DATA+AI SUMMIT 2022

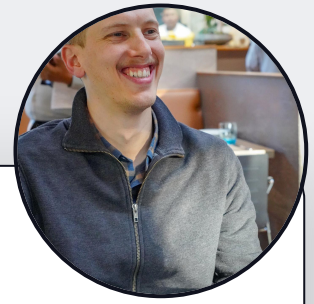
Doubling the Capacity of your Data Platform

Without doubling the cost

R Tyler Croy
Director, Platform Engineering
Scribd

Gavin Edgley
Senior Director, Value Acceleration
Databricks

Introductions



R Tyler Croy

- Director of Platform Engineering
- Data and ML Platform
- Helped bring Delta Lake to Rust
- Open source!

SCRIBD

Gavin Edgley

- Senior Director of Value Acceleration
- Build stories of Data & AI transformation for executives
- Helped 200+ customers

databricks

Three parts to today's talk

1. **Engineering for Scribd's growth**
2. **Remove barriers for engineers**
3. **Optimize infrastructure costs**

1

Engineering for Scribd's growth

SCRIBD

"Change the way the world reads."

Data is how we change the way the world reads

- **Understanding** is key to innovation at Scribd:
  - Understanding who/what/etc content is *about*
  - Understanding what is *interesting* for users to explore

Many **other teams** use our data platform to serve Scribd's mission

Engineering

Finance

Data platform & ML

- Enable users to **discover content** from Scribd's library (one of the world's largest!)
- Understand our document **corpus** - with ML & metadata
- Understand how our **users** are using the product

