The Life-Changing Magic of Tidying Up Data

Does data ‘Spark Joy’? If it does, keep it. If not, dispose of it.

Agenda

01 | The Data Hoarding Problem

02 | Knowing the Data

03 | Locating the Data

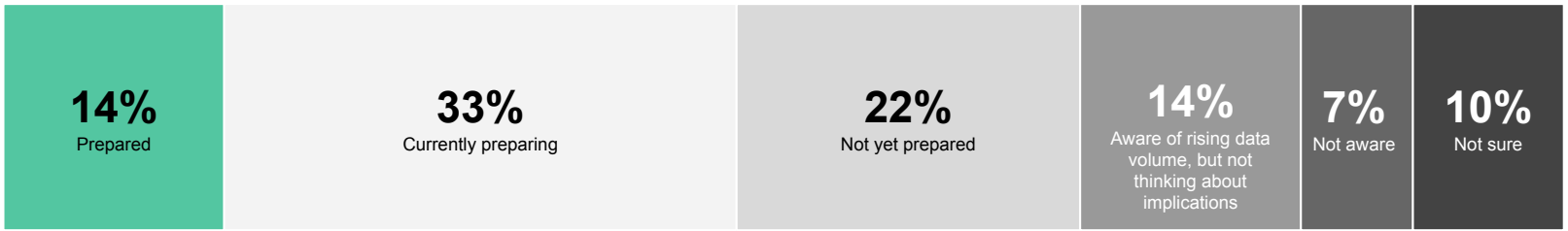
04 | Limiting Data Sharing

05 | Deleting Data that doesn't 'Spark Joy'

Data Growth

Enterprise data volumes are projected to grow nearly **5 times** by 2025. Even if Dark Data does not get proportionally worse, the rising data volume may lead to more dark data. The majority of organizations are not prepared for an influx of data at scale.

Degrees of preparedness for the Data Age



Source: The Data Age is Here. Are You Ready? Splunk 2020 Report

Data Sprawl

As data growth proliferates out of control, sensitive data becomes increasingly vulnerable to leaks, breaches, and compliance fines. When it comes to data sprawl, C-suite IT execs (40%) say their top concern is the increased **risk of data breaches**.

Top Drivers for Data Sprawl

#1	#2	#3	#4	#5	#6
51%	**49%**	**48%**	**42%**	**41%**	**40%**
Increase in Remote Work	Needs of Individual Departments	Use of Multiple File-Sharing Applications	Lack of a Central Repository For Files	Lack of Enforcement for Data-Use Policies	Lack of Company-Wide Data-Use Policies

Source: [Egnyte Governance Trends Report Fall 2020]

DOES DATA ‘SPARK JOY’?

85% of the data is no longer relevant, adequate, or necessary for carrying out the purpose for which it is processed

Source: Gartner Market Guide for File Analysis Software 2018

Privacy's Dependence on Data Governance

Know Your Data

Creation

Analyze schema to identify personal data attributes

Limit personal data collection to what is necessary for the purpose

Locate the Data

Storage

Discover personal data across the enterprise

Identify opportunities to delete personal data in unauthorized locations

Prohibit personal data storage in authorized locations

Limit Data Sharing

Usage

Perform ongoing scanning and validation of data transfers

Restrict non-compliant personal data sharing with third parties

Dispose Data that doesn't 'Spark Joy'

Disposal

Delete/Anonymize Redundant, Outdated, or Trivial data

Dispose of data that is no longer needed

Let’s look at some case studies to demonstrate how practical data minimization was achieved through each stage of the data lifecycle.

Know Your Data

Creation Stage Case Study

Problem Faced

An IoT startup collecting massive volumes of data on their platform wanted to:

- Avoid ubiquitous profiling of users
- Use privacy as a point of competitive differentiation
- Address enhanced due diligence checks from controllers
- Comply with privacy regulations

Constraints included a limited budget and no dedicated privacy resources.

Solution Approach

Using traditional data classification and security tools, conducted an attribute-level analysis of user data schema.

