DATA+AI SUMMIT
BY databricks

Responsible AI on Databricks

Lexy Kassan, Lead Data & AI Strategist
Omar Khawaji, Field CISO

Generative AI is taking the world by storm

91%

of organizations are experimenting with or investing in GenAI [1]

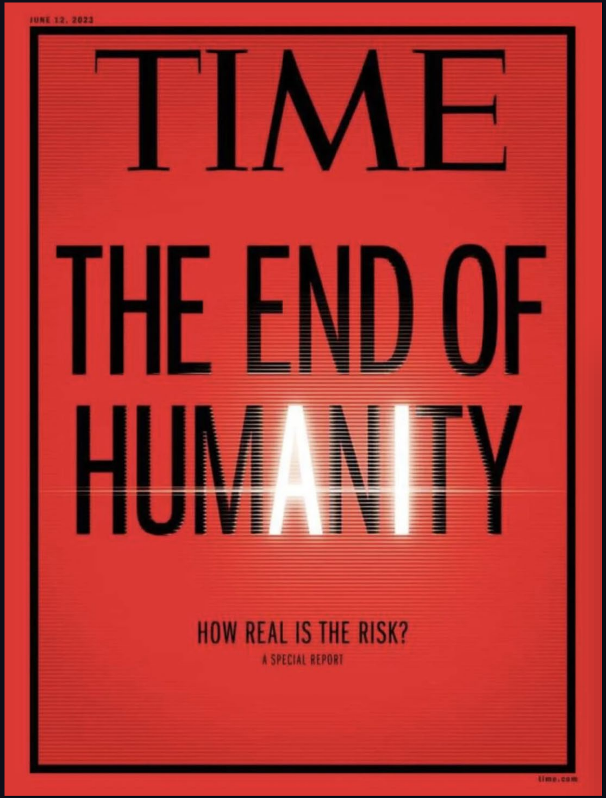
75%

of CEOs say companies with advanced GenAI will have a competitive advantage [2]

40%

increase in performance of employees who used GenAI [3]

1. Laying the foundation for data and AI-led growth, MIT Technology Review,
2. CEO decision-making in the age of AI, IBM Institute for Business Value
3. How generative AI can boost highly skilled workers' productivity, MIT Management Sloan School,



MOTHERBOARD
TECH BY VICE

The New GPT-4 AI Gets Top Marks in Law, Medical Exams, OpenAI Claims

The successor to GPT-3 could get into top universities without having trained on the exams, according to OpenAI.

With
Generative AI,
30%
of hours worked today could be **automated,** says McKinsey

Challenge:

Building and deploying production-quality Gen AI solutions

90%

of enterprises not confident going to production

Responsible AI Brings Value

Market leaders in AI are generating 50% more revenue growth than competitors.

High achievers are 53% more likely to develop responsibly by design.

43% of leaders believe that responsible AI attracts and retains talent

Accenture Study

80% of companies plan to increase investment in Responsible AI

Accenture: From AI Compliance to Competitive Advantage

Data & AI Ethical Ladder

SHOULD	Ethical
CAN	Responsible
MUST	Compliant

Key Concerns Eroding Trust in AI

Quality

Security

Governance

Model Quality

Challenges to Developing High Quality AI

Transparency

More effective models are often less explainable and interpretable

↓

Models cannot move to production

Effectiveness

Different tools by model type, evaluation metric, and development stage

↓

Operational expense, complex to maintain

Reliability

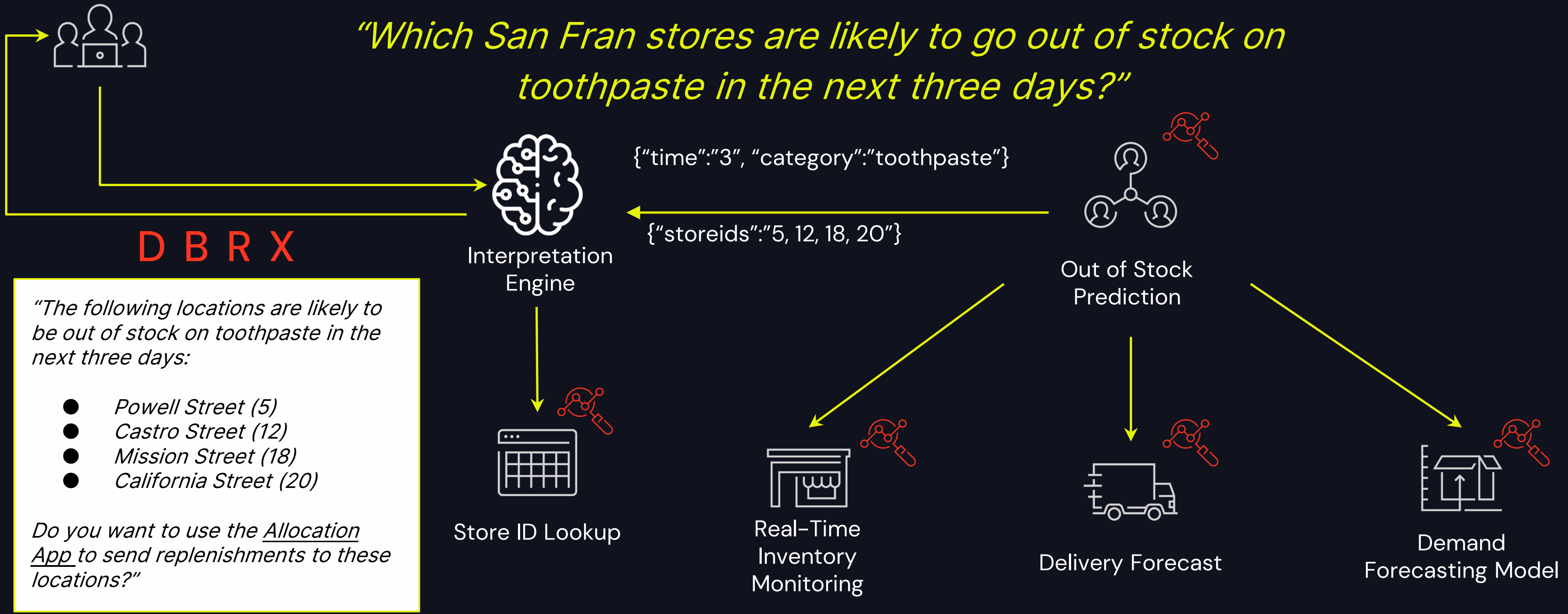
Iterating on models to keep them effective is disruptive to the business

