
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

PRT321

SPONSORED BY DATABRICKS

How Corning built end-to-end ML on a data lakehouse platform with Databricks

Jessica Cornell

Enterprise Account Executive
Databricks

Prem Prakash

Senior Director, Product Marketing
Databricks

Denis Kamotsky

Principal Software Engineer,
Corning

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Databricks

The Lakehouse Company

Inventor and pioneer of the
data lakehouse

Gartner recognized leader in both
- Database Management Systems
- Data Science and Machine Learning Platforms

Creator of highly successful OSS data projects: Delta Lake, Apache Spark, and MLflow

Raised over $3B in investment

4500+ employees across the globe

Global adoption

Over 7000 customers, from F500 to unicorns









Data, analytics, and AI enabled tech's leaders to disrupt industries

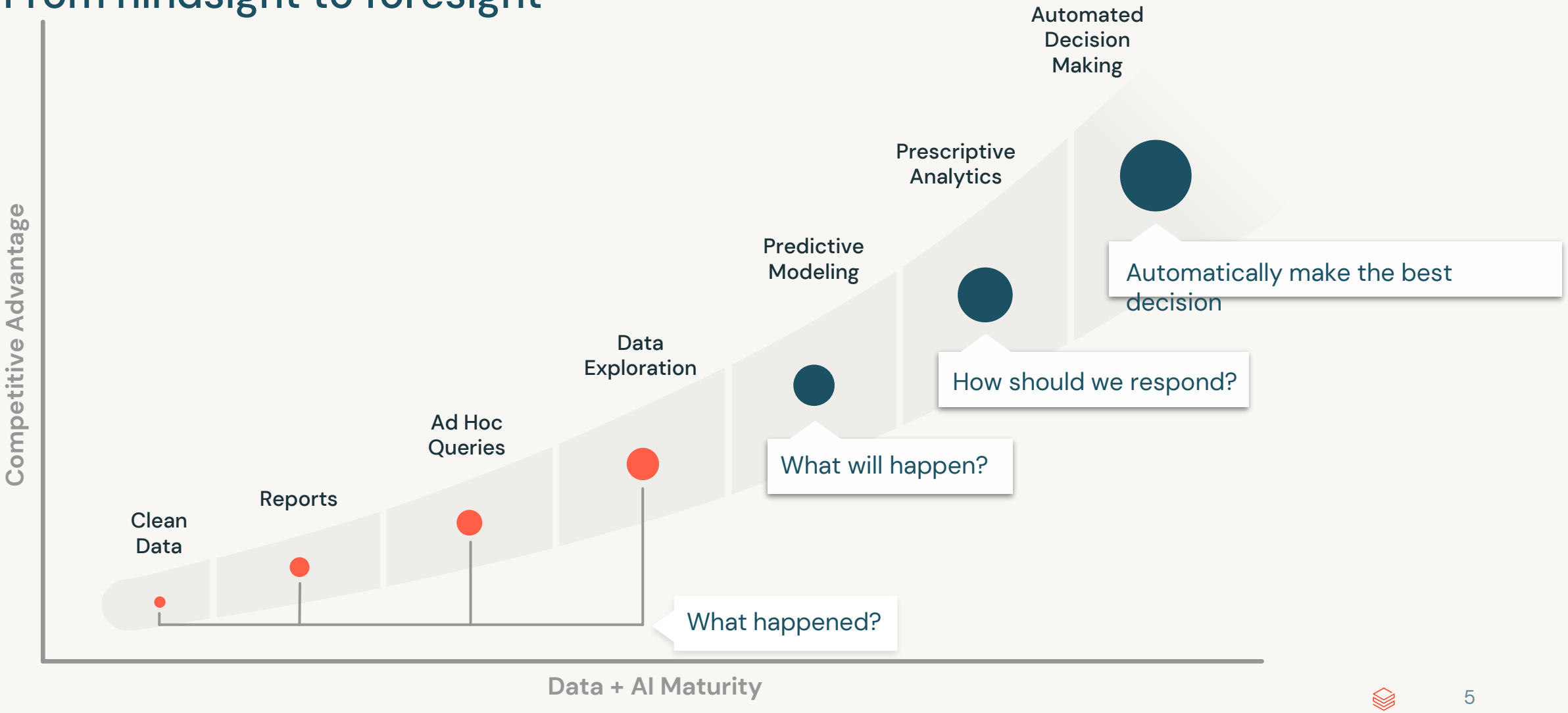






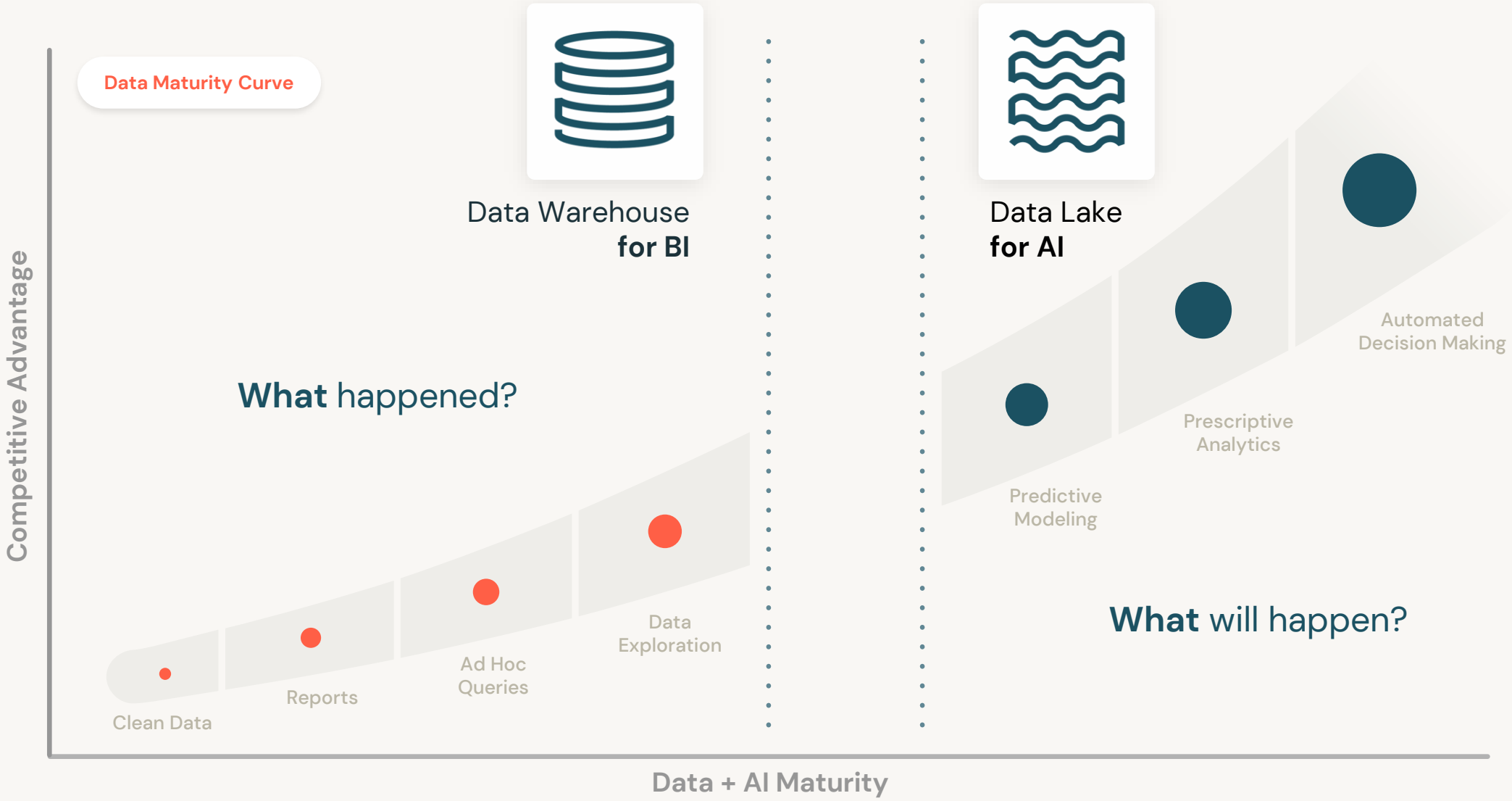
Tech leaders are to the right of the Data Maturity Curve

From hindsight to foresight

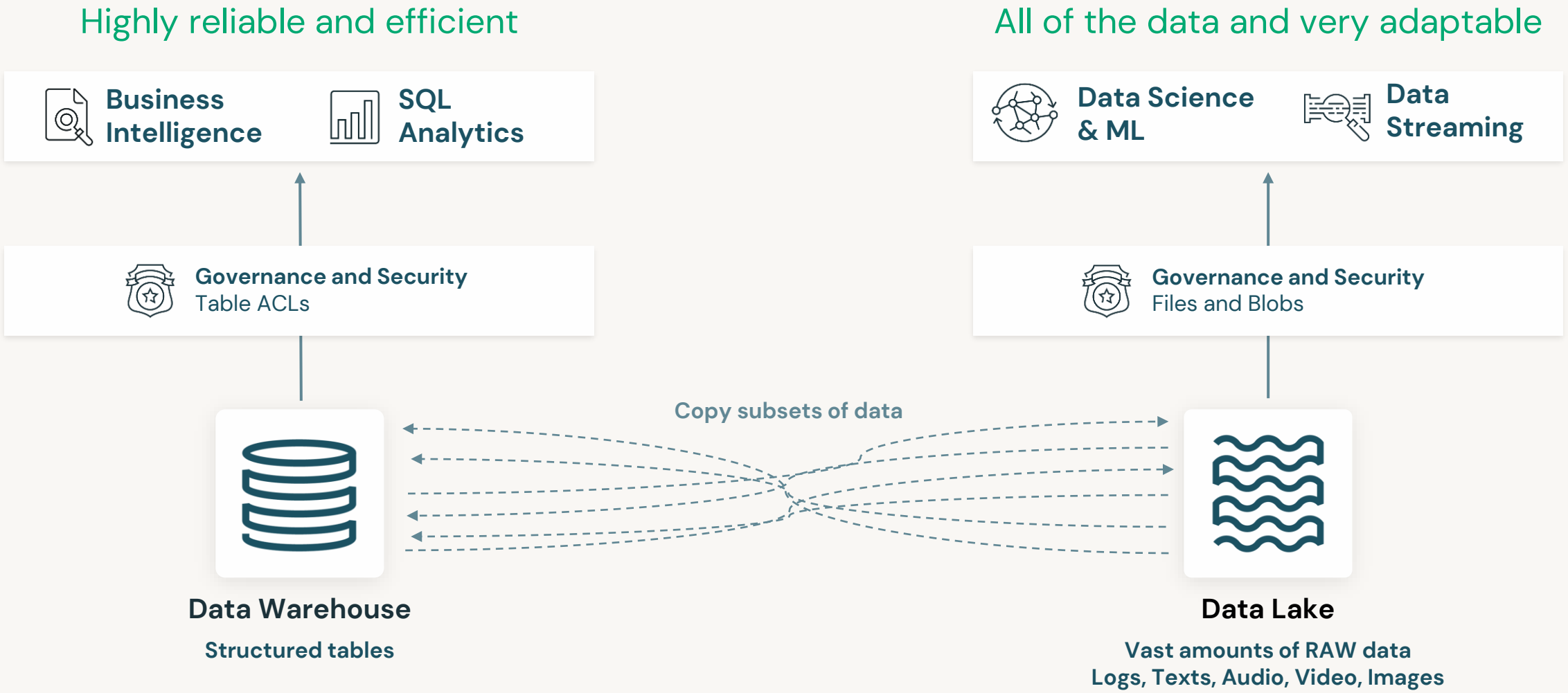


Most enterprises still struggle with data, analytics, and AI

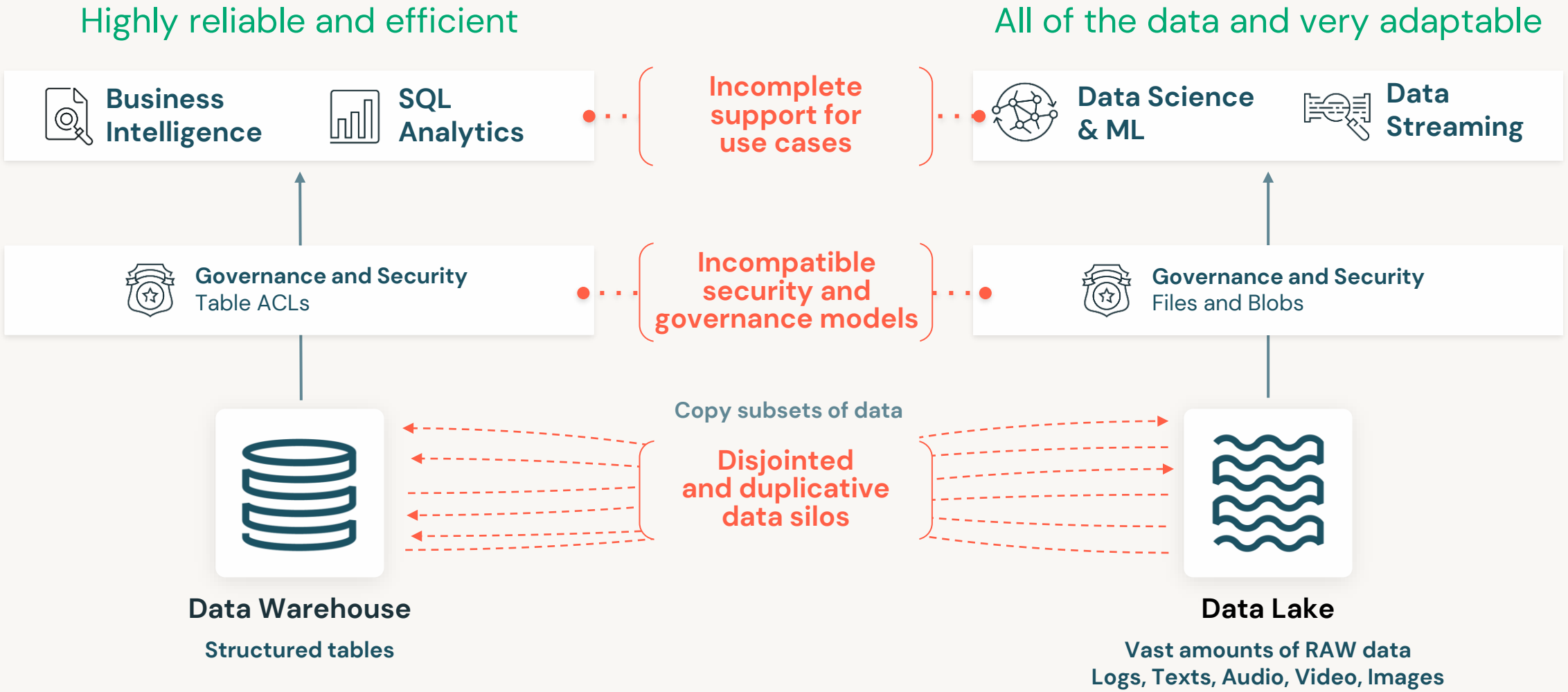
Realizing this requires two disparate, incompatible data platforms