Marketing

Customer Support

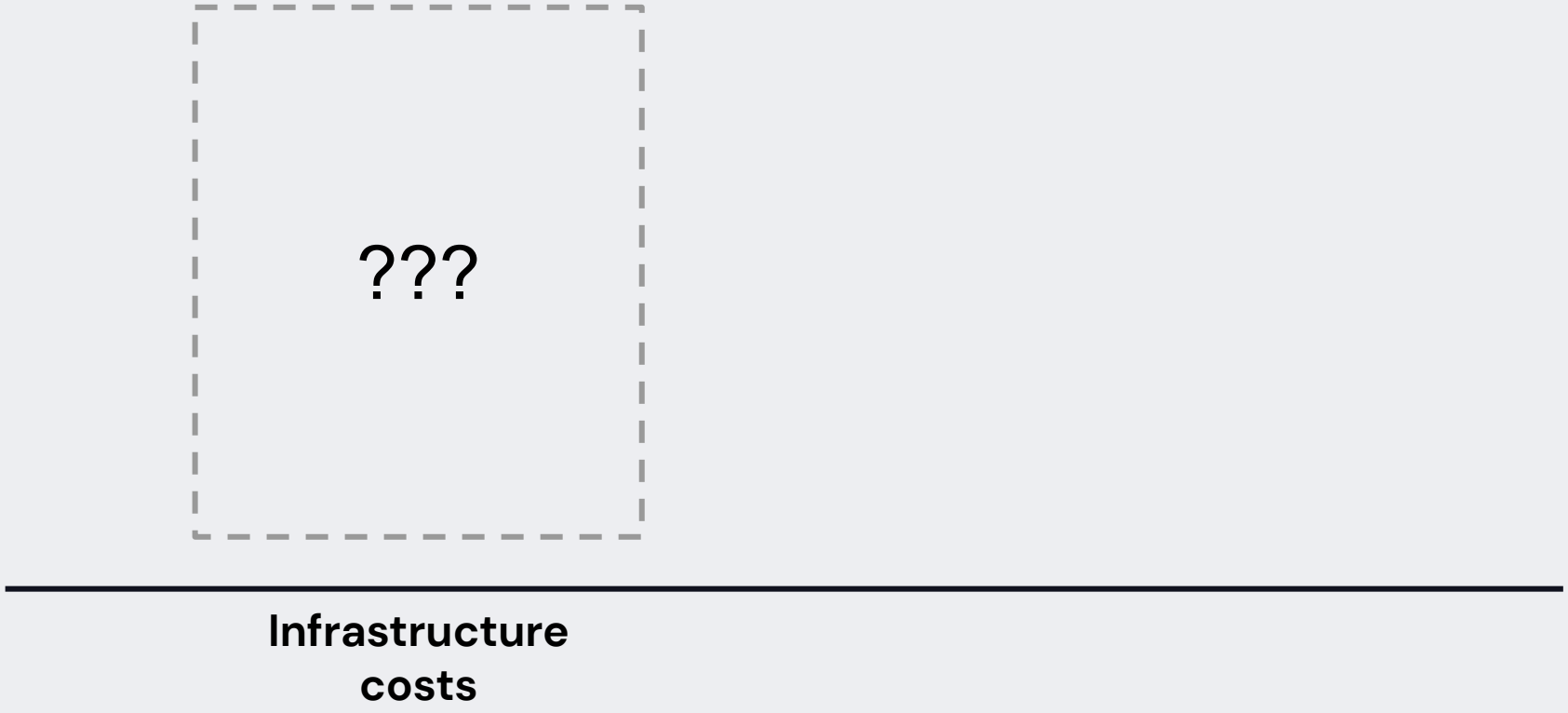
Business Analytics

Product

2

Removing barriers for engineers

We asked our engineers – how much do you think we spend on **data infrastructure**?



We asked our engineers – how much do you think we spend on **data infrastructure**?

We're a data company, our data costs must be huge!

???

We've just moved to the cloud, that must be expensive

Hundreds of thousands of dollars each year! Maybe millions!

The costs are high, so query optimization should be our top priority

Infrastructure costs

What about the cost of our people?

We have a large engineering team

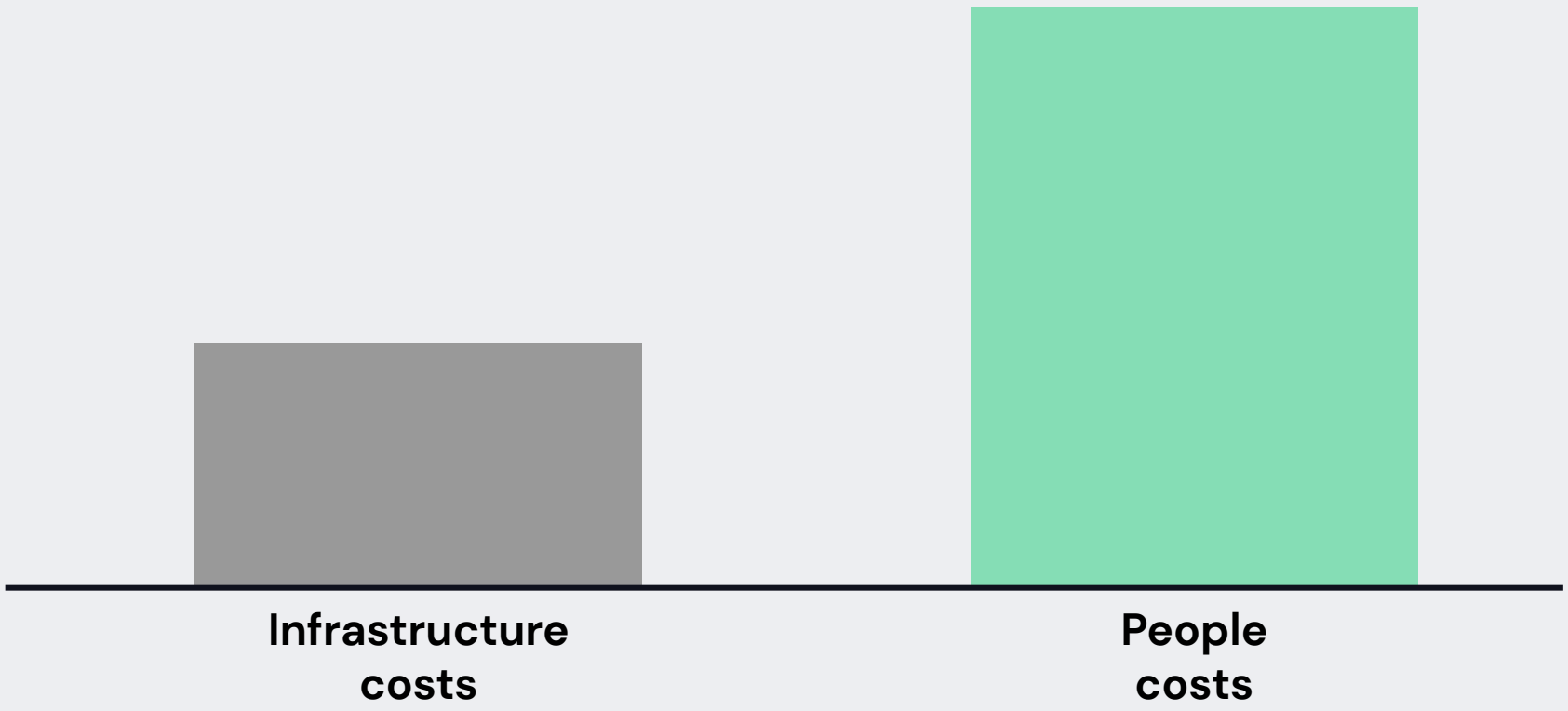
???

