DATA+AI
SUMMIT 2022

Streaming on the Lakehouse

Wednesday, June 29 - 2:50 PM

Ray Zhu
Senior Staff PM, Databricks

Steven Yu
Principal SA, Databricks

Erica Lee
VP ML, Upwork

ORGANIZED BY databricks

Product Safe Harbor Statement

This information is provided to outline Databricks' general product direction and is for informational purposes only. Customers who purchase Databricks services should make their purchase decisions relying solely upon services, features, and functions that are currently available. Unreleased features or functionality described in forward-looking statements are subject to change at Databricks discretion and may not be delivered as planned or at all.

Streaming Data

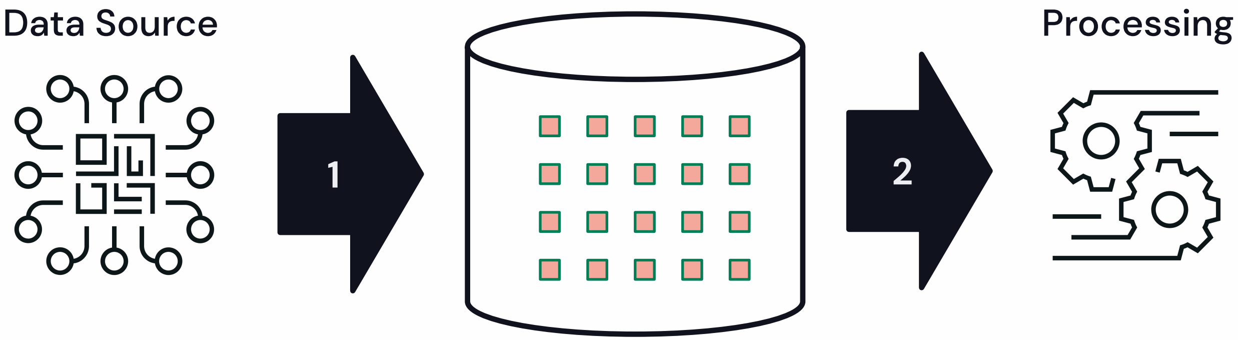
Continuously generated and unbounded data

- DB Change Data Feeds
- Clickstreams
- Machine & Application Logs
- Application Events
- Mobile & IoT Data

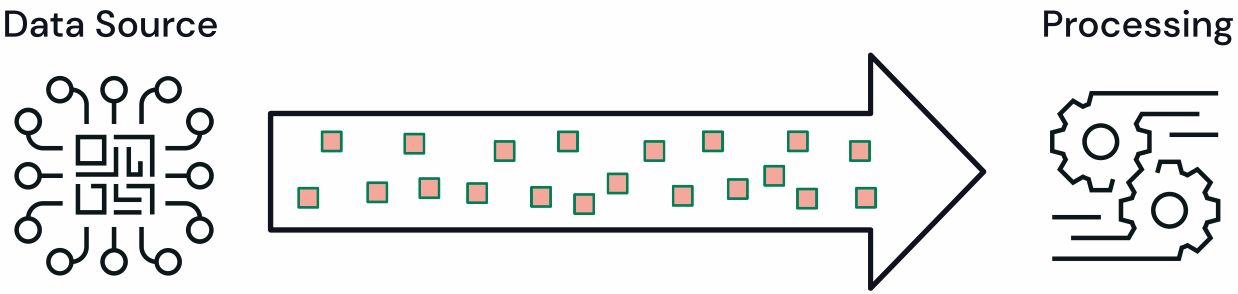
The vast majority of the data in the world is streaming data!

Stream Processing

Traditional Processing is one-off and bounded



Stream Processing is continuous and unbounded



Technical Advantages

A more intuitive way of capturing and processing continuous and unbounded data

Lower latency for time sensitive applications and use cases

Better fault-tolerance through checkpointing

Higher compute utilization and scalability through continuous and incremental processing

Business Benefits

BI and SQL Analytics

Fresher and faster insights

→

Quicker and better business decisions

Data Engineering

Sooner availability of cleaned data

→

More business use cases

Data Science and ML

More frequent model update and inference

→

Better model efficacy

Event Driven Application

Faster customized response and action

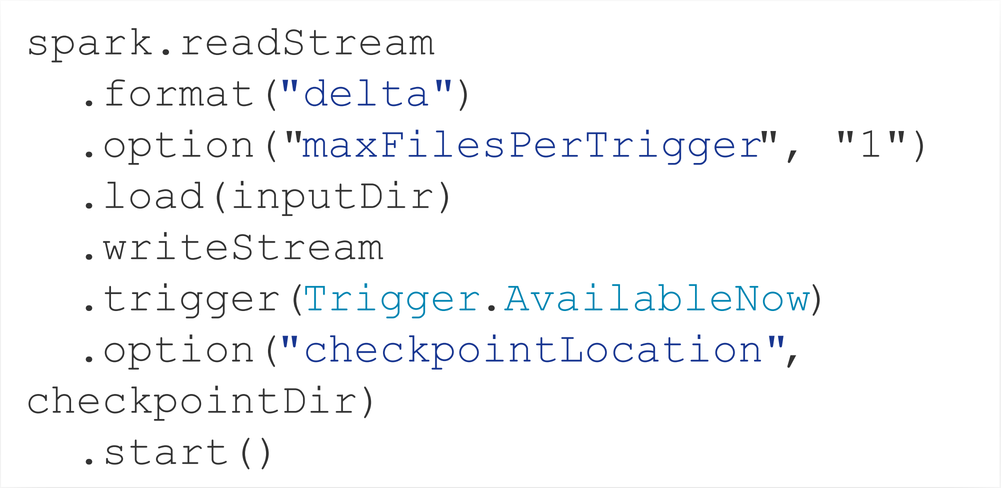
→

Better and differentiated customer experience

Common Misconceptions

Misconception #1

X Stream processing is only for low latency use cases



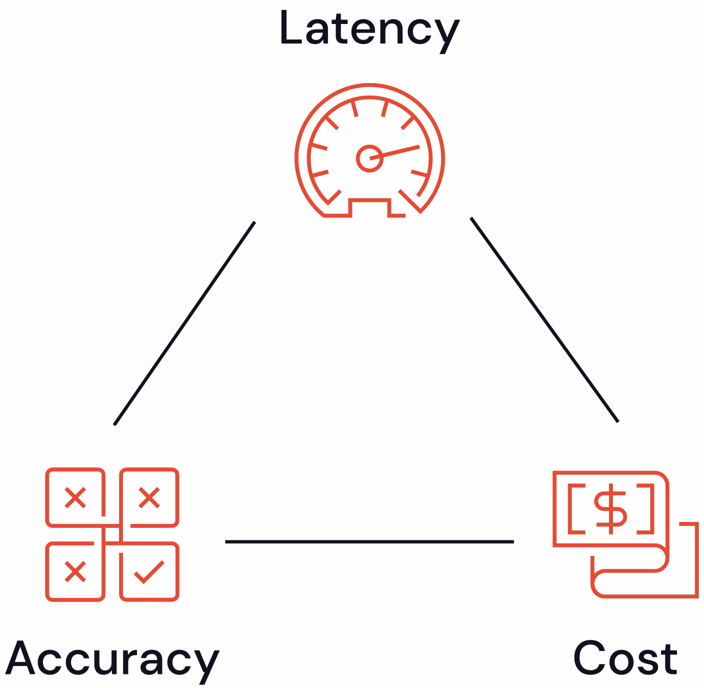
```
spark.readStream
   .format("delta")
   .option("maxFilesPerTrigger", "1")
   .load(inputDir)
   .writeStream
   .trigger(Trigger.AvailableNow)
   .option("checkpointLocation",
checkpointDir)
   .start()
```