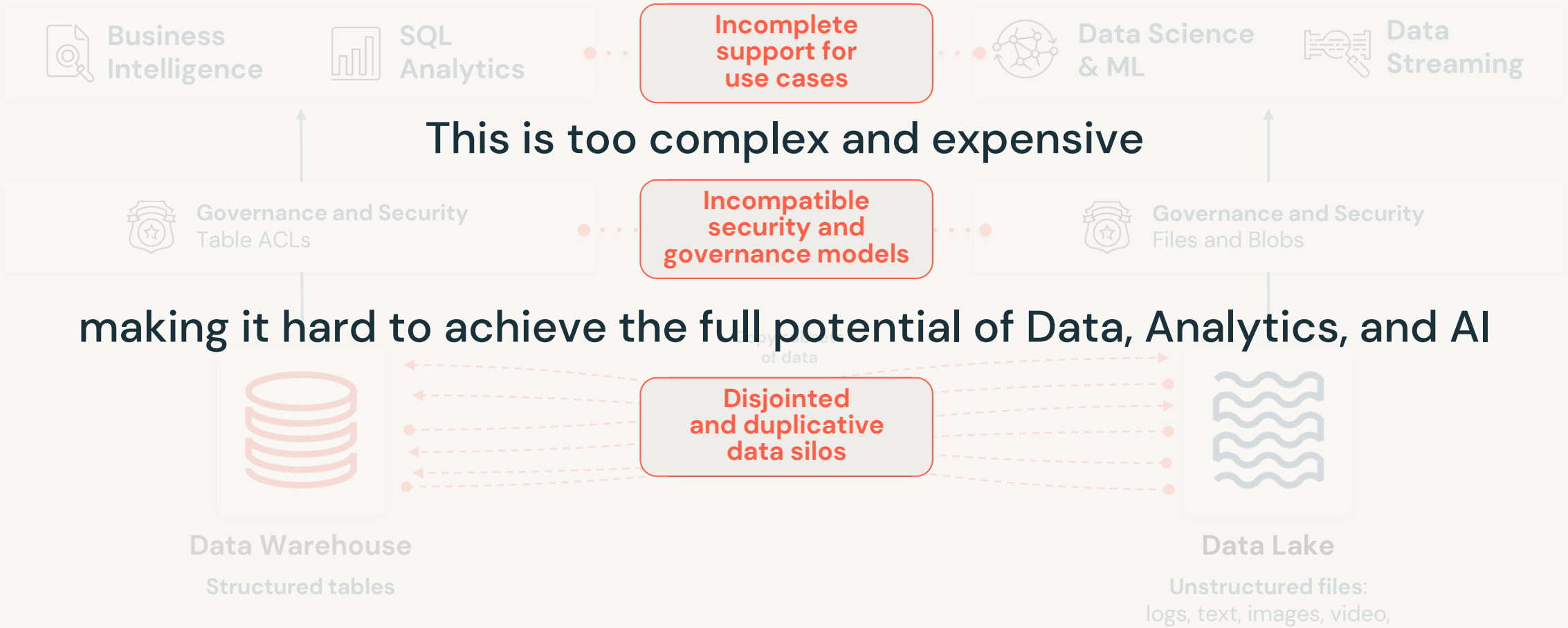
Realizing this required two disparate, incompatible data platforms



Realizing this required two disparate, incompatible data platforms



Realizing this requires two disparate, incompatible data platforms



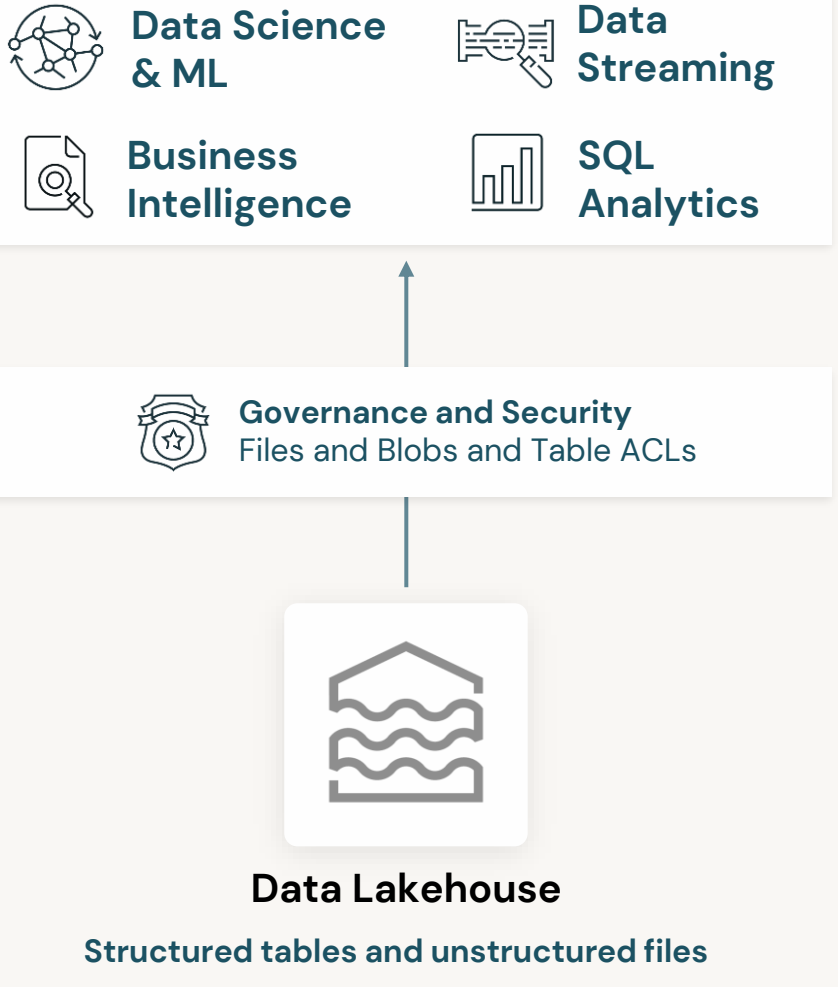
©2022 Databricks Inc. — All rights reserved

This is the lakehouse paradigm

All ML, SQL, BI,
and Streaming use cases

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

A **reliable**
data platform
to **efficiently** handle all data types



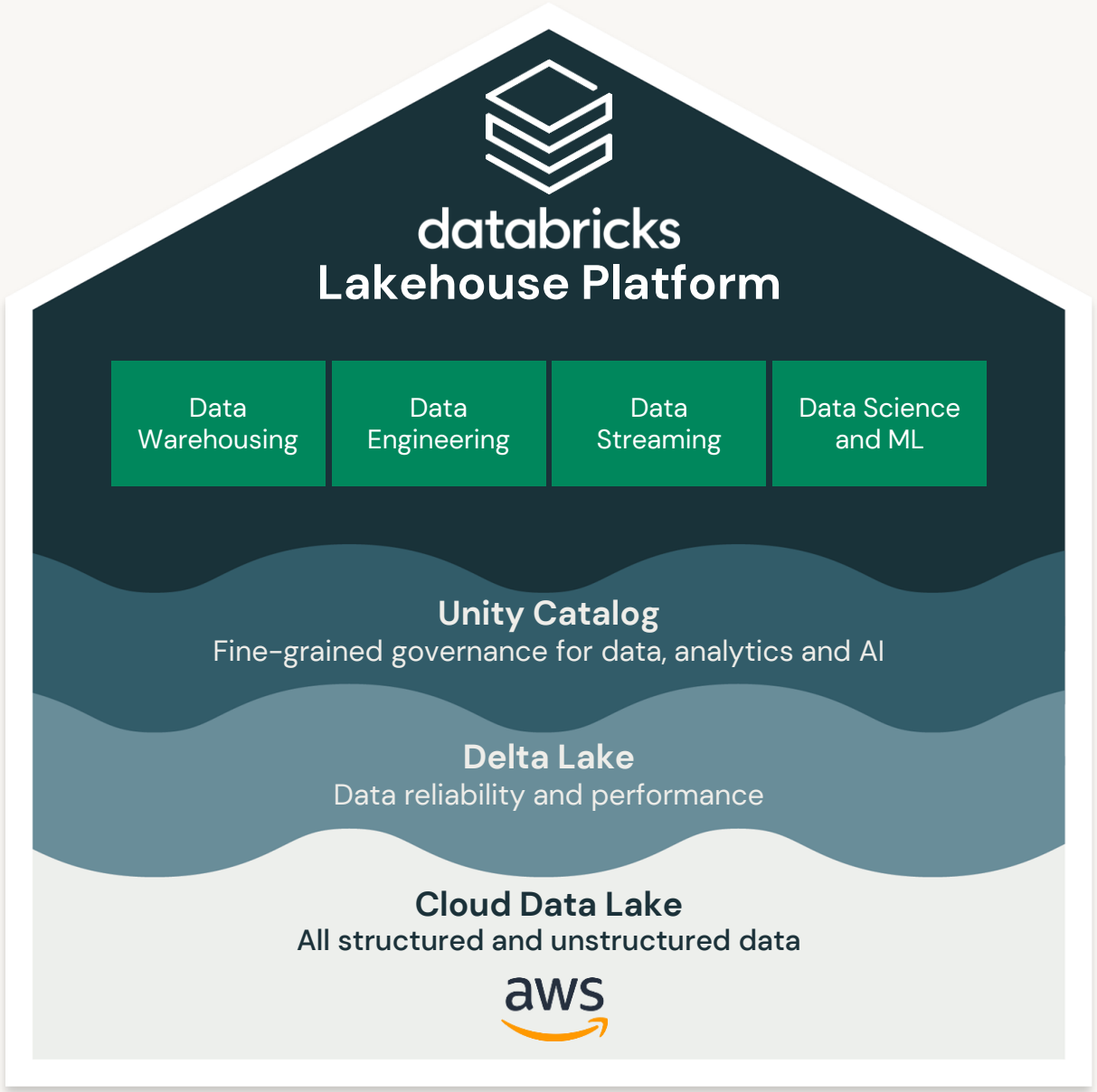
Technologies

Data Applications

Unity Catalog
Fine-grained governance
for data and AI

Delta Lake
Data reliability and performance

©2022 Databricks Inc. — All rights reserved



Databricks Lakehouse Platform

Simple

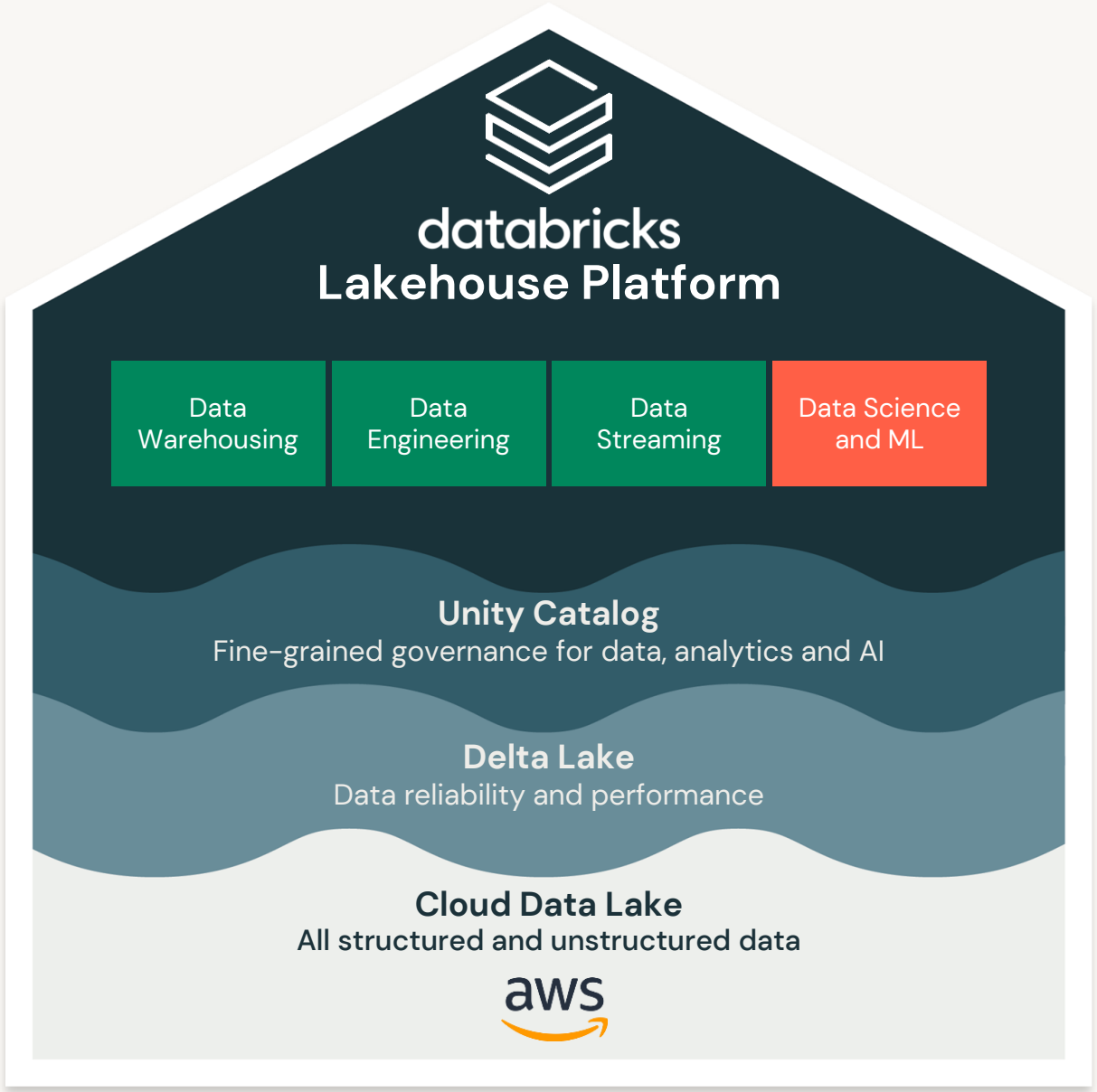
Unify your data warehousing and AI use cases on a single platform

