Data Rakshaq

From consent to compliance - seamlessly

Data Discovery & Classification

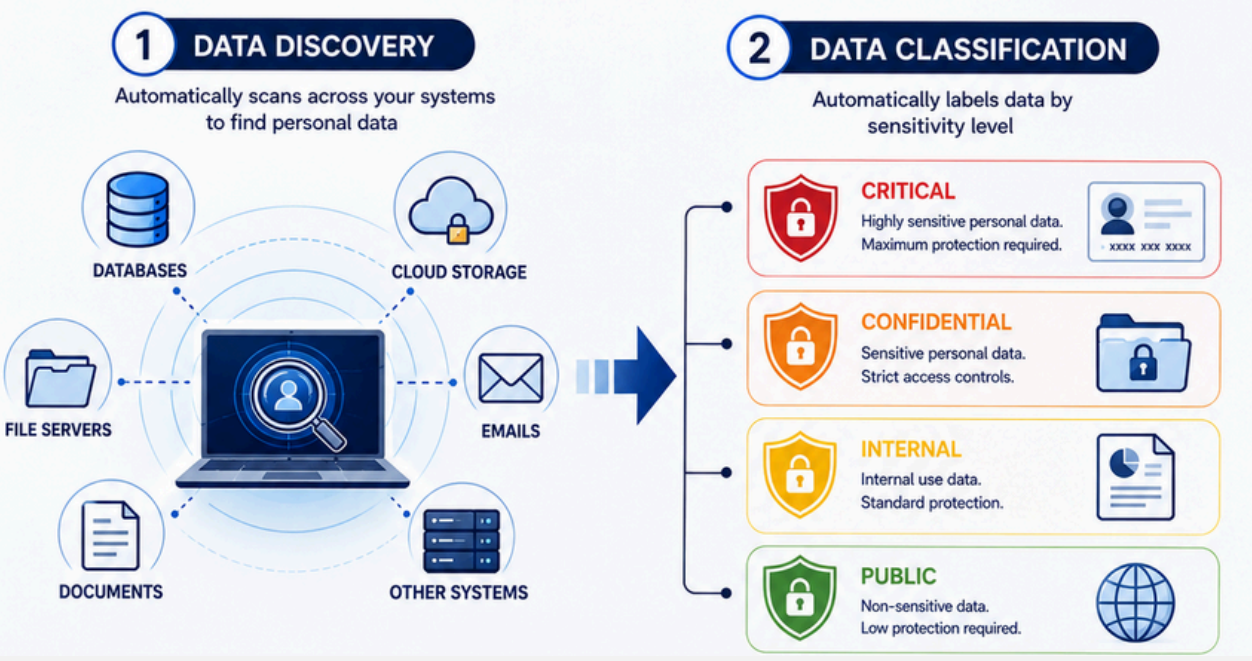
A Complete Enterprise Guide

Understanding How DataRakshaQ Finds, Scans, and Classifies Your Data

Table of Contents

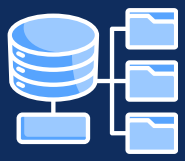
What is Data Discovery & Classification?

Before you can protect personal data, you first need to know where it lives. Data Discovery is the process of automatically scanning all your organization's systems — databases, files, cloud storage, and more — to find personal data wherever it exists. Data Classification is the next step: once data is found, it is labelled by its sensitivity level — so your teams know which data needs the highest protection and which can be handled with standard controls.



Under India's Digital Personal Data Protection (DPDP) Act 2023, organizations must know exactly what personal data they hold, where it is stored, and how it is used. DataRakshaQ's Data Discovery & Classification engine makes this possible — automatically, accurately, and at scale.

How the DataRakshaq Scanner Works



Connect to Data Sources

You connect DataRakshaq to your databases, cloud storage, file systems, and other data repositories using ready-made connectors. No coding is required.

Scan & Extract Data

The scanner reads data from every connected source — tables, columns, files, documents, audio, and video — and extracts text content for analysis.

OCR for Non-Text Files

For scanned PDFs, images, and documents, Optical Character Recognition (OCR) converts visual text into machine-readable text so it can be analyzed.

AI/ML Pattern Recognition

Artificial Intelligence and Machine Learning models analyze the extracted text to identify personal data patterns — names, phone numbers, Aadhaar numbers, financial data, health records, and more.

