Doing the Data Preparation Heavy Lifting

Every resource that humans seek to use for an advanced purpose (oil, metal, etc.) must be processed in some way before it can be transformed into a new, more valuable form. The newest precious resource, data, is no different. Entire IT departments and teams of skilled data engineers are charged with this important duty every day, spending hours of their time (another precious resource) assembling data pipelines to bring data into BI and data science platforms where it can be used to make decisions. That involves cleaning the data and getting it ready for analysts, business users, developers, data scientists, and even customers accessing their own data through internal embedded analytics.

Every advanced manufacturing process develops specialized tools, and in the analytics realm, the most advanced tools available are machine learning (ML) and artificial intelligence. These powerful systems can be used to help data engineers and anyone else involved in data preparation simplify their lives by automating large-scale, repetitive tasks. One of the biggest opportunities for AI, ML, and augmented analytics to help is in the realm of data deduplication.

As the name implies, data deduplication is the act of removing redundant data from a dataset, which can help build a more coherent dataset or reduce dimensionality. A data deduplication tool has a lot of moving parts and can bring a complex array of computing power and assistance to a basic-sounding task. In this paper, we'll delve into why data deduplication is important, some ways it can work, and a sampling of some other ways AI can assist with data prep.

Why Is Deduplication Important?

According to Forbes, 80% of data scientists' work is data preparation, and 76% of those polled reported that it was their least favorite task! These reasons would be sufficient to do whatever it takes to automate these processes. The data experts (engineers, scientists, IT teams, etc.) who are in charge of data preparation are most interested in connecting to data sources, building stable pipelines, and other more complex tasks. Forcing them to deal with rote, annoying tasks like data preparation reduces morale and takes data experts away from the higher-value work that they could be doing to benefit the company.

Proper deduplication can have a huge impact on a company's bottom line. While the rise of countless cloud sources has coincided with a lowering of costs per unit of stored data, there are still costs associated with maintaining data volumes, and having duplicate data drives those costs up. The extra data can also slow down query response times, delaying the time it takes to make decisions. Additionally, duplicate data can return false results that lead to incorrect business decisions. In today's fast-paced modern business environment, delays and errors like these are costly. The proliferation of data storage types and locations introduces a new range of accompanying errors.

To mitigate that risk, deduplication is a vital step in the data-cleansing process. Removing duplicate information, either inside an individual database or as part of a data model (with a deduplication scrubber or other tool) is key to ensuring accurate, timely results from analyses. Next, let's go into performing deduplication with a machine learning algorithm.

Deduplication via ML Algorithm

One of the most common ways to deploy a deduplication tool in a BI tool is by aggregating data along a certain dimension in order to perform analytics. In many cases, data is coming from various platforms (different cloud databases, software systems like Salesforce, etc.) and the same information can be represented differently, which produces misleading data aggregations that could lead to false conclusions. Deduplication detects different instances of the same information and automatically applies logic to ensure uniform entry as part of the data transformation.

