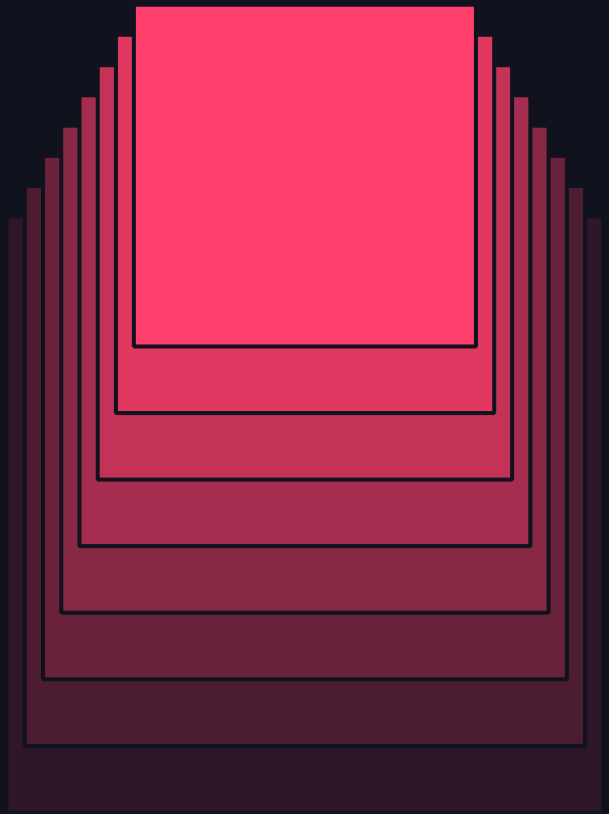
DATA+AI SUMMIT BY databricks



UNLOCKING THE LAKEHOUSE WITH EFFICIENT DATA PIPELINES

June 2024

Prabodh Mhalgi
Capital One
Usman Zubair
Databricks

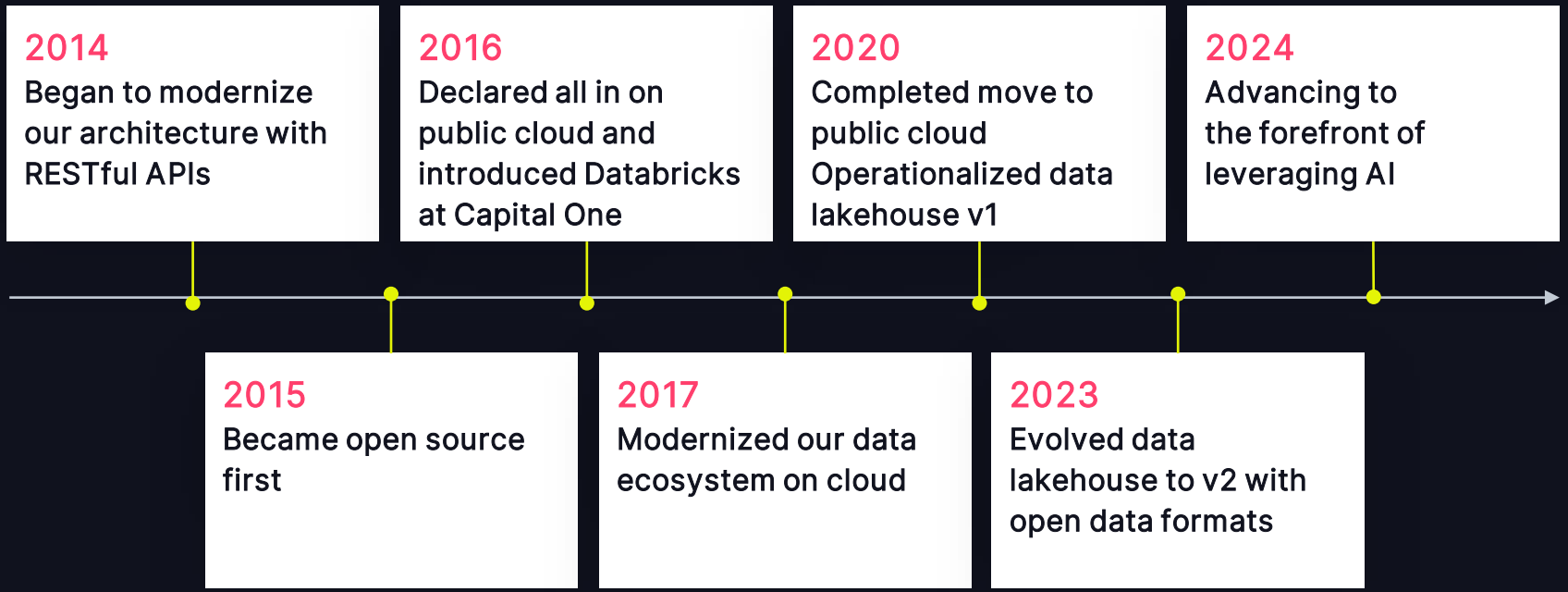
Prabodh Mhalgi
Sr. Lead Data Engineer
Capital One

