Multimodal eDiscovery

Pentaho Data Platform

Hitachi Vantara Federal
Date: July 2023

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

Table of Contents

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

Multimodal eDiscovery

eDiscovery, short for electronic discovery, refers to the process of identifying, collecting, and producing electronically stored information (ESI) for legal proceedings or investigations. It involves the use of technology to search, review, and analyze vast amounts of digital data to find relevant information that can be used as evidence in litigation, regulatory compliance, data breach impact analysis or internal investigations related to misconduct or policy violations. The term was coined in early 2000s with increasing reliance on software IT systems and the digital information being generated. It has since become an established field within the legal domain.

In today's digital age, a significant amount of information is created and stored electronically, including emails, documents, databases, social media posts, instant messages, and other forms of digital communication, including audio, video and images. eDiscovery helps legal professionals and investigators efficiently manage and sift through this multimodal[1] electronic data to find relevant evidence. While more convenient in contrast to traditional paper-based investigation, it also poses unique challenges due to the volume, complexity, and evolving nature of electronic information.

In the public sector, including the US Federal Government, there are prescribed guidelines, frameworks, and best practices for the investigation process. *eDiscovery and the Freedom of Information Act (FOIA) are connected through their shared objective of accessing and retrieving electronic information.* The data collection process is where the similarity ends. The distinction is that they operate within distinct legal frameworks and have different purposes and scopes. eDiscovery is focused on the specific legal case or investigation, while FOIA provides the citizens with the right to access government records and information. Sometimes they get collated and thus, the need to highlight the contextual difference here.

eDiscovery	FOIA
eDiscovery is guided by specific rules and procedures established by the legal system, such as the Federal Rules of Civil Procedure (FRCP) in the United States. The process is typically governed by court orders, and parties involved in the legal matter have obligations to preserve and produce relevant electronic evidence.	FOIA has its own set of procedures and timelines for submitting requests to government agencies. Agencies must review the requested records and determine if any exemptions apply, such as for national security, personal privacy, or confidential business information. They are required to provide the requested information or justify any denials.

The Federal Judiciary, including the United States Supreme Court and various district and appellate courts, establishes and interprets the rules governing eDiscovery in federal litigation. The Federal Rules of Civil Procedure (FRCP), specifically Rule 26 "Duty to Disclose; General Provisions Governing Discovery" and Rule 34 "Producing Documents, Electronically Stored

[1] *Multiple modalities or types of data analysis combining various methods, such as text analytics, visual analytics, audio/video/image analytics, metadata analysis, machine learning and more.*
Dictionary definition: also multi-modal; involving several ways of operating or dealing with sosmething

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

Information, and Tangible Things, or Entering onto Land, for Inspection and Other Purposes", provide guidelines for eDiscovery processes in federal courts.

National Institute of Standards and Technology (NIST), a non-regulatory agency within the U.S. Department of Commerce, has published guidelines and frameworks related to eDiscovery. For example, NIST Special Publication 800-209 "Security Guidelines for Storage Infrastructure" provides best practices for the preservation of digital evidence.

Agencies such as Department of Justice (DOJ), Federal Trade Commission (FTC), Federal Trade Commission (FTC), Securities and Exchange Commission (SEC), Internal Revenue Service (IRS), Internal Revenue Service (IRS), Federal Communications Commission (FCC), Department of Defense (DoD), and various regulatory bodies plays a significant role in criminal investigations and enforcement actions.

Here are a few examples of eDiscovery-related initiatives and guidelines from the DoD:

DoD Discovery Metadata Specification (DDMS) is a standardized metadata framework to facilitate the exchange of electronic discovery metadata among the DoD components, agencies, and legal entities. It is a XML data-encoding specification for Intelligence Content Discovery and Retrieval (CDR) that defines detailed implementation guidance for using result sets in service responses applicable to the IC and Department of Defense (DoD) and information produced by, stored, or shared within and between the IC and DoD.

Defense Office of Hearings and Appeals (DOHA) handles personnel security clearance appeals during which eDiscovery processes may be conducted to collect, review, and produce electronic evidence during security clearance adjudications.

U.S. Army Legal Services Agency (USALSA) is the legal services agency of the U.S. Army, provides legal support to Army activities. It may handle eDiscovery processes in matters involving the Army, such as military justice proceedings, administrative actions, or contract disputes.

Defense Contract Management Agency (DCMA) is responsible for contract management and administration. In the context of contract disputes and investigations, eDiscovery processes may be utilized to gather and analyze relevant electronic evidence.