Multicloud

One consistent data platform across clouds

Open

Built on open source and open standards



Databricks Lakehouse Platform

Simple

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

Multicloud

One consistent data platform across clouds

Open

Built on open source and open standards



Machine Learning on the Databricks Lakehouse Platform

ML is going prime time but challenges remain

90%

Enterprise applications will be AI-augmented by 2025 —IDC

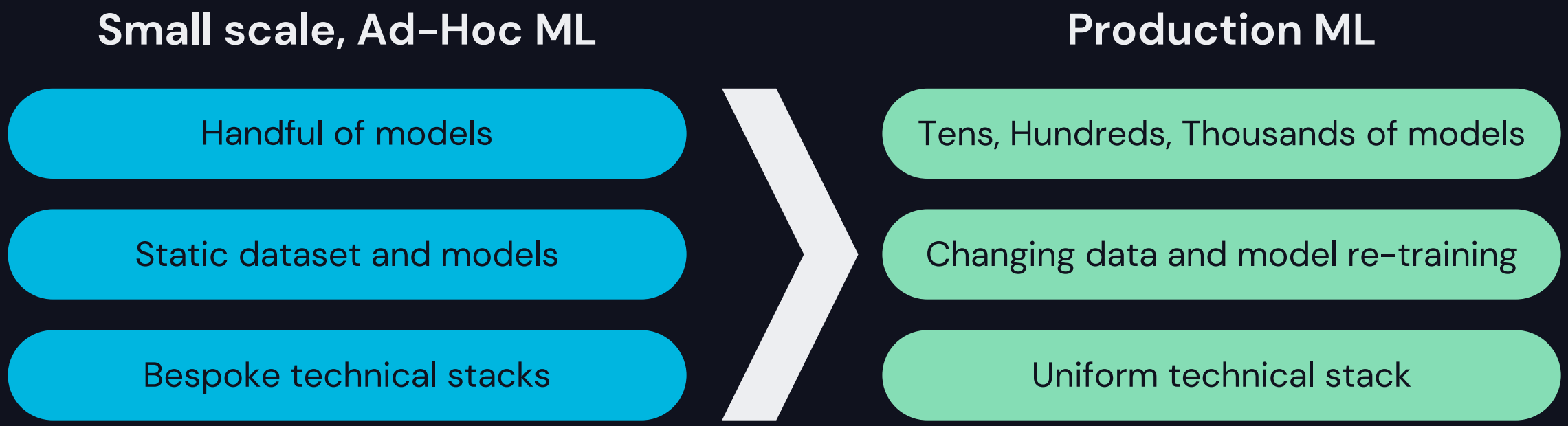
$10T+

Projected business value creation by AI in 2030 —PWC

But—huge challenges getting ML to work at scale!

Most ML projects fail before getting to production

Where ML gets stuck

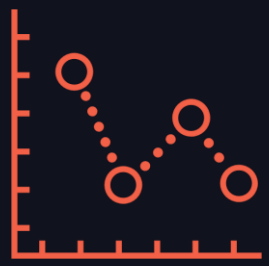


Challenges with doing ML at scale

Siloed development stack across DataOps, ModelOps and DevOps



Data Engineering Stack



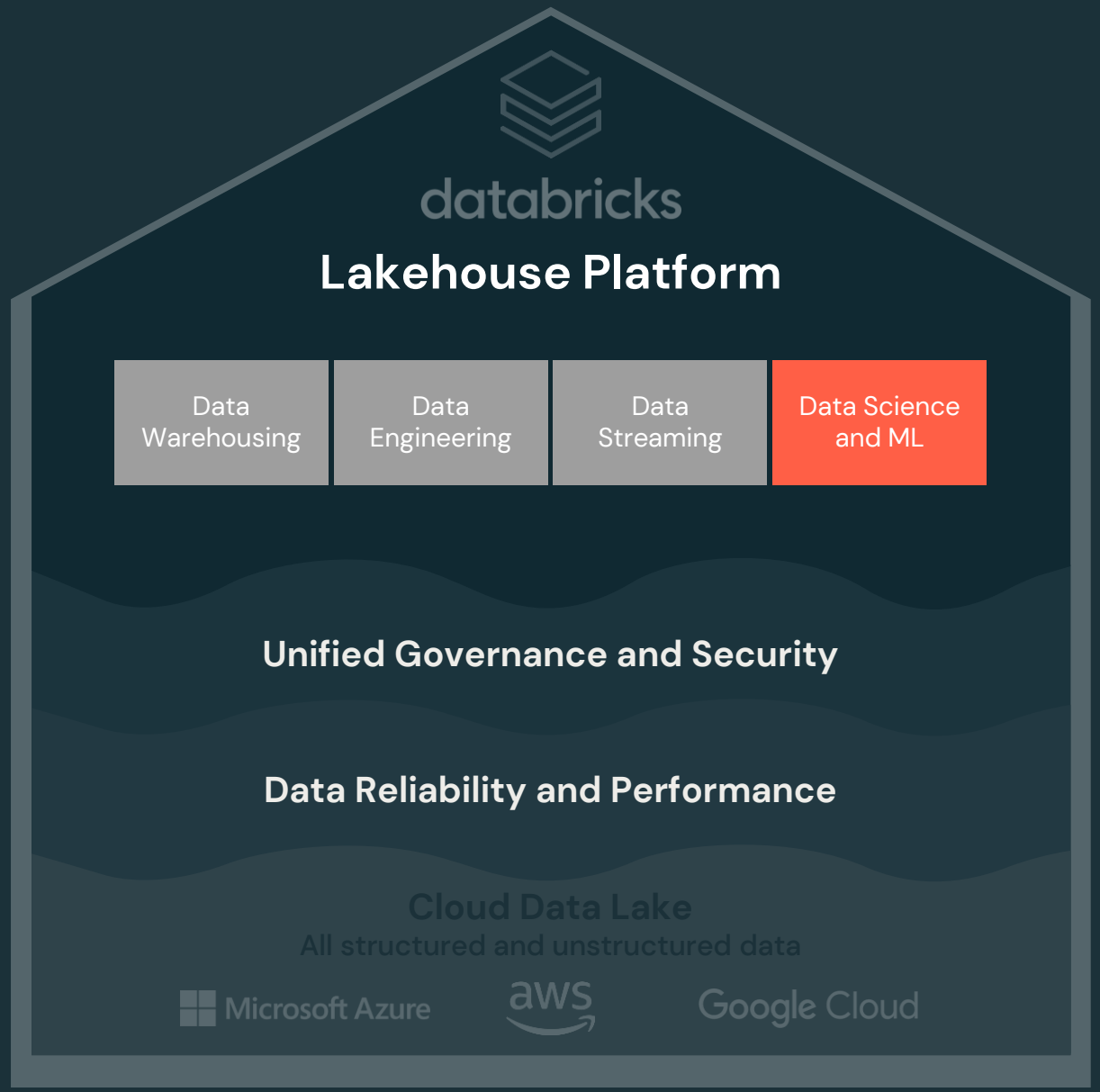
Analytical ML Stack



Online Service Stack

Difficulties with reproducing model, code and data | Manual monitoring | Challenges with governance

End to end Machine Learning on the Lakehouse

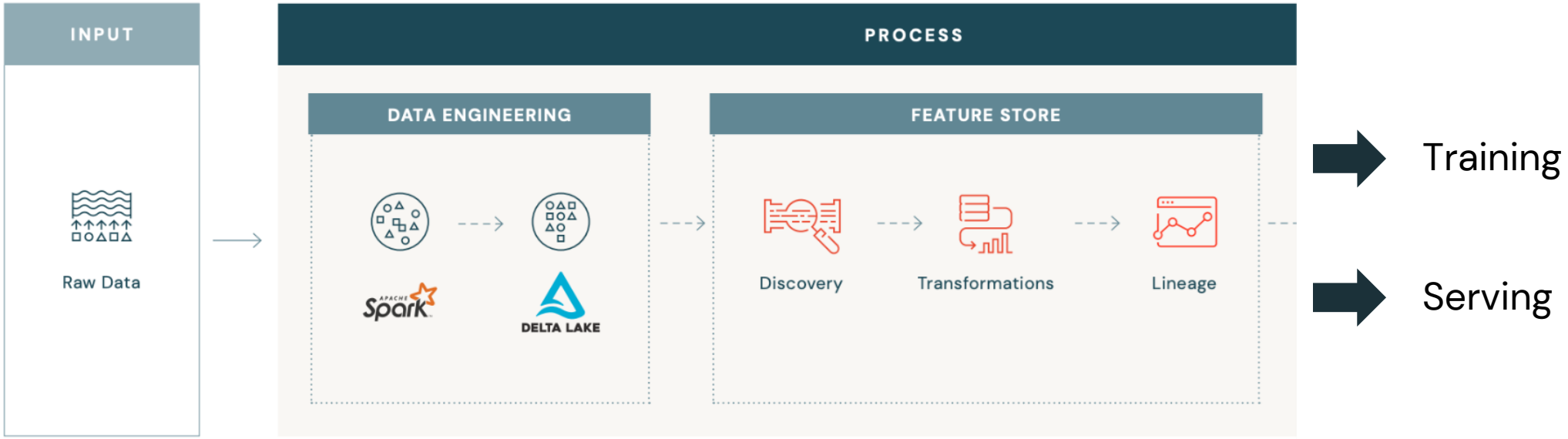


databricks
Machine Learning



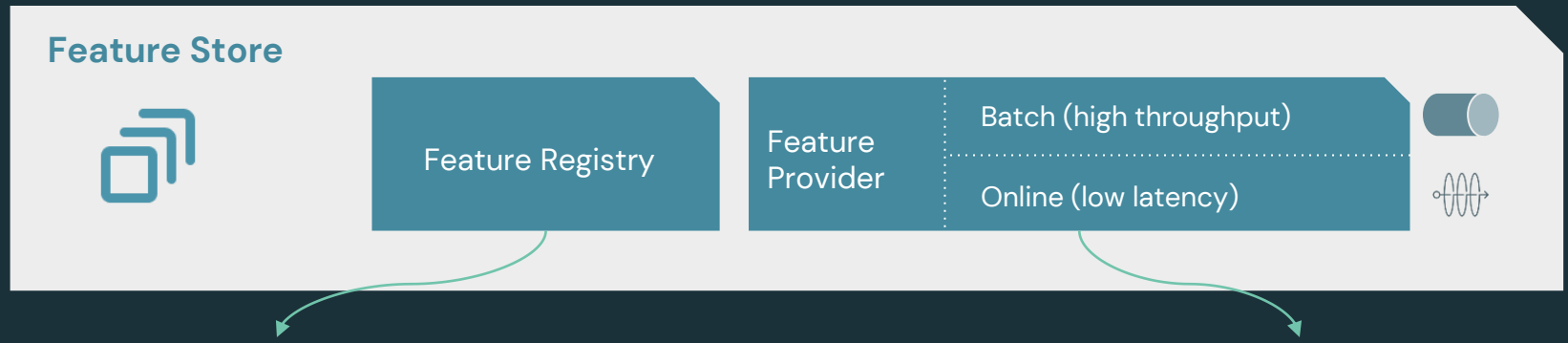
Data Prep and DataOps

Built-in data prep



Databricks Feature Store

The first feature store codesigned for a data and MLOps platform



The Feature Registry allows reuse

- Discoverability and reusability
- Versioning
- Upstream and downstream lineage

The Feature Provider prevents offline/online skew

- Batch and online access to features
- Feature lookup packaged with models
- Simplified deployment process

Co-designed with DELTA LAKE

- Open format
- Built-in data versioning and governance with Unity Catalog
- Native access through PySpark, SQL, etc.