↓

Reduced value to the business, disruption, risk

Transparency with Compound AI Systems

Use purpose-built and explainable agents, data, and tools



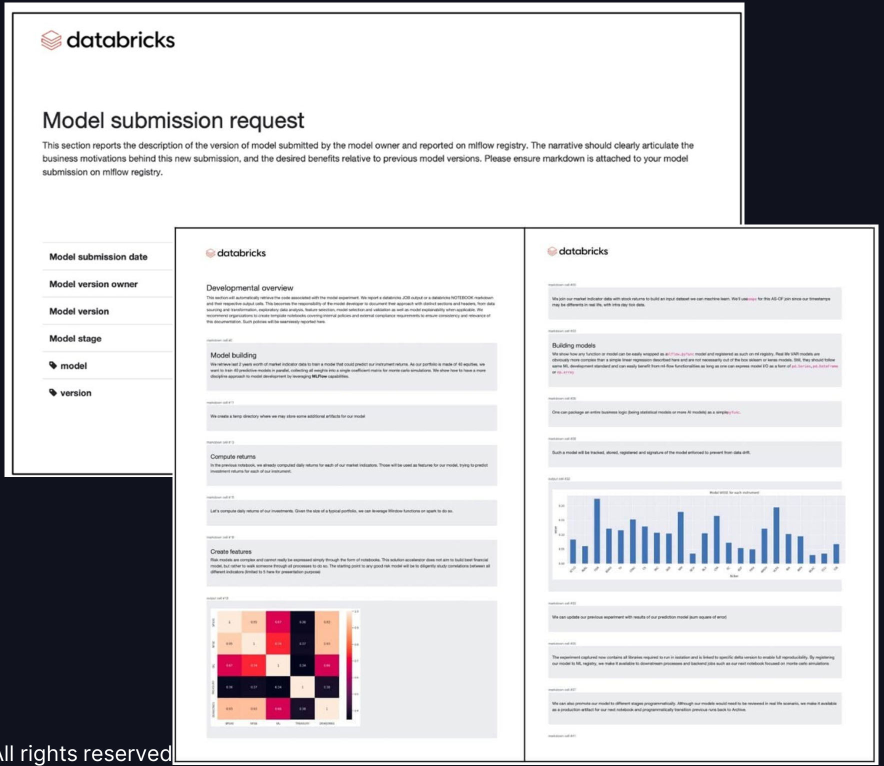
Automate Model Documentation

Generate the explanations you need to deploy with confidence

Leverage metadata you already have:

- Notebooks
- Unity Catalog
- MLflow
- Logging

[Model Risk Management Solution Accelerator]



ML Effectiveness

Automate evaluation of appropriateness of use

mlflow

- ML statistical measures
- Built-in and custom metrics
- Extensions for bias checking
- LLM evaluation metrics
- LLM-as-a-Judge for RAG responses

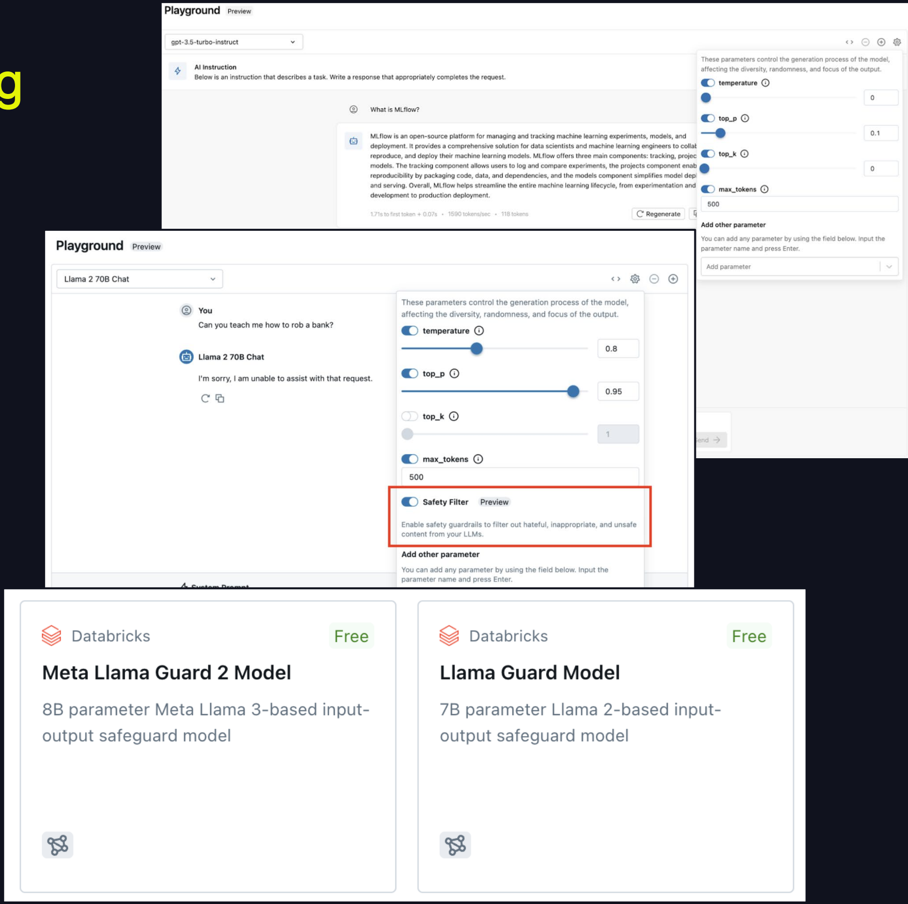
```
with mlflow.start_run(run_name='keras'):
  # log model and datasource
  mlflow.keras.autolog()
  mlflow.spark.autolog()

  sig = infer_signature(X_train, y_train)

  mlflow.shap.log_explanation(model, X_train[:100])
```

```
from mlflow.metrics.genai.metric_definitions import answer_relevance

answer_relevance_metric = answer_relevance(model="endpoints:/gpt-4")

results = mlflow.evaluate(
  model,
  eval_df,
  model_type="question-answering",
  evaluators="default",
  predictions="result",
  extra_metrics=[answer_relevance_metric, mlflow.metrics.latency()],
  evaluator_config={
    "col_mapping": {
      "inputs": "questions",
      "context": "source_documents",
    }
  }
)
print(results.metrics)


results.tables["eval_results_table"]
```

LLM Effectiveness

AI Playground: Selecting and Safeguarding Generative AI

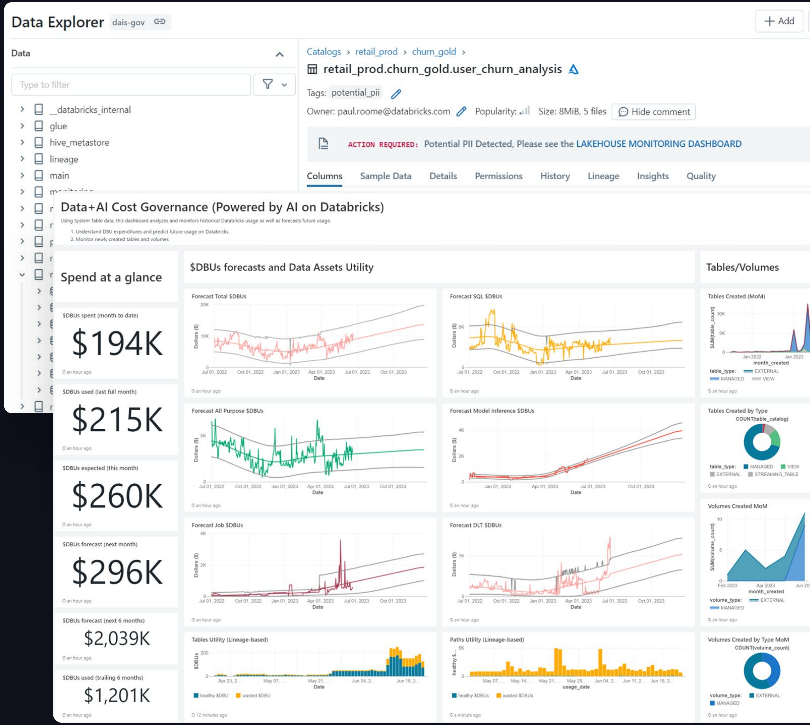
- Test & compare model responses
- Add filters to foundation models with AI Guardrails
- Further enhance LLM safety with Marketplace-hosted models



