Time Is on My Side: A Case Study of Bulk Metadata Curation using Alation

Naveen Kalyanasamy, Andrea Levy, Aaron Kalb

Alation Inc.
Redwood City, CA
Info @ https://www.alation.com/

Abstract: This case study quantifies 211 work-days of time saved by one organization using an automated bulk-curation feature of the Alation Data Catalog. Data stewards are charged with keeping data documentation up-to-date. Often this means manually editing the individual metadata for which they are responsible. With bulk curation, data stewards save time and can more easily keep the metadata current. We study how the Alation platform assists data-driven companies in saving time and effort through faster, better, and broader data documentation.

Nearly half of organizations using Alation utilize the catalog set feature to perform bulk metadata curation, and this case study highlights the impressive time savings that one organization achieved. The time savings and benefits of using the platform extend well beyond this area but are not covered in the present work.

Keywords: Stewards, Metadata Management, Data Curation, Data Culture, Data Governance

1. Background

For organizations striving to be data-driven and aiming to promote a data-culture, providing users with contextual information about the data (i.e. metadata) in their organization is key [1][2]. Metadata that is fully described and up-to-date is the cornerstone of effective metadata management. Cataloging the data that an organization uses, enabling access to this data across the organization, and providing context and analytics around these data assets is essential to achieving a return on investment from asset management [3]. Curating accurate and extensive metadata across the data sources enables users to find, understand, trust and use the right data, and can contribute to enhanced net output of individual teams.

This case study investigates one biotechnology company's time savings and efficiency gains in the metadata curation process through bulk metadata updates in Alation – a software platform which (among other things) enables bulk metadata curation using the "Catalog Sets" feature. 49% of organizations[1] using Alation leverage this feature to curate their metadata.

2. Theory

Alation provides the "Catalog Sets" feature to help data curators group data objects together using custom rules. This tooling aids curators by enabling them to update metadata in bulk across all of the grouped objects. The flexibility and capabilities of catalog sets enable companies to employ them in the context of data governance, protected data (e.g. personally identifiable information – PII), and risk management. Catalog sets generally group data objects of the same type together, so one might have catalog sets of schemas, tables, or attributes (columns). There is also a custom catalog set that allows the user to manually group specific objects together. The alternative to performing these bulk metadata updates is that curators would have to navigate to each data object that they would want to update metadata for and manually update each one. Depending on the scale of the data within the organization and the different data sources that exist, this would require a significant amount of resources.

Even though there are broader benefits to using the Alation platform, the aim of this case study is to measure the effort that the catalog set feature, by itself, would save compared to the scenario where Alation does not exist and curators needed to update the metadata manually.

[1]Organizations reporting usage data to Alation

3. Data

We use anonymized catalog set data logs that quantify the time taken to create a catalog set and the number of members (data objects) grouped under each set. Details of the individual sets and data objects cannot be identified through these logs, but the scale of the sets does allow us to quantify their time savings over the alternative of manually curating the metadata of these set members. In this case study, we explore how two stewards at a biotechnology company leveraged this feature to curate metadata across their organization.

4. Method

4.1. Catalog Set Configuration Time

We use the data logs generated during the configuration of each catalog set to measure the time that a user spends creating and customizing each catalog set. The standard catalog set creation process involves the following steps:

- Naming and describing the set
- Establishing rules/conditions to group data objects together
- Specifying the metadata updates that will be propagated to all members of the set

