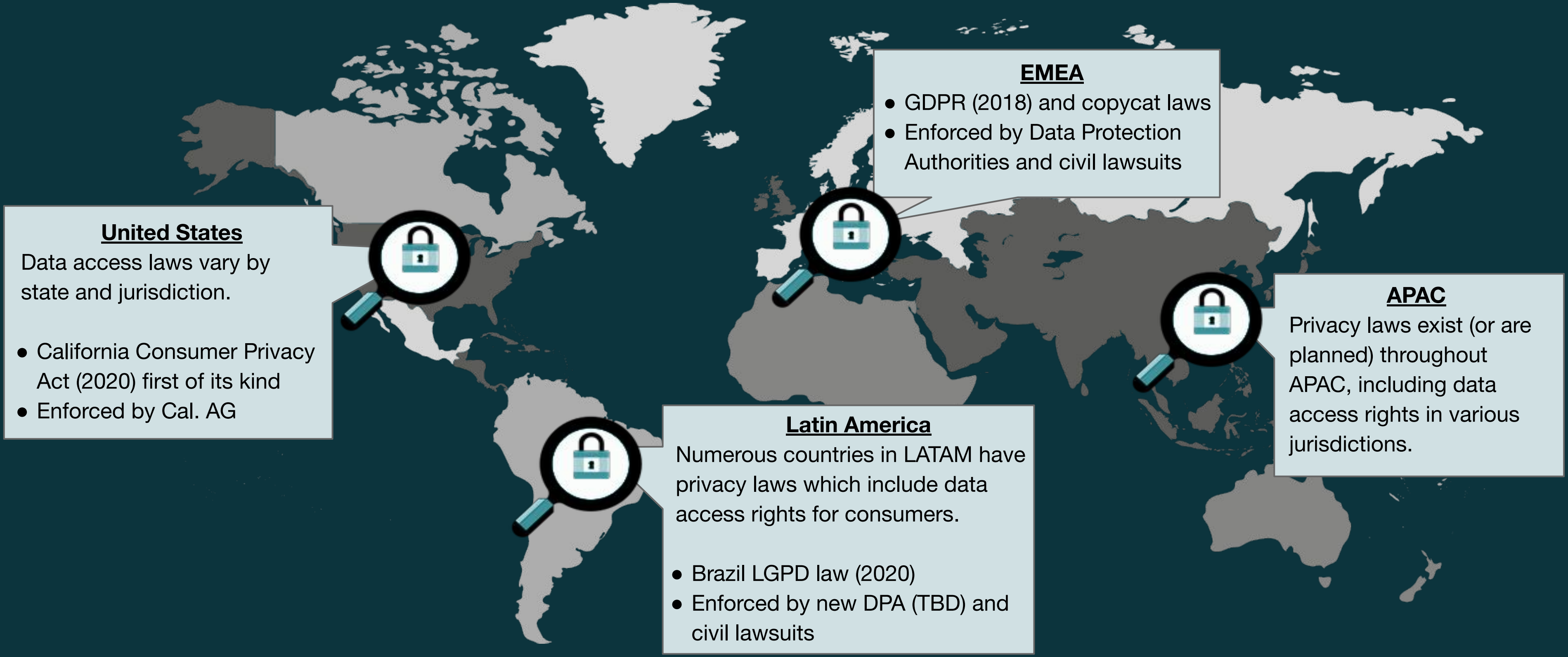
Privacy Engineering: Assurance and Trust

Nishant Bhajaria

Regulatory Landscape

Users in many jurisdictions (and growing) have rights



United States

Data access laws vary by state and jurisdiction.

- California Consumer Privacy Act (2020) first of its kind
- Enforced by Cal. AG

EMEA

- GDPR (2018) and copycat laws
- Enforced by Data Protection Authorities and civil lawsuits

Latin America

Numerous countries in LATAM have privacy laws which include data access rights for consumers.

- Brazil LGPD law (2020)
- Enforced by new DPA (TBD) and civil lawsuits

APAC

Privacy laws exist (or are planned) throughout APAC, including data access rights in various jurisdictions.

Some privacy laws can do more harm than good

Makes the privacy teams look like blockers

AND

Creates the narrative that engineers do not care about privacy.

