



Data Security and Privacy Anti-Patterns

September 25, 2019

Steve Touw

Co-Founder and CTO, Immuta

Agenda

Stage Setting

- Privacy vs Security
- Importance of privacy
- What is an anti-pattern?

Anti-Patterns

- Snowflakes
- Conflating Who, What, & Why
- Copy & Paste
- Start from scratch

Conclude

- Summary
- Questions

Defining Security and Privacy

Content

cust_last_name	credit_card_number	transaction_time	transaction_amount	action_location
Hollerin	4026087576796554	2018-10-12T20:09:29	176.14	
Putterill	6100079544545679	2018-10-25T23:45:58	679.19	
Thews	357102393131851	2019-03-19T03:51:48	704.07	
Scrannage	491388359027006	2018-07-21T19:09:25	51.32	
Sarjeant	38470670033720	2018-05-31T10:59:13	518	ring
Kohrsen	60187677318808	2018-07-15T15:11:45	664.45	amukti
Helleker	20147699255379	2018-05-12T22:51:26	592.9	
Thews	357102393131851	2019-03-19T03:51:48		Campo Verde
Scrannage	491388359027006	2018-07-21T19:09:25		Angers
Sarjeant	38470670033720	2018-05-31T10:59:13		Caen
Kohrsen	60187677318808	2018-07-15T15:11:45		Zijin
Hollerin	Helleker	20147699255379	2018-05-12T22:51:26	Mozga
Putterill	Bergot	40	New Hampshire	529109
Thews	Claris	75	Texas	159889
Scrannage	Leesa	73	Arizona	194812
Sarjeant	Jennifer	33	Texas	89032
Kohrsen	Greg	42	Maryland	108600
Helleker	Claude	44	New York	92000

Context



Analyst

Does marketing analyst need social security number?



Analyst

Does credit card analyst need gender?



Analyst

Does HR analyst need everything?

Privacy



I want to steal your data!

Security

An Innovation timeline

2000s

Generating and Storing Data

Expanding digital footprint turned business activity, consumer behavior, social lives, etc. into data.

2010s

Deriving Value From Data

Democratized tools, machine learning and interactive analytics made data-driven decision-making possible and valuable.

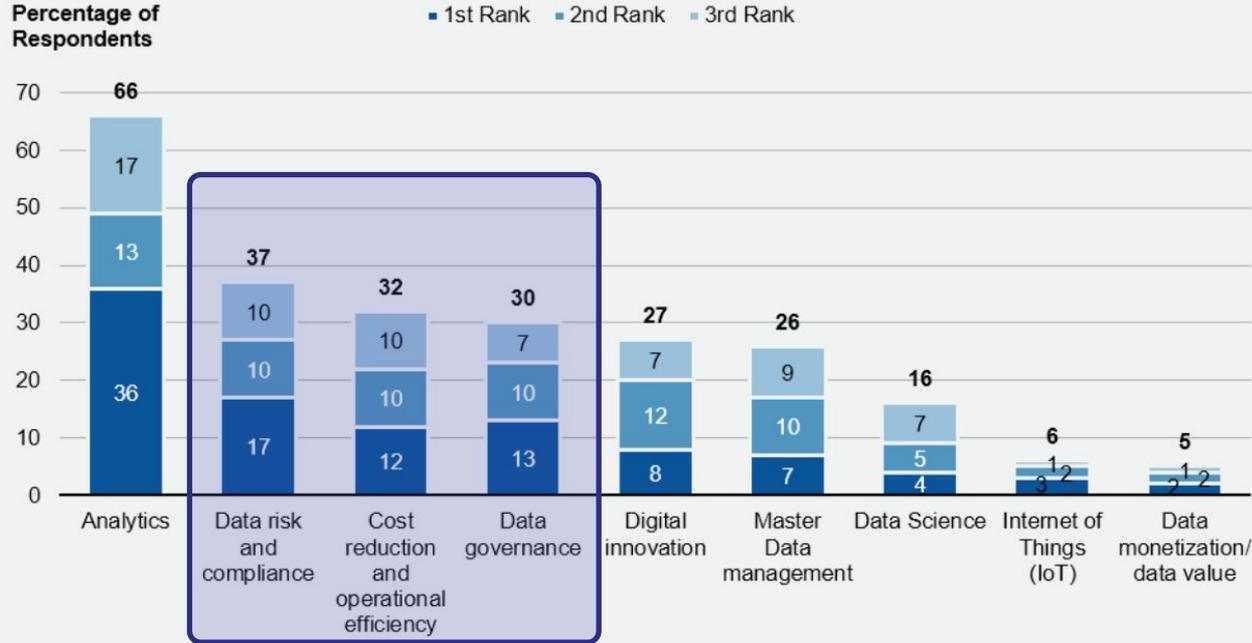
2019

Responsibly Leveraging Data

Increased risk associated with data utilization will generate new legal, ethical, and management tools and standards.

Importance of Current Data Management Initiatives

Top 3 most important (ranked)



Base: n = 104 Gartner Research Circle members/only asked of initiatives "have invested in" at Q02/excludes "not sure."

Q. Which of these current initiatives are most important to your organization's data management strategy today?

Please click to rank your top 3 in order of importance.

ID: 351994

© 2018 Gartner, Inc.

NO:

How much data can I get?

YES:

How much information do I need?

OUR MISSION

Ensure the legal and ethical
use of data.



What Data/Privacy Policies...

Entitlements to Data

- Open to everyone
- Manual approval flows
- Logic prescribed: what groups or attributes must you have to be given access
- Manual add: essentially how tables work today, someone GRANTS access

Row Level Security

- By comparing user attributes to data attributes
- Moving time windows
- Limited percentage of data
- Purpose
- Differential Privacy

Column Masking (cell or column level)

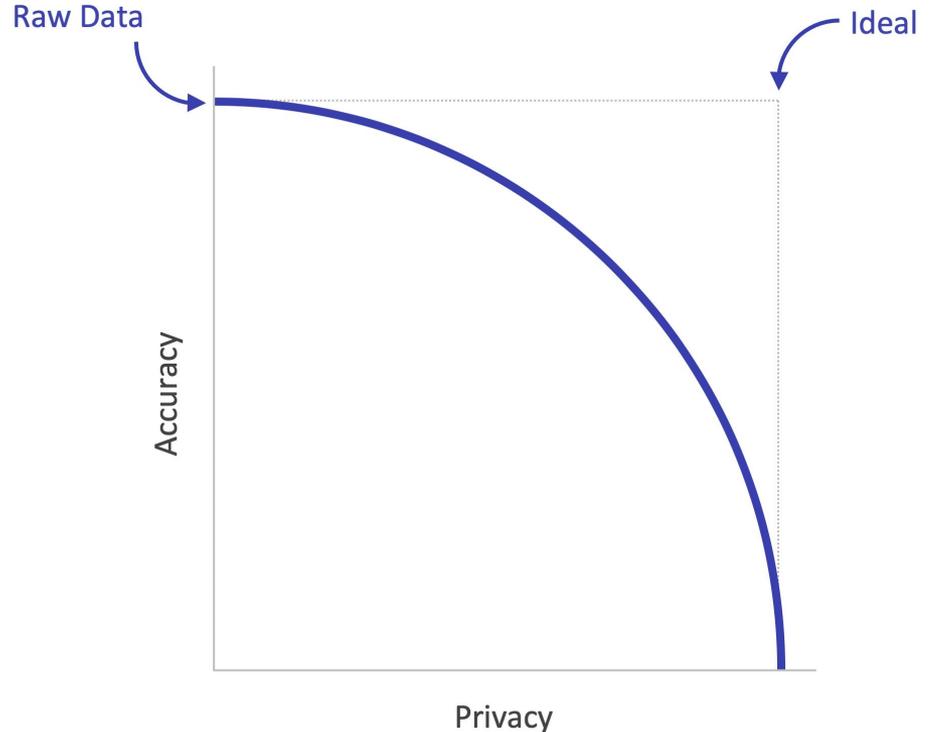
- Salted hash
- Replace with constant or null
- Regular expression
- Rounding / k-anonymization
- Format preserving masking
- Reversible masking

