White Paper

Solving DSARs’ Big Data Problem

Four recommendations plus the one thing you should never do
Table of Contents

Addressing Data Subject Rights is a Big Data Problem .......................................................... 3
Five Critical Data Subject Request (DSR) Fulfillment Capabilities ..................................... 4
Why is identifying data subjects and their sensitive data so complex? ............................... 6
  Master data management to the rescue? Not so much. ......................................................... 6
  Privacy’s universal data subject view ................................................................................. 7
Solving Data Subject Rights’ Big Data Problem: Four Recommendations, Plus the One
Thing You Should Never Do .................................................................................................. 8
  REDUCE YOUR PI SURFACE AREA: A THREE-STEP APPROACH ..................................... 8
  Step 1: Use sampling scans to discover which systems contain personal information .............. 8
  Step 2: Use deeper scans to identify the tables that contain personal data and any data handling issues . 10
  Step 3: Remediate data handling issues ............................................................................ 11
  PREPARE FOR A DSR “DENIAL OF SERVICE ATTACK” .................................................. 11
  ADHERE TO THE DATA HANDLING BEST PRACTICE OF DE-IDENTIFICATION ............. 12
  COMPLY WITH PRIVACY AND SECURITY BY DESIGN PRINCIPALS ............................... 13
  Apply the Concepts of Identity Resolution to Improve Accuracy ........................................ 14
  NEVER MAKE THE PROBLEM WORSE BY CREATING ADDITIONAL COPIES OF CUSTOMER DATA ................................................................. 15
About Integris Software ........................................................................................................ 16
Integris DSR Solution Brief .................................................................................................. 16
Addressing Data Subject Rights is a Big Data Problem

Consider the following:

- A single bank transaction may get replicated across 100 systems.
- Storage is so cheap that enterprises collect petabytes of data each year and keep almost all of it.
- Data is routinely propagated across the enterprise to support a wide variety of users and business initiatives.

Unfortunately, the massive growth in data collection and proliferation has not been accompanied by an equally matched effort in data management and governance.

The consequences have been painful. Data breaches. Misuse of private data. Loss of consumer trust. In response, companies have poured resources into implementing security controls to block or restrict access to their data. But whereas Security is focused on who is using the data, Privacy is about how the data is being used and for what purpose.

Meanwhile, regulations like GDPR and CCPA are obligating companies to respect and respond to data subject rights. But achieving basic compliance requires that companies understand what personal information they have, where it’s located, and its purpose. Up until now, the basic data inventory process has been a manual one consisting of application data owner surveys and spreadsheets.

DSRs push these manual processes to their breaking point. Not only in people resources required to manually search those 100 systems in the bank example for each DSR, but also in the accuracy and completeness required to be defensible with the regulators. It’s a big data problem and a new approach is required to process petabytes of data, extract key data points and derive the relationships between them. Meanwhile, companies have been left scrambling to meet their obligations.
Five Critical Data Subject Request (DSR) Fulfillment Capabilities

The five critical Data Subject Request process and fulfillment capabilities are intake, verify, search, deletion, and response. DSR fulfillment is critical being in compliance with both the California Consumer Privacy Act (CCPA) and the General Data Protection Regulation (GDPR). While CCPA and GDPR have their own unique take on DSR fulfillment, these five critical capabilities are a must:

1. **Intake**

   During intake, a data subject makes a request via email, an online form, or other communiqué. The enterprise then needs to verify the requestor’s identity and existence within the data ecosystem and track the request fulfillment through to resolution. All within the required timelines (30-45 days depending on the regulation).

2. **Verify**

   The next step is verification of the identity of the requestor. For companies which provide services online, this step may require customers to login and verify their identity. For regulations like GDPR, which may include employees and vendors, this requires that the enterprise confirm the existence of the data subject anywhere in their ecosystem and then identify corresponding info to include in the response.

