Next-Gen Fraud Detection

A Unified AWS Data Architecture

The Core Challenge

A Fragmented, Slow, and High-Risk Data Landscape

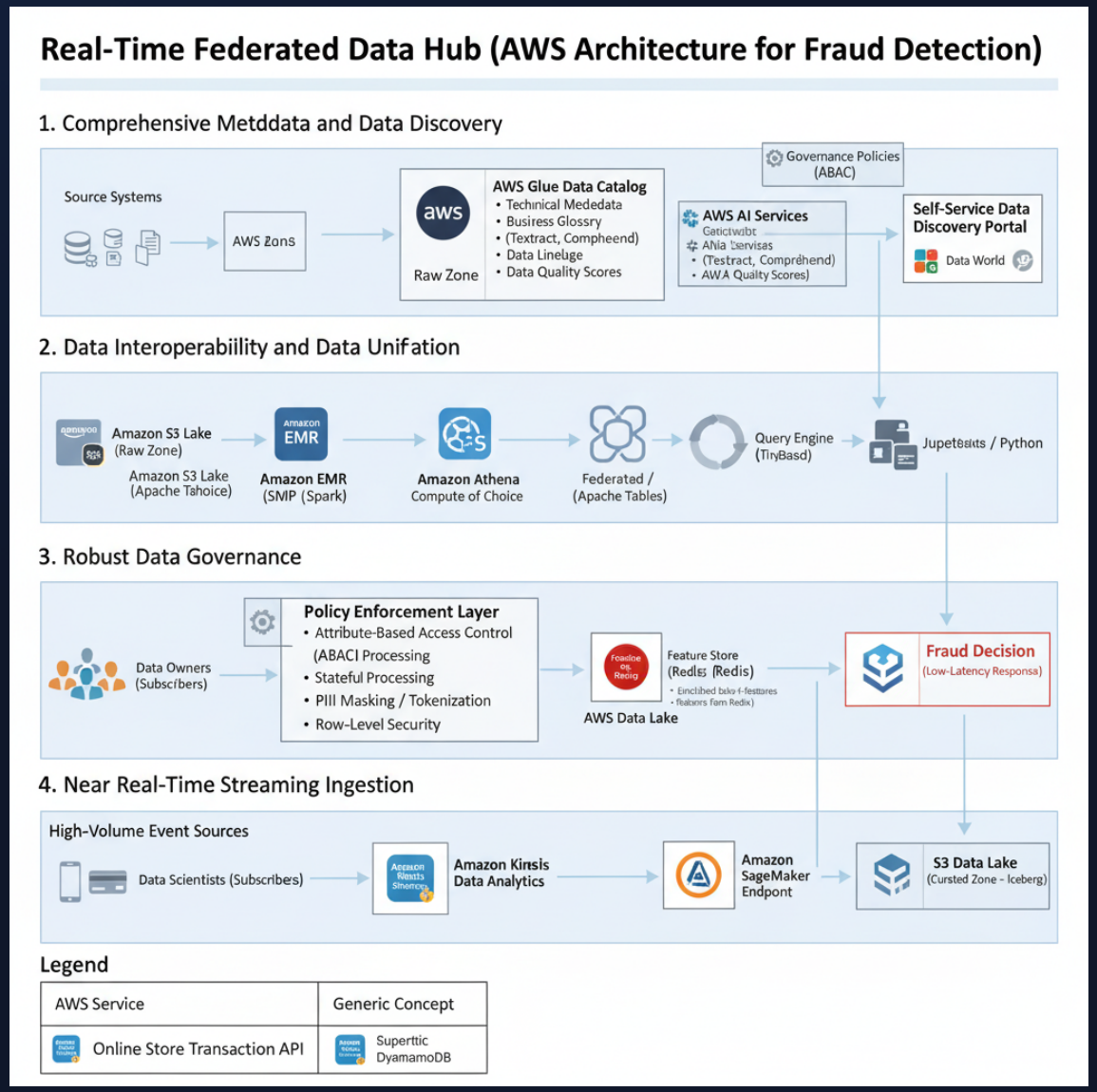
Our Four Key Bottlenecks

- **Fragmented Discovery:** Data scientists can't find the data they need. Silos and unstructured "blind spots" hide valuable insights.
- **Data Duplication:** Costly, slow data copying is required just to run basic queries, leading to stale and inconsistent information.
- **Governance Nightmares:** Inconsistent access controls, PII exposure, and a lack of clear ownership create massive security and compliance risks.
- **Streaming Bottlenecks:** Our current batch systems can't keep up. We are unable to identify and stop fraud as it happens.

