AWS re:Invent

NOV. 29 – DEC. 3, 2021 | LAS VEGAS, NV

STP204

Scalable data architectures for digital natives

Vinay Wagh
Director of Product
Databricks

Cullen Dejean
Sr. Startup Solutions Architect
AWS

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Vinay Wagh
Director of Product



Ihor Leshko
Director of Engineering



Jeff Pang
Principal Engineer

©2021 Databricks Inc. — All rights reserved

About databricks

5,000+ customers

Creators of





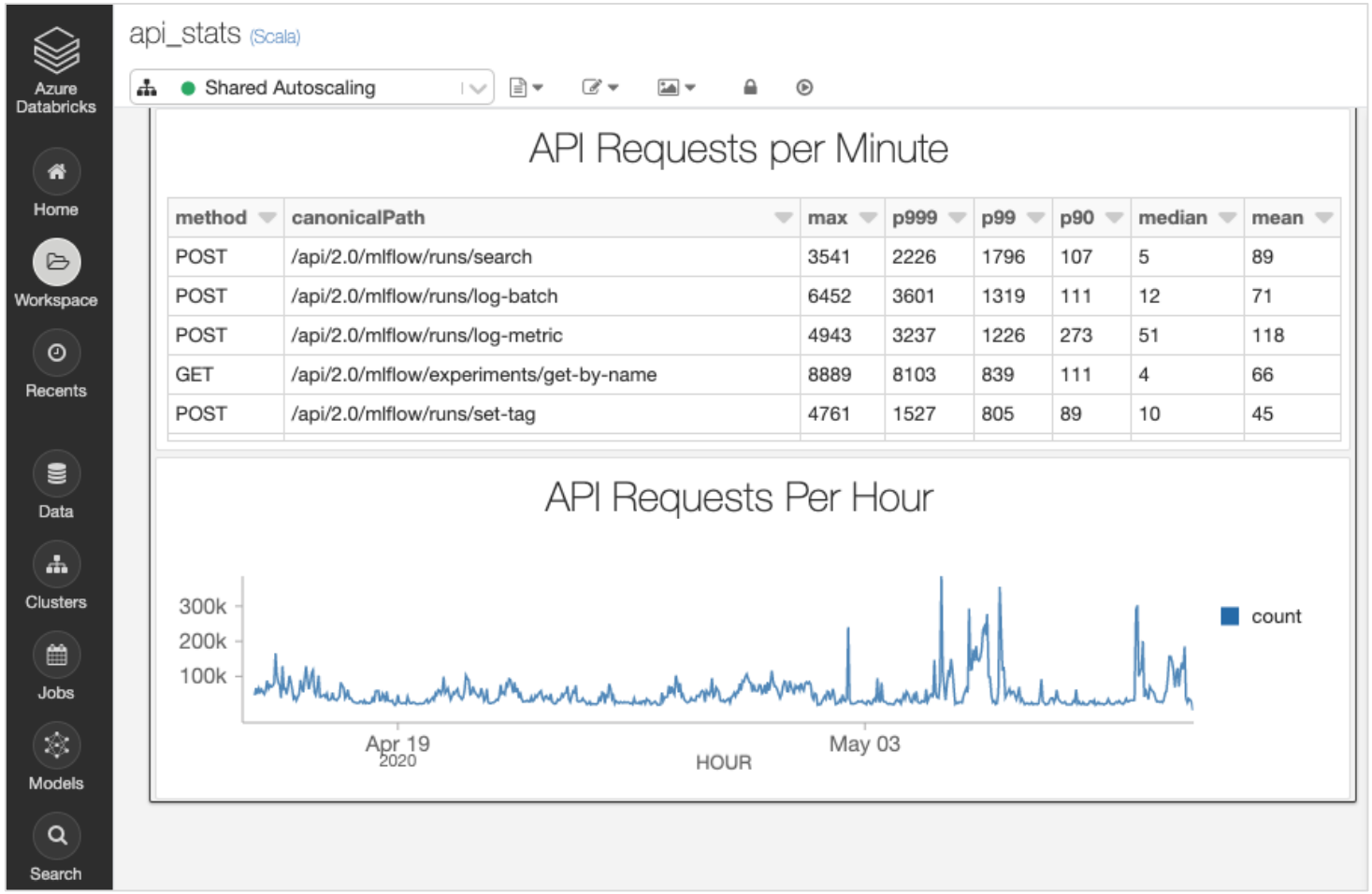




2,300+ employees

©2021 Databricks Inc. — All rights reserved

Our product

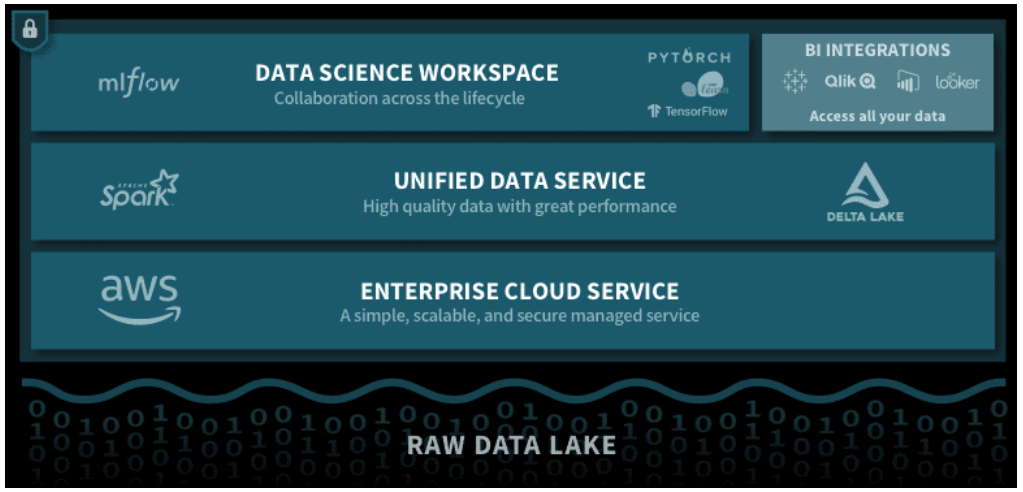


method	canonicalPath	max	p999	p99	p90	median	mean
POST	/api/2.0/mlflow/runs/search	3541	2226	1796	107	5	89
POST	/api/2.0/mlflow/runs/log-batch	6452	3601	1319	111	12	71
POST	/api/2.0/mlflow/runs/log-metric	4943	3237	1226	273	51	118
GET	/api/2.0/mlflow/experiments/get-by-name	8889	8103	839	111	4	66
POST	/api/2.0/mlflow/runs/set-tag	4761	1527	805	89	10	45





Data scientists Data engineers Business users



©2021 Databricks Inc. — All rights reserved

Agenda

The databricks architecture

Inside the Unified Analytics Platform

Challenges & lessons

Growing a SaaS data platform on AWS

Accelerating a data platform with data & AI

Serverless that scales the use of data & AI

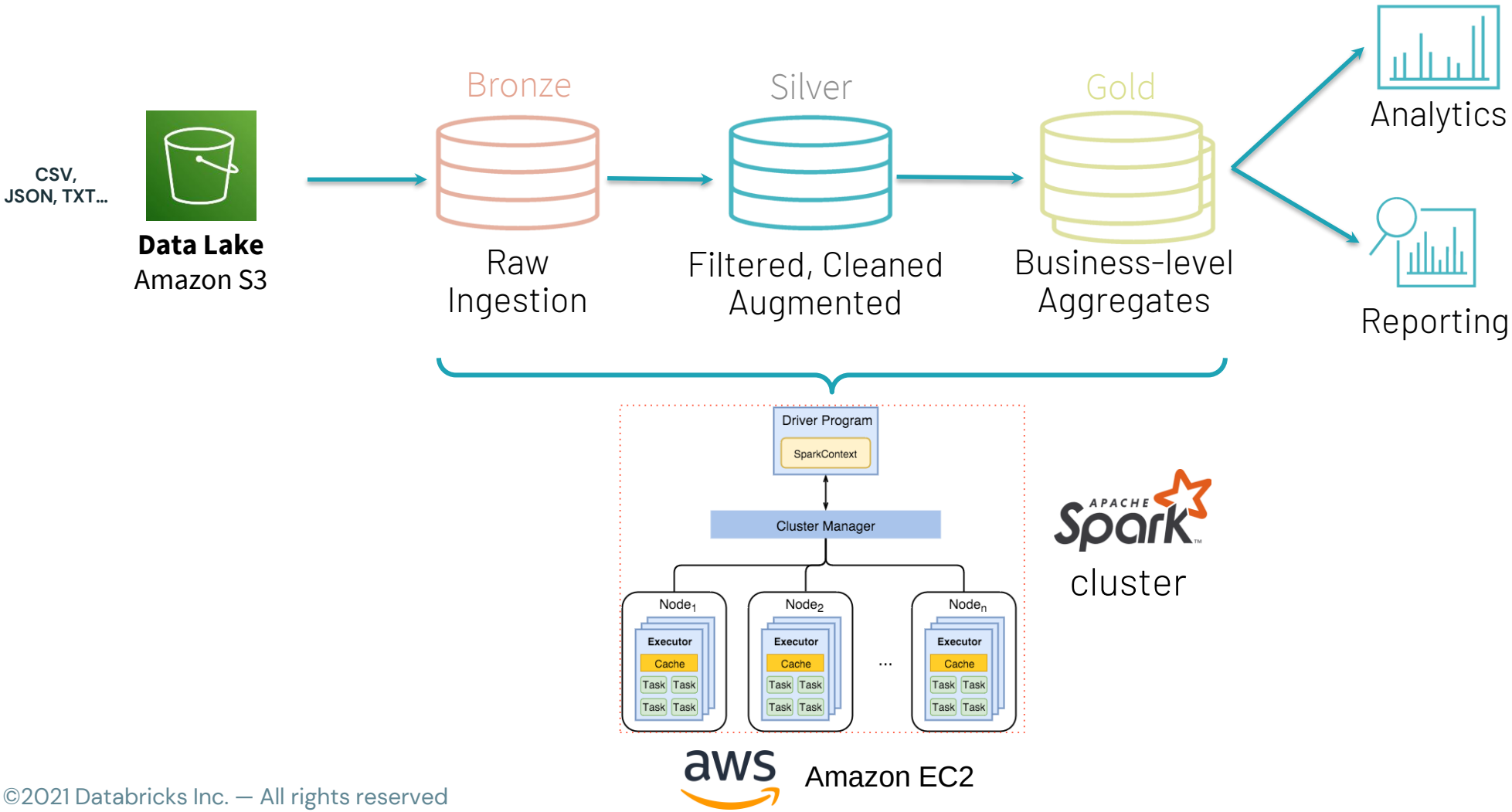
©2021 Databricks Inc. — All rights reserved

