OCTOPAI

The 3 Must-Haves for Every DATA CATALOG

Building a data catalog is a strategic milestone for any data-driven organization.

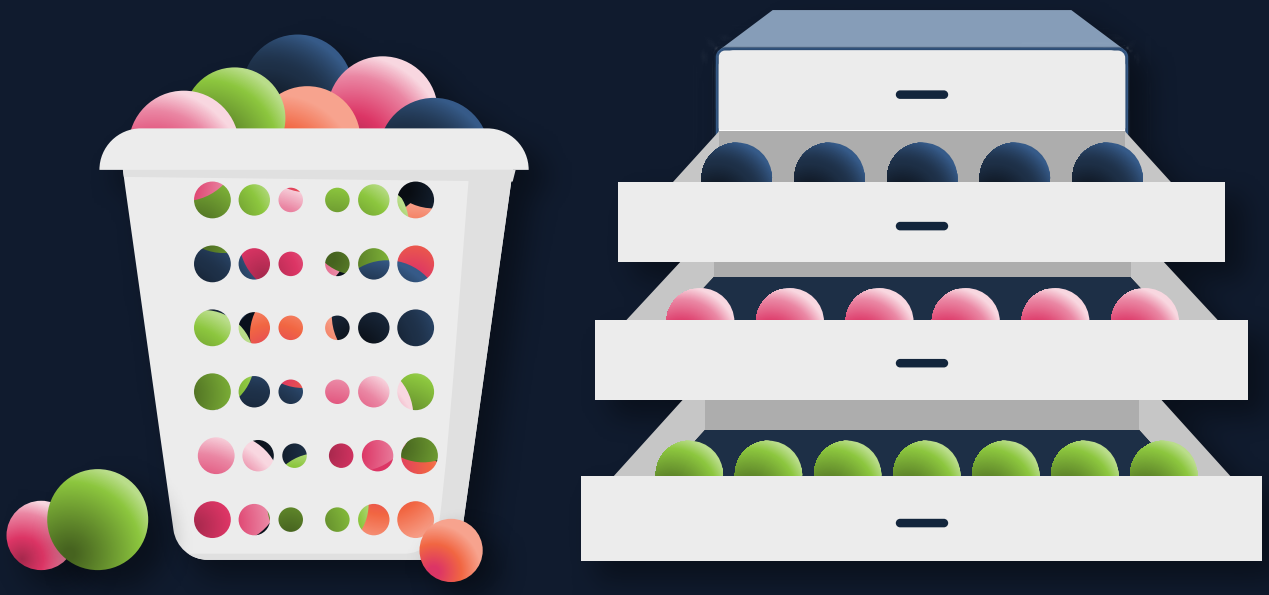
So if your company has decided it's time for one, you need to make sure the catalog has all of the necessary capabilities that will save time and help create value from your data.

Data Catalog: The Need

So many errors, false assumptions, and general misunderstandings result from ambiguous terminology.

When team members incorrectly assume that everyone has a shared understanding of the terms being used, chaos can ensue. This is why most contracts have a section right at the top that defines the terms used in the document, even seemingly trivial terms such as "us" and "you."

A common understanding of the terms and their underlying data used in a business is essential, especially in large, diverse, global companies where not everyone has the same technical, cultural, or business background. The key to this common understanding is what's known as a data catalog.



What Is a Data Catalog?

We all know what a catalog is: An alphabetical listing of terms and their definitions, usually consisting of the terms used in a book or by a particular field of interest. A data catalog is similar—in essence, it's an alphabetical listing of data assets, their technical definitions, and their associated business terms and meanings as used in a specific business. It provides an authoritative source of meaning for the data used in all of a company's operations.

It sounds simple, right? After all, everyone knows what a "customer" is, so why do we need a formal definition for it?

Dig a little deeper in any good-sized company, however, and you'll see why it's needed.

Depending on where the data is coming from, fields that mean the same thing could have completely different names, or the same concept might be represented in different ways.

Even something as simple as a person's name can get complicated; in different systems, it might be represented as:

- A single name field, formatted first_name last_name
- A single name field, formatted last_name, first_name
- Separate fields for first and last names, with a wide variety of names and labels:

  firstName, Name_F, FName…

Wait, I was looking for the customer's name - is this even the customer's name? Maybe this is the salesperson's name?

Throw in data sources from foreign countries, where field names and labels might be in foreign languages, and the possibilities are quite literally endless. Having a data catalog ensures that a business term and its underlying data mean the same thing across all departments, business units, and geographies. It serves to reduce or eliminate confusion and assumptions regarding the meaning of a term as applied by all data citizens.

