Informatica

White Paper

# Intelligent Data Privacy

Discover, Classify, and Protect Personal and Sensitive Data

**About Informatica**

Digital transformation changes expectations: better service, faster delivery, with less cost. Businesses must transform to stay relevant and data holds the answers.

As the world’s leader in Enterprise Cloud Data Management, we’re prepared to help you intelligently lead—in any sector, category or niche. Informatica provides you with the foresight to become more agile, realize new growth opportunities or create new inventions. With 100% focus on everything data, we offer the versatility needed to succeed.

We invite you to explore all that Informatica has to offer—and unleash the power of data to drive your next intelligent disruption.

## Table of Contents

## Executive Summary

With global privacy regulations, explosive growth of personal data, escalating data losses, and growing customer expectations, organizations must automate and optimize their intelligence and protection of personal and sensitive data. With strict regulations and heightened enforcement[1], organizations can face staggering costs for non-compliance of privacy regulations or breaches[2]. They must develop data privacy and security governance policies and strategically integrate the enforcement and communication of these policies with a comprehensive protection solution.

Protecting expanding data environments calls for data security that uses next-generation tools to discover, classify, and monitor data movement and access; map data subjects and identities; continuously analyze and track sensitive data risk; automate and orchestrate the dynamic application of security controls; and effectively communicate the status of data privacy actions. To deliver more proactive and intelligent sensitive data risk monitoring, investment prioritization, and protection, these tools must include the following capabilities:

- Privacy policies that link business and technology processes to meet regulatory requirements
- Automated and integrated sensitive data discovery, proliferation analysis, AI-driven detection of anomalous user activities, and multifactor data risk analytics
- Mapping identities to sensitive data to support data subject access requests and integration with consent management systems
- Actionable intelligence that leverages data-centric security to support the orchestration of remediation in a single platform and respond to data subject rights and consent requests
- Broad coverage across today's complex, on-premises and multi-cloud environments

This paper explores how these capabilities reduce the risk of data loss and misuse, while improving compliance with data privacy regulations such as GDPR, HIPAA, CCPA, and FINRA, especially when provided by a comprehensive single platform that also lowers total cost of ownership. In addition, the paper describes what organizations should consider when evaluating functionality in these six critical areas:

1. Managing data governance policies
2. Discovering and classifying personal and sensitive data
3. Mapping individual identities
4. Analyzing and tracking risk
5. Protecting personal and sensitive data and managing subject requests and consent
6. Communicating privacy actions and status

Finally, this white paper addresses what it takes to implement a data privacy solution and highlights the benefits of a properly deployed data-centric security solution.

[1] Breach Level Index (https://www.breachlevelindex.com/) reports an 87.5% increase in breached data records in 2017.
[2] GDPR fines for noncompliance can be up to $20 million or 4% of annual revenue.

## Growing Data Volumes and Privacy Regulations

By 2025, the global data volume is predicted to be 10 times the size of data generated in 2016[3]. This exponential data growth presents a host of challenges:

- Diversity in the form of structured, semi-structured, and unstructured data
- Proliferation as data moves around the globe from data stores, including Hadoop nodes, cloud instances, file servers, and relational databases
- More users from more locations, functions, and geographies
- Exponential growth in onsite, remote, and mobile access points

This data growth is further complicated by the emergence of national and regional privacy regulations across the globe, with differing requirements for compliance. This wave of privacy legislations will require organizations to precisely, and continuously, understand and manage personal and sensitive data.

To achieve the enterprise data intelligence necessary to comply with changing global privacy regulations, organizations can no longer rely on siloed and legacy approaches to finding, analyzing, monitoring, and protecting sensitive data. With the threat of sizable fines and the need to maintain customer trust, they urgently need to shift to a new data security paradigm that takes a holistic approach with:

- Centralized data governance management that efficiently handles policy changes and compliance reports
- Continuous sensitive data monitoring and risk analysis that makes it easier to prioritize security programs and investments
- Identity-based intelligence, which provides global and granular analytics of sensitive data based on identity to support data subject access requests and integration with consent management
- Rich, interactive visualizations that provide a complete understanding of data movement both inside and outside the enterprise and between partner and client organizations
- Integrated protection that connects discovery, risk, and monitoring to automated remediation

## Key Requirements for Data Privacy

Chief Information Security Officers (CISOs), enterprise architects, and privacy officers know they need continuous, contextual, and responsive protection for enterprise data assets to spot new threats as they arise. However, to achieve this goal at enterprise scale, they must tackle several challenges, discussed below.