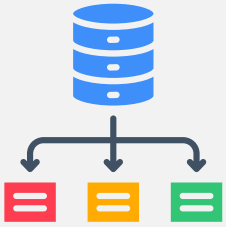
Data Classification

Each piece of identified data is automatically labeled: Critical, Sensitive, Standard, or Non-Sensitive — based on predefined rules and AI models.

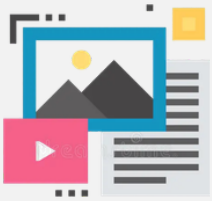
Data Inventory & Risk Report

A complete Data Inventory is generated — showing what personal data was found, where it is, its classification, and risk level. This feeds directly into DPDP compliance workflows.

Structured Data



Structured data is organized in rows and columns with a fixed format, making it easy to store, search, and analyze. It is commonly found in database tables, spreadsheets, and CSV files, and is typically stored in relational databases, NoSQL databases, and data warehouses. **DataRakshaq** scans structured data using direct query-based connectors, making the process predictable and efficient. This type of data often contains core HR, customer, and financial personal information that is **critical** for **DPDP compliance.**



Unstructured Data

Unstructured data refers to information that does not follow a predefined format or structure, making it more complex to process and analyze. It includes PDFs, Word documents, images, emails, audio, and video files, and is commonly stored across file servers, cloud storage, email systems, and shared drives. **DataRakshaq** scans this data using OCR combined with AI/ML-based text extraction and analysis to identify sensitive information. Since the data varies by file type, language, and quality, scanning is more challenging. Common examples include contracts, KYC documents, medical records, and call recordings, all of which often contain **personal data subject** to **DPDP compliance.**

Ready-Made Connectors
Structured Data

Relational Databases (SQL)

Relational databases store data in tables with rows and columns. They are used by most enterprise applications for customer records, transactions, HR data, and more.

- PostgreSQL
- MySQL
- Microsoft SQL Server
- Snowflake

NoSQL Databases

NoSQL databases store data in flexible formats — documents, key-value pairs, graphs — suitable for modern applications and large-scale data.

- MongoDB

Object & File Storage

Object storage systems hold unstructured files — documents, images, backups — at massive scale. These often contain hidden personal data in files.

- Amazon S3
- Google Cloud Storage
- Local Filesystem

PostgreSQL

PostgreSQL is a widely used open-source relational database that stores structured data in tables. **DataRakshaq** scans PostgreSQL databases to identify customer, employee, financial, and other personal data for compliance.

MySQL

MySQL is one of the world's most popular open-source relational databases, commonly used by websites and business applications. **DataRakshaq** scans MySQL databases to detect user accounts, contact details, transaction records, and other sensitive personal information.

Microsoft SQL Server

MySQL is one of the world's most popular open-source relational databases, commonly used by websites and business applications. **DataRakshaq** scans MySQL databases to detect user accounts, contact details, transaction records, and other sensitive personal information.

Snowflake

Snowflake is a modern cloud data warehouse designed for large-scale analytics and business intelligence. **DataRakshaq** scans Snowflake to identify customer, financial, marketing, and analytical datasets containing personal information.

MongoDB

MongoDB is a document-based NoSQL database that stores data in flexible JSON-like formats. **DataRakshaq** scans MongoDB collections to detect PII in user profiles, application logs, session data, and document records.

Amazon S3

Amazon S3 is AWS's cloud object storage service used to store documents, backups, and media files. **DataRakshaq** scans S3 buckets to identify personal data in KYC documents, contracts, scanned IDs, backup files, and reports.

Google Cloud Storage

Google Cloud Storage is a scalable cloud object storage platform for storing files and datasets. **DataRakshaq** scans stored documents, media files, exported datasets, and scanned forms to detect personal information.

Local Filesystem

A Local Filesystem stores files on on-premise servers and network drives within an organization. **DataRakshaq** scans these locations to identify personal data in shared documents, spreadsheets, scanned files, and archived records.

Ready-Made Connectors — Vector Databases

Vector databases are a new class of database designed to store AI-generated data (called embeddings). As enterprises adopt AI and large language models (LLMs), personal data can end up stored in vector databases without anyone realizing it. DataRakshaQ scans these too.