In practice, privacy is a continuum

To preserve privacy, organizations have to make the data less closely resemble the raw data (or full data).

Moving along this curve, data become more robust against certain types of privacy risks.

The actual trade-off is highly coupled with analytical context.



Anti Patterns

What are they?

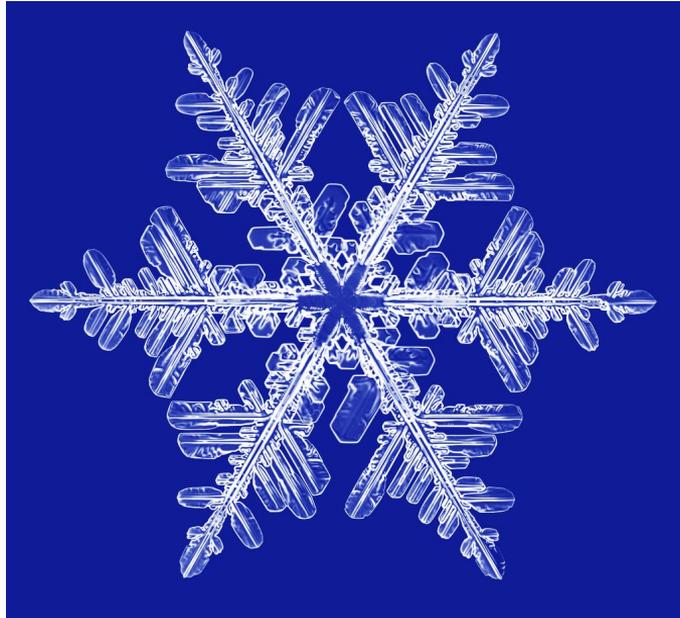
- Seemed like an obvious solution at the time
- Rarely were they obviously a bad idea from the start
- Bad ideas are realized when the world changes underneath you: policies become more complex (think GDPR and the California Consumer Privacy Act [CCPA]) or the organization needs to be more data driven but analytical efforts are stymied

Why you should listen to me?

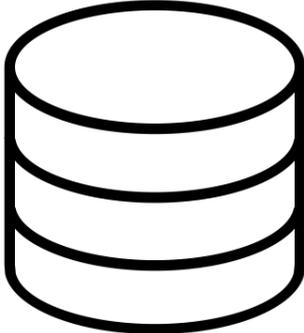
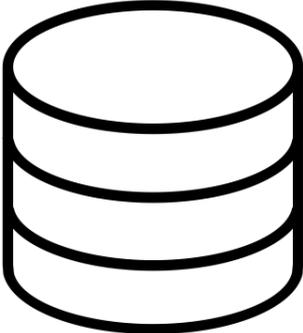
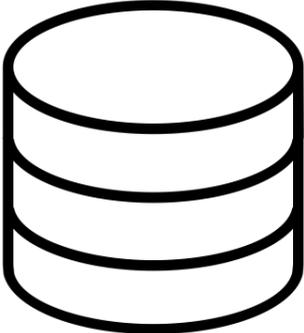
- Worked with 100s of customers across a heterogeneous set of verticals
- Very consistent anti-patterns have emerged

Anti-Pattern 1:

Data Policy “Snowflakes”



You Have Data Everywhere



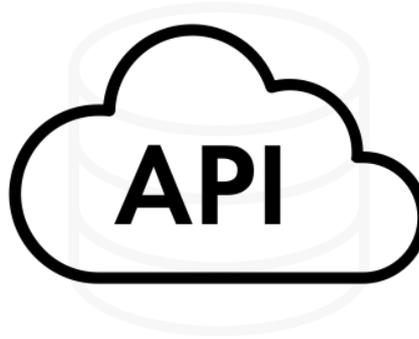
You Must Protect That Data



How Do You Protect It (The Anti-Pattern)



Protected with a custom or commercial app in front of the data



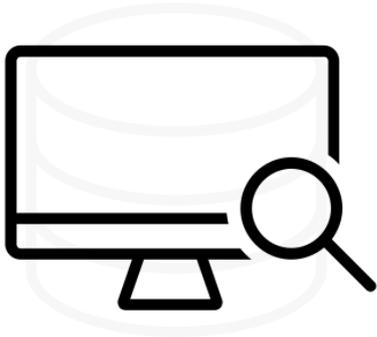
Protected with a unique API specific for the data



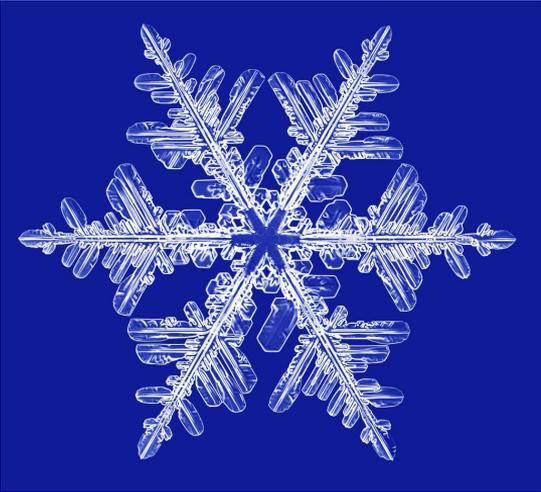
Protected by a slew of custom database views and manual GRANT processes

