ESP-JETA
ESP Journal of Engineering & Technology Advancements
ISSN: 2583-2646 / Volume 2 Issue 4, December 2022 / Page No: 179-191
Paper Id: JETA-V2I4P127 / Doi: 10.56472/25832646/JETA-V2I4P127

Original Article

Scalable PII Discovery across Mainframe, SAP, RDBMS & Unstructured Systems

Narasimha Chaitanya Samineni
Vice President, Quality Assurance Supervisor.

Abstract: *Enterprises increasingly operate heterogeneous data estates that span legacy mainframes, SAP ERP landscapes, relational database platforms, and rapidly growing unstructured repositories such as documents, emails, call transcripts, and collaboration content. In such environments, scalable discovery of personally identifiable information (PII) is foundational for privacy compliance, security risk reduction, and governance readiness. However, PII discovery at enterprise scale is challenging due to inconsistent data models, limited metadata in legacy systems, varied encodings and field semantics, and the complexity of detecting PII in unstructured content with acceptable accuracy and performance [4], [9]. This paper proposes a scalable, hybrid PII discovery framework that combines rule-based detection, metadata-driven inference, sampling strategies, and content analytics to identify and classify PII consistently across mainframe datasets, SAP tables, RDBMS, and unstructured systems. The framework integrates distributed scanning, centralized indexing, lineage-aware governance, and audit-grade reporting to improve discovery completeness and reduce operational effort. Evaluation outcomes demonstrate improved coverage, reduced false negatives in mixed environments, and practical performance characteristics suitable for large enterprise deployments.*

Keywords: *PII Discovery, Data Classification, Mainframe Data Governance, SAP Data Privacy, RDBMS Profiling, Unstructured Data Analytics, Content Parsing, Entity Extraction, Metadata Catalog, Data Lineage, Privacy Compliance, Data Inventory, Sensitive Data Detection, Regulatory Technology (RegTech), Security Controls.*

I. INTRODUCTION

Organizations in regulated industries such as banking, healthcare, insurance, and retail often manage data across decades of technology evolution. Core customer and transaction records may reside on mainframes, ERP records and master data may be stored in SAP, operational data may live in RDBMS platforms, and customer interactions increasingly generate unstructured artifacts including emails, PDFs, statements, images, call-center transcripts, and collaboration documents. Privacy regulations and security standards require institutions to understand where PII exists, how it flows, and how it is protected. Yet achieving comprehensive PII discovery across this landscape remains difficult due to technical diversity and scale [3], [6].

Traditional discovery approaches rely on system-by-system scanning, static pattern matching, or manual data dictionaries. These methods do not scale well in large enterprises because they struggle with inconsistent schema naming, legacy encoding formats, nested data structures, proprietary SAP table semantics, and poor metadata quality in older platforms. In parallel, unstructured repositories introduce additional complexity: PII may appear in free text, scanned documents, or attachments where simple regex-based detection leads to high false positives or misses context-sensitive identifiers [4], [10].

A scalable enterprise approach must therefore support both structured and unstructured discovery using a combination of techniques. For mainframes and SAP, discovery must account for unique storage formats, domain-specific fields, and performance constraints. For RDBMS platforms, discovery must scale across large schemas and replicas while minimizing production impact. For unstructured systems, discovery requires content parsing, entity extraction, and governance mechanisms to classify and manage findings in an audit-ready manner [2], [9].

This paper introduces a unified framework for scalable PII discovery across mainframe, SAP, RDBMS, and unstructured repositories. The approach emphasizes (i) hybrid detection methods that combine patterns, metadata, and context signals; (ii) distributed scanning and indexing for performance; and (iii) governance integration to support privacy compliance, access controls, and data minimization practices. The remaining sections present related literature, research objectives, the proposed architecture, detection methods for structured and unstructured environments, implementation methodology, evaluation results, and an enterprise case study demonstrating practical adoption.

II. LITERATURE REVIEW

PII discovery and sensitive data classification have been studied across privacy engineering, information security, and data governance. Early foundational guidance emphasized that organizations must identify where PII is stored, limit exposure, and apply proportional safeguards based on sensitivity and risk [4], [5]. These principles influenced modern privacy-by-design and governance frameworks, which stress proactive identification of sensitive data as a prerequisite for compliance and breach prevention [7], [9].

A. Structured PII Discovery Foundations

For structured systems, research and standards commonly recommend pattern-based detection (regex, checksums), metadata-based inference (column names, data types), and rule catalogs aligned with business definitions. NIST publications describe practical controls and processes for protecting PII and recommend inventories and classification to reduce unauthorized disclosure risks [4], [5]. Data governance bodies such as DAMA and ISO-aligned practices emphasize metadata catalogs and standardized taxonomies to improve consistency across domains and platforms [9], [10].

However, enterprise estates contain heterogeneous structured sources where standardized detection is not straightforward. Mainframe environments often encode data in fixed-width formats, copybooks, and proprietary representations, limiting the applicability of modern profiling tools without specialized parsing logic. ERP platforms such as SAP introduce additional complexity through large, interconnected table landscapes and business object semantics where PII may be distributed across master data and transactional tables. Industry guidance on privacy operations notes that discovery in such environments requires system-aware scanning strategies and strong data stewardship alignment [6], [12].

