



A 4-STEP APPROACH TO

Next-Generation Data Classification

Revolutionize data discovery & data classification with a privacy-centric approach, designed for today's data environment.

Table of contents

Introduction	3
Regular Expressions & Pattern Matching	5
Contextual Classification	7
File Classifiers by Type	8
Policy-based Classification	10
Classification Anywhere	11



In order to manage and protect your data, you need to know not just where it is, but whose it is, and what it is.

Traditional approaches to classification focus on either manual tagging or resource-intensive pattern matching (which isn't always reliable). Given the rate and diversity of data growth – whether it's data in a big data repository or data moving between cloud storage and data lakes – these traditional approaches aren't scalable or sustainable, and don't provide the context necessary to address the privacy and security challenges of today's environment.

“Traditional approaches to classification aren't able to address the privacy and security challenges of today's data environment.”

That's why BigID approaches classification differently: not based on what's worked in the past, or identifying what's covered by specific legislation, but with a privacy-centric approach designed from the ground up.

That means looking at the data – all the data, wherever it lives – and finding ways to classify, tag, and connect disparate data points into meaningful relationships, identities, and profiles.

BigID embraces a discovery in-depth approach that goes deep and wide: finding data wherever it is, and layering in context and correlation for classification. This approach builds on (and extends) more traditional classification methods, and expands coverage to various types of sensitive information – from personally identifiable information to profile information to broader sensitive information.

So how does BigID do it?

We've got a few ways – all purpose-built for today's data environment (and volume and variety).



BigID approaches classification differently: not based on what's worked in the past, or identifying what's covered by specific legislation, but with a privacy-centric approach designed from the ground up.



Regular Expression and Pattern Matching

The most traditional of classification methods for data, this technique matches known regular expressions (regex) and patterns to information that lives inside your data.

MasterCard credit card numbers, for instance, are a sixteen-digit number that starts with 5262. Therefore, it's entirely plausible that any sixteen-digit integer string that begins with 5262 can be labeled as a MasterCard credit card number.

“

If you already know the structure of the information you're trying to match, you'll be able to identify similar patterns within a set of data.

”

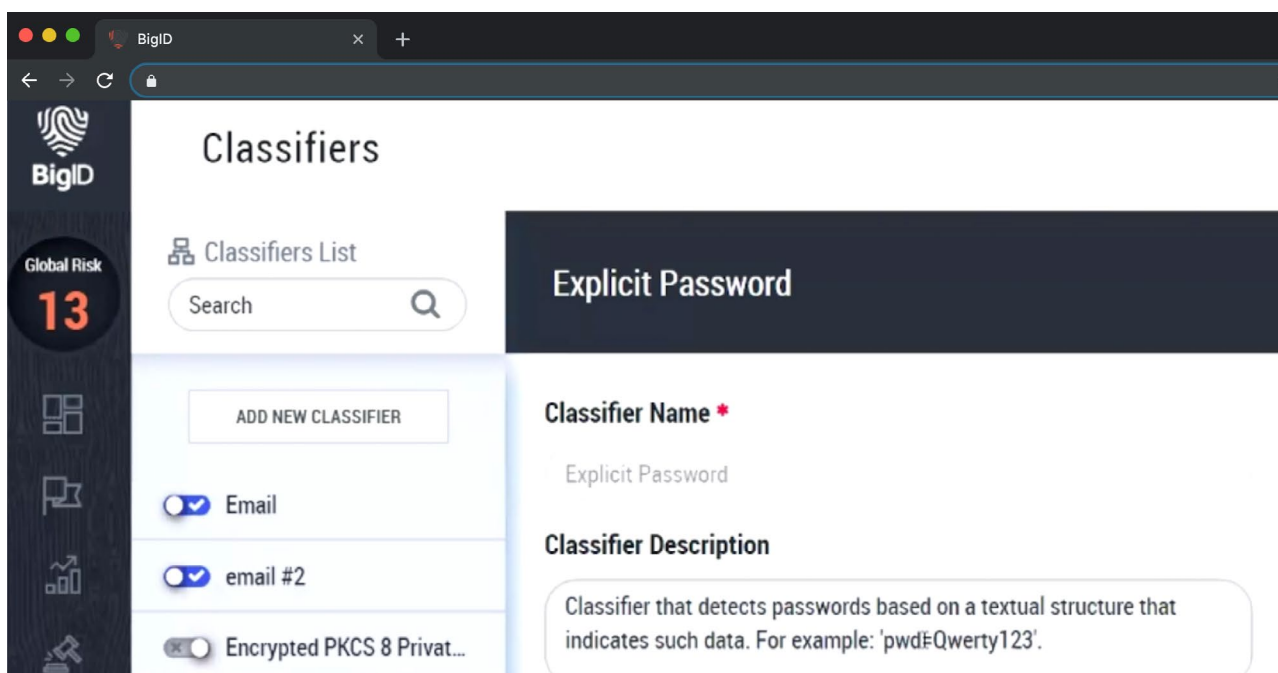
Similarly, pattern-based identifiers like zip codes, IBAN numbers, Social Security numbers, and more can all fall into this category: if you already know the structure of the information you're trying to match, you'll be able to identify similar patterns within a set of data.