So Now You Want To Do Data Analytics

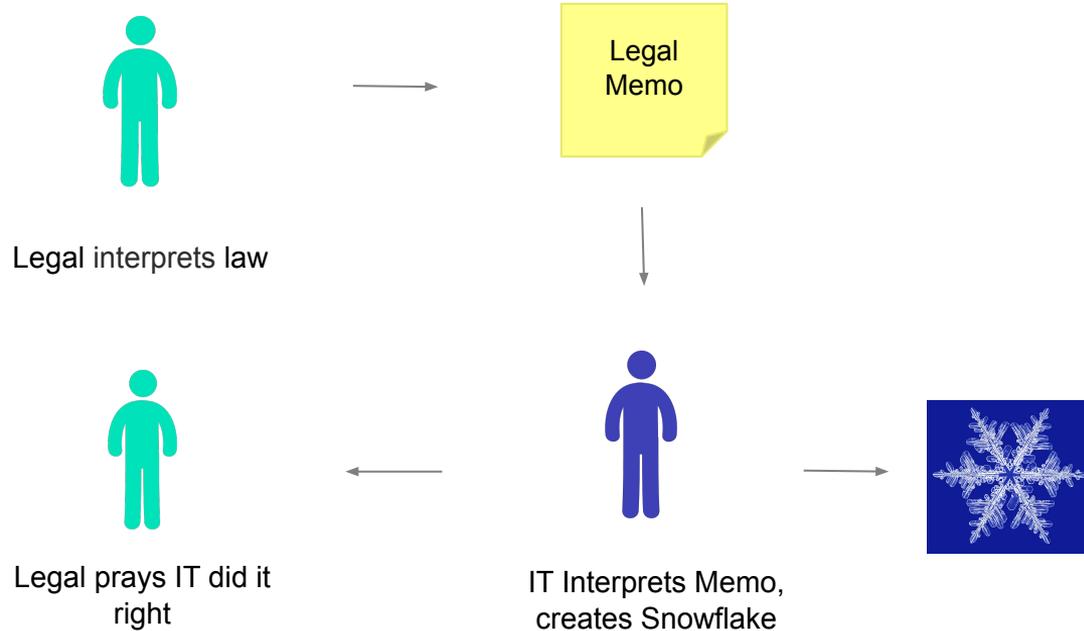
How the heck do I get to all the data? You expect me to stay within the confines of all these various custom access patterns??!?



You've Created Snowflakes



How Snowflakes Are Made



How Do We Fix It?



First Critical Need: Industry Standard Access and Processing Patterns/Paradigms



SQL

Excellent, I can connect in ways I'm used to that works across tools and data with no new code



Spark



HDFS

Filesystem/S3



Second Critical Need: Enforce Data Policies Consistently and Abstractly



Entitlements

Row Level

Masking

Purpose

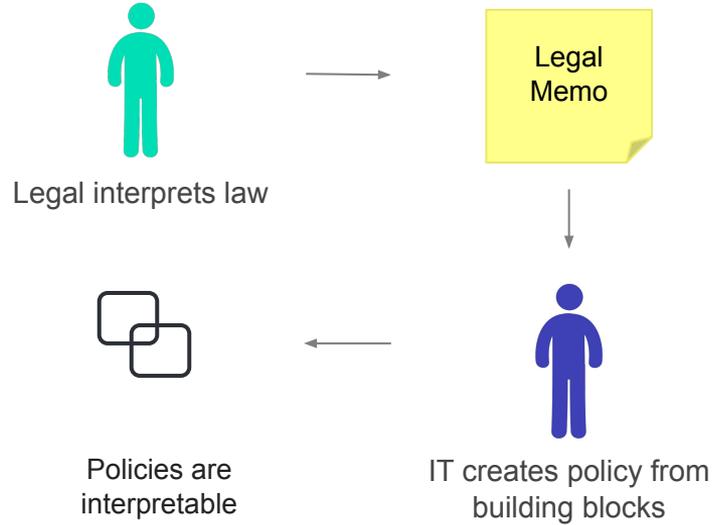
Excellent, I can enforce controls consistently no matter where my data lives or how it's being processed



BENEFIT 1

Interpretable Enforcement

Consistent and understood
policy building blocks



BENEFIT 2

Eliminates Mistakes Through Consistent Policy Building Blocks

Consistent and understood policy building blocks



Legal interprets law



IT creates policy from building blocks



Policies are Correct and visible

BENEFIT 3

Agility: You Can Change a Policy In Seconds

Consistent and understood
policy building blocks



Policies Abstracted and Thus Can
be Enforced Across Many
Sources Through One Change

IT updates policy from
building blocks

BENEFIT 4

Recognized “Digital Handshake” For Data Exchanges

Consistent and understood
policy building blocks




Legal interprets law



Legal
Memo



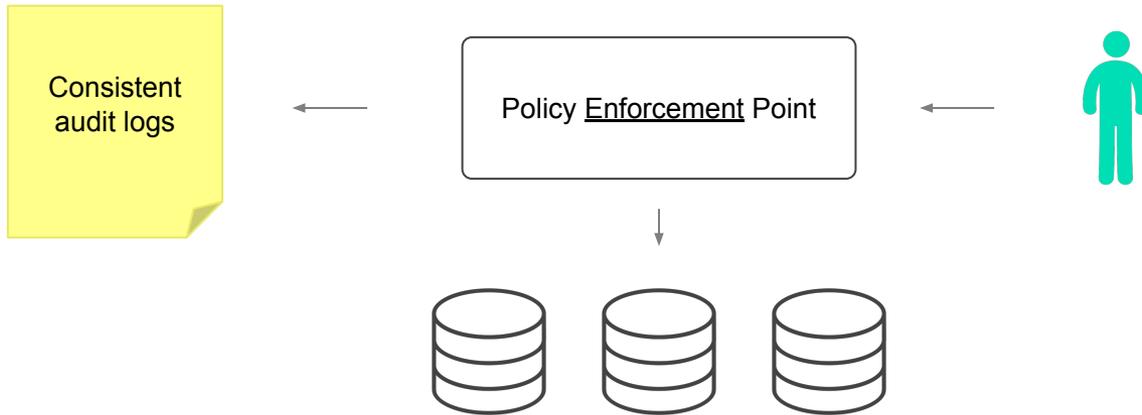
IT creates policy from
building blocks



Everyone knows how to get access
to data. IT understands how to
share data in a recognized way