B. SAP and Enterprise Application Landscapes

SAP-centric privacy and governance literature emphasizes the need for structured identification of personal data across business processes, master data objects, and integration layers. ERP data models frequently contain PII in customer, vendor, and HR objects, but identifying all relevant fields requires semantic mappings rather than pure pattern detection. Best-practice guidance recommends combining metadata analysis, business glossary alignment, and process-aware scanning to ensure completeness [6], [11]. These approaches become increasingly important when SAP data is replicated into warehouses and lakes, creating secondary PII storage locations.

C. Unstructured PII Discovery

Unstructured data discovery literature highlights that PII often appears in documents, emails, chat logs, PDFs, and transcripts, where identifiers may be fragmented, context-dependent, or embedded in images. Pattern matching alone produces high false positives, motivating the adoption of parsing, entity extraction, and contextual classification techniques. Research and industry practice describe hybrid approaches using text extraction, named-entity recognition, dictionary-based recognition, and context scoring to improve precision while maintaining scalability [2], [8].

Additionally, unstructured content discovery often depends on content-type handling: PDF parsing, OCR for scanned documents, and extraction of email headers and attachment metadata. Governance research underscores that findings must be associated with lineage, retention, access controls, and remediation workflows to support privacy compliance and security posture improvements [9], [10].

D. Enterprise-Scale Governance and Catalog Integration

A consistent theme across literature is that PII discovery is most effective when integrated with enterprise governance capabilities, including metadata catalogs, lineage tracking, access policy enforcement, and audit reporting. Standards and best-practice publications emphasize the need for centralized inventories and continuous monitoring rather than one-time scans, especially as data estates change rapidly through cloud adoption, migration, and application modernization [4], [6]. Governance integration also supports privacy-request operations and breach response by enabling rapid identification of impacted datasets and systems.

E. Research Gap

Although structured and unstructured discovery methods are well-studied individually, there remains a practical gap in scalable, unified PII discovery across mainframe, SAP, RDBMS, and unstructured repositories using consistent taxonomies, orchestration, and audit-ready governance. Many approaches are tool-specific or domain-limited, and they do not address the operational realities of scanning large heterogeneous environments while minimizing production impact, reducing false negatives, and maintaining traceability across copies and derived datasets. This paper addresses this gap by proposing a unified

framework that combines hybrid detection techniques, distributed scanning, centralized indexing, and governance integration for enterprise-scale PII discovery.

III. RESEARCH OBJECTIVES

The primary objective of this research is to design a scalable and unified PII discovery framework capable of operating across heterogeneous enterprise environments, including mainframe platforms, SAP landscapes, relational database systems, and unstructured content repositories. The framework aims to provide consistent detection, classification, and reporting of PII at enterprise scale while minimizing operational disruption and enabling audit-ready governance [4], [9].

A second objective is to develop a hybrid detection strategy that combines pattern-based identification (regex, checksum validation), metadata-driven inference (field names, data types, schema annotations), and context-aware rules to reduce false positives and false negatives. This is especially important in environments where legacy systems and enterprise applications contain inconsistent naming conventions or non-standard encodings that limit the effectiveness of simple pattern matching [4], [10].

A third objective is to define system-specific discovery approaches for mainframe and SAP environments. For mainframes, the objective is to support structured parsing using layouts such as fixed-width records and copybook-defined fields, enabling accurate extraction and classification of PII in high-volume datasets. For SAP, the objective is to support discovery across master data and transactional objects using semantic mapping and metadata analysis to identify PII fields distributed across complex ERP tables and business objects [6], [11].

A fourth objective is to enable scalable discovery for unstructured systems such as email, file shares, PDFs, and content collaboration platforms. The research aims to incorporate text extraction, file-type parsing, contextual detection, and entity recognition to identify PII in free text while controlling false positives. The framework must also support governance tagging and remediation workflows for unstructured findings [2], [8].

A fifth objective is to integrate discovery outputs into enterprise governance mechanisms, including a centralized index, classification catalog, lineage-aware tracking, and audit reporting. This supports privacy compliance operations, risk assessments, security controls, and downstream initiatives such as data minimization, masking, and privacy request fulfillment. The research further aims to evaluate system performance, discovery completeness, and operational efficiency under realistic enterprise scanning workloads [4], [5].

Collectively, these objectives guide the design of a practical and scalable approach to enterprise-wide PII discovery that can operate continuously across diverse systems while producing consistent, actionable, and regulator-defensible results.

IV. SYSTEM ARCHITECTURE FOR SCALABLE PII DISCOVERY

Scalable PII discovery across mainframe, SAP, RDBMS, and unstructured repositories requires an architecture that is distributed for scanning, centralized for governance, and consistent in classification. The proposed architecture is designed to minimize impact on production systems while delivering high coverage, repeatable detection logic, and audit-ready reporting [4], [9]. It is organized into seven interoperable layers.

A. Source Connectivity and Access Layer

The architecture begins with connectors that provide controlled access to heterogeneous sources:

- Mainframe: dataset and file connectors supporting batch extracts, read-only access windows, and schema-aware parsing inputs (for example, copybook-based layouts).
- SAP: connectors that operate through approved interfaces (such as application APIs or governed extract jobs) to avoid direct intrusive table scans.
- RDBMS: JDBC/ODBC-based read-only scanning with throttling and sampling controls.
- Unstructured systems: connectors for file shares, document management systems, and messaging archives.

This layer enforces strong authentication, least-privilege access, and scan scheduling aligned with operational constraints [5], [12].