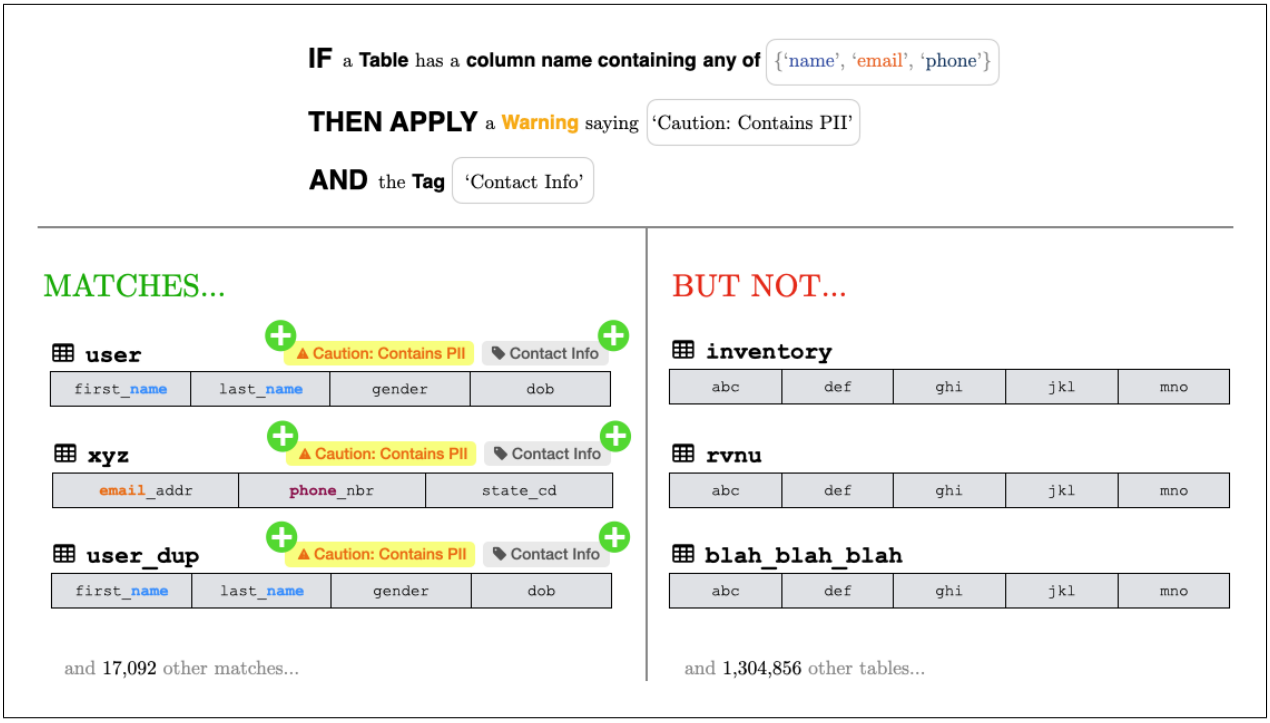

Fig. 1: Example catalog set creation process for identifying PII data

Type	Set count	Total creation time (minutes)	Member Count	Average creation time (minutes)
Schema	2	2	18	1
Attribute	8	20	174,975	2
Table	11	40	935,737	4
Total	**21**	**62**	**1,110,730**	**3**

Table 1: Catalog Set creation time and grouped members count

Fig. 1 depicts a fictional example catalog set scenario to demonstrate their utility. This example catalog set identifies and tags all columns containing PII data across an organization's data sources. The catalog set groups all the columns that contain "Name", "Phone", or "Email" and tags them in the catalog and also adds a warning for users consuming the data.

Table 1 shows the total catalog set creation times for each catalog set type at the specific organization in this study. On average, it took three minutes to create a catalog set.

4.2. Time to Update Metadata Manually

We use logs containing the number of data objects in each catalog set to estimate the time it would require a steward to update the metadata manually.

Type	Average time to search and find (seconds)	Number of searches performed
Schema	7.5	144
Attribute	9.3	83
Table	5.1	3719

Table 2: Average time to search and find data objects by type

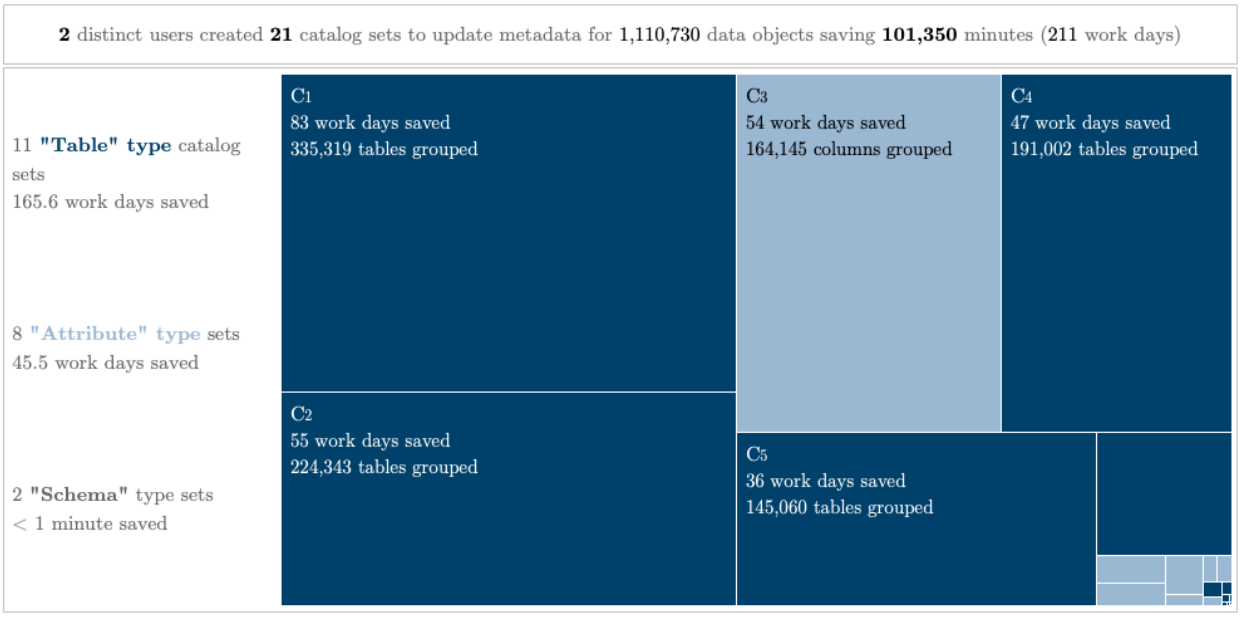

Fig. 2: Break down of 211 work days of time savings from catalog sets. Depending on the number of data objects grouped into them, catalog sets contribute significant time savings