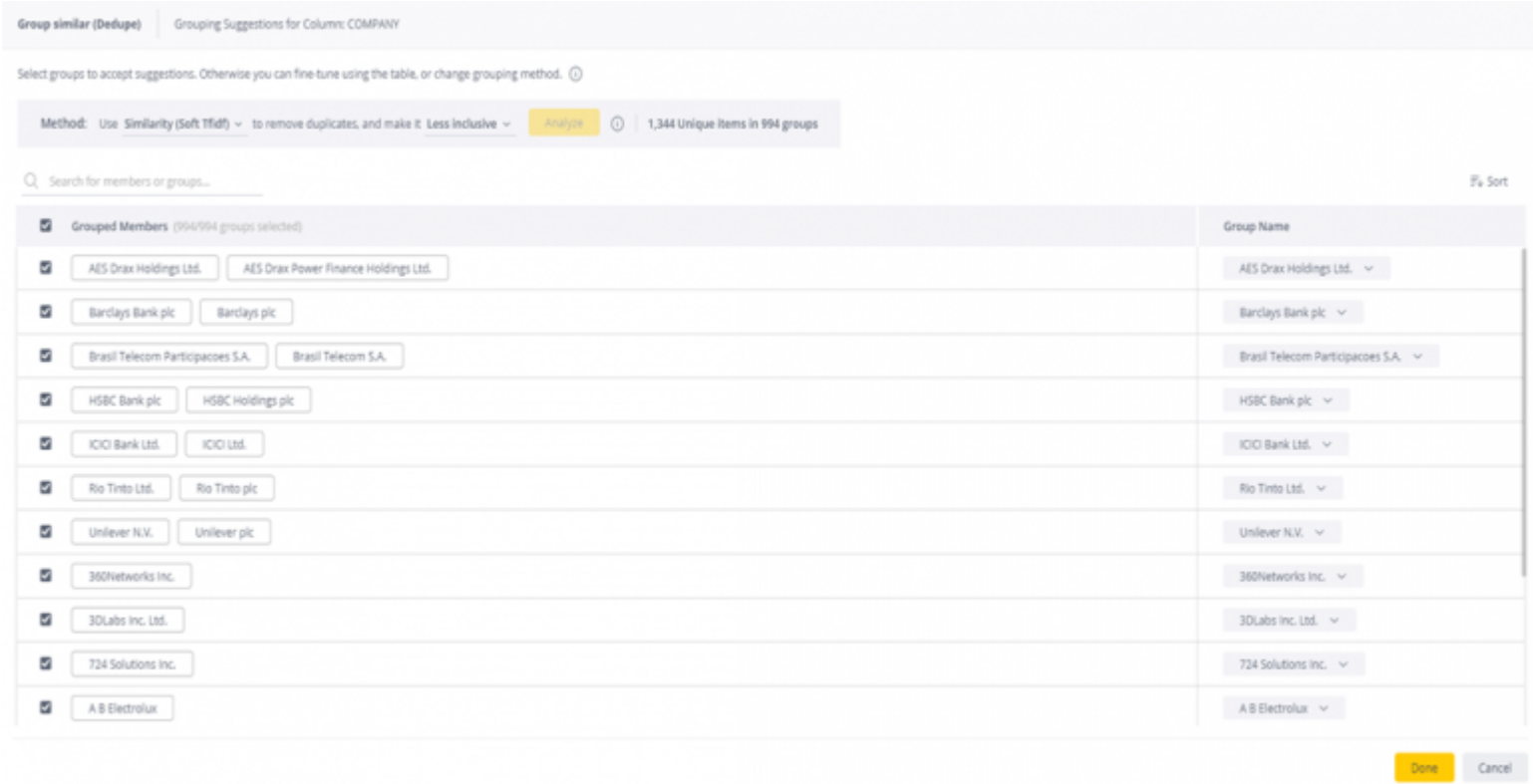
String correction is part of many data onboarding processes. BI support personnel spend large amounts of time manually grouping and replacing redundant data, sometimes using lookup with predefined dictionaries accumulated from their personal experience. Another option is using custom SQL CASE statements, like so:

```
case [Name]
when 'John Smith' then 'J. Smith'
case [Country]
when 'Afrika' then 'Africa'
```

The manual process of correcting these duplication opportunities can be slow, frustrating, and error-prone! However, a superior option for accelerating and automating this process is to use ML. The right algorithm can handle these tasks quickly and help construct a clean dataset, ready for dashboard production. Let's look at how that works:

When executing deduplication with an algorithm, the system takes a given data column and divides it to a group of information (clusters) that relate to a single attribute (an attribute that has the same meaning — this can be phonetic properties, distance, etc.). Then the algorithm allows the user to define the desired attribute per group (frequency of appearance in the cluster, etc.). The first time the system analyzes a column, the human user chooses how it behaves and how inclusive the results should be. Once the algorithm has done its work, the final information can be used in the dashboard for analytics as a regular field. Easy!

