

Unified ML Monitoring Hands-on Workshop

Presented by

Max Fisher, *Solution Architect*

Amy Wang, *Senior Solution Architect*

Meet the Presenters



Max Fisher
Solution Architect
max.fisher@databricks.com



Salma Mayorquin
Solution Architect
salma.mayorquin@databricks.com

Agenda

- Introductions, 5m
- Databricks Overview & Lakehouse, 5m
- Unified ML Monitoring Overview, 15m
- Hands-on lab, 55m
- Q&A, 10m



Lakehouse

One simple platform to unify all of
your data, analytics, and AI workloads

CUSTOMERS

5000+

Across the globe

ORIGINAL CREATORS



We're working with enterprises in every industry

Healthcare & Life Sciences

Humana

AMGEN®

OPTUM



Biogen

NHS

REGENERON

AstraZeneca

Manufacturing & Automotive

Schneider
Electric

DAIMLER



Media & Entertainment

CONDÉ NAST

RIOT
GAMES

SHOWTIME

COMCAST



VIACOM



Energy & Utilities

TOTAL

devon



ExxonMobil

aggreko

Quby

Financial Services

Financial Services

HSBC

BNP PARIBAS

Nasdaq



Nationwide

CREDIT SUISSE

ABN AMRO

Public Sector

cfpb

CMS



U.S. Citizenship
and Immigration
Services

LOS ANGELES
FIRE
DEPARTMENT

New York Power
Authority

Retail & CPG

sam's club

CVS Health



Foot Locker

7-ELEVEN

Henkel

MARS

H&M

Digital Native

Digital Native

gousto



DOLLAR SHAVE CLUB

Grab

zalando

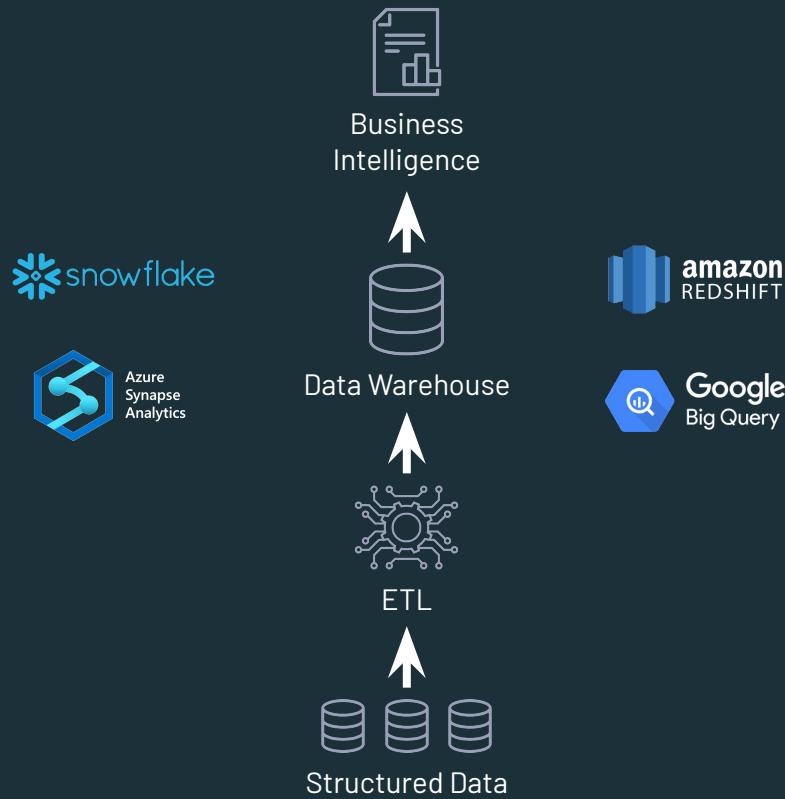
BUTCHER
BOX

wehkamp

SCRIBD

SEVATEC
INNOVATION TO SERVICE

What is the “Lakehouse” Paradigm?



Data Warehouses

Pros

- Great for Business Intelligence (BI) applications

Cons

- Limited support for Machine Learning (ML) workloads
- Proprietary systems with only a SQL interface