Usman Zubair
Lead Technologist
Databricks

AGENDA

- Road to the Lakehouse
- Building Efficient Data Pipelines
- Key Takeaways
- The Road Ahead

CAPITAL ONE JOURNEY



CAPITAL ONE SCALE

Petabyte scale data lake

Terabytes added every day

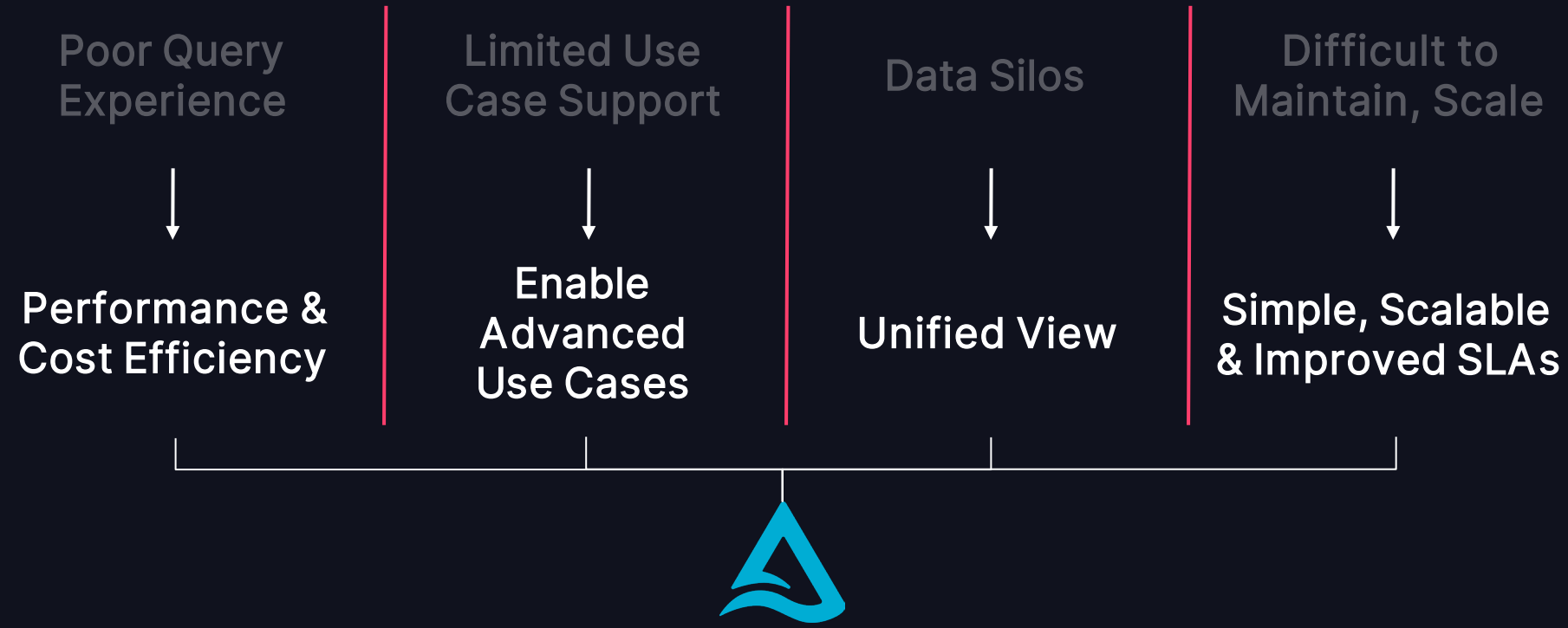
1000s of datasets

Near real-time ingestion

1000s of users

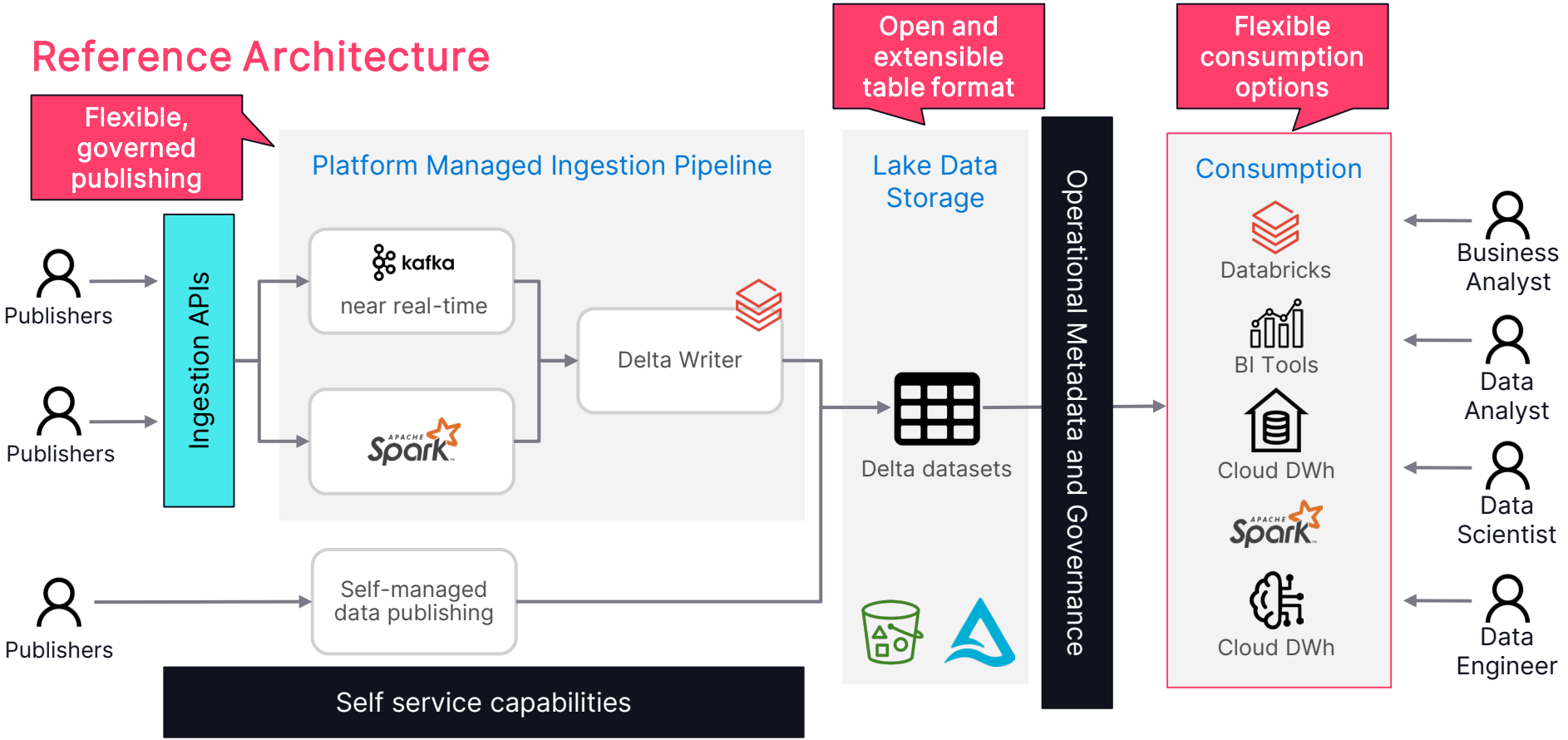
ROAD TO THE LAKEHOUSE

Challenges and Objectives



ROAD TO THE LAKEHOUSE

Reference Architecture



BUILDING EFFICIENT DATA PIPELINES

Challenges

Challenges of Scale	Lack of Versatility	Compute Utilization	Reliability
↓	↓	↓	↓
Exponential data growth after initial success	Simple batching logic Single config for all clusters	Under utilization of clusters lead to high costs	Sudden spikes caused backlogs

BUILDING EFFICIENT DATA PIPELINES

HOW DO WE OPTIMIZE OUR DATA LOAD PROCESS?