BI systems with this kind of functionality are able to analyze and group data from a data model at the column level. This means that the model looks at each column, identifies similar strings, and groups them together into a single value, like so:



Three methods for analyzing data (similarity functions) include:

Similarity: The system applies a soft TFIDF (term frequency-inverse document frequency) algorithm where similar strings are grouped together by rank. Terms that appear too often, such as “company” and “inc” are scored lower since they are less unique strings.

Spelling: Grouping strings similar in spelling.

Phonetic Similarity: Groups strings terms that sound similar, for example, strings that use ‘f’ or “ph.”

When using a BI tool with AI deduplication features, also test each method to find the results that work best for you, adjusting the level of inclusiveness or granularity until you create a dataset that fits your needs. Most systems are set at a default level that is less inclusive, meaning results are more precise. The more inclusive the results, the less precision there is. If your results are too inclusive, you may need to spend time editing the results that don't match your needs. No deduplication tool is perfect, especially not the first time you're using it.

AI and Data: Possibilities

Machine learning and AI have broader capabilities to clean data beyond just acting as a fancy deduplication scrubber. Again: any rote human task can be performed by a trained algorithm in an AI system. They can clean and standardize data as it's being moved, extracted, and loaded, even spot outliers in the dataset and alert human users.

First off: standardizing data. It's easier to analyze and work with data that's all in a similar format. For instance, an AI system can be trained to sift through a dataset and convert every entry to lowercase (often the preferred format for writing queries). Eliminating errant spaces before or after entries is another useful standardization practice, since BI tools can be very literal when responding to query requests. For example, "users" is not the same as " users" to a computer, so analysis of those two entires will return incorrect results. Rather than spending hours of human time to comb through data, AI can handle these repetitive tasks with ease and deliver a cleaner dataset.

Today's organizations deal with data in a wide array of locations and cloud databases, so loading it all into a central BI tool is a must. This is another place that AI can help, scanning and suggesting cleansing actions while the data is being moved. In addition to acting as a deduplication tool, AI can also fill in blank values and ensure that the end product is ready to be analyzed. Depending on the training dataset, the AI can also give human users additional options to improve the quality of the new dataset.

Outlier detection is a classic AI task. Since algorithms are good at pattern recognition, they're also good at spotting values that fall outside the expected results. As a system scans the dataset, it's essentially asking "Does this column look okay?" and checking for things like accepted values, null values, numerical outliers, uniqueness (e.g. every user should have their own user_id, etc.), and referential integrity (for values that are keys to different tables). When an AI comes up against a value that doesn't make sense with what it's been trained to expect, it can surface that rogue result to human users, who can take action.

These are just a few ways that AI systems could one day help data engineers and all users get more out of their data and spend less time on the manual (but necessary) processes of data cleansing.

AI and Engineers: Partners in Building the Future

The future is being built on data *right now*! That future will bring together countless datasets from the cloud, apps, and anywhere else that builders collect and store data. It'll be up to data engineers to build enduring, stable pipelines that deliver that precious data to BI tools and allow analysts, users, and customers to get the most out of it.

Those intrepid engineers won't be alone. The future will be built on data and powered by augmented analytics. Artificial Intelligence systems integrated throughout the powerful analytics platforms of the future will clean and prep data as it's ingested into the platform. These systems will get rid of troublesome duplicate data, fill null values, identify outliers, suggest joins, and even guide human users to find new insights that they might never have identified on their own.

Sisense is leading the way into this new era with powerful machine learning and AI components that make life easier for data engineers and everyone else involved in the BI process. See for yourself what Sisense could do for your organization and build the future of your company on a platform that’s powerful enough to handle anything your imagination can throw at it.

Dig Into Sisense for Data Engineers