The  databricks architecture

Inside the Unified Analytics Platform

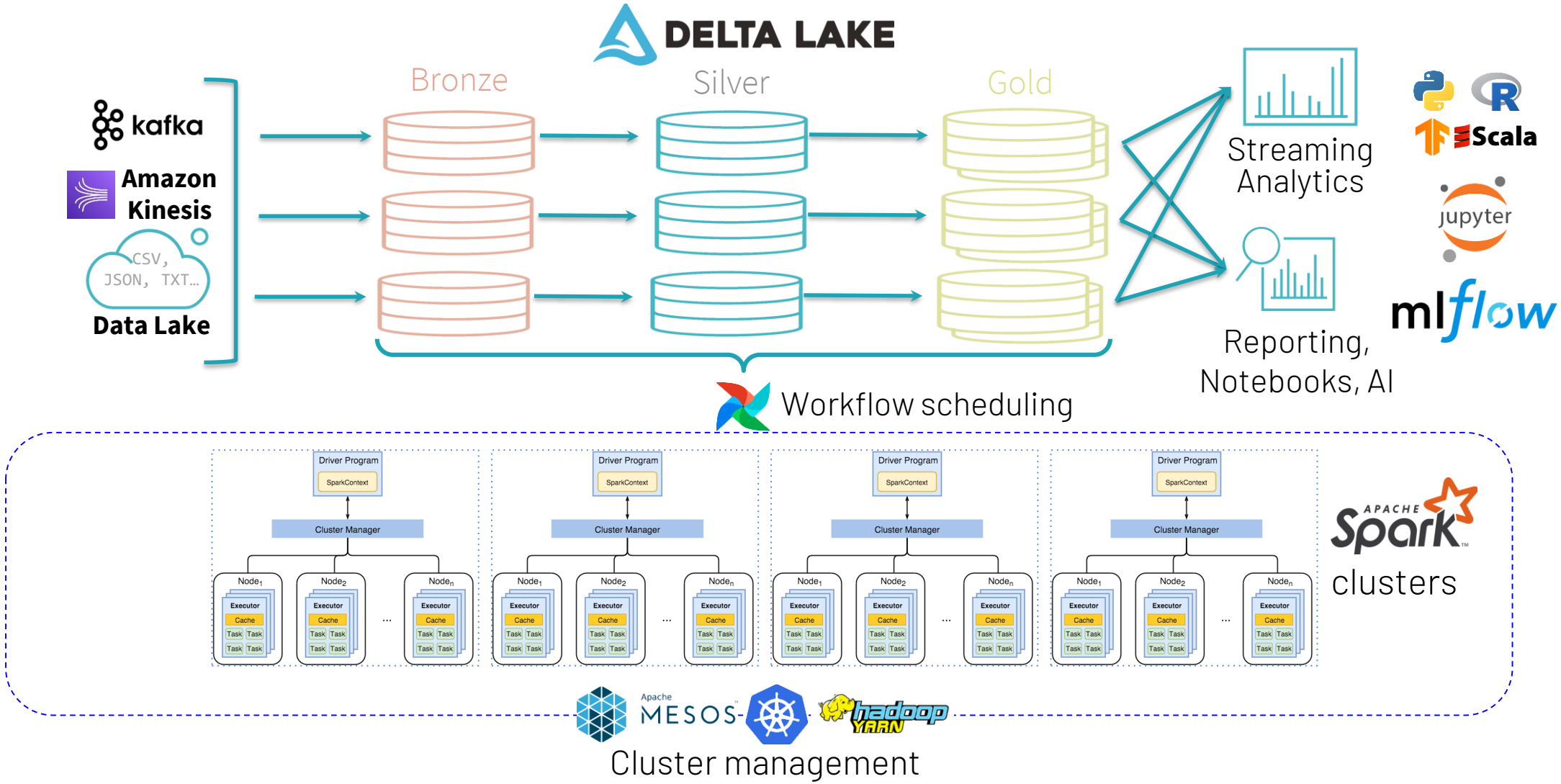
©2021 Databricks Inc. — All rights reserved

Simple data engineering architecture



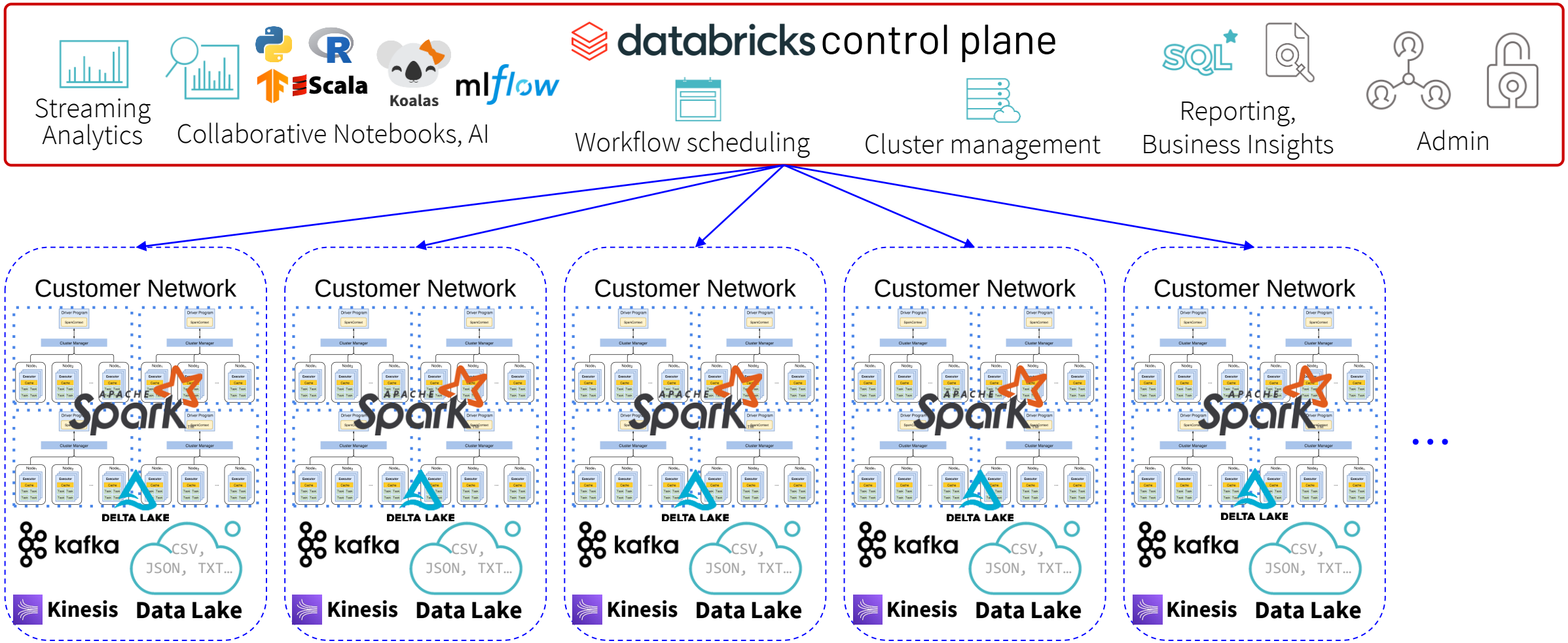
©2021 Databricks Inc. — All rights reserved

Modern data engineering architecture



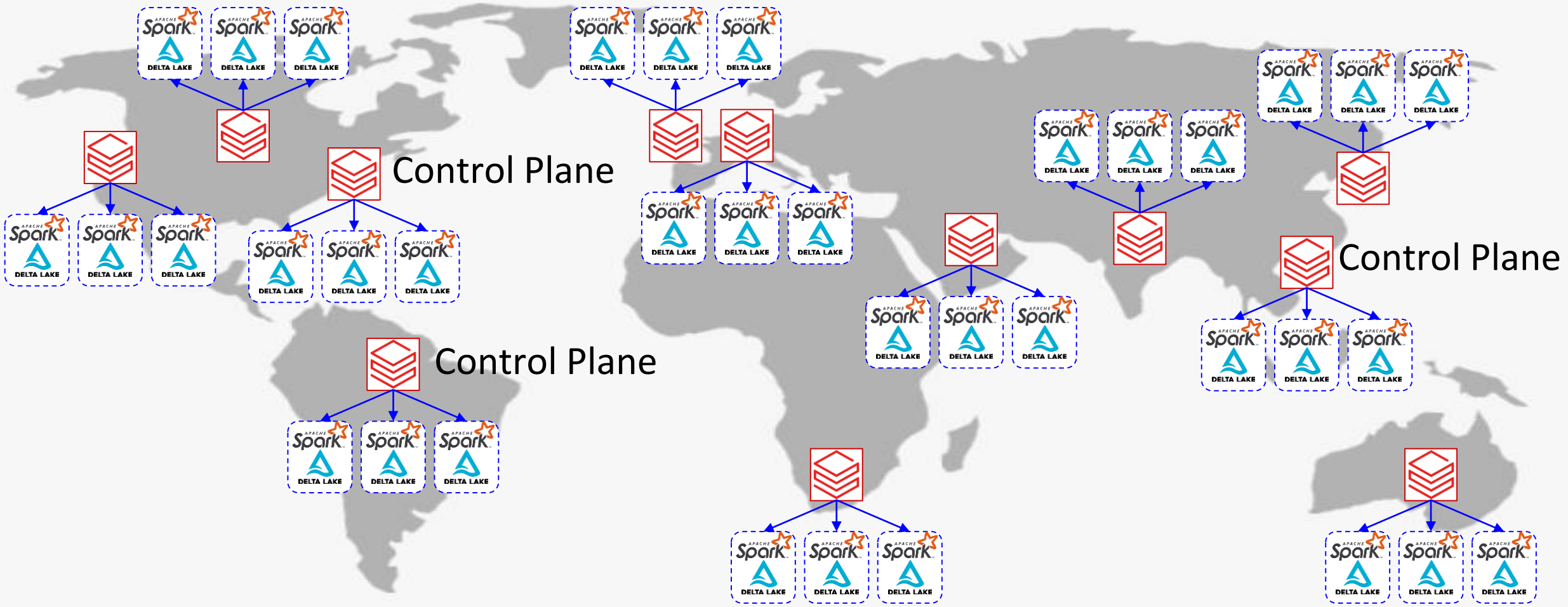
©2021 Databricks Inc. — All rights reserved

Multiply by thousands of customers . . .



©2021 Databricks Inc. — All rights reserved

. . . across many regions . . .



©2021 Databricks Inc. — All rights reserved

→ the Databricks control plane

100,000s of users

100,000s of Spark clusters per day

Millions of VMs launched per day

Exabytes of data processed per day

What did we learn from building a large-scale, multi-cloud data platform?