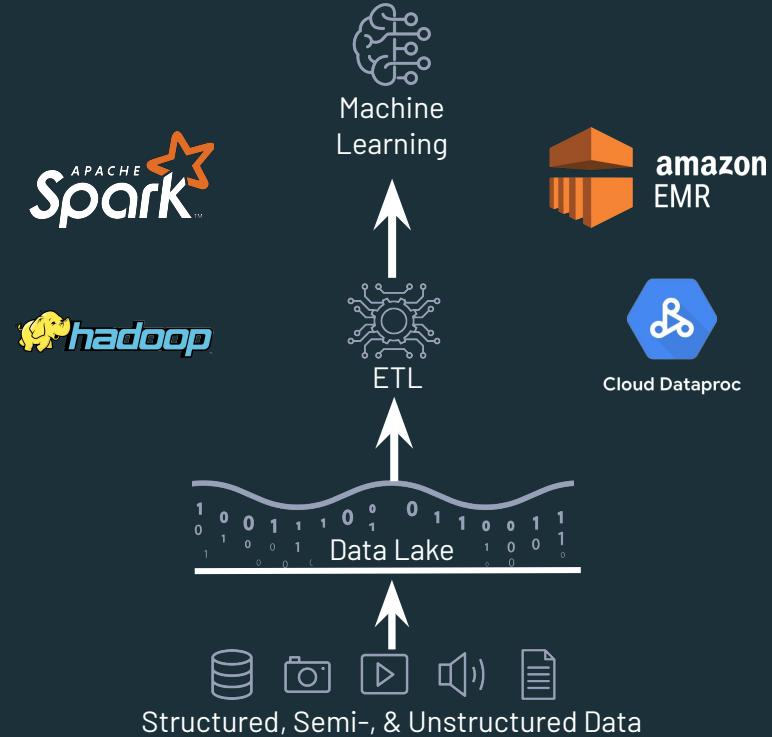
Data Lakes

Pros

- Supports ML
- Open formats and big ecosystem

Cons

- Poor support for BI
- Complex data quality problems



Today, most enterprises struggle with data

Data Warehousing



Data Analysts



Data Engineering



Data Engineers

Streaming



Data Engineers

Data Science & Machine Learning



Data Scientists

Siloed data teams decrease productivity

Amazon Redshift

Azure Synapse

Snowflake

SAP

Teradata

Google BigQuery

IBM Db2

Oracle Autonomous Data Warehouse

Hadoop

Amazon EMR

Google Dataproc

Apache Airflow

Apache Spark

Cloudera

Apache Kafka

Apache Flink

Azure Stream Analytics

Tibco Spotfire

Apache Spark

Amazon Kinesis

Google Dataflow

Confluent

Jupyter

Azure ML Studio

Domino Data Labs

TensorFlow

Amazon SageMaker

MatLAB