... could our team cost more than our infrastructure?

Infrastructure costs

People costs

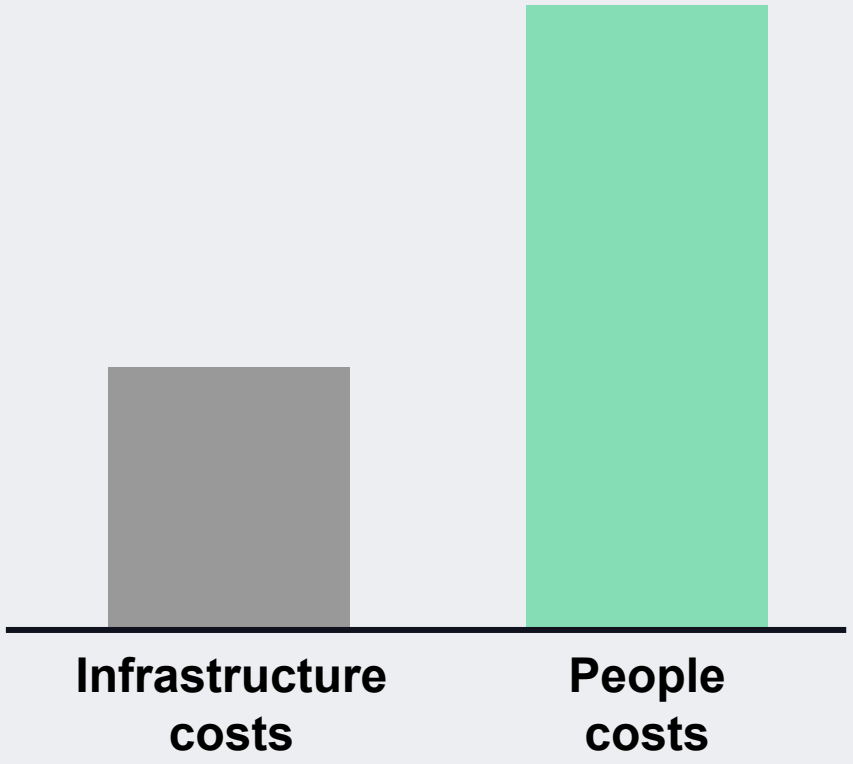
We spend much more on people than on infrastructure



!

How does this insight guide our approach?