### Standardization, Consistency, and Auditing

To protect thousands of data stores, organizations need to standardize data definitions and associated protection and governance policies. Without centralized policy management, security practitioners and architects cannot consistently manage policy changes or audit and track compliance.

[3] IDC, "Data Age 2025: The Evolution of Data to Life-Critical," David Reinsel, John Gantz and John Rydning, April 2017.

### Accuracy and Effectiveness

To maintain a high, reliable level of accuracy in a constantly changing environment, a data privacy solution requires data context, subject identity mapping, user behavior analytics, and risk analysis. Without them, the solution will generate a high number of false positives, thus rendering itself ineffective at continuous private data protection.

### Scalability and Time to Value

Achieving risk reduction within tight deadlines calls for an automated data privacy solution that can scale to protect many data stores in far less time than it would take to protect each data store individually. This is a pressing concern for security and compliance officials who must respond immediately to potential data security risks. In fact, Gartner predicts "The amount of time it takes for security teams to detect and respond to excessive risk (with a goal of preventing or minimizing the financial impact) will be one of the most critical security metrics over the next decade."[4]

Without intuitive usability, orchestration of actions, workflows, artificial intelligence (AI), and automation, a data privacy solution can neither scale nor deliver fast enough time to value.

### Continuous Protection Without Interruption

A data privacy solution must be flexible enough to maintain data protection and compliance insights in an environment where data, usage, users, and regulations are in constant flux. If a solution cannot support policy customization and regular reassessment of risk, it cannot provide truly continuous protection of private data.

### Out-of-the-Box Deployment Readiness

Not only do today's CISOs and privacy officers need a data-centric approach to security with new capabilities driven by automation and AI, they also require the ability to derive more value from current security controls wherever possible. Any addition to the security ecosystem should be API-driven for easy integration with existing enterprise security controls, such as SSO, password management, Ranger™, Sentry™, ServiceNow®, shield, encryption, and tokenization solutions.

### Support for On-Premises and Multi-Cloud Environments

Organizations in every industry, from financial services and health care to transportation and entertainment, are at some stage of digital transformation. At this stage, their world is a complex one. Any data protection solution they choose must cover all their data sources, whether managed in traditional on-premises systems, multi-cloud environments, or big data anywhere—consisting of either structured or unstructured data formats.

[4] Gartner, "Seven Imperatives to Adopt a CARTA Strategic Approach," Neil MacDonald, 10 April 2018.

## Evaluating Data Privacy Solutions

To protect personal data, organizations must start with the basics:

1. Setting policies for properly handling sensitive data
2. Understanding where sensitive data resides and consistently classifying it
3. Mapping individual identities to sensitive data
4. Analyzing and tracking data risk
5. Protecting sensitive data and remediating risk
6. Measure, communicate, audit response

To support these basic requirements, a platform that manages sensitive data must include the following functional areas:

### Define and Manage Governance Policies

This capability defines sensitive data elements and the process for protecting them in compliance with all applicable regulations and laws.

