BigID

Your Customers' Privacy, **Protected!**
www.bigid.com • info@bigid.com • @bigidsecure

Building a Modern Data Registry: From Content to Context

Introduction

For organizations, understanding what data they store and analyze is gaining increasing urgency owing to new privacy regulations, security imperatives and pressure to extract more value from the information they store. Historically, organizations invested in a variety of technologies to inventory their physical assets such as servers and PCs but lacked adequate technology to find, map and inventory data assets.

What technology did exist within the data protection or data governance fields lacked the necessary granularity, context or data source coverage to meet contemporary problems like data privacy compliance. Contemporary problems require contemporary solutions. This white paper will examine approaches to building a modern decentralized, data registry that looks beyond either data classification or cataloging to give organizations the data intelligence necessary to support modern data privacy, protection, and governance use cases that require detailed data knowledge.

Beyond Data Classification and Cataloging

Historically, organizations wanting to gain insight into the data assets populating their servers and applications had to rely on decade-old security-centric data classification or governance-centric data cataloging technologies. Neither is adequate for building an inventory or registry of data assets because they lack sufficient content granularity, usage context or data source coverage to cover all the places modern enterprises keep data.

Data classification, found in data protection tools like DLP, DRM, DAM and Access Governance, was rooted in regular expression based pattern matching to find data of a specific type. These tools originated when organizations had fewer places where they kept data (like a relational database, file share or mainframe) and also more limited types of data they wanted to classify (like credit card or social security data PII). However, these tools prove inadequate for contemporary problems and environments where you want to scan across a wider universe of cloud, big data, NoSQL and applications; find more than just PII; and also understand what person or entity the data is associated with.

Conversely, data governance tools evolved data cataloging technology based on extracting technical metadata to help make data discoverable for data professionals. However, like a data Yellow Pages, data cataloging can provide some information to where data of a particular type can be found without shining any light on what specific data is actually kept there. Moreover, like their data classification cousins, data cataloging lacks broad coverage since many systems do not surface or capture metadata. These tools also lack usage context since they do not capture operational or business metadata that can indicate access, purpose or consent.

Neither data classification or cataloging alone can accomplish the goal of providing a comprehensive decentralized data registry. Something else is needed.

A Data Registry How To

A data registry is a comprehensive and inclusive list of what data is kept where and why. A data registry will need to take on certain characteristics if it's meant to satisfy both data privacy, protection, and governance goals.

Some of these characteristics are functional:

1. **Content Granularity**

   You can't protect what you can't find, so security of data requires detailed knowledge of what data you keep where. You can't protect the social security data you keep if you first don't know that you keep 25 instances of that data in 25 places. Moreover, privacy compliance requires every organization to account for what data they collect and use for every individual. This requires not only that an organization knowsknow what type of data they collect and process, but also whose data they collect and process. Privacy is about people, so knowing the "people" context of data is essential to meet privacy use cases.

2. **Usage Context**

   Beyond just knowing what data is kept where, gathering data intelligence requires context as to who, what and why the data is used. This requires operational, technical and business knowledge such as who can access this data; what applications are consuming this data; with what 3rd parties is this data shared; for what purpose are we collecting this data; do we as an organization have adequate consent to collect or process this data.

3. **Data Source Coverage**

   A registry that only covers unstructured file shares can't provide a complete data inventory or map. A registry that only covers relational databases similarly gives inconclusive data intelligence. As data sources and applications proliferate inside the enterprise, enterprises wanting to get complete knowledge of the data they collect, and the process will need complete coverage that spans unstructured file shares, structured databases, Big Data, cloud, NoSQL, logs, mail, messaging, applications and more.

Some of these characteristics are operational:

1. **Decentralized**

   A registry should not be a warehouse. Companies don't want registries that are duplicates of the data they map. Keeping the registry more of an index-like map minimizes cost, complexity, and potential security hazard.

2. **Scale**

   Organizations increasingly keep tens if not hundreds of petabytes of data. The registry needs to provide an efficient index of what data is kept where along with associated usage context in a way that can scale to a global enterprise.

3. **Dynamic**

   Data is not static. Its get created, changed and moved on a constant basis. The registry, therefore, needs to be self-updating, accommodating to change, and near real-time in providing a picture of what data is kept where, and when.

To meet these functional and operational requirements, a modern data registry will need a fresh approach to gathering and surfacing intelligence about what data an organization collects and processes. This requires a fresh approach to discovery that supersedes either classification or cataloging alone.