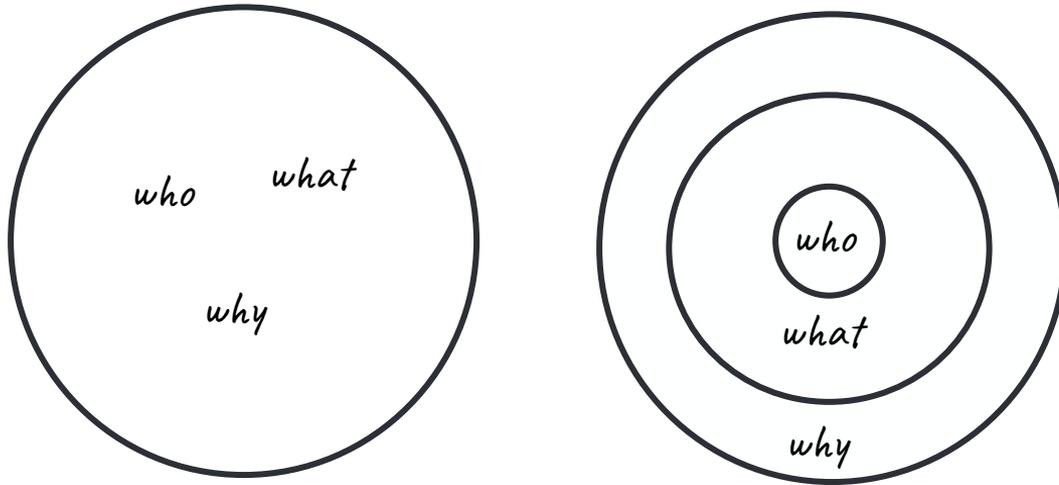
BENEFIT 5

Complete and Consistent Audit Log of All Data Activities



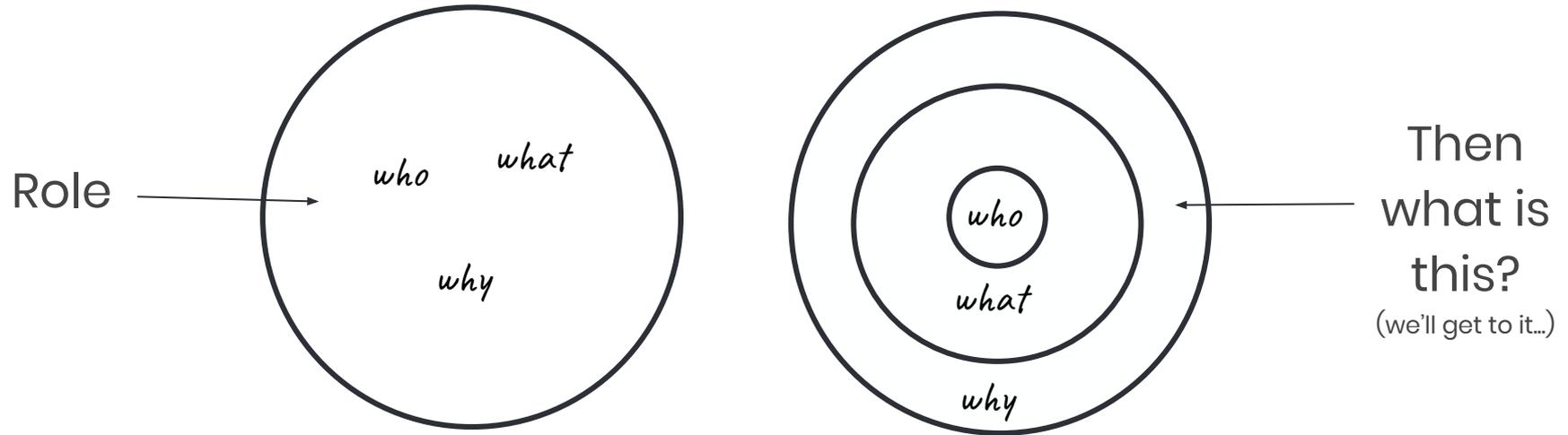
Anti-Pattern 2:

Conflating Who, What, and Why



Also Known As...

Role Based Access Control (RBAC)



A couple views. What's the challenge?

Set a view for each role.

ROLE	FIRST NAME	LAST NAME	PHONE #	SSN	ADDRESS
Marketing					
Product					
Credit					
Fin Crime					

Four roles
establishes
four views?



Marketing



Product

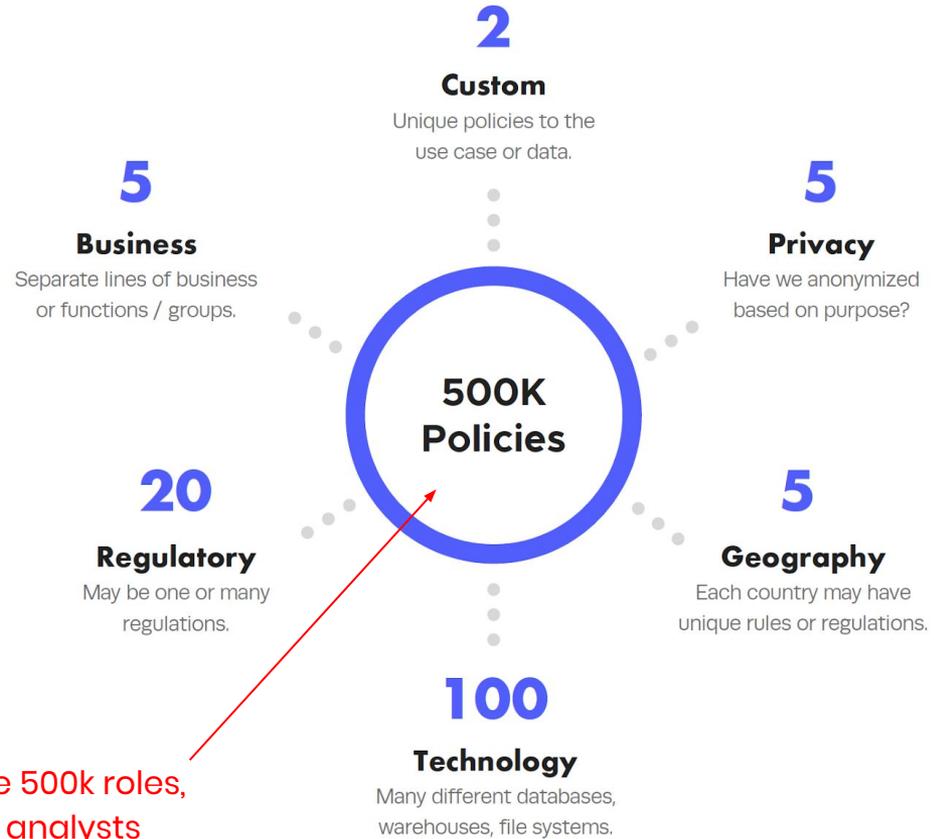


Credit



Financial crime

Humans can't handle so many policies.



Which means you have 500k roles,
more than tables or analysts

Let's Run Through An Example



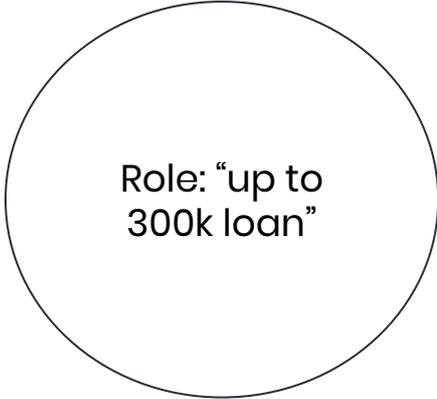
I want to buy a 300k house...

Components to get a loan

- Credit score
- Debt vs. income
- Down payment
- Employment history
- Loan size

Wrong Way / Right Way

conflate who and what - a ROLE



Nobody knows WHY they have access to the 300k loan [role]...

instead describe WHY
someone should get a loan

who

if Credit score > 550
and Debt vs. income (including loan size) < 30%
and Down payment > 5%
Employment years > 1

Then, allow Loan

what

The WHY is clearly prescribed and executable without subjectivity

Let's Beat this RBAC Horse Some More...

Now Bob wants a loan...