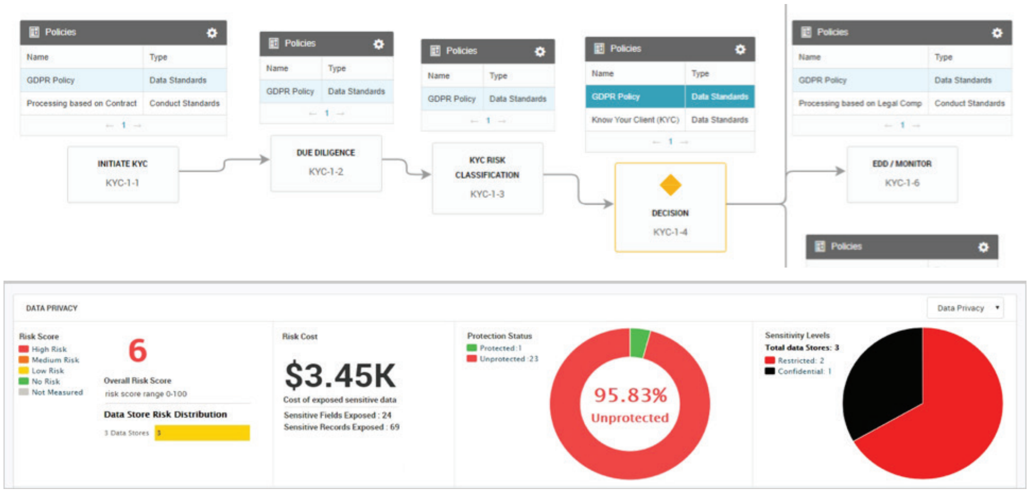
Many data governance programs are born out of the world of regulatory compliance. For organizations, regulations provide a required purpose for the program, while also providing the executive support and funding to grow the program. Often, new data governance programs need to start small, taking on one project at a time to work through the policies, processes, and procedures for their new program. For compliance, this means that programs focus on the groundwork that's needed around a privacy regulation, while also looking to the future of holistically governing data as a high-value enterprise asset as the program grows.

While setting up their new program or working through the guidelines of a new privacy regulation, data governance program leaders will look to identify several key items. They'll need to build out high-level, internal policies such as who can safely access the data and how team members will work together with the data across organizational boundaries. These same program owners will also build out a business glossary or data dictionary. The glossary will become the single point of truth for a data element (e.g., defining for all team members what exactly a "customer" is).

In practice, this means that organizations will be empowered to collaborate across business functions to capture and align policy definitions and requirements, document the processes for responding to subject registry requests, as well as determine and document a consent management framework. Then, assign ownership and accountability for enforcement, management, response, and remediation of each data governance and privacy policy.

In addition, consider looking for solutions with the ability to:

- Manage policies as part of an integrated platform, rather than a separate application
- Map system and process flows for sensitive data
- Visualize privacy and protection outcomes in context
- Report outcomes to all affected stakeholders
- Provide an audit trail over all governed artifacts
- Automate responses to data subject requests at the point of interaction



Policies

| Name | Type |
|---|---|
| GDPR Policy | Data Standards |
| Processing based on Contract | Conduct Standards |

Policies

| Name | Type |
|---|---|
| GDPR Policy | Data Standards |

Policies

| Name | Type |
|---|---|
| GDPR Policy | Data Standards |

Policies

| Name | Type |
|---|---|
| GDPR Policy | Data Standards |
| Know Your Client (KYC) | Data Standards |

Policies

| Name | Type |
|---|---|
| GDPR Policy | Data Standards |
| Processing based on Legal Comp | Conduct Standards |

Figure 1: Sample data governance definitions mapped in the Informatica Data Privacy platform with graphical summaries of data privacy, risk scores, risk cost, protection status, and sensitivity levels.

### Discover, Classify, and Understand Personal and Sensitive Data

This capability automatically locates, ranks risk, and classifies sensitive data to provide a broad and deep understanding of where it resides and how it proliferates throughout an organization. Traditional data discovery and classification tools identify databases that contain sensitive data by listing the table/column names where that data resides. While this approach offers some insight into where protection may be needed, it does not provide any guidance on where to start, where the sensitive data proliferates through other data stores, or indicate whether the data has already been protected.

### Scan Data Stores to Discover Sensitive Data

The solution should meet an organization's specific needs around scalability and it should be able to scan across multiple vendor platforms and different data repositories, such as:

- Structured data across traditional relational databases, including mainframes
- Cloud applications
- Semi-structured data (e.g., CSV, XML, JSON) on HDFS and Amazon S3
- Unstructured data on CIFS NFS
- SharePoint
- Traditional structured data stores

The solution should also be able to import data protection methods and status from data stores that are already protected.

