USING AI AND MACHINE LEARNING TO POWER DATA FINGERPRINTING

Waterline **Data**

Introduction

In the era of Big Data, a data catalog is essential for organizations to give users access to the data they need. But it can be difficult to deploy this important tool if populating the catalog entails a significant amount of manual work. The key to a successful data catalog implementation is the rapid population of the data catalog so it can be quickly put into use. This can be achieved with an automated data fingerprinting process that uses a combination of machine learning augmented with human generated curation plus ratings and reviews.

data catalogs are ciritical to the success of big data deployments.

The #1 Challenge in Successfully Deploying a Data Catalog

One of the biggest challenges in deploying a data catalog is getting it populated with useful information. While many organizations have a business glossary with defined terms and definitions, connecting this kind of business metadata down to the technical metadata, which contains statistical demographics and the actual location where that data lives, is a challenge. In the data cataloging space, people will refer to the ability to "tag" data. This is the act of connecting a physical instance of a column of data with the associated business glossary term. In most organizations, one business term will have tens if not hundreds of physical instances of that term deployed across an organization.

For example, "First Name" is a business term that is located in all kinds of systems. The problem is that locating all instances of "first name" is a very tedious task. This is made more difficult because very often data is not always nicely labeled. So a column could have a name of "First Name", "fname", "fn", "given name" or even "C01", and all of those columns could contain first names.

This translates into a big problem for implementing a data catalog. How do you connect all of the business terms to the actual implementation of data

associated with that term? While some catalogs try to do this by crowd sourcing, practical experience shows that this doesn't work because this approach doesn't scale to deal with the growing amount of data that is coming into organizations.

Critically, crowd sourcing doesn't deal with so-called "dark data" or data with which no one is familiar. With the incredible velocity with which new datasets are being created, it's not possible to track all the instances in an organization. This can also occur because either the data is new to an organization—for instance if you purchase data from an outside supplier—or because no one has touched it in a long time.

This is the reason why automation is vital to deal with the problem of data tagging.

Data Tagging & Fingerprinting Using AI and Machine Learning

Waterline Data addresses the challenge above by using artificial intelligence and machine learning to analyze data and do what we call "data fingerprinting." Fingerprinting works on the concept that a column of data has a signature, or a fingerprint, and that by examining the data values in a column of data, we can identify what that data is and determine two things: which other columns share this same fingerprint, and what is the business term or label that can be connected to this data.

On this second point—connecting a business term to an unlabeled or mislabeled column of data—Waterline Data fingerprinting can do this for lots of business terms, but not for everything out of the box. For some terms, it has to be trained. For example, it knows what a first name or last name is, or what a credit card number is. But it doesn't know what "Claim Number" is for ACME Insurance because the format of a claim number would be unique to ACME. However, once a knowledgeable business user or data steward tags just one column as "Claim Number," the system now knows what a claim number is. The tag for this business term gets propagated automatically to all

Manually populating a data catalog using crowdsourcing doesn't scale with the growing volume of data

of the other unlabeled columns of data that have the same fingerprint.

The reason it is powerful is because you only have to tag a unique attribute once, and the computer learns and propagates the tags automatically. Curation can even be carried over to a brand new data source. Suppose a new s3 bucket with terabytes of new data was just brought online. How you manually sift through it? With Waterline Data, existing fingerprints can be used to automatically match against the new body of data.

a bottle neck in the organization, but becomes an agile group responsible for quickly getting the right data into the hands of the right users in a governed, compliant manner.

How Does Data Fingerprinting Work?

So how does the fingerprint itself get formed? There are two main techniques that are used connect a business term to a column of data. The first is using regular expressions. This works well for something like a credit card number, where we expect a certain number of digits and those digits have a consistent set of repeatable information.

The second is what we will call "value based" fingerprinting. This is what would be used in the "Claim Number" example above. By taking basic profiling information about the column, the data type, the value distributions, etc. in a specific combination, we have established a mathematical formula

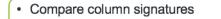| Data Profiling (basic statistics)	Fingerprinting
• Field length • Nulls • Min/max • Cardinality • Data type(s) • Value distribution	• Create a column signature based on combination of statistical factors (details are Waterline trade secret)

for converting that profiling information into a unique identifier, earlier referred to as a "column signature." We can then compare identifiers and find other columns that match. The trick here is getting the algorithm for creating the signature just right to avoid both over and under matching.

This is the approach that is used to find a column that contains "cities", for example. We simply take a list of cities, use that to create a signature, and then whenever the system sees a column that shares a similar signature, it gives it the same label.

Now what happens if for your data set, you have a more granular understanding of cities? Perhaps you have different columns for "French Cities", "German Cities", "California Cities". The first-time Waterline Data Fingerprinting sees these columns of data, it would think they were just "Cities", because that is all it knows. However, a user can then tell it that the column it just labeled isn't just "Cities", but it is more specifically "California Cities" and the machine learning algorithm updates itself and from then on, the system will properly label other columns that have only Californian cities based on the updated and more precise signature.

Match Fingerprints
- Compare column signatures
- Tag a column of data to a business term
- Propagate tags to all similar signatures
- Manage over and under-matching (Waterline trade secret)

Human curation + Machine Learning
- Users accept/reject tags
- Teaches the system to reduce false positives

Example
Waterline tags a column as "Cities". User corrects it as U.S. cities. System learns

Examples
How do you know:
- a 16 digit code is a credit card?
- first name is not just a word?
- a column with 0's and 1's that is marked gender is different or the same from another column with 0's and 1's

Value based fingerprinting is very powerful because it isn't fooled by poorly labeled data. It doesn't care if a column is labeled "C01", "CLM" or "Claim Number". Note that it will take advantage of good labels to validate its initial

conclusion. So, a column with the exact same data that has a column labeled "Claim Number" will get a higher confidence score than the same data that is labeled "C01". The algorithm will identify both columns and label them "Claim Number", but it might rate the nicely labeled column with a confidence of 96% and the one with the generic label with a confidence of 92%. But if you reverse the column labels for first and last name, it will ignore the column labels in that case.

The idea of examining the actual data values, whether via regular expression or value based tagging, is critical because data is often mislabeled. And while only a small percentage of your data might be mislabeled, the only way to figure that out is by looking at the data values themselves and manual inspection doesn't work because you can't manually inspect one million rows of data. Another important aspect of examining the data values is that it also allows you to automate the identification of sensitive data that needs to be either secured, or even just cataloged to comply with government regulations like the EU GDPR.

Conclusion

Waterline makes data fingerprinting is simple. Data fingerprinting automates the tagging of structured and semi-structured data with associated business terms from a business glossary. The devil however, as usual, is in the details that make the system scale, make it robust, and keep it from both over- and under-matching.