3. **Search**

   In order to fulfill the request, the enterprise will need to locate a requestor’s personal data by searching across its data ecosystem. The type of information the enterprise will be searching for will differ based on data subject type. For example, is the data subject a current customer or a former employee? CCPA only applies to ‘California consumers’ whereas GDPR also includes employees and contractors (privacy by design would look to encompass current and potential future scenarios). The search process identifies relevant PI attributes, categories, and the company’s purpose for collecting and processing the subject’s information. The search then needs to identify the specific systems and locations that contain the data subject’s personal data.
4. Deletion

For deletion requests, the enterprise will need to validate which systems the data can be deleted from, based on regulatory or business constraints. An example of a business constraint might be a warranty registration database that contains personal information. The enterprise cannot delete customer information from this database because it impedes the ability to fulfil a legal obligation to provide a customer with, say, an extended warranty on his purchase.

Next, the enterprise will need to initiate a process to delete or obfuscate the customer's data from the relevant systems, as well as request the same from third-party data processors. Lastly, the enterprise will need to audit and confirm the deletions.

5. Response

Templates help ensure an efficient and consistent fulfillment data subject request process. All communications and activities should roll into a reporting dashboard and audit trail to demonstrate accountability, compliance, and progress towards resolving requests.

Which of these five capabilities is the most challenging?

For many organizations, the most complex, tedious, and resource-intensive step in the process is finding PI and tying it back to the data subject.
Why is identifying data subjects and their sensitive data so complex?

Not only has data proliferated, but it’s also mutated into derivative forms. Customer data is often collected across multiple channels without being linked to a master identifier. Also, when downstream systems aren’t updated there can be discrepancies between primary and secondary systems.

To make matters worse, both the regulatory environment and what’s considered sensitive data is changing. CCPA defines personal information that “could reasonably be linked, directly or indirectly, with a particular consumer or household.” The word “household” is not found in GDPR. It implies that personal information does not have to be tied to a specific name or individual (think home address, home devices, geolocation data, home network IP addresses, and the like).

Resolving identities across hundreds of sources is a data processing and data quality nightmare. The vast majority of companies simply do not have the tooling in place to access and monitor the volume, variety, and velocity of personal data flowing in, out, and across their organizations.

Master data management to the rescue? Not so much.

Many medium and large enterprises have implemented master data management systems (MDM) to resolve identities and create a golden record for interacting with a customer. MDM and customer data platforms hold the promise of delivering a 360-degree view of the customer to improve sales, service, and growth.

However, “customer” is often defined in different ways across an enterprise and that definition does not always equate to an individual. Also, data subjects can look different across data sources and business scenarios because of:

- Nicknames
- Middle initials and suffixes
- Maiden names
- Different email addresses, phone or postal addresses
- Address changes
- Typos in addresses or abbreviations

Even when companies build master data management processes, they typically identify a few trusted sources from which to provide inputs.

And of course, not all personal data is tied to a user ID. Even without an ID the individual can still be identified in a data set. By simply mapping IDs to pre-existing metadata, the enterprise can run
the risk of creating a false sense of security about the data it has, which security parameters are being applied, and whether it is in compliance with regulatory mandates.

Finally, while CCPA applies to California consumers only, GDPR applies to all data subject types such as customers, employees, vendors, and partners.

Privacy’s universal data subject view

MDM and other customer platforms and processes are not tuned to a ‘universal data subject’ view which includes multiple data subject types.

The result is that companies have to either enhance their mastery of ‘the customer’ to include other data subject types or build a new identity resolution process. The business case for embarking on such projects has not been compelling up until CCPA and GDPR. Outside of privacy, there aren’t any business use cases that require consolidating massive amounts of personal information on different types of data subjects.

Most enterprises simply do not have the intent or willingness to spend valuable resources and embark on an expensive journey to create a master data subject view.
Solving DSARs’ Big Data Problem:
Four Recommendations, Plus the One Thing You Should Never Do

Reduce Your PI Surface Area: A Three-Step Approach

To help solve the inherent nature of the DSR big data problem we recommend a three-step approach for finding your PI and tying it back to your data subject. The end result of the three steps described below is that you’ll dramatically reduce both the surface area for your DSR fulfilment process, as well as what would otherwise have been in scope for CCPA, GDPR, and other regulations.

Step 1: Use sampling scans to discover which systems contain personal information

Solutions that attempt to scan large swaths of your data ecosystem will either keel over or get stuck in a protracted scan. It simply takes too long to do comprehensive data subject searches right out of the gate, especially when you’ve got 30-day timelines to meet. To solve this bottleneck, reduce your surface area by creating an inventory of personal information, regardless of the data subject. This helps you identify your high-risk systems, locate PI, and detect classification and labeling issues.