BUILDING EFFICIENT DATA PIPELINES

Optimizing the Lakehouse

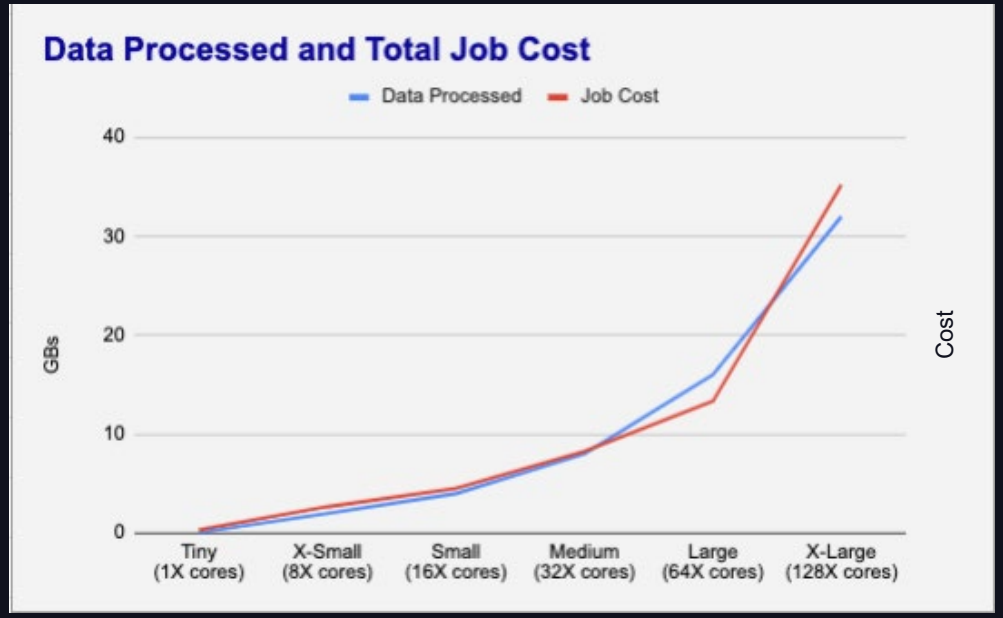
Compute Size	# of Worker Cores	# of Datasets	Total Data Size
Tiny	1X	10	10MB - 100MB
X-Small	8X	1 / 2 / 5	2GB
Small	16X	1 / 5	4GB
Medium	32X	1 / 5 / 10	8GB
Large	64X	1 / 5 / 10	16GB
X-Large	128X	1	32GB