Co-designed with mlflow

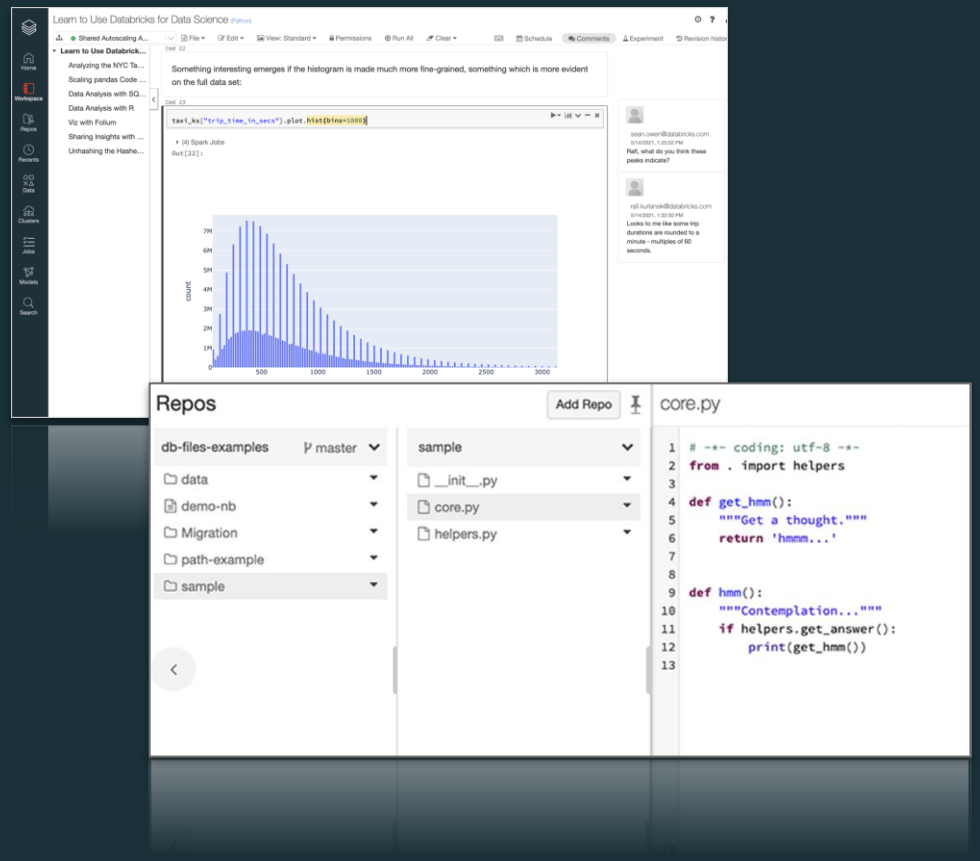
- Open model format that supports all ML frameworks
- Feature version and lookup logic hermetically logged with Model

Experimentation and ModelOps

Enabling you to build models your way

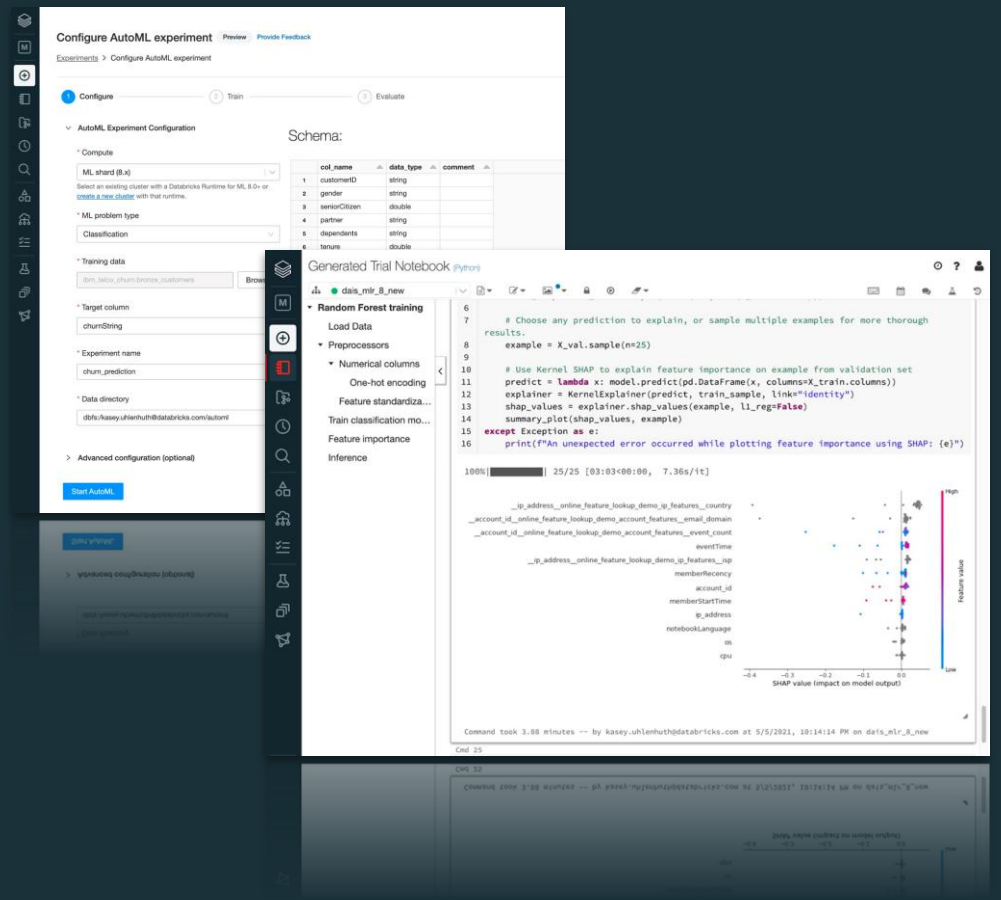
Multi-language Notebooks and APIs

Co-edit Notebooks in Python, R, Scala, and SQL



Low code ML with AutoML

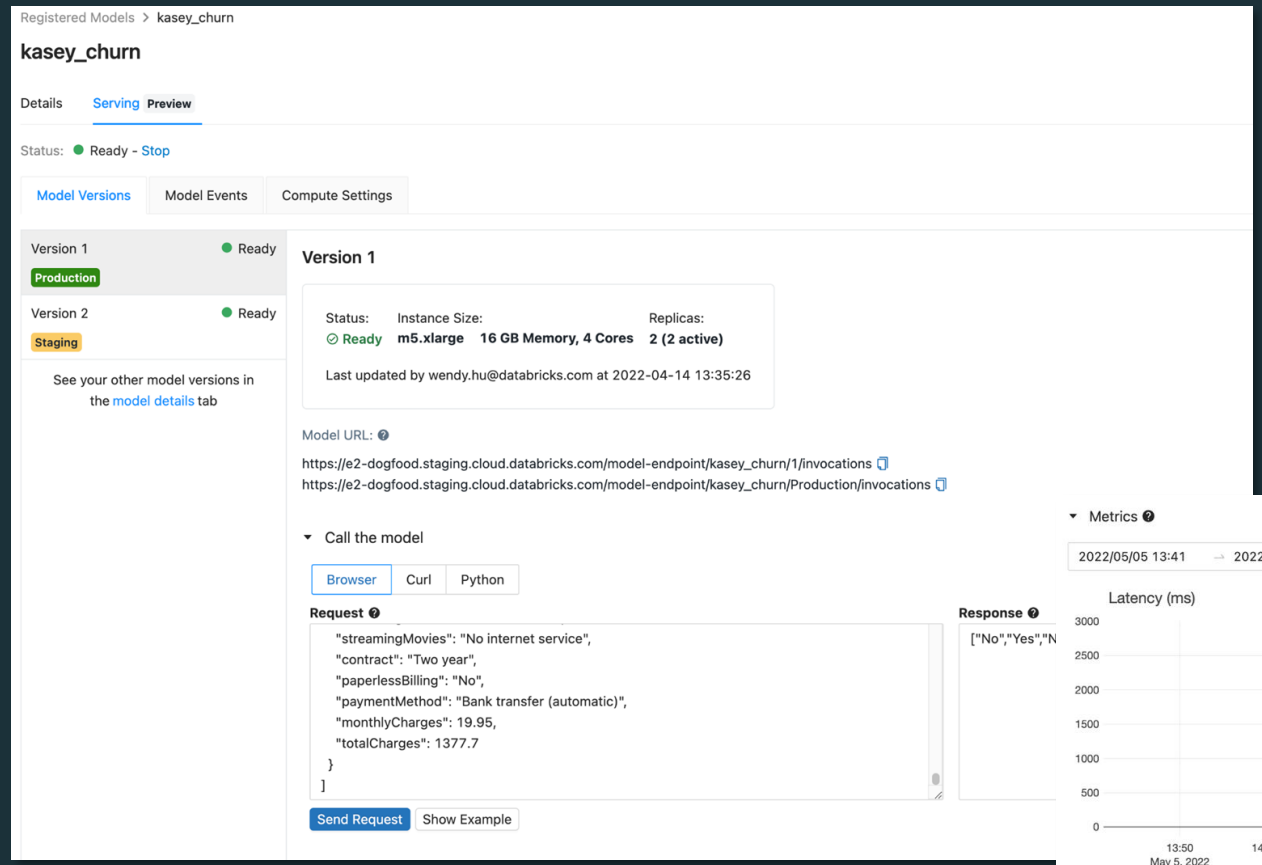
UI based ML development with a glass box approach



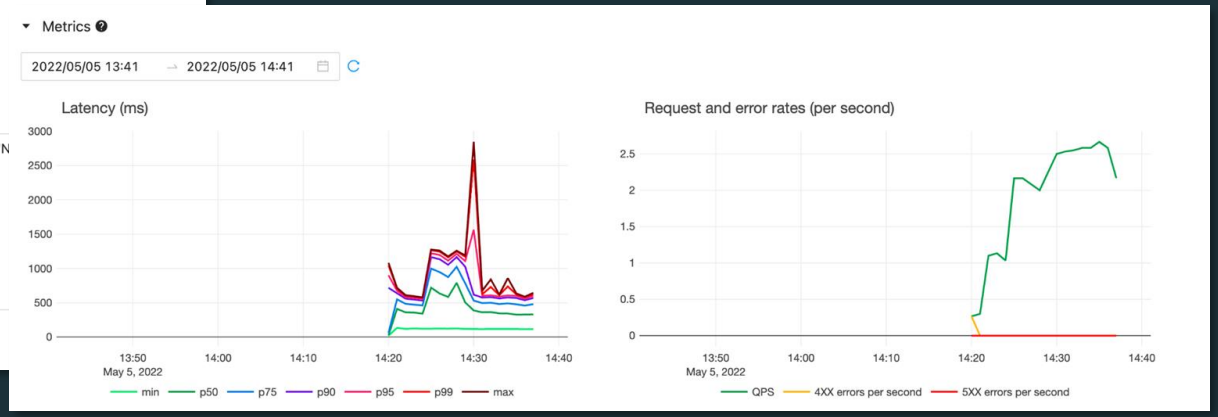
Deployment and DevOps

Serverless Real-Time Inference

Production-grade model serving hosted by Databricks



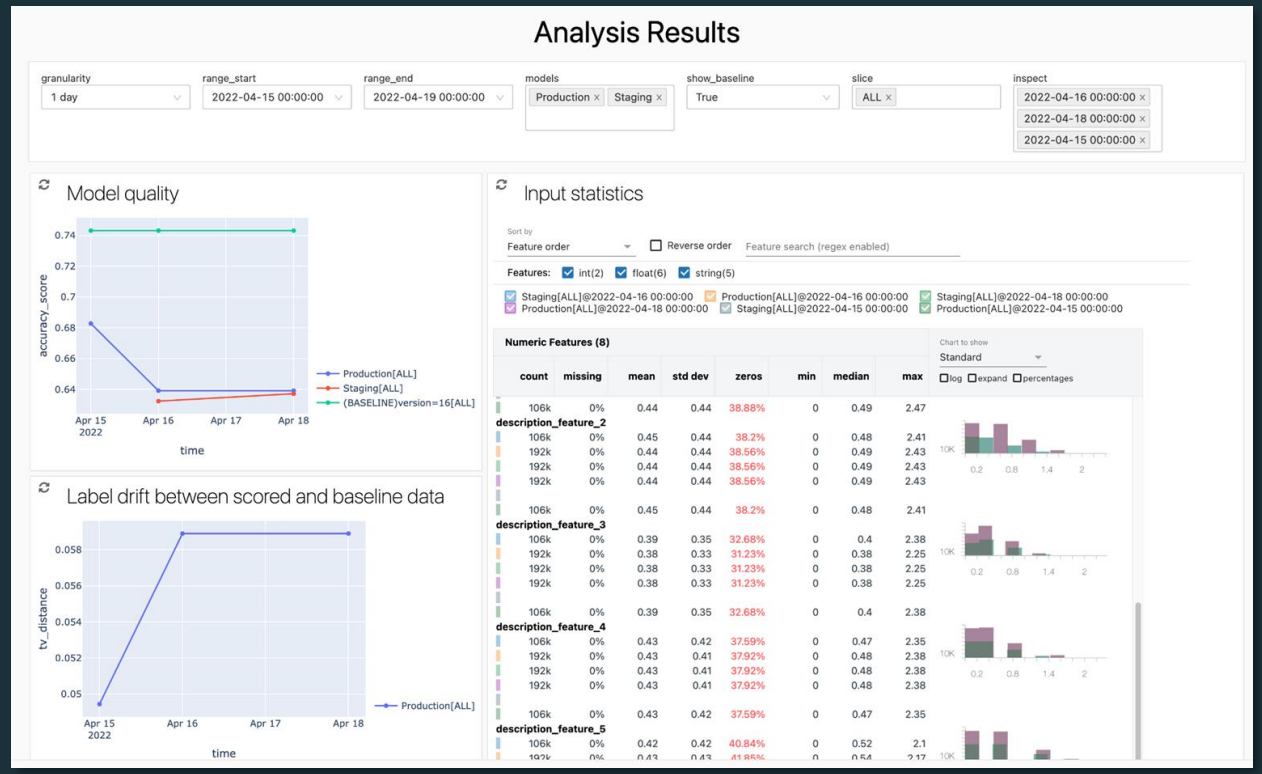
- Low latency REST endpoints
- High availability SLAs
- Autoscaling (including to 0)
- Built-in observability



Lakehouse Monitoring

Model monitoring

Monitor data distributions and model performance

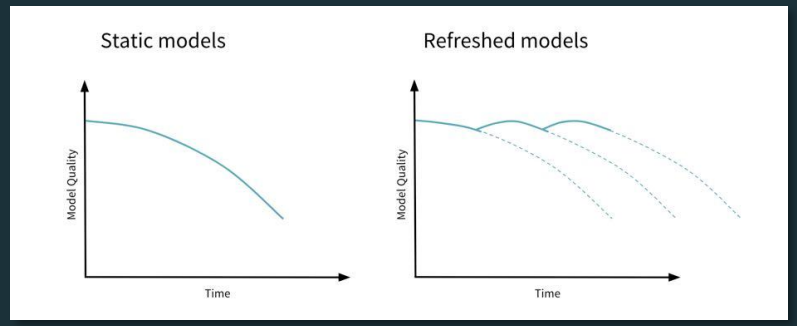


- Centralized and standardized mechanism for monitoring models in production
- Auto-generated dashboard with flexibility to add custom metrics
- Alerting to get notified on drift or quality issues

