Alation

Consuming Data Productively

The five ways a data catalog drives productive self-service analytics

90% of the time that is spent creating new reports is recreating information that already exists. You have plenty of data at your fingertips, but how do you turn it into a meaningful insight? Has this inquiry been made before? How do you find the right data to fit your analysis? And when you do find data sets that look relevant, how do you make sure that the data you have is trustworthy and accurate?

Finding & Using the Right Data

Without a way to effectively share prior work & identify verified data sources, analysts and other data consumers spend most of their time re-creating information that already exists. Rather than uncovering new insights, too often we're re-hashing old findings.

The problem is that despite the fact that most organizations have adopted a wide variety of powerful analytic, BI and visualization tools, sharing data knowledge through those tools is actually quite difficult. In understanding data, context is key. But despite the world-class tools that have been deployed, very few of them store the meaning of the data next to the insights that they uncover.

Let's spend a few minutes looking at why that's the case.

90%

of the time that knowledge workers spend in creating new reports is recreating information that already exists.

The Challenge of Context

Data in most organizations is stored in a variety of repositories, some physical such as paper-based reports or inside the heads of an organizational brain trust and others electronic (spreadsheets, databases, or Hadoop). Asking a question of the data requires knowing or discovering which repository is most likely to store the data that you need.

Each of data store has its own format for storing data. In the case of a database, the format is a relational table. In the case of Hadoop, it is a file. Encryption and compression techniques used to optimize the physical footprint of data can add another layer of complexity to the format that the data is stored in. For example, data stored in files in Hadoop can be further stored in different formats on Hadoop like Avro or Parquet. So understanding where to look for data requires understanding both a physical location and the technical format that the data is stored in.

In addition to understanding the location and format of the data, in order to effectively navigate data, you also need to know how to interpret the data. Data is often stored with a limited repository of descriptions of the data. Data stores typically capture only the physical metadata of the data- things like the field type is a string, numeric, etc. Users have access to the physical data, but what's often lacking is the business context. How do you know what the data means or how it is meant to be used?

To understand the business context of the data, most organizations have evolved a complementary but very ad hoc collection of wikis, email threads, and chat sessions. These unconnected tools are used to identify subject matter experts (SMEs) and to communicate about data. Sometimes more formal documentation is produced by a data steward or data curator in a technical repository, or by a developer in the source code. Even in the most advanced analytic organizations, analysts often rely on physical notebooks to store business metadata notes.

If every organization had a fully annotated version of their data, the world would be a better place. But, in reality, navigating all of these informal, disparate systems to access data knowledge can be a full time job. And, without a system that can automatically populate full technical and business metadata, the vast amount of data flowing through a typical enterprise can become overwhelming. Information about data can get lost, and finding reliable answers from among all these

possibilities can be difficult. Which is probably why 58% of business managers still make decisions based on intuition, even when data is available[1].

Executives make big decisions frequently and review them often. Even when data is available to inform a decision, many times it is not used because the process of accessing, validating and applying the data is too cumbersome. Hence the need for data systems and processes to evolve to a point at which they can operate at the speed of business.

58%

of business managers still make decisions based on intuition, even when data is available

The Cost of Tribal Knowledge

What you often see is that, even in the best cases, where executives are using data to make decisions, they fall back on organizational tribal knowledge to prepare their analysis. They set up a series of face to face or phone interviews to inform the process of analysis by gathering expertise. Social networks in the organization provide the trust in data that the executive needs to make a good decision. In essence, this natural process is a technique for both validating the veracity of the data and creating the business metadata necessary to fully understand the context of the data used for analysis. All good best practices for data-driven decisions.