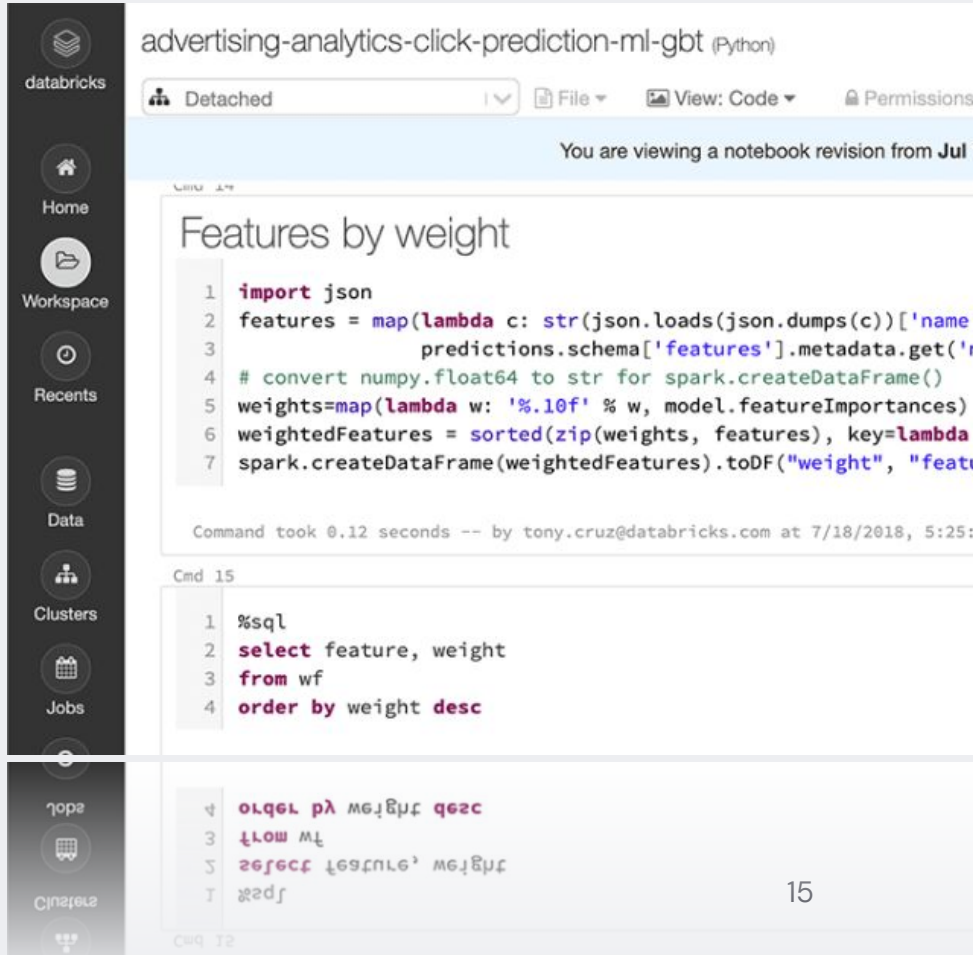
How does this insight guide our approach?



- Our time is our most valued asset
- Removing barriers for people is **most important**
- Reducing workload costs secondary

How do you remove barriers?

- Access to data means people can find what they need.
- Introducing Databricks Notebooks to the organization as a way to
  - Organize work around data
  - Do data development work
  - Easily collaborate across parts of the organization
- Giving people a query interface is nice, but if they don't know how to use it..they won't use it.
  - Enabling users requires thoughtful guidance



We gave **everybody** access to Databricks notebooks

What this means for...

Engineering

Customer Support

Marketing

Business Analytics

Product

Not giving everybody your credit card

- Shared interactive clusters and **cluster policies**
- Read-only access to production data via **instance profiles**
- Default choices for **Databricks Runtimes (DBRs)**
- Guidance to developers on *right-sizing* their resources

Engineers started solving problems **we didn't know existed**

- Shared notebooks for **Incident Response**
  - Better understanding of impact of incidents in real-time
  - Accelerating mean-time to resolution
- Shared notebooks for **validating feature success**
  - Product teams using data to immediately determine success of deployments
- Databricks SQL Queries for **edge-case alerting**
  - Needle in the haystack style problems affecting users which warrant deeper analysis