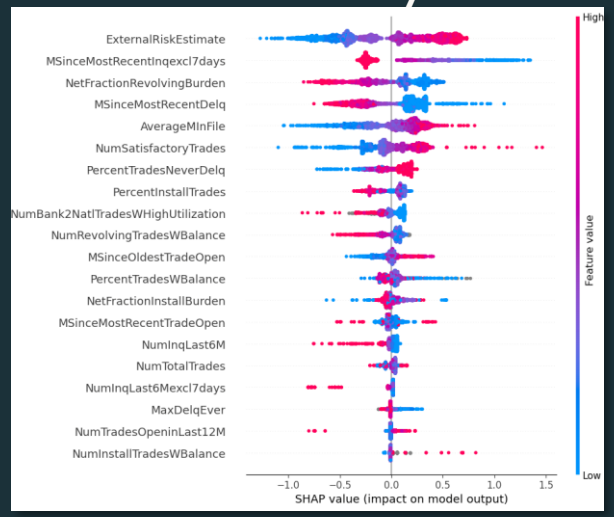
Model monitoring

Assess business impact and ROI from your ML projects

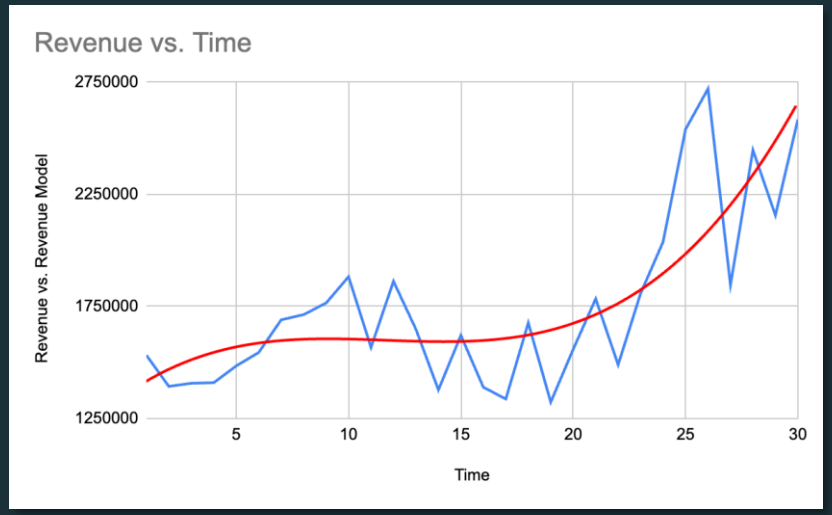
Skew/Drift and Model Refresh



Explainability and Root Cause Analysis



Join with Business Metrics



databricks
Machine Learning





aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Founded:
1851

Headquarters:
Corning, New York

Employees:
~50,000 worldwide

2020 Core Sales:
$11.5 billion (at rate of 107 ¥/$)

Fortune 500 Ranking (2020):
277

CORNING

© 2021 Corning Incorporated

Corning's mission is another 170 years of innovation and independence. But we never take our eyes off the big picture. Our ultimate goal is a world that is better because of our efforts: a world with cleaner air, healthier lives, richer entertainment experiences, and more efficient communication.

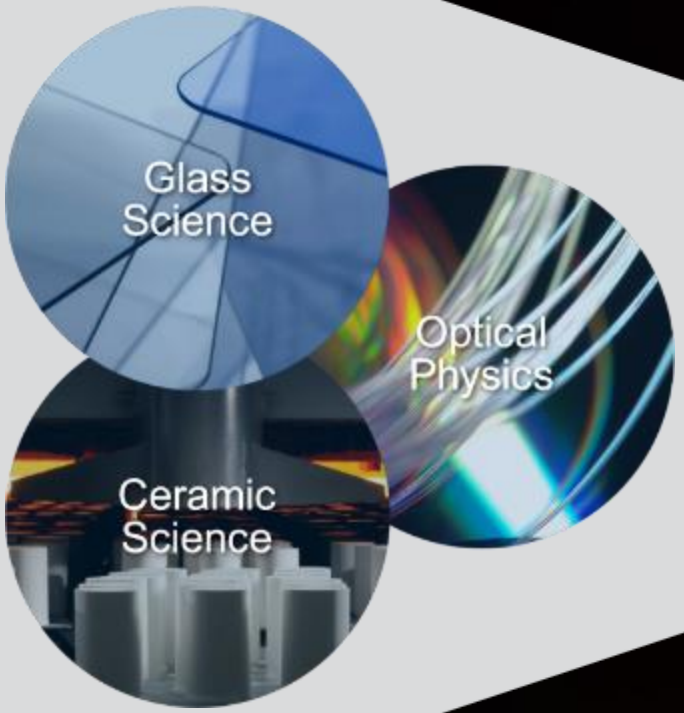
https://www.corning.com/media/worldwide/global/documents/corning_at_a_glance.pdf

aws

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Best-in-the-World Capabilities

3 Core Technologies



4 Manufacturing & Engineering Platforms



5 Market-Access Platforms



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

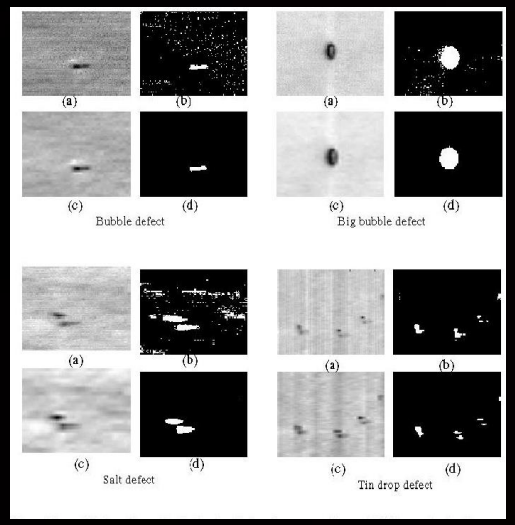
“Our aspiration is to apply the power of digital to create competitive advantage.”

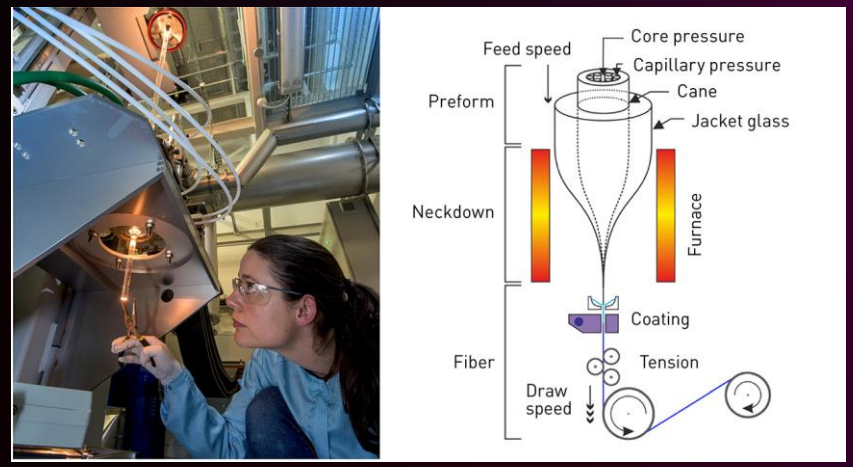
Glenn Bleiler

Emerging Technologies, Corning IT

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

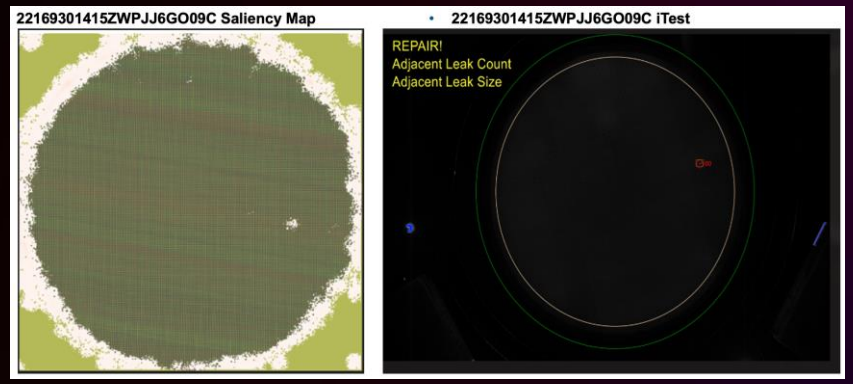
Computer Vision in Manufacturing





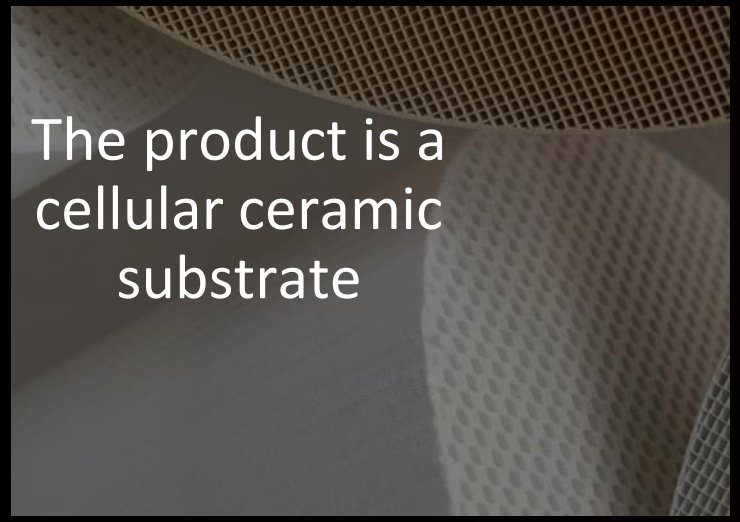
[2] Processing of Soot Deposition I G.E Jason et al., Opt. Express 27, 20567 (2019); doi: 10.1364/OE.27.020567

