Versa's Approach to scaling an AI/ML data pipeline for DLP

July 2024

General Disclaimer

Although Versa Networks has attempted to provide accurate information in this guide, Versa Networks does not warrant or guarantee the accuracy of the information provided herein. Versa Networks may change the programs or products mentioned at any time without prior notice. Mention of non-Versa Networks products or services is for information purposes only and constitutes neither an endorsement nor a recommendation of such products or services or of any company that develops or sells such products or services.

The Importance of Data Classification and Processing for Artificial Intelligence in Data Loss Prevention

Artificial Intelligence (AI) can be a powerful tool in supporting Data Loss Prevent initiatives. AI-powered tools can scan vast amounts of data to detect patterns and categorize information based on predefined criteria. AI can continuously monitor data flows, ensuring that newly created or modified data is classified correctly in real-time

The effectiveness of an AI approach to support these initiatives will be highly dependent on effective data classification processes – anchored to categorizing data based on its level of sensitivity and the potential impact of its unauthorized disclosure. This foundational step is crucial for implementing robust DLP strategies, as it helps organizations understand what data needs protection and prioritize security efforts accordingly.

Understanding and dealing with data issues are further enhanced through categorization approaches. For instance, questions like “How do I find Personally Identifiable Information (PII)?” or “How do I secure my intellectual property?” are common starting points for discussions on confidential data. These questions underline the need for clear data classification policies, which often include predefined classes to streamline the categorization process.

Typically, data will map to pre-defined classes, which simplifies the task of identifying and securing sensitive information. For example, data classification policies might categorize data into classes such as public, internal, confidential, and restricted, providing specific examples for each category (see Table 1 below). This structured approach ensures that all data, especially the most sensitive and confidential types, is appropriately identified and protected.

Classification	Description	Examples
Public	Non-sensitive data	Marketing materials, publicly available reports
Internal	Data intended for internal use only	Internal emails, internal reports
Confidential	Sensitive data	Employee records, financial statements, strategic documents
Restricted	Highly sensitive data	PII, intellectual property, legal documents, trade secrets

Table 1: Example Data Classification Categories

By adopting a robust data classification and processing framework, organizations can better manage and protect their sensitive information. This is crucial not only for compliance with data protection regulations but also for safeguarding against data breaches and ensuring the overall security of the organization's digital assets.

AI/ML data pipeline overview for DLP:

Designing a data pipeline for image processing and classification involves several stages, from data ingestion to serving the results. A high-level breakdown of such a pipeline would involve the following

- **Data Ingestion.** This is the stage where data is acquire or imported. Typical data for training models for data loss prevention is not available publicly and difficult to procure. As data available in public datasets is minimal and templatized.
    - Data sources: From public datasets, cloud storage (S3/GCP), APIs.
    - Data synthesis: Generating synthetic datasets from real ones using GenAI.

- **Data Preprocessing**
    - **Cleaning:** Identify and handle missing, corrupt or unusable images. Re-run the data sourcing step if a significant percentage of images in a category are unusable.
    - **Upscaling:** Use advanced algorithms for upscaling images and some images maybe useful but not high resolution.
    - **Normalization:** Scale pixel values to a range (e.g., 0-1).
    - **Augmentation:** In case the dataset is small, a process to artificially enlarge the dataset by creating modified versions of images by running transformations on the images would be required.
    - **Splitting:** Divide the data into training, validation, and test sets.

- **Feature Extraction.** Depending on the method or model, the process might need to extract specific features from the images.
    - Techniques like edge detection, color histograms, bounding box detection would commonly be used.

- **Model Training**
  - **Architecture Selection:** For our use-case of classifying images into categories as well as performing text extraction like OCR, algorithms help to identify text bounding boxes, text extraction using transformer models. Avoid using Tesseract as it is slower and doesn't perform as well as transformer models tuned for vision applications.
  - **Training:** Feed the training data and adjust model weights using backpropagation and optimization techniques. This would be done as part of the fine-tuning step where to update weights in the last layer.
  - **Validation:** Use a separate validation set to tune hyperparameters and avoid overfitting.

- **Model Evaluation.**
  - **Testing:** Evaluate the model's performance on a held-out test set.
  - **Metrics:** Compute accuracy, precision, recall etc., against a known labeled image test dataset.
  - **Dashboard:** The training and validation pipeline is augmented with a dashboard to collect metrics, tag it as well as visualize it using tools such as Prometheus/Grafana, WandB, MLFlow.

- **Deployment.**
  - **Model Export:** Convert the model to a format suitable for deployment.
  - **Serving:** Deploy the model on an on-prem cluster or public cloud cloud platform. The model artifacts can be stored in artifact registry. When the model is deployed in an on-prem Kubernetes cluster or GKE or EKS, the corresponding model with the appropriate version is pulled from the artifact registry and loaded for inference.
  - **API:** An API endpoint (e.g., using FastAPI) can be deployed so that devices or API Data Protection services can send images for categorization of the image and receive prediction as to whether the image contains any sensitive/confidential information and the relevant category etc. This service may be made available as a standalone service or integrated into a cloud service.

- **Monitoring & Maintenance:**
  - **Performance Monitoring:** Regularly check the model's real-world performance
  - **Retraining:** Periodically retrain the model with fresh data in case the model can be updated (partially training step) or on a monolithic dataset (with additional/fresh data samples) or if its performance drops.

- **Feedback Loop:** Augment the data pipeline by implementing a system/dashboard to collect and visualize metrics from the inference tasks on the served models. This improves the models based on training with the additional data which was found to be incorrectly classified.

AI/ML DLP service deployment components

Versa offers a multi-engine, multi-stage AI/ML Data Loss Prevention (DLP) service designed for flexible deployments, ensuring high accuracy and scalability. The data pipeline begins with initial file analysis and invokes advanced DLP actions if no known patterns are detected. It handles a spectrum of file types, securely transferring them to a cloud-based DLP service. Files are then queued for batch processing, utilizing Kubernetes clusters for efficient detection and classification in both cloud and on-prem environments.

To learn more about best practices or how to integrate Versa's advanced DLP service into your securitye architecture, please reach out to a Versa representative.

Learn more at www.versa-networks.com

Follow us @versanetworks.

2550 Great America Way, Suite 350 | Santa Clara, CA | 95054