SAS

PyTorch

Disconnected systems and proprietary data formats make integration difficult

Siloed stacks increase data architecture complexity



Analytics and BI



Data marts



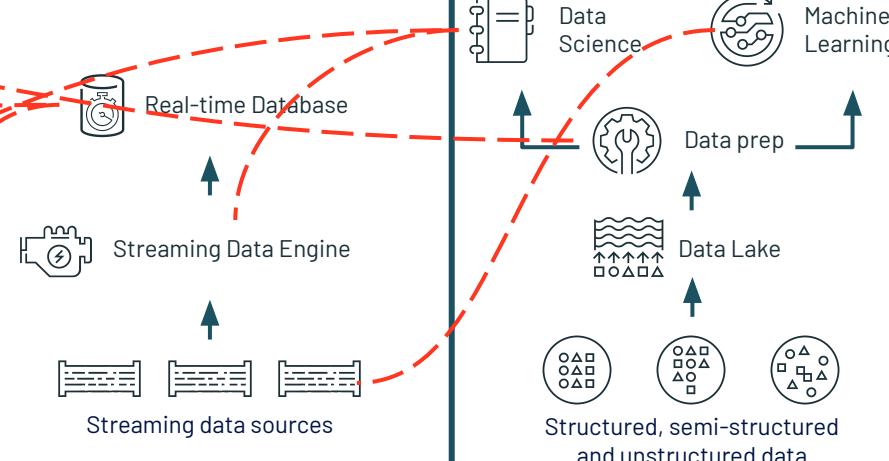
Data warehouse



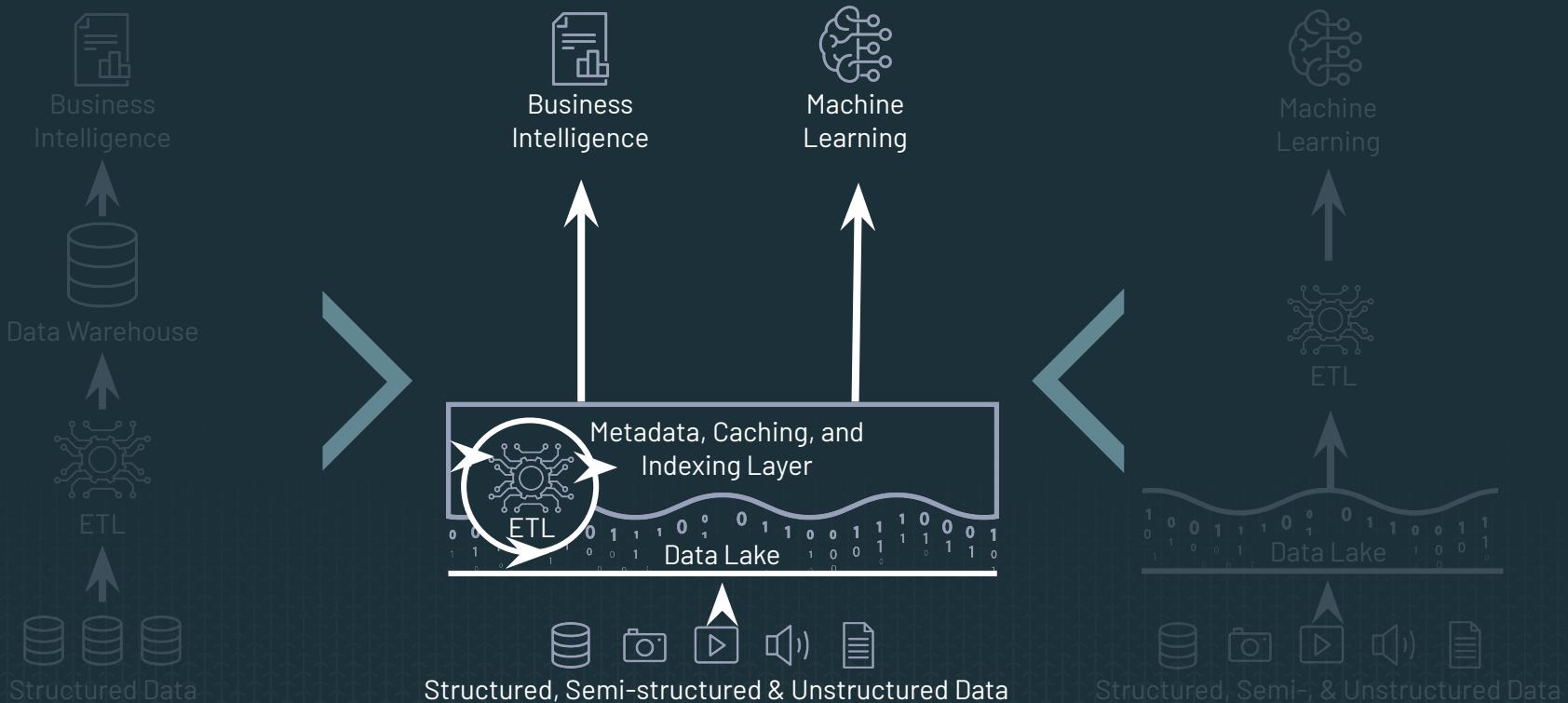
Structured data



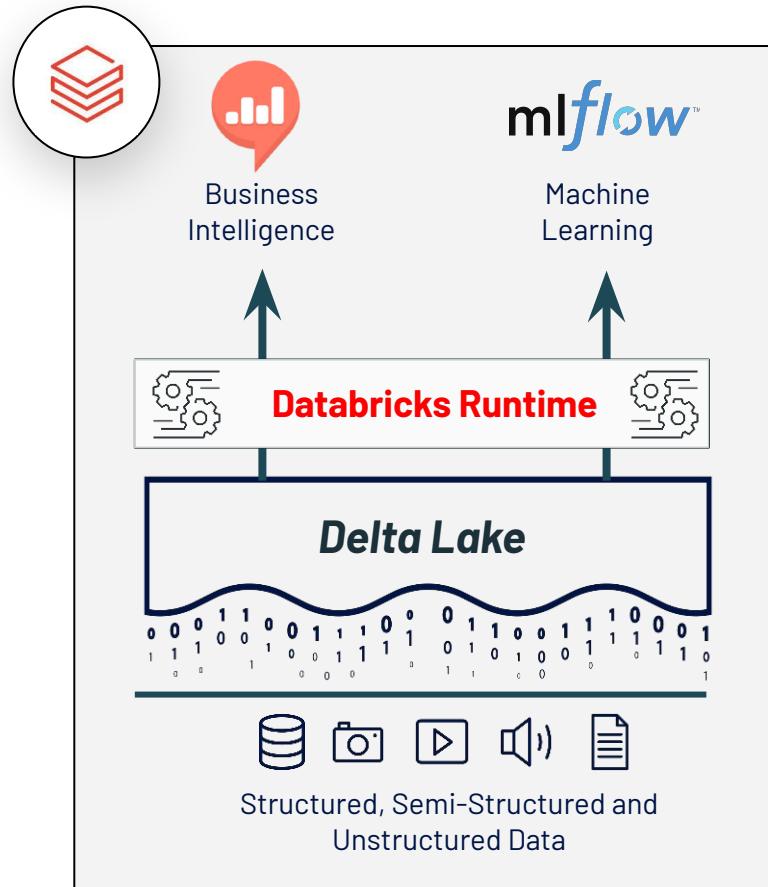
Structured, semi-structured and unstructured data



New Way Forward: lakehouse



The Databricks Lakehouse Platform



Introduction to the Lab



Quick Poll: What is the biggest challenge you face with your Machine Learning Projects?

- Figuring out the right algorithm to use to train your model?
- Tracking all the data associated with training and testing your model?
- Operationalizing your finished models?
- Governing the lifecycle of your machine learning models?

