NLP (Natural Language Processing) Driven Data Mining and Data Breach Review Application

athenaworks

Index

Client Background

Industry leader specializes in Data Breach Response, PII & PHI Detection, Data Security, Sensitive Information Detection, Privacy, Data Subject Access Requests, Incident Response, Data Protection, GDPR, CCPA and Cybersecurity.

Team Makeup

Diversity Metrics

5 Women | **13** Men
Geographically Located: **India & USA**
Ages: **27 - 45 yrs old**

Team

1 Project Manager
1 UX/UI
2 Backend Dev
5 Frontend Dev
2 Data Sciences
6 QA
1 Cloud Infra & DevOps
18 Engineers

Technology Stack

Framework: React, Node.js, .NET

Programming Languages: Python, JavaScript, C#

Environment & Operating Systems: Ubuntu, Window, Docker

Web Server: Nginx

Databases: MongoDB, Elasticsearch

Project Overview

The project consisted of building a **Software-as-a-Service App**, considered as one of the first known platforms of its kind dedicated to the discovery of sensitive information, data breach review, and building lists of affected individuals for data breach notification. The app is designed for service providers and law firms responding to data breaches in compliance with GDPR, FERPA, PCI, HIPAA, CCPA or other privacy laws or regulations.

It's a cloud-based app that allows law firms, legal service providers, and incident response teams **upload massive amounts of breached data for analysis**. When the documents are uploaded, they are analyzed, PII is detected, and a report is generated. Users then have the ability to go through affected documents to look for affected individuals.

It comes with a sensitive information discovery capability to help teams respond faster, more accurately, and with less risk and resources than would otherwise be possible using traditional discovery approaches.

The app combines many new and existing discovery techniques into a unique workflow specifically designed to solve the problems discovering personally identifiable information (PII), protected health information (PHI), and student education records.

Challenges Presented / The Goals

Business Challenges

A Global cyber research firm predicts that cybercrime damages will cost the world $6 trillion annually by 2021, doubling from $3 trillion in 2015.

A couple of data breach & cybersecurity facts and figures for the year 2019 to 2021:

- Data Breach & Cybersecurity damage costs are predicted to hit $6 trillion annually by 2021
- Cybersecurity spending will exceed $1 trillion from 2017 to 2021

The Problem; users were forced to manually enter information found about affected individuals. They had no way of ensuring the data was accurate and no way to prevent duplicates, often relying on products such as Excel to hold the information.

The Goals

Mindful of the crippling cost and staggering volume of data breaches and cybersecurity events, the client's ultimate goal was to:

1. Develop a platform that transforms the way cyber teams and review companies perform data breach review.
2. Build a cloud-native app on the fastest tech stack available using the latest advances in AI, Machine Learning, and NLP to make Data Breach Discovery more tolerable.
3. Overcome the data processing using "legal tools" and relying on spreadsheets, custom forms, and hundreds of wasted man-hours trying to do the work manually.
4. Ultimately develop a purpose-built app for data breach discovery using AI with super user-friendly UI.

The Solution

Backed with a combined experience of 30 years in the field of eDiscovery, Information Governance, and Data Protection for the world's most sophisticated organizations, we collaborated with the client's founders to leverage their knowledge and produce an incredible platform to data mine and extract sensitive information from unstructured data.

Our team was well aware of the fact, that when an incident occurs, time is of the essence—and to combine millions of documents to find affected PII requires a new approach altogether and thus Emergent was designed to reduce time, cost, risk, and effort associated with the defensible discovery of personally identifiable data.

This **Cloud-Native, Scalable, & Secure App** can be briefly summarized:
Deployed on AWS | Autoscale Capacity | Encrypted at Rest | Encrypted in Transit | Containerized | Any Jurisdiction.

Solution Overview

- **Robust Upload:** Upload or import from S3, Google Drive, Dropbox, Office 365, or SFTP
- **Defensible Processing:** Process uploaded information using standard e-discovery methods
- **Active Lookahead:** Actively associate individuals with documents as the system learns
- **Detect & Classify:** Automatic PII/PHI detection using Machine Learning
- **Anomaly Detection:** AI-based detection of data anomalies, such as typos in Social Security Number
- **Entity Relating:** Link and relate entities to build lists of affected individuals/data subjects
- **MapAccel:** Map spreadsheet columns to entities and import entities while reviewing documents
- **Entity Resolution:** De-duplicate and normalize related entities, even with maiden/nicknames

Results

1

Accelerate Detection of Personally Identifiable Information

The app processes electronically stored information using proprietary algorithms pre-trained for sensitive information detection and extraction that goes well beyond the capabilities of regular expressions.

2

Assess Privacy Impact Within 72 Hours

The app's reports are designed to help assess the impact of the breach before reviewing the data. Use the analytics to cull and organize documents in preparation for data mining affected individuals.

3

Generate a List of Unique Individuals

The app's coding technology and workflow are designed to resolve the relationships between individuals and their elements found across multiple documents. Its ML model helps quickly extract, relate, and export a list of unique individuals.

Contact Us!

athenaworks.com

Athenaworks