©2021 Databricks Inc. — All rights reserved

Growing a SaaS data platform

©2021 Databricks Inc. — All rights reserved

Evolution of the Databricks control plane

We didn't start with a global-scale, multi-cloud data platform

Challenge: Scaling a data platform from one customer to 5,000+

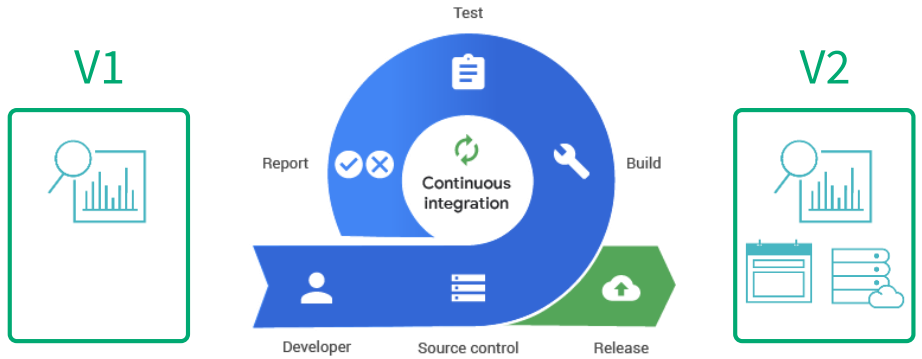
Lesson: The *factory* that builds and evolves the data platform is more important than the data platform itself

©2021 Databricks Inc. — All rights reserved

Fast time to market

Databricks control plane “in a box”

- Need to deliver value quickly
- Need to iterate quickly
- Can’t break things while iterating!

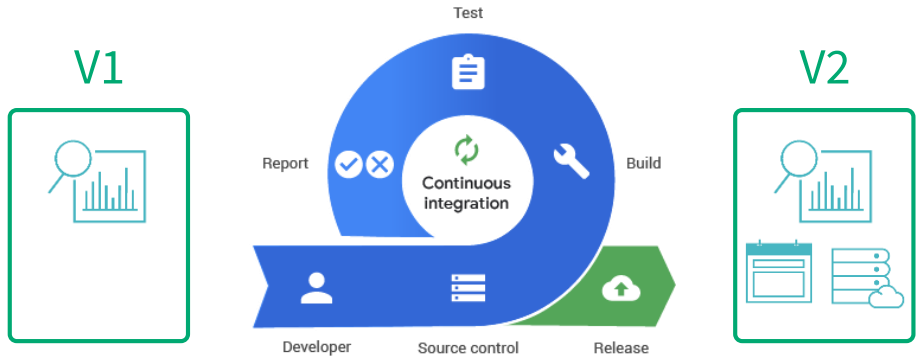


©2021 Databricks Inc. — All rights reserved

Fast time to market

Databricks control plane “in a box”

- Need to deliver value quickly
- Need to iterate quickly
- Can’t break things while iterating!



Scala Bazel React

Keys to success

- Modern CI
- Fast developer tools
- Testing, testing, testing

Up to 500x Scala build speedups

Tens of millions of tests per day

Hundreds of Databricks “in-a-box” test environments per day

©2021 Databricks Inc. — All rights reserved

Expand the total addressable market

Replicating control planes quickly

- Need different configurations for different environments
- Need to update many environments
- Can’t slow down platform development!



©2021 Databricks Inc. — All rights reserved

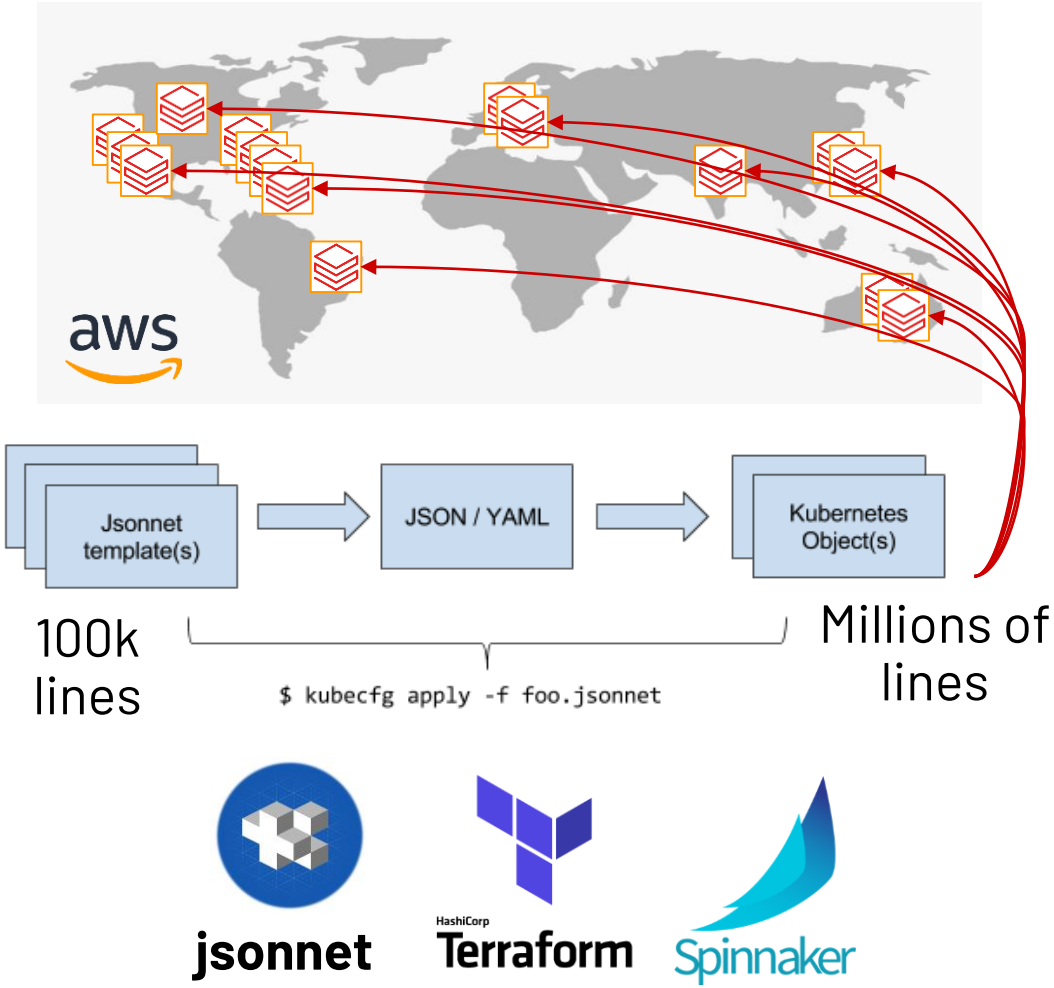
Expand the total addressable market

Replicating control planes quickly

- Need different configurations for different environments
- Need to update many environments
- Can't slow down platform development!

Keys to success

- Declarative infrastructure
- Modern CD infrastructure



©2021 Databricks Inc. — All rights reserved

Land and *expand* workloads

Scaling the control plane

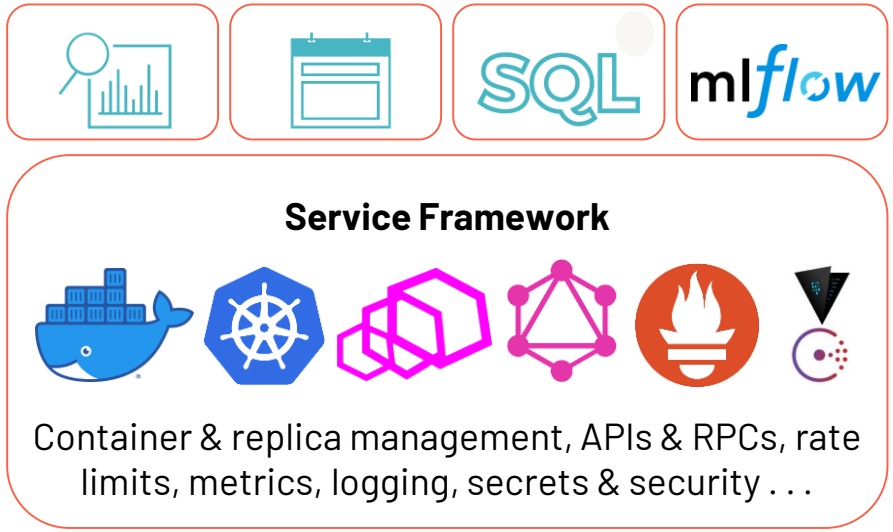
- Need to support more users & workloads
- Need to build more features that scale
- Don’t want devs to reinvent the wheel!

©2021 Databricks Inc. — All rights reserved

Land and *expand* workloads

Scaling the control plane

- Need to support more users & workloads
- Need to build more features that scale
- Don’t want devs to reinvent the wheel!



Keys to success

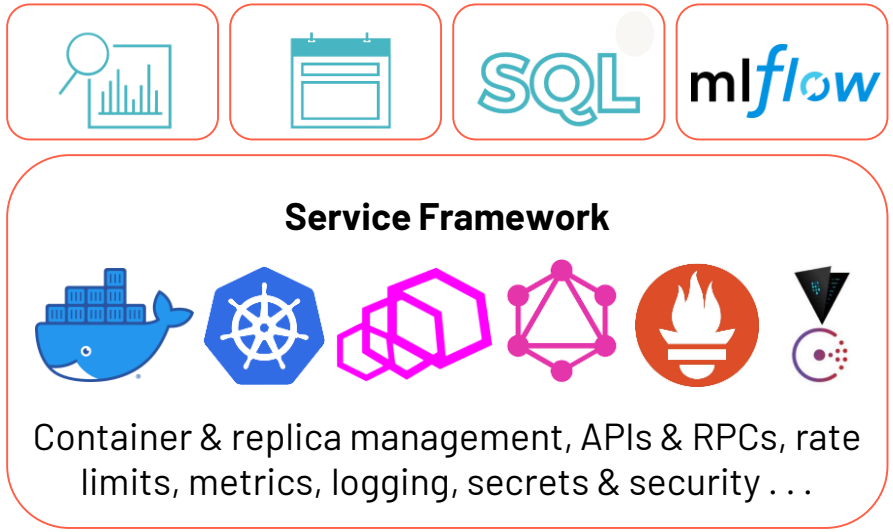
- A service framework to do the hard stuff

©2021 Databricks Inc. — All rights reserved

Land and *expand* workloads

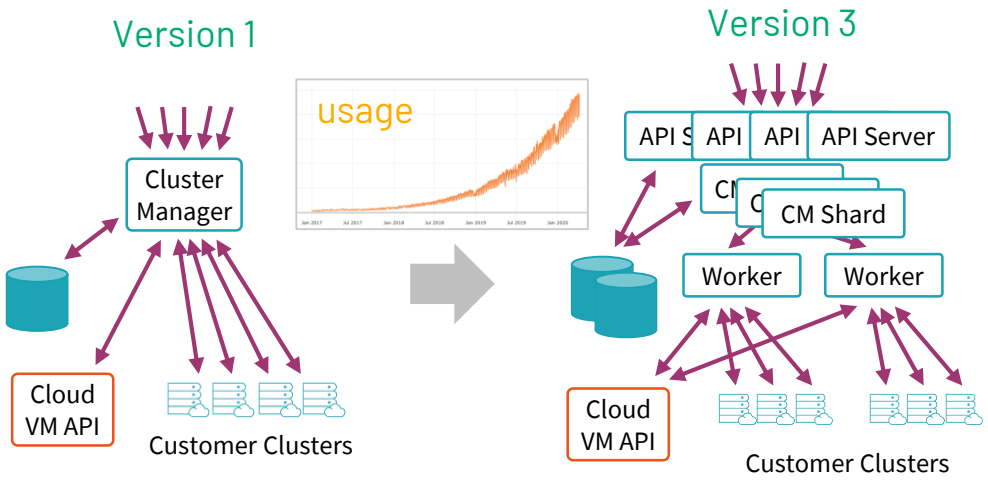
Scaling the control plane

- Need to support more users & workloads
- Need to build more features that scale
- Don't want devs to reinvent the wheel!