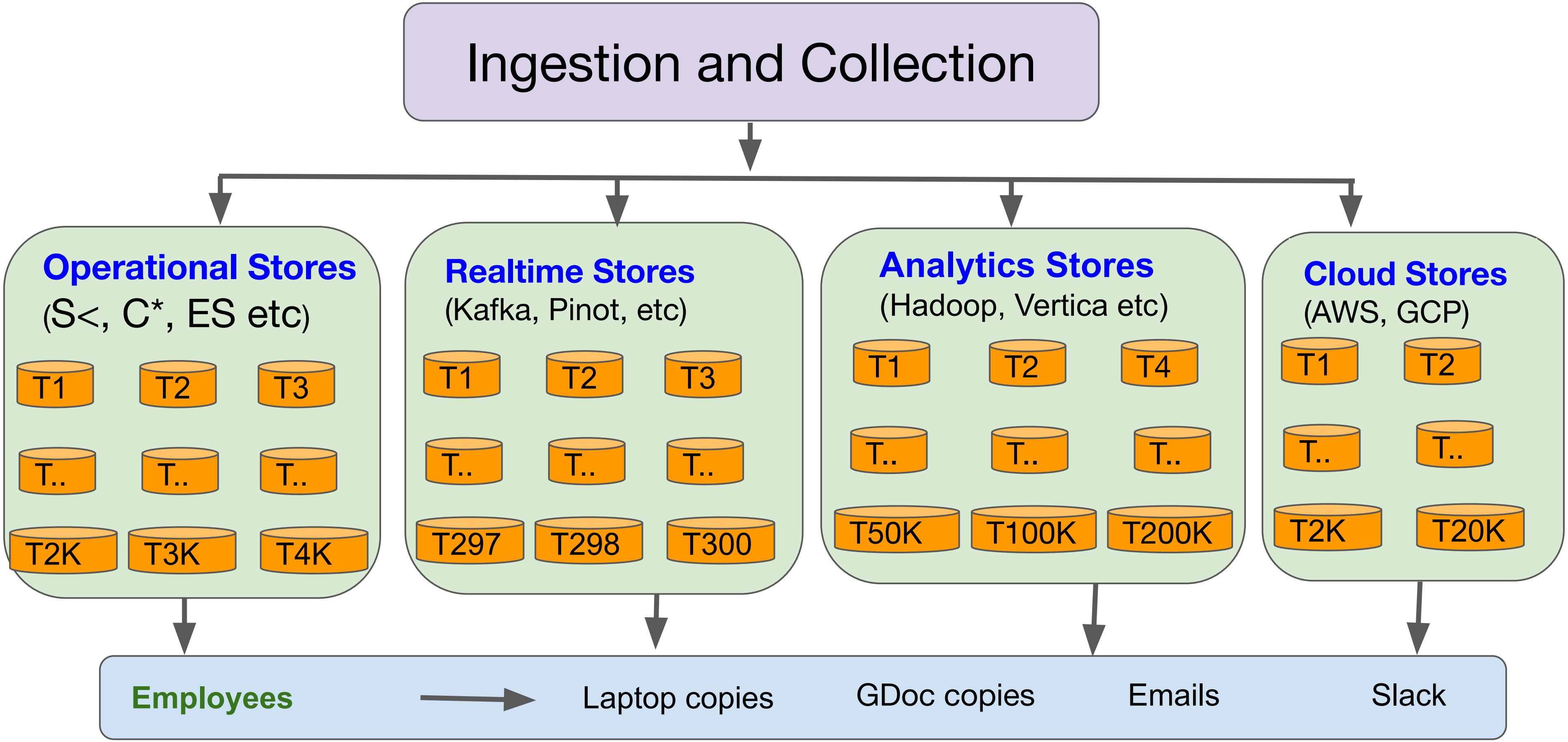
POLITICO

February 24, 2021

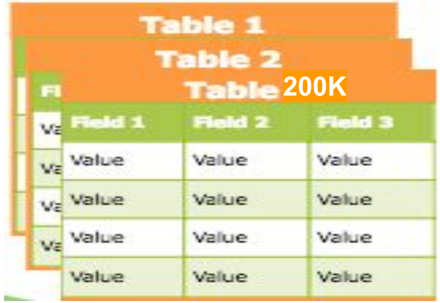
‘Millions of people’s data is at risk’ — Amazon insiders sound alarm over security

“Data is at risk because Amazon has a poor grasp of what data it has, where it is stored and who has access to it.”

