WHITEPAPER

# Mitigating Data Privacy Risk Using Insight AnalytiX

## Efficiently Identifying PII in Unstructured Data Using Analytics

DATA DYNAMICS

## The Growing Problem with PII

In today's world, driven by the use of big data and online transactions, the use of PII (Personal Identifiable Information) has exploded, which has created major concerns with consumers about the privacy of their data. These concerns have translated into a number of regulations enacted into law by various regulatory bodies around the world.  The amount of data that can be classified as PII is fairly broad and can generally defined as the following:

> Any representation of information that permits the identity of an individual to whom the information applies to be reasonably inferred by either direct or indirect means. Further, PII is defined as information: (i) that directly identifies an individual (e.g., name, address, social security number or other identifying number or code, telephone number, email address, etc.) or (ii) by which an agency intends to identify specific individuals in conjunction with other data elements, i.e., indirect identification.
>
> U.S. Department of Labor

PII-related regulations may be driven at a state, national, and international level, with each state and country issuing their own rules regarding the handling of PII data and where that data can be stored (sovereignty laws, and while PII privacy laws and regulations are written for all businesses within a jurisdiction, some are written for specific industries.  Here is a list of the main PII regulations at the present time:

GDPR - Developed by the European Union, the General Data Protection Regulation was written in 2016, with enforcement beginning in 2018, in an effort to harmonize data privacy across Europe. It essentially establishes that data privacy is a fundamental right of all EU

CCPA  - The California Consumer Privacy Act  was signed into law in June 2018 and went into effect in January 2020 and is a state statute that is intended to improve the privacy rights of California residents.  While the law went into effect relatively recently, a more recent law, the California Privacy Rights Act, was created to tighten up and improve the protections.  The new initiative will take effect in January 2023, with an enforcement date of July 2023.

HIPAA  - The Health Insurance Portability and Accountability Act was signed into law during the Clinton Administration in 1996, and Title II of the Act was created to provide guidelines to help digitize the flow of healthcare data.  The key from a regulatory standpoint is that this law creates the policies and procedures for maintaining privacy and security of PII for healthcare and establishes criminal and civil penalties and fines for offenses.

## The Regulatory Costs of PII Regulations

Figure 1: The real costs of PII breaches

Organizations must handle the costs from multiple areas when dealing with the fall-out from multiple areas, as laid out in Figure 1. In a 2020 study from the Ponemon Institute and IBM, the average out-of-pocket cost for a security breach was almost $4 million. Additionally, it should be noted that 80% of those breaches included PII and, according to the study, healthcare was the top target of hackers. While the cost included out-of-pocket costs for lost business, fines, and litigation, it does not include the cost of the damage to the organization 's reputation.

So, what does this mean from a reputational cost perspective? While calculating that type of hard dollar cost can be difficult, there is some evidence to point that the long-term reputational costs could be significantly higher. For example, in 2017 the credit ratings agency Equifax was hacked, exposing PII for approximately 146 million people, and while the estimated out-of-pocket expenses is around $600 million, the company's stock price lost nearly $4 billion in market capitalization, a loss of nearly 7x the immediate cost.

Noting that most of the privacy regulations are relatively new and that the enforcement and penalties paid by companies are just beginning to be known, recent trends show that regulators are beginning to ramp up their efforts and that fines are beginning to be assessed. Under the rules set forth by the European Union, fines for violations related to improper handling of data under the GDPR can be substantial: "Art. 83(5) GDPR, the fine framework can be up to 20 million euros, or in the case of an undertaking, up to 4 % of their total global turnover of the preceding fiscal year, whichever is higher."

Here are a few recent examples that highlight the seriousness of the fines:

British Airways – was fined more than $240 M by British regulators due to exposure of PII during a data breach.
Marriott – was fined almost $130M by British regulators for another breach and theft of PII data.

In a recent article in Compliance Week, the situation was summed up this way: "The world's most stringent privacy law, the European Union's General Data Protection Regulation (GDPR), turned 2 years old on May 25. In those 24 months, the rules have put data privacy compliance on every board's agenda and have given Big Tech notice that their activities—and revenue streams—are under review."[1]

[1] Compliance Week, May 27, 2020, "Two Years in, GDPR Defined by Mixed Signals and Unbalanced Enforcement ", Neil Hodge