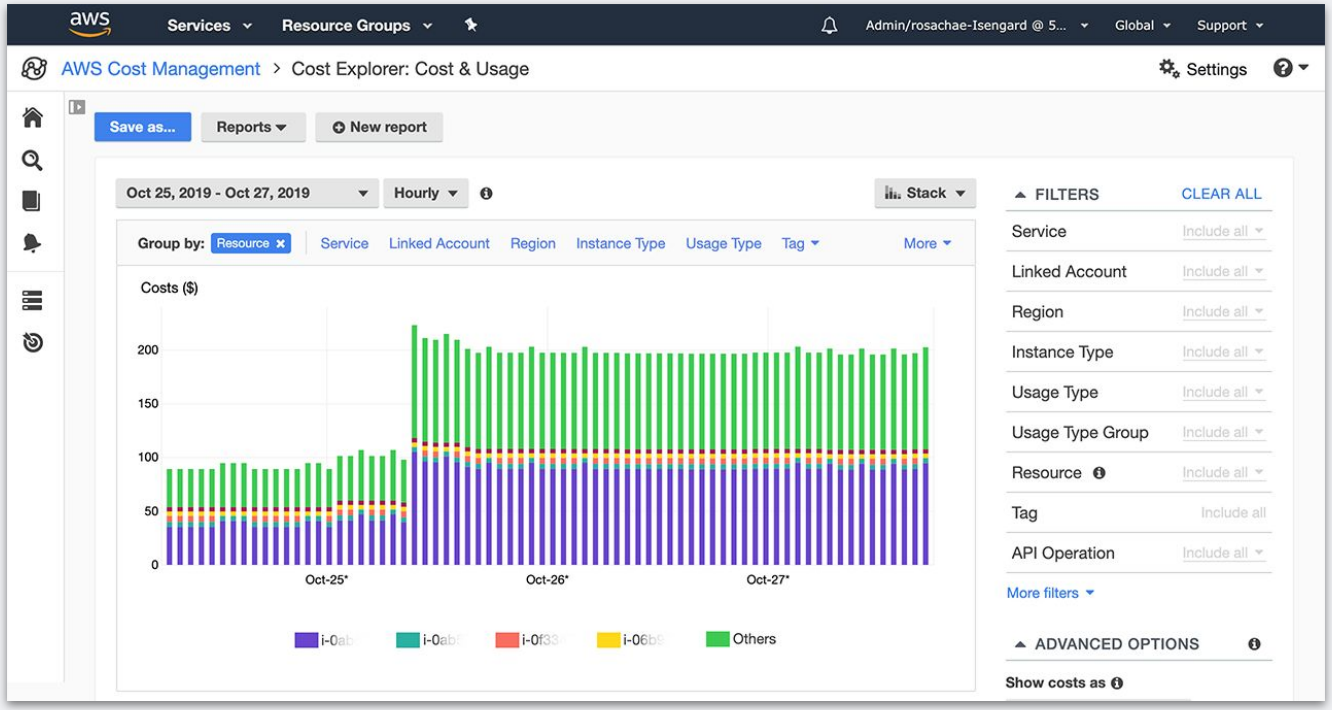
3 Optimizing infrastructure costs

Looking from the **AWS** perspective…

The biggest cost is the cloud, not Databricks

- Tagging *everything*
  - "owner"
  - "department"
  - "task_id"
  - "dag_id"
- AWS Cost Explorer
- AWS Budgets

Example AWS Cost Explorer



… and from the **Databricks** perspective

- Databricks Usage Log Delivery
  - Understand which workspaces are using which resources, for how long, etc
- Overwatch
  - Source available
  - Gives us massive insight into real resource utilization of jobs and notebooks.
  -

Example OverWatch dashboard

Altogether now!

With analysis we get faster results for users with less effort

- Platform teams can identify and suggest ideal "defaults" for cluster sizing
- Identify workloads which will benefit massively from adopting **Photon**
- Assessing and migrating to **Databricks SQL (Serverless)**
- Finding classes of overprovisioned workloads
  - Too much EBS, memory, CPU
- Assuming best intentions!
  - Nobody is *trying* to burn our cash, but not everybody has the insight we do

Bringing this success to you

Why TPC-DS?

TPC-DS is a **decision support benchmark** that models several generally applicable aspects of a decision support system, including queries and data maintenance.[1]

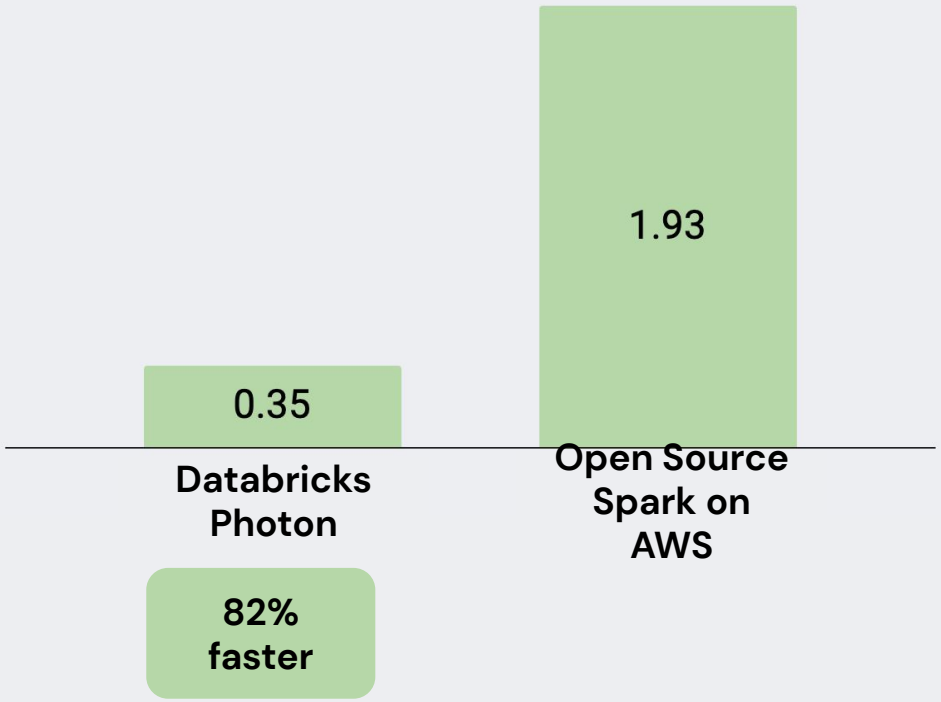
An **obstacle course** to test your data platform