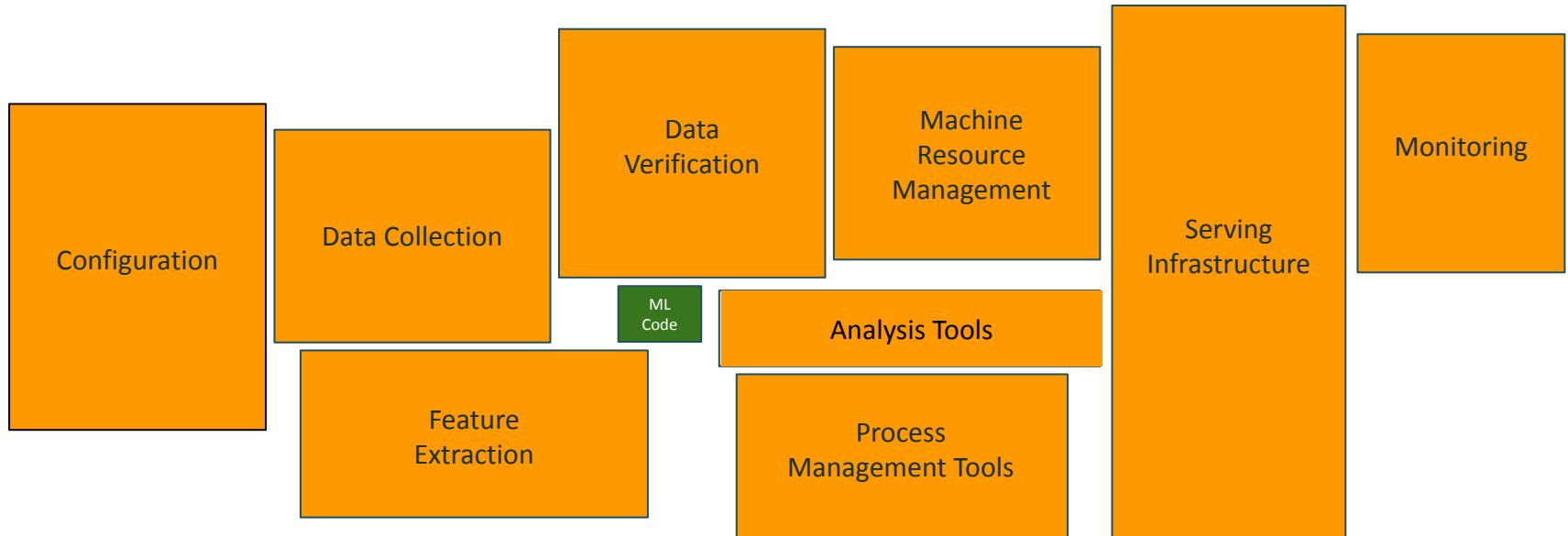
Quick Lab Setup

- Start Your Engines (Clusters)!
 - Go to the Databricks Workspace and Create / Start the following:
 - Interactive Cluster | Databricks Runtime 7.5+ ML
 - SQL Endpoint | Small+
- Upload the Notebooks

Unified ML Monitoring on Databricks

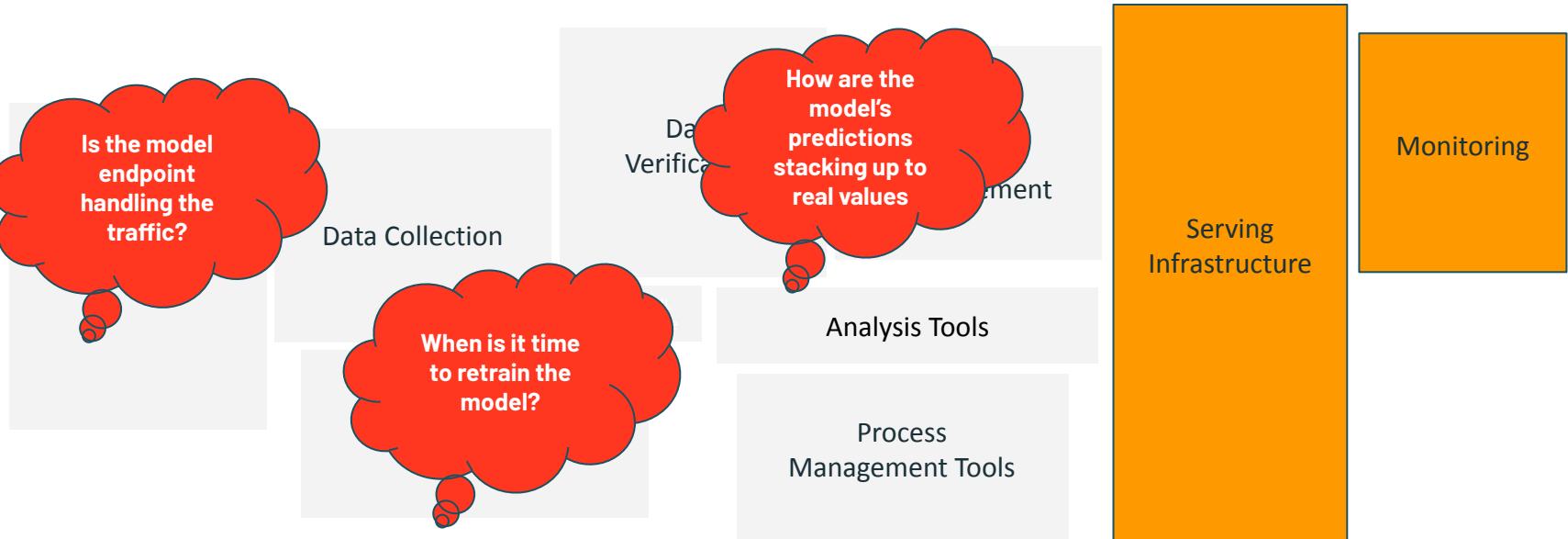
- Why Monitoring Models Is Hard
- How Databricks Completes the Picture
- **Lab:** Unified ML Monitoring Dashboard

Hardest Part of ML isn't ML, it's Data



Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small green box in the middle. The required surrounding infrastructure is vast and complex.

