ataccama ONE

How AI is Transforming Machine Learning

Whitepaper

At Ataccama, we believe that **machine learning can transform data management systems**. Over the past few years, machine learning has proven useful in the field of data analytics. With this in mind, Ataccama has been investing resources into leveraging the same cutting-edge machine learning techniques in the context of data management. By developing AI-powered data cataloging and master data management solutions, modern business challenges can be remedied in a more effective and efficient way.

Discover how our Ataccama ONE platform **automates manual** and **time-consuming tasks and optimizes performance**.

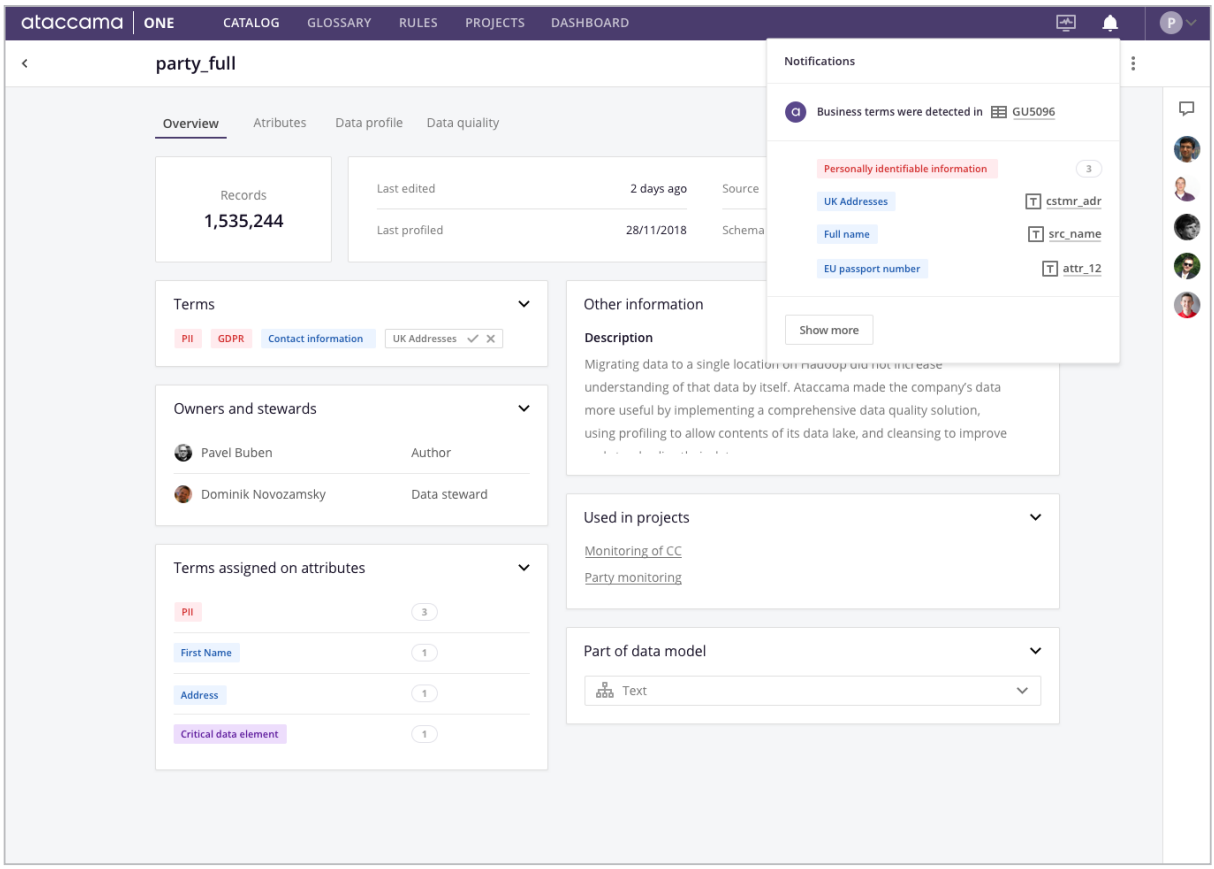
THE IMPORTANCE OF AN AI-POWERED DATA CATALOG