Stream processing can be applied to use cases of any latency

“Batch” is a special case of streaming

Misconception #2

X The lower the latency, the better



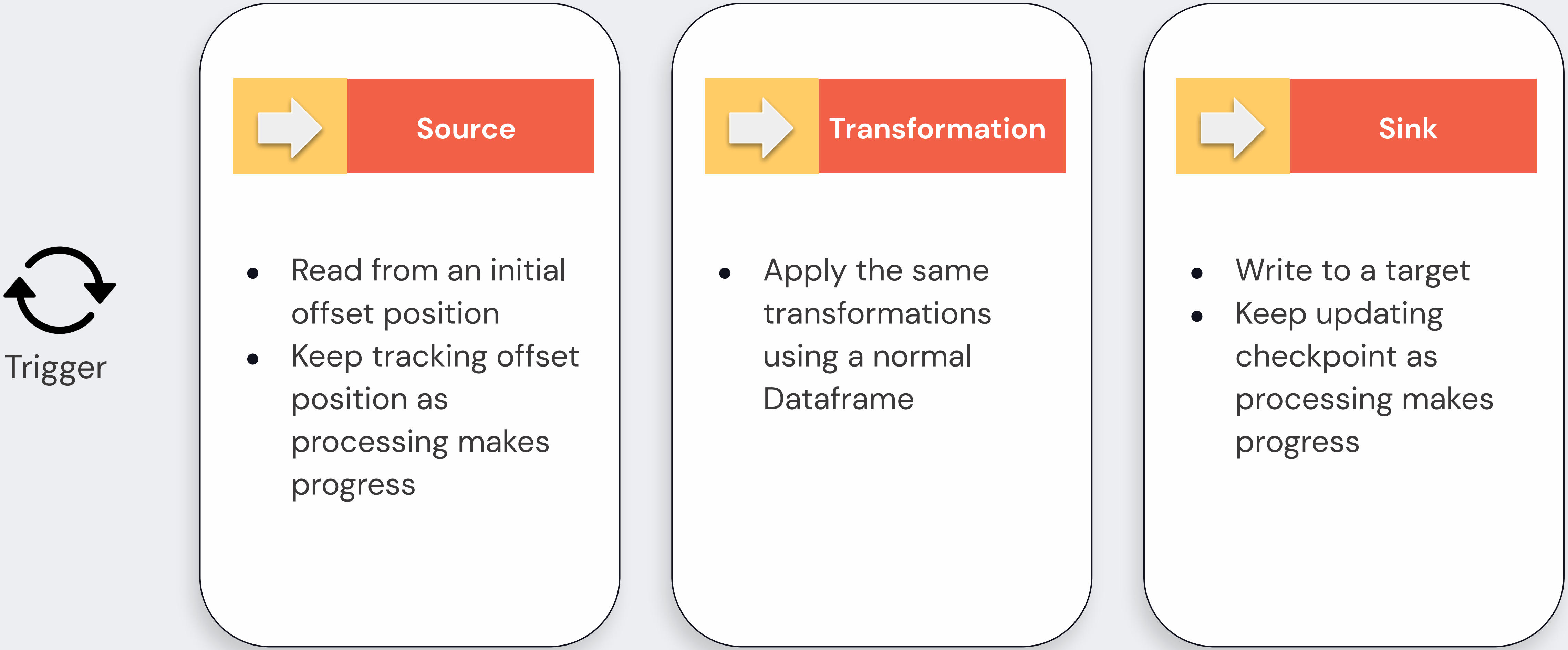
Choose the right latency, accuracy, and cost tradeoff for each specific use case

Stream Processing with Structured Streaming

Structured Streaming

A scalable and fault-tolerant stream processing engine built on the Spark SQL engine

Structured Streaming



Trigger

Source

- Read from an initial offset position
- Keep tracking offset position as processing makes progress

Transformation

- Apply the same transformations using a normal Dataframe

Sink

- Write to a target
- Keep updating checkpoint as processing makes progress

Source

```
spark.readStream.format(<source>)
.option(<>,<>)...
.load()
```

Transformation

```
spark.readStream.format(<source>)
.option(<>,<>)...
.load()
.select(cast("string").alias("jsonData"))
.select(from_json($"jsonData",jsonSchema).alias("payload"))
```

Sink

```
spark.readStream.format(<source>)
.option(<>,<>)...
.load()
.select(cast("string").alias("jsonData"))
.select(from_json($"jsonData",jsonSchema).alias("payload"))
.writeStream
.format("delta")
.option("path",...)
```

Configuration

```
spark.readStream.format(<source>)
.option(<>,<>)...
.load()
.select(cast("string").alias("jsonData"))
.select(from_json($"jsonData",jsonSchema).alias("payload"))
.writeStream
.format("delta")
.option("path",...)
.trigger("30 seconds")
.option("checkpointLocation",...)
.start()
```

Trigger Types

- **Default:** Process as soon as the previous batch has been processed
- **Fixed interval:** Process at a user-specified time interval
- **One-time:** Process all of the available data and then stop

Output Modes

- **Append (Default):** Only new rows added to the result table since the last trigger will be output to the sink

- **Complete:** The whole result table will be output to the sink after each trigger

- **Update:** Only the rows updated in the result table since the last trigger will be output to the sink

Structured Streaming Benefits

Unified Batch and Streaming

Unified API makes development and maintenance simple

High Throughput

Optimized for high throughput and low cost

Exactly Once Semantics

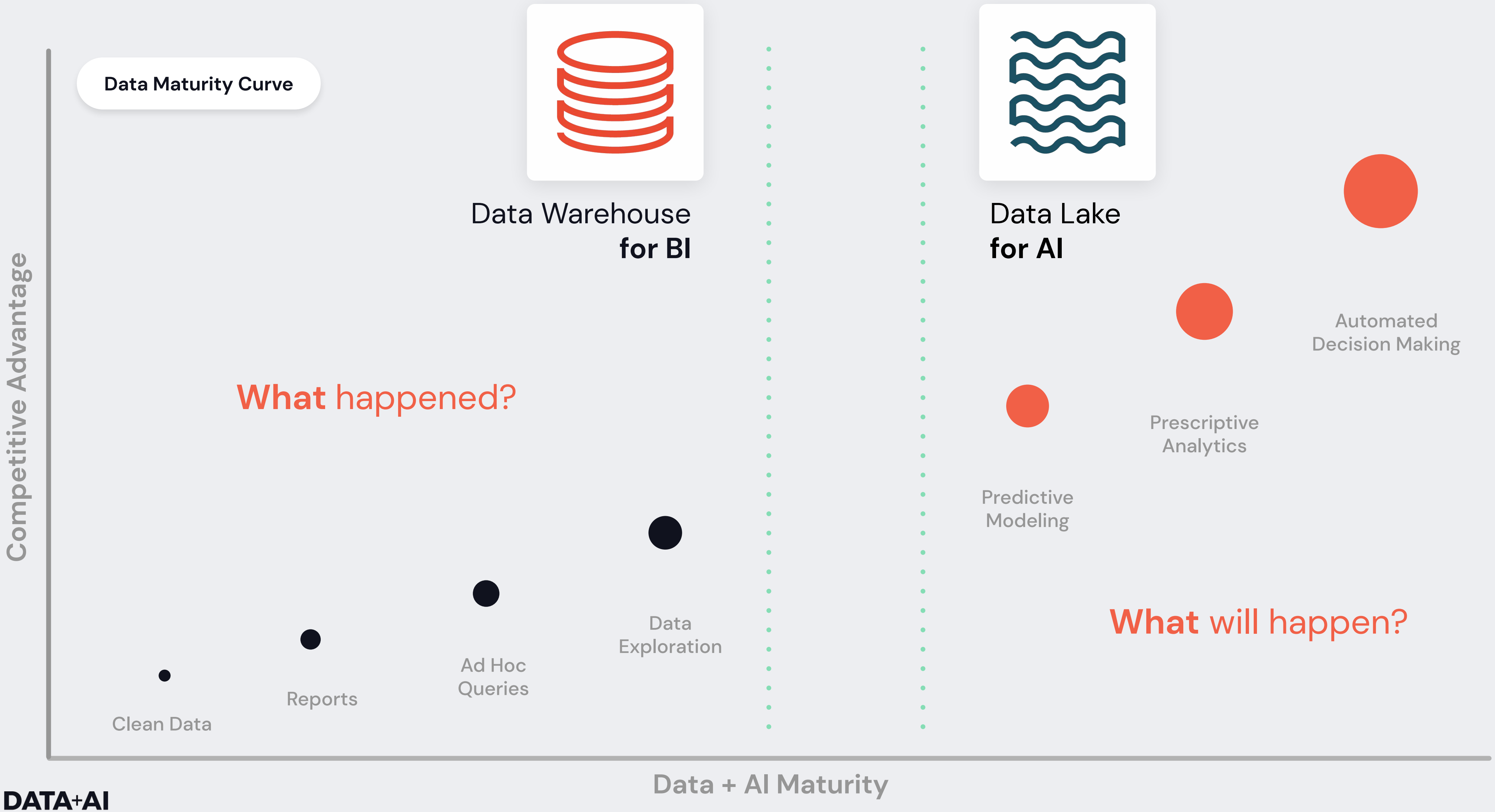
Fault-tolerance and exactly once semantics guarantee correctness

Rich Connector Ecosystem

Streaming connectors ranging from message buses to object storage services

Streaming on the Lakehouse

Realizing this requires two disparate, incompatible data platforms



Data Maturity Curve

Data Warehouse
for BI

Data Lake
for AI

What happened?

