DATA+AI SUMMIT 2022

Scaling Privacy

Practical Architectures and Experiences

Aaron Colcord
Sr. Director, Center of Excellence, Privacera

Mei Gui
Software Engineer, Databricks

About us

Aaron Colcord

Privacera

Aaron is an Adaptive technical leader with 20+ years' experience in spearheading enterprise data solutions and enabling scalable, secure processes which lead to powerful insights from complex data systems.

He has spent the last couple of years working passionately inside evolving technologies such as Lakehouse, Data Mesh, the ever-evolving modern data stack, and acquiring 7 patents in this area.

He joined Privacera because of a belief in the mission and technology to advance customers and their data management programs.

Mei Gui

Databricks

Mei is a software engineer working on DBFS Fuse and files in repos at Databricks. Mei is the lead engineer for integrating Privacera and DBFS Fuse on Databricks. Prior to Databricks, Mei was a software engineer at Uber.

“If we accept as normal and unavoidable that everything in our lives can be aggregated and sold, then we lose so much more than data. We lose the freedom to be human.”

Tim Cook
CEO, Apple

Review from last year's presentation

The Main Points

Have you ever…

Gone to a website and read their privacy policy, clicked accept cookies, accepted terms of service, or EULA?

Collecting information about your customers can

- Improve the experience.
- Allow the company to understand their business better.

At the core, privacy is a policy and legal obligation

- You have the data. It used to be your business to just secure it.
- Do you want your information monetized? Sold? Traded?
- Most companies don't do this. But the privacy policy is there for you.

Clicked 'accept all' on website, used a digital assistant..

Review from last year's presentation

The Main Points

What is the difference between these?

Security

- Preventing unauthorized usage of systems
- Ensuring users don't see the incorrect information
- Creating boundaries to enforce the right action in the system

Compliance

- The process of making sure your company and employees follow all laws, regulations, standards, and ethical practices that apply to your organization

Privacy

- Data privacy may be defined as the authorized, fair, and legitimate processing of personal information
- Consent rights
- Do not share
- Slippery space

Review from last year's presentation

The Main Points

Ideal Scalable System

User Rights

- Revocation of consent
- Portability
- Erasure
- Rectification
- How is data used?
- Rights follow data reuse
- Flexible to change

Classification

- Should align with a Data Governance program
- Should adapt to changing data
- Proactive
- Reclassification

Access

- Authorization of user may change
- Supports agile access
- Business use is preserved
- Automated systems obey privacy

Audit/Governance

- How was it used?
- How was it accessed?
- How was it protected?
- Did it cross borders?