Traditional pattern matching is often set by regulation: if something like PCI-DSS determines that organizations have to be able to identify credit card numbers, patterns for credit card numbers can be quickly analyzed and added to a set of dictionaries.

Classification by pattern matching is by no means obsolete – but it's important to address more than the bare minimum requirements.

We've added security identifiers, for instance, so that organizations are able to identify security-focused data points like API keys, credentials, tokens, and even common passwords.

So, for some types of data, pattern matching does the job.



Contextual Classification

A much trickier set of data to classify is the type that doesn't necessarily follow any given or consistent pattern: it's difficult to identify "friendly names" – much less see them in the context of a specific identity. Context is also critical to distinguish between two data values that have similar formats, but are two different types of information (a Social Security number and an account number, for example).

Can your traditional classification tools correlate a specific Social Security number with a first name, eye color, geolocation, and education information – all related to a single person or identity?

Nope. But BigID can.

“

Context is critical to distinguish between two data values that have similar formats, but are two different types of information.

”

BigID leverages Machine Learning (ML) and Named Entity Recognition (NER) to not only automatically identify sensitive information like voting records, social media activity, or height based on inference or other techniques – but also to link that specific instance of sensitive information into an individual identity or profile.

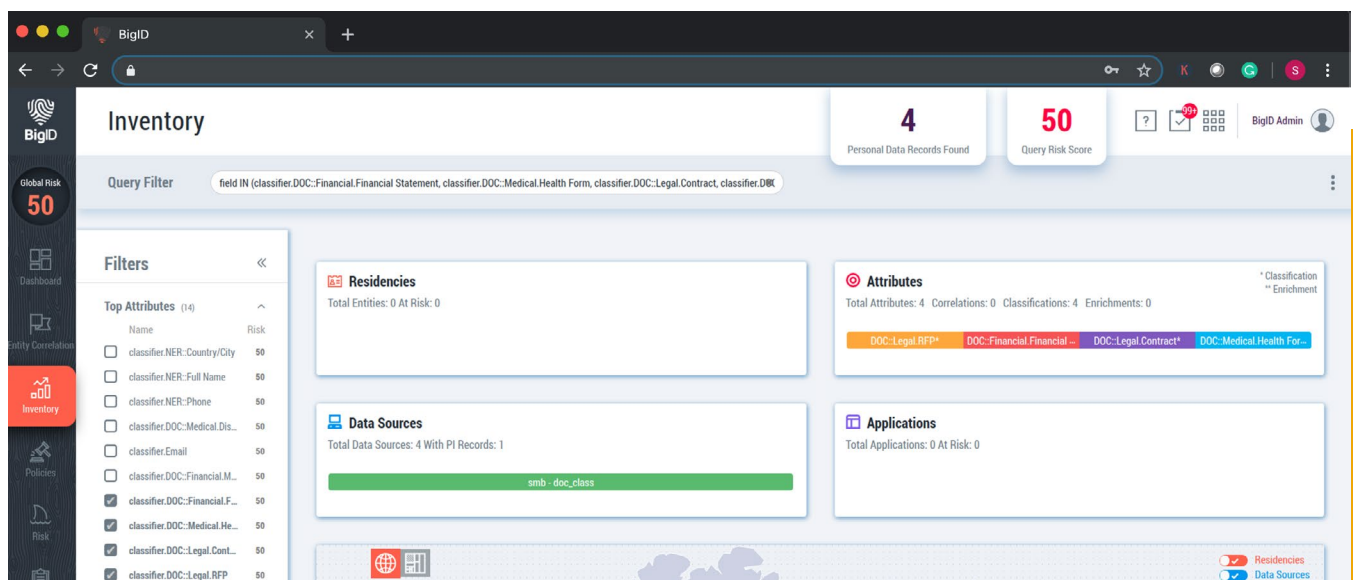
Data privacy and protection legislation is expanding the definition of personal information (and subsequently what type of information needs to be protected), and so should your classification and discovery solutions.