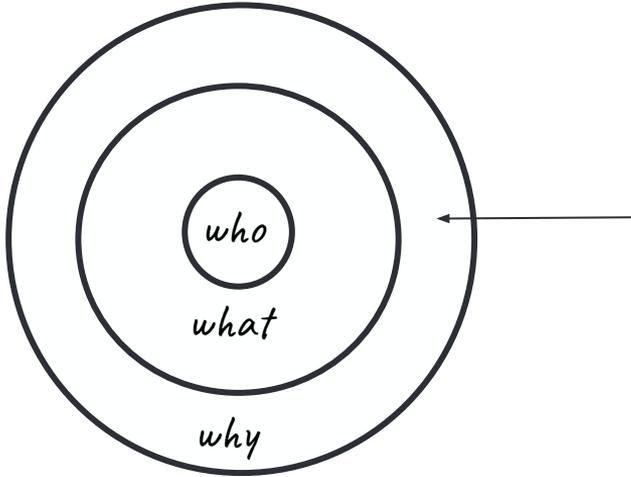


- Do I let him into the “up to 300k loan” role?
- Do I know for sure what that gets him?
- It's a manual decision process (WHY) each time
- We see manual approval flows of 10s of people and backlogs of months!

How Do We Fix It?

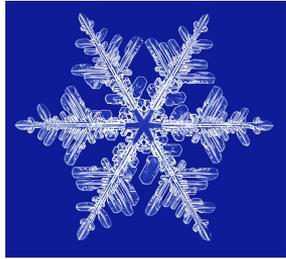


Separation

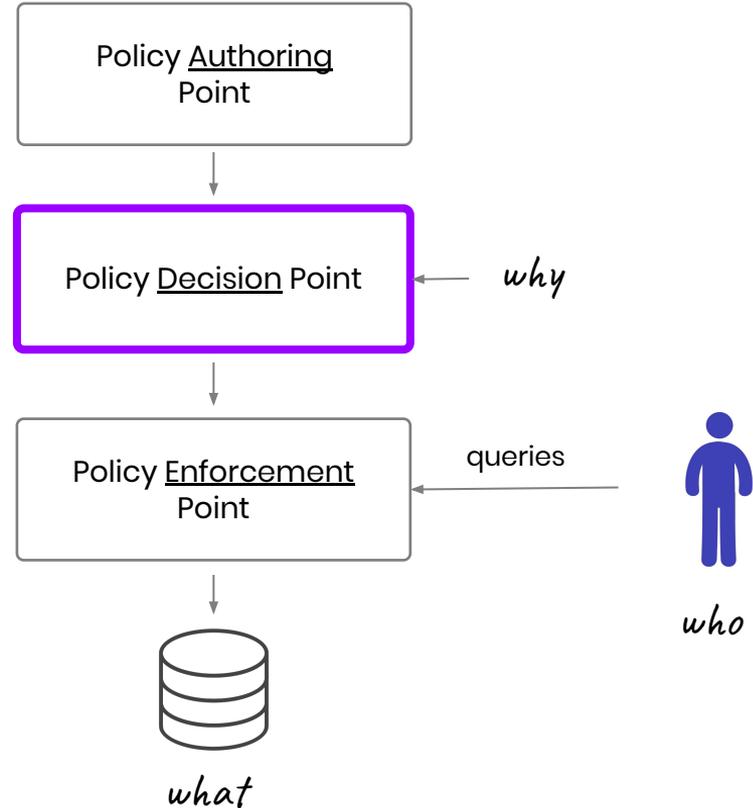


- **Define your users (WHO)**
- **Define your data (WHAT)**
- **Define decisions with logic (WHY)**

What Does This Mean In Practice?



Break Out the Snowflake to allow scalability, flexibility, comprehension

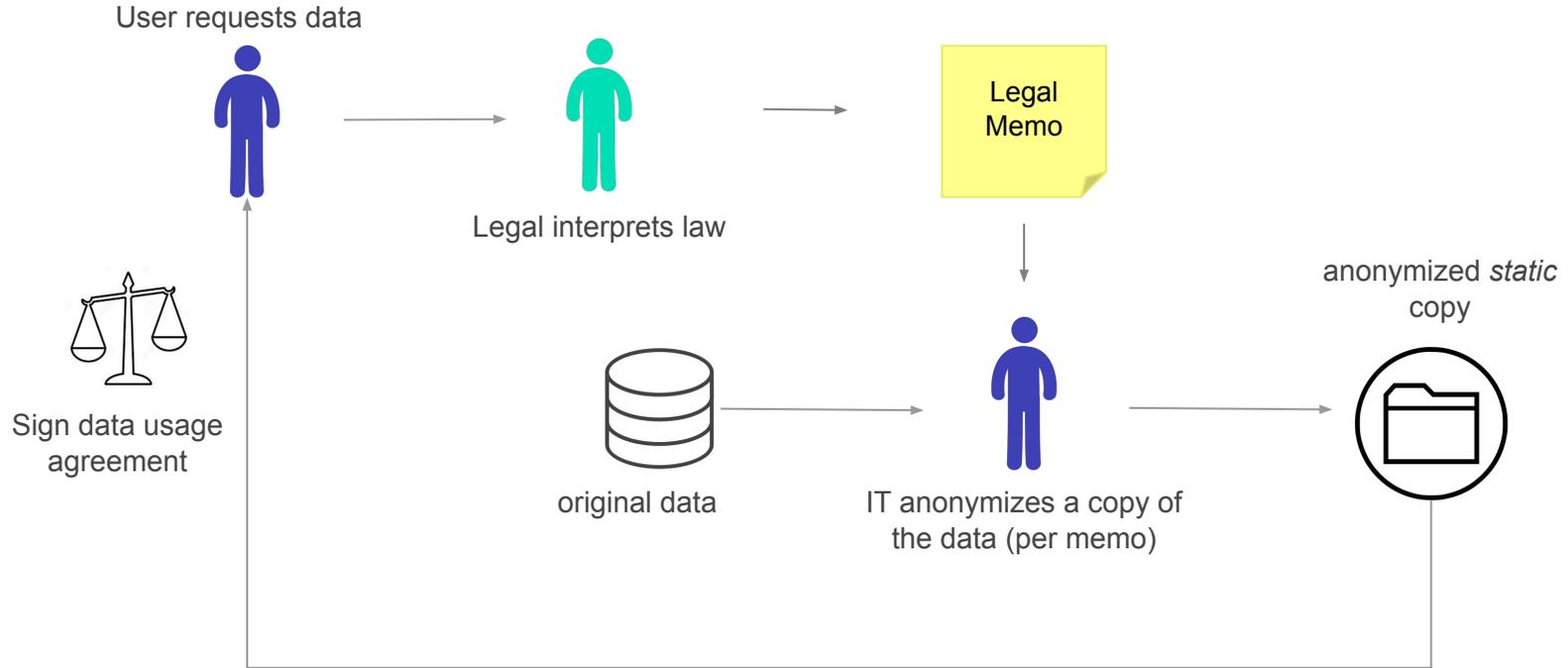


Anti-Pattern 3:

Copy & Paste



I NEED DATA!!