Data is (and gets) Everywhere



Data is Stored in Different Data Stores



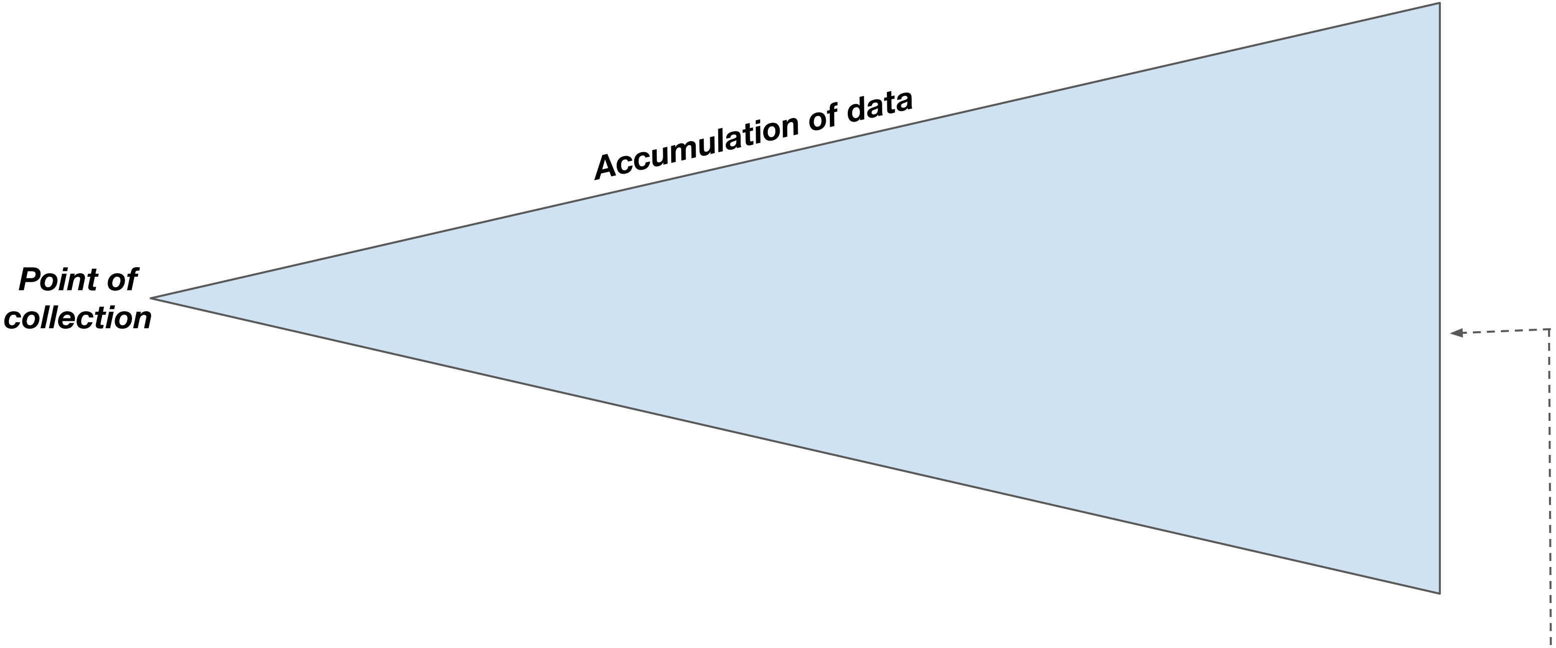
# of Datastores : 10+
Hadoop, Schemaless,
Cassandra, etc.

# of Tables ~ 250K

# of Files : 800M+

Total Volume of Data Stored ~ 800PB

Growing Data Size (through Copies, Inferences & Data Joins)



Most companies classify and inventory data here

- Data size causes delays and inaccuracies
- Attack surface increases and hurts detection

