data.world

WHITEPAPER

Data Discovery Without Limits

How a knowledge graph unlocks the comprehensive spectrum of search

By Ole Olesen-Bagneux, PhD and Juan Sequeda, PhD

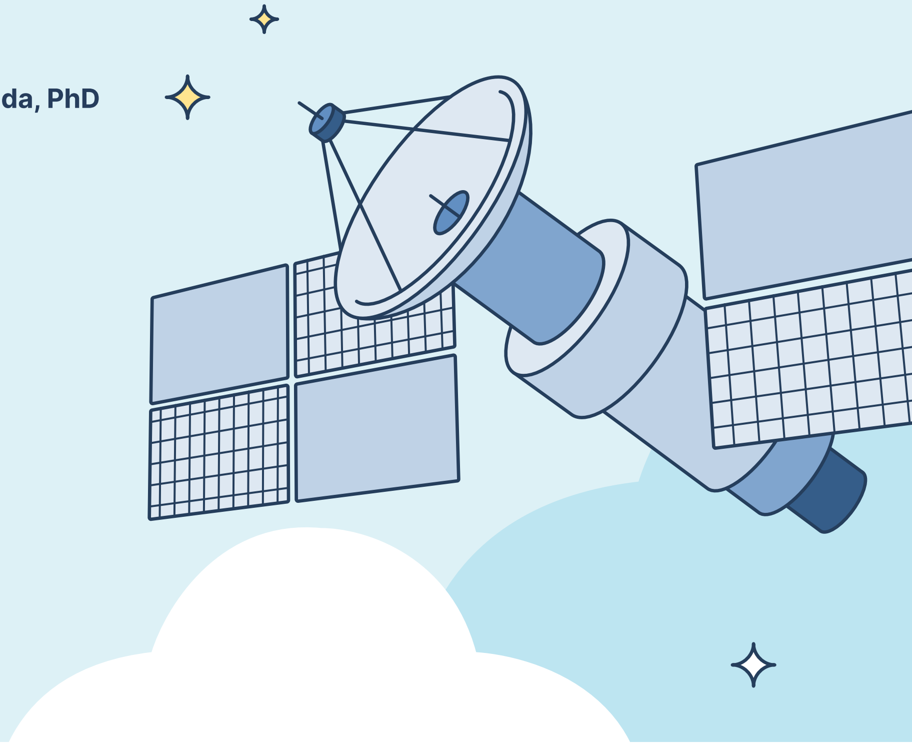
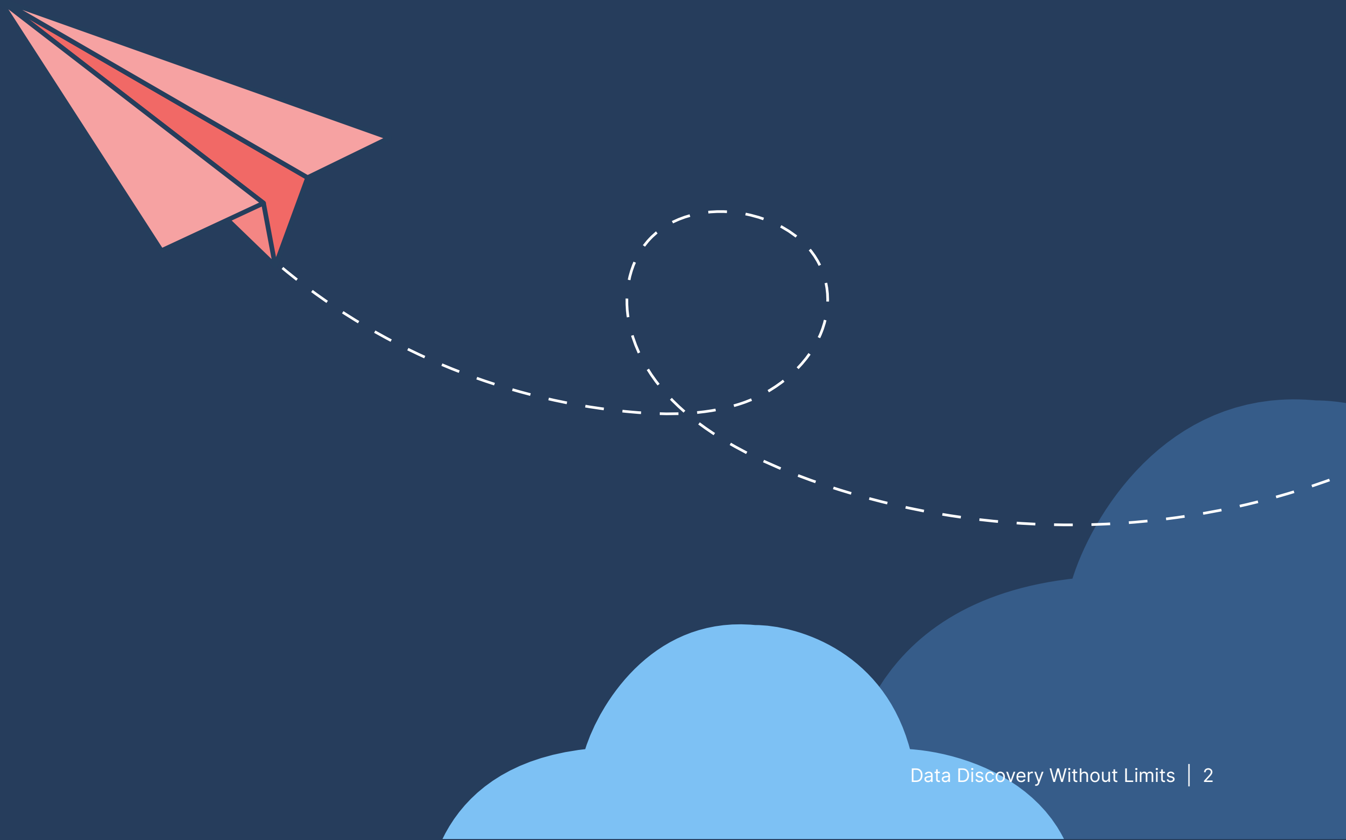