Vector DB	Type	Typical Use Case	Personal Data Risk
Pinecone	Managed Cloud	Enterprise AI search, LLM memory	High — may store customer query PII
ChromaDB	Open Source	Local AI apps, prototype LLMs	Medium — developer environments with real data
pgvector	PostgreSQL Extension	Hybrid structured + vector storage	High — combined with existing personal data tables

Unstructured Data Scanning — OCR & AI/ML

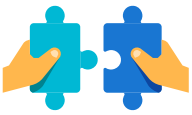
What is OCR and Why Does It Matter?

OCR (Optical Character Recognition) is a technology that reads text from images and scanned documents. When a form is scanned and saved as a PDF, the computer does not 'see' the text — it just sees an image. OCR converts that image into readable text that can then be analyzed for personal data.

File Type	How DataRakshaQ Processes It
Scanned PDFs 	OCR extracts text → AI identifies PII
Images (JPG, PNG) 	OCR reads embedded text → pattern matching
Native PDFs 	Direct text extraction → AI classification
Word Documents 	Text extraction → NLP analysis
Spreadsheets (CSV/XLSX) 	Structured parsing → PII detection
Emails (EML/MSG) 	Header + body parsing → PII flagging

AI/ML Capabilities for Unstructured Data

DataRakshaQ's AI/ML engine goes beyond simple keyword matching. It understands context, language patterns, and data types to accurately identify personal data even when it is written in different ways.

AI/ML Capability	What It Does
Named Entity Recognition (NER) 	Identifies names, organizations, locations, dates
Pattern Matching 	Detects data by format (regex + AI)
Contextual Understanding 	Understands surrounding text to confirm PII
Multi-language Support 	Detects PII in regional languages
Confidence Scoring 	Assigns a confidence % to each detected item
Audio Transcription 	Converts speech in audio/video to text
Video Frame Analysis 	Extracts frames and reads text from video

Data Classification — How Data is Labelled

🔴 Critical (Very High Risk)

- Data whose exposure can cause severe harm, legal liability, financial loss, or direct harm to an individual.
- Requires the highest level of protection.

Examples: Aadhaar number, PAN card, passport, biometric data, financial account numbers, medical records, passwords, encryption keys.

Required actions:

- Encrypt all data.
- Restrict access on a need-to-know basis.
- Audit and log every access.
- Retain data only for the shortest necessary period.
- Continuously monitor for unauthorized access.

🟠 Sensitive (High Risk)

- Personal data that can cause significant harm, identity theft, financial fraud, or distress if exposed.
- Requires strong security and privacy controls.

Examples: Full name + phone number, email address, date of birth, salary information, HR records, CCTV footage, location history.

Required actions:

- Implement role-based access control.
- Maintain access logs and activity monitoring.
- Obtain consent where applicable.
- Conduct periodic access reviews.
- Store data securely.

🟡 Standard (Medium Risk)

- Personal data that is not highly sensitive on its own but may become sensitive when combined with other information.
- Requires standard organizational security measures.

Examples: First name, job title, employer name, city of residence, general preferences, publicly available contact details.

Required actions:

- Apply basic access controls.
- Follow data retention policies.
- Obtain consent when combining with other personal data.
- Monitor data usage to prevent misuse.

🟢 Non-Sensitive (Low Risk)

- Data that does not identify a specific individual or is already publicly available.
- Requires only basic security measures.

Examples: Aggregated statistics, anonymous survey responses, publicly published information, generic business data.

Required actions:

- Maintain standard security practices.
- Ensure data integrity.
- No special data protection obligations beyond general security requirements.

Getting Started with DataRakshaq

Identify Your Data Sources

List all the databases, file systems, cloud storage, and applications that may hold personal data in your organization.

Configure Connectors

DataRakshaq connects to your data sources using ready-made or custom-built connectors. It requires only read-only credentials, ensuring your data is never modified.

Define Classification Rules

Review the default DPDP classification rules and add any organization-specific rules (e.g., your internal employee ID is Critical data in your context).

Run Initial Scan

Trigger the first full scan across all connected sources. DataRakshaQ will automatically discover, classify, and inventory all personal data.

Schedule Continuous Scanning

Set up automated recurring scans (daily, weekly, or monthly) to detect new personal data as it enters your systems.

Review & Validate Results

Review the Data Inventory — validate high-confidence findings, investigate flagged items, and manually classify any uncertain records.

Integrate with DPDP Workflows

Connect the Data Inventory to DataRakshaQ's DSR, Consent, and Grievance modules for end-to-end DPDP compliance.