Various criminal investigative organizations within the DoD, such as the **Defense Criminal Investigative Service (DCIS)** and the **Naval Criminal Investigative Service (NCIS)**, engage in eDiscovery activities during investigations of criminal offenses under military jurisdiction.

Query Categorization

The questions posed as part of the eDiscovery or FOIA request will be broken down further into one or more queries against various data sources, and the dots are connected to be compiled into a presentable, formatted response for user purview or further investigation. The queries will fall broadly into the following categories:

Keyword queries search for specific words or phrases within the electronic data. For example, a query might include keywords related to a specific subject matter, individuals involved, or events of interest. This can help retrieve documents containing those keywords, such as emails, documents, databases, or other forms of digital content.

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

Date-range queries help narrow down the search results by specifying a specific timeframe.

File type queries involve searching for specific file types. This can be useful when focusing on specific types of electronic evidence, such as searching for documents in emails, specific format such as PDF, images, or audio files.

Boolean queries allow for more complex search criteria by combining multiple keywords or conditions using logical operators such as "AND," "OR," or "NOT." or specific criteria for inclusion or exclusion.

Proximity queries search for terms or phrases that occur within a specific distance or proximity to each other. This can be useful when looking for specific contexts or relationships between terms.

Conceptual queries employ techniques like concept clustering or concept searching to identify documents or data related to specific concepts or themes. It helps uncover information beyond exact keyword matches by leveraging semantic analysis or machine learning algorithms.

eDiscovery Process

eDiscovery Process involves these tasks:

⇒ Identification
⇒ Preservation
⇒ Collection
⇒ Processing
⇒ Review
⇒ Analysis

(repeat from Identification to refine and/or gather additional relevant information)

⇒ Production

Identification: Determining where potentially relevant ESI exists, such as computer systems, servers, email archives, cloud storage, or mobile devices.

Preservation: Taking steps to ensure the integrity and preservation of the identified ESI to prevent tampering, alteration, or loss.

Collection: Gathering the identified ESI from various sources, including making forensic copies or acquiring data through remote access.

Processing: Transforming the collected data into a searchable and reviewable format, such as extracting metadata, text, and attachments from files.

Review: Analyzing the processed data to identify relevant information, often involving the use of specialized software or platforms for keyword searching, filtering, and categorization.

Analysis: Evaluating and organizing the relevant data to understand its context, relationships, and implications for the case or investigation.

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

Production: Presenting the identified and reviewed ESI in a suitable format for use in legal proceedings, which may include exporting, redacting sensitive information, and converting files to the required format.

eDiscovery Solution Architecture

Designing an ideal eDiscovery solution architecture depends on several factors, including the specific requirements of the organization, the volume of data involved, and the desired level of automation. While different organizations may have unique needs, the following components are generally considered important in an effective eDiscovery architecture:

Data Sources: Identify and understand the various data sources within the organization, such as email servers, file shares, databases, cloud storage, collaboration platforms, and individual devices. Ensure that the architecture can integrate with and collect data from these sources efficiently.

Data Collection: Implement mechanisms to collect data from the identified sources while maintaining data integrity and preserving metadata. This may involve techniques like forensic imaging, targeted collection, or remote data acquisition, depending on the situation.

Data Processing: Establish a scalable and robust data processing mechanism that can handle large volumes of data. This step involves extracting metadata, indexing content, and converting files into a searchable format handling different file types and extracting data from various formats.

Data Storage: Determine how the processed data will be stored and managed. This may involve the use of on-premise servers, cloud, or a combination of both. Ensure appropriate security measures are in place to protect sensitive data.

Data Analysis and Review: Integrate tools and platforms that enable efficient data analysis, review, and search capabilities. This can include advanced search functionalities, machine learning techniques for document clustering or predictive coding, and visualizations to aid in data understanding and decision-making.

Case Management: Implement a centralized case management system to track and manage the progress of each eDiscovery case. This system should provide features for assigning tasks, tracking timelines, maintaining audit logs, and facilitating collaboration among legal teams.

Security and Compliance: Incorporate robust security measures to protect sensitive and confidential data throughout the eDiscovery process. Consider encryption, access controls, data loss prevention, and compliance with relevant regulations like GDPR or HIPAA.

Reporting and Production: Develop mechanisms to generate reports and export data in a suitable format for production during legal proceedings. This may involve redacting sensitive information, converting files to specified formats, and ensuring chain of custody documentation.