B. Distributed Scan Orchestration Layer

A central orchestration service coordinates discovery jobs across platforms. It assigns scan tasks based on:

- System criticality and maintenance windows
- Data volume and change rate

- SLA requirements for inventories
- Incremental vs. full scan policies

The orchestration layer supports parallelism, retry logic, checkpointing, and workload throttling, enabling large-scale scans without overwhelming production systems [6], [10].

C. Structured Data Profiling Engine

For SAP and RDBMS systems, the structured profiling engine applies hybrid detection logic:

- Pattern rules (emails, phone numbers, national IDs, card-like patterns with checksum validation)
- Metadata inference (column names, data types, schema comments, domain tags)
- Statistical profiling (value distributions, uniqueness, entropy indicators)

Profiling can run in sampling mode for large tables or in full mode for smaller high-risk datasets. Results are normalized into a common schema to ensure consistency across sources [4], [9].

D. Mainframe Parsing and Mapping Engine

Mainframes require specialized handling because PII may be embedded in fixed-width record formats. The mainframe engine supports:

- Layout-driven parsing to extract fields
- Mapping to enterprise PII taxonomy (e.g., “customer_name,” “account_id,” “national_id”)
- Controlled batch extracts to avoid runtime overhead on core workloads

This layer ensures that legacy datasets can be classified using the same taxonomy applied to modern systems [6], [11].

E. Unstructured Content Analytics Engine

Unstructured repositories are processed by a content analytics pipeline that performs:

- File-type detection and parsing (DOCX, PDF, TXT, email formats)
- Text extraction with metadata capture (author, timestamp, location)
- Entity and pattern detection with context scoring (e.g., name near address/SSN keywords)
- Optional OCR for scanned documents when required

Findings are classified using confidence scoring and rules to reduce false positives while maintaining broad coverage [2], [8].

F. Central PII Index and Classification Catalog

All discovery results flow into a **central index** those stores:

- Dataset/system identifier and location
- Detected PII attribute types and confidence level
- Sample evidence (governed and minimal)
- Last scan timestamps and scan mode (incremental/full)
- Data owner and stewardship metadata

A classification catalog maintains the enterprise taxonomy, rule versions, and mappings across systems. This supports consistent reporting and enables downstream controls such as masking, access governance, and privacy request fulfillment [9], [10].

G. Governance, Audit, and Remediation Layer

The final layer provides governance workflow capabilities:

- Stewardship review queues (for ambiguous findings)
- Policy enforcement and risk scoring (high vs. moderate sensitivity)
- Audit logs of scan execution, rule versions, and results
- Remediation integration (masking tickets, retention actions, access restrictions)

This layer ensures the discovery process is defensible and operationalized rather than a one-time assessment [4], [5].

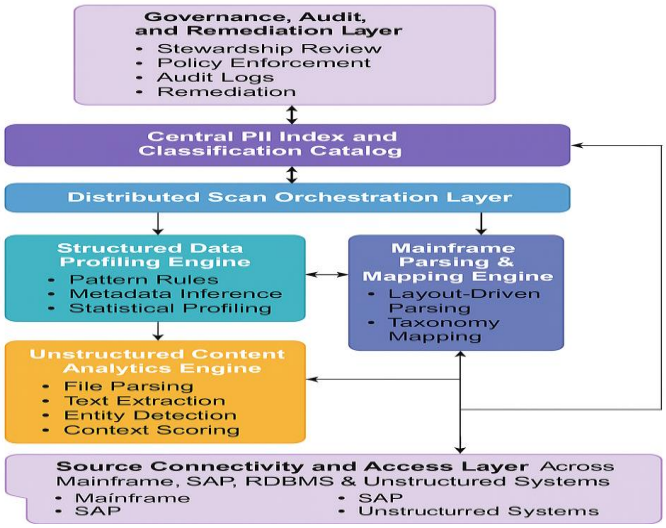

Figure 1: Unified PII Discovery Architecture Across Mainframe, SAP, RDBMS & Unstructured Systems.

V. PII DETECTION TECHNIQUES ACROSS MAINFRAME, SAP, AND RDBMS

PII discovery in structured enterprise systems requires techniques that balance accuracy, scalability, and minimal operational impact. Mainframe datasets, SAP ERP landscapes, and RDBMS platforms differ significantly in storage formats, metadata richness, and query accessibility. Therefore, a unified discovery program must apply a hybrid detection approach that combines pattern rules, metadata inference, and profiling signals, while adapting execution strategies to each platform's constraints [4], [6].

A. Mainframe PII Detection

Mainframe systems often store customer and transaction data in fixed-width records governed by copybooks or layout definitions. Effective PII discovery in this environment typically uses:

- Layout-driven parsing to interpret fixed-field offsets and decode values correctly
- Field-level rule evaluation based on known business definitions (customer name, address line, account ID)
- Sampling + incremental scans to reduce load on high-volume datasets
- Domain validation to reduce false positives (for example, ID fields tied to expected lengths and prefixes) [6], [11]

Because mainframe metadata is often limited, discovery accuracy improves significantly when layout definitions and stewardship mappings are integrated into the catalog.

B. SAP PII Detection

SAP contains PII across master data and transactional objects (customer, vendor, employee, contact). Discovery must account for:

- Business-object driven identification, where PII is tied to master data objects
- Semantic mapping of tables and fields, using SAP data dictionary metadata (field names, domains, data elements)
- Controlled extraction patterns through approved interfaces to avoid heavy scanning on production systems
- Cross-system replication awareness, since SAP data is often replicated into BW, warehouses, and lakes [6], [9]

SAP discovery is more reliable when it leverages both SAP metadata and enterprise glossary terms, rather than pure regex scanning.