However, a data catalog isn't just a convenient listing of terms to keep everyone on the same page.

It's an important reference for business analysts, data analysts, data scientists, data governance managers, everyone on the BI team, and all data citizens. That's why it's key that the data catalog is user-friendly for all different types of users, from IT teams to the least technical report user. The catalog must be something your data citizens actually want to use, and they will if its interface is intuitive and helps them to meet deadlines, independently.

How Do Data Citizens Use the Data Catalog?

Think of the data catalog as an enabler - it's the key to unlocking the value in data.

How? By creating transparency into:

- What data is available
- What the data represents and means in technical and business terms
- Where the data exists
- Where the data originated
- Who is responsible for the data
- How the data could be, should be, and is being used

Common uses of the data catalog include:

- **Consistency** - A good data catalog helps data professionals ensure that data assets, reports, and dashboards represent data consistently, both within and among the various objects.
- **Management** - A common understanding of business terms makes it easier to create, maintain, and integrate new sources of data in the environment.
- **Development and delivery of new reports and dashboards** - Standardization enables data professionals to make the connections needed between data elements that mean the same thing but have different names.
- **Deriving value from data** - Enables using data correctly, efficiently and effectively to complete data-driven initiatives.

In short, a data catalog is useful not only to the day-to-day business users but also to the BI technical staff and all data citizens.

Side note: Don't confuse the term "data catalog" with "data dictionary." Although both are based on metadata, a data dictionary is a low-level reference describing the attributes of columns and tables in a database. It has little to do with the higher-level business terminology. Business users will consult the data catalog but will probably never see a data dictionary.

Building a Data Catalog

To build a data catalog, a good place to start is existing reports, databases, and processes.

Why?

Because these are the places where important business terms and their relation to their underlying data end up. If "customer," "cost of goods sold," or "full-time employee equivalent" are important terms for a business, they're going to be represented in reports somewhere and those terms tie back to the data assets in the databases etc.

There could be a fair amount of detective work involved, or something akin to archaeology in understanding data when you don't have effective literacy in place. Some of the data sources could be from 10 or 20 years ago, if not longer, and the people who originally worked on them are now long gone, leaving little or nothing in the way of documentation. In some cases, figuring out what certain things mean might be left to making the best guess.

The trouble with the manual approach is that it is a tedious, painstaking process that, depending on the size and complexity of the business, will take a whole BI team or professional services group several months, and that's before considering the maintenance involved in the manual approach. Simply tracking down and interpreting all the metadata requires a significant effort by a small army of analysts.

It's a process fraught with frustrations, such as:

- Incomplete, incorrect, or obsolete metadata
- Data transformations and calculations that don't mean what people think they do
- Conflicting assumptions within data citizens regarding the meaning of a given term
- Time-consuming for many different people in the organization

Happily, there's a better way.

The 3 'Must-Haves' for Your Data Catalog

1 Automation

Building a data catalog the old-fashioned way can become mired in the details, intricacies, false leads, and dead ends of your labyrinthine data environment. In doing this manually, many BI teams get lost in the weeds and eventually get burned out on pursuing a project that doesn't seem to make much progress over time.

To shorten the project cycle and take your data catalog to the next level, you can have your data catalog generated automatically.