Tracking Model Reliability

Lakehouse Monitoring: AI-powered monitoring and observability

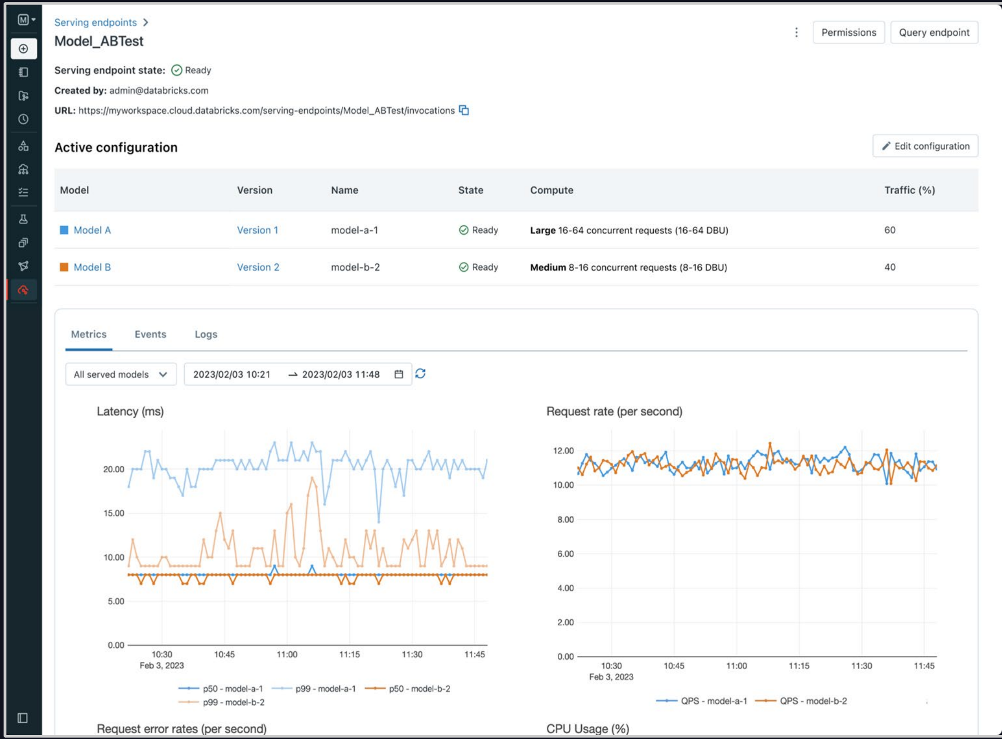
- Auto-Generated, auto-updated, customizable dashboards
- Proactive alerts for quality issues including model drift and degradation
- Monitor fairness, bias, and other measures of appropriateness



Keeping Models Reliable

Model Serving: Iterate without disruption

- Stable model endpoints
- A/B testing or canary deployments
- Automatic version tracking



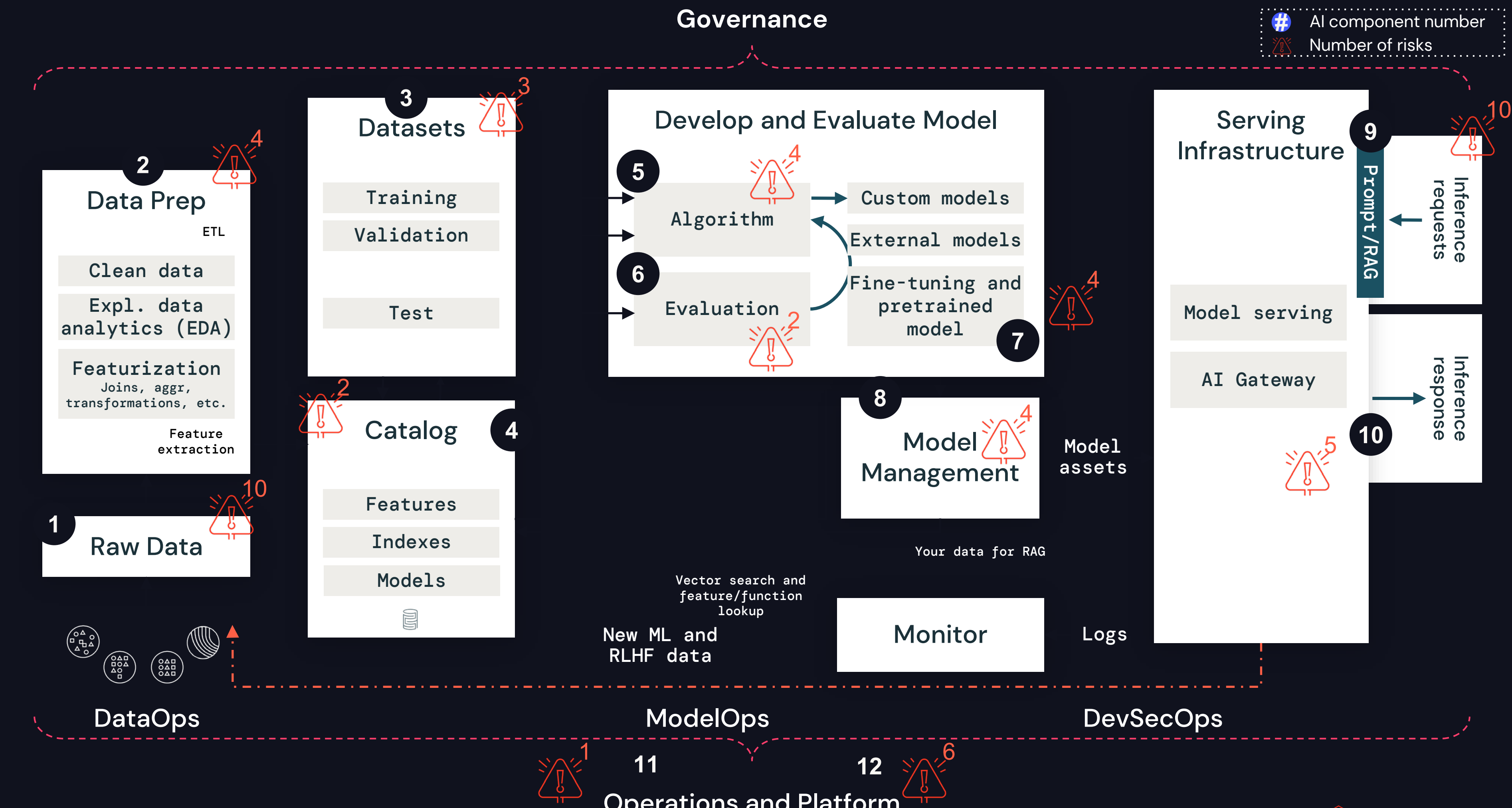
AI Security

How do we secure traditional tech?

1. **Tech:** understand the components and data flows within the system.
2. **People & Process:** define clear roles and establish a structured operating model.
3. **Risks (all):** identify and understand potential harms that AI can cause.
4. **Architecture:** be proficient in various deployment models and understand their associated risks.
5. **Threats:** consider known classes of threats.
6. **Risks (contextual):** conduct risk analysis for specific use cases to identify risks worth mitigating.
7. **Controls:** understand where to implement controls that effectively mitigate risks.