Integration and Automation: Seek opportunities to integrate the eDiscovery architecture with other existing systems, such as legal case management, document management, or incident response tools. Automate repetitive tasks and workflows to improve efficiency and reduce manual effort.

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

Scalability and Flexibility: Design the architecture to be scalable, accommodating future growth and changing requirements. Consider the ability to handle increased data volumes, support new data sources, and adapt to evolving technology and legal standards.

Implementation Challenges

eDiscovery Challenges	
Data Volume and Variety	• Managing and processing large volumes of data from various sources, including structured and unstructured data, can be complex and time-consuming. Handling data from disparate or legacy system with non-standard access interfaces is often a non-trivial challenge. • The exponential growth of the data and it may continue to change as the information retrieval process is being applied. Properly preserving data and implementing legal holds to prevent data loss and intended\unintended modification prevention is essential.
Data Security and Privacy	• Ensuring proper access controls, encryption, and compliance with data protection regulations handling sensitive and confidential data can be complex to manage and to stay current. • The legal and regulatory landscape around data ecosystem is continually and rapidly evolving.
Ecosystem	• Vendor selection process has to be exhaustive in order to account for data processing capabilities, resource utilization (CPU, GPU, Storage) and customer support. • Training users on the functionalities and workflows of eDiscovery software is crucial for its continued and effective use.
Cost and ROI	• Licensing fees, implementation costs, and ongoing maintenance expenses can build up easily, particularly for organizations with limited or tight budgets. • Technical expertise required to deal adeptly with the data volume and variety are hard to find and expensive.

Table 1 Categories of Challenges

The challenges increase manifold in multimodal eDiscovery environments. The complexity of the data and metadata (“data about data” – that is, contextual, structural, administration, rights and preservation attributes associated with the data content) renders the pre-processing and normalization as computationally intensive. Organizations must have the necessary infrastructure and scalability to handle the data volume and complexity effectively.

Validation of the collected information for eDiscovery response may involve complex algorithms and models that are not always easily interpretable. Not understanding how the different modalities interact may introduce potential data/algorithmic biases and inaccuracies. Irrespective, the organization should ideally operate within the bounds of legal and regulatory requirements, such as data protection laws, confidentiality obligations, and restrictions on data use. Organizations must incorporate robust security measures to protect the multimodal data, including encryption, access controls, and automated monitoring of data flows.

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

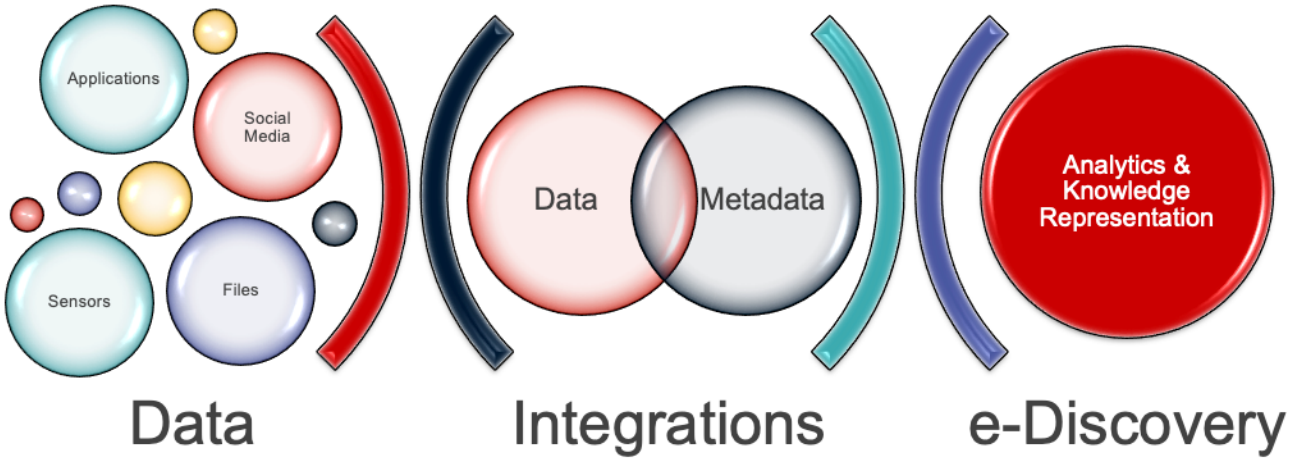

Figure 1 Data, Metadata and Knowledge Representation

Hitachi Analytics for eDiscovery

Hitachi is a global conglomerate that offers a wide range of products and services across various industries. Hitachi provides a variety of technologies and solutions that can be utilized in the streamlined eDiscovery process for law enforcement, legal investigations, FOIA-related analytics and regulatory compliance. Here are some relevant open-architecture low-code/no-code Hitachi products and services:

⇒ **Pentaho Data Platform** accelerates data onboarding with robust dataflow orchestration with broad connectivity to virtually any data source or application, drag-and-drop interface to create data pipelines and templates that execute edge to cloud.