Monitoring Can Be A Complex Part of this Problem



The Answers Live In Many Places



Azure Machine Learning



Azure Log Analytics



Amazon SageMaker



Amazon CloudWatch





ML Model



ML Model



mlflow™

/Shared/UMLWorkshop/sensor_prediction_model

Track machine learning training runs in an experiment. Learn more

Experiment ID: 1678798819384765 Artifact Location: dbfs:/databricks/mlflow-tracking/1678798819384765

Notes

Showing 18 matching runs

Columns Filter Clear

Start Time	Run Name	User	Source	Version	Models	Parameters	Metrics	Tags			
2021-05-18 13:44	Sensor..._maxfl..._01-hr	-	sleem...	True	0.0	mse	0.019	0.008	0.998	sklear...	Rando...
2021-05-18 13:44	Sensor..._maxfl..._01-hr	-	sleem...	True	0.0	mse	0.017	0.002	0.999	sklear...	Rando...
2021-05-18 13:44	Sensor..._maxfl..._01-hr	-	sleem...	True	0.0	mse	0.018	0.004	0.998	sklear...	Rando...
2021-05-18 13:44	Sensor..._maxfl..._01-hr	-	sleem...	True	0.0	mse	0.019	0.005	0.998	sklear...	Rando...
2021-05-18 13:44	Sensor..._maxfl..._01-hr	-	sleem...	True	0.0	mse	0.02	0.007	0.997	sklear...	Rando...
2021-05-18 13:44	Sensor..._maxfl..._01-hr	-	sleem...	True	0.0	mse	0.021	0.012	0.995	sklear...	Rando...
2021-05-18 13:44	Sensor..._maxfl..._01-hr	-	sleem...	True	0.0	mse	0.021	0.014	0.994	sklear...	Rando...
2021-05-18 13:44	Sensor..._maxfl..._01-hr	-	sleem...	True	0.0	mse	0.018	0.003	0.998	sklear...	Rando...
2021-05-18 13:44	Sensor..._maxfl..._01-hr	-	sleem...	True	0.0	mse	0.019	0.007	0.997	sklear...	Rando...
2021-05-18 13:44	Sensor..._maxfl..._01-hr	-	sleem...	True	0.0	mse	0.017	0.002	0.999	sklear...	Rando...
2021-05-18 13:44	Sensor..._maxfl..._01-hr	-	sleem...	True	0.0	mse	0.019	0.005	0.998	sklear...	Rando...
2021-05-18 13:44	Sensor..._maxfl..._01-hr	-	sleem...	True	0.0	mse	0.018	0.003	0.998	sklear...	Rando...
2021-05-18 13:44	Sensor..._maxfl..._01-hr	-	sleem...	True	0.0	mse	0.019	0.007	0.997	sklear...	Rando...
2021-05-18 13:44	Sensor..._maxfl..._01-hr	-	sleem...	True	0.0	mse	0.019	0.007	0.997	sklear...	Rando...
2021-05-18 13:44	Sensor..._maxfl..._01-hr	-	sleem...	True	0.0	mse	0.019	0.007	0.997	sklear...	Rando...
2021-05-18 13:44	Sensor..._maxfl..._01-hr	-	sleem...	True	0.0	mse	0.019	0.007	0.997	sklear...	Rando...
2021-05-18 13:44	Sensor..._maxfl..._01-hr	-	sleem...	True	0.0	mse	0.019	0.007	0.997	sklear...	Rando...
2021-05-18 13:44	Sensor..._maxfl..._01-hr	-	sleem...	True	0.0	mse	0.019	0.007	0.997	sklear...	Rando...