Why is it hard to secure AI?

1. **Tech:** missing a mental model of complete AI components.
2. **People & Process:** unclear roles and operating model.
3. **AI Risks (all):** lack of a comprehensive AI risks catalog.
4. **Architecture:** unaware of security implications of various AI deployment models.
5. **Threats:** unclear which AI threats to be concerned with.
6. **AI Risks (contextual):** unsure which particular risks to focus on mitigating.
7. **Controls:** unsure which controls to apply and where to apply them.


Governance
AI component number
Number of risks
Raw Data
1
10
Data Prep
2
4
ETL
Clean data
Expl. data analytics (EDA)
Featurization
Joins, aggr, transformations, etc.
Feature extraction
Datasets
3
3
Training
Validation
Test
Catalog
4
2
Features
Indexes
Models
Develop and Evaluate Model
5
Algorithm
4
6
Evaluation
2
Custom models
External models
Fine-tuning and pretrained model
7
4
Model Management
8
4
Model assets
Your data for RAG
Vector search and feature/function lookup
New ML and RLHF data
Monitor
Logs
Serving Infrastructure
9
Prompt/RAG
10
Inference requests
Model serving
AI Gateway
5
10
Inference response
DataOps
ModelOps
DevSecOps
11
1
12
6
Operations and Platform

55 risks across 12 components of AI

Raw data
1.1 Insufficient access controls
1.2 Missing data classification
1.3 Poor data quality
1.4 In effective storage and encryption
1.5 Lack of data versioning
1.6 Insufficient data lineage
1.7 Lack of data trustworthiness
1.8 Data legal
1.9 Stale data
1.10 Lack of data access logs

Algorithms
5.1 Lack of experiment tracking and reproducibility
5.2 Model drift
5.3 Hyperparameters stealing
5.4 Malicious Libraries

Data Prep
2.1 Preprocessing integrity
2.2 Feature manipulation
2.3 Raw data criteria
2.4 Adversarial partitions

Datasets
3.1 Data poisoning
3.2 Ineffective storage and encryption
3.3 Label flipping

Evaluation
6.1 Evaluation data poisoning
6.2 Insufficient evaluation data

Model
7.1 Backdoor machine learning / trojaned model
7.2 Model assets leak
7.3 ML supply chain vulnerabilities
7.4 Source code control attack

Governance
4.1 Lack of asset transparency and traceability
4.2 Lack of end-to-end ML lifecycle

Model Management
8.1 Model attribution
8.2 Model theft
8.3 Model lifecycle without HITL
8.4 Model inversion

Model Serving – Inf response
10.1 Lack of audit and monitoring inference quality
10.2 Output manipulation
10.3 Discover ML model ontology
10.4 Discover ML model family
10.5 Black box attacks

Operations
11.1 Lack of MLOps repeatable enforced standards

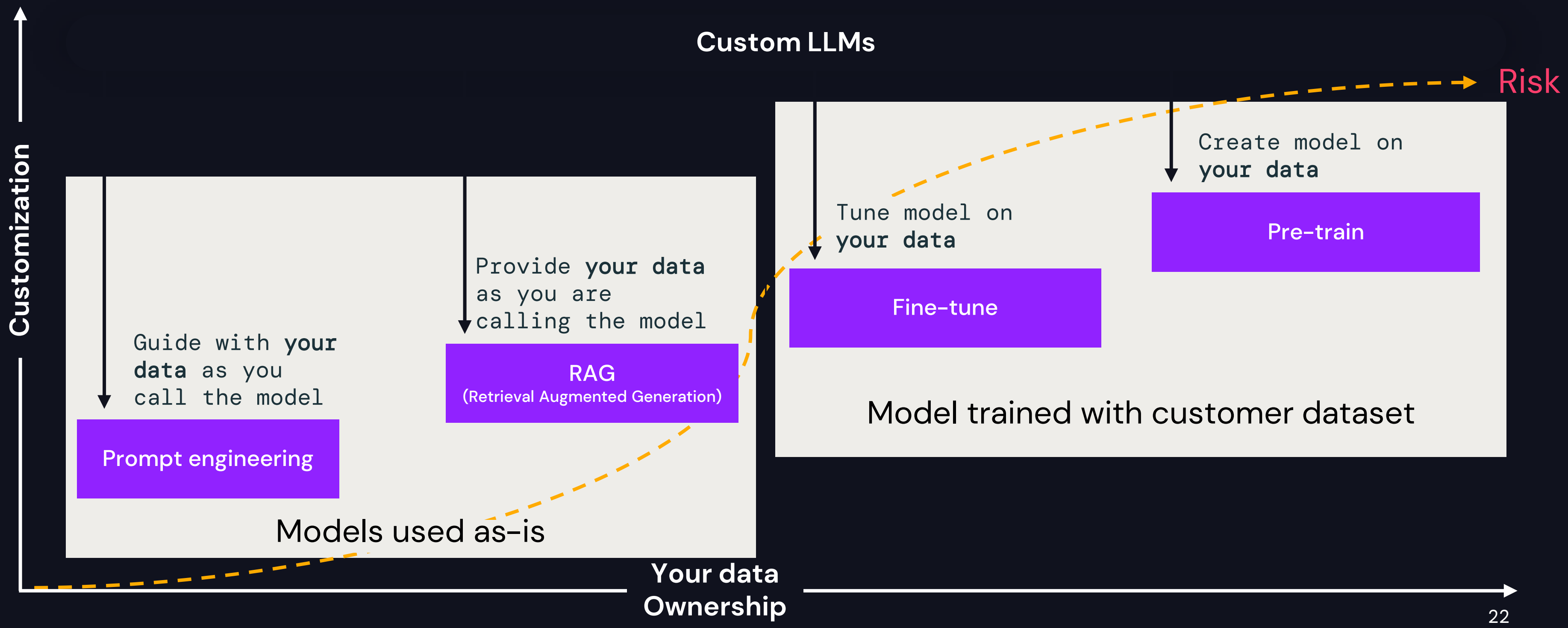
Model Serving – Inf requests
9.1 Prompt inject
9.2 Model inversion
9.3 Model breakout
9.4 Looped input
9.5 Infer training data membership
9.6 Discover ML Model Ontology
9.7 Denial of Service
9.8 LLM hallucinations
9.9 Input Resource Control
9.10 Accidental data exposure

Platform
12.1 Lack of vulnerability management
12.2 Lack of penetration testing and bug bounty
12.3 Lack of Incident response
12.4 Unauthorized privileged access
12.5 Poor SDLC
12.6 Lack of compliance