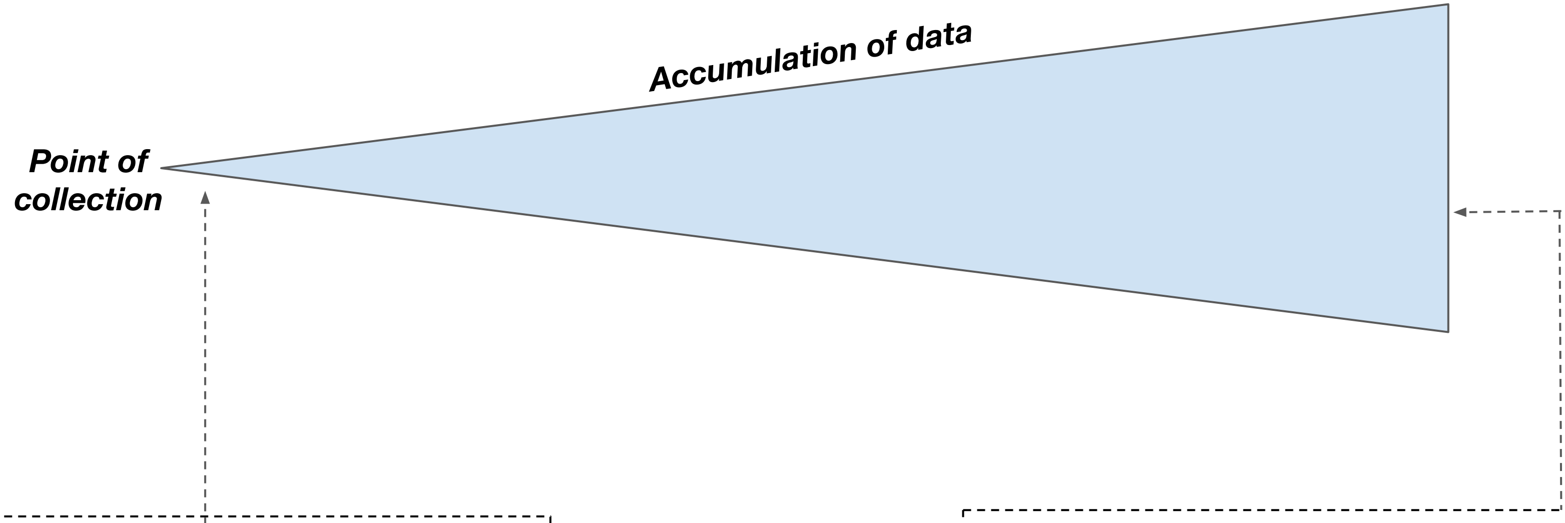
This Makes it Difficult To….

1. Ensure compliance with legal obligations
2. Identify who is collecting data and for what
3. Enforce access control
4. Validate data deletion
5. Secure data, given the vast attack surface

It also hurts the business overall

1. Suboptimal data quality
2. Data storage costs
3. Wasted compute resources

Results of Early Inventory & Data Classification



Inventory and classification of data at point of collection creates downstream efficiencies

Lesser data inventoried and classified down the funnel (easier deletion, better security, improved data quality)

What does “Data Privacy Governance” Mean?

Engineers and Data Custodians will need to own:

1. Ownership and Transparency - Tag data upon collection
2. Data Ingestion Minimization - Collect only what you need
3. Purposeful Usage and Retention - Delete data upon usage completion
4. Data Protection - Manage data access based on privacy risk

Data Governance

Data Governance

“Data Governance refers to the ability to collect, identify, track and protect customer and business data. It includes risk analysis around the data, policies to manage the risk and tools to enforce those policies consistently at scale”

Let’s define “Data Classification”

Data Classification

“Data Classification is an iterative process to identify privacy risk posed by individual data as well as various combinations of data, so as to protect them with risk-appropriate tools and processes”

Key Observations

- Process is continuous and iterative
- Purpose is to help protect data
- Require a combination of processes and tools

Data Classification: A Real-World Scenario

Data classification requires context and collaboration

Warm-up exercise

Scenario: An online retail that collects IP addresses so as to analyze how shopping behaviors map to geographic locations.

Classification Tiers

- Level 1 for data that is highly privacy sensitive
- Level 2 for data that is less privacy sensitive
- Level 3 for data that is available in the public domain

What considerations will you account for when you classify the data?

Solution: Data classification scenario

Let’s revisit the key details.

Classify “IP address” based on privacy risk
along 3 levels ranging from level 1 to level 3.

Top Considerations

- IP address maps to physical location
- Knowing physical location creates identification risk
- The privacy risk seems high.
- Should “IP address” be classified as Level 1?
- Answer: It depends!

When it comes to IP addresses, several considerations may affect privacy risk.

How confident are you in the correctness of the IP?

Accuracy depends on VPN, Wi-Fi usage, etc.

IP accuracy is directly correlated to privacy risk. Classifying IP addresses as level 1 without considering accuracy may lead to unintended consequences.

What physical location does the IP address map to?

If it maps to a college dorm or a hotel, the privacy risk may be low.

If it maps to home in a small town with a population of less than a thousand people, the privacy risk may be high.

Do you have other data in your system that could be combined with the IP address data?

Joining the IP addresses with phone numbers or email addresses may end up increasing re-identification risk.

Is the IP address representative of the user’s location?

Key takeaways

- Privacy is contextual, and so is data classification
- Data classification for a data element could depend on other data
- Classification will require iteration; the privacy risk will change as the underlying data changes
- Finally, classification will require collaboration across silos

Data Classification is the foundational starting point for governance.