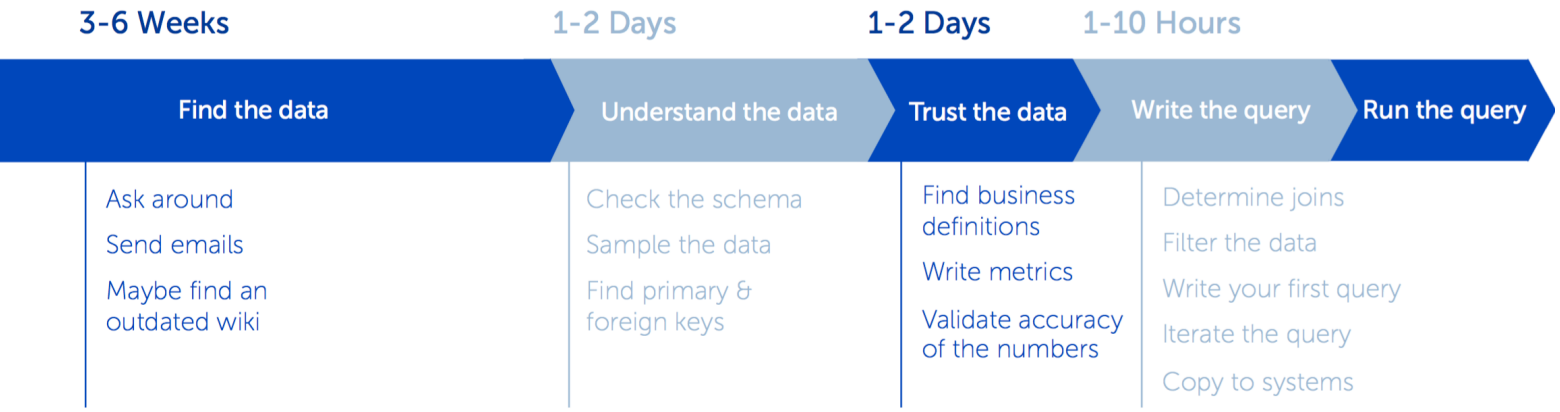
But unfortunately what happens only too often is that after collecting this valuable metadata, it is documented in a physical notebook, placed in a desk drawer, and forgotten. Sometimes its forgotten forever, although in the luckier cases it is only forgotten until the next time that a data-driven decision needs to be made in this same subject matter area.

As you might imagine, it takes a significant amount of time to find subject matter experts, gather information and confirm the validity of data sets. Part of this is the time it takes to meet in person or via phone, and part of this is the time it takes to navigate the silos between organizational units that make it difficult to find experts or even simple answers. This impacts productivity and turns analysis into a time consuming and expensive proposition.

---

[1] according to *Gut & gigabytes: Capitalising on the art & science in decision making*, a September 2014 survey report by the Economist Intelligence Unit

In research that we've done with clients, getting to the point of performing analysis can take anywhere from 3 weeks to 2 months. For a typical organization, the journey of finding the right data looks something like this:



What is the solution to this time wasting and unfocused activity? In working with our customers, we've found that software that creates a living, collaborative catalog can make data easier to find, understand and trust. When you externalize and aggregate tribal knowledge from the heads of all your subject matter experts, you can create a comprehensive portrait of enterprise data knowledge that extends understanding to everyone in an organization. Combined with new self-service tools for analyzing and visualizing data, modern catalogs provide a new way to maintain and apply data for insights.

What Is a Data Catalog?

You're likely familiar with lists or inventories that help source the products and services you need. Before the Internet existed, there was a giant yellow phone book that contained page after page of products and services available in a specific geographical area. Or you might recall the help-wanted section of a local newspaper that listed available jobs. Due to space limitations, these inventories offered only the most basic information about a product, service, or job.

If you wanted additional context about the value of a product or service or company — for example, if you wanted to know how well the offering stacked up to competitors or whether the suggested price was reasonable — you'd have to talk to neighbors, friends, or colleagues. Or spend an hour or two in the library researching the backgrounds of specific companies to determine if you wanted to

do business with them or work for them. Though the inventory provided basic factual data, you needed to supplement it with information from trusted sources to better understand the value of the goods or services to you.

Today modern catalogs such as Amazon, Google, LinkedIn, Yelp or Airbnb have largely replaced their physical counterparts, while offering much more in the way of context about the data contained within. They record both the data itself and a social graph of interactions with the data.

Think about Google for instance. Google has an inventory of all of the websites published to the internet. This inventory is enriched by Google through an algorithm called PageRank which tracks what the relationships are between web pages. In addition to tracking the relationships between different pages, Google's catalog of the internet also uses the behaviors of individuals searching for web pages to enrich the catalog with information on relevance of web pages to human interest. And Google is continually enhancing the catalog with more and more information automatically gleaned from both machine and human interactions with the objects of interest - in their case, web pages.

A data catalog works the same way. By starting with a sample of the data and the physical metadata an inventory is created. That inventory is then enriched by observations that are made as machines and humans interact with the data. This creates a catalog that is the single point of reference for your data, the physical metadata, and the business metadata associated with it.

Modern catalogs are also updated as soon as the information changes, in real time, and with very little friction. Instead of being outdated, these catalogs actually become more useful over time. And they provide richer contextual information through a combination of both machine learning and data verified by humans.

To understand how human verification works to enrich these catalogs, think of a modern catalog such as the restaurant guide Yelp. The number of

> "Finding the right data that mattered to me would take **30 minutes** in conversation with someone. I'm looking to…find the data that matters to me in **5 seconds.**"
>
> — **Data Scientist**
> Financial Services Company

reviews and the rating system (calculated by machine) tells you something about the quality of the listing. While the ability to dig deeper into information about the restaurant by reading reviews, provides further confirmation by your fellow humans.

Like Yelp, a data catalog contains the business rules and best practices associated with your data. This information can be shared by analysts themselves or defined by an IT Steward or Chief Data Officer to ensure the truth and accuracy of the data in question.