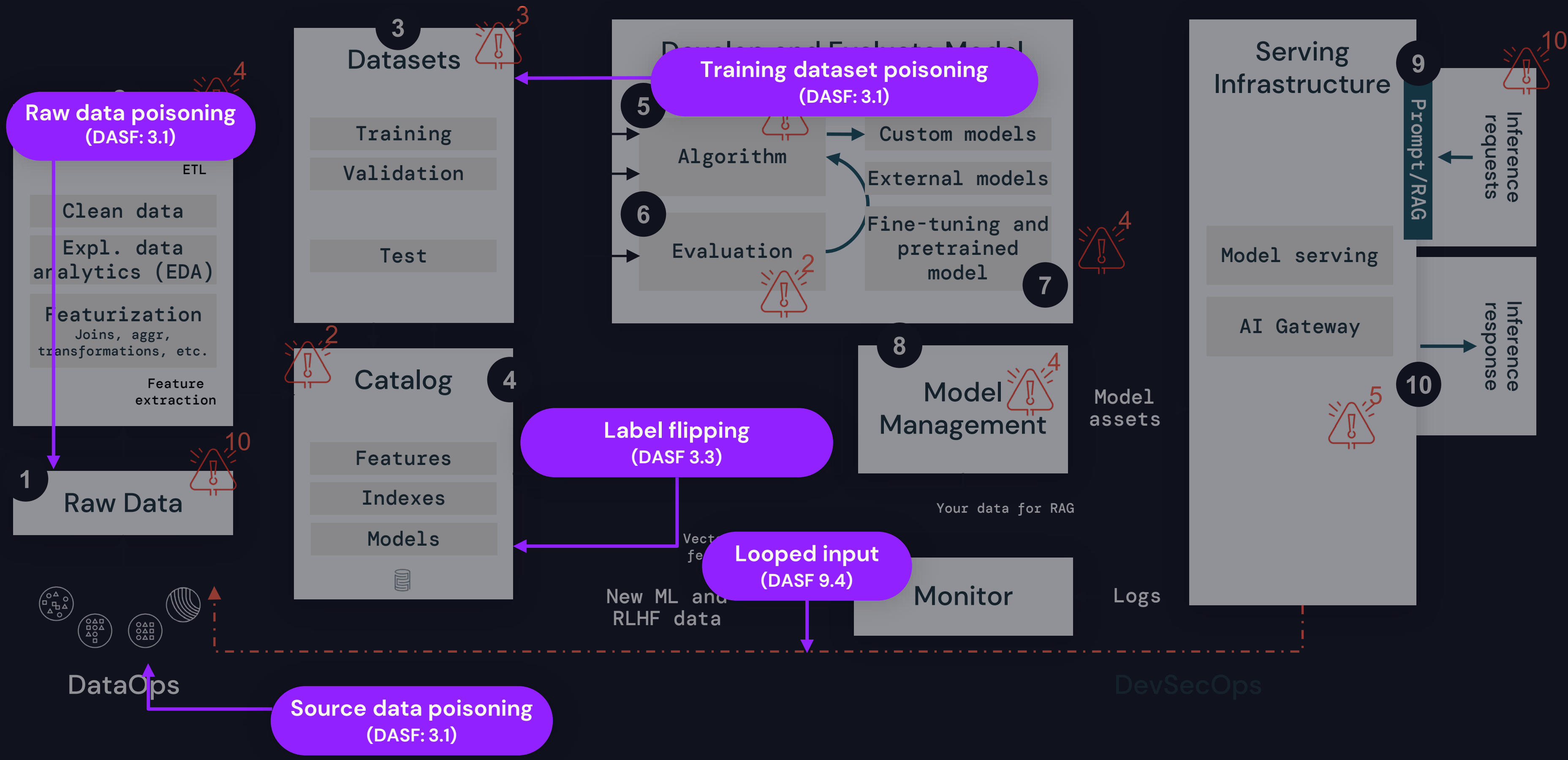
Green = Novel Risk
White = Traditional Risk

LLM deployments models

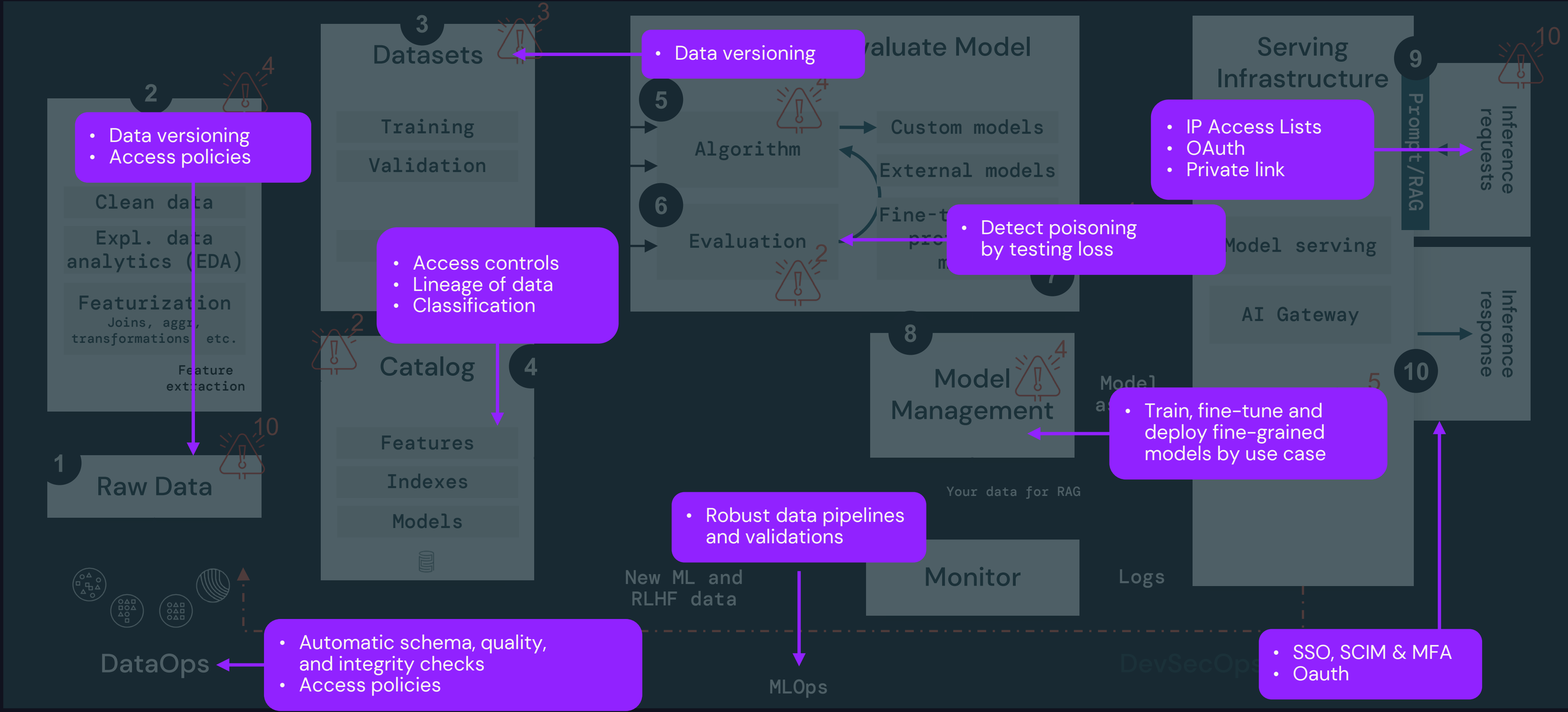
The more you customize models with your data, the more security you need.