Source: 1. TPC.org

With Databricks, jobs run **fast**

Time to insights – 30TB TPC-DS

Hours to complete jobs



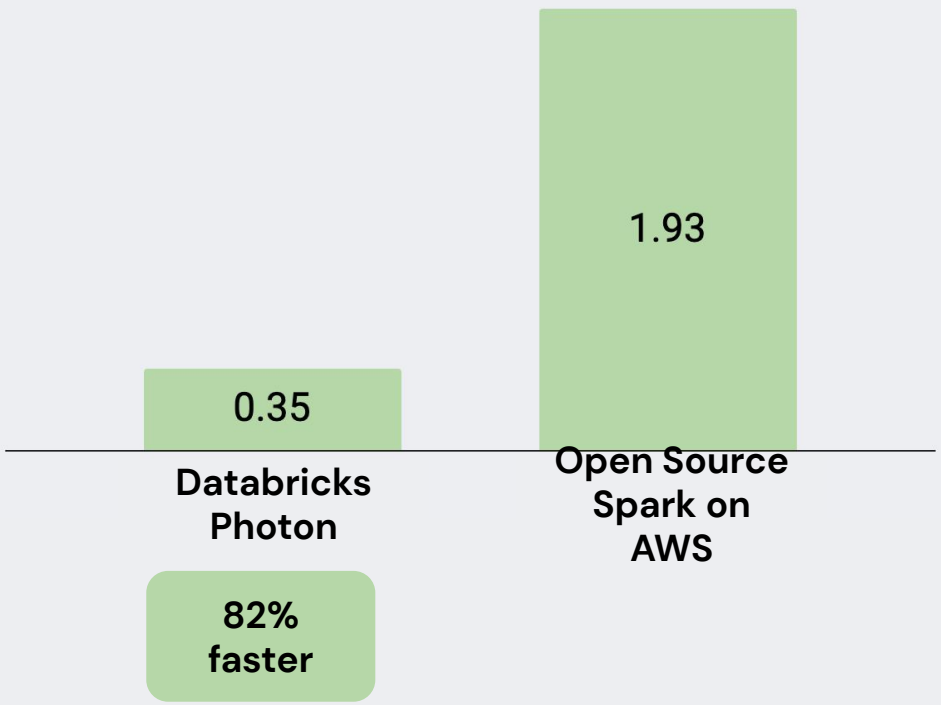
DATA+AI SUMMIT 2022

Source: TPC-DS benchmarks (April 2022)