What will happen?

Clean Data

Reports

Ad Hoc
Queries

Data
Exploration

Predictive
Modeling

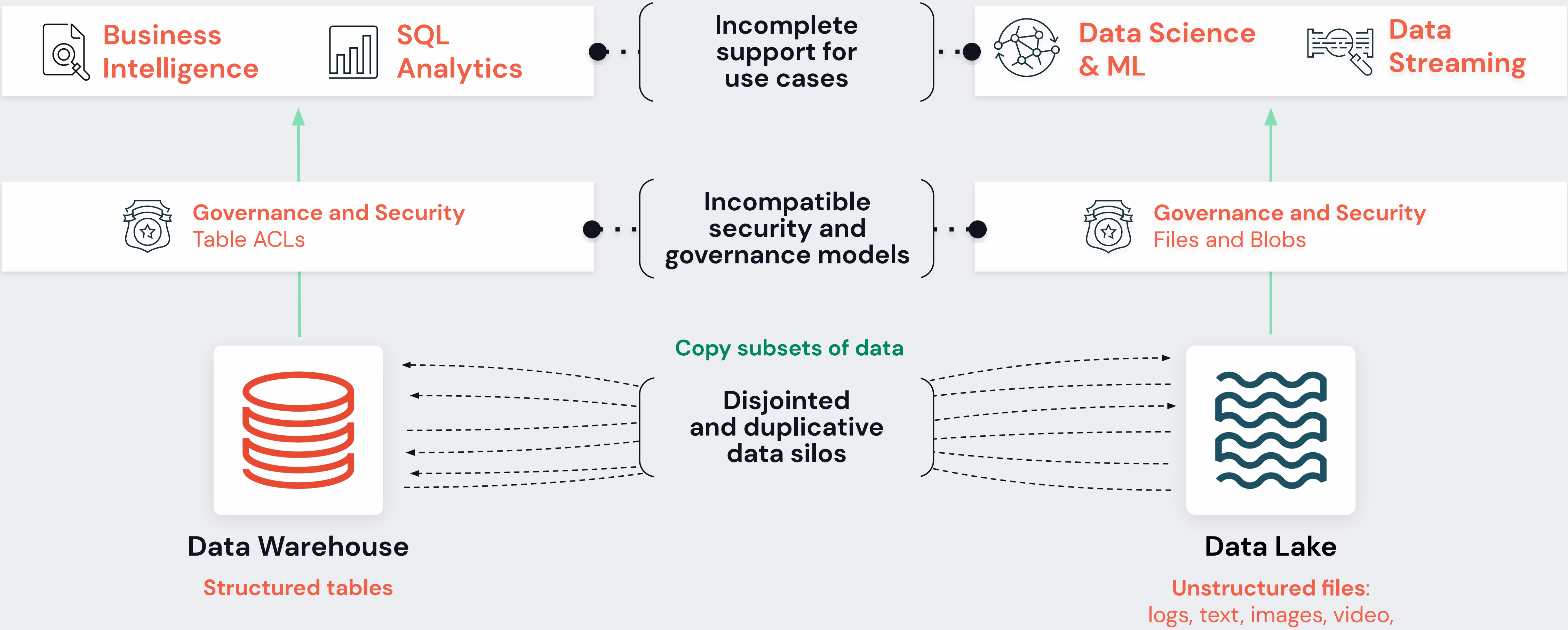
Prescriptive
Analytics

Automated
Decision Making

Competitive Advantage

Data + AI Maturity

Realizing this requires two disparate, incompatible data platforms



Business Intelligence

SQL Analytics

Incomplete support for use cases

Data Science & ML

Data Streaming

Governance and Security
Table ACLs

Incompatible security and governance models

Governance and Security
Files and Blobs

Copy subsets of data

Disjointed and duplicative data silos

Data Warehouse

Structured tables

Data Lake

Unstructured files:
logs, text, images, video,

DATA+AI
SUMMIT 2022

Realizing this requires two di incompatible data platforms

Business Intelligence

ence

Data Streaming

Incomplete support for use cases

Governance and
Table ACLs

ernance and Security
and Blobs

Incompatible security and governance models

Disjointed and duplicative data silos

arehou

Structured tabl

Data Lake