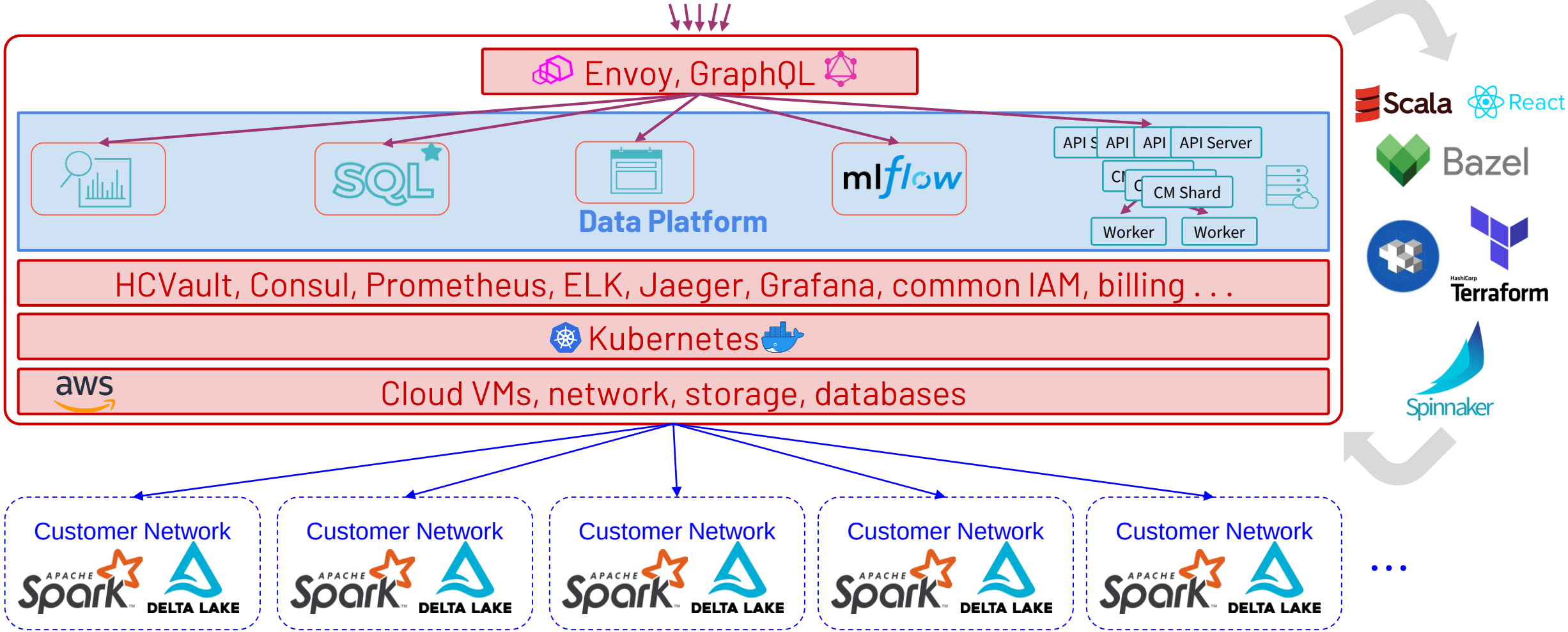
Keys to success

- A service framework to do the hard stuff
- Decompose monoliths to microservices



©2021 Databricks Inc. — All rights reserved

The Databricks data platform factory



©2021 Databricks Inc. — All rights reserved

Accelerating a data platform with data & AI

©2021 Databricks Inc. — All rights reserved

Inception: Improving a data platform with data & AI

We are one of our biggest customers

Challenge: Building a data platform is hard without a data platform

- Need data to track usage, maintain security
- Need data to observe and improve how users use the data platform
- Need data to keep the data platform up and running

Lesson: Data & AI can accelerate data platform features, product analytics, and DevOps

©2021 Databricks Inc. — All rights reserved

How we use Databricks to accelerate itself

Key platform features

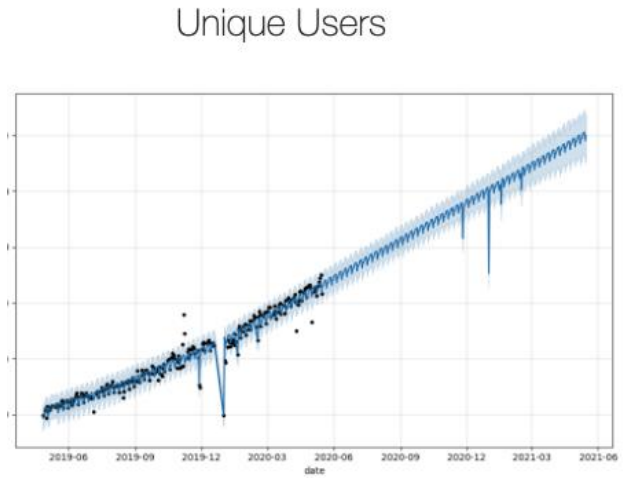
- Usage and billing reports
- Audit logs

Essential product analytics

- Feature usage, trends, prediction
- Growth and churn forecast, models

Mission-critical DevOps

- Service KPIs and SLAs
- API and application structured logs
- Spark debug logs



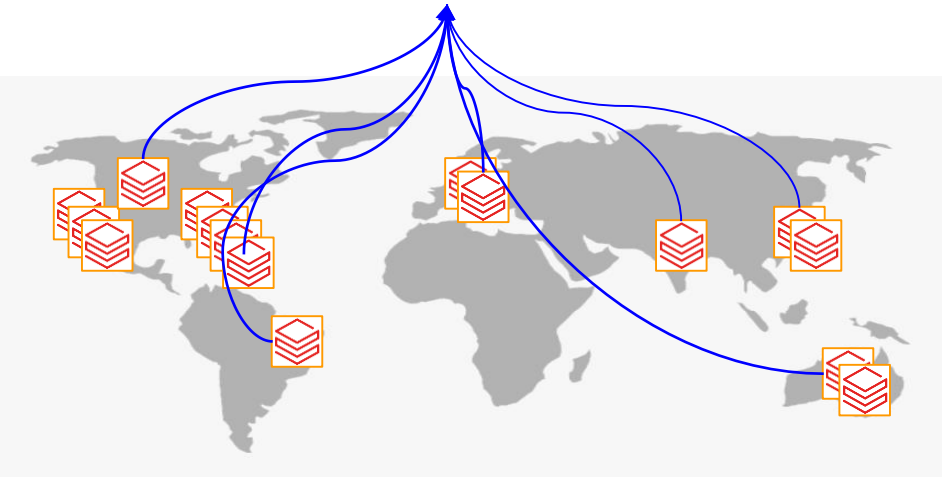
Target	Actual	Sample Size	Latency 95p (ms)	Latency 99p (ms)
99.95	100	28,800	N/A	N/A
99.95	99.99	6,355,888	N/A	N/A
99.95	100	157,258,264,810	N/A	N/A
99.7	99.8	412,581	N/A	N/A
99.95	100	237,160	N/A	N/A
99.9	99.98	1,771,736	3,722	15,827
99.9	99.99	556,785	8,492	10,048
99.95	100	129,360	N/A	N/A