Speed leads to **compute savings**

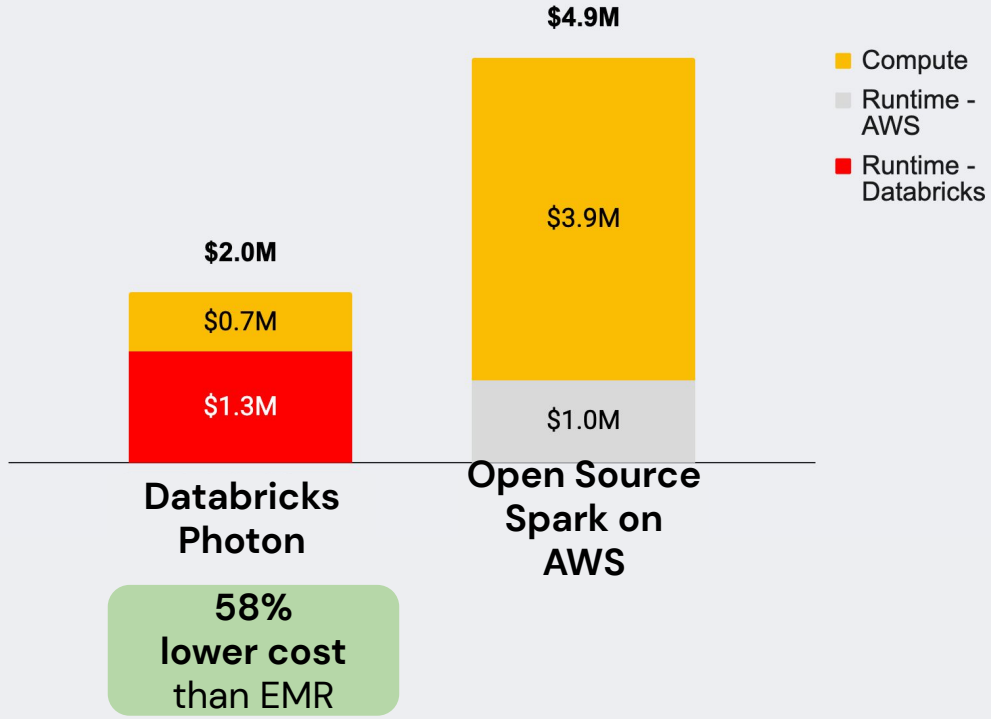
Time to insights – 30TB TPC-DS

Hours to complete jobs



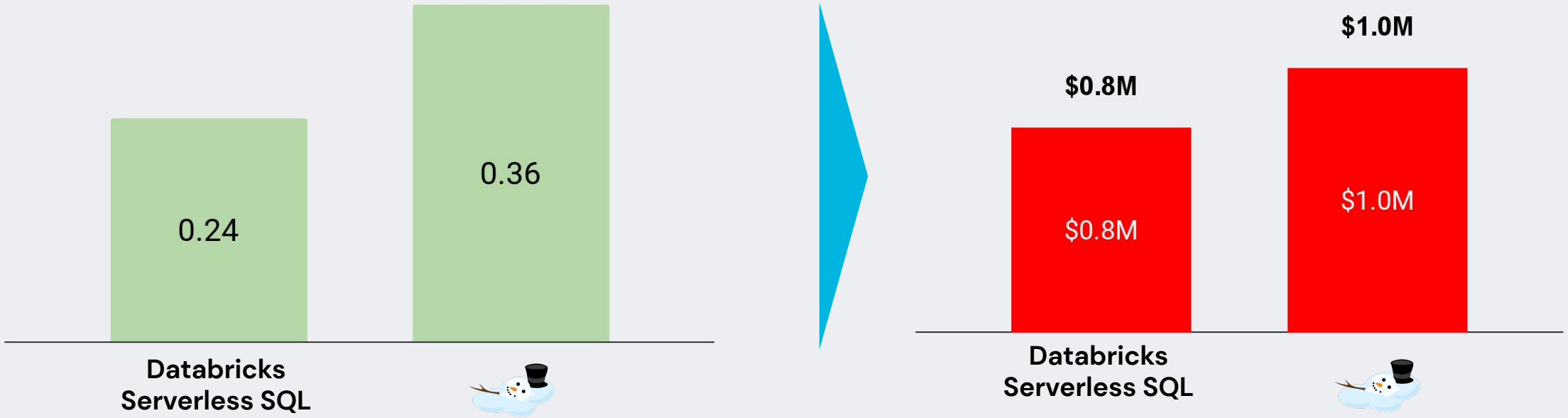
Total infrastructure cost

($Ms per Year)



Source: TPC-DS benchmarks (April 2022), AWS & Databricks list pricing (2022)

This applies to **SQL** too



Source: TPC-DS benchmarks (April 2022), Databricks, AWS & competitor list pricing (2022)

Summary

1. **Engineering for Scribd's growth**
2. **Remove barriers for engineers**
3. **Optimize infrastructure costs**

Next steps

Not a Databricks customer?

Try Databricks
databricks.com/try-databricks

Already a Databricks customer?

Explore...

- Photon
- Serverless SQL
- Overwatch*

* Overwatch is a Databricks Labs project

DATA+AI
SUMMIT 2022

Thank you

Now ask us questions!

We might have answers?