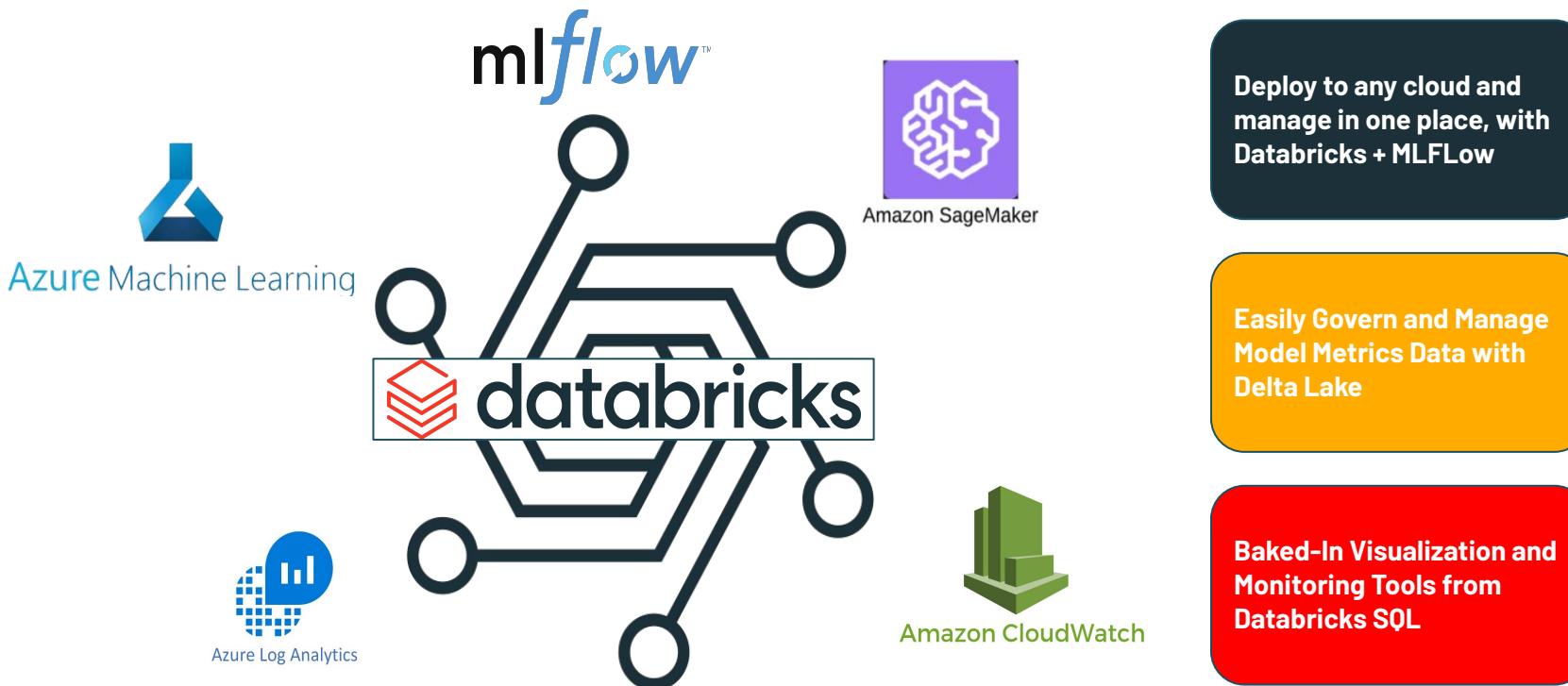
Support + troubleshooting
Get support request

How can I bring
all of this
information
together?

The screenshot shows a Databricks workspace interface. On the left, there's a sidebar with 'sensorprediction' selected. The main area displays MLflow artifacts for an experiment, including parameters, metrics, and tags. Below that is a 'Network bytes transacted' chart. To the right, there are two tabs: 'Logs' and 'Databricks Metrics'. The 'Logs' tab shows a table of log entries with columns for timestamp, log type, message, and source. The 'Databricks Metrics' tab shows a chart of metrics over time. At the bottom, there's a 'Logs' table from Azure Data Explorer with columns for timestamp, log type, message, and source.