⇒ **Pentaho Data Catalog** with its patented AI-Driven Discovery utilizes unique data fingerprinting to automate discovery and classification of structured, semi and unstructured data. It establishes via build and/or import a taxonomy of business terms in a glossary format and establishes relationships. Data Quality is assessed via data provenance and lineage, critical quality metrics and business rules across the business. The user-friendly search provides a simplified experience across all data assets.

⇒ **Hitachi Content Platform (HCP)** is an object storage platform that enables organizations to store, manage, and access large volumes of unstructured data. It provides scalable and secure storage infrastructure for retaining and accessing data during eDiscovery.

⇒ **Hitachi Data Ingestor (HDI)** is a cloud-based file services platform that allows organizations to securely access and share files across multiple locations and devices. It can facilitate data collection and collaboration during the eDiscovery process.

⇒ **Hitachi Content Intelligence (HCI)** is a data analytics and search platform that enables organizations to gain insights from their unstructured data. It can help with data analysis, content indexing, and searching relevant information during eDiscovery.

⇒ **Hitachi Digital Evidence Management (DEM)** is a comprehensive solution for managing digital evidence in law enforcement and legal environments. It provides secure storage, advanced search capabilities, and evidence tracking features.

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

⇒ **Hitachi Unified Compute Platform (UCP)** is a converged and hyper-converged infrastructure solution that integrates storage, computing, and networking resources. It can be leveraged to support the underlying infrastructure required for eDiscovery processes.

⇒ **Hitachi Consulting Services** assist organizations in various aspects of data management, information governance, and compliance. These services can include guidance on eDiscovery strategies, process optimization, and technology implementation.

An example architecture to demonstrate high-level orchestration of both structured and semi- / unstructured digital content for eDiscovery is included below. *Note: This is only provided for reference – it will vary for different ecosystems based on additional factors such as available toolsets, data sources, required analytics and typical nature of the posed queries.*