Practical Architectures

Project Mercury

Data eco-systems

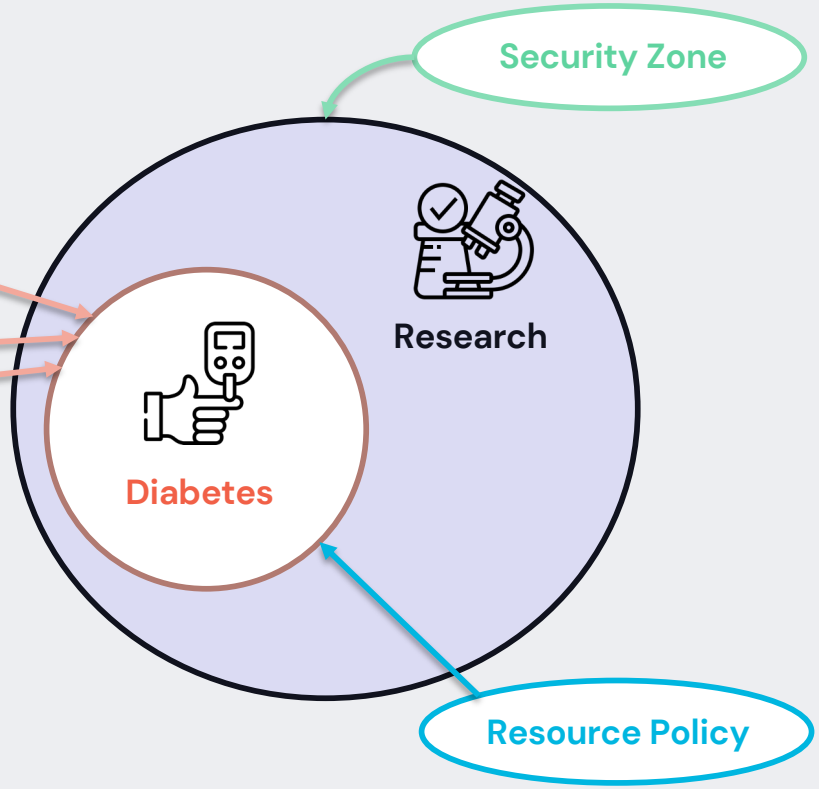
Evolving Architectures

Our story

Being Democratized

We want to build a ML Model

- 2-3 Datasets → Training Set
  - External Purchased
  - Internal Generated
- Machine Learning Model
- Logs for the ML Model
  - Which contain real data – which we might not be able to predict



Privacy Impact Assessment (PIA)

Project by Project Process

Define the Scope

Data Flow Analysis -

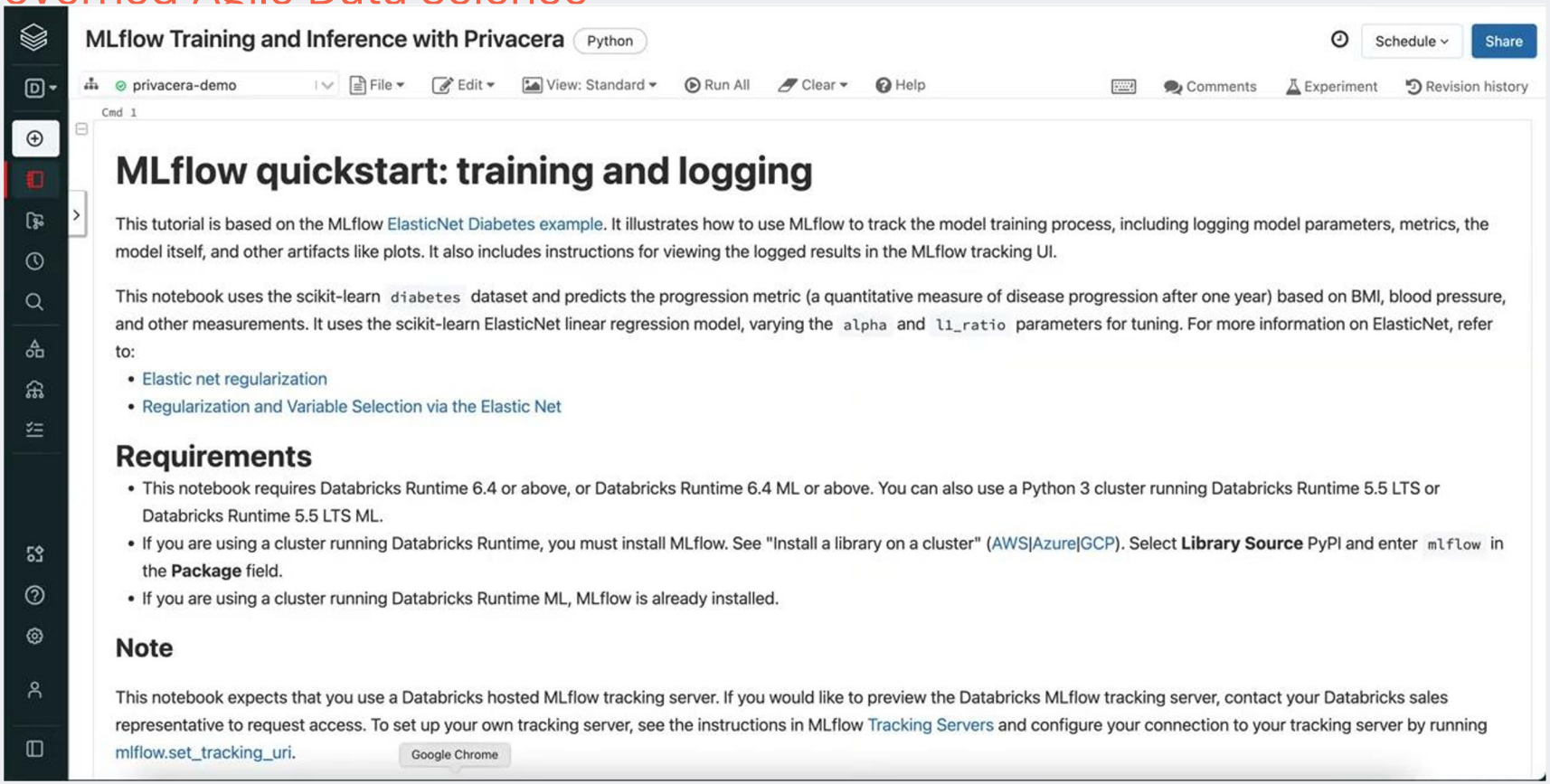
- Show the lineage
- Identify the PII
  - Why is it being collected (Business Reason)