There’s no need to boil the ocean right out of the gate. Start with a light sampling scan, and based on what you find, you can follow-on with deeper scans.
Why sample? Because even with best-in-class architecture, deep scanning takes time. Consider the example below which compares how most vendors handle PI scanning compared to Integris’s sample scanning approach.

<table>
<thead>
<tr>
<th></th>
<th>Competitor Full Scan</th>
<th>Integris Sampling Scan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data environment</td>
<td>1 petabyte (1,000 TB)</td>
<td>1 petabyte (1,000 TB)</td>
</tr>
<tr>
<td>Type of scan</td>
<td>Identify problem</td>
<td>Describe the solution.</td>
</tr>
<tr>
<td>Scan rate per terabyte</td>
<td>Deep crawling</td>
<td>Sampling</td>
</tr>
<tr>
<td><strong>Time to scan 1,000 TB</strong></td>
<td>12,000 hours (500 days)</td>
<td>167 hours (7 days)</td>
</tr>
</tbody>
</table>

**Note:** Illustrative and based on structured data sources. YMMV based on number of systems, data volumes, and sampling types selected.

Working 24/7, the competing approach can take 500 days to do a deep scan of one petabyte of enterprise data. With a sampling scan approach, we can finish scanning that same one petabyte in a week.

The competing approach is to do a deep scan to find and map PI back to a set of IDs. In 90 days, the competing approach might only get through a few systems (the small outlined area in the upper left on the graphic below). In that same time period Integris has done a much more complete scan of your data ecosystem.
Next, we reduce the haystack down to the systems where we discovered PI. As illustrated below, we expected to find PI in our customer, vendor, and employee systems. But we also found PI in our Business Intelligence systems.

Step 2: Use deeper scans to identify the tables that contain personal data and any data handling issues

In addition to the timely fulfillment of data subject requests, regulations like GDPR and CCPA also require good data handling practices. Continuous defensibility to meet compliance requirements boils down to understanding where your sensitive data resides across all data sources, and then mapping that data back to your data handling obligations.

To continue with our example, we expect to see CCPA data in our customer, employee, and vendor data systems. But as we dive deeper into these systems, we notice that:

- We’ve got data that’s supposed to be encrypted, but it’s in clear text.
- We find a trove of data that’s been mislabeled; its sitting in the wrong part of a database.
- Within many of the systems that we know contain CCPA data, we find retention issues.

These data handling issues present a challenge, because when we execute our DSR process, we’re going to want to limit the number of tables to those that we know contain sensitive data. We want to make sure the data is where it’s supposed to be.
Step 3: Remediate data handling issues

Continuously monitor your sensitive data against your data handling policies and raise these issues so that you can take the appropriate actions. Fixing these issues has the added benefit of further reducing your DSR surface area.

Continue to drill down into your systems to see which attributes, or types of PI elements they contain. And from there, you should drill all the way to the data element level, which allows you to be precise in your remediations.

Ideally, you’ll want to tie directly into your broader ticketing and InfoSec ecosystem to tell other tools and/or people where to go to remove, encrypt or minimize a data set.

Prepare for a DSR “Denial of Service Attack”

If you get flooded with thousands of DSRs at once the impact is a denial of service attack that overwhelms your CSR and IT staff. Under this scenario, your manual processes reach the breaking point and you can’t respond to requests within the required timelines (usually 30 to 45 days depending on the regulation).

Consider solutions that will allow you to automate your DSR processes so you can fulfill thousands of requests automatically. Today’s DSR solutions need to be completely API driven and generate detailed metadata. Detailed metadata like system location, owner, business metadata, and system classification is crucial to support a variety of deletion workflows.

You need to be able to find exactly where the data subject resides within your ecosystem, then trigger deletion and portability workflows. Of course, you’ll want an audit trail of all deletion activities including deletion confirmations for internal audits and compliance needs.

The end goal is to decrease the pressure on IT teams to find data subjects and validate deletions.
Adhere to the Data Handling Best Practice of De-identification

De-identification prevents data analysts from connecting an individual to their personal information. This enables the data analyst to access useful data without compromising customer privacy. Many organizations are de-identifying data for continued analytics after a right-to-forget (as shown in the illustration below).