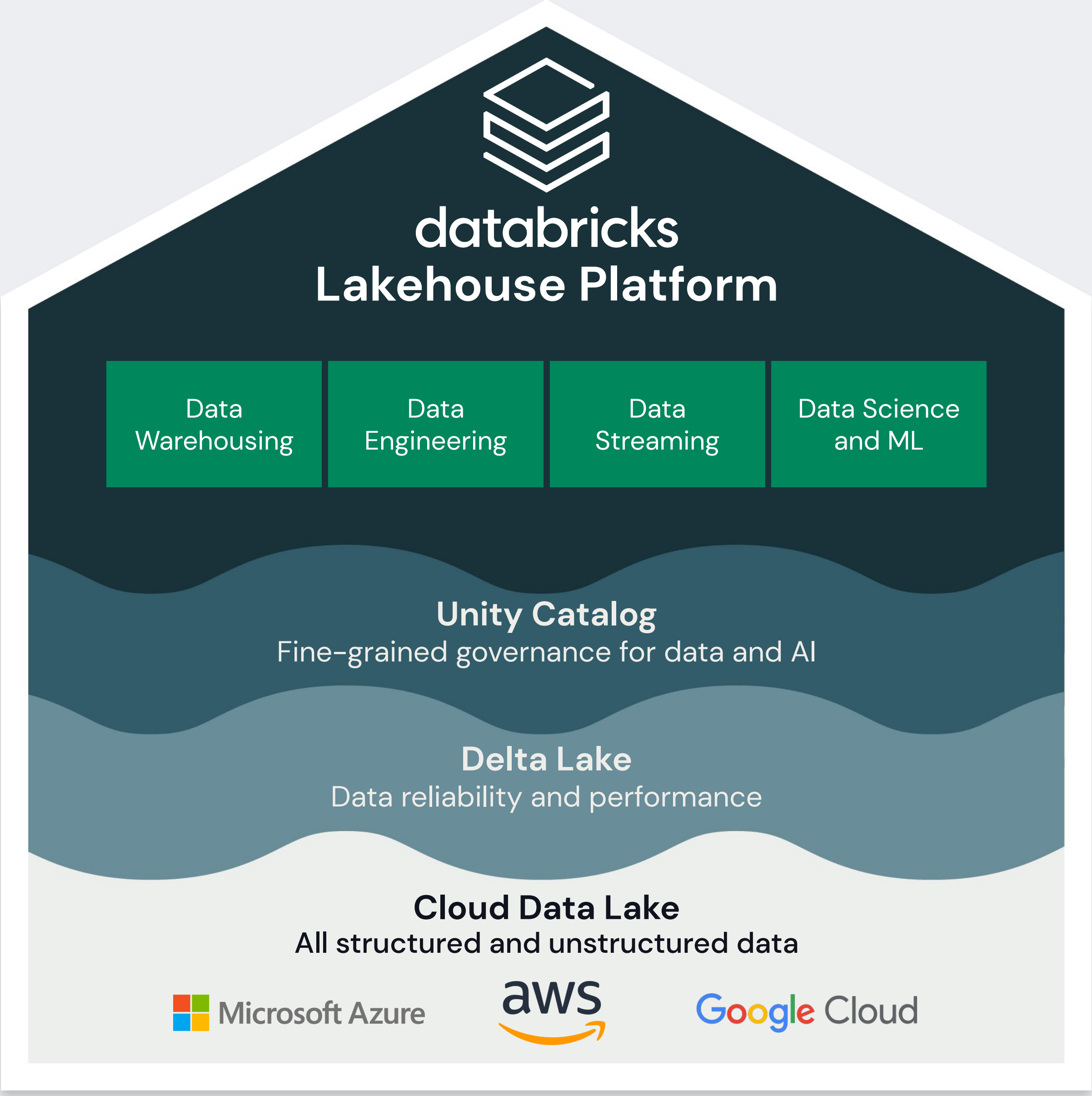
structured files:
text, images, video,

Lakehouse Platform

All machine learning, SQL, BI, and streaming use cases

One security and governance approach for all data assets on all clouds

An open and reliable data platform to efficiently handle all data types



Databricks Lakehouse Platform

Simple

Unify your data warehousing and AI use cases on a single platform

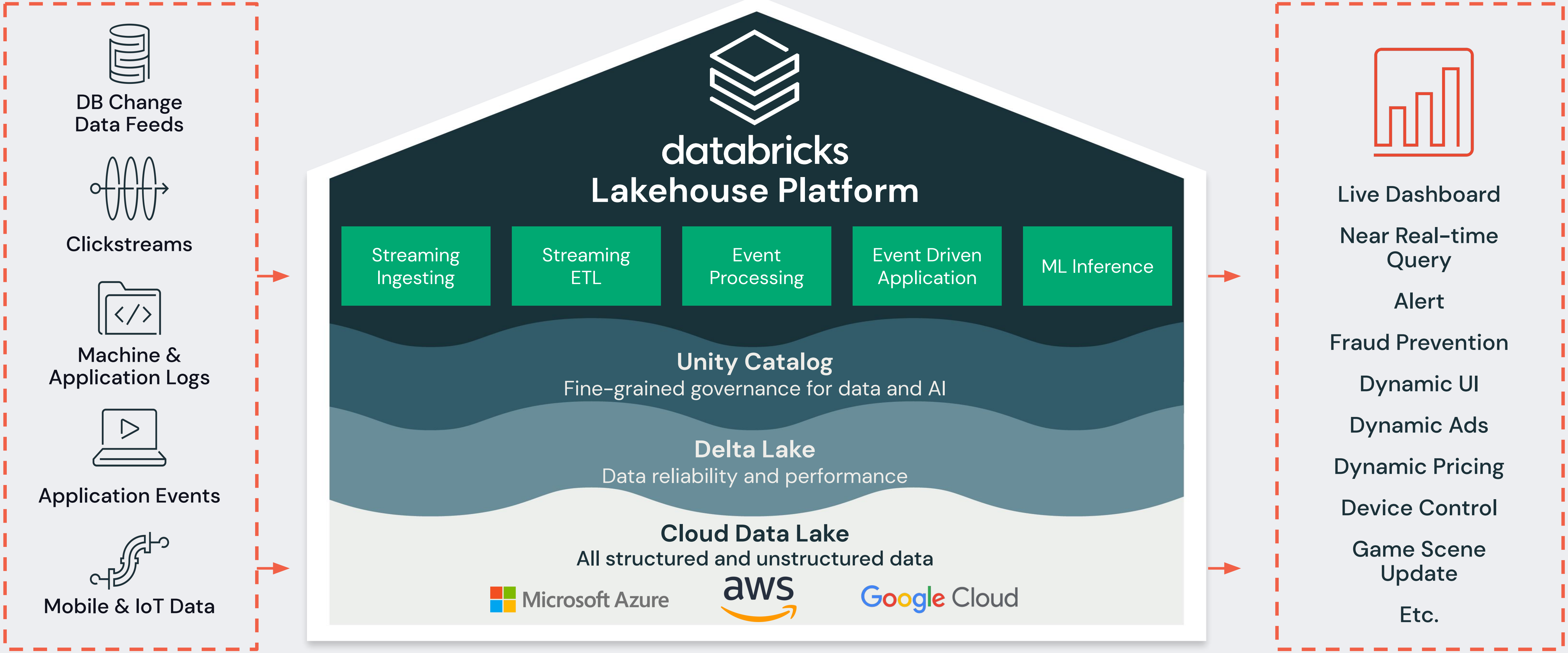
Open

Built on open source and open standards

Multicloud

One consistent data platform across clouds

Streaming on the Lakehouse



Lakehouse Differentiations

Unified Batch and Streaming

No overhead of learning, developing on, or maintaining two sets of APIs and data processing stacks

Favorite Tools

Provide diverse users with their favorite tools to work with streaming data, enabling the broader organization to take advantage of streaming

Optimal Cost Structure

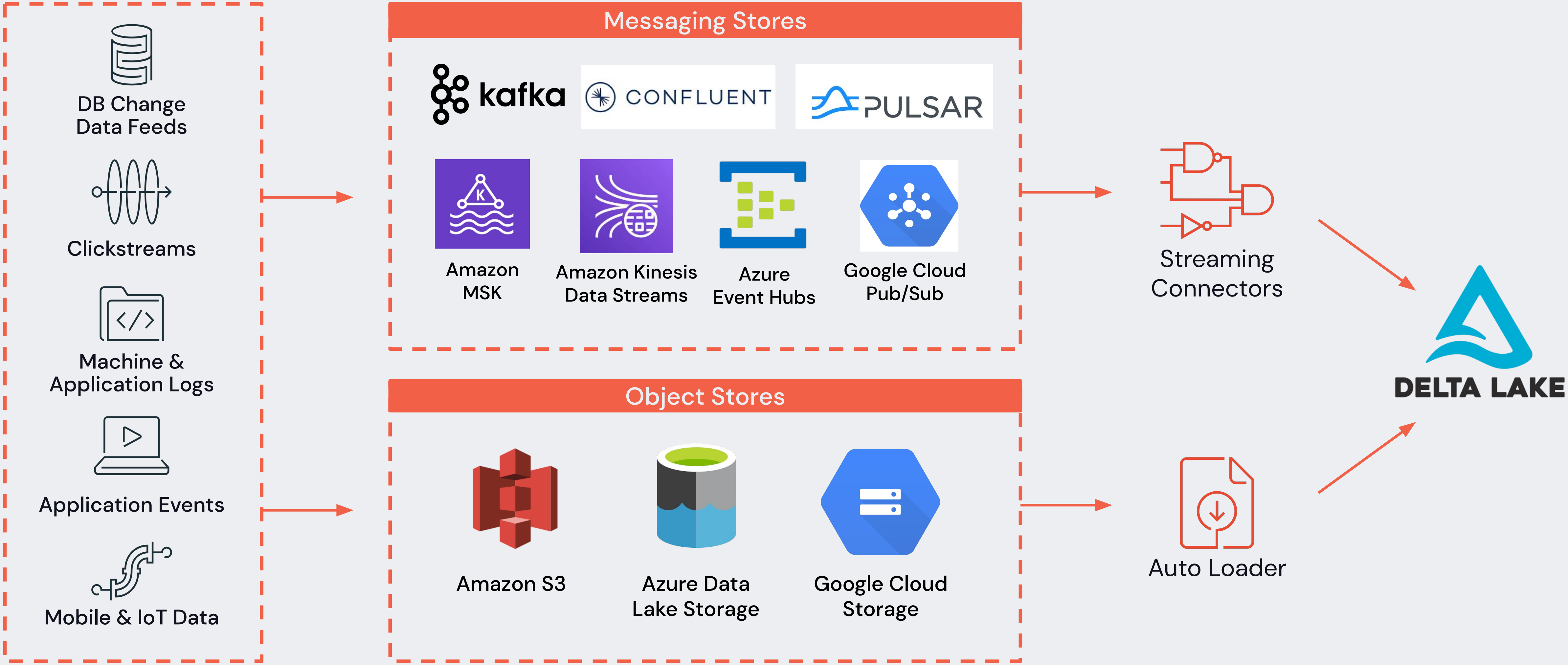
Easily configure the right latency-cost tradeoff for each of your streaming workloads