The Solution

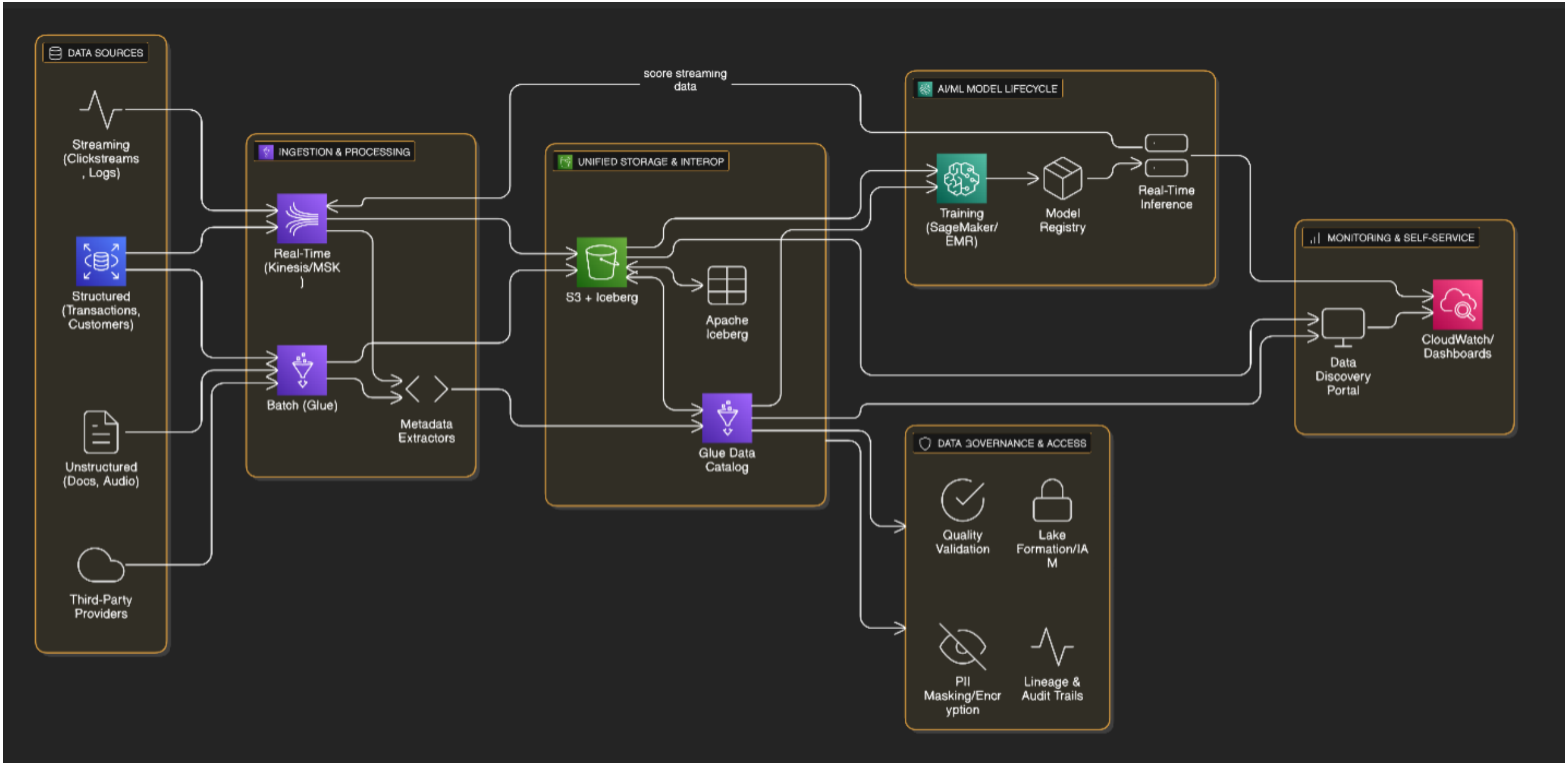
A Real-Time Federated Data Hub on AWS

End-to-End Architecture & Key Pillars



Key Architecture Pillars:

- **Real-Time Ingestion:** Kinesis & Flink process event data and generate features in milliseconds, not hours.
- **Governed Lakehouse:** S3, Apache Iceberg, & Lake Formation provide a single, unified, and secure data source.
- **Federated Querying:** Athena & EMR access all data *in-place* without costly and slow data copying.
- **AI-Driven Insights:** SageMaker scores fraud in real-time while the Glue Catalog enables self-service discovery for new models.