But not all discoverable sensitive information is linked to an identity. In fact, 87% of the US population can be identified using only their Zip Code, Gender, and Birthdate. Each of these data points is benign on its own, but when combined become toxic. Data lakes, data warehouses and other systems that support data analysis ingest disparate pieces of customer data from a variety of sources. When combined, this data has the potential to reveal customer identities along with highly sensitive personal information (as illustrated below).
That’s why it’s critical to inspect down to the data element level to inform you exactly what’s in your data lake, not just what the metadata implies. When you operate at the data element level you can also identify highly sensitive combinations of data across your data ecosystem. For example, capabilities like multi-label classification (aka correlated labeling) can validate that a dataset can’t be used to identify an individual (as illustrated below).

Comply with Privacy and Security by Design Principals

Any DSR fulfillment process and associated systems must comply with privacy and security by design principles.

In August 2019, the BBC reported that a security expert contacted dozens of UK and US-based firms to test how they would handle a "right of access" request made in someone else's name. In each case, he asked for all the data that they held on his fiancée. In one case, the response included the results of a criminal activity check. Other replies included credit card information, travel details, account logins and passwords, and the target's full US social security number.
Annie Bai and Peter McLaughlin’s IAPP article sounded the alarm bell in that “The terrible beauty of the California Consumer Privacy Act is that innumerable companies will soon be required to undertake totally novel consumer-facing responsibilities...It is a new door for improper data access — not a back door, but an actual, legit front door — for fraudsters to obtain all manner of valuable personal information.”

Companies in highly regulated industries like financial services already have sophisticated ID verification systems in place. If your firm doesn’t have one of these systems in place, then you may want to explore adding this capability into your DSR workflow (Evident ID is one such system).

Another vulnerable area is the personal information that a CSR may have access to when they are responding to a DSR. To protect sensitive information, it’s best to obfuscate PI from the UI so CSRs can’t see sensitive data as they respond to DSRs.

**Apply the Concepts of Identity Resolution to Improve Accuracy**

It’s also important to apply the concepts of identity resolution to identify your data subjects across multiple data sources. Why? Data subject information changes over time and your data subjects may use different information in their interactions with your company (e.g. nicknames, maiden name, address changes, initials, and Jr./Sr.).

Upon receiving a data subject request, it may be helpful to run a quick data subject search to confirm their existence within your data ecosystem. This helps validate that your data subject exists and also provides instant access to additional attributes that help disambiguate the data subject (e.g. John Smith and John Smith Jr. living in the same household with the same address and family email). This can also tell you if the data subject exists in various data subject types, such as customer, employee, or vendor in your data ecosystem.

For example, when John Smith submits his request you can quickly confirm that he exists, identify John Smith the customer vs John Smith the employee, and look up typical attributes found across your data ecosystem. Based on the privacy by design principles, this information is adequate for most enterprises to respond to access requests.
Never make the problem worse by creating additional copies of customer data

Finally, never make the problem worse by creating additional copies of customer data or violating your own security policies by moving sensitive data between secure network zones. Not only are copies of data inherently outdated, but they exacerbate data sprawl, and open you up to additional security risks. We think Paige Bartley, Senior Analyst at 451 Research said it best:

“Modern data privacy and protection regulations such as GDPR and CCPA have inadvertently created a paradox in which organizational attempts to fulfill data subject requests (DSRs) can occasionally result in the exposure of sensitive data to additional unnecessary parties: particularly when duplicate data is generated.

Integris Software’s privacy-by-design DSR methodology minimizes this risk, and provides a highly-defensible approach to meeting data privacy requirements via its ability to pinpoint sensitive data – regardless of whether or not it is tied to a user ID – across data sources both in-motion and at-rest.”

Paige Bartley
Senior Analyst, Data, AI & Analytics
451 Research
About Integris Software

Integris Software, the global leader in data privacy automation, helps enterprises discover and control the use of sensitive data in a way that protects privacy and fuels innovation.

Privacy is now critical to an effective data protection strategy. By sitting upstream from security, Integris tells you what data is important and why so you can be precise in your InfoSec controls.

Integris works securely, at scale, no matter where sensitive data resides. You get a live map of your sensitive data where you can apply policies, surface issues, fulfill data subject access requests, and automate remediations via your broader ticketing and InfoSec ecosystem.