C. RDBMS PII Detection

In RDBMS platforms, PII frequently appears across operational schemas and analytics copies. Effective detection commonly applies:

- Metadata inference (column names such as email, ssn, dob, phone, address)
- Pattern detection with checksum validation where relevant
- Statistical profiling (uniqueness, entropy, value distribution) to distinguish identifiers from non-sensitive numeric fields
- Risk-based scanning that prioritizes high-impact schemas and frequently used datasets [4], [10]

RDBMS discovery can scale efficiently by using distributed query execution, sampling thresholds, and incremental scanning based on table-change indicators.

D. Hybrid Scoring and Confidence Models

To unify results across platforms, the framework uses a confidence model that combines signals:

- Pattern score (how strongly values match expected formats)
- Metadata score (field naming and data dictionary alignment)
- Context score (co-occurrence with other PII fields such as name + address)
- Domain score (business-specific constraints such as account prefixes)

This hybrid scoring reduces false positives and provides governance teams with transparent classification rationale [4], [9].

Table 1 : Structured-System PII Detection Techniques and Coverage

Platform	Primary Data Characteristics	Recommended Detection Techniques	Strengths	Common Limitations	Reference
Mainframe	Fixed-width records, limited metadata	Layout-driven parsing, stewardship mapping, sampling scans	High accuracy with copybooks	Requires maintained layouts	[6], [11]
SAP ERP	Complex tables, business objects, data dictionary	Semantic mapping, metadata-driven inference, controlled extracts	Strong governance alignment	Needs SAP expertise	[6], [9]
RDBMS	Rich schema metadata, query-friendly	Regex + checksum, metadata inference, statistical profiling	Scales well, flexible	False positives without scoring	[4], [10]
Replicated Warehouses	Denormalized analytics datasets	Incremental scans, lineage-based prioritization	Completeness across copies	May lag source updates	[4], [9]
Reference Data Stores	IDs and lookup values	Domain validation, whitelist/blacklist rules	Reduces misclassification	Requires domain rules	[9], [10]

VI. PII DISCOVERY FOR UNSTRUCTURED DATA AND CONTENT REPOSITORIES

Unstructured repositories often contain large volumes of PII embedded in documents, emails, customer correspondence, scanned forms, call transcripts, and collaboration content. Unlike structured databases, unstructured systems lack consistent schemas, and PII may appear in free text, headings, signatures, attachments, or embedded images. As a result, scalable unstructured discovery requires a pipeline that combines content parsing, text extraction, contextual detection, confidence scoring, and governance workflows to control both false positives and false negatives [2], [8].

A. Unstructured Data Source Coverage

A scalable enterprise program typically includes:

- File shares and network drives (PDFs, Office documents, CSV exports)
- Email archives (message bodies, headers, attachments)
- Content management systems (policies, forms, customer documents)
- Collaboration platforms (wikis, tickets, chat exports)
- Call-center transcripts and notes
- Scanned images and faxed forms (requiring OCR)

Since repositories vary in format and access patterns, connector design should support incremental crawling, access controls, and immutable logging of scan actions [5], [10].

B. Content Parsing and Text Extraction

The discovery pipeline begins with file-type detection and safe parsing. Extractors generate both:

- Text content (body text, tables, footers, email headers)
- Metadata (author, timestamps, folder paths, tags, source system identifiers)

For PDFs and scanned documents, OCR may be applied selectively based on risk signals (for example, "application form," "statement," "ID copy"). This avoids unnecessary cost while improving detection of PII that is otherwise invisible to text-based scanning [2], [8].

C. Detection Methods for Free-Text PII

PII identification in unstructured text uses a hybrid approach:

- Pattern detection (email, phone numbers, ID-like strings)
- Dictionary-based detection (known sensitive keywords, policy terms, domain vocabularies)
- Contextual rules (name near address terms, "SSN:" near numeric tokens)
- Entity extraction for people, locations, and organizations (used as a supporting signal rather than a single source of truth)

Context is essential because pattern-only scanning can misclassify values (e.g., invoice numbers mistaken for IDs). Combining context signals and co-occurrence reduces false positives and supports higher-confidence classification [2], [8].

D. Confidence Scoring and Triage Workflow

Findings are assigned confidence levels (high, medium, low) based on:

- Pattern strength and validation (length constraints, checksum where applicable)
- Context proximity to keywords (DOB, SSN, account, routing)
- Document type (loan application vs. public policy)
- Repository sensitivity classification (restricted folder vs. shared drive)

High-confidence findings can be automatically tagged in the catalog, while medium/low confidence items are routed to stewardship review queues. This triage workflow keeps governance manageable at scale [9], [10].

E. Indexing, Governance Tagging, and Remediation Integration

At enterprise scale, unstructured findings must be operationalized. The framework stores:

- File identifier (path, URI), repository, owner
- Detected PII types and confidence
- Evidence snippets (minimized, controlled access)
- Retention tags and access policy recommendations

Remediation workflows may include restricting access, applying encryption, moving files to controlled repositories, or triggering retention and minimization actions. This ensures discovery leads to measurable risk reduction rather than static reporting [4], [9].