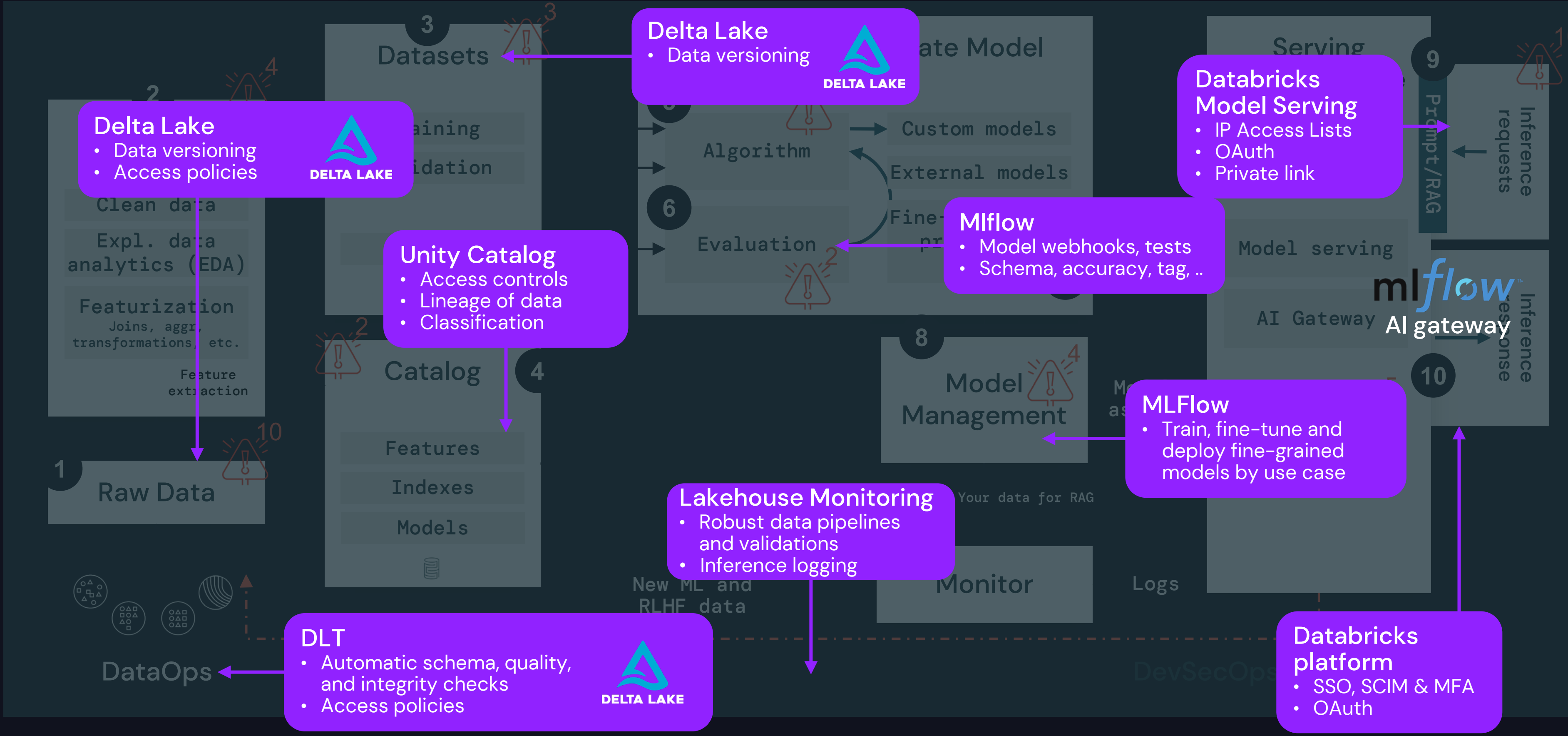
Ex.: Training Data Poisoning: *threats*



Ex.: Training data poisoning: *mitigating controls*



Ex.: Training data poisoning: Databricks controls

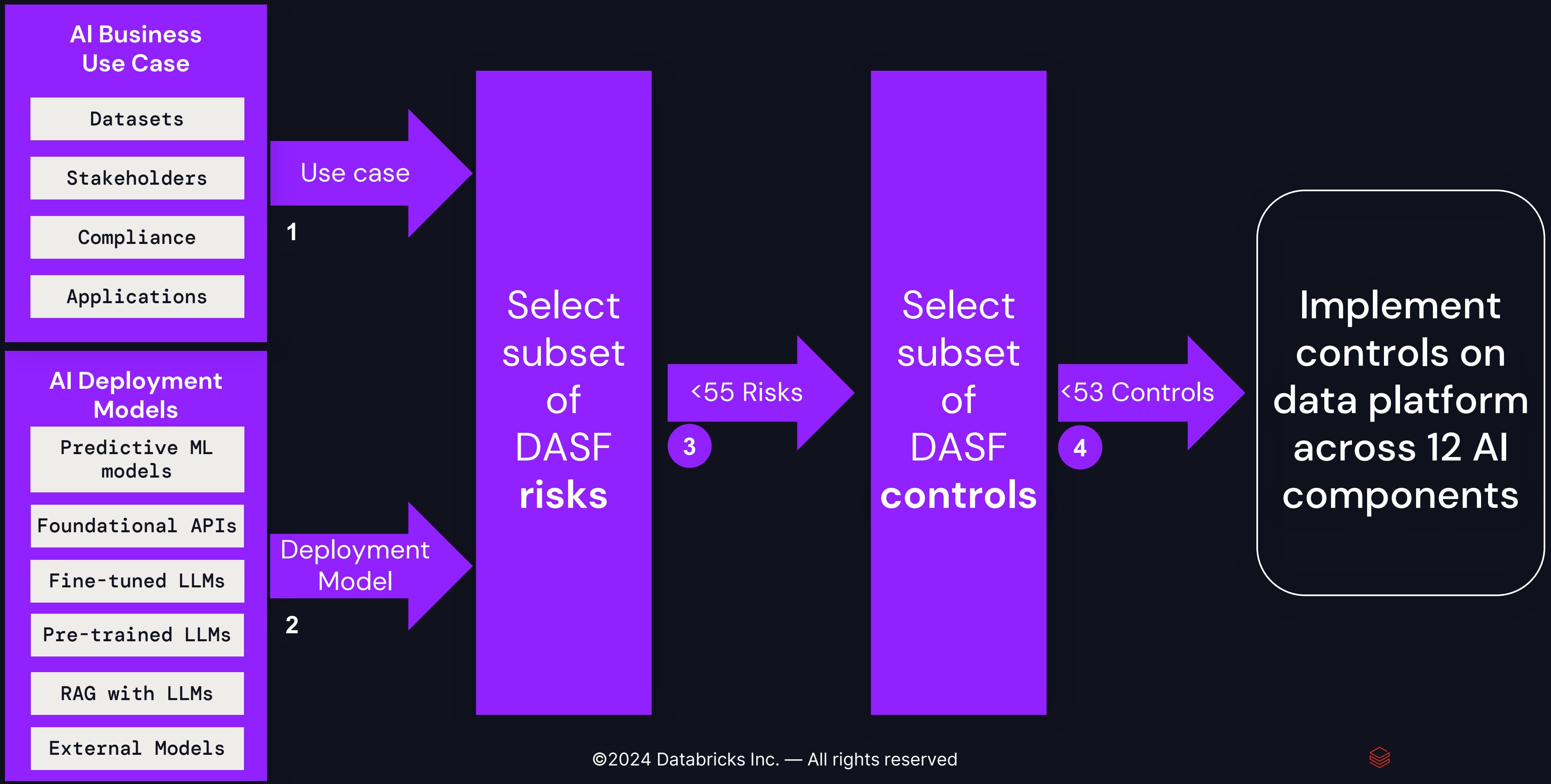


Top 10 controls for mitigating AI risks

AI Novelty

Controls	Data poisoning	Prompt injection	Model theft	Trojaned model	Trustworthiness
Authentication and authorization	●	◐	●	◕	◐
Data and model encryption	●	○	●	○	●
Data governance	●	○	○	○	●
Model governance	○	◐	◐	◕	●
Secure MLOps	●	◐	◐	◕	●
Testing and detect loss after (re)training	◐	○	◕	●	●
Securely serve models	○	●	●	○	◐
Zero Trust/Model Segregation	○	○	◐	●	●
Secure with Model Gateway	○	●	●	◐	◐
Audit & monitor	●	●	●	◐	◐