©2021 Databricks Inc. — All rights reserved

Our distributed data pipelines

Real-time streaming

Kinesis Event Hubs kafka



DELTA LAKE

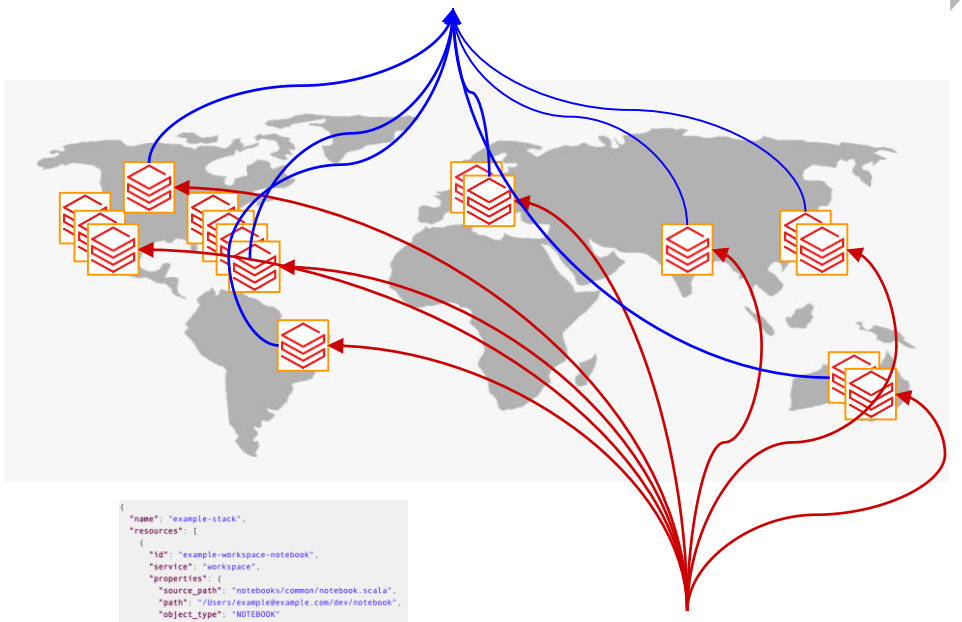
Data foundation & analytics

©2021 Databricks Inc. — All rights reserved

Our distributed data pipelines

Real-time streaming

Kinesis Event Hubs kafka





Declarative data pipeline deployments

DELTA LAKE

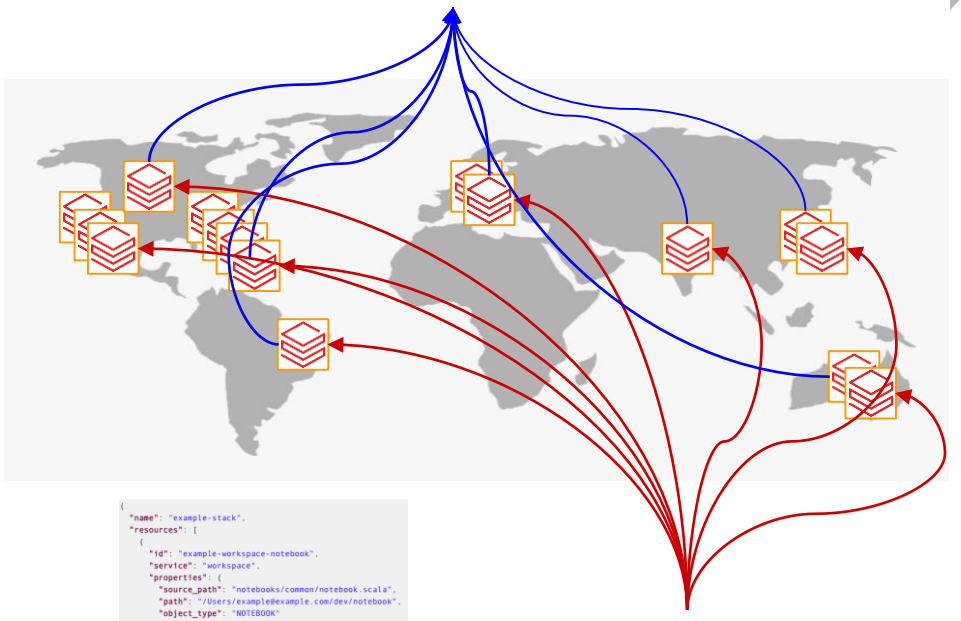
Data foundation & analytics

©2021 Databricks Inc. — All rights reserved

Our distributed data pipelines

Real-time streaming

Kinesis Event Hubs kafka





Declarative data pipeline deployments

DELTA LAKE

Data foundation & analytics

Grafana Prometheus M3

elasticsearch + logstash + kibana

JAEGER

Time-series, raw logs, request tracing, dashboards

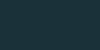
Hundreds of TB logs per day

Millions of time series per second

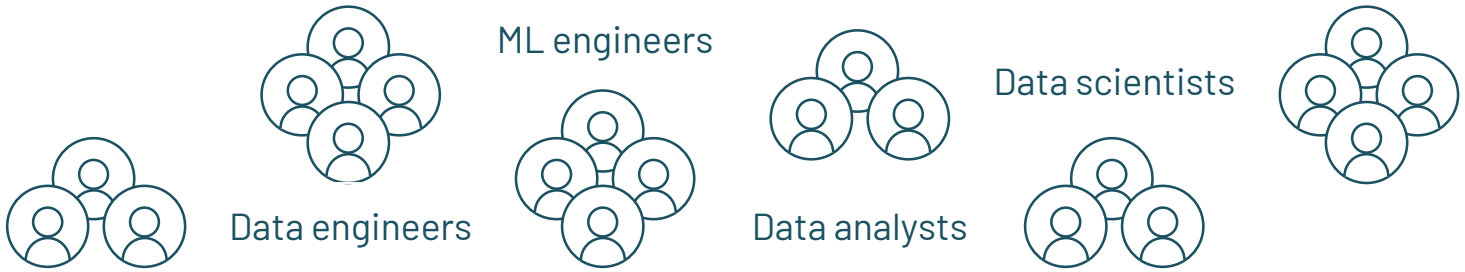
©2021 Databricks Inc. — All rights reserved

Scaling the use of data & AI

©2021 Databricks Inc. — All rights reserved



LAKEHOUSE



ALL THE USERS

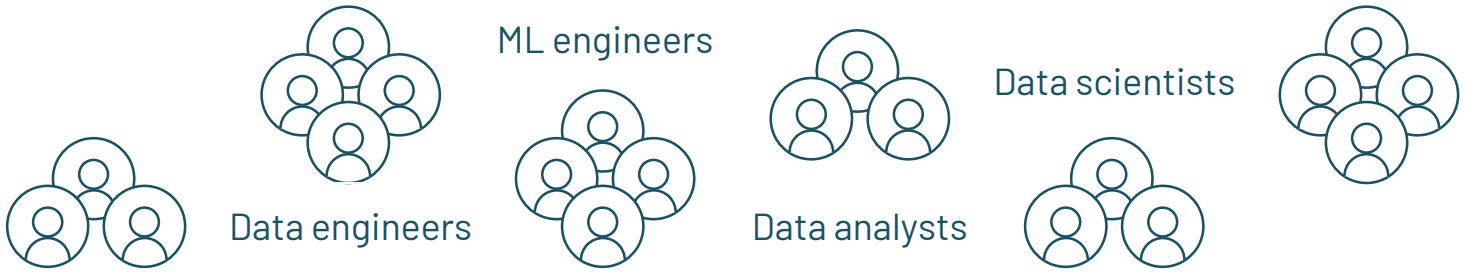
Scaling the use of data across all users is challenging



ALL THE DATA

Structured, Semi-structured, and Unstructured Data

LAKEHOUSE



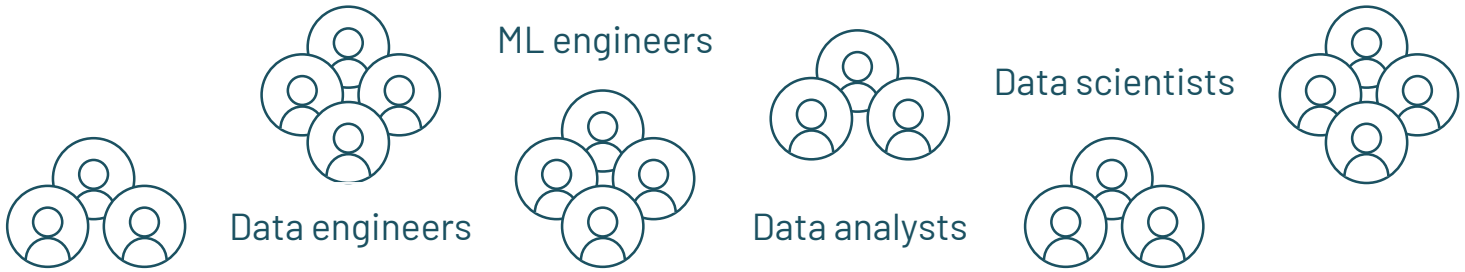
ALL THE USERS



ALL THE DATA

Structured, Semi-structured, and Unstructured Data

LAKEHOUSE



ALL THE USERS

ELASTIC INSTANT COMPUTE, SECURITY & COMPLIANCE,
PERFORMANCE/COST MANAGEMENT



ALL THE DATA

Structured, Semi-structured, and Unstructured Data

Elastic instance compute

It takes ~3-5 minutes to launch an instance



Engineers and analysts don't like to wait for clusters to boot up

5 minutes

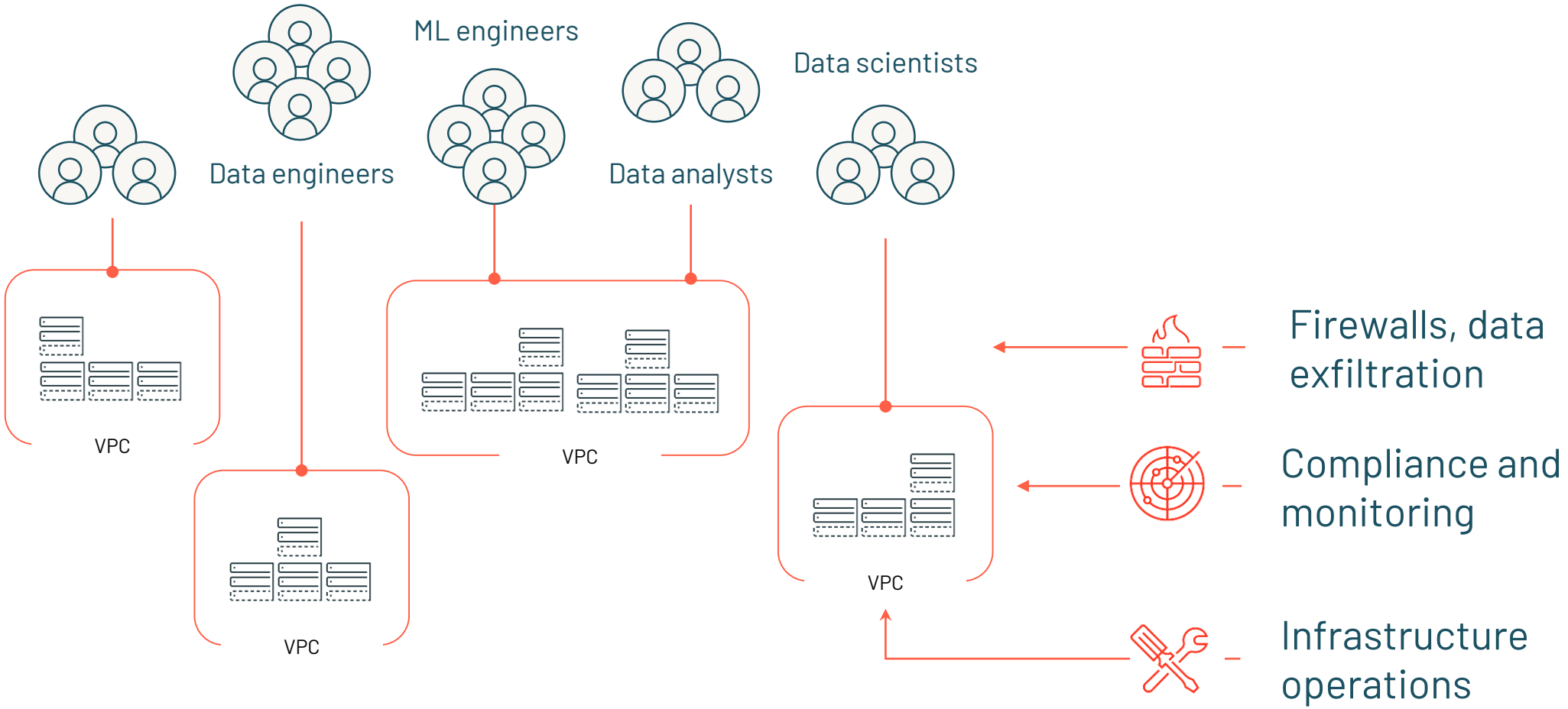
80th percentile of jobs complete in less than 5 minutes

SLA

In some cases, data teams have to provide a query SLA of under 1 minute for production tables