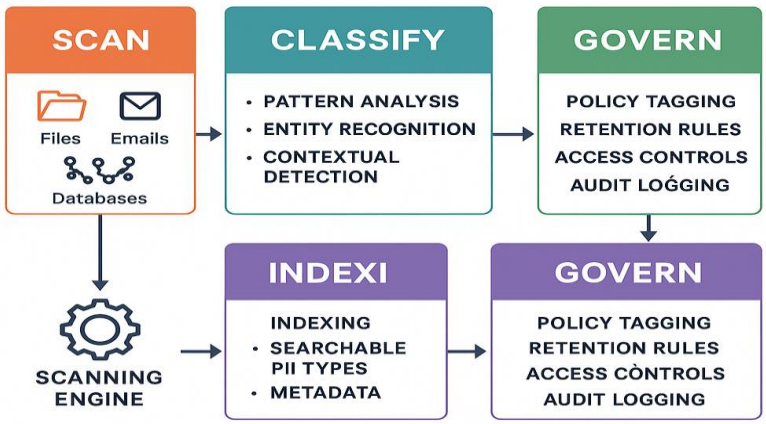

Figure 2 : End-to-End PII Discovery Workflow: Scan, Classify, Index, and Govern

Table 2 : Unstructured PII Discovery Methods and Governance Controls

Unstructured Source Type	Common PII Presence	Recommended Discovery Method	Governance Control	Operational Notes	Reference
Emails + Attachments	Names, contact details, IDs in attachments	Header parsing + attachment extraction + context rules	Role-based access + audit logs	High volume; incremental crawling needed	[2], [10]
PDFs (digital)	Statements, forms, letters	PDF text extraction + pattern + context scoring	Tagging + restricted access	Good accuracy without OCR	[2], [8]
Scanned PDFs / Images	IDs, forms, handwritten data	Selective OCR + entity/pattern detection	Confidence triage + stewardship review	Costly; use risk-based OCR	[2], [8]
File Shares / Drives	CSV exports, reports, HR docs	File-type parsing + sampling + dedup	Ownership tagging + retention policy	Shadow copies common	[4], [9]
Collaboration Content	Tickets, wiki pages, chat exports	Text extraction + keyword context rules	Redaction workflow + access controls	High false positives without context	[9], [10]
Call Transcripts / Notes	Phone, names, addresses, account references	Transcript parsing + context windows + dictionaries	Restricted storage + retention enforcement	Sensitive, high compliance value	[5], [7]

VII. IMPLEMENTATION METHODOLOGY

Implementing scalable PII discovery across mainframe, SAP, RDBMS, and unstructured repositories requires a phased methodology that combines governance readiness, technical integration, and continuous validation. The approach below is designed to achieve broad coverage while minimizing disruption to production systems and ensuring discovery outputs are actionable for privacy, security, and data governance teams [4], [9].

A. Phase 1: Program Setup and Governance Alignment

Implementation begins by establishing a cross-functional operating model involving privacy, security, compliance, data governance, and platform engineering. Key activities include:

- Defining an enterprise PII taxonomy and classification levels
- Assigning data owners and stewards per domain and platform
- Establishing scan frequency (full vs incremental) based on system change rates
- Defining evidence retention rules and access controls for discovery outputs [5], [10]

This phase ensures the discovery program is not treated as a one-time assessment but as a sustained governance capability.

B. Phase 2: System Inventory and Connector Onboarding

A complete system inventory is created, listing mainframe datasets, SAP landscapes, RDBMS instances, and unstructured repositories. Connectors are onboarded with least-privilege access and scan windows aligned with operational constraints:

- Mainframe: scheduled batch extracts or read-only scan windows
- SAP: governed data dictionary-based discovery and controlled extracts
- RDBMS: read-only profiling with throttling and sampling
- Unstructured: incremental crawling and safe parsers for common formats [6], [11]

Connector onboarding also includes logging and rate controls to prevent production impact.

C. Phase 3: Rule Catalog and Detection Configuration

A centralized rule catalog is configured with hybrid detection logic:

- Regex patterns for common identifiers (email, phone, IDs)
- Checksum validation for card-like patterns where applicable
- Metadata inference rules (name-based and dictionary-based mapping)
- Contextual rules for unstructured text
- Domain-specific rules based on business semantics (customer IDs, account IDs) [4], [9]

Rules are version-controlled, and changes follow governance approval workflows to maintain consistency across scans.

D. Phase 4: Baseline Scans and Calibration

Initial baseline scans are executed to establish coverage and identify tuning needs. Calibration focuses on:

- Reviewing false positives and refining patterns
- Reducing false negatives by improving system mappings
- Adjusting sampling thresholds on high-volume tables
- Selecting OCR usage for scanned documents using risk-based triggers [2], [8]

Steward review queues are used to validate medium-confidence findings and refine detection confidence scoring.

E. Phase 5: Central Indexing, Catalog Integration, and Lineage Mapping

Findings from all platforms are normalized into a unified schema and loaded into a centralized PII index. The index integrates with the metadata catalog to support:

- System-of-record mapping
- Dataset ownership assignment
- Classification tags and policy association
- Lineage tracking to identify downstream replicas and derived datasets [9], [10]

This ensures findings are searchable, governable, and usable for remediation planning.