In addition, consider looking for these features to ease administration and scaling:

- Agentless scanning
- Ability to schedule scan jobs
- False positive detection and handling
- Ability to configure scans to use metadata, data, or both

### Classify Sensitive Data

This capability involves creating and modifying search definitions for sensitive data domains, such as credit card numbers, addresses, names, or other personal information. Classifying sensitive data also includes the ability to create custom classification policies as necessary to comply with privacy regulations.

Ideally, the solution should come with out-of-the-box data domains for compliance with major regulations such as GDPR, HIPAA, PII, PCI, and PHI, as well as the ability to create custom domains and classification policies to cover additional requirements. It should also provide the abilities to:

- Leverage classification policy templates to quickly create private data definitions and specify the data's sensitivity level
- Access all data domains without needing to import them
- Create search definitions for both metadata and data, using pattern matching (including regular expressions), reference tables, and complex combination rules (e.g., only match if Name + Address + SSN are present)
- Create advanced search rules, such as the ability to do logical checks against data (e.g., check sums) or rules beyond regular expressions, such as prioritization rules when search results conflict
- Minimize false positives through conformance scoring (e.g., only accept if > xx% of selected rows match), blacklists, and whitelists
- Estimate the cost of data loss for each row that matches a classification policy

### Understand Where Sensitive Data Proliferates

Data proliferation involves determining and tracking where each sensitive data record also exists in other data stores throughout an organization and providing drill-down capabilities to understand, monitor, and protect the record wherever it proliferates.

Ideally, the solution should come with out-of-the-box tools to:

- Track sensitive data movement through data stores from third-party metadata providers (e.g., Microsoft, Cloudera, Hortonworks)
- Ingest custom proliferation data
- Provide rich visualizations of an organization's global sensitive data and its proliferation patterns across geographies and departments

- Continuously update a ranked list of the identities, data stores, data domains, and departments with the most activity on sensitive records
- Provide a drill-down capability to view which sensitive domains and columns are proliferating between two selected systems

### Map Identities

This capability supports GDPR compliance for data privacy by creating an index of data subjects, or individuals, whose personal data is retained in an organization's data stores.

The solution should also include these features:

- A mapping of which data stores contain personal data about each identified subject
- The geographic location of the subjects and the data stores that hold their personal data
- Searchability for each data subject to view an individual's comprehensive data footprint and risk
- Support for data breach notifications, a subject's right to be forgotten (RTBF), and subject requests for data portability and access
- Integration with consent management; many organizations rely on master data management to manage how data can be used

### Analyze and Track Data Risk

This capability calculates sensitive data risk costs based on defined policies in a common risk framework that compares data stores to each other to generate a risk score that drives the protection effort.

The solution should include multiple adjustable factors in its risk framework to calculate the risk for each data store. Look for factors such as:

- Classification/sensitivity level of the data
- Amount of sensitive data
- Protection status of the data
- Number of sensitive fields
- Breach cost of the data
- Sensitive data movement to other data targets
- Number of users with access to the data
- User activity count against the sensitive data
- Custom fields defined by users to tune their risk model to their operating environment

Organizations should be able to fine-tune the individual weights of each of these factors based on the organization's operating environment. Ideally, the platform should also include the ability to simulate a risk score calculation for each individual data store, with details about how the risk is calculated, before implementing protection controls.

### Monitor User Access and User Activities

This capability makes the platform more dynamic and improves risk scoring by capturing who has access to sensitive data and how it is used. To correlate user activities against sensitive data, this capability must be able to ingest raw user activity log files in real time, then integrate them with LDAP and Active Directory data, such as users, user group memberships, user aliases, and user access to data stores, tables, and columns. It should be able to report on users and groups with access to sensitive data by user(s), data store, time of occurrence, and other criteria, and incorporate this user activity into each data store's overall risk scoring.

It should also leverage AI to capture anomalous user behavior and automation to generate alerts and send notifications to the appropriate security team stakeholders when anomalies are detected.