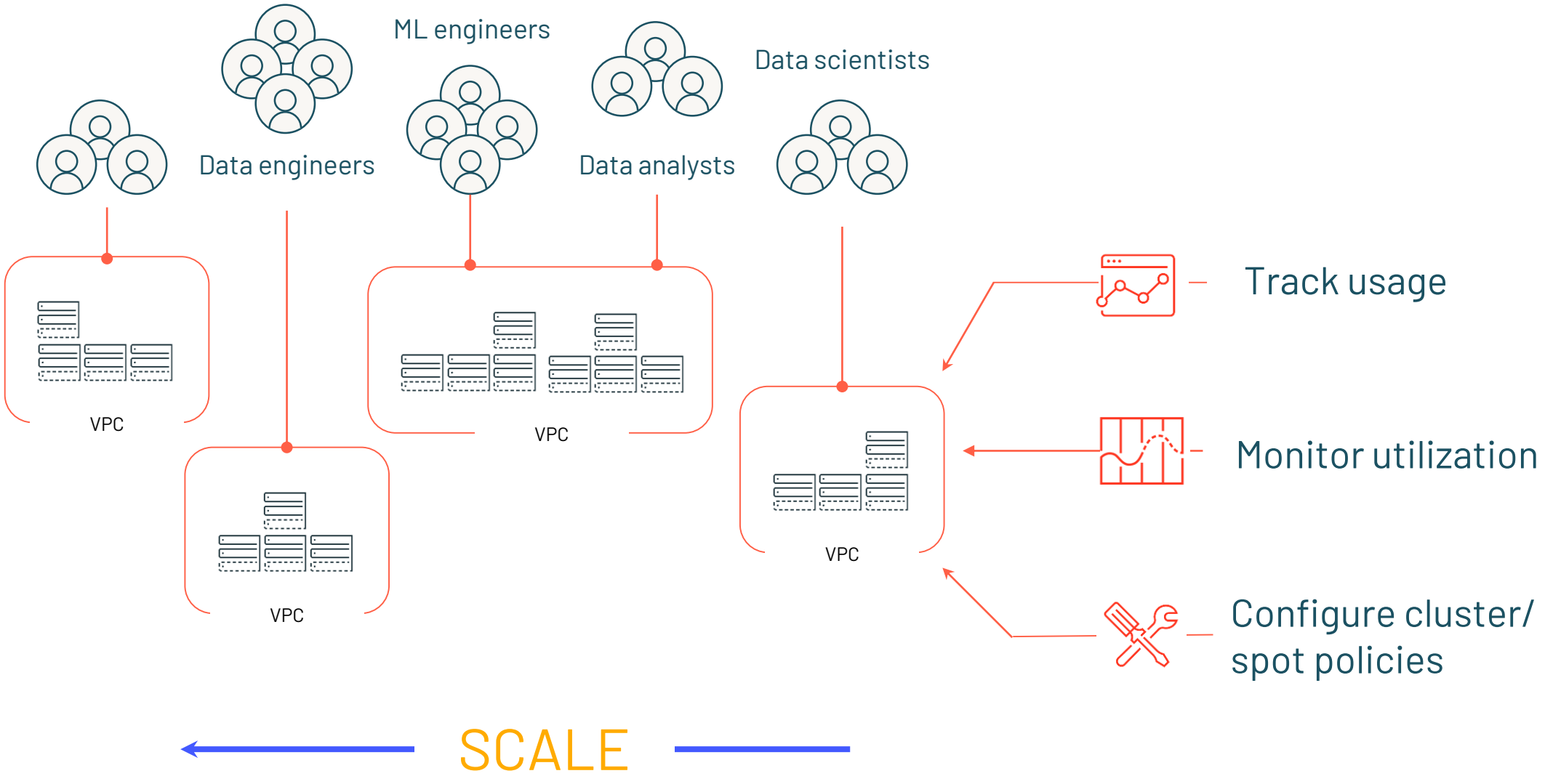
Need compute instantly without having to run it all the time

Security & compliance



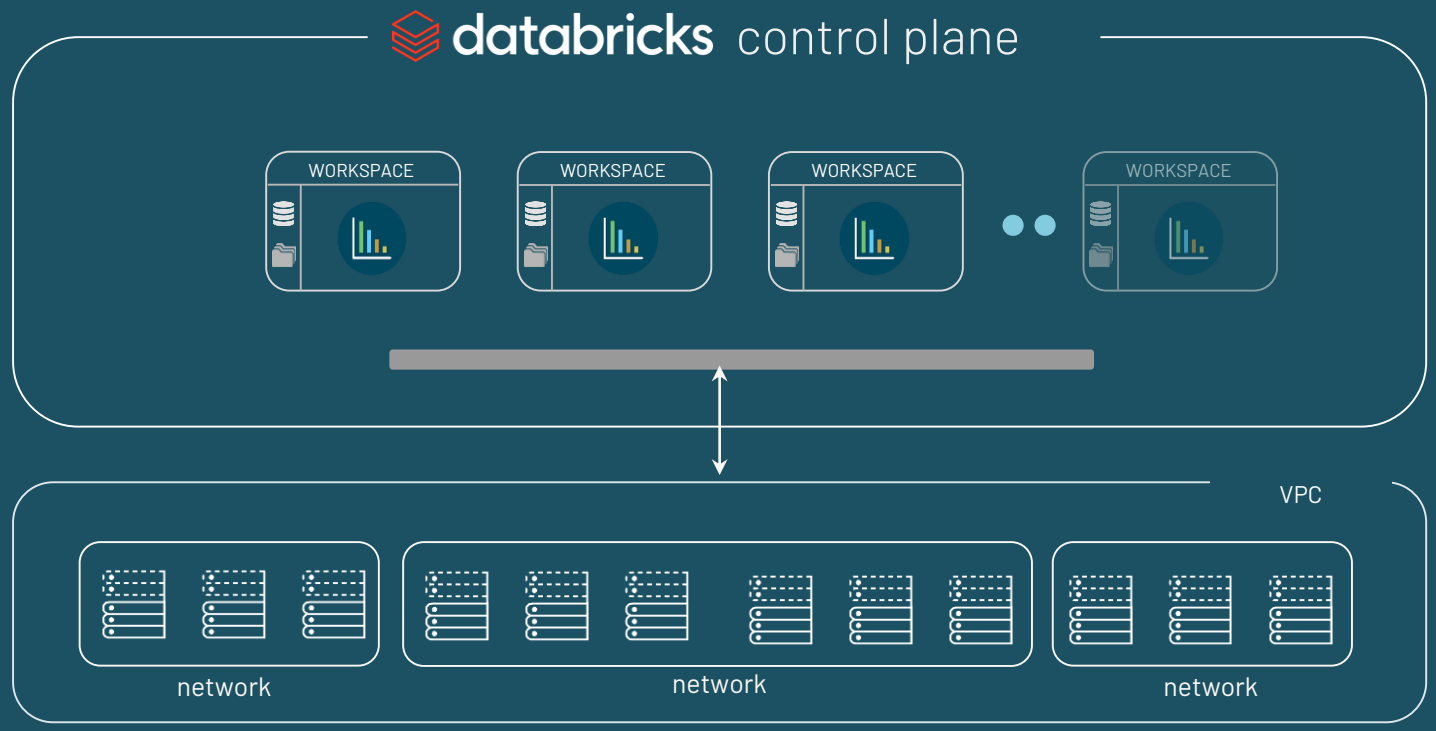
Scaling to all users requires scaling security & operations