End-to-End Streaming

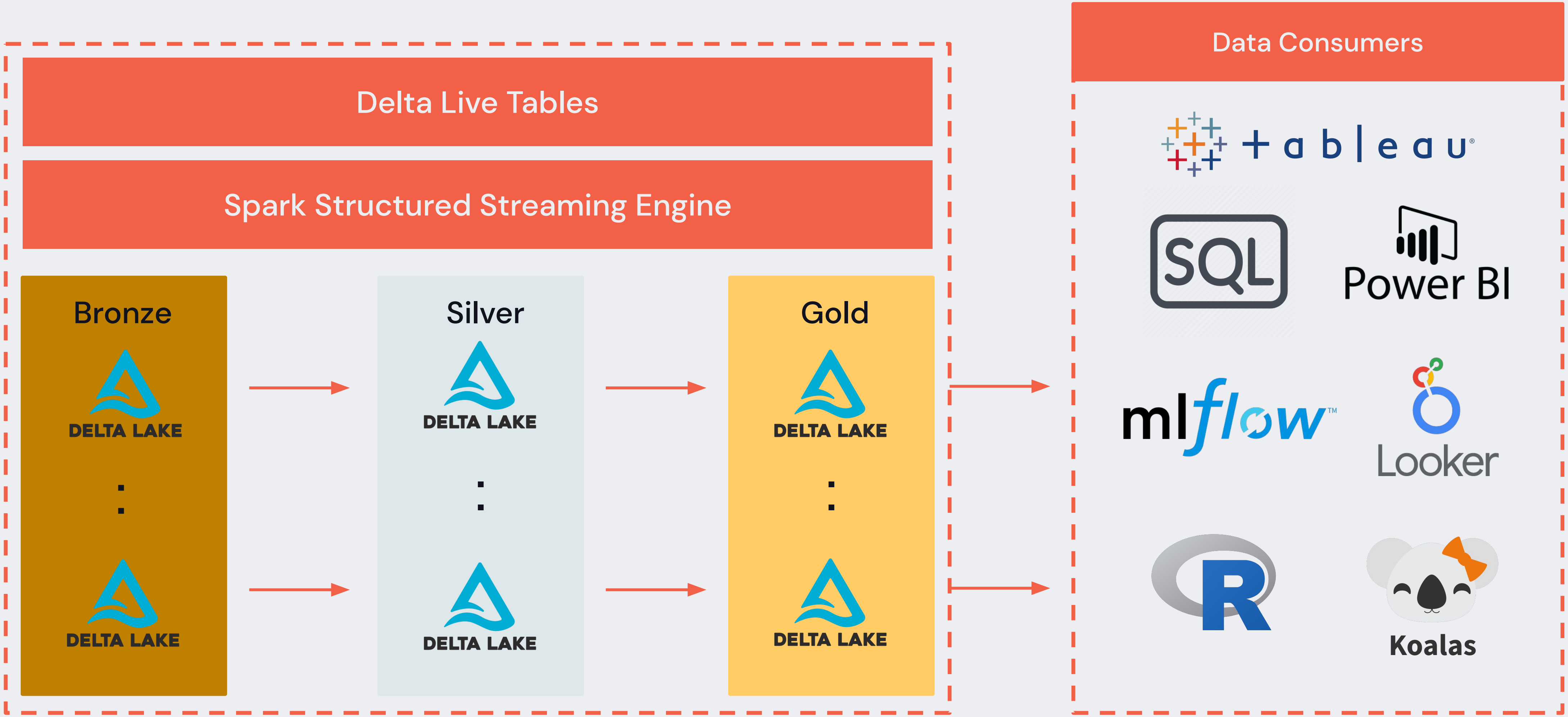
Has everything you need, no need to stitch together different streaming technology stacks or tune them to work together

Streaming Patterns on the Lakehouse

Streaming Ingestion



Streaming ETL



Streaming ETL Choices

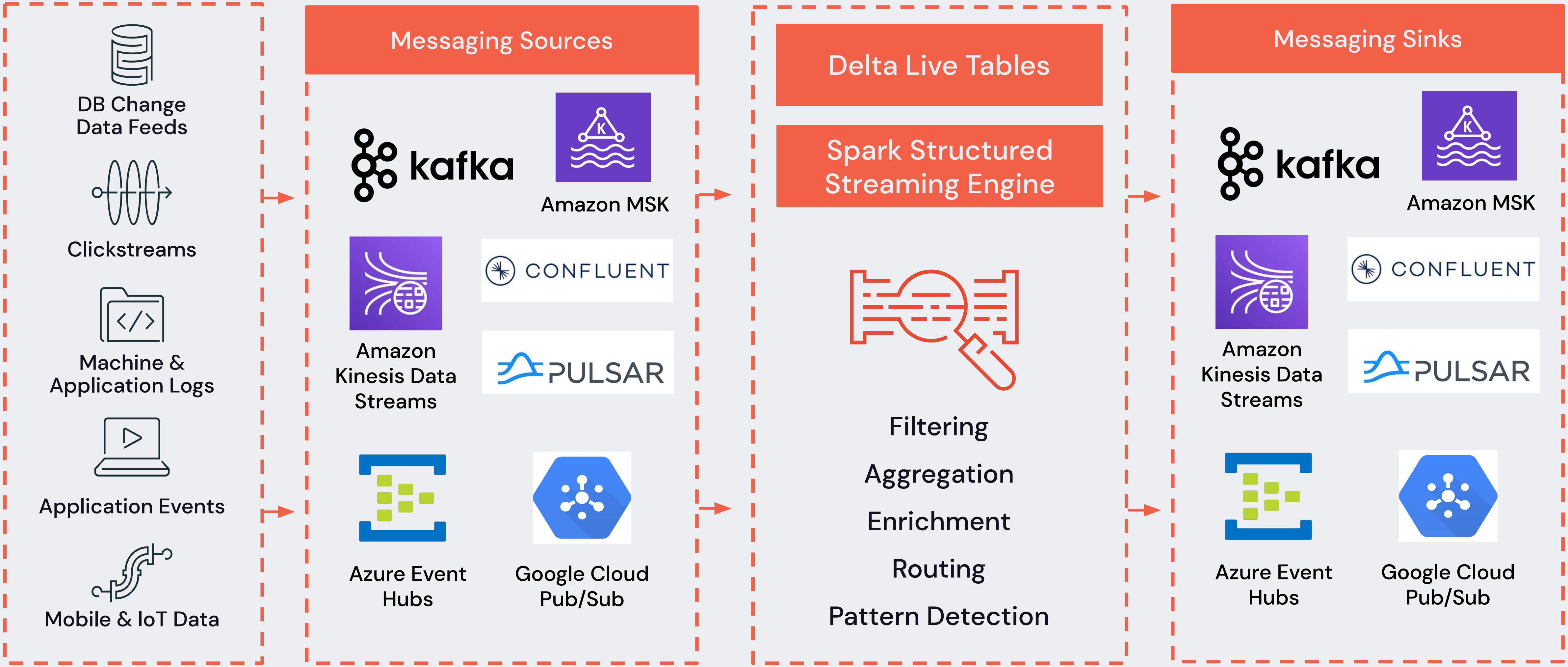
Delta Live Tables (DLT)

- Fully managed ETL service (batch + streaming) by Databricks
- The preferred way of doing streaming ETL in the Delta Lake
- Focus on ease of use
- Top choice for any new streaming ETL workloads

Structured Streaming

- The same Spark Structured Streaming API you have been using
- Roll your own ETL pipelines with Structured Streaming + Delta Connector + Workflow/Jobs
- Focus on flexibility
- Top choice for migrating existing Structured Streaming workloads

