WHITEPAPER

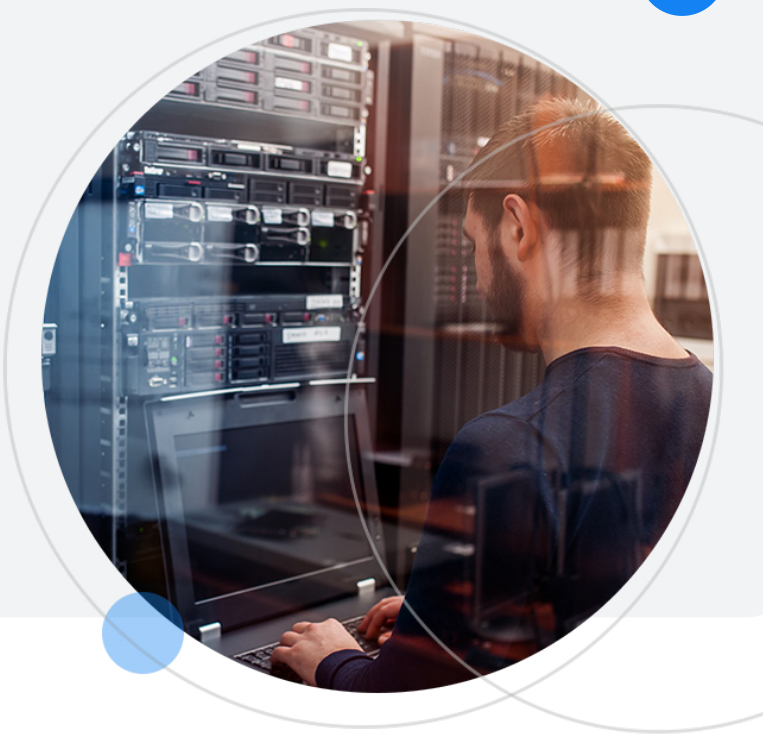
Data masking tools for data-centric enterprises: **Secure, agile, mass-scale**

Securing and protecting sensitive and personal data is imperative for today's enterprise

Data masking tools are critical for business success in our data driven economy where enterprises face increasing cyber threats and global data protection laws.

Introduction

With the proliferation of personal data – collected by enterprises, financial institutions, and medical services providers, to name just a few – the need for protecting individual privacy is paramount. One way to protect privacy is to mask the data, by consistently changing names, or including only the last 4 digits in a credit card or social security number. This whitepaper explores today's data masking techniques, the challenges they post for enterprises are faced with, and a new, real-time approach, based on business entities, that promises to address these challenges in the most comprehensive manner.



The need for data masking

Data masking has been around for decades. Today, it is needed more than ever, in order to effectively protect sensitive data, and to address:

Regulatory compliance

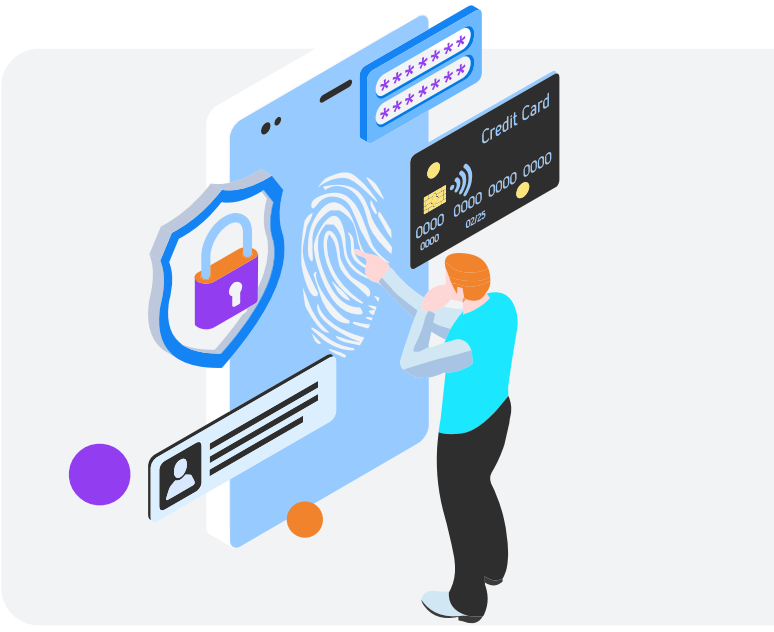
Highly regulated industries, like financial services and healthcare, already operate under strict privacy regulations, including the Payment Card Industry Data Security Standard (PCI DSS), and the Health Insurance Portability and Accountability Act (HIPAA). Since the introduction of Europe's GDPR in 2018, there has been a proliferation of privacy laws across the globe including CCPA and CCPR in California, LGPD in Brazil, and PDPA in the Philippines and Singapore. Such privacy laws seek to protect Personally Identifiable Information (PII) by, and restrict access to it whenever possible.