Performance/cost management



Databricks Serverless

Databricks serverless architecture



Databricks Account

Databricks account

Security & monitoring

Regulatory compliance

Instant compute

Customer AWS Account

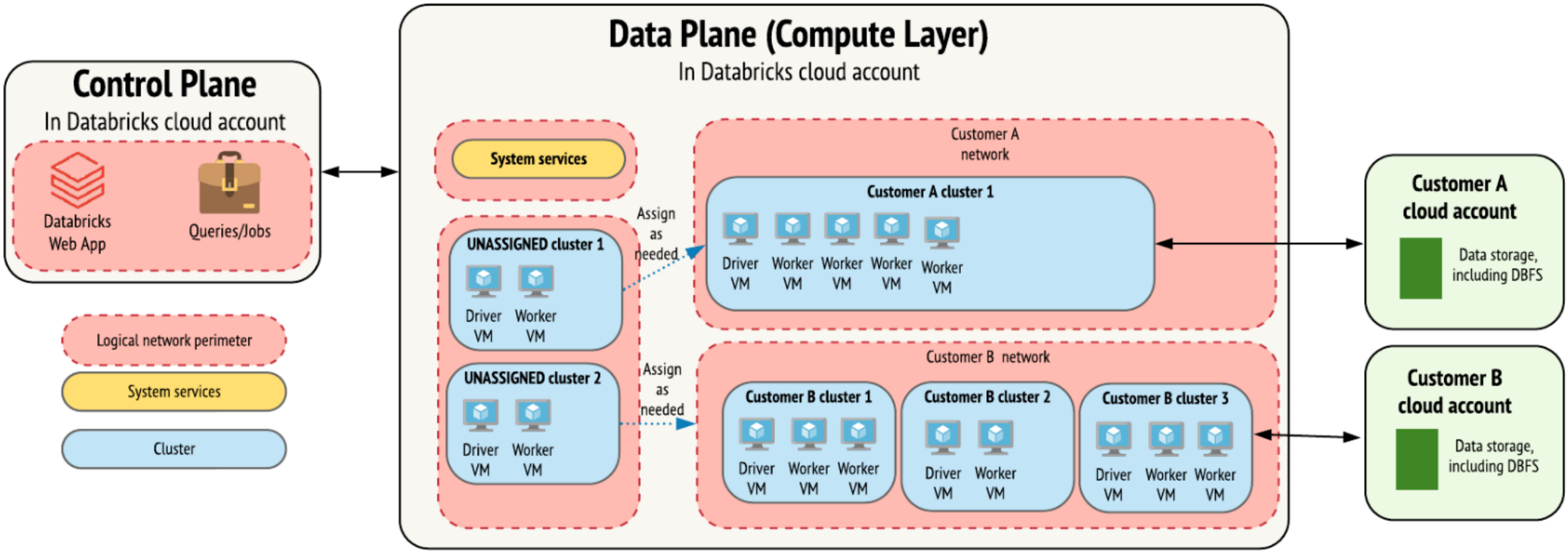
Customer Account



Amazon S3

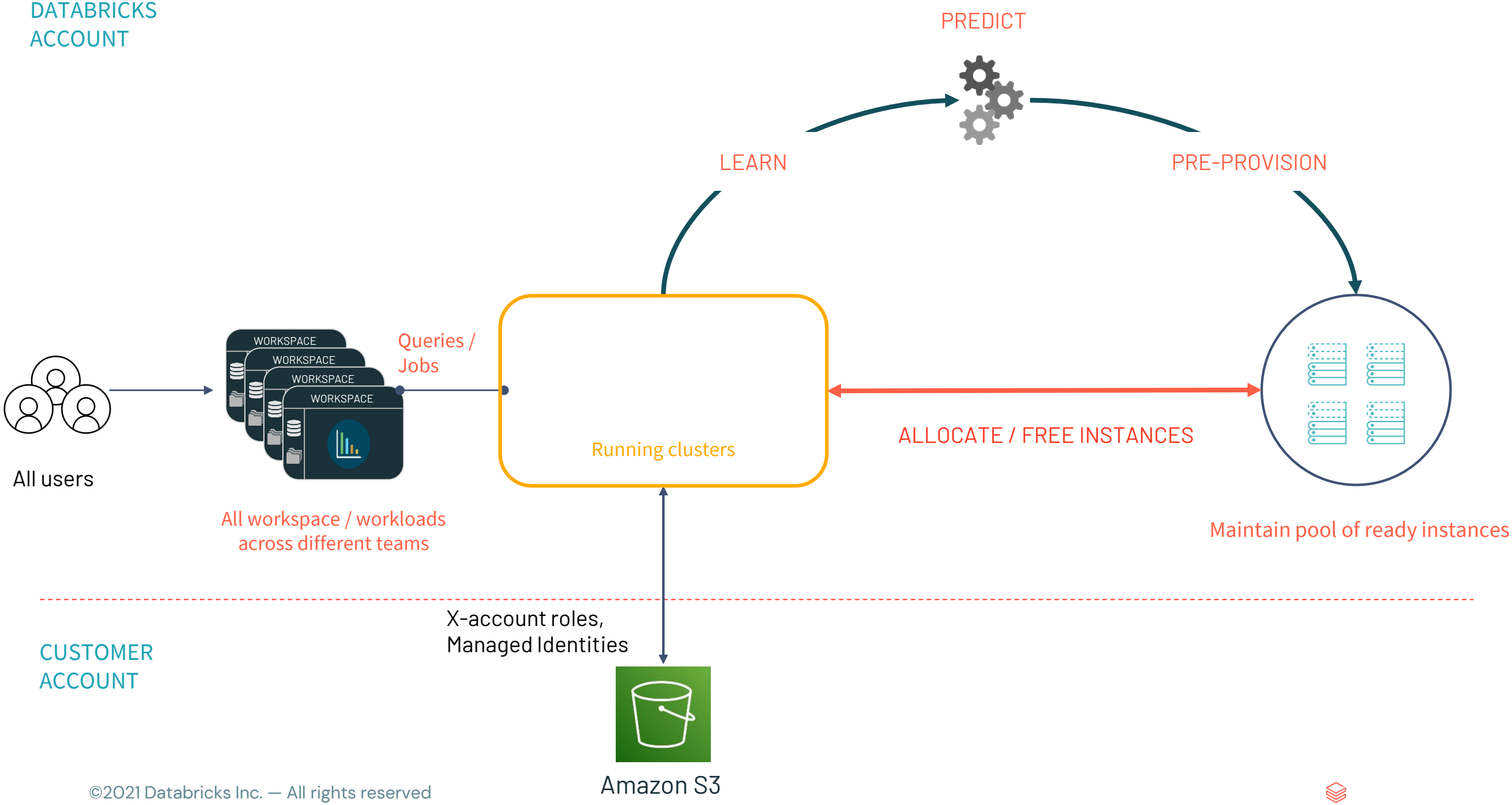
databricks

Databricks serverless architecture



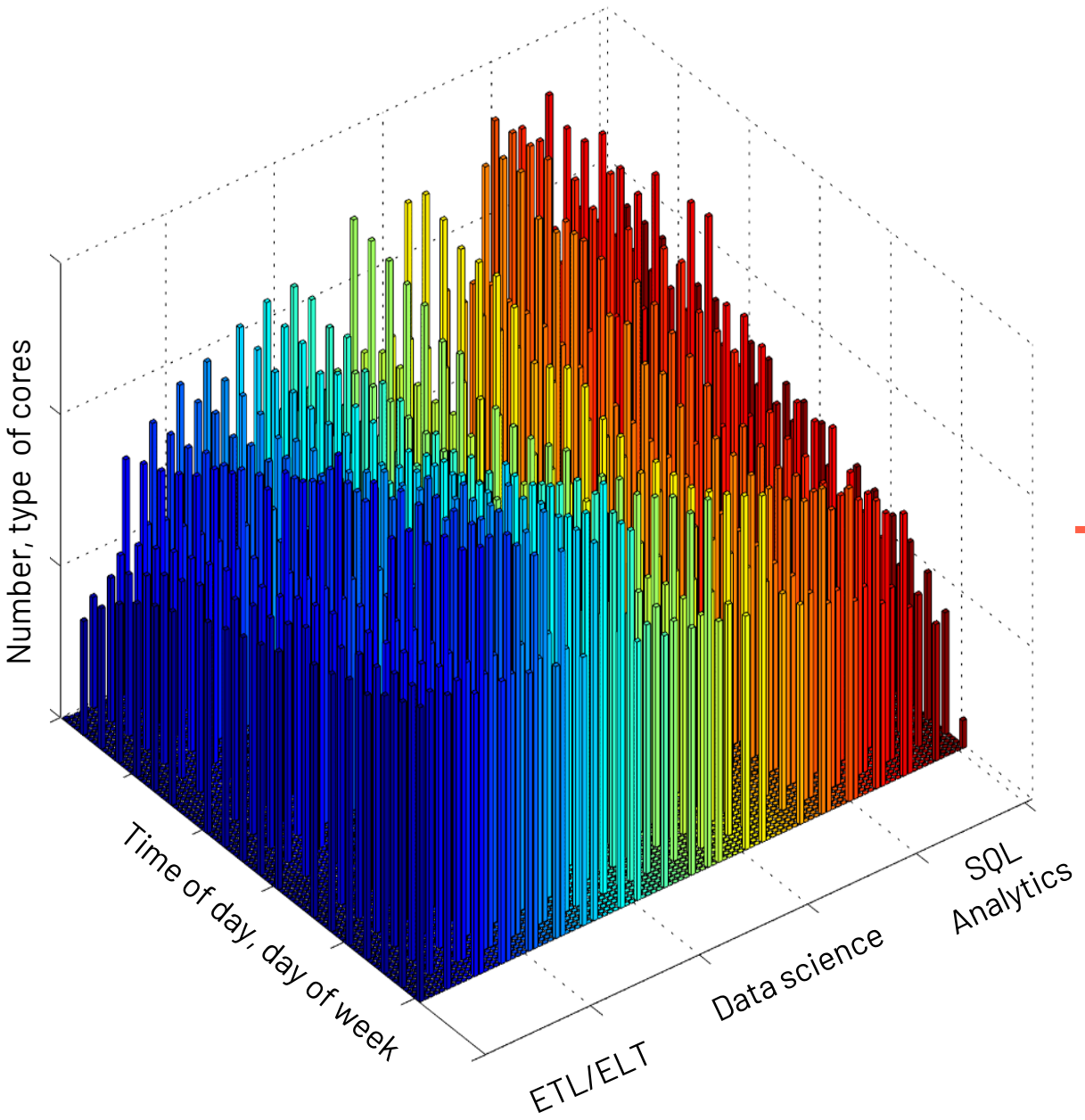
Amazon EC2, Amazon EKS

DATABRICKS ACCOUNT



CUSTOMER ACCOUNT

©2021 Databricks Inc. — All rights reserved



NEAR TERM

Scale clusters based on usage

Manage spot fleets/policies

Auto-terminate frequently to save cost

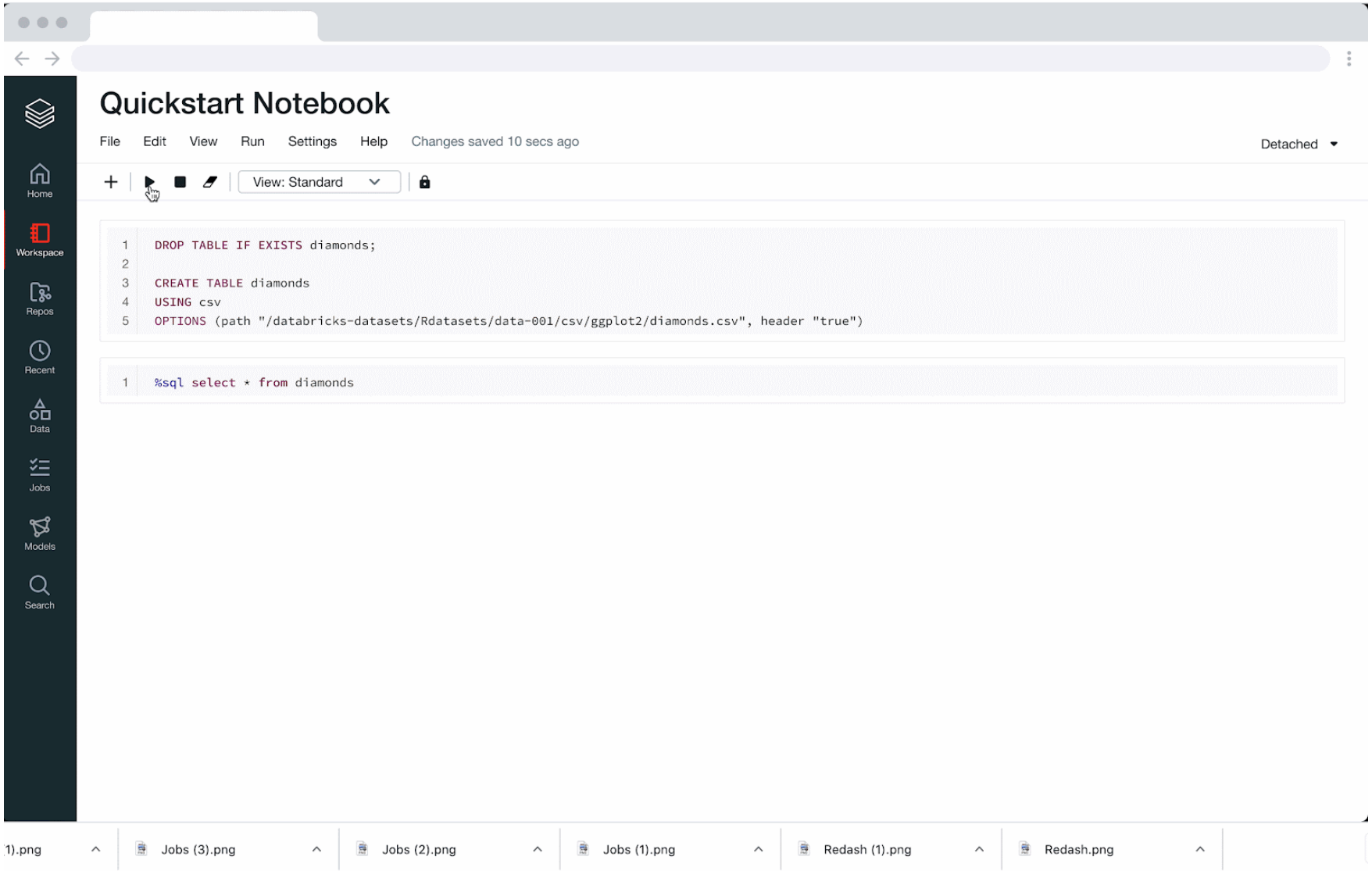
FUTURE

Automatically adjust instance/ cluster size

Efficiently schedule jobs to reduce cost

Simplicity helps you scale

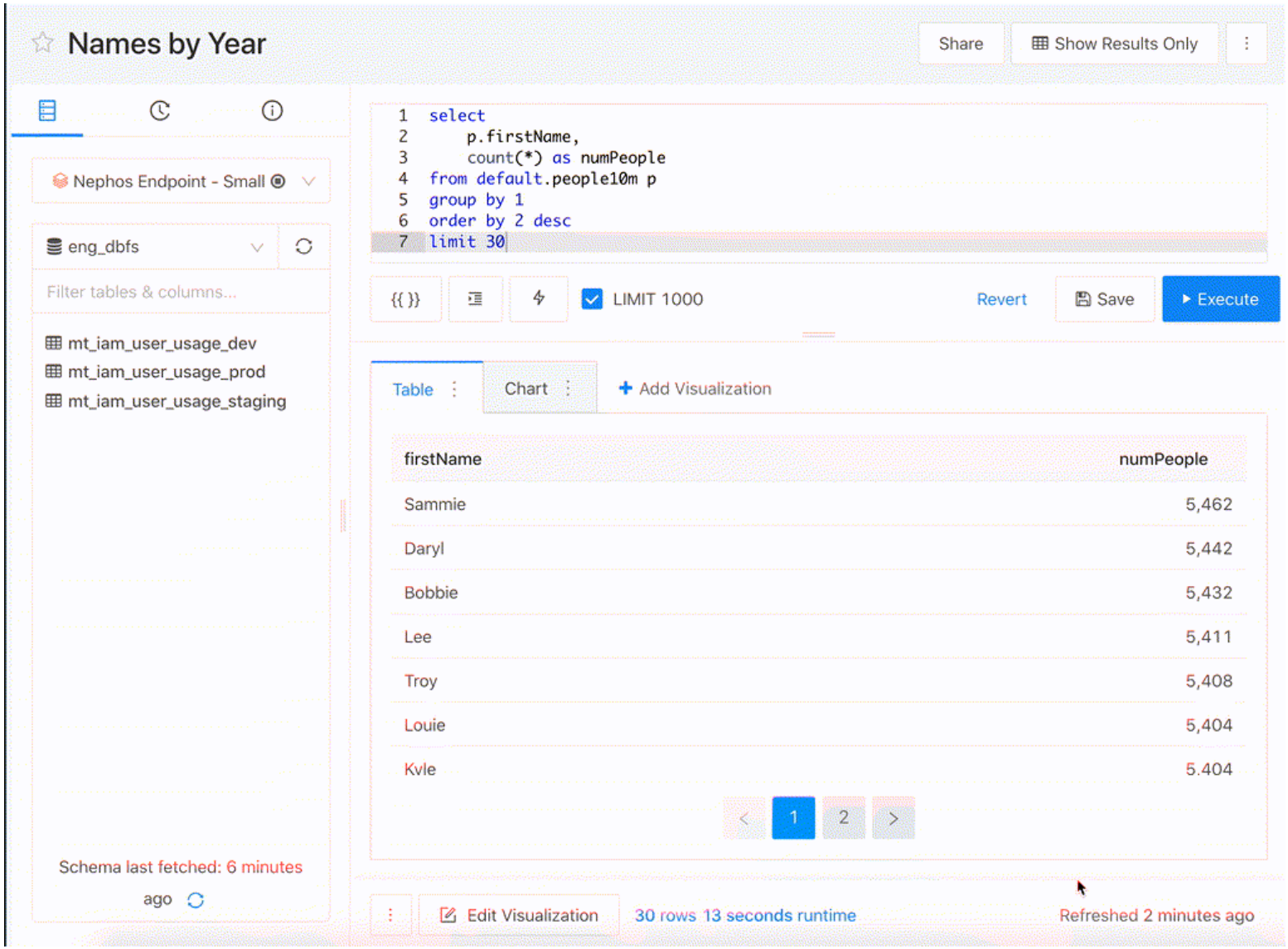
Simplifying notebooks



Single node, single user experience is the default

©2021 Databricks Inc. — All rights reserved

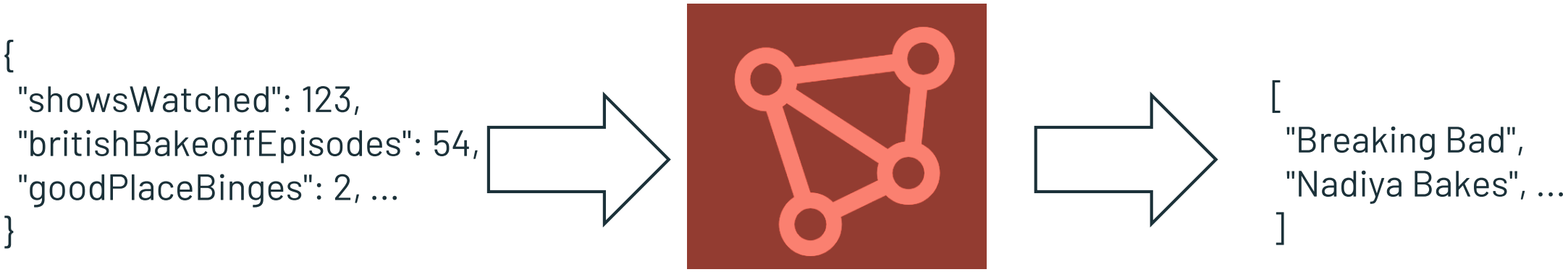
Simplifying SQL



Compute available in seconds

©2021 Databricks Inc. — All rights reserved

Simplifying scalable model serving



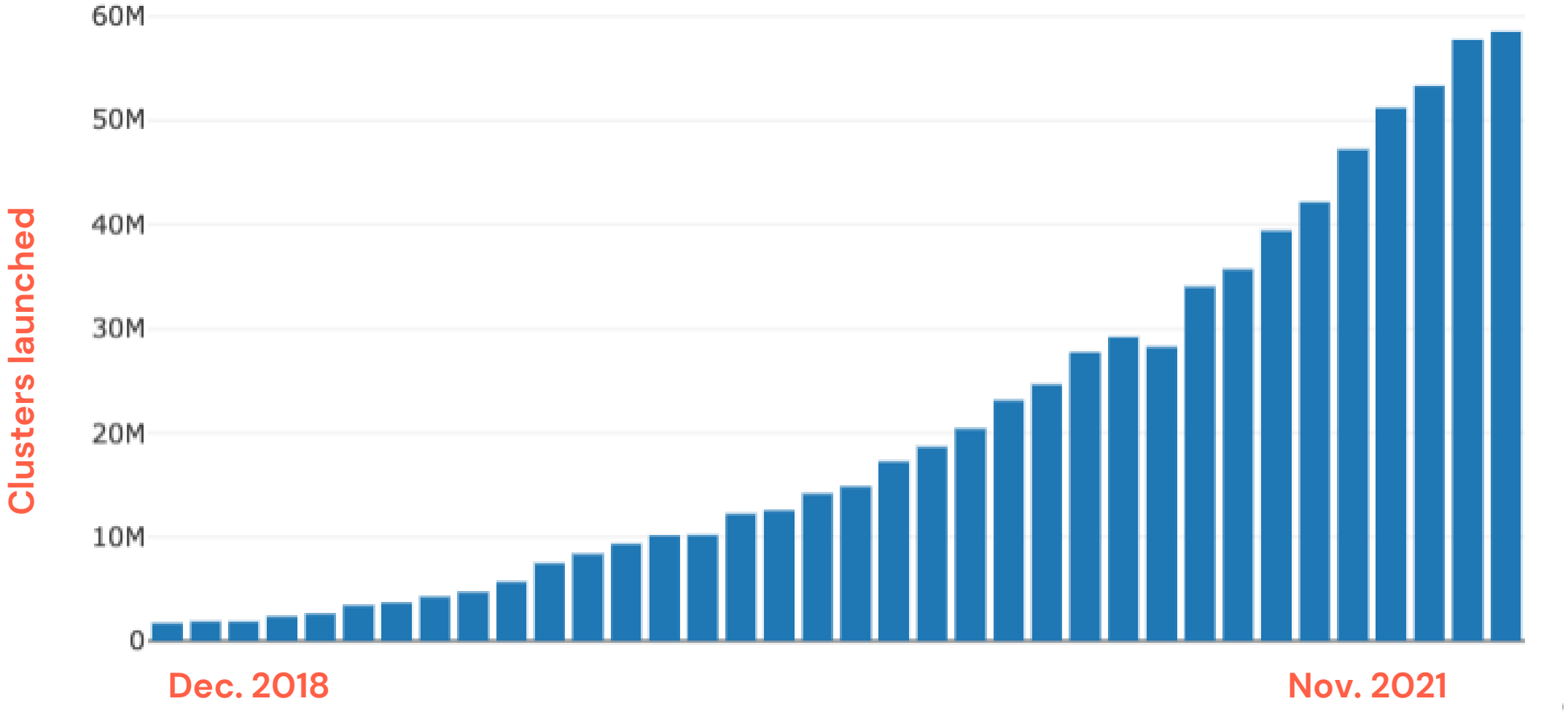
- Abstract away type of compute
- Managed availability, scalability

©2021 Databricks Inc. — All rights reserved

Simplifying data science and machine learning

- First hard part of DS & ML platform
  - Scalability, availability, security
  - Core to Databricks product
- Serverless gives ability to apply expertise to new domains
- Also provides fundamentally faster compute spin-up

©2021 Databricks Inc. — All rights reserved



©2021 Databricks Inc. — All rights reserved

Takeaways

The databricks architecture

Data processing on millions of VMs around the world in multiple clouds

Challenges & lessons

A data platform that integrates with cloud standards and quirks is key

The factory that builds and evolves the data platform is more important than the data platform itself

Operating at scale requires cost and capacity management

Data & AI accelerates data platform features, product analytics, and DevOps

Try Databricks!

www.databricks.com/try-databricks-aws

©2021 Databricks Inc. — All rights reserved



databricks

©2021 Databricks Inc. — All rights reserved



Thank you!

© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.