Other catalogs work similarly. LinkedIn is a catalog that replaces a rolodex of contacts or a file folder of resumes. By collecting and analyzing connections between individuals, LinkedIn provides trusted information about your professional network. And in many cases, it replaces the manual work that you likely hired a recruiter to do in the past.

Airbnb is a housing catalog based on a two-way marketplace of property suppliers and consumers. It captures user and supplier feedback to create a trusted information hub for vacation experiences. In the process, it replaces with software recommendations the manual work of a travel agent.

Similarly, Amazon is a catalog of products that includes rankings and recommendations from consumers to help you make a more informed decision as you select your purchase. It replaces recommendations you may have received from friends with suggestions from other consumers - and from your own online behavior - to increase your confidence in the value of your selection. You have the opportunity to provide feedback, validating or confirming your experience with the product.

Google may actually be the most interesting of these catalog examples because of its use of the algorithm called PageRank. Before Google, advertising dollars influenced the search results of any website. With Google's innovative PageRank algorithm - which mapped the relationships between web pages and consumer

links to those pages - the top consumer-used pages showed up higher in rankings than paid content in a Google Search. Google users received more immediately useful content based on the actions of other users.

A modern data catalog mirrors these catalog features: it collects and analyzes metadata and the connections between the individuals who are using data within your organization. Depending on the algorithms used, you'll receive trusted contextual information in a self-service format that aids the speed and accuracy of your data usage, including search, querying, and collaboration with your colleagues.

> "[I had a] physical notebook I kept on my desk with notes about all the tables I used and all the ways to join them."
>
> — **Product Manager**
> Online Retailer

The Value of AI in a Data Catalog

Catalogs observe and analyze the relationships between people and data. At a time where the size of stored data is increasing at a monumental rate, the speed by which machine-based algorithms can capture these relationships is critical. Hiring more people to manually document data is no longer a scalable solution. The more we automate the collection, cleansing, preparation, documentation and dissemination of data, the faster we can derive value from our data.

Does this mean that machine-based artificial intelligence (AI) has more value than human input? Quite the opposite in fact. In the catalog examples referenced above, human input through crowd-sourcing or direct confirmation creates singular value for the catalog's recommendation system. How valuable would Amazon be if you only saw product listings without the user ratings or recommendations? Ratings and recommendations help you determine which products are most attuned to your needs. In other words, they provide the context that makes the information valuable. And this human feedback helps the catalog filter products so that the most popular shift to the top of page listings for a more efficient and rewarding shopping experience.

Machine algorithms are essential because they automate information capture. But to be most effective, machines must learn from people and about people and

their preferences. Creating and implementing accurate AI systems requires the input of human knowledge.

Behavior I/O: How observations drive better behavior

Algorithms get better as they receive more human feedback. A well-known example of this phenomenon is LinkedIn's "People you may know" feature. As users confirm a match made by LinkedIn's algorithm, they're also teaching the machine what a good match looks like. The machine subsequently gets better and better at suggesting matches. In essence, humans "train" the machine through a feedback mechanism that can be termed "Behavior I/O."

In computing, input/output (or I/O) is the communication between an information processing system - such as a computer - and the outside world, possibly a human or another information processing system. In the consumer world, companies like Fitbit learn from the behaviors of people to help train other people to behave better, in Fitbit's case to be more physically active to improve your health. In the world of data, we can use machine learning to observe the behavior of analysts. And then use those observations as Behavior I/O to allow other data consumers to use data in a better way based on the direct input from analysts and from practices observed.

So how does a data catalog use observations of analyst behavior and human input to provide value to other users? The human-machine learning system inherent in a Behavior I/O feedback loop provides the following benefits:

> "[When teams] collaborate and share queries with each other, we reduce the amount of repeated work by
>
> 20%
>
> — **VP of Analytics**
> Analytics Firm

Behavior Correction

One of the most common problems in data analysis is incorrect use of data. People often use the wrong report for analysis repeatedly, even when the new correct report has been available for weeks. This is not only true for reports, but also for the data assets that make up the report, including filters, data transformations and calculations. By observing user behavior, a data catalog with Behavior I/O might surface a warning to alert a user that they have selected an outdated data asset. This alert could contain both a notice that a more accurate data asset is available and provide a direct link to the correct asset, improving time to success.

Detailed Data Context

If you're using a Tableau workbook that everyone else uses, how do you know if the underlying data is accurate? Even if you know that 323 people ran the same query in the last 20 days, do you know who used it and why? What if 90% of those queries were run from a different workbook in another subject area? A data catalog with Behavior I/O observes user actions to provide the history of how all of the assets linked to a workbook were used - from the workbook to the sheet, the underlying SQL query and back through every database and Hadoop system that touched that data to the very source where the data was originally stored. The value? Better conclusions - since you understand the full lineage of your data and how others have used that data before.