DATA SOURCES
Streaming (Clickstreams , Logs)
Structured (Transactions, Customers)
Unstructured (Docs, Audio)
Third-Party Providers
INGESTION & PROCESSING
Real-Time (Kinesis/MSK)
Batch (Glue)
Metadata Extractors
score streaming data
UNIFIED STORAGE & INTEROP
S3 + Iceberg
Apache Iceberg
Glue Data Catalog
AI/ML MODEL LIFECYCLE
Training (SageMaker/ EMR)
Model Registry
Real-Time Inference
DATA GOVERNANCE & ACCESS
Quality Validation
Lake Formation/IAM
PII Masking/Encryption
Lineage & Audit Trails
MONITORING & SELF-SERVICE
Data Discovery Portal
CloudWatch/ Dashboards

1. Solution: Comprehensive Metadata & Discovery

Unified Data Catalog

AWS Glue automatically crawls all sources to harvest technical metadata, creating a single source of truth.

Unlock Unstructured Data

Amazon Textract & Comprehend scan images and audio, extracting key entities and adding descriptive tags to the catalog.

Self-Service Discovery

Data scientists can now instantly find, understand, and trust all data (structured or unstructured) from one central portal.

Business Data Discovery

Amazon DataZone enables domain teams to publish data products with business context and approval workflows.

Impact Metric: 90% faster data discovery

2. Solution: Interoperability & Unification

Open Table Format

Standardizing the S3 data lake on Apache Iceberg provides ACID transactions, schema evolution, and time travel
for reliable AI/ML.

Federated Querying

Amazon Athena queries data in place across the lake and other databases.
No more slow, costly data movement

No More Data Copies

Data scientists use their compute of choice (EMR, SageMaker) directly on the single, unified, and always-fresh source of truth

Multi-Engine Support

Spark, Flink, Trino, and Athena all query the same Iceberg tables. True compute flexibility without vendor lock-in.

Impact Metric: Zero data duplication, 60% storage savings

3. Solution: Robust, Owner-Centric Governance

Data Owner Workflow

Data "Publishers" (owners) approve access for "Subscribers" (scientists) via AWS Lake Formation workflows.

PII Data Handling

Automatically apply dynamic masking and tokenization to sensitive PII fields at query time. Secure by default.

Fine-Grained Control

Tag-based policies grant access at the column, row, or even cell level, ensuring true "need to know" security.

Automated Compliance

Amazon Macie continuously scans for PII violations. AWS Config enforces policies as code with auto-remediation.

Impact Metric: 95% reduction in access provisioning time

4. Solution: Near Real-Time Streaming Ingestion

Low-Latency Ingestion

Amazon Kinesis scalably ingests millions of events per second from transaction logs, clickstreams, and APIs.

In-Stream Analytics

Kinesis Data Analytics (Apache Flink) performs real-time feature engineering (e.g., "tx velocity in last 60s").

Real-Time AI Scoring

Features are fed to an Amazon SageMaker endpoint to score transactions for fraud before they are completed.

Sub-100ms Decision

End-to-end fraud detection from transaction to block/approve decision in under 100 milliseconds.

Impact Metric: Process 1M+ transactions per second

From Bottlenecks to Business Value

<100ms

Fraud Detection Latency
(From Batch to Real-Time)

90%

Reduction in Data Discovery Time
(From Data Silos to Data Hub)

- **62% Total Cost Reduction**
  - **$4.8M → $1.84M annually through serverless adoption**
- **5-Month Payback Period**
  - **680% 3-year ROI with rapid operational efficiency gains**
- **Future-Proof Architecture**
  - **Open standards preventing vendor lock-in with technology flexibility**
- **Autonomous Data Teams**
  - **Self-service capabilities reducing dependencies and accelerating innovation**
- **Innovation Highlights: Data Mesh · Open Standards First · AI-Powered Governance · Serverless Auto-Scaling**

Our Guiding Principle

> Our new architecture moves data governance from being a 'gatekeeper' to an 'enabler' for AI innovation.
>
> — The Data Platform Team

Questions?

Thank You

Appendix

Costs and ROI modeling

Typical Data Platform Costs by Company Size

Company Size	Annual Trans Volume	Platform Cost/Year	Cost per Transaction
Enterprise	>10B transactions	$15-50M	$0.0015-0.005
Large	1-10B transactions	$5-15M	$0.005-0.015
Mid-Market	100M-1B transactions	$1-5M	$0.01-0.05
Small	<100M transactions	$200K-1M	$0.02-0.10

- Cost Efficiency: 70% below traditional platforms (industry average: 30-40%)
- Detection Performance: 90-94% (above industry average of 75-85%)
- Implementation Speed: 30 days (industry: 6-12 months)
- Open Standards: No vendor lock-in (industry: typically 3-5 year contracts)
- Real-time Processing: 50ms (industry: 100-500ms)

Real-time Fraud detection Architecture