Insider threats

Many employees and third-party contractors access enterprise systems on a regular basis. Production systems are particularly vulnerable, because sensitive information is often used in development, testing, and other pre-production environments. With insider threats rising 47% since 2018, according to the Ponemon Institute report, containing sensitive data costs companies an average of more than $200,000 per year.

External threats

In 2020, personal data was compromised in 58% of the data breaches, states a Verizon report. The study further indicates that in 72% of the cases, the victims were large enterprises. With the vast volume, variety and velocity of enterprise data, it is no wonder that breaches proliferate. Taking measures to protect sensitive data in non-production environments will significantly reduce the risk.



Data masking techniques

Over time, a variety of data masking techniques have been created. Selecting the right approach is dependent on the intended data use. The goal is to maximize data protection, while minimizing data exposure.

Persistent, static data masking

Non-production environments, such as those used for analytics, testing, training, and development purposes, often source data from production systems. In such cases, private data is protected with static data masking, a one-way transformation ensuring that the masking process cannot be undone.

When it comes to testing and analytics, repeatability is a key concept because using the same input data delivers the same results. This requires the masked data values to persist, over time, and through multiple extractions.

Dynamic data masking

Dynamic data masking is used to protect, obscure, or block access to, sensitive data. While prevalent in production systems, it is also used when testers or data scientists require real data.

This type of masking is done on the fly, in response to a real-time event. When the data is located in multiple source systems, masking consistency is difficult, especially when dealing with disparate environments, and a wide variety of technologies.

Data masking on the fly

When analytics or test data is extracted from production systems, staging sites are often used to integrate, cleanse, and transform the data , before masking it. The masked data is then delivered to the analytics or testing environment. This multi-stage process is slow, cumbersome, and risky due to the possible exposure of private data.

Unstructured data masking

Scanned documents and image files, such as insurance claims, bank checks, and medical records, contain sensitive data stored as images. Many different formats (e.g., pdf, png, csv, email, and Office docs) are used daily by enterprises in their regular interactions with individuals. With the potential for so much sensitive data to be exposed in unstructured files, the need for unstructured data masking is obvious.

Data masking challenges

Enterprise IT landscapes typically have many production systems, that are deployed on-premise and in the cloud, across a wide variety of technologies.

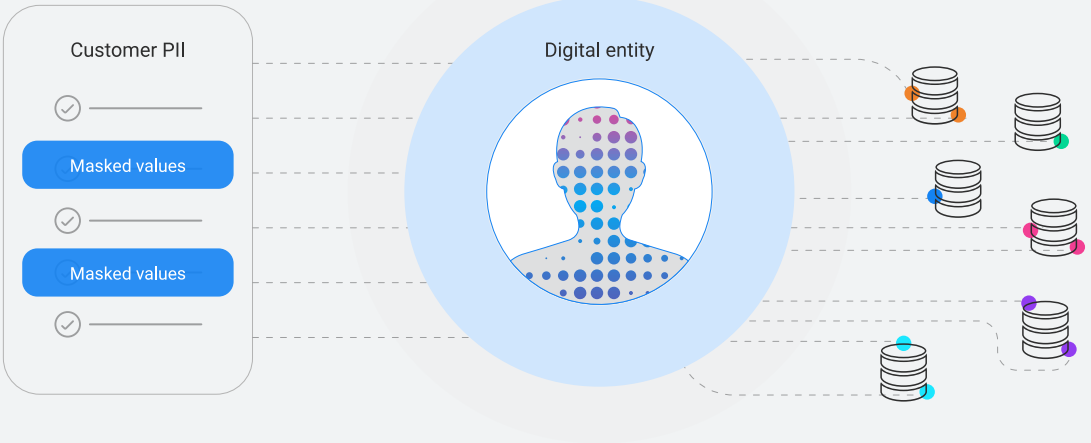
To mask data effectively, an enterprise needs to:

1. **Identify** the sensitive data and PII that require protection
2. **Resolve** identities to ensure the same values across systems (e.g., If Rick Smith is masked as Sam Jones, that identity must be consistent wherever it appears)
3. **Comply** with company governance policies for role, location, and permissions-based data access
4. **Scale** for real-time access and mass-batch data extraction
5. **Manage** growing volumes of unstructured data

Entity-based data masking

K2View Data Masking is based on the company's operational Data Fabric, which organizes fragmented data from disparate systems according to Digital Entity™ data schemas – customer, order, device, or anything else that's important to the business.

The digital entity unifies everything a company knows about the business entity – including all interactions, transactions, and master data. This individualized data organization simplifies data protection and privacy compliance. It accelerates enterprise-scale dynamic data masking for operational use cases, and persistent data masking for non-production environments.

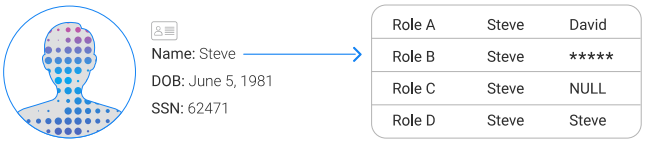


No staging needed

K2View in-flight Data Masking eliminates the need for slow, cumbersome and risk-prone staging areas, where unmasked data is exposed to potential breaches. Using our graphical data orchestration tool, data, from multiple production systems, is integrated, cleansed, and masked on the fly.

The entity-based data model simplifies the complexity, ensuring that individual customer data is consistently:

- Integrated, across multiple sources
- Persistent, over time and multiple extractions
- Referential integrity fully preserved



Dynamic masking

K2View Data Masking transforms, obscures, or blocks access to sensitive information fields based on user roles and testing environment privileges. Using data orchestration, a wide variety of in-line masking functions can be invoked to protect the data. Java functions can be used to implement additional masking functionality.

Unstructured data masking

Protect unstructured data including images, PDFs, XML, CSV, text-based files, and more, with static and dynamic masking capabilities. Replace sensitive photos with fake alternative ones; use OCR to detect content and enable intelligent masking, synthetically generate digital versions of receipts, checks, contracts and other items for testing purposes. By managing unstructured data within the digital entity data schema, referential integrity is ensured, and consistency maintained across structured and unstructured data.

Extensive and extendible masking functions

K2View Data Masking has an extensive library of masking functions designed to provide realistic but fake data. The chart below provides a number of examples including masking to valid social security numbers (SSN), selecting names from name directories, random number generation, and address-based zip codes. The library can be easily extended by with Java functions that implement additional masking functions.

Field	Masking function
SSN/National ID	Generate valid SSN
Credit card	Generate valid number based on card type
First name/Last name/Zip code	Select from collection
Date of birth	Shuffle (preserve statistical diversity)
Any String/number	Random String/number
Email	Concatenation based on new first and last
Constant	Static masking based on a pre provided
Address	Based on the provided Zip

Summary

Data masking has become a pillar technology that large organizations use to comply with privacy protection regulations. Although the practice of masking data has been around for years, the sheer volume of data – and the ever-changing regulatory environment – have presented enterprises with many serious challenges. This article outlines the current data masking techniques, which are proving to be largely insufficient. It also introduces a novel, real-time data masking methodology, based on business entities, that many of the world's largest enterprises have already implemented to great effect.

To find out more, we invite you to visit our website.