Event Processing



DB Change Data Feeds

Clickstreams

Machine & Application Logs

Application Events

Mobile & IoT Data

Messaging Sources

kafka

Amazon MSK

Amazon Kinesis Data Streams

CONFLUENT

PULSAR

Azure Event Hubs

Google Cloud Pub/Sub

Delta Live Tables

Spark Structured Streaming Engine

Filtering

Aggregation

Enrichment

Routing

Pattern Detection

Messaging Sinks

kafka

Amazon MSK

Amazon Kinesis Data Streams

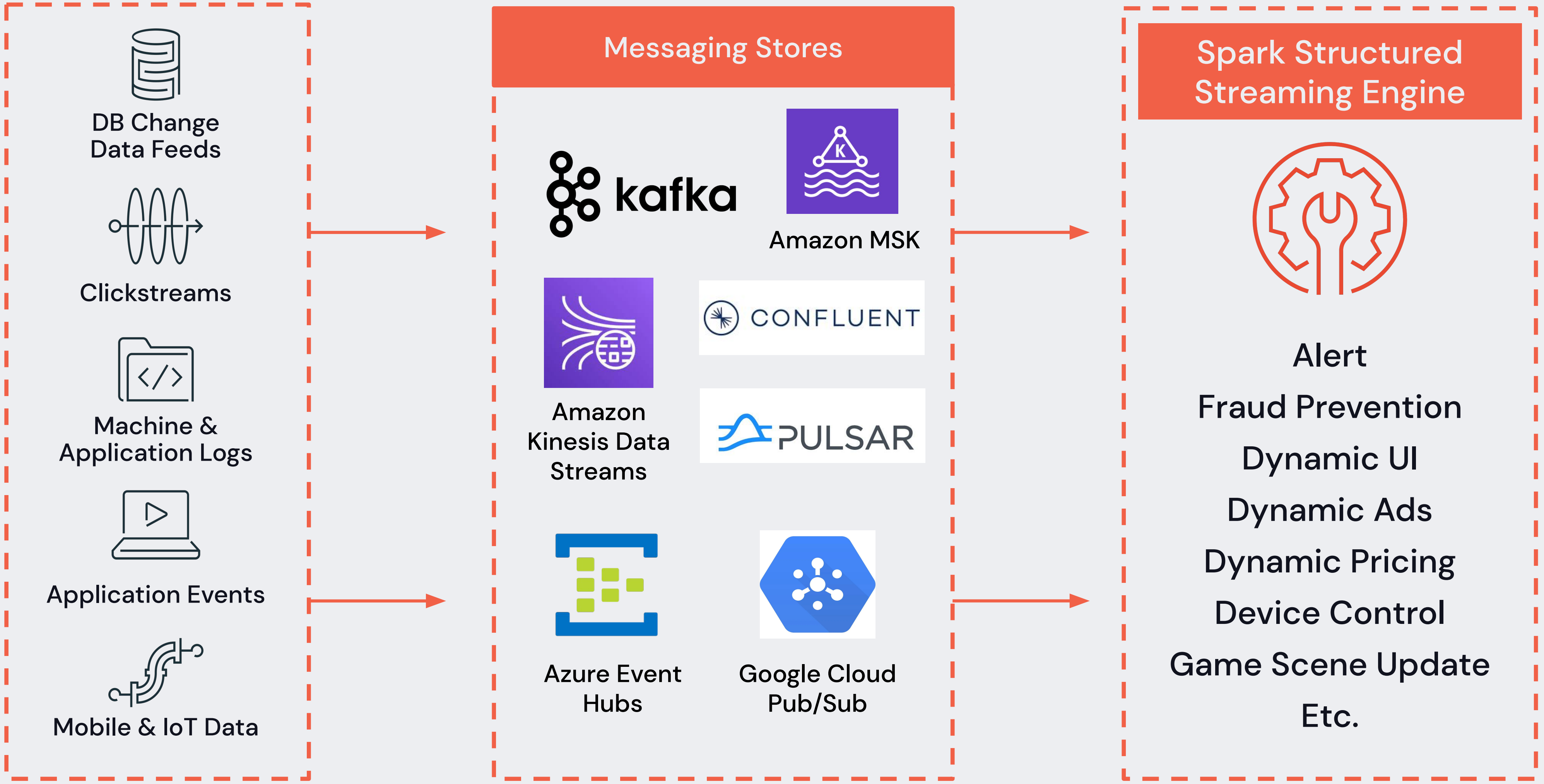
CONFLUENT

PULSAR

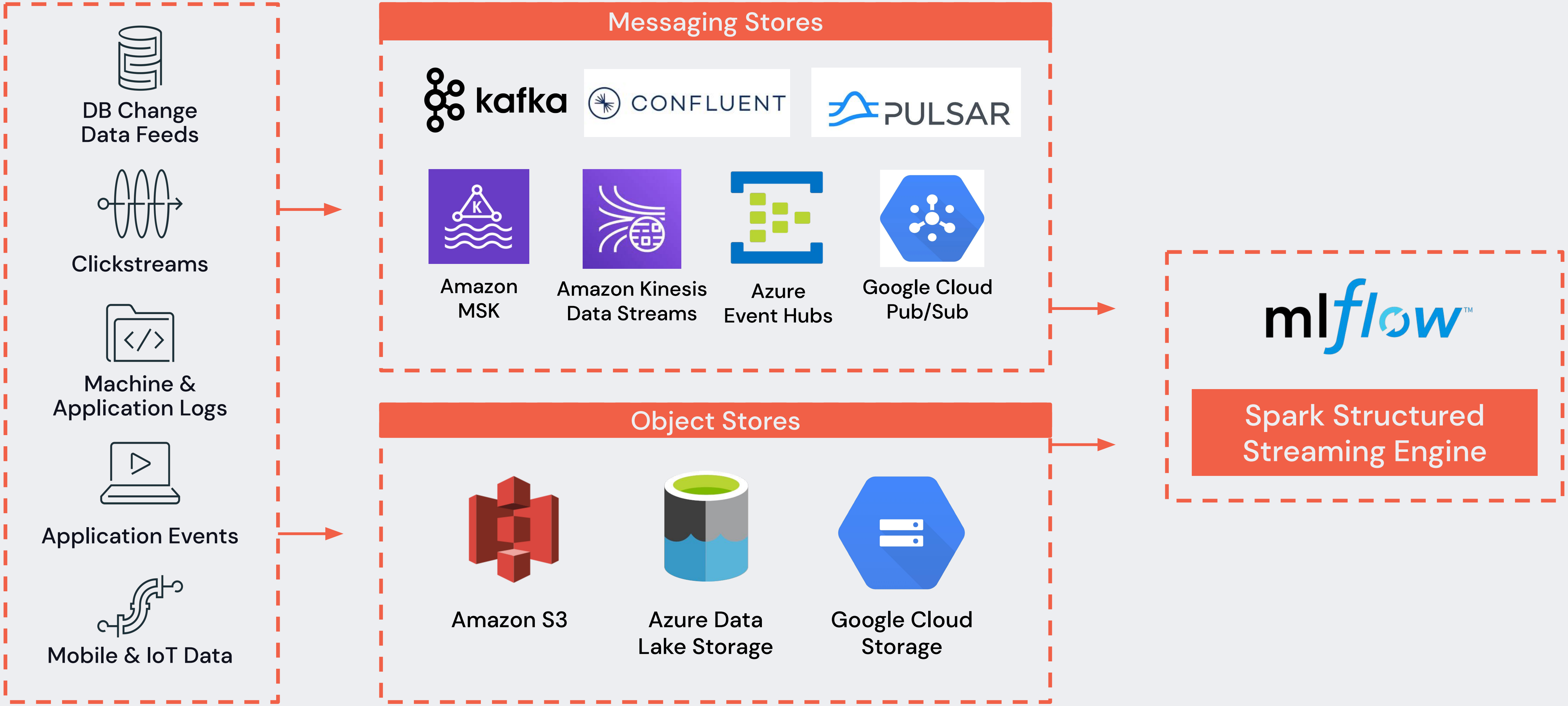
Azure Event Hubs

Google Cloud Pub/Sub

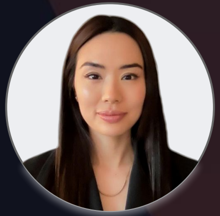
Event Driven Application



ML Inference



Upwork ML: Streaming Pipelines for Search & Discovery



Erica Lee

VP, ML at Upwork

Oleksii Diagiliev + Le Gu (contributors)