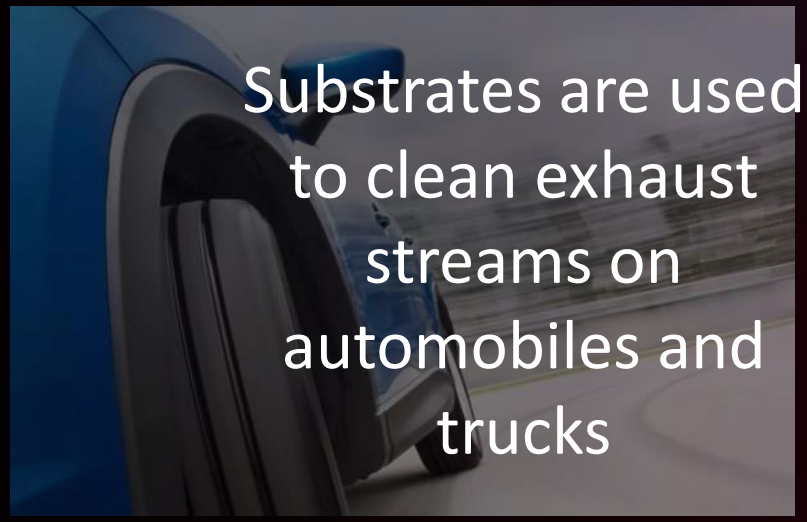


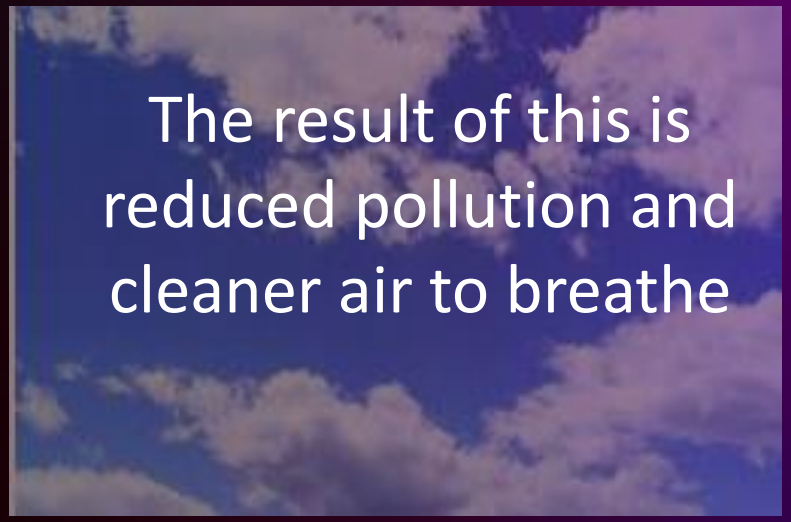


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

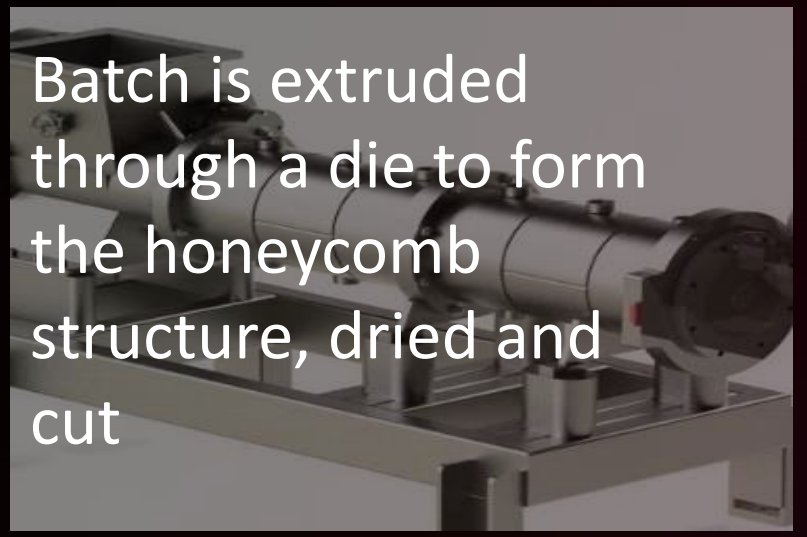
Business Context









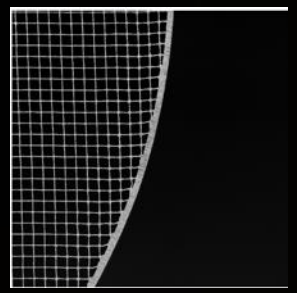


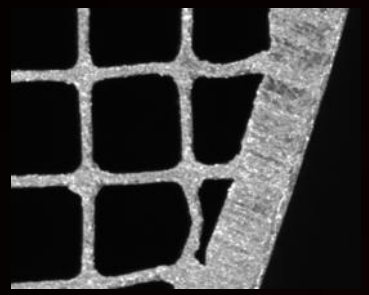


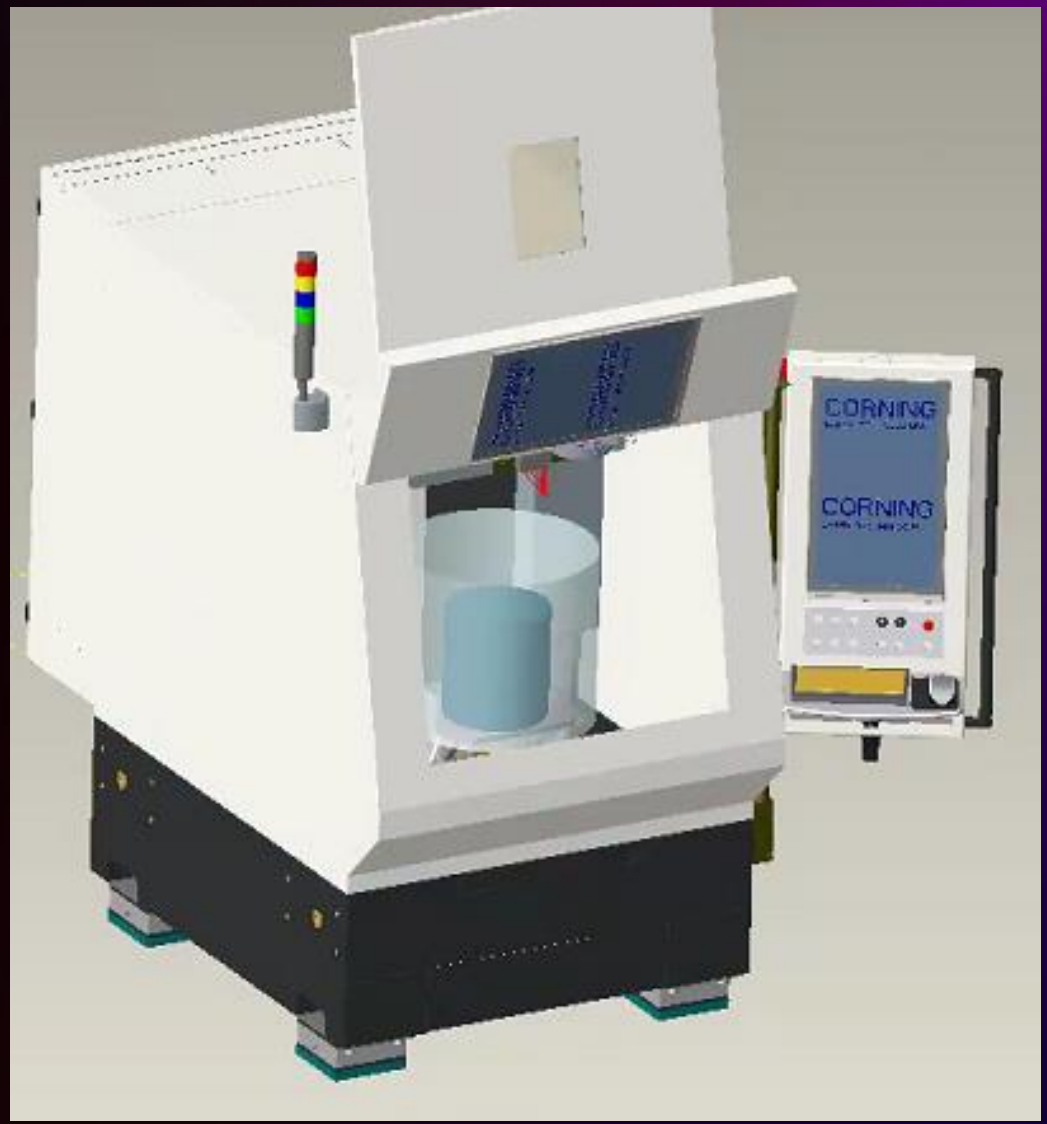
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What is Veridi?

- Area scanning technology with bright/dark field lighting
- Captures images at a high resolution
- Performs defect detection
- Provides real-time forming feedback in plants with short cycle time







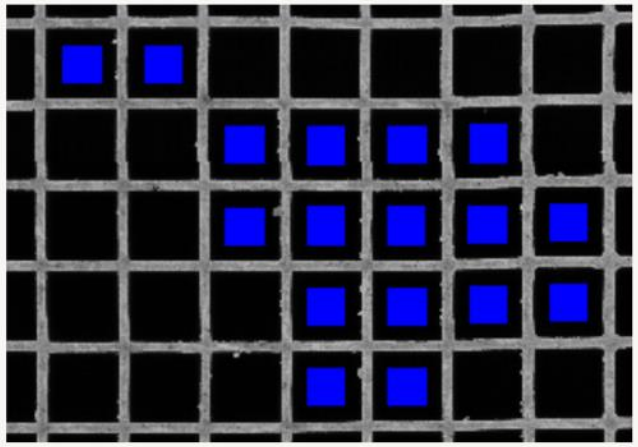
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Effects of Debris

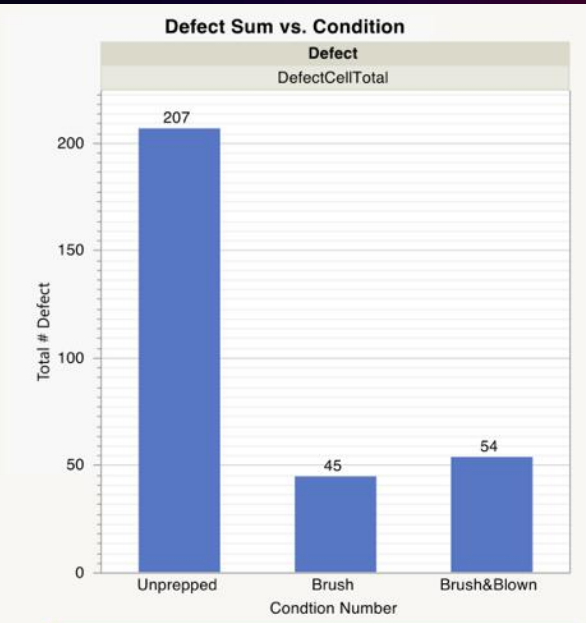
Dust can cause inaccurate measurements that will lead to wrong decisions.

Brushing and blowing out the part is required before each part scan.

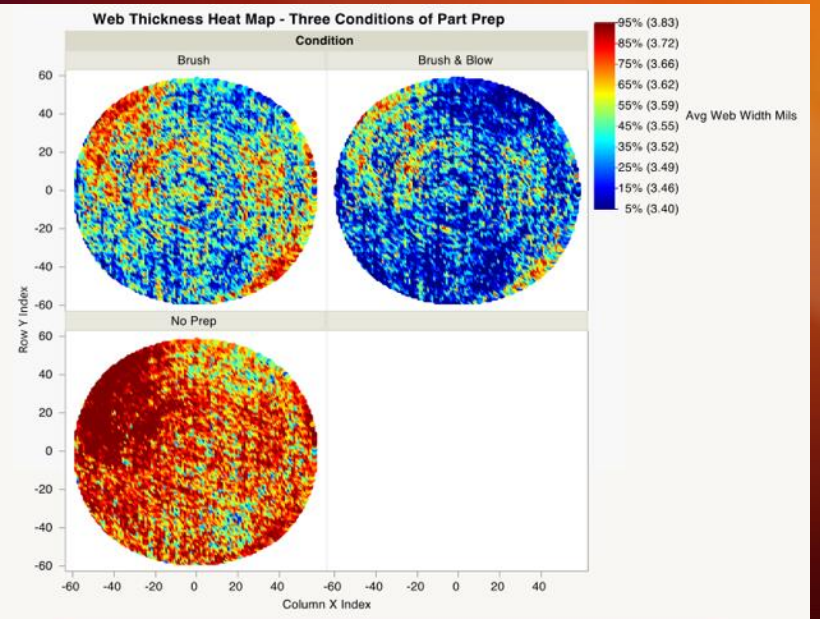
We can have 4x the number of defects flagged on our parts and a 0.2 mil web thickness increase without part prep.



False Distortion that was flagged due to debris



Defect Sum of Different Part Preps



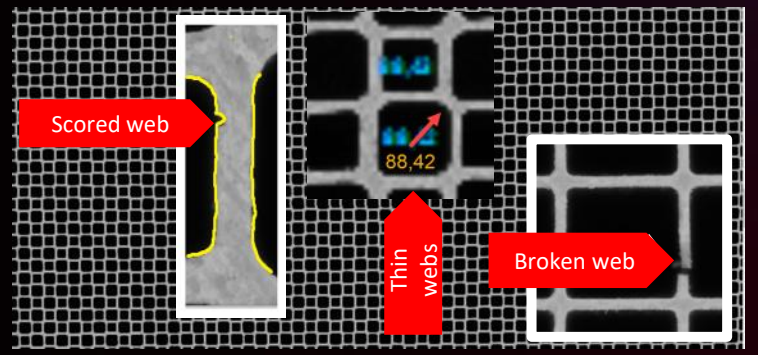
Increase in Web Thickness without Part Prep

aws

Business Need

Tighter emission regulations result in higher product quality requirements

Need significantly more product measurements with higher accuracy



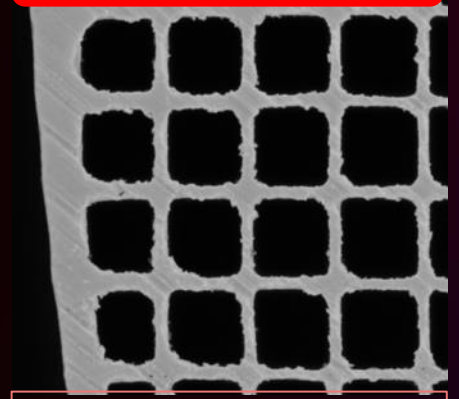
Deployed a new high resolution imaging system



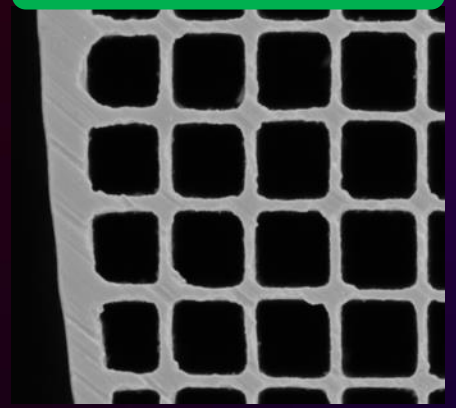
Veridi Production Unit

The manufacturing process creates dust leading to inaccurate results

Debris



No Debris



False Distortion that was flagged due to debris

Inaccurate results led to incorrect and costly decisions in manufacturing

- High-resolution measurement cycle time is long
- Detect debris, stop the cycle
- Improve consistency of the manual part preparation

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Team



Amanda Yoder
Business Stakeholder and Mastermind



Mengying Xiao
Application Engineer



Max Dutkin
Platform Engineer



Leslie Lowman
Agile Coach



Denis Kamotsky
ML Engineer



Xiaotian Zou
Data Scientist

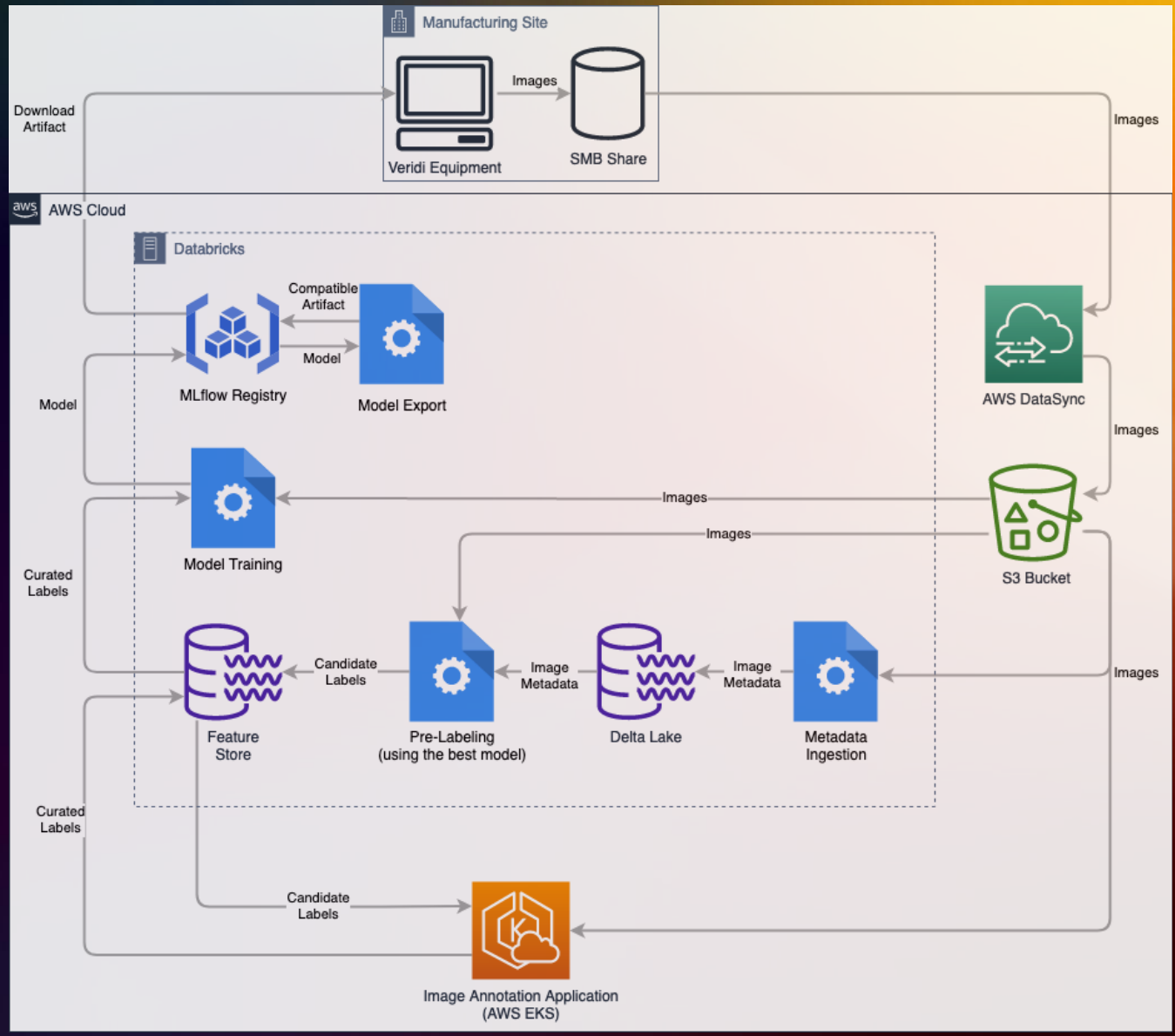
aws
© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Solution: Learning Loop

Supervised deep learning model was trained in the cloud and deployed at the edge (in factory).