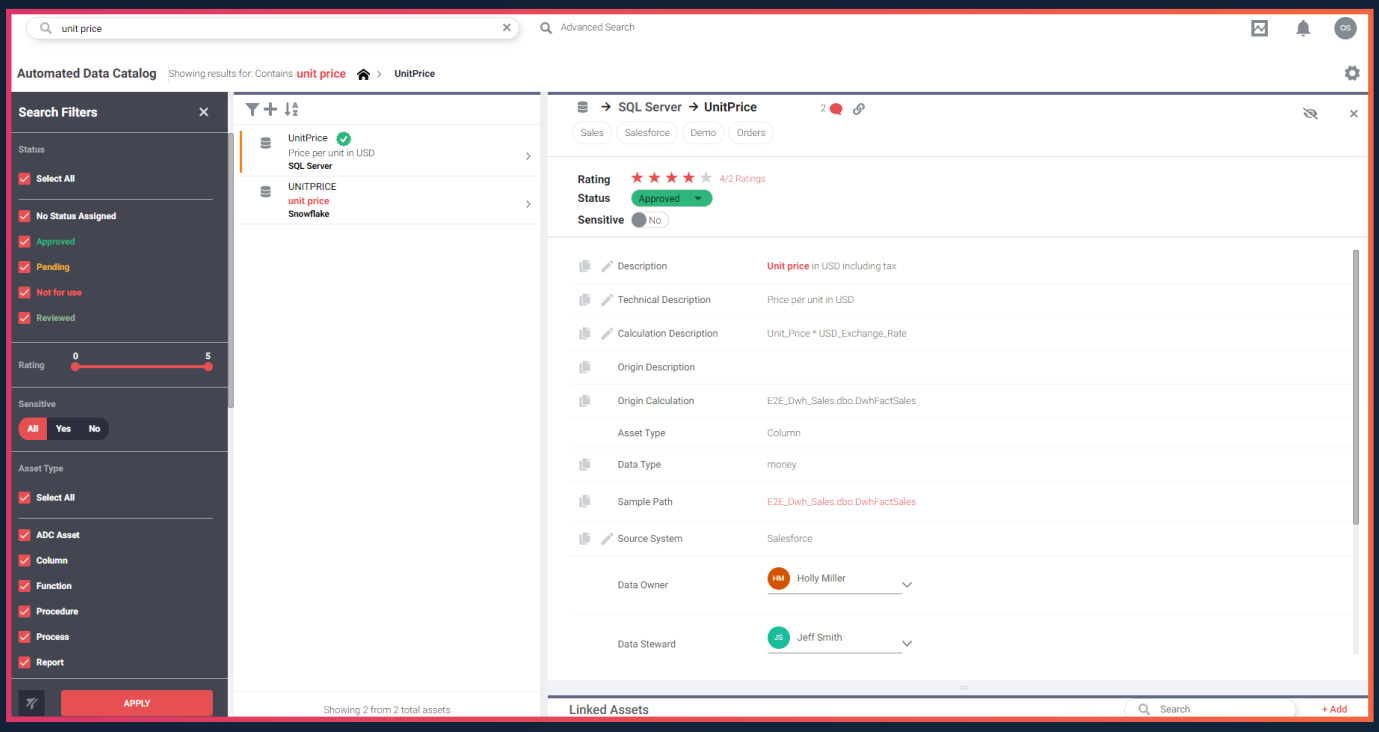


Figure 1: Results generated within Octopai's Automated Data Catalog after searching the term "unit." **Left panel**: Search results of all places this term appears within physical, semantic, and presentation layers of each reporting system. **Top right panel**: All the attributes for the selected term (in this case the 'Unit Price' column in the SQL Server) such as Descriptions, Source System, Owners, Tags and more. **Bottom right panel**: Linked assets for the selected asset which help users understand that those assets are related.

Generating your data catalog automatically will simplify and streamline the execution of the following tasks:

- Centralize data layers from the entire BI landscape including reporting systems, databases and processes:

  In this context, the term data layer refers to any component of the data structure model that includes the physical, semantic, and presentation layers:

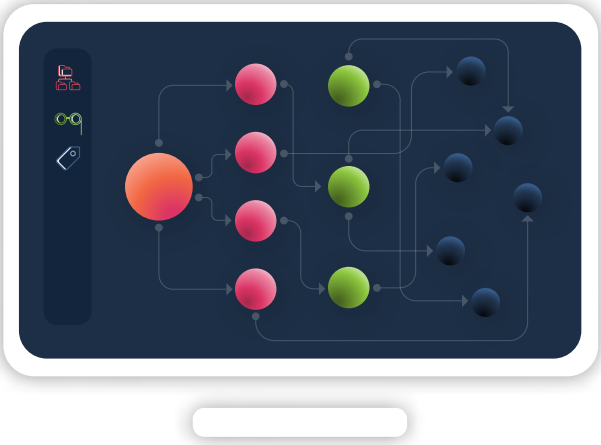
  - Physical layer: This layer includes the actual tables and columns that define the data structure at the database level.
  - Semantic layer: The semantic layer represents the data from the user's perspective. It hides the details and complex relationships of the underlying physical layer and packages them into data objects that correspond with the organization's understanding of the information in common business terms. (This is one of the reasons why a solid data catalog is so important.)
  - Presentation layer: This layer is what shows up on columns in reports and dashboards, and it often involves some kind of summarization or aggregation into a small number of key performance metrics. It can also be in multiple languages.
- Standardize data layers across different reporting systems: Centralizing the data layers involves standardizing them across all of these systems.
- Reduce or eliminate large-scale data entry projects: No one wants more tedious grunt work, and there are few activities as mind-numbing (or error-prone) as manually entering data assets, especially data assets that already exist somewhere in some form.

The key to success for all of these tasks is automation.

2 Integrated Data Lineage

While having all definitions in place for data assets has huge benefits, to complete the picture you will need integrated data lineage as well. Lineage will provide traceability of the data to understand where the data originated and where it is being used. If you're thinking there are different types of lineage and wondering which of them is important for traceability, the answer is all of them.