With the increasing amounts of data in an organization's possession, data catalogs are more vital than ever. A data catalog is a place where stakeholders store and share metadata about all data assets from multiple data sources. The metadata contains information about where the data is stored, who the owner is, what its classification is, as well as some statistical characteristics. The data catalog is the right place for any analyst, developer, or data scientist looking for available data assets to be used in their project. A successful data catalog is built around two main areas: **automation** and **collaboration**. Automation ensures that technical metadata is enriched with business metadata so users have more information about an asset. Since not everything can be enriched automatically, it is also crucial to provide a user-friendly interface so users can enrich some metadata manually, discuss it, and share the knowledge.

We leverage machine learning in our Ataccama ONE Data Catalog in the following ways:

BUSINESS TERM SUGGESTION

A user adds a new item to the data catalog (e.g. a table from a relational database) to enrich the metadata by assigning business terms from a company-wide business glossary. The definition of a product code might differ from enterprise to enterprise as organizations may follow different patterns. Detecting business terms can be automated easily using regular expressions or by comparing data with some reference data sets. This approach is not very flexible, though. It does not reflect user interactions and matches



With machine learning, the data catalog actively learns from user input and suggests terms.

only the same data (described by the same metadata) with the same regular expression all the time. **Machine learning-based business term tagging** is able to **actively learn** from user input and **suggest terms based on previous user actions** without the need of explicitly predefined business term assignment logic. The system recognizes the similarity between items in the data catalog and infers suggestions on business term assignments based on the level of that similarity.