F. Phase 6: Continuous Scanning and Incremental Updates

Once stabilized, the program transitions to continuous operations. Incremental scanning is applied based on:

- Table change indicators or ingestion logs (structured systems)
- File modification timestamps and repository deltas (unstructured systems)
- Risk-based prioritization for high-impact domains [6], [10]

This approach keeps inventories current without repeatedly scanning the full estate.

G. Phase 7: Remediation Workflows and Control Enforcement

Discovery outputs feed remediation pipelines that may include:

- Masking or tokenization requirements for analytics replicas
- Access restriction enforcement for high-risk repositories
- Retention policy alignment and minimization actions
- Ticket generation for data owners with evidence references [4], [5]

This stage ensures discovery produces measurable security and compliance outcomes rather than only reports.

H. Phase 8: Reporting, Audit Readiness, and Metrics

The program produces periodic governance reports including:

- Coverage metrics by platform and domain
- PII type distribution and high-risk findings
- Scan freshness and system compliance status
- Trend metrics on false positives/negatives and remediation closure [6], [9]

Audit readiness is supported through immutable scan logs, rule version tracking, and governed evidence capture.

VIII. PERFORMANCE EVALUATION AND RESULTS

The scalable PII discovery framework was evaluated against four primary dimensions: coverage completeness, detection accuracy, runtime performance, and operational impact across heterogeneous sources (mainframe, SAP, RDBMS, and unstructured repositories). The evaluation goal was to verify that enterprise-scale discovery can be executed continuously with acceptable resource usage while producing consistent, governance-ready outputs [4], [9].

A. Coverage and Completeness

Coverage was measured as the percentage of onboarded systems and repositories successfully scanned within the defined scan cadence (full or incremental). The framework achieved strong coverage across structured systems due to standardized connector onboarding and scan orchestration. The largest improvements were observed in environments where PII previously remained undiscovered in analytics copies and replicated datasets, demonstrating the effectiveness of catalog integration and lineage-driven prioritization [4], [10].

For unstructured systems, completeness improved when incremental crawling and repository delta detection were applied, reducing missed content caused by file movement or new folder creation. High-risk repositories (shared drives, customer correspondence folders) benefited most from risk-based scanning policies [2], [8].

B. Detection Accuracy (Precision and Recall Trends)

Detection accuracy was evaluated using steward-reviewed sampling sets across each source type. Structured systems achieved the highest precision for well-defined identifiers (email, phone number formats, national ID patterns) when pattern rules were combined with metadata inference and domain constraints. SAP environments saw notable reductions in false positives when semantic mapping via SAP dictionary metadata was used rather than pattern-only scanning [6], [11].

In mainframe scans, accuracy depended significantly on the availability and correctness of record layouts. Where layout definitions were maintained, field-level identification accuracy improved substantially, reducing misclassification caused by fixed-width encodings and overloaded fields [6], [10]. In cases with incomplete layouts, confidence scoring and stewardship review queues were essential to avoid over-classification.Unstructured systems exhibited the greatest variability. Pattern-only detection produced higher false positives, but contextual rules and confidence scoring improved precision by requiring proximity to sensitive keywords and cross-entity co-occurrence (for example, name with address and date of birth indicators). Selective OCR improved recall for scanned documents but introduced additional processing cost, reinforcing the value of risk-based OCR triggers [2], [8].

C. Runtime Performance and Scalability

Runtime performance was assessed under typical enterprise scanning patterns:

- Sampling-based scans for high-volume structured tables reduced resource usage while maintaining high detection reliability for high-confidence identifiers.
- Incremental scanning reduced overall runtime by prioritizing changed datasets rather than repeatedly scanning full estates.
- Parallel task execution enabled high throughput without saturating any single system, supported by throttling and scan window controls [6], [10].

For unstructured repositories, performance was primarily influenced by file parsing complexity, attachment depth in email archives, and OCR usage. The framework maintained stable throughput when parsing was staged (metadata extraction → text extraction → detection scoring), allowing expensive steps such as OCR to be applied only where needed.

D. Operational Impact and Production Safety

A critical evaluation goal was minimizing disruption to production systems. The framework achieved low operational impact through:

- Read-only access patterns and governed scan windows
- Query throttling and row-level sampling for RDBMS scans
- Batch-based extraction strategies for mainframe datasets
- Controlled SAP discovery approaches aligned with approved interfaces and extraction jobs [6], [11]

This reduced the likelihood of performance degradation on core transactional workloads.

E. Governance Readiness and Actionability

The findings produced governance-ready outputs by normalizing results into a centralized PII index, with consistent taxonomy tags, confidence scores, ownership assignments, and scan freshness timestamps. This enabled:

- Prioritization of high-risk datasets for remediation
- Faster security reviews and privacy compliance assessments
- Improved data access control alignment
- Better downstream support for masking and privacy-right operations [4], [9]

Overall, the evaluation confirms that scalable PII discovery is achievable across highly heterogeneous enterprise environments when detection methods are hybrid, scanning is orchestrated and incremental, and findings are integrated into governance workflows.

IX. ENTERPRISE CASE STUDY

This section presents a representative enterprise deployment of the proposed PII discovery framework across a heterogeneous environment consisting of mainframe systems, SAP landscapes, multiple RDBMS platforms, and large unstructured repositories. The case study demonstrates how unified discovery improves coverage, reduces compliance risk, and enables governance teams to operationalize findings through consistent taxonomy and remediation workflows [4], [9].