Privacy Analysis - How are we using this data and for what purpose?

Report - Risks and Implications

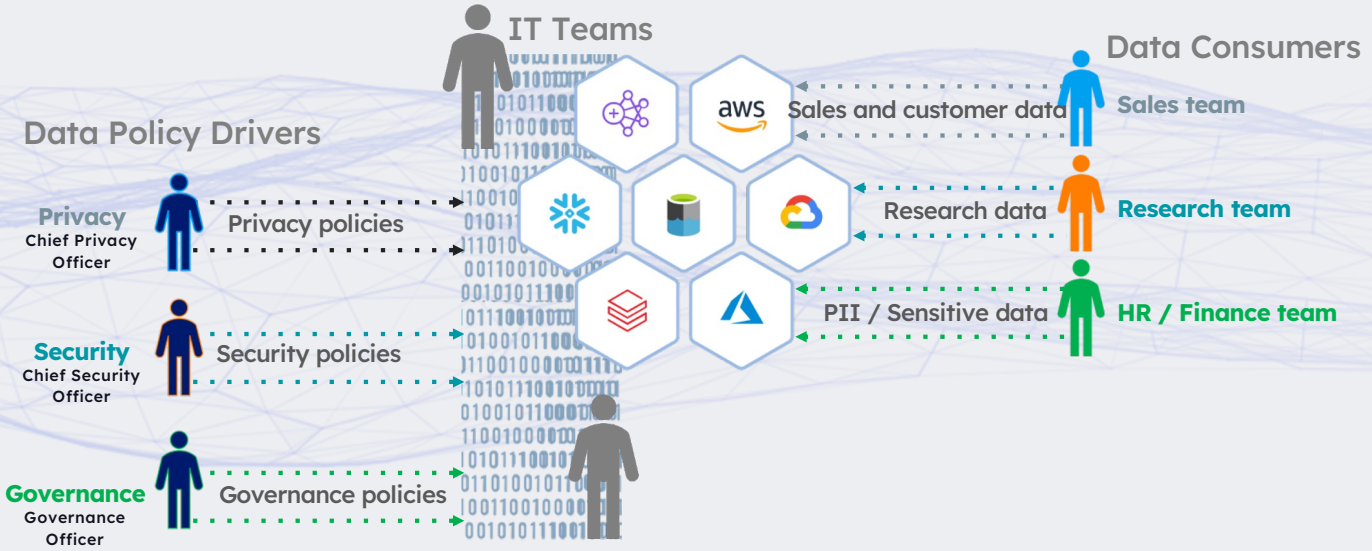
Demo

Governed Agile Data Science



Data Governance

Blink and you missed it.



Did we pass the PIA?

Did we balance Democratized with Governance?