Data classification will help you apply risk analysis to data dynamically so that you can enforce privacy on an ongoing basis.

Classification is important not just for privacy.

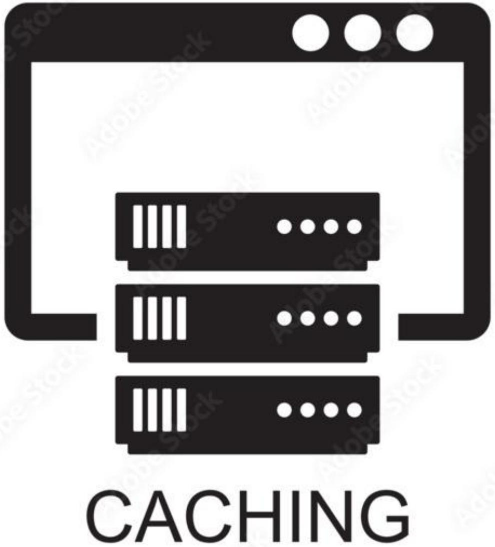
Helps achieve other goals

- Security
- Efficiency
- Quality
- Costs

Data Sharing Challenges

Often occurs in the background without the user, whose data it is, being aware of it.

The other big challenge is a lack of or broken guardrails.

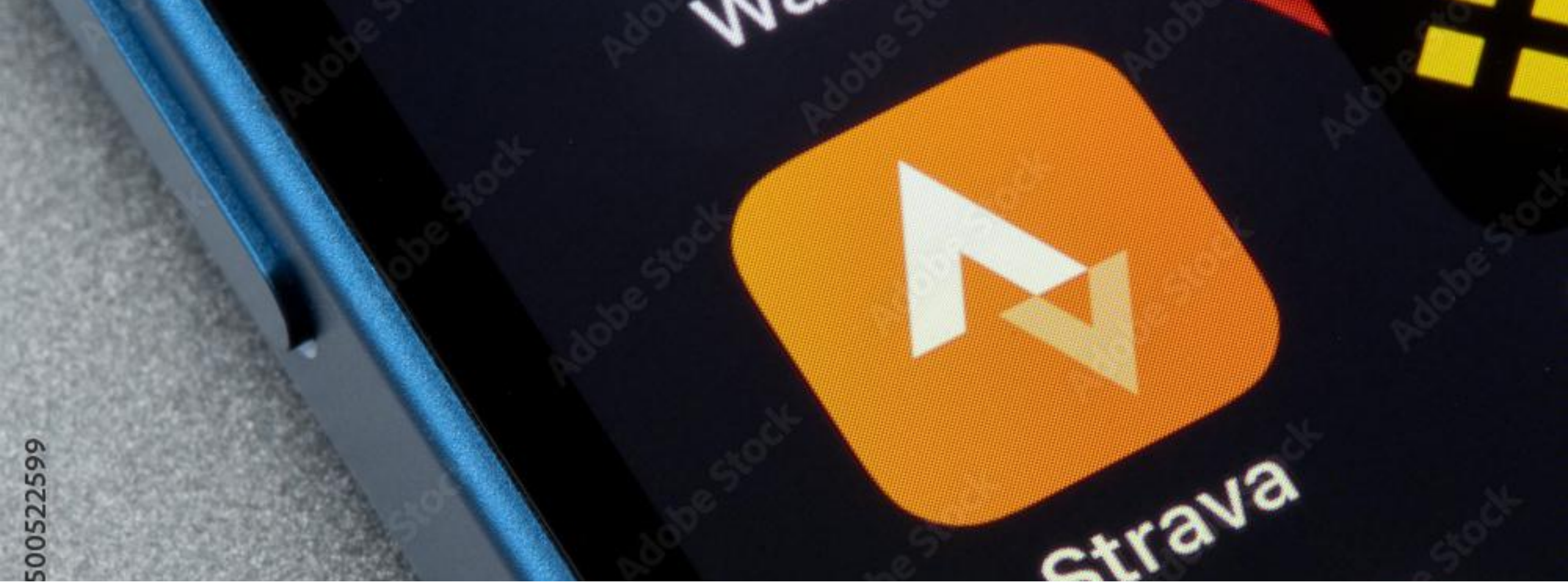


Caching reduces the burdens on backend databases.

This is an example of technical guardrails protecting innovators and consumers from bad outcomes.