Developed models that prohibit aggregation of specific attributes to prevent 'Profiling' users.

Visualized personal data generated and identified personal data that should not be collected, enforcing the 'collection limitation' principle from the get-go.

“Before you start, visualize your destination.”

PI Catalog
Identified 1000+ privacy attributes collected and established the initial scope of the privacy program.

Compliance Register
Data-driven granular compliance register to map industry specific and international privacy related requirements.

CPRA Readiness
Avoiding the risk of ‘automated decision-making’ with aggregation limits.

How do we keep the catalog current?
Establish processes for periodic scans and privacy approval requirements for changing/adding new attributes.

Locate the Data

Storage Stage Case Study

Problem Faced

An automotive company with a complex technical environment but established privacy program wanted to:

- Have a single source of truth of all personal data
- Enhance data maps
- Automate GDPR data subject rights

The company had low confidence in the data maps previously developed using survey and interview-based approaches.

Solution Approach

Performed metadata scans for preliminary data categorization to identify ROT data and identify files in-scope for the content scan.

Performed full content scans to identify personal data and contextually classify files based on content.

Enriched discovered data with input from stakeholders to create consolidated data maps with business context.

“The aim of storage is to give every item a home.”

Excessive Data Sprawl

Identified that ~20% personal data was in unapproved locations and over-provisioned repositories.

Living Data Maps

Created data maps anchored in data sources that are current and can help privacy risk decision making; a significant improvement over the manual versions.

Secure Enclave

Once personal data locations were identified, data protection efforts could be focused on these areas. Secure enclaves were established to prohibit data exfiltration outside these approved locations.

Limit Data Sharing

Use Stage Case Study

Problem Faced

A consumer goods conglomerate with more than 50 distinct brands, each with its own consent management process wanted to:

- Gain visibility into third-party data flows
- Implement 'Do not Sell' requirements under CCPA
- Consolidate all customer data into an integrated marketing solution while respecting consent

In majority of the cases, consent was brand specific and each brand used 100s of vendors.

Solution Approach

Conducted data flow discovery via web tracker managers and APIs that monitor data in motion to validate what attributes are being shared.

At the attribute level assessed whether the flows are consistent with the legally defined purpose of processing.

Correlated opt-out preferences from existing consent management solutions to action on 'do not sell' requests and monitor data flows on an individual level.

“Tidying is just a tool, not the final destination”

Excessive Data Sharing Detection
5% data flows to third parties were inappropriate with terminated contracts, disabled accounts, personal email accounts, or sharing more PII than approved.

Third Party Reporting Compliance
Documented GDPR record of processing activity based on actual data and enabled continuous CCPA compliance by detecting new data flows.

Enhanced Vendor Risk Management
Ongoing scanning and discovery of data flows provided insights to identify discrepancies in vendor risk assessments.

Dispose Data that doesn't 'Spark Joy'

Disposal Stage Case Study

Problem Faced

A healthcare company faced with ~50% YoY growth in storage wanted to:

- Reduce storage costs
- Limit HIPAA compliance risks
- Enhance the maturity of its privacy program

Solution Approach

Discovered PII using traditional regex and pattern matching techniques.

Used context-based classifiers to discover inferred personal data by identity, type, and sensitivity.

Defined data retention policies and created automated workflows to take action.

“Storage experts are hoarders.”

Storage Reclamation

Less than 50% of the total data scanned contained PII. Redundant Files (<10%) and Outdated Files (<10%). A significant portion of the scanned data was Abandoned Files that needed further analysis. $2M/yr cost savings by deleting toxic data.

Enhanced DLP Rules

Leveraged insights from personal data inventory and analysis to enhance out-of-the-box DLP classifiers.

Takeaways

1. The survey and spreadsheet based approaches are starting points but not sustainable.
2. Data governance is the foundation for building privacy programs at scale.
3. Cross-functional teaming is critical for privacy initiatives.
4. Beyond GDPR compliance, leverage data minimization to guide strategic decisions.

Thank you.