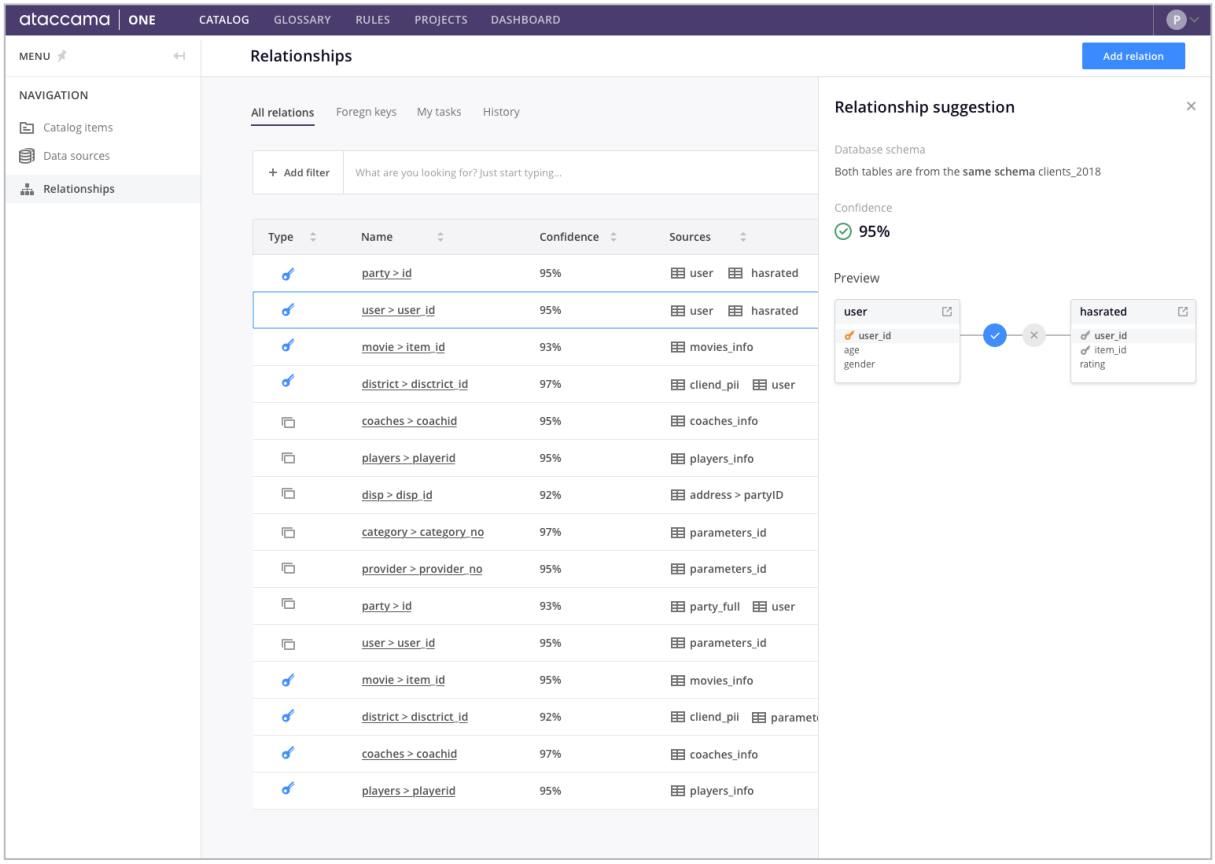
Relationship Discovery

The ability of **machine learning-based recognition** of similar items in a data catalog enables much more than that. Generally, an enterprise has a lot of related data in different systems. However, there is a lack of options to formally capture these relationships across multiple sources. Often, due to a number of independently chained and undocumented ETL jobs, it is hard to draw a clear picture of which data is related. Imagine a group of analysts starting a project and being given a specific data set. Due to a shortage of information about the data set (probably caused by employee churn), they are not able to understand certain codes in the data (e.g. their data set contains a code value that might refer to its description, and unfortunately they don't know where to find it). With **relationship discovery**, the data catalog is able to identify whether the code value refers to another data set within the organization. This will bring the team of analysts more relevant information on data understanding and help them to deliver more accurate reports.

Based on **similarity detection**, the data catalog is also able to identify other types of relationships besides foreign keys. Relations between tables in different environments and duplicated data caused by multiple overlapping extracts can be identified by a machine learning-enabled data catalog, and can shed light on systems' overarching data lineage and structure.

OPTIMIZING MDM WITH MACHINE LEARNING

Another great opportunity for the application of machine learning lies in any standard master data management use case. A central