Model stops the physical measurement and prevents using incorrect product data.

Supervised drift detection through periodic evaluation of manually labeled samples.



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Why Databricks

Data Engineer



- Scala / Python
- Spark / SQL
- Jobs / Pipelines
- Ingestion Patterns
- Data Versioning
- Delta Lake
- Scalable Storage



ML Engineer



- Python
- ML Libraries
- Notebooks
- Experimentation
- Model Versioning
- MLflow
- Scalable Compute (GPU)

aws

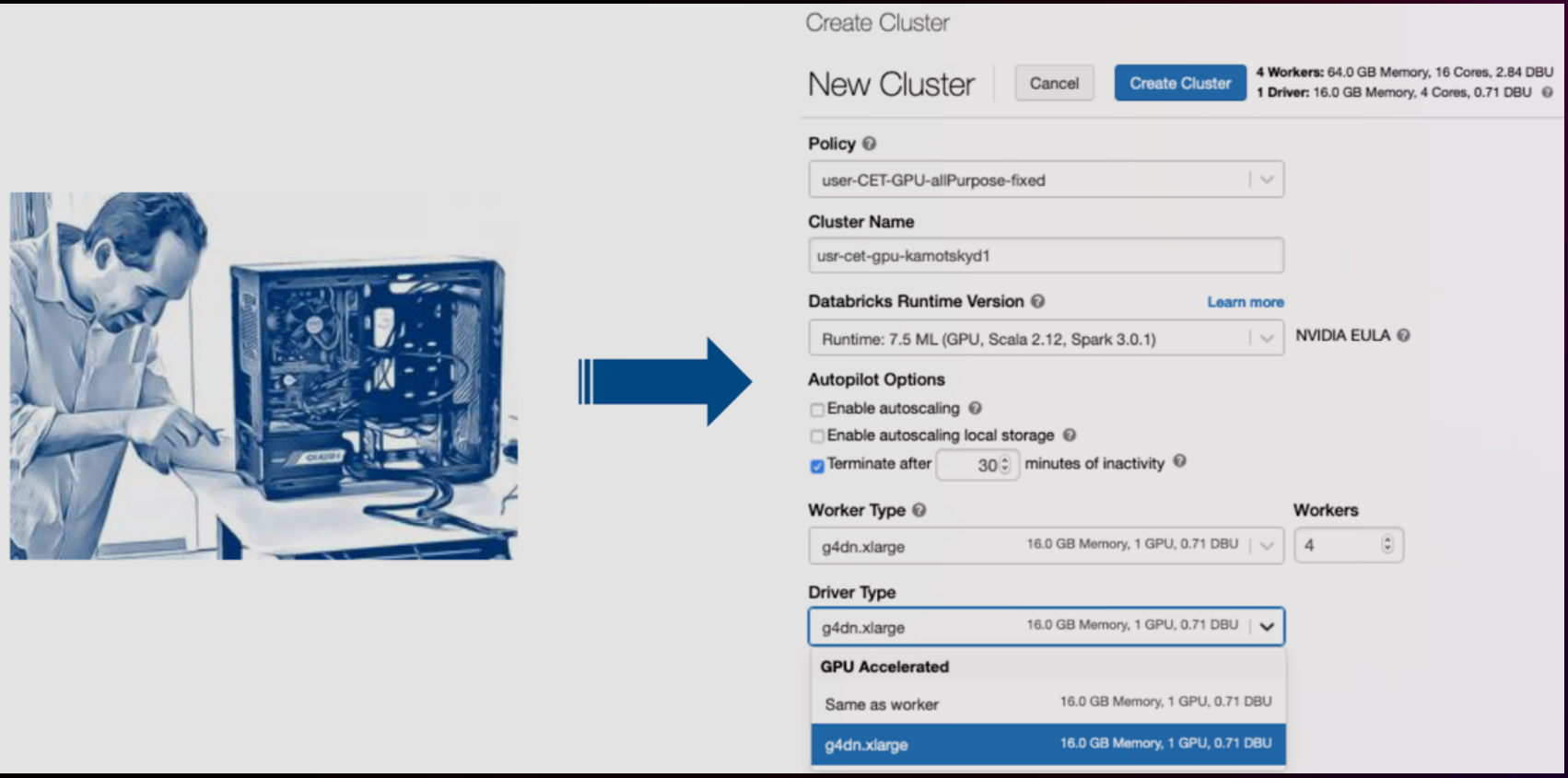
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

General - Corning (L4)

Corning's Advanced Analytics Platform

- Databricks on AWS
- Open-Source on premises
- Distributed Computing based on Apache Spark

- Easy to store large volume of images in S3 Buckets
- Optimized data storage in Delta Lake for faster access
- Rich Python ecosystem preinstalled and preconfigured
- MLflow integration



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Typical Keras Computer Vision Classifier Training

- Organize images on the local filesystem into directories corresponding to class labels
- Configure Keras ImageDataGenerator to flow images from these directories using train/validation split and, potentially, some image augmentations like zooming, shifting, or brightness changes
- Create Keras model (potentially loading weights from pretrained)
- Compile Keras model
- Create LR scheduling and early stopping callbacks
- Fit Keras model on the ImageDataGenerator
- Save the model
- Create evaluation ImageDataGenerator, load model, evaluate

Example of using `.flow_from_directory(directory)`:

```
train_datagen = ImageDataGenerator(
        rescale=1./255,
        shear_range=0.2,
        zoom_range=0.2,
        horizontal_flip=True)
test_datagen = ImageDataGenerator(rescale=1./255)
train_generator = train_datagen.flow_from_directory(
        'data/train',
        target_size=(150, 150),
        batch_size=32,
        class_mode='binary')
validation_generator = test_datagen.flow_from_directory(
        'data/validation',
        target_size=(150, 150),
        batch_size=32,
        class_mode='binary')
model.fit(
        train_generator,
        steps_per_epoch=2000,
        epochs=50,
        validation_data=validation_generator,
        validation_steps=800)
```

https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/image/ImageDataGenerator#flow_from_directory

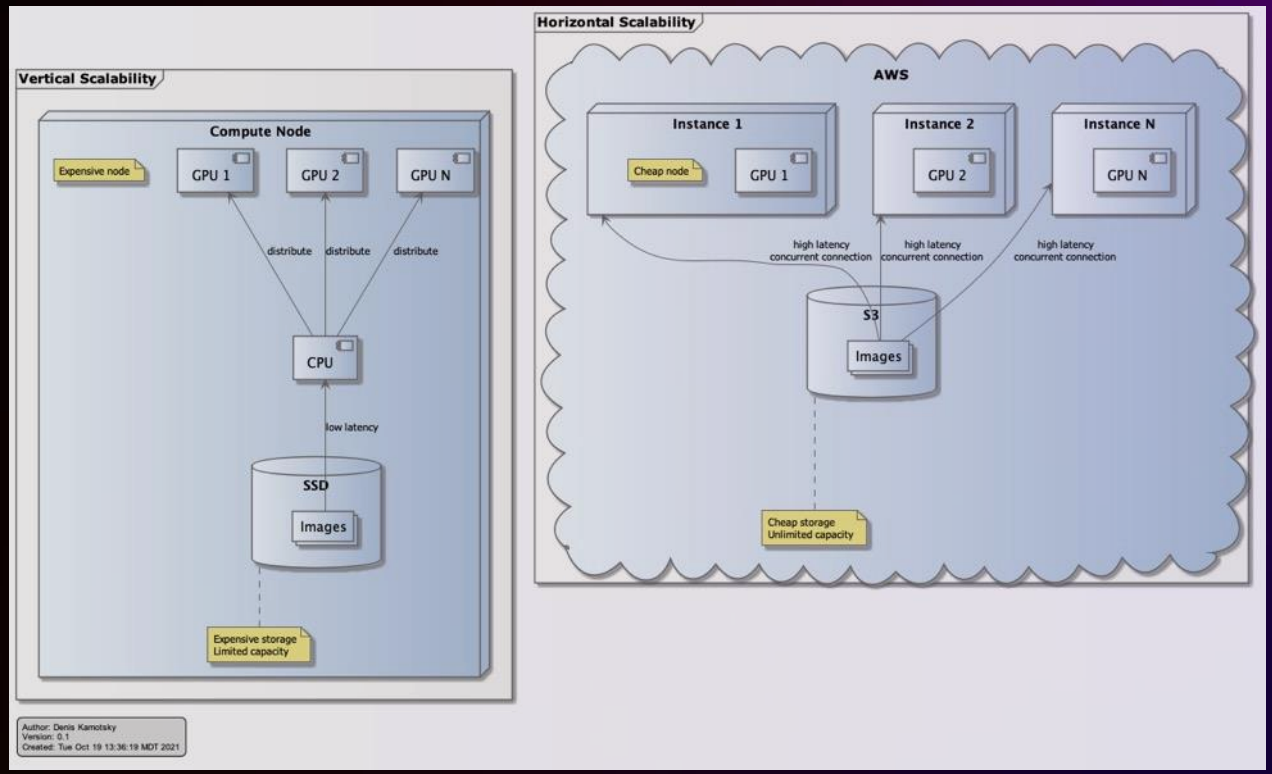
aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Challenges

- Not Cloud-friendly: biased toward vertical scalability (single large node)
  - Easy to use multiple GPUs on the same node. Harder to set up a distributed environment. In the Cloud, cheaper to have multiple single-GPU nodes.
  - Images flow from the local filesystem, which is assumed to be fast (SSD). In the Cloud cheaper to store images in S3, which is slow but distributed.

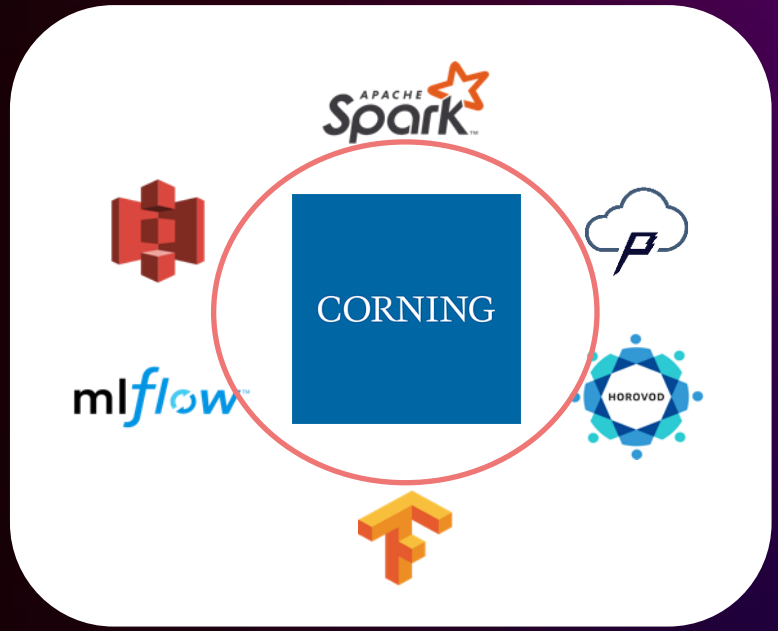
- Tracking training progress requires custom callbacks
- Tracking hyperparameters and metrics across multiple training sessions is a challenge
- Model evaluation generates only basic metrics
- Model is not interpretable
- Save format is not portable between Tensorflow versions



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Reusable Computer Vision Library

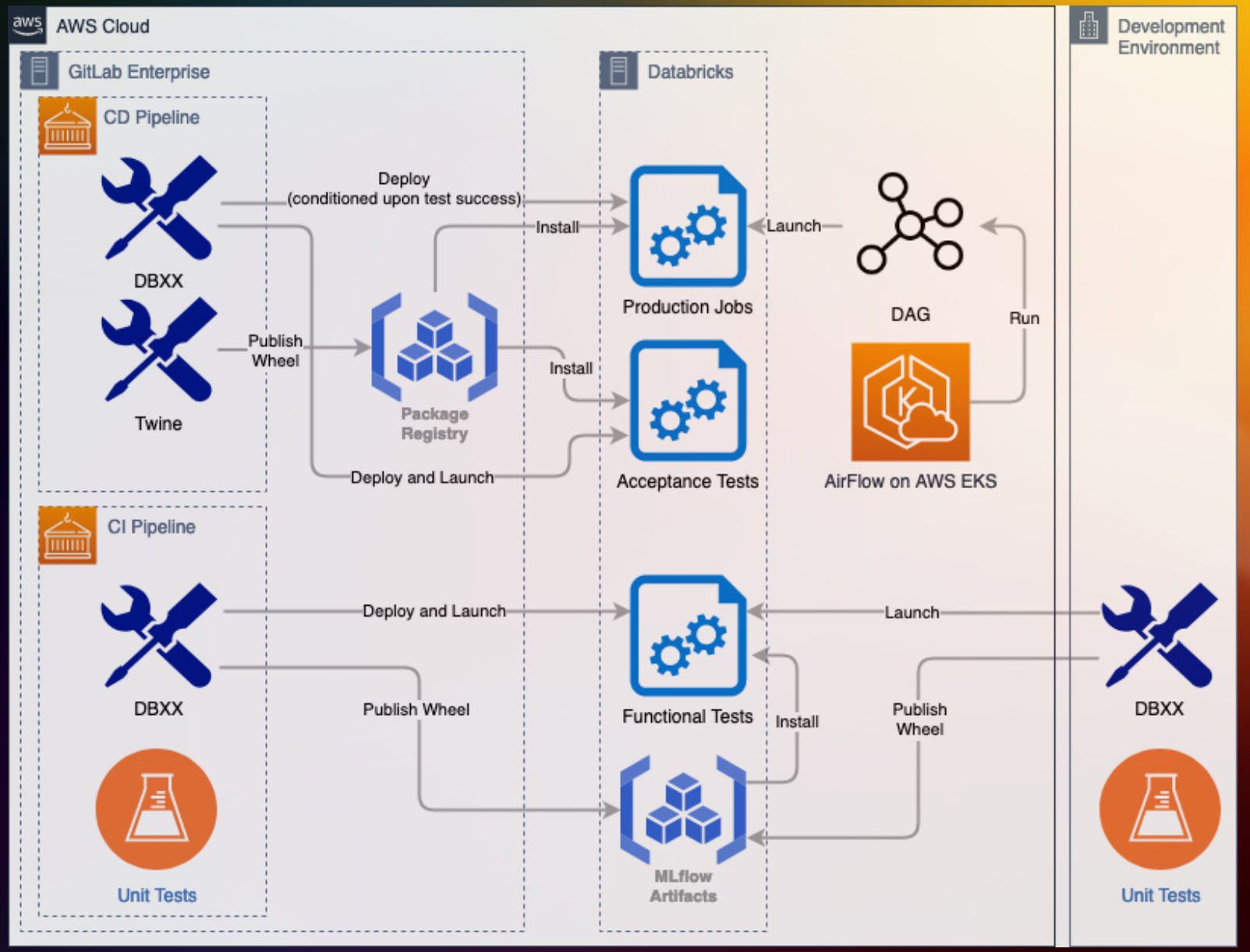
- Low-code Python library developed in Corning: integrates Horovod, Petastorm, Spark, Tensorflow, Mlflow
- Distributed via “pip install” from the Corning Toolchain
- Will be open-sourced, stay tuned…
- Designed for Databricks, applicable to open-source Spark
- API inspired by Keras API, has a familiar feel
- Trains Keras-based computer vision models
- Trains and evaluates models on distributed GPU-enabled clusters, out of the box
- Tracks everything it does in MLflow, out of the box
- Includes additional features:
  - Multiple model formats
  - Interpretability with GradCam and LIME
  - Advanced image augmentation pipelines
  - Many more…