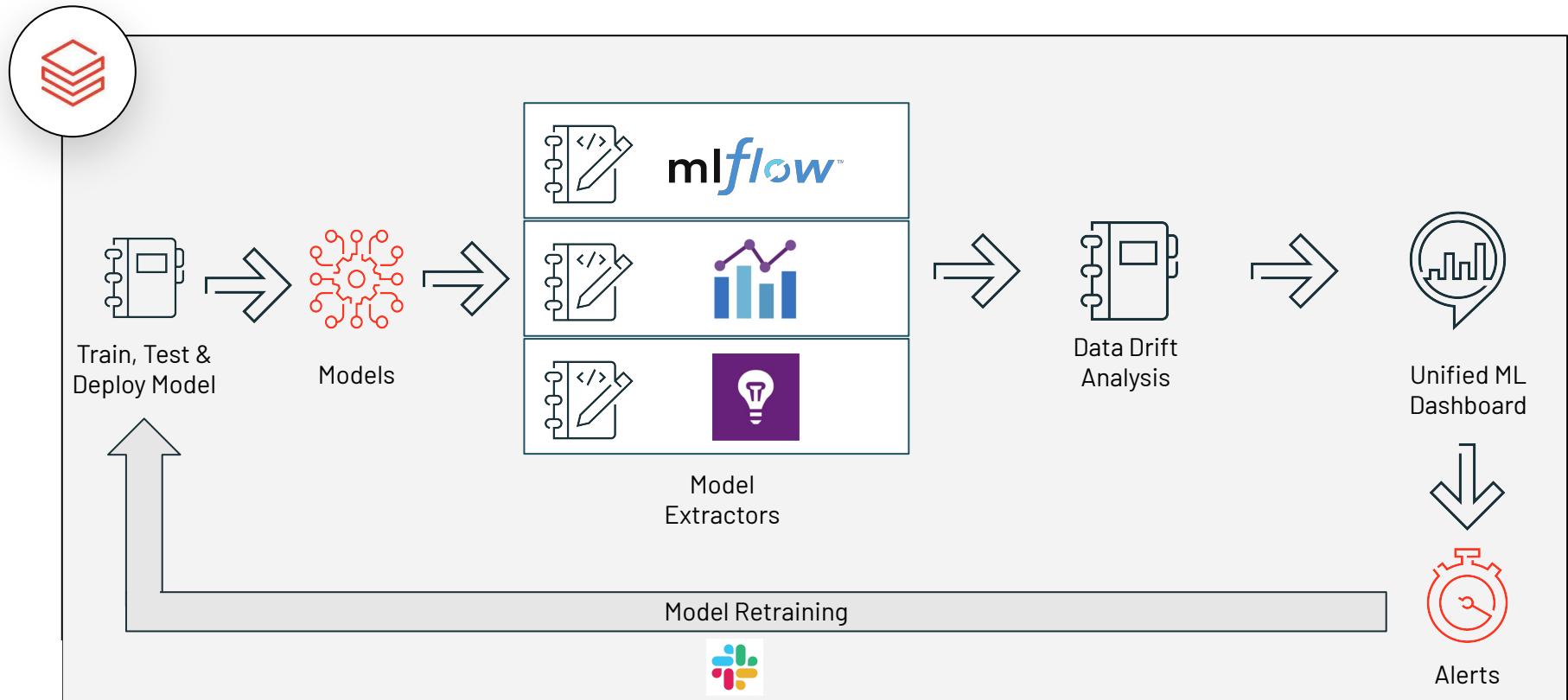
The Solution



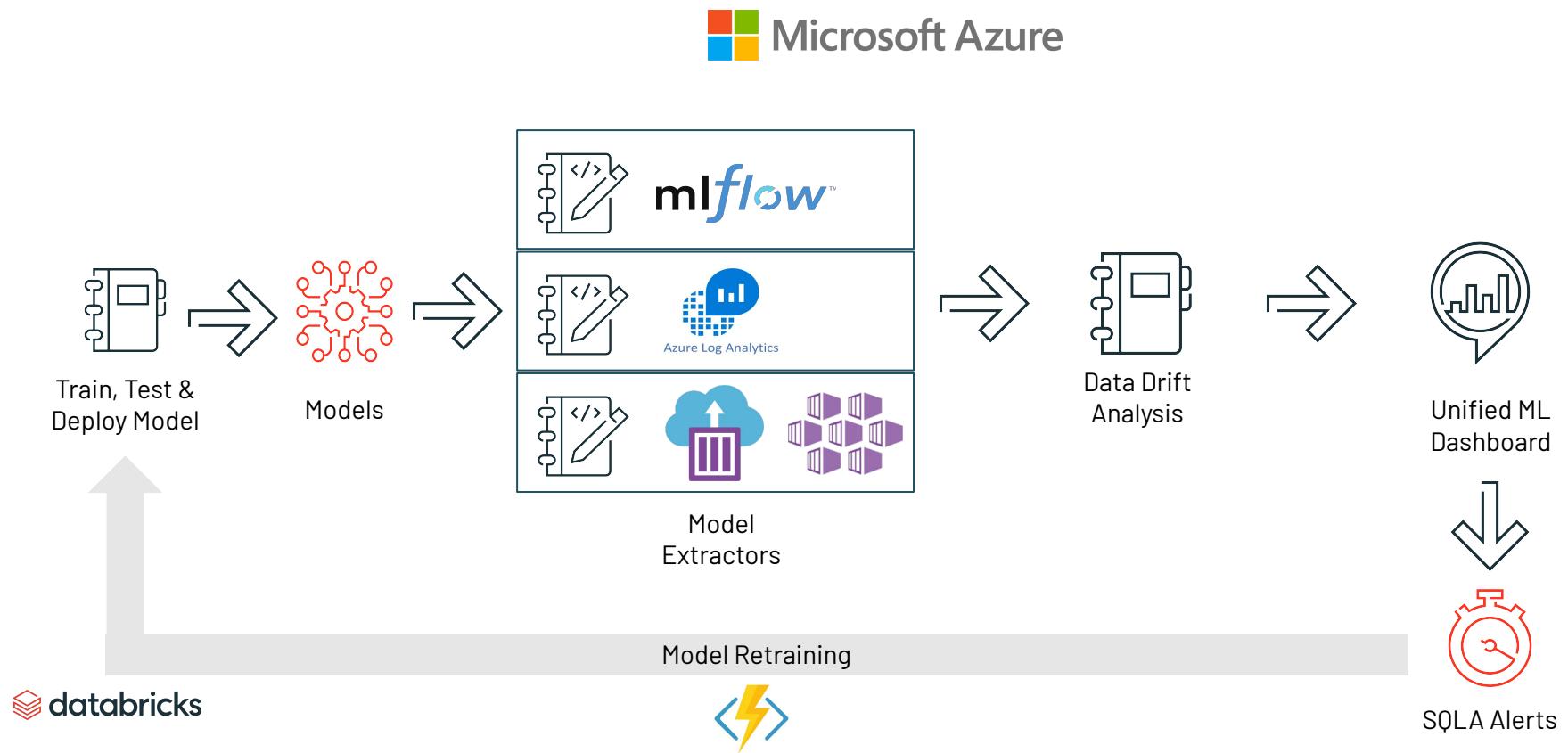
A photograph of a man with dark skin and short hair, wearing black-rimmed glasses and black over-ear headphones. He is looking down at a laptop screen, which is partially visible at the bottom right. A large, semi-transparent hexagonal grid pattern covers the background of the slide.

How Can I Do This in Databricks?