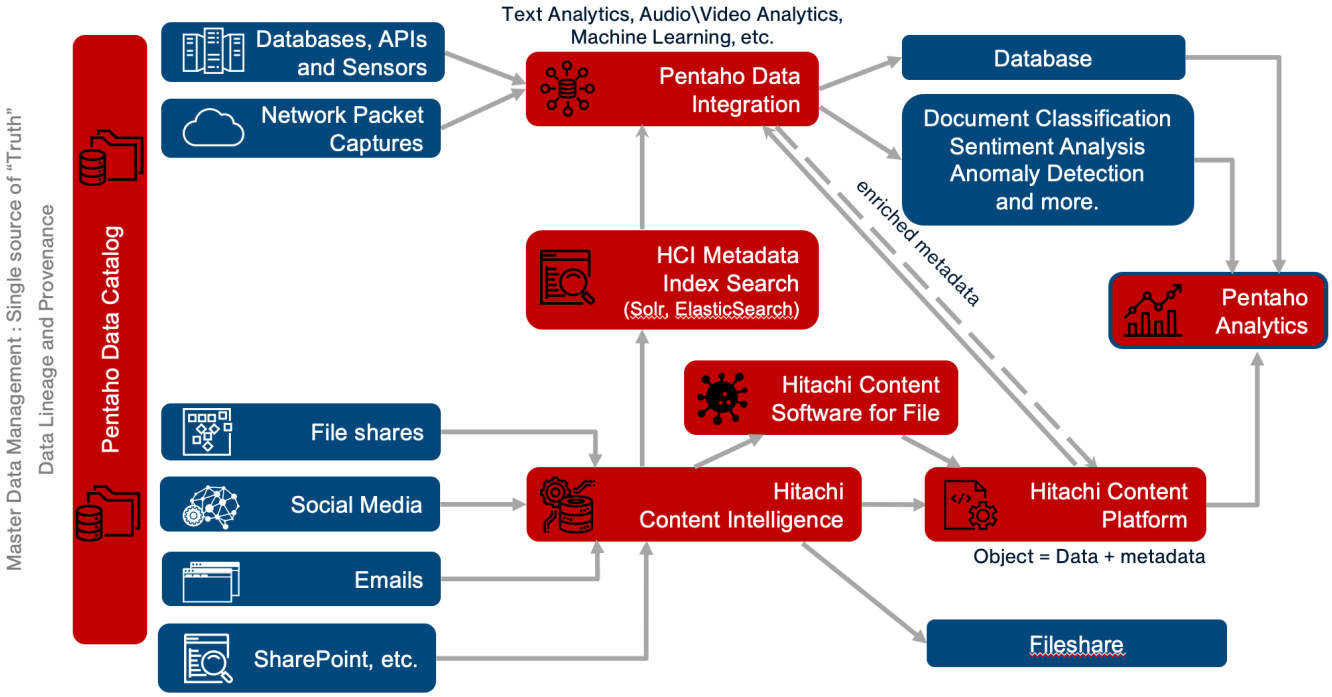

Figure 2 Structured and Unstructured Data Integration and eDiscovery

Machine Learning (ML) for eDiscovery++

Machine Learning algorithms can play a significant role in expediting multimodal eDiscovery by automating certain tasks and assisting with data analysis. Here are some machine learning algorithms commonly used in multimodal eDiscovery to enhance efficiency and whose models can be both trained and used as part of the dataflow in the Pentaho Data Platform data transformations:

Document Classification: Machine learning algorithms, such as Support Vector Machines (SVM), Naive Bayes, or Random Forest, can be employed to classify documents based on their relevance or responsiveness. These algorithms learn from labeled training data and can prioritize document review, reducing manual effort and accelerating the identification of relevant documents.

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.

Text Analytics: Natural Language Processing (NLP) techniques, combined with machine learning algorithms, can extract relevant information from text-based data. Named Entity Recognition (NER), sentiment analysis, topic modeling, and language models (e.g., word embeddings, Transformers) can help uncover key entities, sentiments, and themes within documents, aiding in faster analysis and understanding of textual data.

Clustering: Unsupervised learning algorithms like k-means clustering or Latent Dirichlet Allocation (LDA) can group similar documents together based on their content. Clustering assists in organizing large datasets, identifying document clusters, and discovering hidden patterns or topics within the data.

Image and Video Analytics: Convolutional Neural Networks (CNNs) are often employed for image and video analysis in multimodal eDiscovery. These deep learning algorithms can detect objects, identify faces, extract visual features, and analyze image or video content to support investigations involving multimedia evidence.

Anomaly Detection: Machine learning algorithms can be utilized for anomaly detection to identify outliers or abnormal patterns within multimodal data. Anomalies may indicate potential data breaches, unusual behavior, or fraudulent activities, enabling organizations to promptly address such incidents.

Recommender Systems: Recommender systems leverage machine learning algorithms, such as collaborative filtering or content-based filtering, to suggest relevant documents, similar cases, or potential connections based on the user's interactions or data characteristics. This assists in knowledge discovery, improving efficiency and aiding decision-making.

It's important to note that the successful application of machine learning algorithms requires proper training data, algorithm selection, and careful validation to ensure accuracy, fairness, and completeness. Domain expertise, data quality, and ongoing monitoring of the algorithms are essential to maintain the reliability and effectiveness of the results irrespective of data or model shift.

About Hitachi Vantara Federal

Hitachi Vantara Federal is the trusted leader in mission-centric data solutions for the Federal government. We're a collaborative, full-service company with longstanding OT/IT roots. We empower data-driven insight with a deep bench of integrated partners — advancing Federal customer missions regardless of their data maturity levels. Hitachi Vantara Federal is a FOCI-mitigated subsidiary of Hitachi Vantara and holds a TS facility clearance.

We are familiar with and adhere to the NIST AI Risk Management Framework, DoD Ethical Principles for Artificial Intelligence, and the fast-emerging Responsible and Trustworthy AI Guidelines from various Federal agencies in the development and implementation of our solutions. Our analytics products are listed within the DoD CDAO Tradewinds Marketplace as well as part of DIA Needipedia (ID# = 06APR2023_01MAR02023_HITACHI-VANTARA_6-1-3_ENG_SUM).

Visit us at: www.hitachivantarafederal.com

© 2023 Hitachi Vantara Federal, Corporation. All Rights Reserved.