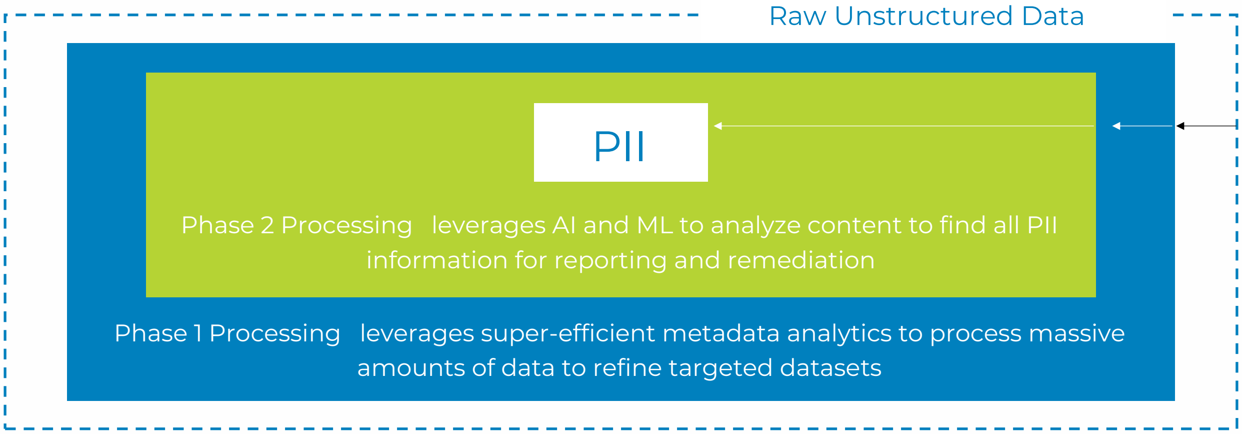
## Needle in the Haystack

In most organizations, unstructured data represents the vast majority of their data, has been growing between 20% to 30% per year, and will account for an estimated 80% of their storage capacity by 2025, according to IDC. To put it into perspective, the scale of unstructured data growth, the same study estimates the amount of data annually produced will reach 175 exabytes. This growth is driven by the continued digitization of business processes away from paper and the increase of different sources of information — IoT endpoint solutions such as sensors, imaging, and others — that are now being used for a wide variety of purposes to drive better, smarter business decisions. As CIO magazine recently noted: "Where a typical enterprise once took in structured data from mission-critical applications and stored it away, that same business will now need to handle many new varieties of unstructured data — think sensors, video feeds, and hardware telemetry, to name just a few."[2]

[2] CIO Magazine, July 8, 2020 "Drive Business Advantage from Your Unstructured Data ", Sandeep Singh

This massive growth compounds the issue of finding and protecting PII for unstructured data. Historically, most organizations have utilized point solutions that have been deployed on a departmental basis, creating silos of protection that leads to complexity, inconsistency, and inefficiency that can then drive up costs without really mitigating the risk. In today's digital world, the flow of information disperses data across the entire enterprise environment, and new regulations are forcing organizations to rethink their old ways of doing things and look for solutions that can meet these challenges.

## Better Efficiency: First, Separate the Wheat from the Chaff



Phase 2 Processing leverages AI and ML to analyze content to find all PII information for reporting and remediation

Phase 1 Processing leverages super-efficient metadata analytics to process massive amounts of data to refine targeted datasets

Figure 2: Efficient phased approach to PII identification

A better approach to tackling this problem is to deploy a simple, multi-step process to "separate the wheat from the chaff" and process the data more efficiently. This can be done by leveraging different processes that are best suited to handle the massive scale of the total unstructured dataset and identify a much smaller subset of the data for the next phase to identify the PII data, which requires more processing power. The benefits are impressive: 1) much faster processing of information dramatically reduces project time; 2) employing more intensive Phase 2 processing only for the smaller subset of identified data requires much less processing and infrastructure to find PII, reducing costs; and 3) by eliminating complex siloed solutions and tackling the entire unstructured dataset, you are able to employ a best practices approach that dramatically

## Flexible Architecture: Proven Enterprise Performance & Scalability

### User Interface

Application interface

User can login and run the scan on an analysis set by using multilevel logical expressions

Analyze processed analysis set

Export Report

### Orchestration

Coordinate between IA, SX, DSE, and CPE

Monitor Scans

Create Alerts

Generate Report

List NAS devices from StorageX

List files for an analysis from StorageX

### Data Science Engine

Analyze file content for Personal Identifiable information (PII).

Runs on top of Advanced Pattern recognition, Keyword search, and artificial Intelligence.

### Content Processing Engine

Streams files from NAS devices with CIFS(v2&v3)/NFS(v3) protocols.

Built-in OCR engine for images and scanned doc to extract content from it.

Figure 3: Flexible Insight AnalytiX stack

To process the massive unstructured datasets for most enterprises, the solution must provide a significant amount of flexibility to scale performance and capacity as needed.  Built on VMs, the Insight AnalytiX infrastructure can easily scale while maintaining central control to manage across the entire enterprise. Simply scale up or down the processing engines (no agents) to add more horsepower that is needed to handle the workload.

## Insight AnalytiX Delivers Privacy Risk Classifier

The Insight AnalytiX Privacy Risk Classifier was built to meet the needs of large enterprise customers, processing their massive amounts of unstructured data efficiently to help them identify and mitigate risks associated with handling their customer's PII.  Easy to deploy, manage, and maintain, the solution can be deployed with other components that are part of the Unified Unstructured Data Management Platform to add additional value for infrastructure modernization to create a holistic approach for data and storage optimization. The latest upgrade is focused on enhancing the product's Data Protection and Security Functionalities. It will help customers with flexible, scalable PII discovery, deep analytics, and reduced data vulnerability.