The User is **Frustrated**

- Have **no process** to discover or request access to the data across the organization.
- They typically have to **wait months** (yes, months) for the data to arrive.
- They're working with a **static** snapshot of the data, which is typically **months old** – if not simply outdated – because of the above process.
- They're required to sign **data usage agreements** and therefore must be very careful about how they subsequently share the data with their colleagues



The Organization is **Frustrated**

- They **lose insight** into who has what data and how it's being used.
- They significantly **increase storage costs** as many anonymized versions of data need to be stored for various different user scenarios.
- They have **complex** ETL “spaghetti” to manage the creation of all the anonymized copies.
- It **isn't clear** how the anonymization policies are actually implemented (or if they are correct) across different data systems (see anti-pattern 1).
- Biggest of all, their **data science initiatives are stymied** because frankly, none of this works and nobody can access data in the way they need.



How Do We Fix It?



A Perfect Analogy?

- A great analogy is how Netflix led to the demise of Blockbuster. **Blockbuster was the copy and paste method.** After searching through the store on foot, you got the raw video, watched it, remembered to rewind, and returned it.
- Netflix changed that. Instead of copy and paste, **they provided a live feed to the movie over the internet.** The value here was discovering the movie you wanted through a web search then having immediate access from your living room, without moving from your couch.



Data analytic programs need Netflix, not Blockbuster. With Blockbuster, they fall apart.

Be

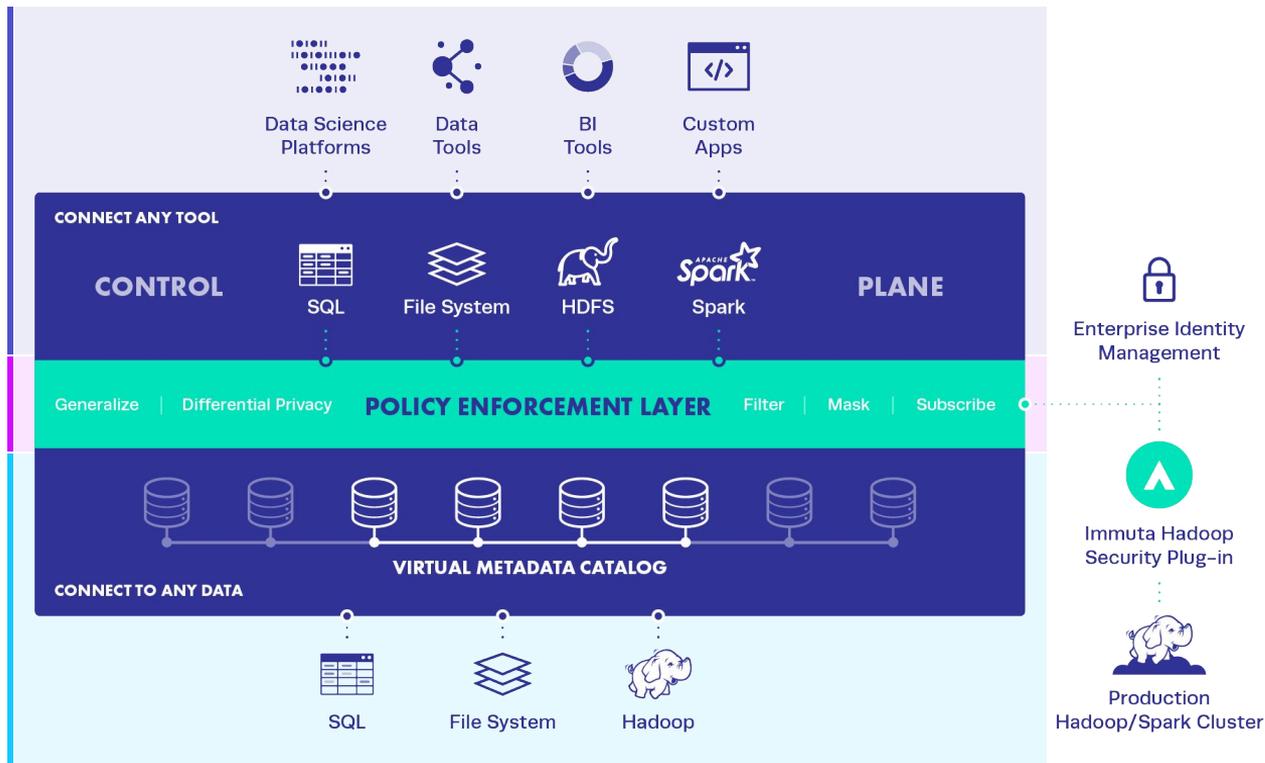
NETFLIX

Through a Data Control Plane...

Data Analysts
Play Here

Legal & Compliance
Plays Here

Data Owners
Play Here



The User is Happy

- They can rapidly discover data, request access, and immediately be entitled to the data based on the logic of the access policy.
- They are accessing live, up-to-date data through industry standard access patterns.
- They are 1) comforted knowing they followed a recognized access process and 2) can share work (code) with their colleagues, knowing they'll be able to access the data through this same control plane.



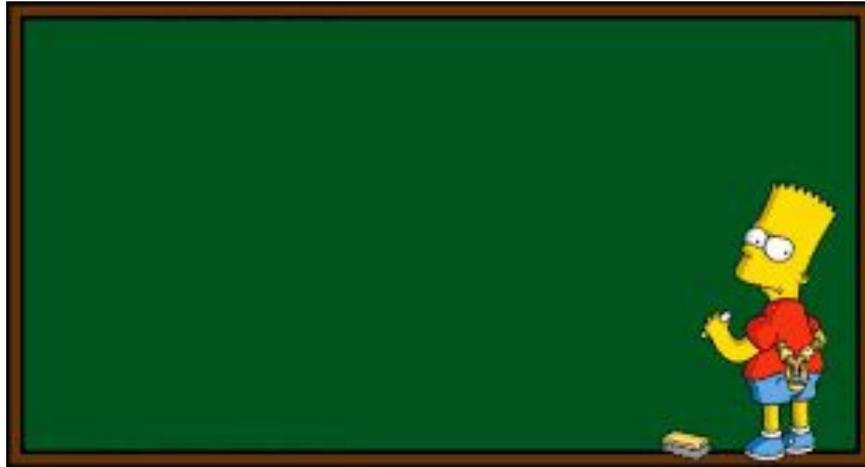
The Organization is Happy

- Data owners can share their data with complex access and anonymization policies that are enforced at request and query-time. No ETL jobs, no extra storage.
- Policies can be reviewed or enhanced by legal and compliance – the policies can be written in a way that's simple to understand by all employees (who, what, why).
- The control plane, which reduces your surface area for a security breach, and acts as an abstraction to your database.
- Full audit of all actions are captured with reporting capability to fully understand who is using what data for what purpose.