Databricks AI Security Framework (DASF)



AI Governance

Challenges to Governing AI

Control

Disjoint tools for access management to data & AI

Increased data breach risk, operational expenses

Privacy

Inconsistent classification and protection of data

Risk of data leaks for both PII and IP

Audit

Incomplete insight into access & usage

Non-compliance risk, reputational harm

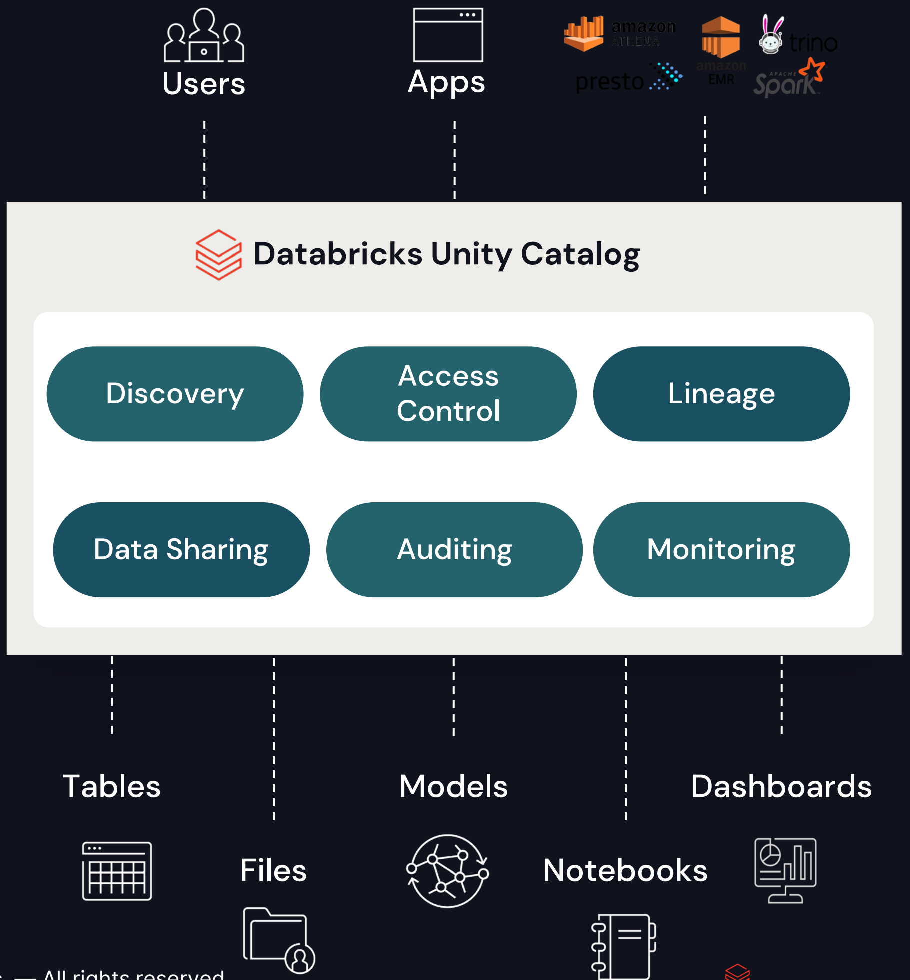
Databricks Unity Catalog

Unified visibility into data and AI

Simple permission model for data and AI

AI-powered monitoring and observability

Open data sharing



Unified Governance

Selected Data & AI features in Unity Catalog

Controls

Single plane of **fine grained access** across:

- AI Features
- AI Models
- Tables
- Filesystems

+

Privacy

Default **privacy preservation:**

- Column masks
- Row filters
- Data obfuscation
- Data tokenization
- Classification
- Attribute based policies

+

Audit

Single plane of **audit** across **data** and **AI:**

- Usage
- Discovery
- State of entitlement
- Lineage of data

→

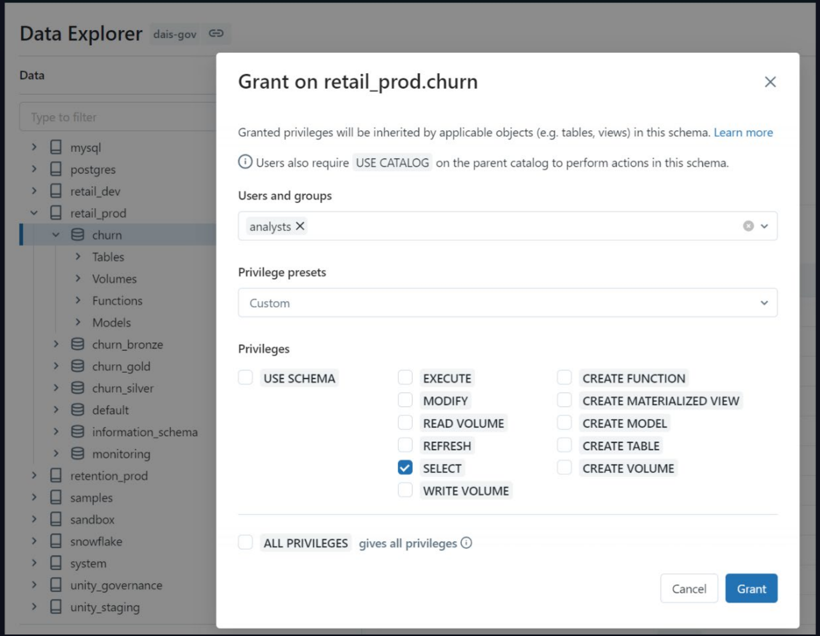
Compliance

- Data Science teams have access to requisite data only
- PII data cannot be used to train models
- Compliance team understands data used to train AI
- Audit/Governance team able to audit access and usage in real time

Control Access

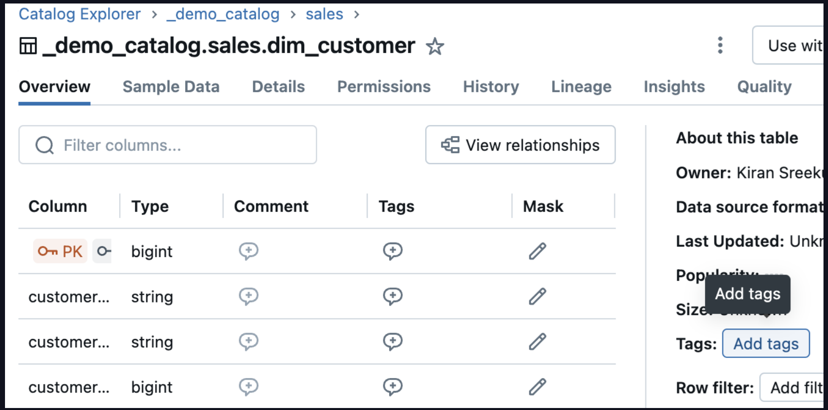
For all data & AI assets

- Unified interface for managing and auditing access policies
- Fine-grained access controls
- Open interfaces with consistent permissions