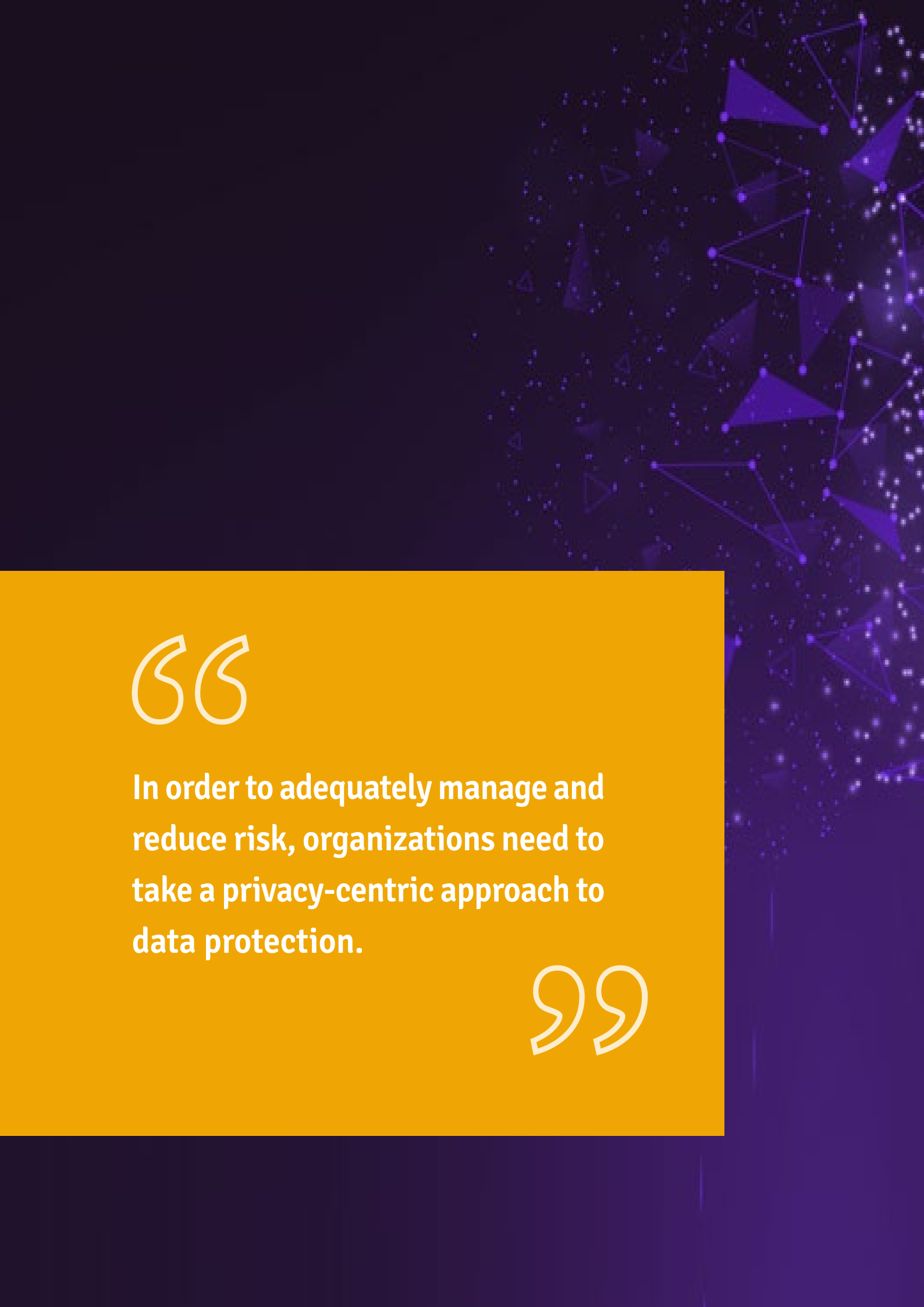
File Classifiers by Type

As data volume continues to grow, it's important to protect the right type of information with the right policies: legal documents should follow one policy, financial another, and so on and so forth.

That's why we added file classification by type to our arsenal: BigID has machine-learning models that automatically classify documents based on the content and structure of a file – without being limited to any specific data classifier. These models can recognize sensitive file types: from financial statements to boarding passes to discharge summaries to merger & acquisition documentation and more.

In order for organizations to ensure they have the right data protection in place, they need to be able to easily and accurately identify it first.





“

In order to adequately manage and reduce risk, organizations need to take a privacy-centric approach to data protection.