A New Approach to Building a Data Registry from Data Intelligence

For a data registry to provide a full accounting and inventory of an enterprise's distributed data assets the enterprise will need attribute-level detail. This requires data intelligence down to the discrete entity value, a level not possible with metadata alone. But getting to an attribute or entity level view of data without losing the residual context information contained in technical, operational and business metadata requires a hybrid approach for content discovery and contextualization. This can be summarized as follows

Entity Discovery & Resolution:
To get the kind of data intelligence necessary for privacy and protection use cases, organizations need to be able to identify what data entity is where. This requires a data discovery mechanism that can extract and resolve data entities based on data values whether the data resides in structured, unstructured or semi-structured stores. This also requires the ability of the scanning system to disambiguate identical looking data based on context (for example separating a social security number from an account ID even though both may have the same value).

Entity Correlation and Contextualization:
Privacy introduces the need to understand data entities in the context of whom the data belongs to. Privacy is about people, and the emerging privacy regulations like the General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA) require organizations to account for what data they hold on what individual. This requires a new form of data intelligence not previously required in security or governance tools: the need to correlate data to a data subject entity ie a person. Therefore a modern data registry will need to also show correlation or association of data to an entity like a data subject. While this is essential for privacy, this can also provide a new dimension for understanding the connectedness of data to high-value identities like transactional IDs, account IDs, patent IDs or similar.

Entity Classification By Type and Category

A data registry should indicate what type of data is kept where. Data governance tools have traditionally determined "type" from metadata captured in a metadata catalog. However, metadata is often inconsistent and many times inaccurate. Moreover as has already been explained a contemporary data registry needs to have entity level granularity requiring more refined entity level classification. Traditional classification tooling, as has been noted, relies on regular expressions for pattern detection. Modern approaches need to move beyond this and build on AI or ML to expand how data can be identified based on heuristics and inferred categorizations. But immaterial of how the data is classified, an ability to classify and categorize an entity at scale is essential for delivering on a vision of entity-level registry.

Metadata Capture and Cataloging

While pure-play metadata catalogs leave much to be desired from a registry perspective they still provide value in that they can locate where categories of data can be found. This is useful in both helping to classify data entities correctly and identifying places to prioritize a deeper entity search. The challenge with metadata catalogs, however, is that rely on human tags and annotations making them often inconsistent from one store to another, language dependent, human error prone and limited in coverage to data sources where labels or annotations exist. Moreover, today's metadata catalogs limit themselves to technical metadata, missing an opportunity to provide usage context by capturing operational and business context like access rights, purpose-of-use, or consent.

Modern Data Knowledge Begins With A Data Registry

You can't know your customer unless you first know their data. Data is the digital currency of the modern enterprise, and yet organizations lack detailed insight into what data they collect and process. This creates problems for data compliance, security and ultimately extracting value from data.

For a decade or more, organizations made do with more narrowly defined attempts at data discovery to meet requirements like PCI or HIPPA compliance. Similarly, to gain a picture of their customers they built Master Data Management (MDM) records that amalgamated information from ten or so data systems vs the thousands most organizations actually possessed. However these early efforts at lighting-up dark data are increasingly inadequate to meet new regulations, security complexity and data governance needs like preparing data for BI and AI.

New privacy regulations like GDPR and CCPA are forcing companies to examine their data in ways they have not previously had to. Privacy regulations revolve around knowing what data an organization collects and processes on every individual or data subject. This requires an ability to understand every element (or entity) of data content by type and person, and to understand the usage context of that content in terms of access, purpose-of-use, and consent. And since no privacy regulation limits itself to just unstructured or structured data, it means that companies need an ability to find, inventory and map their data assets across any kind of data source. Neither classification-based data protection tools nor catalog-centric data governance tools allow for complete coverage.

BigID's Third Way: Correlation, Classification, Cataloging

BigID's Big Idea is to introduce a third way to interrogate data to gather detailed knowledge around its content and content over a wide data source coverage area. Before classifying or cataloging data, BigID uses advanced ML to pick out what data is connected to what entity. This smart correlation, allows BigID to first find all the data associated with an entity like a data subject. This simplifies scanning across any data source and reduces a hundreds of petabyte problem to a more manageable terabytes problem. This patented data discovery approach gives BigID the unique ability to know what data is correlated to what entity. Only after correlation is the data classified and the metadata collected to provide a complete 360 view of content and context. Using BigID organizations don't need to abandon classification or cataloging. They merely add a new dimension for understanding and mapping their data across all their data sources.

Using BigID an organization can get a complete picture of their data without copying or duplicating it. The resulting registry combines the best of correlation, classification, and cataloging but over any data source either in the data center or cloud.