Objective

Processing time should be <= ~5 mins

BUILDING EFFICIENT DATA PIPELINES

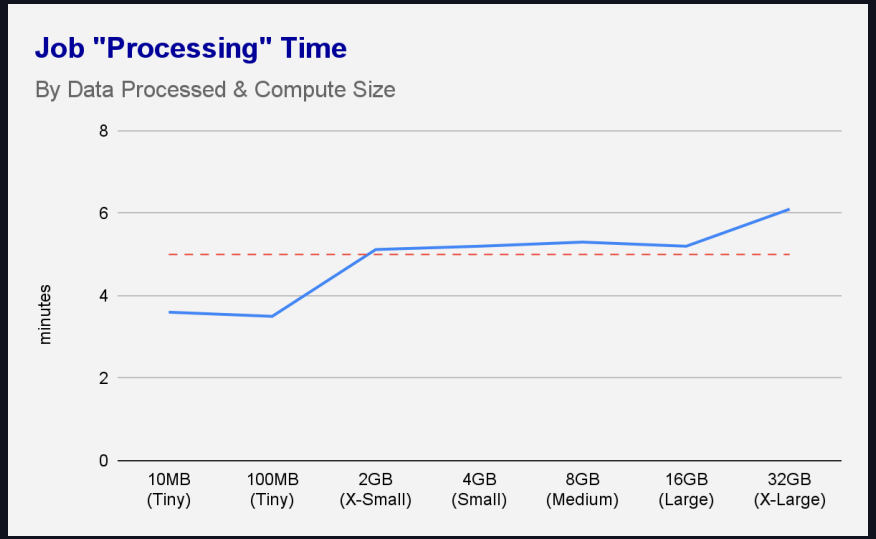
Optimizing the Lakehouse

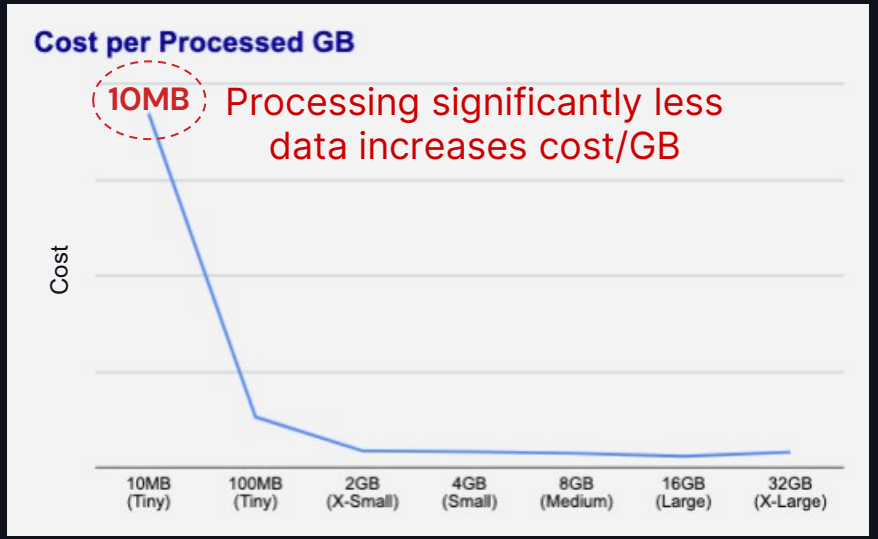


Job cost is closely proportional to amount of data processed

BUILDING EFFICIENT DATA PIPELINES

Optimizing the Lakehouse





BUILDING EFFICIENT DATA PIPELINES

Optimizing the Lakehouse

60%

reduction in compute costs

CAPITAL ONE SCALE

Petabyte scale data lake

Terabytes added every day

1000s of datasets

Near real-time ingestion

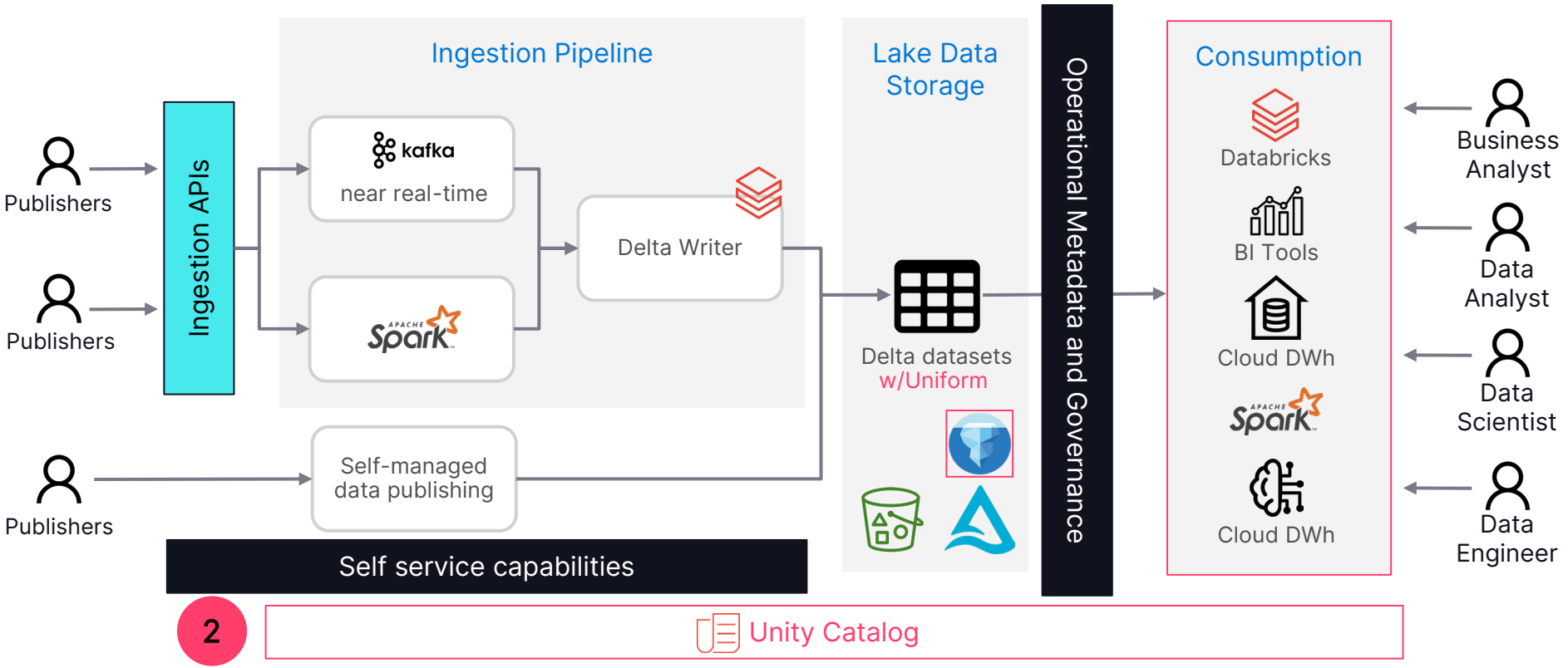
1000s of users

KEY TAKEAWAYS

- Optimize your compute utilization
- Leverage Databricks Workflows with job clusters
- Ensure interoperability with open table formats
- Build single, unified view of data

THE ROAD AHEAD

1 Delta Uniform



DATA+AI SUMMIT BY databricks

QUESTIONS?

DATA+AI SUMMIT BY databricks

THANK YOU