Engineering Manager, ML Infra + Director, ML S&D

DATA+AI SUMMIT 2022

ML Applications at Upwork

upwork™ is the world's **work marketplace** to solve complex work serving 30% of Fortune 100 and Enterprise customers

Search

Discovery

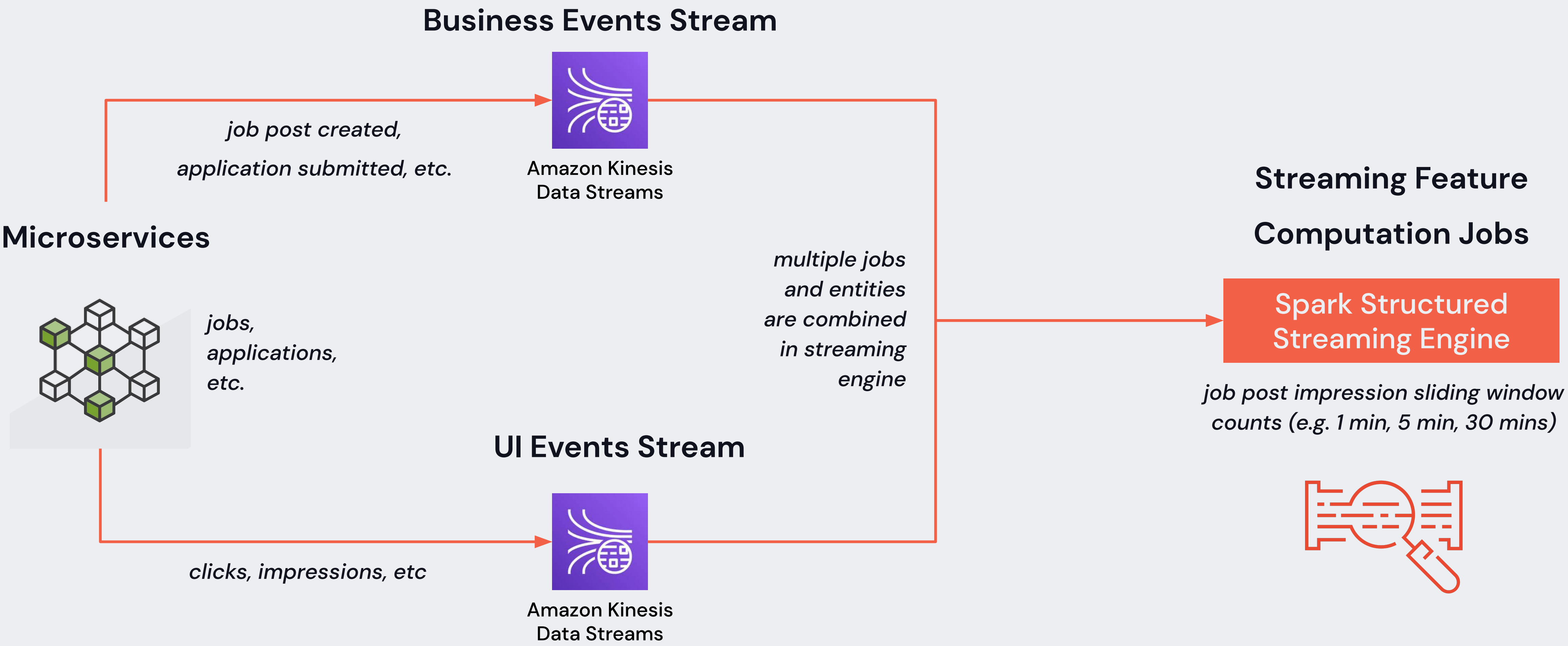
Trust & Safety

Marketing

Growth

We use ML to automate and scale capabilities to our users.

Streaming Feature Computation



Business Events Stream

job post created,

application submitted, etc.

Amazon Kinesis
Data Streams

Microservices

jobs,
applications,
etc.

UI Events Stream

clicks, impressions, etc

Amazon Kinesis
Data Streams

multiple jobs
and entities
are combined
in streaming
engine

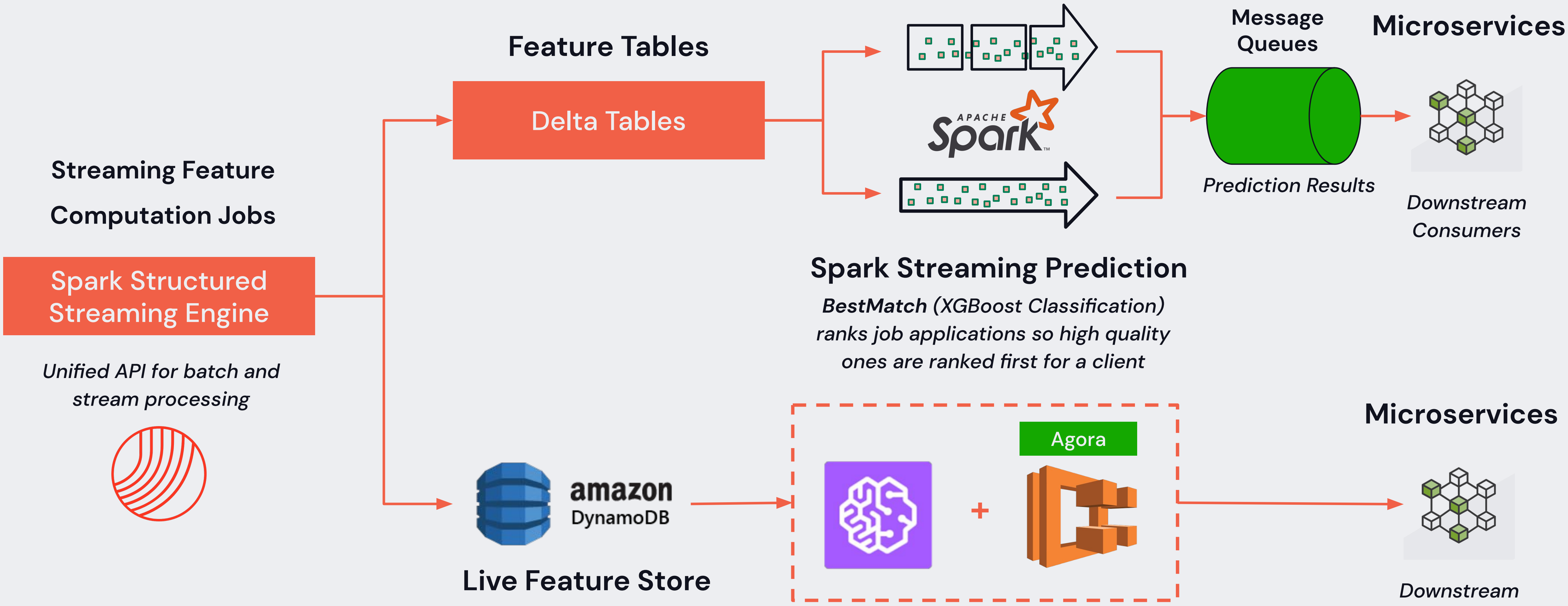
Streaming Feature

Computation Jobs

Spark Structured
Streaming Engine

job post impression sliding window
counts (e.g. 1 min, 5 min, 30 mins)

Model Serving



Spark Batch Prediction

Client Hiring State *(XGBoost Classification) predicts whether client is actively hiring or not*

Prediction Results

Downstream Consumers

Unified API for batch and stream processing

Spark Streaming Prediction

BestMatch *(XGBoost Classification) ranks job applications so high quality ones are ranked first for a client*