A. Environment Overview

The enterprise operated across multiple business units with legacy and modern platforms:

- Mainframe: core customer and transaction datasets stored in fixed-width formats, supporting operational workloads.
- SAP ERP: customer, vendor, and HR master data distributed across complex table structures and replicated reporting environments.
- RDBMS: operational databases supporting digital channels, servicing platforms, and internal applications, plus replicated analytics copies.
- Unstructured repositories: shared drives, document management platforms, policy repositories, email archives, and scanned document collections.

The enterprise faced growing privacy obligations and security controls requiring accurate identification of PII across both structured and unstructured assets.

B. Pre-Implementation Challenges

Before adopting the framework, the organization's PII inventory was incomplete and inconsistent due to:

- Siloed discovery efforts performed independently by system teams with different definitions of PII.
- Limited mainframe metadata and missing layout documentation for some datasets, leading to uncertainty about which fields contained PII.
- SAP semantic complexity, where PII appeared across master data and transactional objects without clear mappings.
- Shadow PII in analytics copies, where data replicated into warehouses and marts was not systematically scanned.
- Unstructured PII sprawl, particularly in shared drives and archived emails, where PII existed in documents and attachments without governance tagging [2], [8].
- These gaps made privacy compliance efforts slow, audit evidence difficult to assemble, and remediation inconsistent.

C. Deployment Approach

The enterprise deployed the framework in staged phases:

a) Phase 1: Structured systems baseline

Onboarded RDBMS and SAP systems first due to richer metadata and easier scanning controls. Established an enterprise PII taxonomy and rule catalog, then executed baseline scans with sampling for high-volume tables.

b) Phase 2: Mainframe integration

Introduced mainframe layout-driven parsing using available record definitions, then prioritized high-risk datasets for deeper scanning. Steward review workflows were used to validate ambiguous findings and refine mappings [6], [11].

c) Phase 3: Unstructured expansion

Onboarded file shares, document repositories, and email archives. Implemented staged parsing and confidence-based classification to control false positives, using selective OCR for high-risk scanned documents [2], [8].

d) Phase 4: Governance operationalization

Integrated findings with a central catalog and index, assigned owners, and launched remediation workflows such as access restriction, retention tagging, and masking requirements for downstream replicas [9], [10].

D. Outcomes and Observed Benefits

After implementation, the enterprise reported:

- Improved completeness across the estate, particularly in analytics replicas and SAP reporting environments that were previously excluded from scanning [4], [10].
- Higher consistency of classification, as all systems adopted a single taxonomy and confidence scoring model.
- Reduced operational overhead, because teams no longer ran separate discovery processes per system; instead, scans were orchestrated centrally.

- Faster compliance readiness, enabling quicker responses for risk assessments and privacy operations due to centralized indexing and ownership mapping [9], [10].
- Better unstructured risk visibility, with tagged repositories and prioritized remediation for high-risk document collections and shared drives [2], [8].

E. Key Lessons Learned

Two practical lessons were particularly significant:

- Lineage awareness drives completeness. Many PII risks originated in replicated datasets and derived marts. Discovery improved notably once replicas were treated as first-class scanning targets rather than optional systems.
- Confidence triage is essential for unstructured data. Steward review workflows helped prevent over-classification while ensuring that high-confidence findings drove immediate remediation actions.

Overall, the case study confirms that scalable PII discovery becomes operationally sustainable only when it is unified across platforms, orchestrated for minimal production impact, and integrated into governance workflows for actionability and audit readiness [4], [9].

X. DISCUSSION

The results and case study indicate that scalable PII discovery is most effective when treated as an enterprise operating capability, rather than a one-time scan. A key observation is that heterogeneous environments demand platform-aware discovery, because mainframe fixed-width encodings, SAP semantic object models, RDBMS schema breadth, and unstructured content variability require different scanning strategies and validation controls. The hybrid detection approach, combining patterns, metadata inference, and context scoring, reduces false positives and improves overall coverage compared with pattern-only scanning, especially in SAP and unstructured repositories where semantics matter [4], [6].

A second insight is that discovery becomes materially more valuable when integrated with central governance services, including a taxonomy-driven catalog, ownership mapping, confidence scoring, and remediation workflows. In practice, discovery outputs that are not linked to ownership, risk priority, and control enforcement often remain unused. Catalog integration and lineage-aware tracking help address a common enterprise issue: shadow PII in replicated datasets, analytics marts, and exports that are not covered by system-by-system scanning [9], [10].

The evaluation also highlights the importance of production-safe scanning. Read-only access patterns, throttling, sampling for high-volume tables, and controlled batch extraction windows for mainframes can achieve continuous discovery without impacting mission-critical workloads. For unstructured sources, staging the pipeline (metadata → text extraction → detection) keeps throughput stable and allows expensive steps such as OCR to be applied selectively [2], [8].

Finally, governance teams benefit from a confidence-based triage model. Unstructured PII discovery is inherently noisy, and stewardship review queues are necessary to prevent over-classification while still capturing high-risk content. This combined model improves operational trust in discovery results and supports audit defensibility by documenting rule versions, scan timestamps, and classification rationale [4], [9].

XI. LIMITATIONS

The results and case study indicate that scalable PII discovery is most effective when treated as an enterprise operating capability, rather than a one-time scan. A key observation is that heterogeneous environments demand platform-aware discovery, because mainframe fixed-width encodings, SAP semantic object models, RDBMS schema breadth, and unstructured content variability require different scanning strategies and validation controls. The hybrid detection approach, combining patterns, metadata inference, and context scoring, reduces false positives and improves overall coverage compared with pattern-only scanning, especially in SAP and unstructured repositories where semantics matter [4], [6].