Data sharing, however, can cause privacy harms without guardrails due to the nature of user data when combined with other data on social media

JUST ASK THE UNITED STATES MILITARY.



Strava, the fitness tracking app, uses satellites to record its users' runs, bike rides, and other workouts. (Source)

Strava’s Global Heatmap shows where people around the world go running and cycling

US service members had been recording their runs and the Strava heatmap and unknowingly revealed their locations.

Twitter users figured out they could identify outlines and activity patterns on US military bases in places like Syria, Afghanistan, and Somalia

The map showed supply and patrol routes, as well as the precise location of facilities

You could find out which service members were in which locations at a given point in time.

The companies building the products will own the outcomes, not the users

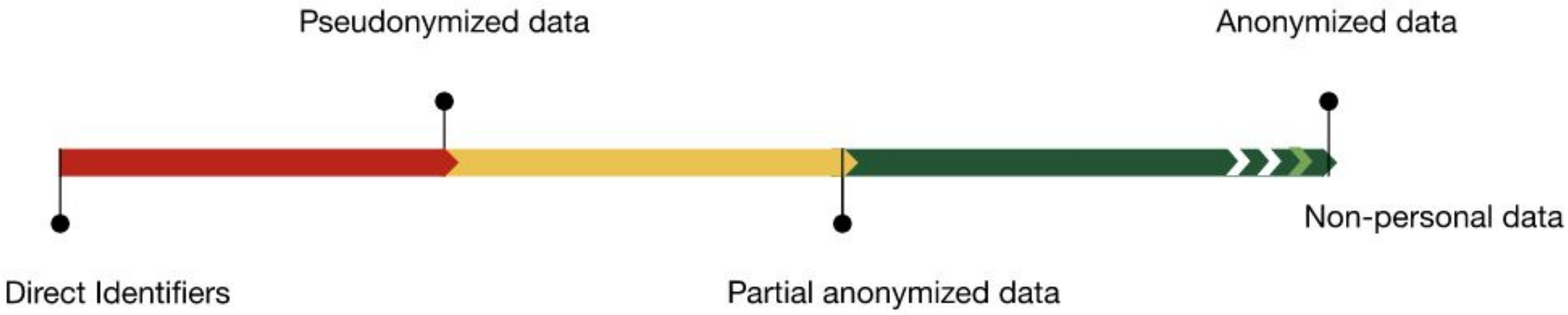
This is how Strava and the United States military were caught by surprise due to data sharing.

I empathize….really

Data sharing + Privacy issues = Major problems.

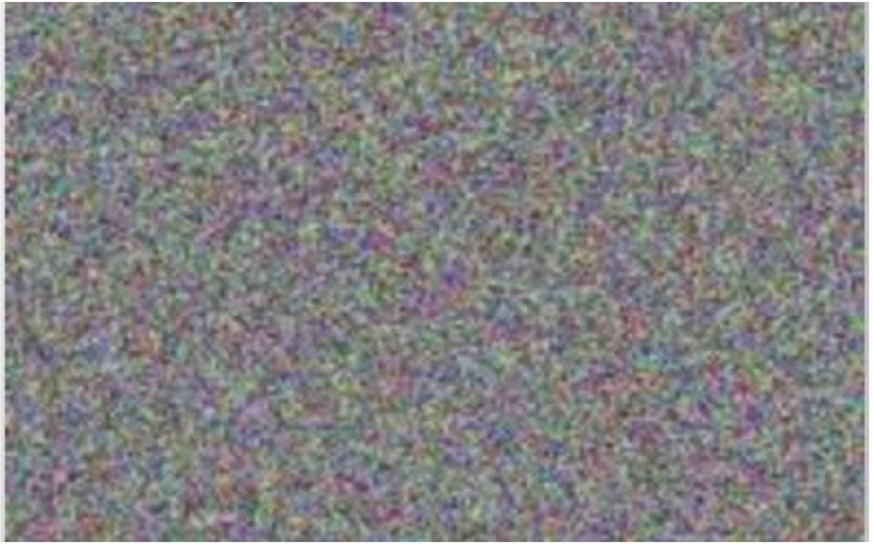
Watch out for red flags

Anonymization techniques



Encryption

What the sender and the authorized recipient see



What everyone else sees

K-Anonymity

A case study: 40,000 Boston trips

Overlay:
https://www.youtube.com/watch?v=d2SF2_RrKg8&t=0s

The goal : how can we reduce the uniqueness of each ride by coarsening the data?