In addition, consider looking for a solution that allows you to enrich your view of user activity and behavior:

- Analyze user activities across all supported data store types
- Provide detailed user profile pages to quickly view and report on an individual's activity
- Leverage a system log server to ingest industry standard activity logs
- Import user access rights
- Provide rich visualizations for user behavior analytics

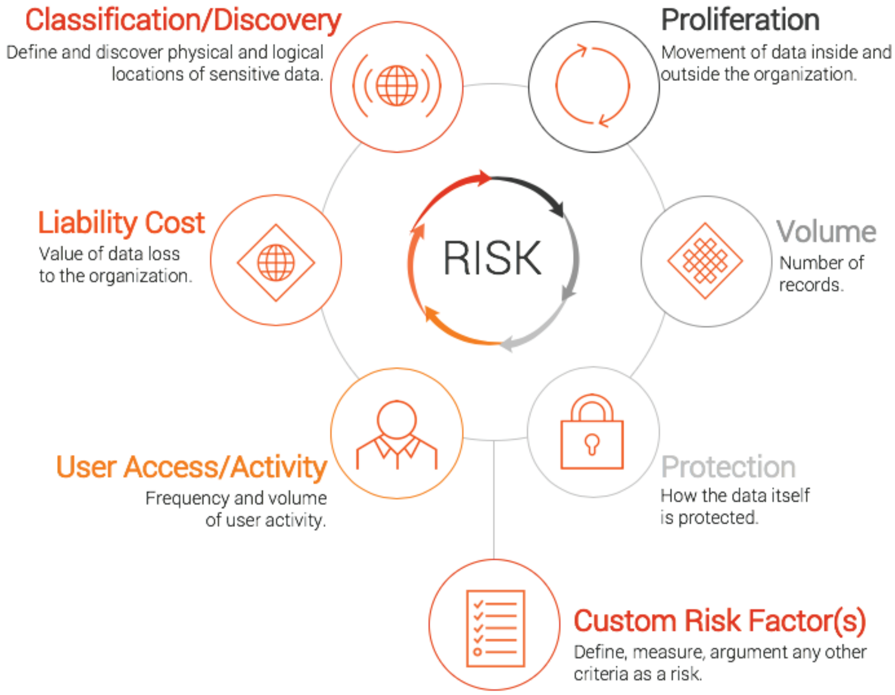

Classification/Discovery
Define and discover physical and logical locations of sensitive data.
Proliferation
Movement of data inside and outside the organization.
Liability Cost
Value of data loss to the organization.
RISK
Volume
Number of records.
User Access/Activity
Frequency and volume of user activity.
Protection
How the data itself is protected.
Custom Risk Factor(s)
Define, measure, argument any other criteria as a risk.

Figure 2: Risk analytics achieve highest accuracy when a solution can continuously discover and classify sensitive data and where it proliferates, scale for data growth, apply protection methods at the data level, incorporate custom risk criteria, monitor and track user behavior, and generate estimated costs of data loss.

### Protect Data, Manage Subject Rights and Consent

This capability automatically applies defined protection policies to encrypt, mask, or otherwise protect data. It should be launchable directly within the target data store or by initiating a protection workflow.

**For data protection**, the capability should support manual and automated workflows for persistent data masking, dynamic data masking, encryption, access controls, ticketing systems (e.g., ServiceNow), and other data protection methods. Protection policy creation should be bi-directional: protection rules should be built within the platform and pushed to protection tools, or vice versa.

Remediation should also include capabilities to automate notifications and alerts to inform users of exceptions and other defined conditions that require their attention. It should also include the ability to create scripts that trigger automatic actions in response to an alert (e.g., moving users from one LDAP group to another).

To provide comprehensive data protection, consider looking for these features in a solution:

- Protection tools that operate across all supported data store types
- Data-centric protection, such as encryption, masking, and access controls
- Integration with third-party tools, such as Ranger™, Sentry™, and ServiceNow®
- A common policy framework for discovery and protection

**For data subject rights**, organizations need clear intelligence on all the personal and sensitive information related to customer, employee, and third-party identities it holds. The data must also be actionable to respond to data subject rights requests. Automated workflows should provide the ability to delete or mask information. Further, organizations should have instantaneous intelligence on individuals, such as where their data is used and the privacy risk. Data privacy and compliance centers on the individual (i.e., whether an individual's data is protected and whether individuals can control their own data).