A second insight is that discovery becomes materially more valuable when integrated with central governance services, including a taxonomy-driven catalog, ownership mapping, confidence scoring, and remediation workflows. In practice, discovery outputs that are not linked to ownership, risk priority, and control enforcement often remain unused. Catalog integration and lineage-aware tracking help address a common enterprise issue: shadow PII in replicated datasets, analytics marts, and exports that are not covered by system-by-system scanning [9], [10].

The evaluation also highlights the importance of production-safe scanning. Read-only access patterns, throttling, sampling for high-volume tables, and controlled batch extraction windows for mainframes can achieve continuous discovery without

impacting mission-critical workloads. For unstructured sources, staging the pipeline (metadata → text extraction → detection) keeps throughput stable and allows expensive steps such as OCR to be applied selectively [2], [8].

Finally, governance teams benefit from a confidence-based triage model. Unstructured PII discovery is inherently noisy, and stewardship review queues are necessary to prevent over-classification while still capturing high-risk content. This combined model improves operational trust in discovery results and supports audit defensibility by documenting rule versions, scan timestamps, and classification rationale [4], [9].

XII. FUTURE SCOPE

Future enhancements can strengthen coverage, intelligence, and automation. A key area is improving unstructured detection using more advanced language-based techniques, including entity recognition with contextual validation and improved document-type classification. Combining these methods with rule-based controls can improve recall while preserving auditability.

Another direction is expanding continuous discovery through event-driven triggers, such as scanning newly created or modified datasets and files in near-real-time. This reduces inventory staleness and supports rapid risk detection.Enterprises can also strengthen lineage awareness by integrating discovery results with data pipeline metadata, enabling automated detection of replicated PII across feature stores, reporting extracts, and machine learning datasets. Risk scoring can be enhanced by incorporating access telemetry and exposure context (e.g., public share links, overly broad permissions) to prioritize remediation.

Finally, future work can focus on tighter integration between discovery and controls, including automated masking recommendations, policy-as-code enforcement, and self-service governance workflows for data owners to validate and remediate findings efficiently [4], [9].

XIII. CONCLUSION

This paper presented a scalable, unified framework for PII discovery across heterogeneous enterprise environments, including mainframe datasets, SAP landscapes, RDBMS platforms, and unstructured repositories. By combining platform-aware connectors, distributed scan orchestration, hybrid detection techniques, and centralized indexing with governance integration, the framework provides consistent classification and operationally sustainable discovery at enterprise scale.

Evaluation findings and the enterprise case study demonstrate that the approach improves coverage completeness, enhances detection accuracy through combined signals, maintains production-safe scanning performance, and produces governance-ready outputs that support remediation and audit readiness. While limitations remain in layout dependency, SAP access constraints, and unstructured OCR variability, the proposed framework establishes a practical foundation for continuous PII discovery and privacy risk reduction across modern and legacy data estates.

XIV. REFERENCES

[1] California State Legislature, “California Consumer Privacy Act of 2018 (CCPA),” Civil Code §1798.100 et seq., 2018.
[2] European Union, “General Data Protection Regulation (GDPR),” Regulation (EU) 2016/679, 2018.
[3] NIST, Guide to Protecting the Confidentiality of Personally Identifiable Information (PII), NIST SP 800-122, 2010.
[4] NIST, Security and Privacy Controls for Information Systems and Organizations, NIST SP 800-53 Rev. 5, 2020.
[5] NIST, Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0, 2020.
[6] ISO/IEC 27018, Code of Practice for Protection of PII in Public Clouds Acting as PII Processors, ISO, 2019.
[7] ISO/IEC 27701, Extension to ISO/IEC 27001 and ISO/IEC 27002 for Privacy Information Management, ISO, 2019.
[8] A. Cavoukian, Privacy by Design: The 7 Foundational Principles, Information and Privacy Commissioner of Ontario, 2011.
[9] DAMA International, DAMA-DMBOK: Data Management Body of Knowledge, 2nd ed., Technics Publications, 2017.
[10] PCI Security Standards Council, PCI DSS: Requirements and Testing Procedures, v3.2.1, 2018.
[11] R. J. Anderson, Security Engineering: A Guide to Building Dependable Distributed Systems, 2nd ed., Wiley, 2008.
[12] M. Bishop, Computer Security: Art and Science, 2nd ed., Addison-Wesley, 2018.
[13] D. Loshin, The Practitioner’s Guide to Data Quality Improvement, Morgan Kaufmann, 2010.
[14] Apache Tika Project, “Apache Tika: Content Analysis Toolkit,” Apache Software Foundation Documentation, 2021.
[15] Apache Spark Project, “Apache Spark: Unified Analytics Engine,” Apache Software Foundation Documentation, 2021.
[16] IBM, Data Governance and Privacy Management for Hybrid Cloud, IBM Redbooks, 2020.
[17] Microsoft, Data Protection and Privacy in Azure, Microsoft Documentation, 2021.
[18] Amazon Web Services, Data Protection and Privacy Best Practices, AWS Whitepaper, 2021.
[19] Oracle, Data Governance and Compliance for Enterprise Data Platforms, Oracle Documentation, 2021.
[20] SAP, Data Protection and Privacy Guide, SAP Documentation, 2021.