Just-In-Time Guidance

With machine-human collaboration, Behavior I/O can provide "just-in-time" suggestions at the point of consumption based on the behavior of other users. For example, you can receive a suggestion for the most relevant table based on your input as you are writing a query. This saves time compared to contacting people to confirm that you are using the right table. Similarly, you might receive a warning via email as soon as a column you've used in a query is deprecated by another user may save you from using column filled with inaccurate data that would lead to wrong conclusions. These recommendations inline speed actions and align user activity.

Aggregated Input

When observed human usage inputs are aggregated, the larger data set helps you determine usefulness. You can conclude that a query used by only one person over a year ago is less useful than the query run by 153 people as recently as last week because of the benefit of aggregation. Similarly, knowing which of 16 tables on the same topic were used most frequently in the last year and endorsed by two subject matter experts will help you choose between very similar assets. Aggregated data helps you validate decisions.

Five Data Catalog Features that Improve Time-to-Decision and Analyst Productivity

We've found that customers who embrace the value of Behavior I/O in a data catalog increased the analytical productivity of each analyst and business user by up to 50%, and increased the speed of accurate documentation by up to 40%.

These benefits are within reach of any data consumer in a data-driven organization. To positively influence the behavior of your analysts - or frankly anyone that self-serves data in your organization - our experience shows that following five features of a data catalog lead to improved time-to-decision and productivity:

50%

efficiency boost for data **consumption**

40%

efficiency boost for data **documentation**

1 Unified view for all your data

A data catalog provides the most value when it has a robust search capability that covers captured metadata and observed users behavior across all datasets. Make sure that the catalog you choose offers you a combined view of all your data, not just a view of a subset of your data or one type of data. For example, a catalog for just Hadoop, or just relational databases will have limited functionality. In order to find the right data, you need to be able to search through all your data without exception.

2 Machine-human collaboration to enhance data context

Some data catalogs function as simple inventories - without the important behavioral observation capabilities described in this white paper. It's important to find a catalog that not only provides an automated repository of all your data, but also incorporates a machine-human learning system with algorithms designed to provide context about your data and how it is used. A Behavior I/O feedback loop allows your catalog to become smarter over time. As more and more human user behavior is observed and confirmed, you'll receive remarkably fine-tuned contextual information. Data context derived from machine-human collaboration enhances decision making.

3 Verification of sources so you can trust your data

Most data catalogs will provide a data lineage feature to allow you to trace the sources of your data. As you choose a data catalog, look also for additional measures of verification, such as data flags or annotations. This capability allows users to endorse an asset of value, or provide a warning or deprecation if an asset is outdated or inaccurate. Direct human verification increases trust in data.

4 Just-in-time guidance to help you better understand your data

Significant value is added when a catalog can provide "just-in-time" suggestions based on the behavior of other users. A catalog that is responsive to user input - providing usage-based guidance at the point of consumption - can save you significant amounts of time. And may help you bring new analysts and business users on board more quickly.

5 Collaborative capabilities to break down organizational silos

Collaborative capabilities built into the fabric of your data catalog can transform the way your teams interact. And may materially impact the insights discovered.When teams of analysts work in silos, work is re-created instead of re-used, and a great amount of organizational knowledge remains unshared. Look for Wikipedia-like capabilities to share information across teams, and integrated communication tools to enable direct dialogue between team members and other experts. It's an advantage if these assets are searchable. In this way tribal knowledge can be captured and codified, and shared with others across geographies and across time.

Making the most of your data

When you have large volumes of data at your fingertips, knowing how to find and use data sets that are most relevant - with data you know is trustworthy and accurate - changes the way you do business. A modern collaborative, contextual data catalog can solve the problems brought on by an over-reliance on tribal knowledge.

By deploying a data catalog with the five essential features outlined in this whitepaper, you can gain the most value from your data: increase access to all your data, receive timely and accurate usage-based guidance in-catalog, benefit from collaborating with other users, and free your data workers to uncover new insights.

Alation is the first data catalog built for collaboration. With collaboration, analysts are empowered to search, query and collaborate on their data to achieve faster, more accurate insights. Alation automatically captures the rich context of enterprise data, including what the data describes, who has used it, and the fit between the data and different types of analysis. Alation's catalog is generated and updated using machine learning and improved through human collaboration between analysts, stewards, experts and business users.

Alation is funded by Andreessen Horowitz, Bloomberg Beta, Costanoa Venture Capital, Data Collective and General Catalyst Partners. Customers include eBay, MarketShare and some of the world's largest finance and retail firms. For more information, visit alation.com.

Contact Us

info@alation.com
(650) 799-4440