K-Anonymity with 0 decimal points

K-anonymity

GPS Decimal Points

	2	5	10	50	100	1000
0	100%	100%	100%	100%	100%	100%
1	100%	100%	100%	100%	100%	100%
2	100%	100%	100%	99.9%	99.9%	99.1%
3	99.9%	99.8%	99.5%	97.6%	95.3%	87.9%
4	97.4%	93.2%	89.3%	73.1%	59.3%	17.3%
5	68.4%	35.5%	18.3%	2.5%	1.5%	0.9%

K-Anonymity with 4/5 decimal points

K-anonymity

GPS Decimal Points

	2	5	10	50	100	1000
0	100%	100%	100%	100%	100%	100%
1	100%	100%	100%	100%	100%	100%
2	100%	100%	100%	99.9%	99.9%	99.1%
3	99.9%	99.8%	99.5%	97.6%	95.3%	87.9%
4	97.4%	93.2%	89.3%	73.1%	59.3%	17.3%
5	68.4%	35.5%	18.3%	2.5%	1.5%	0.9%

5-Anonymity for 0-5 GPS decimal points

K-anonymity

GPS Decimal Points

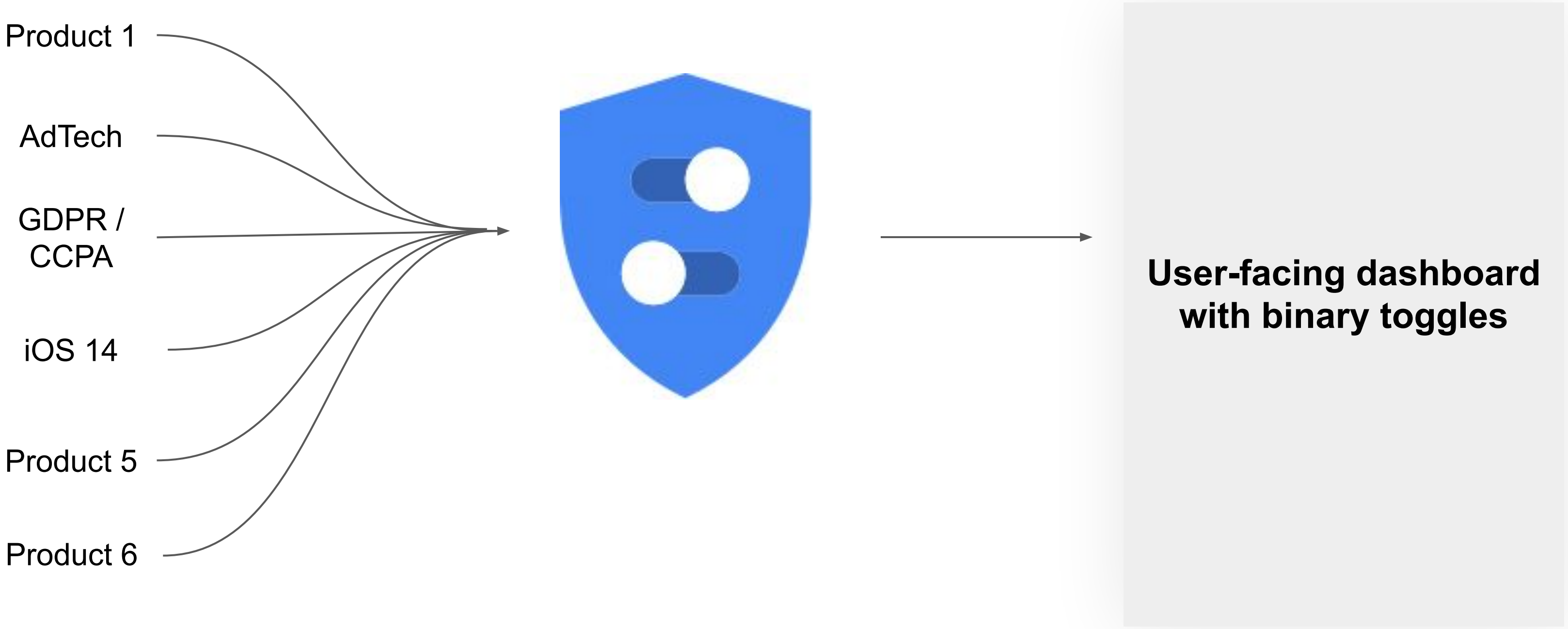
	2	5	10	50	100	1000
0	100%	100%	100%	100%	100%	100%
1	100%	100%	100%	100%	100%	100%
2	100%	100%	100%	99.9%	99.9%	99.1%
3	99.9%	99.8%	99.5%	97.6%	95.3%	87.9%
4	97.4%	93.2%	89.3%	73.1%	59.3%	17.3%
5	68.4%	35.5%	18.3%	2.5%	1.5%	0.9%