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Continuous Delivery

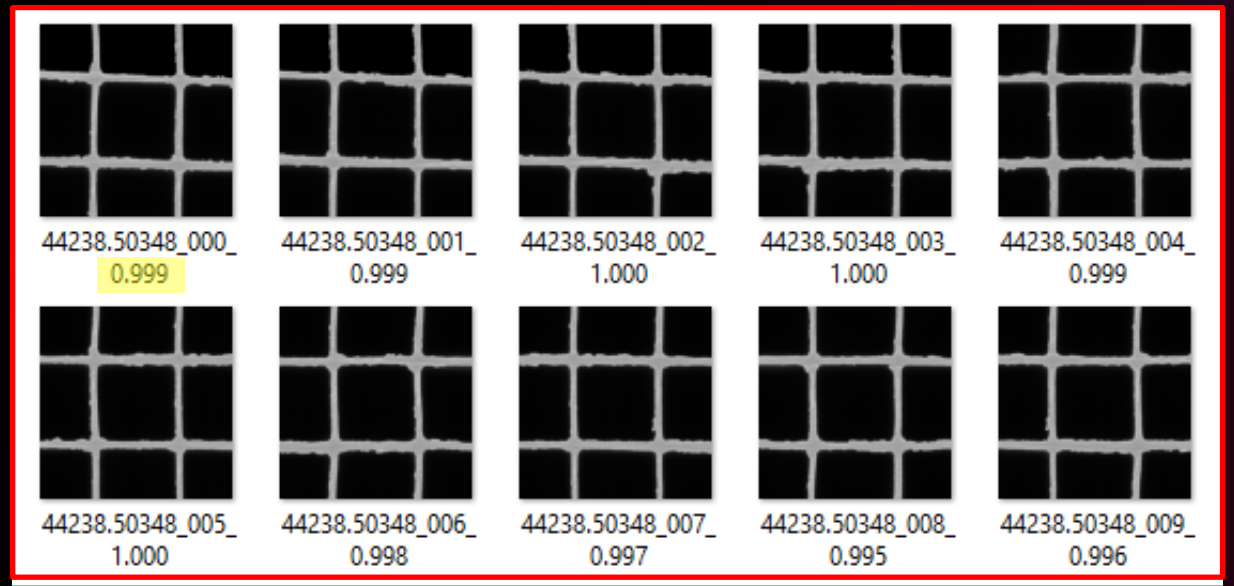
- DBXX tool: DBX adjusted to Corning needs
- Databricks jobs with functional tests
- Two paths for artifacts
  - Package registry for released
  - MLflow for unreleased
- Self-managed AirFlow on EKS for job orchestration in Production



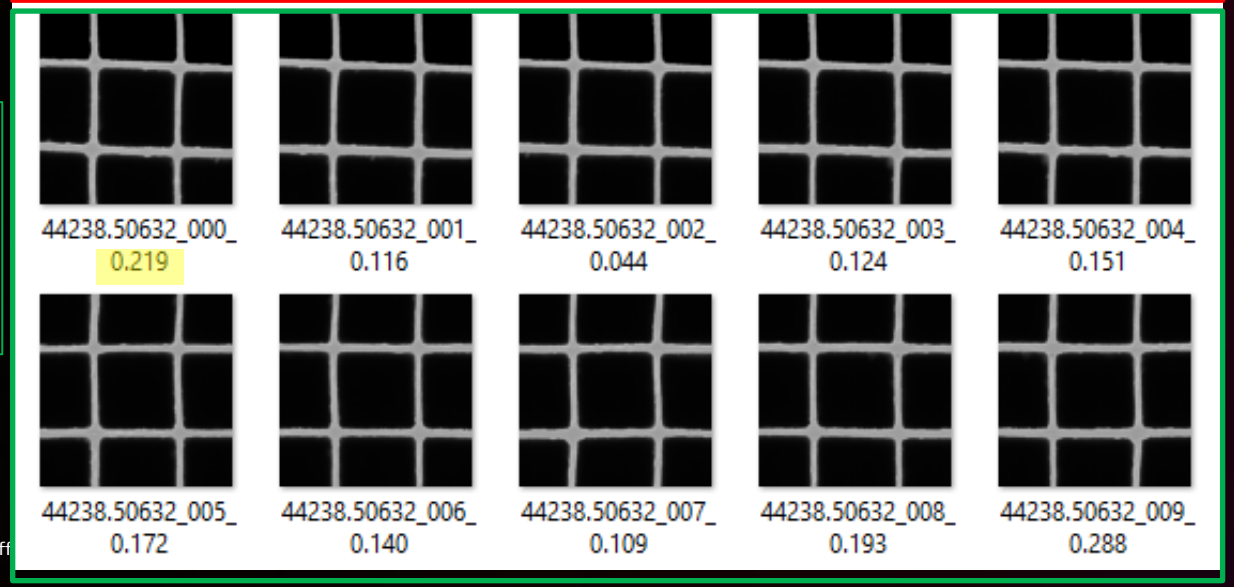
© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Model in Action

Flagged
Part is dirty
Machine will fault and ask operator to rescan.



Accepted
Part is clean
Machine will scan.



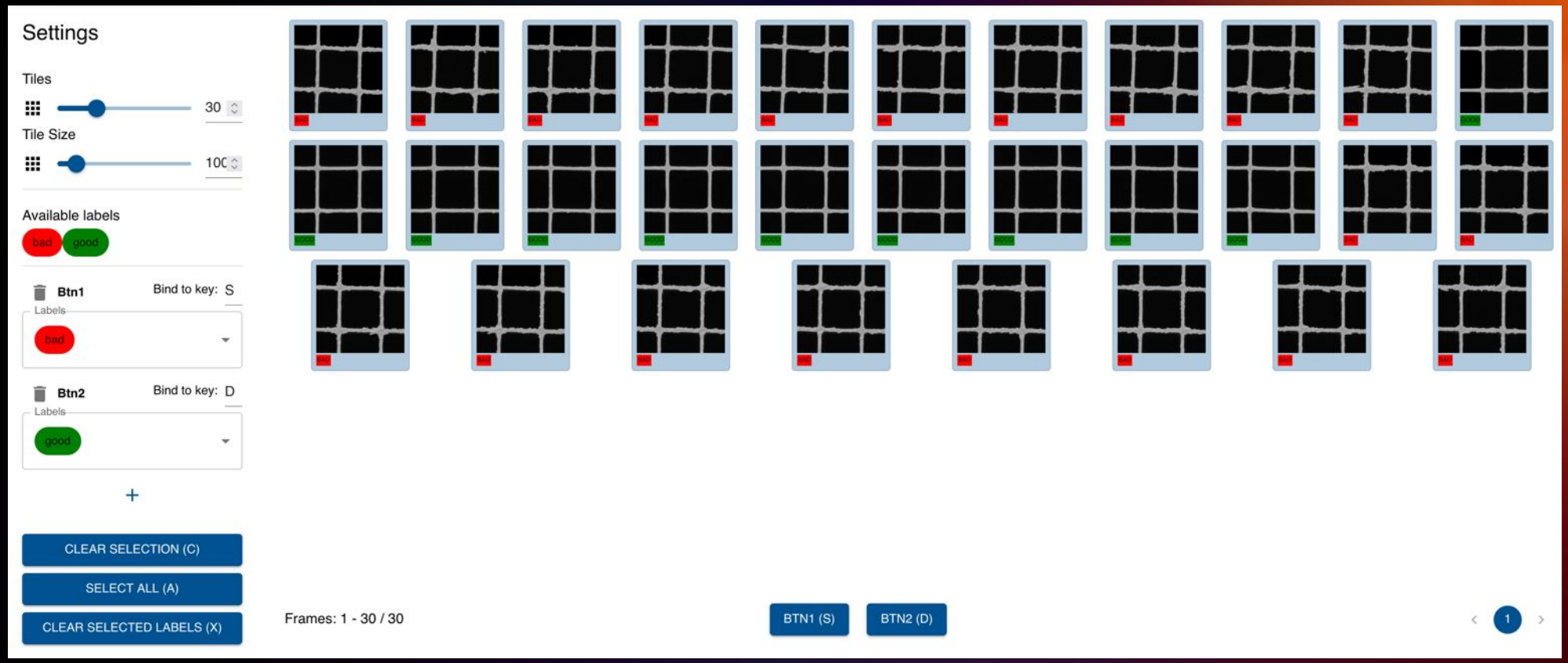
- If part is predicted as unacceptable, then Error Message will appear on Veridi:
  "Failed Debris Check"
- This will fault the machine before scanning which prevents bad data from going to the database.

aws
© 2022, Amazon Web Services, Inc. or its aff
General - Corning (L4

Image Labeling Application

Based on open-source CVAT API

UX optimized for bulk annotation tasks

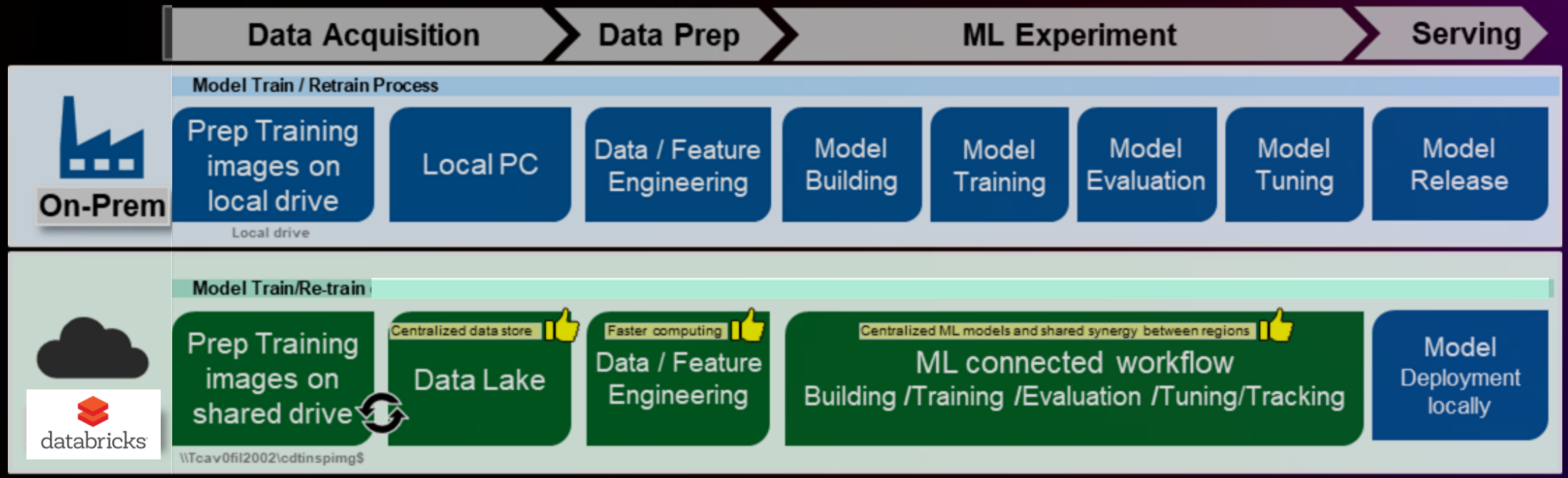


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Business Benefits

Corning CET has benefited for the application of deep learning models for direct use inside new production measurement equipment resulting in $2M of cost avoidance through elimination of finishing defect upset events.

We are excited to deploy deep learning models in other applications throughout the company.



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

2022 Manufacturing Leadership Council Award: AI and Machine Learning



"Corning doesn't simply follow the herd. We look at 'digital' through a Corning-specific lens: What problem are we trying to solve? How can we make our manufacturing execution be more impactful?"

John McGirr
Vice President, Manufacturing, Corning

Manufacturing Leadership Council

"Corning is part of a high-performing contingent of award winners that serve as an example of what's possible in manufacturing's emerging digital era."

Penelope Brown
Senior Content Director, Manufacturing Leadership Council

Winner: Corning Environmental Technologies' Machine Learning for Quality Systems

Machines are helping operators be extra precise when making our ceramic substrates. With the development of high-resolution imaging equipment, we've taken our machinery to the next level to improve product quality.

Machine learning at Corning has enhanced the collection of data, accurately measuring parameters and pinpointing possible problems at the source –with less machine and operator downtime.

The result? Smarter machines. Smarter business.

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!



Please complete the session survey in the **mobile app**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.