Supporting this privacy foundation requires organizations to have identity-centric capabilities to:

- Identify the individuals represented in the organization's sensitive data
- Understand what sensitive data it retains for each individual
- Understand sensitive data by individual identities, with intelligence on various attributes (e.g., location, data stores, risk, protection status, use of each individual's sensitive data)
- Quickly locate an individual's sensitive data to support privacy requests, such as: "What data do you hold on me?" and "Please remove all my data!"

**For consent management**, organizations need a master data management (MDM) solution that functions as the glue that binds your systems and information together. MDM is the single source of truth for your data-driven digital transformation, providing trusted, accurate, complete data for your customer experience program, marketing and sales operations, omnichannel retailing, supply chain optimization, governance efforts, compliance initiatives, and more. Required capabilities include:

- Single view of the data, unifying multiple sources of disparate, duplicate, and/or conflicting information sources
- 360-degree view of critical data and its business relationships with other data
- Complete view of all interactions, linking all transactions for a full view of customer behavior

### Measure, Communicate, Audit Response

This capability provides visualizations and reports to support cross-functional collaboration and satisfy auditors. Moreover, organizations can continuously measure their risk to gauge how well policies and processes impact data risk.

In order to meet the data privacy interests and requirements of business function stakeholders, the platform should provide dashboards and visualizations that track data privacy KPIs, such as:

- **IT**: Audit/compliance support, asset value and risk, DevOps privacy
- **Security**: Compliance, security decisions, data protection
- **Privacy**: GDPR, DPIAs, data subject risk, subject access requests, privacy readiness measurement/tracking
- **Business/Risk**: Risk reduction, data governance, regulatory compliance

This capability should also include support for generating reports on demand to satisfy both typical and unexpected auditor requests.

## Implementing a Solution

An organization cannot achieve a comprehensive view of the risks associated with sensitive data until it can:

- Define and manage governance policies
- Identify data location, volume, and proliferation (i.e., where the sensitive data is created and how it flows through the organization)
- Map subject identities
- Comply with privacy regulations and audit requests
- Understand the data protection status (i.e., how the data is currently protected by data security controls)

The security team needs to confirm what it already knows, but it also needs to understand what it may be overlooking. Regular aggregated risk scoring can deliver a quantitative measure of the security team's blind spots, helping it set priorities and focus 80 percent of its efforts on the 20 percent of data at highest risk.

### Understanding User Activity, Behavior, and Anomalies

Data does not compromise itself—a user is always involved in some way. To determine why and how, the security team needs to leverage analytics, intelligence, and automation to easily and quickly identify unusual user behavior. Suspicious events are then more easily detected and more accurately reported, so IT can defend against or remediate potential threats faster.

### Trigger Alerts for Policy or Behavior Exceptions

Assigning specific protection techniques and tools for each type of policy, noncompliance, or suspicious activity makes the security team more efficient and effective.

In some cases, it might work best to take a fully integrated, automated approach to sensitive data protection. For example, when an intelligent solution detects that a user is accessing more SSN records than normal, a script could automatically activate to dynamically mask the data or move the user to a high-risk user group in LDAP. In other cases, the security team might simply want to generate an alert that requests or requires manual intervention by the data owner or application owner.

As the security team configures and rolls out the data protection solution, it will need to take the following steps:

- Define what protection methods or actions to use for each asset type and user
- Ensure that subjects' personal data is protected in compliance with applicable privacy regulations
- Establish an appropriate level of integration with the protection method
- Monitor the effectiveness of the solution continuously and adapt as required
- Expand the scope to more users and more information asset types

### Focus on the Data

Knowing where personal and sensitive data resides, how it's being used, and who is using it is critical to determining how to protect it. Data servers in a secure location may not require encryption at the disk or file level, but databases hosted by a partner in another location or even another country will need stronger protection to prevent loss of control.

Data lakes accessed by multiple users and applications will need to take location, time, and necessary levels of access into account when setting up access rules.

Data controlled by GDPR will need dynamic access rights that adapt to users' time and location, a map of subject identities to the data stores that contain their personal data, and a workflow that leverages automation to manage consent and respond to subject requests.

Test environments pose their own challenges. Functional testing environments need realistic data to continue operating smoothly. Protection applied to columns of sensitive data needs to ensure that the relationships between tables remain intact.