Manually updating the metadata of a data object involves:

1. Searching and finding the data object to be curated (tables, columns, schema)
2. Performing the required metadata updates to the object

Since the required metadata update is subjective to each scenario, it is not possible to accurately quantify the time taken for manual metadata updates (item 2 above). Hence, we use a conservative lower bound by equating the time it would take to search and find each data object (item 1 above) as the time it would take to update the metadata of each data object manually. We use this organization's search logs to identify the time taken to manually search and find attributes (columns), schema, and tables. Table 2 shows the average time to find each type of data object.

Catalog set	Type	Creation time (minutes)	Member count	Time saved (hours)
1	Table	14	335,319	475
2	Attribute	5	164,145	342
3	Table	11	224,343	318
4	Table	2	191,002	271
5	Table	2	145,060	205

Table 3: Top 5 time-saving catalog sets and the time taken to create them

4.3. Time Saved By Catalog Sets

The time saved through catalog sets is the time it would take to update the metadata manually versus the time it takes in Alation using catalog sets.

The time saved T_S, by a catalog set is calculated as:

$$T_S = (|S| \times T_U) - T_C, \quad (1)$$

where $|S|$ = count of members in the set,

T_U = time to update each member manually,

T_C = time to create the catalog set

5. Results

Using the above methodology, we aggregate the time saved across all of the 21 catalog sets in this organization based on the creation time and number of data objects in each set. Fig. 2 demonstrates the 211 work-days of savings that catalog sets produced.

By spending 62 minutes implementing catalog sets, this organization was able to save 211 work days across the 21 catalog sets that were created for 1,110,730 data objects in the platform. The majority of the savings comes from the 935,737 tables grouped and curated using the 11 "Table" type catalog sets and the 174,975 attributes (columns) curated using the 8 "Attribute" type catalog sets.

Table 3 shows the top 5 sets that have saved the most time at this organization (also annotated in Fig. 2). Catalog set 1 in the table took 853 seconds (14 minutes) to set up, to determine the rules to group, and to update the metadata for 335,319 tables. If we compute Equation 1 with $|S| = 335{,}319$, $T_C = 853$ seconds (14 minutes) and $T_U = 5.1$ seconds (from Table 2) we get time saved as:

$$T_S = ((335{,}319 \times 5.1) - 853) \text{ seconds} \quad (2)$$

which comes out to about 1,709,274 seconds (475 hours).

The 14 minutes of set creation, thus, saved 59 work days (8 hours in a work day) through the automated bulk update. Catalog set 5 on the other hand, was set up with 2 minutes of effort to update metadata for 145,060 tables, saving 26 work days.

The time saved in this analysis depends partly on the estimated time to update each data object manually (in section 4.2). Even if we revise that estimation to say that it takes 1 second to update each data object, this organization would still save 38 work days of effort through the catalog set feature. A 1 second update time would presume that the process was automated in some other way, which would also require extra effort by the curators to set up the automation process.

6. Summary

The Alation platform enables easier and more effective maintenance of comprehensive metadata around data objects. In this case study, we see that:

- The catalog set feature enables curators to save a significant portion of their effort by updating their data objects' metadata in bulk without having to manually update each object. By enabling bulk updates, it is more likely that metadata will be populated and stay up-to-date.
- One organization leveraging Alation's catalog sets to curate their metadata saved 211 work days across their organization with only one hour of effort.

Realistically, when tagging and categorizing data objects would take multiple months to perform by hand, it is more likely that the counterfactual is that the data objects will not be curated at all (or kept up-to-date) and PII data may go unidentified. This lack of curation and metadata management could introduce risk to the company in terms of data protection and privacy issues, which would come with a greater cost than even the time spent to curate the data objects manually.

This case study doesn't cover the additional savings curators would achieve for any updates needed to the metadata or any new data objects that need to be added in after the initial configuration of the catalog set.

The results in this case study are also not exhaustive of all the returns an organization can derive from Alation but do quantify a lower bound of time saved by actual organizations using one component of the Alation Data Catalog. Additional information on the capabilities of the platform can be obtained from the website [4].

References

[1] Dataversity. *Metadata Management Drives the Future of Data Management*. URL: https://www.dataversity.net/metadata-management-drives-the-future-of-data-management/. (accessed: 07.31.2020).

[2] Dataversity. *Demystifying Metadata Management*. URL: https://www.dataversity.net/demystifying-metadata-management/. (accessed: 07.31.2020).

[3] KMWorld. *Metadata Management*. URL: https://www.kmworld.com/Articles/Editorial/Features/Monetizing-digital-asset-management-The-power-of-metadata-management-134815.aspx. (accessed: 11.15.2019).

[4] Alation Inc. *The Alation Product*. URL: https://www.alation.com/product/. (accessed: 11.15.2019).