Regulations like GDPR and the California Consumer Privacy Act (CCPA) are triggering knee-jerk reactions as companies lock down their data for fear of misuse. With Integris, there is finally a way to use your data without fear.

For more information on Integris, visit: www.integris.io or follow @Integrisio on Twitter.

Global HQ in USA
1525 4th Avenue, 5th Floor
Seattle, WA 98101
Phone 1-425-539-2145
info@integris.io

Vancouver, BC Canada Office
450 Southwest Marine Drive
Vancouver, BC
V5X 0C3
Phone 1-425-539-2145
info@integris.io

For an overview of Integris DSR, please see the solution brief that follows.
Integris Software tackles the most challenging aspect of data subject requests (DSR) – finding data subjects across your data ecosystem. By automating the discovery and classification of sensitive data, Integris reduces the burden on IT teams and data source owners.

Deep Search and just-in-time identity matching make it easy for your customer service reps (CSRs) to verify and locate data subjects across thousands of systems. CSRs can capture the request, input data into a standard response template, and share it back out with the data subject. They can preview DSR reports, add private notes, and activate the next step in your workflow.

Worried about a flood of DSRs causing a “Denial of Service” on your CSR and IT staff? Integris gives you the ability to automate your DSR processes to fulfill thousands of requests automatically.

Integris DSR Key Capabilities

**PI Surface Area Reduction**
Our discovery process isolates your systems that contain PI, then maps attributes, categories, purpose, and sources back to each data subject.

**DSR Lifecycle Management**
DSR intake, workflow management, and response generation, as well as integrations with your existing front-end systems.

**Data Subject Validation**
On intake, we help you confirm a data subject exists within your ecosystem and identify multiple types like customer, employee, and vendor.

**Data Subject Deep Search**
We identify the data subject’s relevant PI, as well as the specific systems, tables, and files that contain the subject’s personal information along with purpose, PI categories, system owner, and related information.

**Remediation and Validation**
Additional metadata and event orchestration support end-to-end workflows (e.g. deletion), and you get an audit trail with validation to demonstrate compliance.

The Only DSR Solution that Doesn’t Require Creating Copies of Your Customer Data

Fulfill Data Subject Requests with Unmatched Speed and Accuracy

Integris Software Delivers the Fastest Path to DSR Defensibility

1525 4th Avenue, 5th floor Seattle, WA 98101-1607 USA
+1 (206) 539.2145
sales@integris.io
@integrisio
integris.io
Integris DSR is fast, accurate, and follows the principles of privacy by design

"Modern data privacy and protection regulations such as GDPR and CCPA have inadvertently created a paradox in which organizational attempts to fulfill data subject requests (DSRs) can occasionally result in the exposure of sensitive data to additional unnecessary parties: particularly when duplicate data is generated.

Integris Software’s privacy-by-design DSR methodology minimizes this risk, and provides a highly-defensible approach to meeting data privacy requirements via its ability to pinpoint sensitive data – regardless of whether or not it is tied to a user ID – across data sources both in-motion and at-rest."

Paige Bartley
Senior Analyst, Data, AI & Analytics

**Fast**

Build trust by delivering a rapid response to requests that include up-to-date PI categories and purpose.

Search for data subjects only where you might find them; reducing the PI surface area upfront enables blazing fast, on-demand data subject deep searches.

Future proof your business against DSR “Denial of Service attacks” by automating your end-to-end DSR workflows.

Decrease your dependency on IT teams to find data subjects and validate deletions.

**Accurate**

Advanced identity matching enables high confidence search and identification of data subjects.

Get up-to-date details on the data subject when you need it; stop relying on historical indexes.

Discover all PI even if its not tied to a user ID; machine learning and contextual awareness operate at the data element level on data anywhere - at rest, in-motion, in the cloud, or on-prem.

Find exactly where the data subject resides within your ecosystem, then trigger deletion and portability workflows.

**Privacy by Design**

Reduces the DSR data surface area through inventory risk assessment and remediation.

Enterprise ready; multi-zone support with no need to replicate consumer data across network zones.

Reduces exposure to identity thefts; respond to access requests confidently with attribute types, PI categories, and purpose.

By default, obfuscates PI from the UI so customer service reps can’t see sensitive data as they respond to DSRs.