Still wondering why integrated lineage is a must? Say someone from finance is using a report and needs assistance in understanding a KPI. He refers to the catalog which includes definitions of what data the KPI represents, what it's associated with, and who's responsible for the data asset, in this case, the KPI. Now that the user has an understanding of what the data represents, he has a technical question that has to do with how the data is sourced. The answer will help him better understand how it can be used. To do this, however, he will need to collaborate with the data steward responsible for this KPI. Without integrated data lineage, the data steward would have to spend hours or maybe even days hunting through and for the processes involved in creating the KPI. Integrated data lineage provides built-in, instant visibility into not only how the data is sourced and what's involved in creating it, but also how and where it's being used.



3 Built-In Collaboration

The above anecdote highlighted why built-in collaboration is key. Collaboration about data is an integral part of using data, especially if you're a data-driven company. When collaboration is not built-in, your users are communicating out of context. This causes repetitive conversations with different users and loss of valuable tribal knowledge.

With built-in collaboration, your users know who they should collaborate with, how they can collaborate, and the collaboration takes place in context of the relevant data asset within the data catalog.

What Happens When You Automate a Data Catalog and Add Built-In Lineage and Collaboration?

A data catalog can offer the following benefits:

- **One source of truth**: Having a single source of truth about your company's data enables you to build better, more consistent reports and dashboards and gain more control over your data assets. In turn, this enables effective use of data making sure that any data citizen knows how to locate any data asset, how it should be used, what it represents, and who is responsible for it.
- **Common business language**: Having a common business language means that people in different departments and business units have a common understanding of every important term, and nothing gets "lost in translation." This can prevent embarrassing internal or public miscommunications.
- **Improve self-service BI**: The modern trend of enabling self-service BI provides great benefits by reducing the load on the BI staff and improving the turnaround time for desired reports and dashboards, but it can be a disaster without a centralized marketplace for the data. A data catalog means that any data consumer can efficiently locate any piece of data to use in the most effective way, whether it be to understand an existing report, build a new one, use data for advanced analytics projects or even enable self-sufficiency for that data scientist you just added to your team for the latest AI initiatives your business is longing for.

But when you apply automation to generate a data catalog from your company's data landscape, and add to it built-in lineage and collaboration, the benefits grow profoundly:

- *Time and money savings*: The ability to build a data catalog automatically has the obvious benefit of reducing the time and labor costs required to complete the project. It also has the extra added benefit of reducing wear and tear (read: burnout) on your BI team and the rest of the organization that is involved.
- *A snapshot in time*: As noted earlier, constructing a data catalog manually can take a long time, and to keep everything straight during the project, you have to either freeze all the data assets (i.e., prevent adding new assets or deleting or modifying existing ones) or try to keep up with changes as they happen—neither of which is especially practical. With automation, you can get a "snapshot in time" to serve as your baseline and then continually monitor for changes and update the baseline, automatically.
- *Increase report accuracy and reduce errors*: Automating the data catalog reduces the chances for error inherent in manual processes, meaning that the reports that are built on the basis of the terms in the catalog will be more accurate and have fewer errors.
- *Instant visibility into the data flow*: Integrating automated data lineage into your data catalog enables seamless visibility into the processes and sources involved in creating the data, and where it is being used.
- *Preserving tribal knowledge*: Subject matter specialists exist in every company and in every line of business, since employee turnover is increasing this poses a threat to companies that lose valuable information that later prevents effective use of data to its full potential. A data catalog with its many capabilities provides a platform in which tribal knowledge is preserved, and this has huge benefits for every organization.

The Time for an Automated Data Catalog is Now

The time to decide to automate your data catalog is now—before you embark on a data catalog project. Automating your data catalog pays for itself many times over—not just in the initial data catalog project but in the ongoing maintenance tasks to keep the catalog up to date.

Octopai's metadata management automation platform can be the cornerstone of your data catalog. With Octopai, you can get set up and start the process of building your data catalog within one business day, eliminating the manual drudgery and freeing your BI team up to work on deriving insights—the work they were hired to do.

In addition to importing all metadata items from the data landscape, we also import the associated original descriptions for those items, instead of doing it manually. Huge **savings**.

Thanks to Octopai's built-in collaboration and lineage, users can also communicate in context and get full visibility into the data flow.

If you are planning, starting, or restarting a data catalog project, to get your project across the finish line much faster and more accurately, and bring your company the consistent business terminology it needs to succeed in no time, make sure that the three must-haves of automation, integrated data lineage, and built-in collaboration are a part of it - otherwise you'll be left with an outdated catalog that doesn't provide the entire picture that your data citizens are looking for.