In cross-system business processes, protecting sensitive columns must not break the process. Persistent masking can be deployed in reporting, analytical, and test environments that have little or no need to restore the protected data to its original value.

In production systems, protecting data at rest may be less important than protecting data appropriately for different groups of users. In these cases, data needs to be completely protected from some users and only partially protected from or entirely available to others. A banking call center provides an obvious example of this: database administrators (DBAs) don't need to see a caller's SSN at all; customer service reps need to see the last four digits; and back office users who validate the caller's credit history need to see the full SSN. This fine-grained access control needs to take place dynamically.

### Protection is a Team Sport

Application owners and security analysts must cooperate closely with the DBAs who maintain data operationally. A business process orchestration tool can automate and measure the handoff process among these three groups by ensuring the right level of data security at all times, updating internal systems to reflect that data is protected, and recalculating the risk associated with the data so the security analyst who initiated the protection request can be confident the protection job is complete.

### Intelligence, Analytics, and Automation Lead the Way Forward

A policy-based, intelligent solution that leverages analytics, AI, and automation for compliance reporting, risk analysis, user behavior monitoring, and protection orchestration and remediation ensures that even in a fast-changing environment, data remains up-to-date and secure.

The security team can implement rules that continuously scan new data and changes to metadata. If a data anomaly emerges, the data protection solution can automatically create an alert for a policy violation and send a notification to the appropriate stakeholder to suggest corrective action.

For example, a solution that leverages automation can alert an application owner to take corrective steps to protect data if a new column that contains account numbers appears in an existing system due to a metadata change.

In addition, the solution can combine artificial intelligence with automation for faster response to user behavior that could signal a security threat. When user access, roles, and profiles are clearly defined and user behavior analytics (UBA) are leveraged, an intelligent solution can efficiently spot user activity that diverges from normal behavior patterns or signals unauthorized access to sensitive records, and then provide automated remediation.

For example, a DBA user downloads a report that contains sensitive data, but the user does not have access rights to download that information. The solution could then automatically shift that user into a group without access to the report, create a security alert, and send an email notification to the appropriate team to investigate the potential data risk and ensure that remediation is complete and effective.

While AI-driven anomaly detection and automated remediation are a tremendous help to the security analyst, they can be difficult to get right on the first try. It is generally best to test first in a small group of users or applications, learn from errors, and roll them out gradually.

### Business Benefits

A properly deployed data privacy solution offers business value beyond simply reducing risk. By providing a comprehensive and continuous view of risk for all data assets, subjects, and users, it facilitates governance of security policies and control. This ensures that data owners don't sacrifice flexibility for compliance, keeps data security policies consistent across systems and data sets, and creates a defensible legal position in response to a data breach or audit challenge.

By supporting an approach to data privacy that focuses on applying the right method at the right time to the right type of data, the solution also encourages collaboration between data owners and application owners and improves their buy-in to security measures. Finally, the solution should deliver faster time to value and reduce operational costs. By automating data protection based on risk associated with users, subjects, and/or data stores, it frees IT staff and budget for more strategic tasks.

## Conclusion

According to Gartner, "Risk is not avoided; it is monitored, assessed, balanced with trust, communicated and adapted to acceptable levels—continuously."[5]

Ultimately, the goal is to retain and enhance your customers' trust by continually securing sensitive data across every data source and complying with evolving privacy regulations as personal data proliferates across the enterprise.

To learn more about the Informatica Data Privacy solution, please go to https://www.informatica.com/products/data-security.html.

[5] Ibid.

Informatica

**Worldwide Headquarters** 2100 Seaport Blvd., Redwood City, CA 94063, USA Phone: 650.385.5000, Toll-free in the US: 1.800.653.3871

© Copyright Informatica LLC 2018. Informatica and the Informatica logo are trademarks or registered trademarks of Informatica LLC in the United States and other countries. A current list of Informatica trademarks is available on the web at https://www.informatica.com/trademarks.html. Other company and product names may be trade names or trademarks of their respective owners. The information in this documentation is subject to change without notice and provided "AS IS" without warranty of any kind, express or implied.