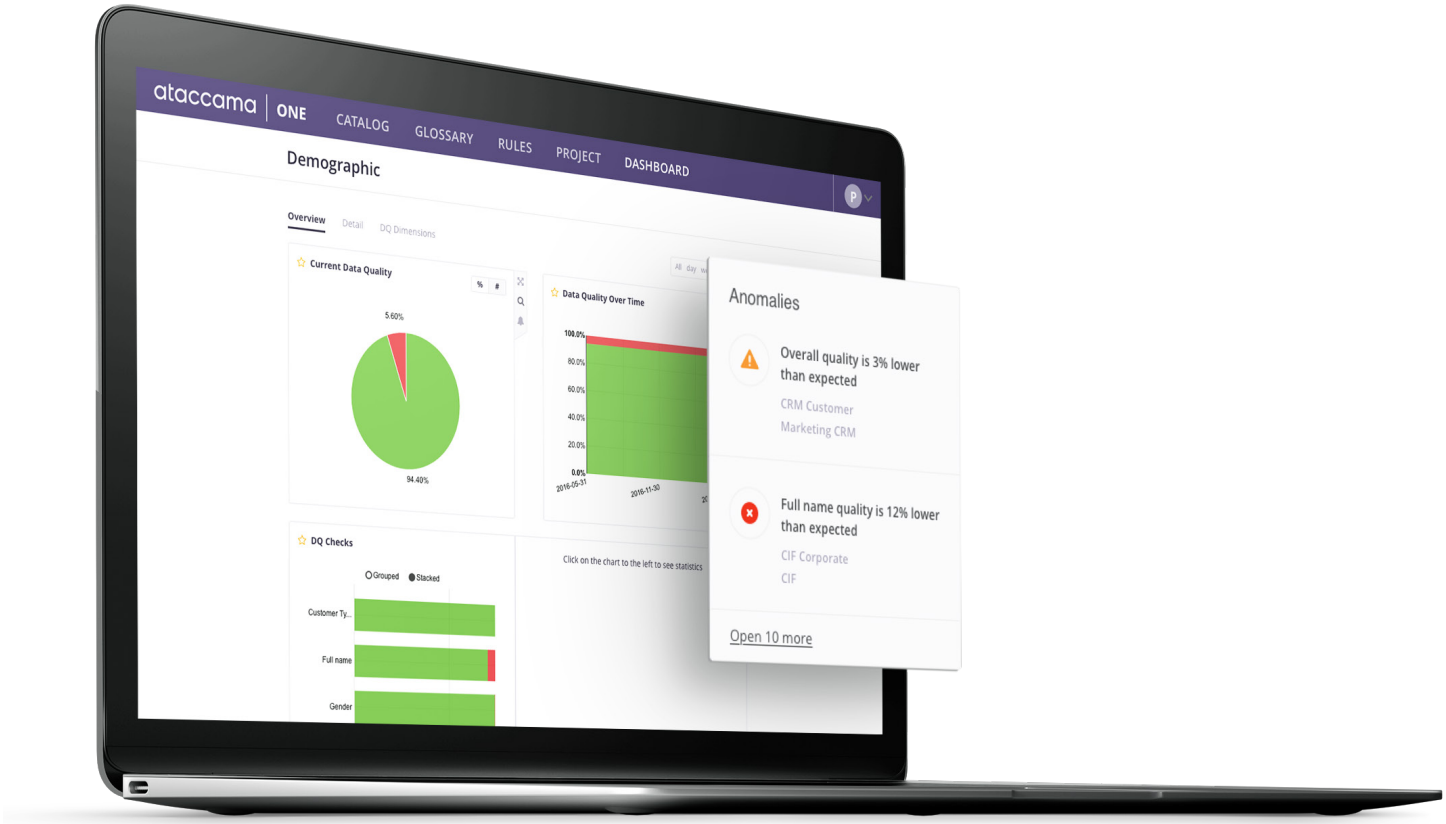
With relationship discovery, the data catalog is able to identify whether an attribute refers to another data set in the organization.

function of any good MDM solution is the process of mastering. Mastering involves the **acquisition** of data from multiple data sources, **matching** records representing the individual instances of the same entity (e.g. same person, same company, same product), i.e., **bucketing**, and creating the best possible record (i.e. **golden record**) out of each bucket by merging them together. Usually, the rules for bucketing and for the creation of a golden record are manual. It is often the case that the rules need to be deterministic and completely transparent (i.e. **a white box**) due to **regulatory requirements**, especially when dealing with personal data. In addition to matching, users have to be able to manually split incorrectly merged records. Due to all these demands, the rules are typically configured upfront and have to be designed thoroughly.

Rule Performance Monitoring

In real world scenarios, the management of extensive sets of data is an iterative process. First, the data from multiple sources is cleansed and unified. Initial rules are prepared, and the system runs on an initial data set. The results are then explored, and rules are tweaked before the process repeats. The output may result in many false positives or negatives. For example, the engine may accidentally match records that shouldn't be matched and it may miss matching records that should be matched. Matching may be done automatically, if the rule is robust and reliable enough; or it may be assisted by producing **matching proposals** to be resolved manually. **Automation** is far less expensive for an organization. Matching rules are typically conservative as manual splits tend to be more expensive from a business perspective than manual merges. Moreover, it is possible to configure explicit overrides which prevent certain records from being merged.

Data stewards resolve matching proposals, handle data quality issues, and label the reasoning for their corrections manually. All these interactions give the opportunity to learn from user feedback. The initial rules may not be effective after some time in production as data has dynamic characteristics. A matching process could start suddenly receiving filled values of an otherwise historically empty attribute, and it may be very helpful for more precise matching. It would be great if the system



Spot anomalies in data quality trends with anomaly detection in the data catalog.

could give suggestions for an update of the rules, whether it's the addition of a new rule, the deletion of an old one, or its alteration. The rules might be redundant, or an order of rule execution might affect performance. The system should then be able to help optimize the performance of the whole process.