Fine-Tune Privacy

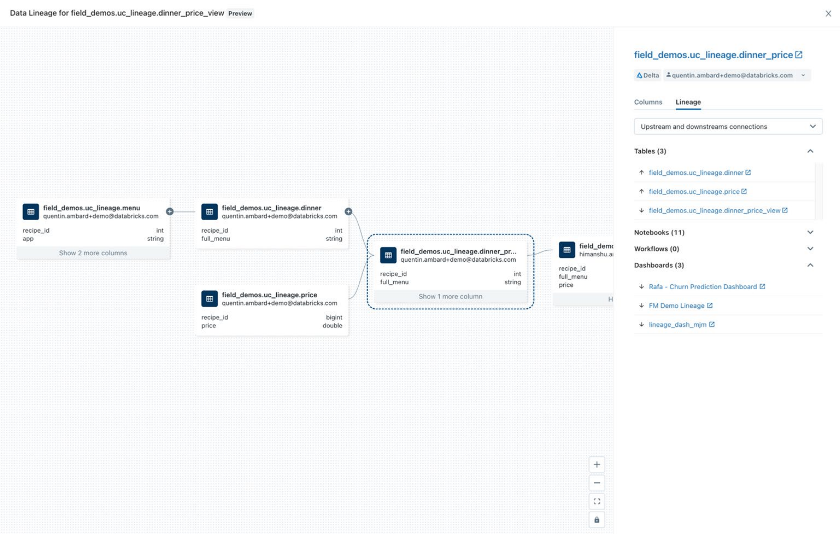
- Classify data & AI with tags (attributes)
- Automate row filters to return only allowed subsets
- Apply masking, obfuscation, and tokenization to refine visibility



In-Depth lineage for all workloads

End-to-end visibility into data use

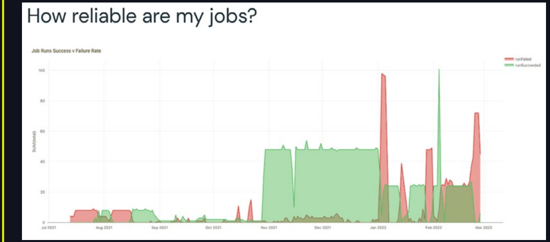
- Auto-capture runtime data lineage
- Track lineage down to the table and column level
- Lineage across tables, dashboards, workflows, notebooks, feature tables, files, and DLT

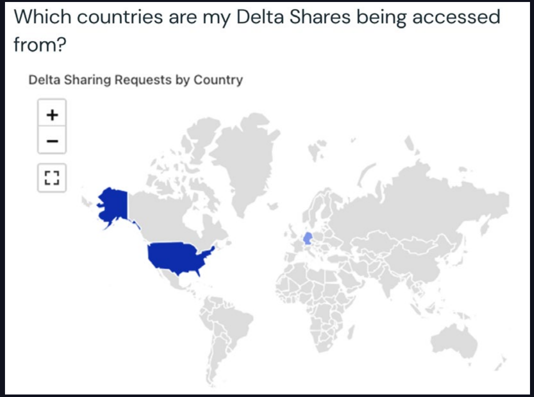


Automatic Audit Logging

Easy observability into user activities

- Comprehensive log of activities
- Monitor detailed usage patterns
- Open interface to other audit tooling
- Analyze audits logs using Databricks
- Configure dashboards and alerts in Lakehouse Monitoring





File System
Clusters
Accounts
Jobs
Notebook
SSH
Workspace
Secrets
SQLPermissions
Instance Pools
SQL Analytics
Genie
Global Init Scripts
IAM Role
MLFlow Experiment
Marketplace
Feature Store
Remote History Service
MLFlow Acled Artifact
DatabricksSQL
Delta Pipelines
Model Registry
Repos
Unity Catalog
Git Credentials
Web Terminal

Further Resources

Implementing LLM Guardrails for Safe and Responsible Generative AI Deployment on Databricks

Mitigating Bias in Machine Learning With SHAP and Fairlearn

The Shift from Models to Compound AI Systems

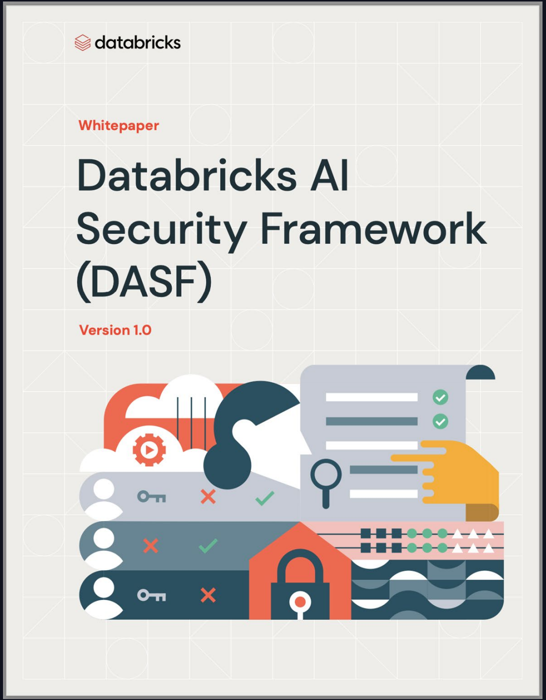
Lakehouse Monitoring: A Unified Solution for Quality of Data and AI

Databricks' Approach to Responsible AI - how we built DBRX

Introducing the Databricks AI Security Framework!

- Securing AI will become easier as we better understand AI
- Each AI use case may have a distinct risk profile
- Be prepared to be wrong... adapt your process
- Adopt an open framework to hasten AI security, e.g.: DASF

How to get it?



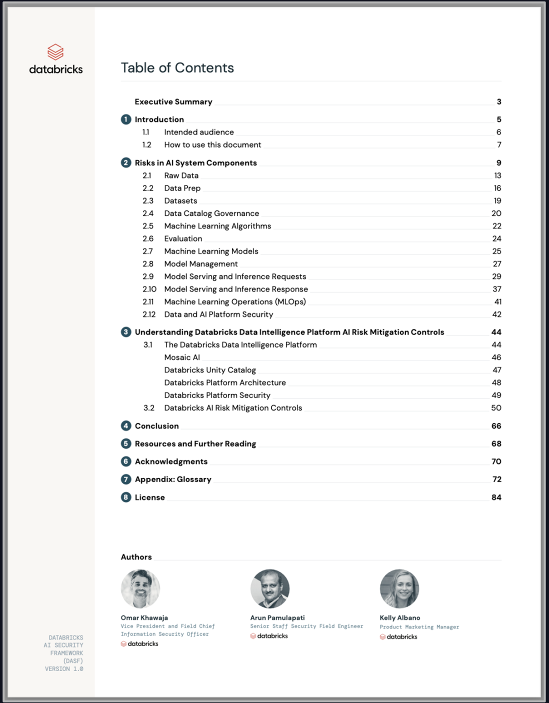

databricks

Table of Contents

Authors

Omar Khawaja
Vice President and Field Chief Information Security Officer
databricks

Arun Pamulapati
Senior Staff Security Field Engineer
databricks

Kelly Albano
Product Marketing Manager
databricks

DATABRICKS AI SECURITY FRAMEWORK (DASF) VERSION 1.0

DATA+AI SUMMIT