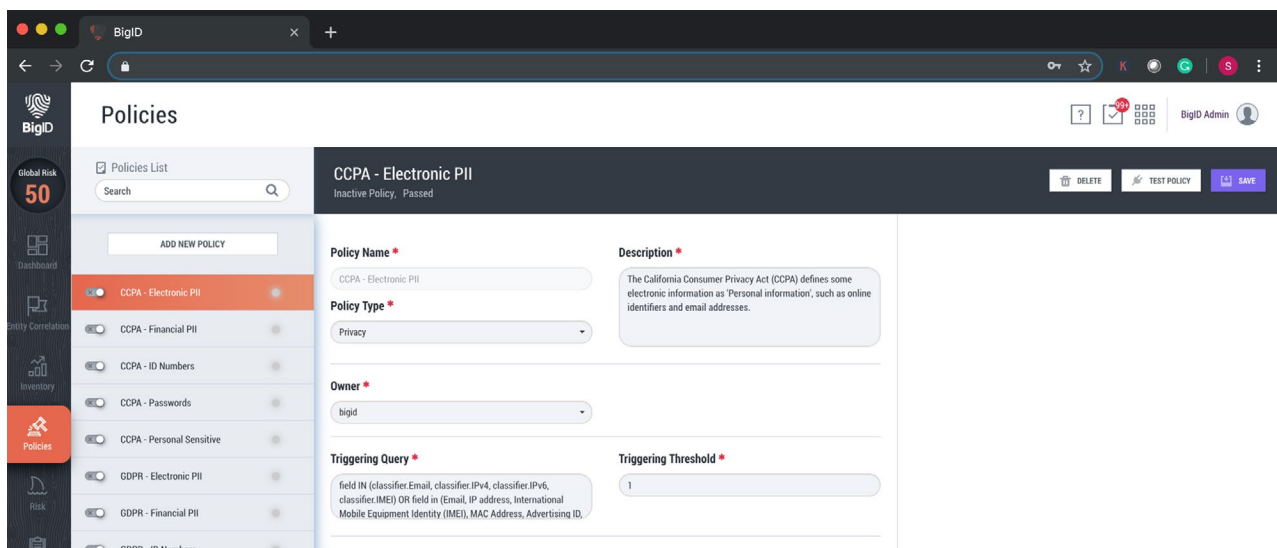
”

Policy-Based Classification

Among all the types of classification and data discovery, the single biggest driver for this is data privacy and protection regulations. From GDPR to CCPA to NYDFS to HIPAA to SOX to GLBA to (the list goes on), organizations need to be able to identify certain types of data that fall under specific regulations, and enact policies to manage and protect that data.

BigID has built-in policy libraries to help classify, manage, and protect specific types of data by policy: this ranges from ID numbers and passwords that fall under the CCPA, national identity schemes for GDPR, to credit card information that falls under PCI.

Classifying and managing data by policy enables organizations to build workflows across that specific type of data, manage access, monitor use, and protect sensitive data that may be under attack.



Classification Anywhere

Data growth, data value, and data meaning is rapidly evolving – and the policies and regulations currently in place are starting to catch up. As the world of data evolves, so does the value of personal data, sensitive data, and the very policies that aim to protect this data.

“

That's why BigID is re-thinking classification: revolutionizing data classification and discovery with an extensible, data-centric approach.

”

That's why BigID is re-thinking classification: revolutionizing data classification and discovery with an extensible, data-centric approach.

Data privacy and protection regulations like the New York SHIELD Act not only extend the definition of “personal information”, but are adding layers on top of traditional classification into their recommendations: organizations need to be able to correlate data – like username & email in combination with a password or security question – in order to apply the recommended security protections. The SHIELD Act, as a harbinger of the next wave of data breach notification laws, both expands what type of data is covered and the definition of what constitutes a data breach.

Successful data protection and privacy automation is dependent on being able to accurately discover, classify, correlate, and catalog all sensitive information, regardless of where it is. Pattern-matching alone is no longer enough: organizations need to be able to correlate data to an identity, establish relationships between single instances of sensitive data, automatically identify both direct and inferred sensitive or personal information, and establish processes and policies to protect and manage that data.

“

By taking a data-driven, innovative approach to classification, BigID intelligently (and automatically) classifies sensitive data & files of any type, wherever they're stored – across your entire organization.

”

BigID's discovery, classification, and correlation extend to unstructured, structured, and semi-structured data at petabyte scale, and applies to everything from Cassandra to Amazon S3 to CIFS to Gmail to Couchbase to Box to Hadoop and everywhere in-between: giving you a unified inventory of your sensitive data – all in one place.

By taking a data-driven, innovative approach to classification, BigID intelligently (and automatically) classifies sensitive data & files of any type, wherever they're stored – across your entire organization.

Want to see it in action? Get a demo to see how BigID does classification differently.