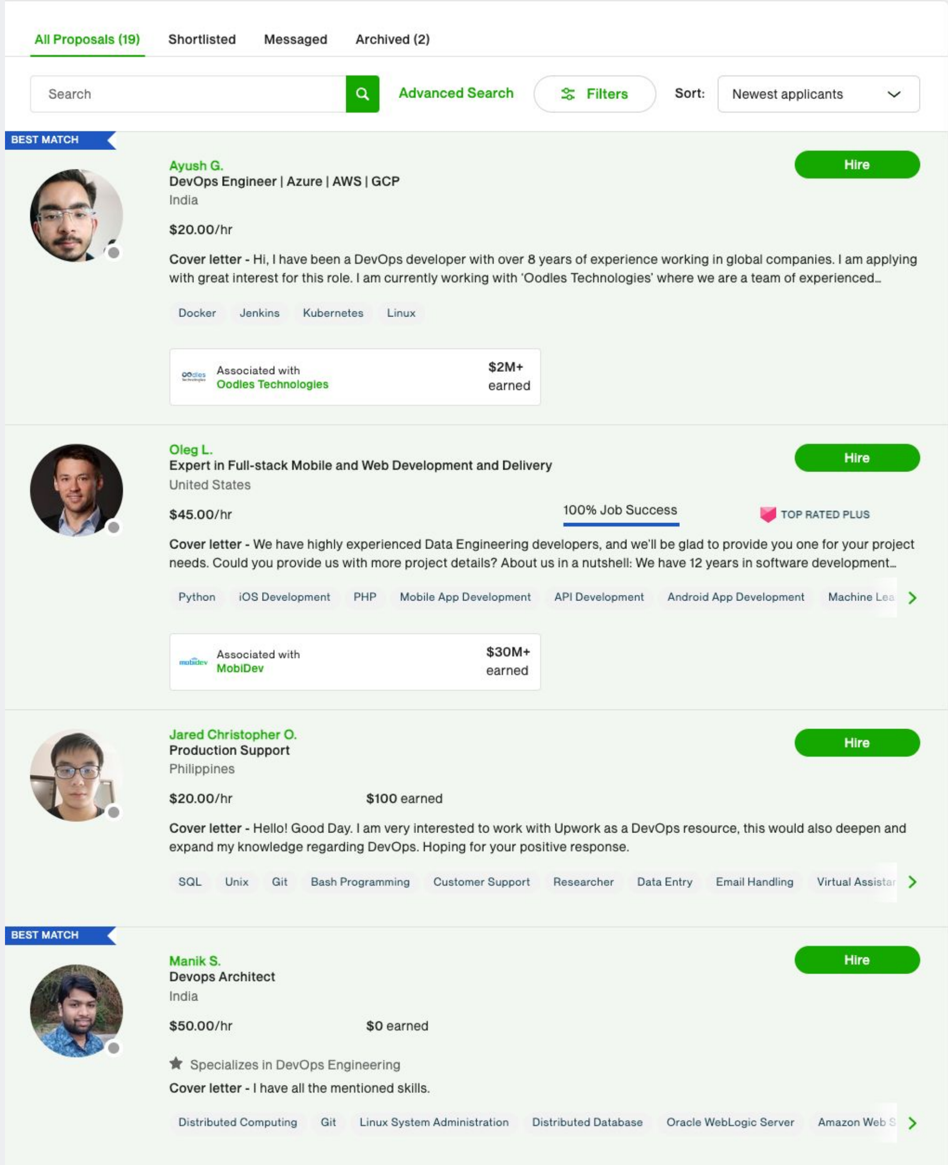
Live Feature Store

Downstream Consumers

AWS SageMaker + ECS

Job Value Prediction *(XGBoost Regression) predicts the job post value in dollars*

BestMatch Ranking Algorithm



BEST MATCH

Ayush G.
DevOps Engineer | Azure | AWS | GCP
India

$20.00/hr

Cover letter - Hi, I have been a DevOps developer with over 8 years of experience working in global companies. I am applying with great interest for this role. I am currently working with 'Oodles Technologies' where we are a team of experienced...

Docker · Jenkins · Kubernetes · Linux

Associated with Oodles Technologies — $2M+ earned

Oleg L.
Expert in Full-stack Mobile and Web Development and Delivery
United States

$45.00/hr — 100% Job Success — TOP RATED PLUS

Cover letter - We have highly experienced Data Engineering developers, and we'll be glad to provide you one for your project needs. Could you provide us with more project details? About us in a nutshell: We have 12 years in software development...

Python · iOS Development · PHP · Mobile App Development · API Development · Android App Development · Machine Lea

Associated with MobiDev — $30M+ earned

Jared Christopher O.
Production Support
Philippines

$20.00/hr — $100 earned

Cover letter - Hello! Good Day. I am very interested to work with Upwork as a DevOps resource, this would also deepen and expand my knowledge regarding DevOps. Hoping for your positive response.

SQL · Unix · Git · Bash Programming · Customer Support · Researcher · Data Entry · Email Handling · Virtual Assistar

BEST MATCH

Manik S.
Devops Architect
India

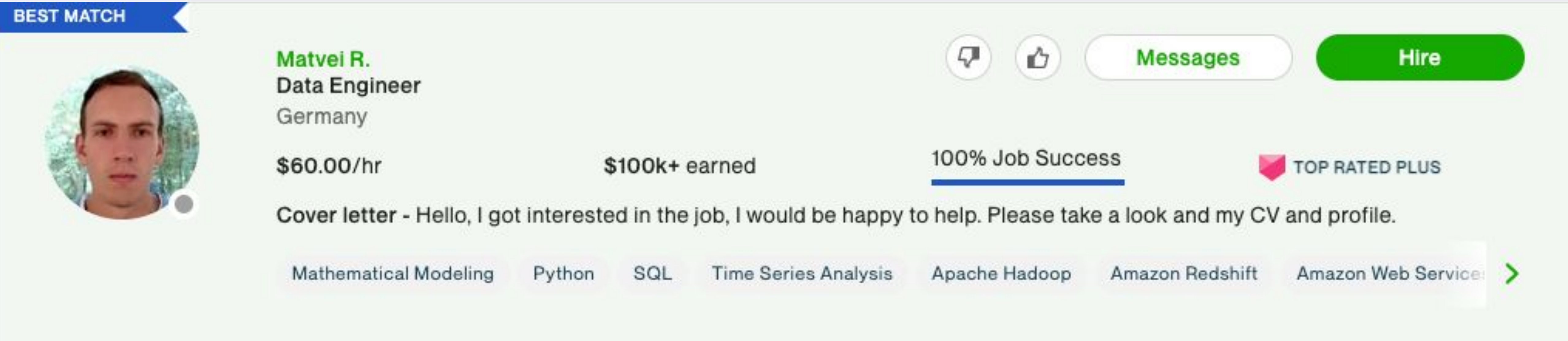
$50.00/hr — $0 earned

★ Specializes in DevOps Engineering

Cover letter - I have all the mentioned skills.

Distributed Computing · Git · Linux System Administration · Distributed Database · Oracle WebLogic Server · Amazon Web S

BestMatch Ranking Algorithm



BEST MATCH

Matvei R.
Data Engineer
Germany

$60.00/hr

$100k+ earned

100% Job Success

TOP RATED PLUS

Cover letter - Hello, I got interested in the job, I would be happy to help. Please take a look and my CV and profile.

Mathematical Modeling | Python | SQL | Time Series Analysis | Apache Hadoop | Amazon Redshift | Amazon Web Service

50%+ of the jobs posted on our platform receive
25+ proposal bids within 24 hours of posting.

Benefits of Databricks Platform

Benefits

- Unified platform that empowers our ML & Data and Engineers with 1 environment to run 2 workloads via Delta Tables (batch & streaming/real-time)
- Interactive and collaborative notebooks reduce dev. time (10%+)

Come Join Upwork ML!

Management

- 1 ML Manager – *Search & Discovery*
- 1 ML Manager – *Trust & Safety*
- 1 ML Manager – *Infrastructure*

Individual Contributors

- 1 ML Ops Engineer – *Search & Discovery*
- 1 Senior ML Engineer – *Trust & Safety*

Contact **Aaron White** (aaronwhite@upwork.com) about openings

DATA+AI
SUMMIT 2022

Thank you

DATA+AI
SUMMIT 2022