What we saw in action

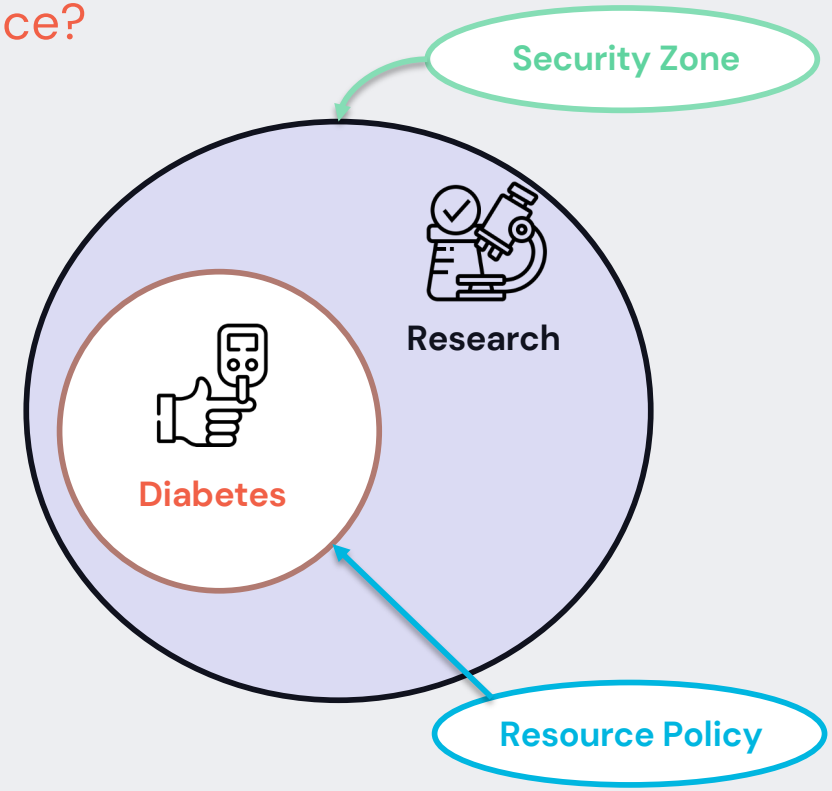
- Fine-Grained Access Control– All Data, Models, Logs were controlled by One Policy
  (Delta, S3, and MlFlow were managed in spot)
- Audit trails showing exactly what was working
- Unstructured Data governed in Development

What we had happening in the background

- Data Movement was being tracked
- What type of Data (PHI) was being used.

Building this represents true Business Value

- Data Lakes power this via 'Borderless Data'
- We 'Scaled Privacy'
- The PIA was a living, breathing enabling element; not a static process requiring high friction coordination.



Practical Experiences

How do you take this and build a program?

Set some principles for architecture

Make your decisions from the principles

- Build around Services, not libraries.
  - Avoid 1 Million Hour Problems
- Implement Policy-Approaches
- Separation of Duties
- Be *Actionable*
- Plan to be Scalable from the start.
- Thread the *Build vs Buy* Equation

Authorization

Authorization is hard. Authentication is not.

Lean into policy-based controls that allow:

- Source access control with the raw data
- Managed control that enforces how data is used
- Rapid updates while the true definition of ‘sensitive information’ is found
- Enablement from the ability to adapt to regulations
- Quality Control to influence

Organizing your Lake, Warehouse, Cloud

Practical Experience

We want to think in terms of the business

- Example: Customer, Field, Product or Customer, Flight, Marketing
- If we align this way, at least our ontology will build the shape

Organize Objects close to these terms and have some way to map

- Losing track of S3 buckets is a problem. Yet it's their flexibility that makes them useful.
- You can't find it later if you didn't plan to find it later.

Tagging Strategy

- Pieces of data may be related to larger subjects
    - Example: Who took the flight is a Customer and stored as a part of Flight.
- We can attach information flexibility about origin, PII, Technical Context, Business Context

How have you managed this?

Engineers vs. the world

All of these processes exist

GSA - A lot of great material

- PIA
- Procedures
- Training Requirements

HIPAA Directives

- How to de-identify
- Safe Harbor
- Expert Determination

"Build it and they will come"

Unify your Stakeholders

- Most of these tools are DIY
- The moment you make the tools, you need adaptability
- Open Source bridges the 'Buy/Build'
- Policy-Approach can solve by connecting Tags to 'Intent to Use'
  - Use Masking Policies to change how data is returned dynamically
  - ABAC

Separation of duties

Don’t let the mouse guard the cheese!

This isn’t that hard.

- It doesn’t always mean people; it can be process.
- Your expert to help defend the organization can’t be the requestor

DATA+AI
SUMMIT 2022

Thank you