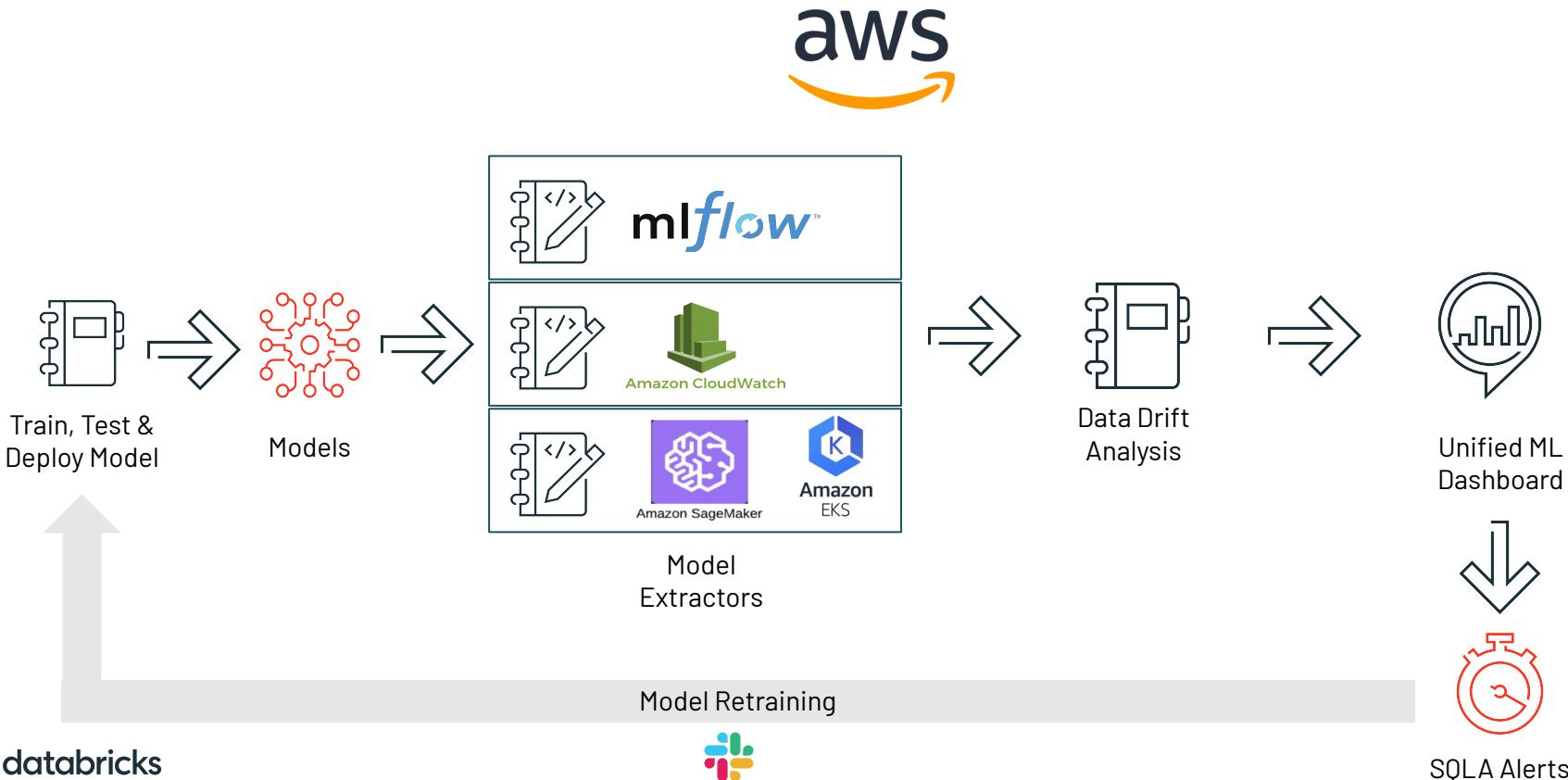
Architecture of the Solution



Architecture of the Solution



Architecture of the Solution



Hands-on Lab

Max will be leading the lab

Amy is available in the chat to answer questions or provide assistance

Content Links

- GitHub Repository for all Lab Materials:
 - [mpfis/unified-ml-monitoring-on-databricks\(github.com\)](https://github.com/mpfis/unified-ml-monitoring-on-databricks)
- [PDF of this Presentation](#)
- [Lab guide](#)

Similar Workshops

- Introduction to Delta and Lakehouse
 - Description
 - First and third Tuesday every month, [sign-up link](#)
- What's new in Databricks?
 - Description
 - First Wednesday every month, [sign-up link](#)

Q&A and Post-workshop Discussion

Thank you for joining us. We're here to answer any questions you have about the content, Databricks, or anything else. We'll send out a copy of this presentation after we wrap up.

We'll send out a survey after the workshop. Let us know what you liked and didn't like so that we can continue to improve.

If you're not currently in contact with your Databricks account team and would like to be, let the presenters know and we can assist.

Contact info for presenters: max.fisher@databricks.com, amy.wang@databricks.com

We're Hiring Solution Architects

If you or someone in your network is interested, we have East and West coast positions available.

Email ericka.styles@databricks.com with your resume or LinkedIn. You can ask your presenters for more information.

databricks.com/company/careers

Looking for applicants with backgrounds in any of these skills: data engineering, machine learning, **or** analytics.

Also hiring in Europe and Asia.



Why is capturing this information important?

83% CEOs say AI is a strategic priority

MIT Sloan
Management Review

\$3.9T Business value created by AI in 2022

Gartner.

85% Of big data projects fail

Gartner.

87% Of data science projects never make it into production

VB