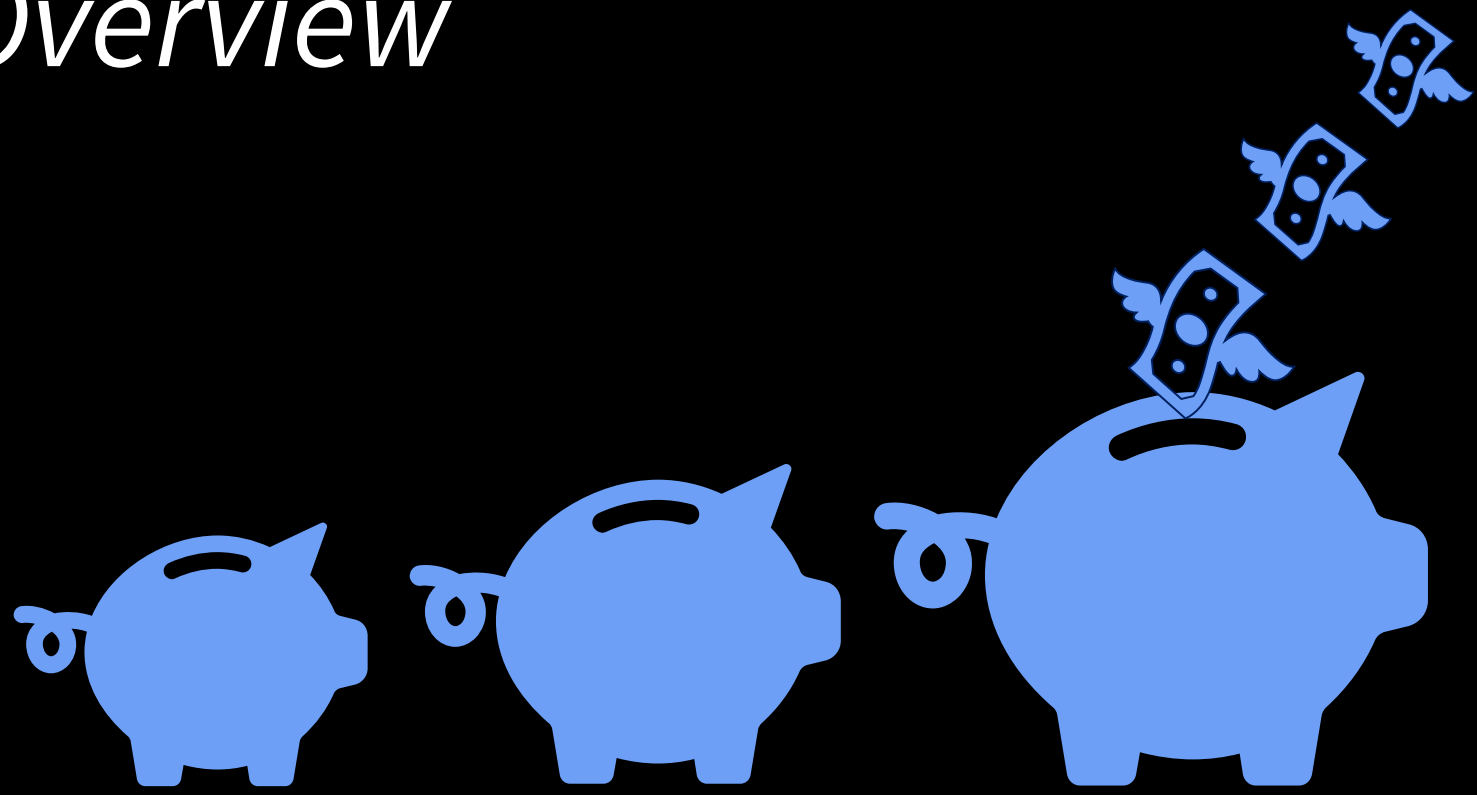
Total Cost

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

55%

4. Solutions

Overview

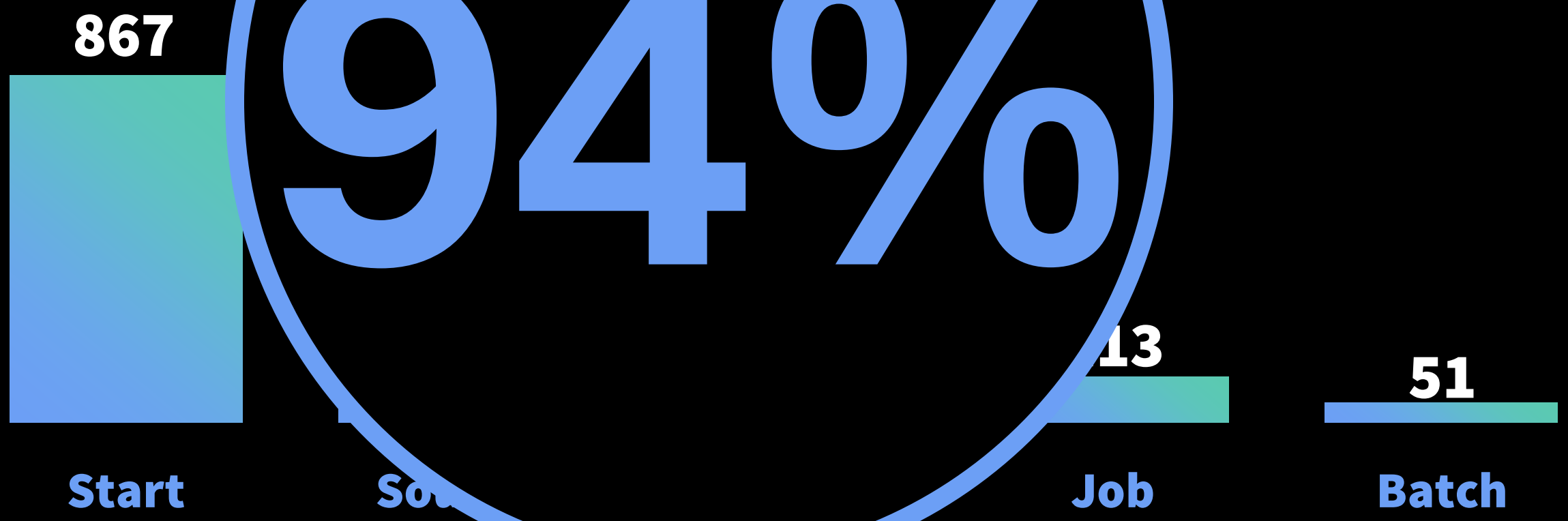


4. Solutions

Overview of cost savings

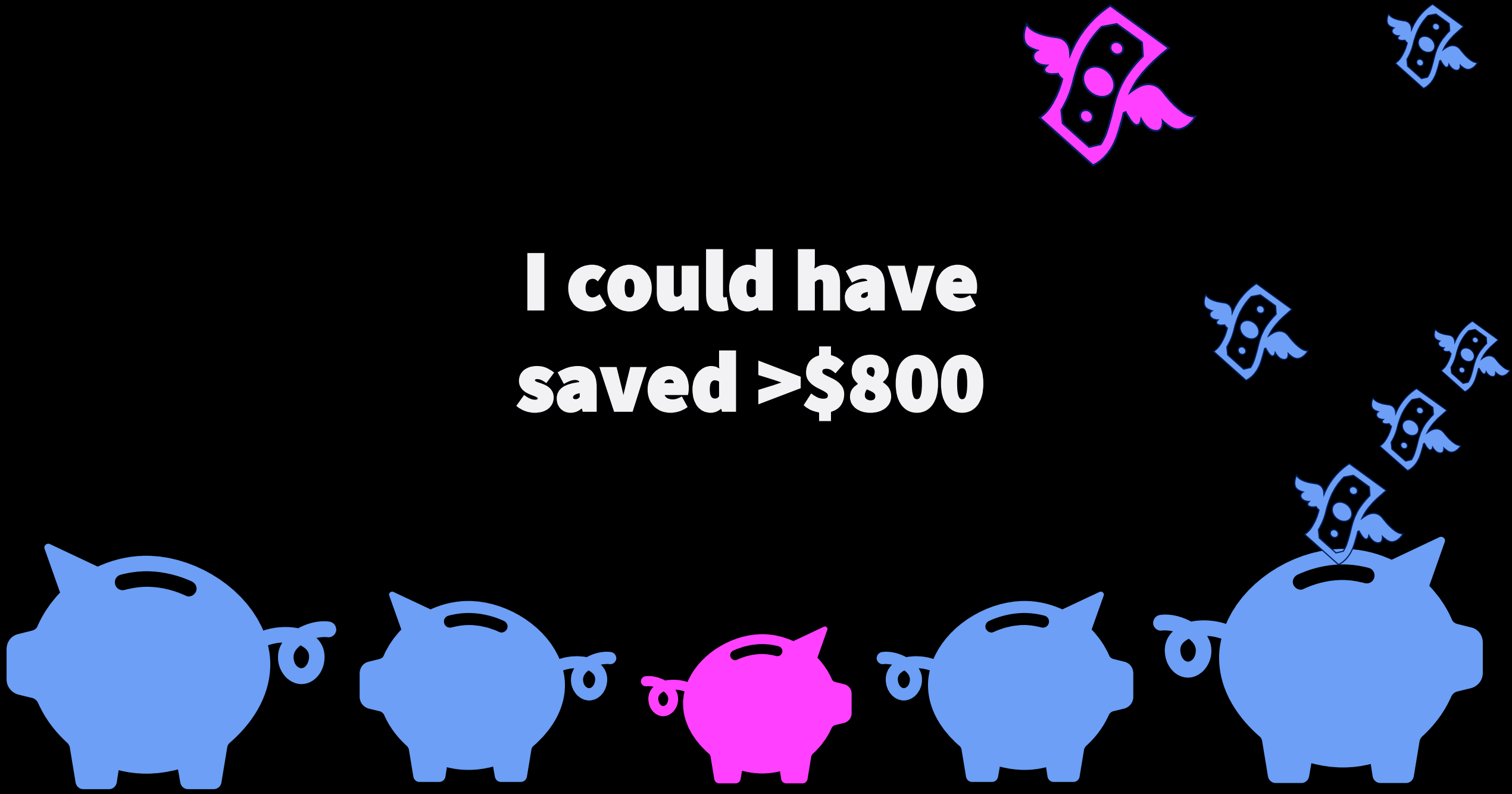
4. Solutio

Overview of c



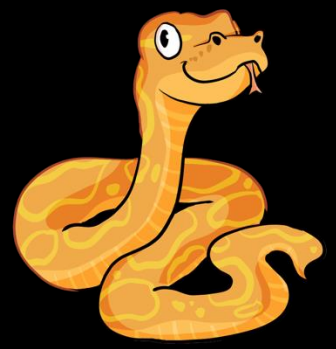
94%

I could have saved >$800



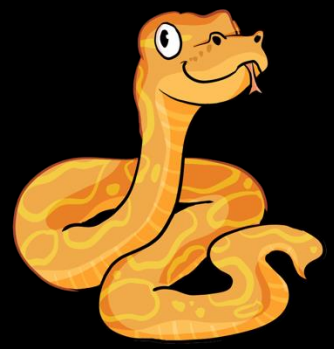
5. Conclusion

- 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 Compute
- Streaming is expensive



Data Engineer

Lisa Hoving