Privacy as engineering enabler and differentiator

New business lines + services

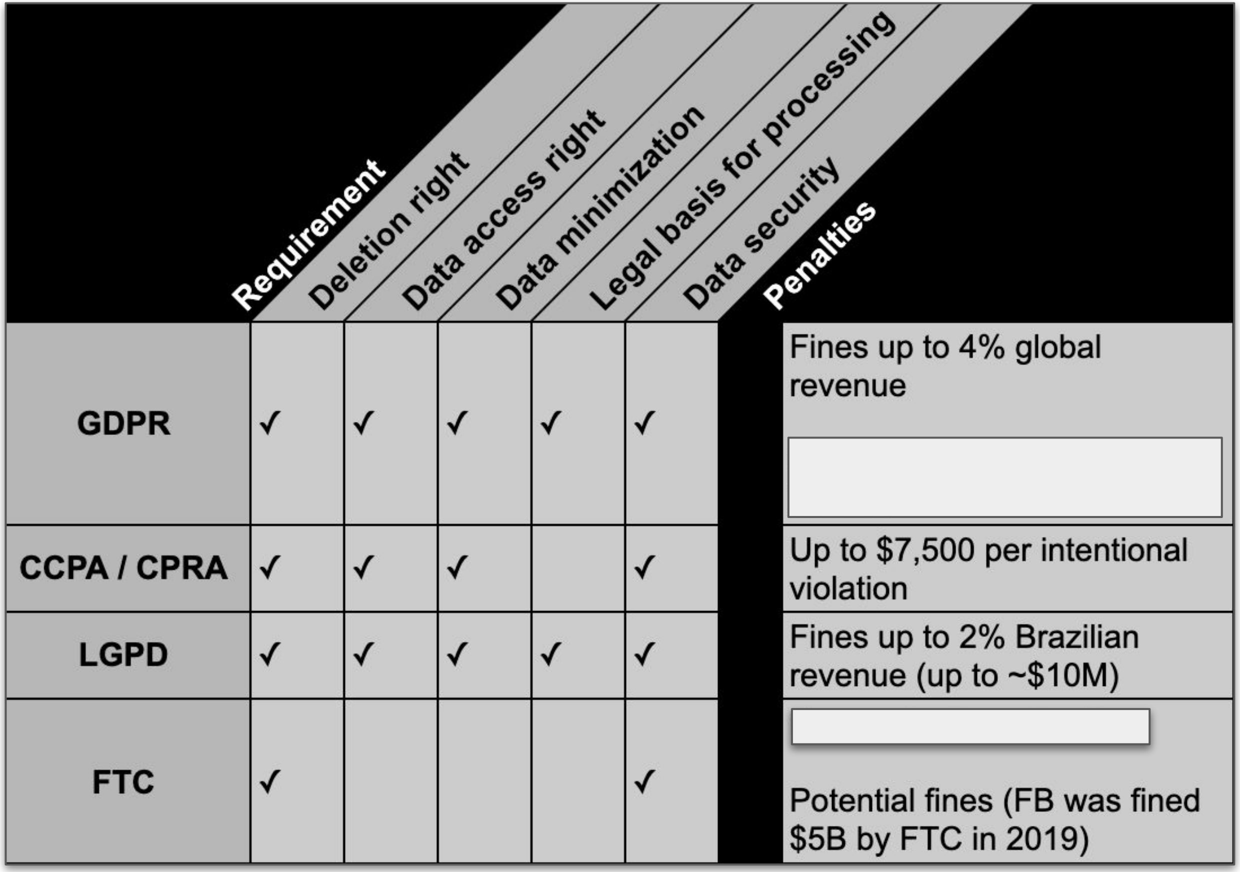
Central APIs for privacy enforcement

Central display for end user



User-facing dashboard with binary toggles

What Legal Risks Could This Mitigate?



Requirement	Deletion right	Data access right	Data minimization	Legal basis for processing	Data security	Penalties
GDPR	✓	✓	✓	✓	✓	Fines up to 4% global revenue
CCPA / CPRA	✓	✓	✓		✓	Up to $7,500 per intentional violation
LGPD	✓	✓	✓	✓	✓	Fines up to 2% Brazilian revenue (up to ~$10M)
FTC	✓				✓	Potential fines (FB was fined $5B by FTC in 2019)