MACHINE LEARNING & ANOMALY DETECTION

Most data governance architecture contains substantial data processing jobs scheduled and triggered by various conditions. These jobs involve **moving and transforming data**, often in chains of multiple jobs dependent on each other. It is essential to monitor the status of the jobs continuously. **Semantic correctness** is an example of why this is important. Certain jobs may succeed even if they encounter an anomaly. Picture a job processing a whole-day batch of transactional data from a branch of an enterprise company. For this particular branch, it is common for there to be far fewer transactions on Wednesday than on Friday, and for there to be none during weekends. One week, Friday's job yields less data than expected and Saturday's job yields some extra data. The job may not fail, but its behavior is unexpected. Machine learning does a good job in the **prediction of reasonable time series** (i.e. time series possessing a trend and seasonality). It is often the case that business processes tend to behave reasonably enough to employ machine learning models for **anomaly detection**.

In Ataccama ONE, we use AI to **detect irregularities** in data loads, including changes to data volumes, outliers, changes in data characteristics, and more. The solution improves over time as it **learns from the user's approach** to reported anomalies.

Data Quality Monitoring

Another example of an **anomaly detection** use case is data quality monitoring, where a set of hand-crafted data quality rules are evaluated according to a scheduled time plan. These rules generate success rates throughout time, which gives us a time series for each of the rules. Often systems allow us to define a static percentage threshold per rule to notify the data stewards involved in the process if the result drops below this threshold. However, the behavior of the results might not be that simple, and it would help to have a system that learns what should or shouldn't be considered an anomaly from

user interaction. Needless to say, it is much more difficult to understand and predict discrepancies in data by observing other aspects besides quantitative variations in data feeds. In such a case, it is necessary to learn, identify, and explain the common root causes of sudden data quality changes in the broader context of multiple data sources as there might be more complex relations among them. To achieve this, the solution has to be able to **learn from user feedback** while generating hypotheses for users to verify and confirm or reject.

In our Ataccama ONE Data Quality Management module, we use **AI and machine learning** to keep information about an organization's current data quality state readily available for faster, improved decision making as well as to deliver **smart, automated monitoring** and project **configuration suggestions**.

Data Increment Monitoring

Similarly, anomaly detection can be applied to master data management use cases. Aggregate numerical characteristics (e.g. amounts of records, attribute value statistics) in delta batches of records being processed on a regular basis are constantly monitored for anomaly detection. If some of these characteristics behave predictably and then start acting unexpectedly at some point in time, the responsible data steward will be notified to resolve the issue while interactively **teaching the system** from that case to **gain better accuracy**.

WHAT SELF-DRIVING DATA MANAGEMENT CAN DO FOR YOU

Finding **valuable insight** in your data is not simply a matter of having vast quantities of it available, but of being able to **find the right data when you need it**. This is why an AI-powered data catalog is essential in organizations today. Not only is machine learning vital in the order and storage of your data, but it's also becoming an important component in the mastering of records.

GET STARTED

ASK FOR A DEMO

See what we can do for your data project.
Get in touch to see a full demo of Ataccama ONE.

→ INFO@ASCENTION.COM

LOOK FOR US IN THE MAGIC QUADRANTS

Gartner Magic Quadrant for Data Quality Tools
Gartner Magic Quadrant for MDM Solutions

This whitepaper complements an Ataccama webinar on the topic which was sponsored by Ascention.

ASCENTION
THE NEXT LEVEL

The use of machine learning in data management helps **stem the tide of data chaos** and ensures that your data is orderly, correct, and easy to find. Interested in learning more about what our Self-Driving Data Management & Governance platform can do for your business? Get in touch with us for a demo.

ATACCAMA ONE: SELF-DRIVING DATA MANAGEMENT & GOVERNANCE


Data Discovery & Profiling


Metadata Management & Data Catalog


Data Quality Management


Master & Reference Data Management


Big Data Processing & Data Integration

Discover more platform modules at **ataccama.com**