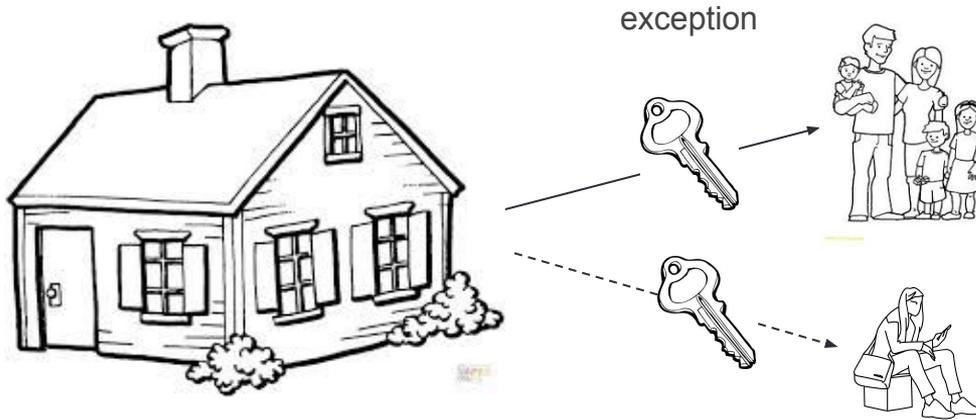
Anti-Pattern 4:

Start From Scratch



Let's Start With An Example...

- The key to your house...you decide who to give it to, right?
- You don't list every person that shouldn't have a key, you instead **lock everyone out**, and then add **exceptions** on who should get the key
- You might even make temporary exceptions for the babysitter, for example



Think About That In The Data World

- Certainly, this is how database **tables** work, you are GRANTED access
- However, this **breaks down in a fine-grained access world** because it's a lot more complicated than just access to the table or not
- You must set more **complex default behavior**

ROLE	FIRST NAME	LAST NAME	PHONE #	SSN	ADDRESS
Marketing					
Product					
Credit					
Fin Crime					

Four roles establishes four views?



Marketing



Product



Credit



Financial crime

Why Must You Have More Complex Default Behavior?

- Well, in short...it's complex!
- If you have to **start from scratch** for every use case / user in your organization on what they can and can't see...you're going to make a lot of mistakes and make a ton of work for yourselves.
- The anti-pattern is starting from scratch on every access decision. You might be doing who, what, and why correctly - but even if you're doing that, starting from scratch each time is just wrong.
- More realistic examples...

We See Customers Define Access Per Use Case

Data sharing agreement X

- Employee table:
 - Mask columns:
 - name
 - phone number
 - address
- Client table:
 - Full access

Data sharing agreement Y

- Employee table:
 - Full access
- Client table:
 - Mask columns:
 - name

Data sharing agreement Z

- Employee table:
 - Mask columns:
 - name
 - phone number
 - address
- Client table:
 - Mask columns:
 - name

duplicated

duplicated

Most Restrictive

How Do We Fix It?



Remember, Think Foundation Then Exceptions

foundation



Day to day data access



WHY: "If in group HR, allow access to column address"

WHO: "user groups: HR, Leadership"



Things like DSAs



EXCEPTIONS: "If acting under agreement y
allow access to column address"

exception



Let's Write It That Way...

Mask Employee.name

- for everyone
- except when acting under data sharing agreement **Y**

Mask Employee.phone number

- for everyone
- except when acting under data sharing agreement **Y**

Mask employee.address

- for everyone
- except when acting under data sharing agreement **Y**

Mask Client.name

- for everyone
- except when acting under data sharing agreement **X**

Benefits:

- **No duplication**
- **No errors**
- **Future proof, e.g. not starting from scratch**
- **Users know what to request**

When Starting From Scratch...

Do I block the name column?

Do I block the phone column?

Do I block the address column?

Do I block the [x] column?



hi compliance leadership, I need access to the Employee table to do this new location analysis



When Setting Foundation and Adding Exceptions...

Mask `employee.address`
for everyone
except when acting under data
sharing agreement Y
or when doing location analysis



hi compliance leadership, I need
access to the **address column** to do
this new location analysis

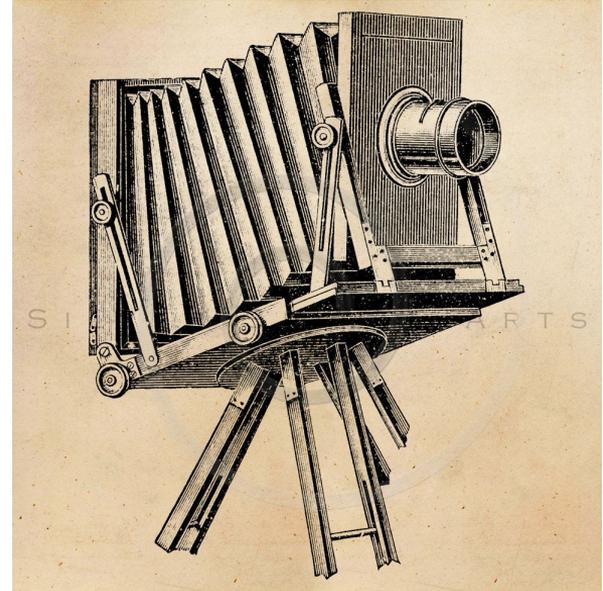


But How Do we Know When Someone is Doing Location Analysis?



Right To Privacy?... a short tangent:

- Early on photography was expensive
- Near the turn of the century the masses had general use of photography
- "instantaneous photographs and newspaper enterprise have invaded the sacred precincts of private and domestic life." – Samuel Warren and Louis Brandeis (U.S. Supreme Court Justice)
 - Proposed right to “be let alone”
- We generally accept being **observed**, but rarely accept being **identified**

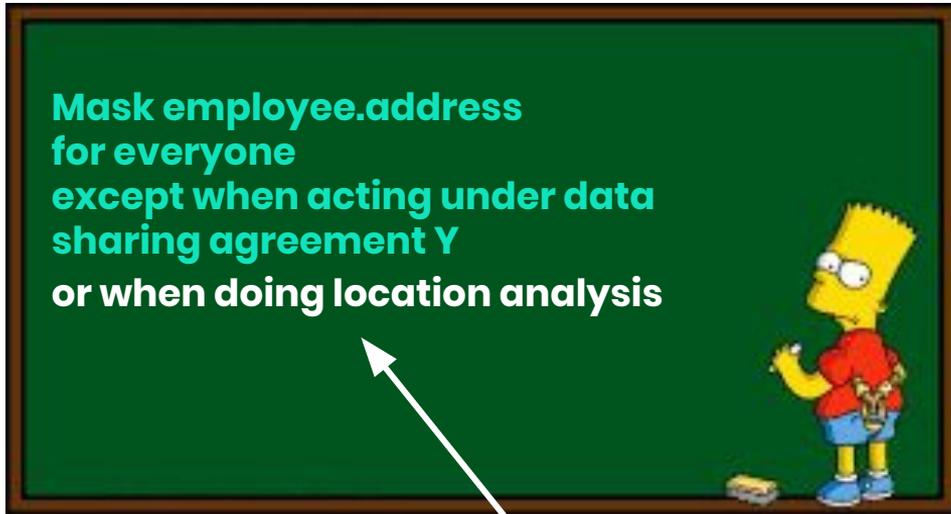


The End of Privacy [as we know it]?:

- Rise of technology and data science has killed privacy as we know it
- Instead of focusing on how and when our data is gathered...
- Privacy should now be **how our data is being used.**



Purpose is a User Attribute



hi compliance leadership, I need
access to the **address column** to do
this new location analysis

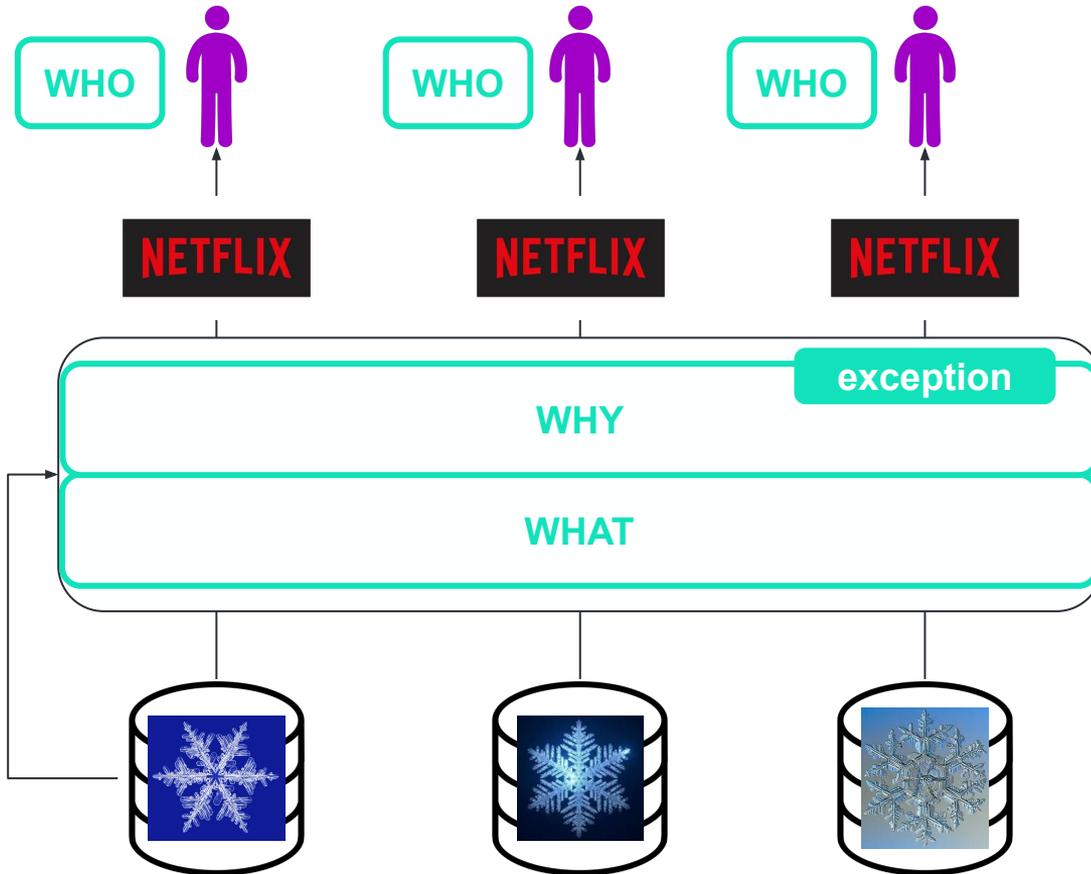


You must treat purpose like a
user attribute!

Let's Summarize



Streamlined Governance, Empowered Analysts



3) Provide “live” access to your “content”, don’t copy

4) Think in exceptions to foundations (and Purpose!)

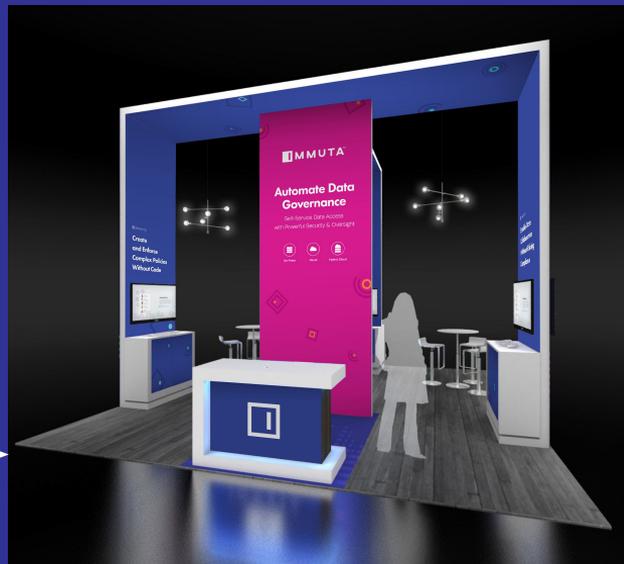
1) Eliminate policy snowflakes, abstract into a consistent control plane

2) Avoid ROLE bloat and separate who, what, and why

Thanks! We're Hiring!

(especially need sales folks)

Booth 1307



Steve Touw

CTO, Immuta

steve@immuta.com



(800) 655-0982

WWW.IMMUTA.COM

[@IMMUTADATA](https://twitter.com/IMMUTADATA)

Rate This Session!

Cyberconflict: A new era of war, sabotage, and fear

See passes & pricing

David Sanger (The New York Times)
9:55am-10:10am Wednesday, March 27, 2019
Location: Ballroom
Secondary topics: Security and Privacy

 Add to Your Schedule
 Add Comment or Question

Rate This Session

We're living in a new era of constant sabotage, misinformation, and fear, in which everyone is a target, and you're often the collateral damage in a growing conflict among states. From crippling infrastructure to sowing discord and doubt, cyber is now the weapon of choice for democracies, dictators, and terrorists.

David Sanger explains how the rise of cyberweapons has transformed geopolitics like nothing since the invention of the atomic bomb. Moving from the White House Situation Room to the dens of Chinese, Russian, North Korean, and Iranian hackers to the boardrooms of Silicon Valley, David reveals a world coming face-to-face with the perils of technological revolution—a conflict that the United States helped start when it began using cyberweapons against Iranian nuclear plants and North Korean missile launches. But now we find ourselves in a conflict we're uncertain how to control, as our adversaries exploit vulnerabilities in our hyperconnected nation and we struggle to figure out how to deter these complex, short-of-war attacks.

David Sanger
The New York Times



David E. Sanger is the national security correspondent for the *New York Times* as well as a national security and political contributor for CNN and a frequent guest on *CBS This Morning*, *Face the Nation*, and many PBS shows.

✓ Attending

Notes

Remove

Cyberconflict: A new era of war, sabotage, and fear

9:55 AM - 10:10 AM, Wed, Mar 27, 2019

Speakers



David Sanger
National Security Correspondent
The New York Times

Ballroom

Keynotes

David Sanger explains how the rise of cyberweapons has transformed geopolitics like nothing since the invention of the atomic bomb. From crippling infrastructure to sowing discord and doubt, cyber is now the weapon of choice for democracies, dictators, and terrorists.

SESSION EVALUATION