Table of contents



Introduction

Think of a data catalog like a search engine for all the data in your company. Search is the entry point for users to find data in order to solve critical business problems. However, searching for your enterprise data can be a tough, if not disappointing experience.

For example, you may be searching for data about a specific customer, but the results delivered by your data catalog consists of a long laundry list of irrelevant resources. Maybe you're trying to locate data within a specific department related to a specific product — but none of the keyword combinations or filters retrieve anything useful. Sound familiar?

We will discuss these use cases in more detail in a moment. But first, we must look at why traditional data catalogs struggle to search for data and how a newer, more modern breed of catalogs — built on a knowledge graph — are disrupting data discovery, data governance, and cloud migration. This whitepaper will explain how knowledge graphs allow people to search for absolutely anything they want — even beyond data and metadata.

Powerful, trusted, and comprehensive search is critical for the following tasks:

01

Discovering relevant, useful data for analysis

02

Understanding the provenance and lineage of sensitive data

03

Identifying data that most urgently needs to be migrated to the cloud

The Spectrum of Search

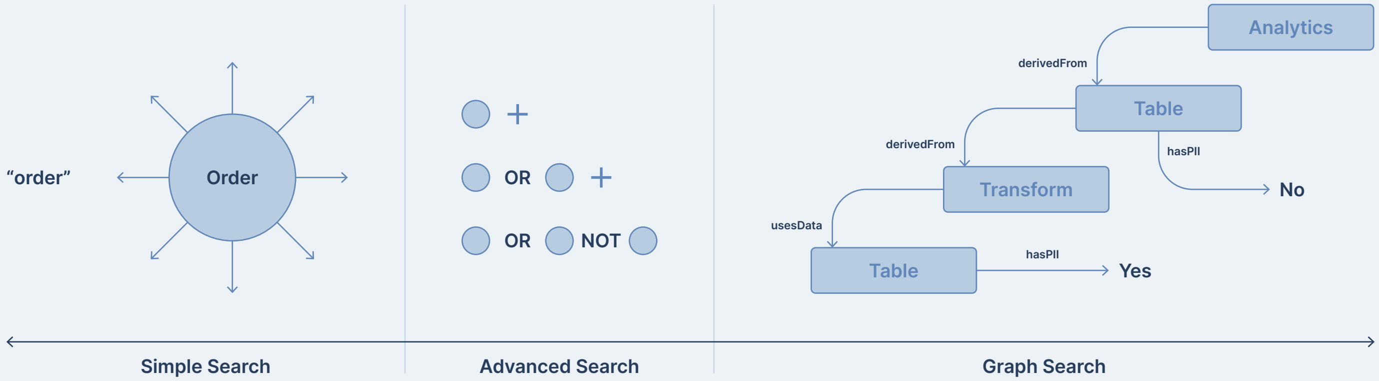
A data catalog is a collection of metadata that represents the IT landscape of an organization It's the map of the data and knowledge of an organization. The catalog tells you what IT systems exist in your organization, what kind of data lives in those systems, and who works with that data. You don't search catalogs like you would a traditional database via SQL but rather as a collection of metadata. In data catalogs, you search for data — it can be data sources or datasets, terms, people, everything really, that can be understood as either a data asset — or in a more modern approach — a data product.

Not all searches are created equal. For example, maybe you're hungry and want to see hamburger recipes. You perform a basic query that returns a laundry list of results that you don't mind scrolling through for inspiration. But what if you're gluten free, live in a small apartment on the upper east side of Manhattan, and don't have time to cook? Now the burger query is a bit more well-defined, and you're searching for a single result to satisfy your needs and quench your hunger.

Accordingly, all searches live within a defined range: a *Spectrum of Search* (Figure 1). This is how search works with a search engine in a browser, and it's how search should work in a data catalog. The more expressive searches you make, the more you include ranges, sets of topics combined with AND, OR, NOT, the better you navigate the relationship between resources.

The spectrum of search consists of the following:

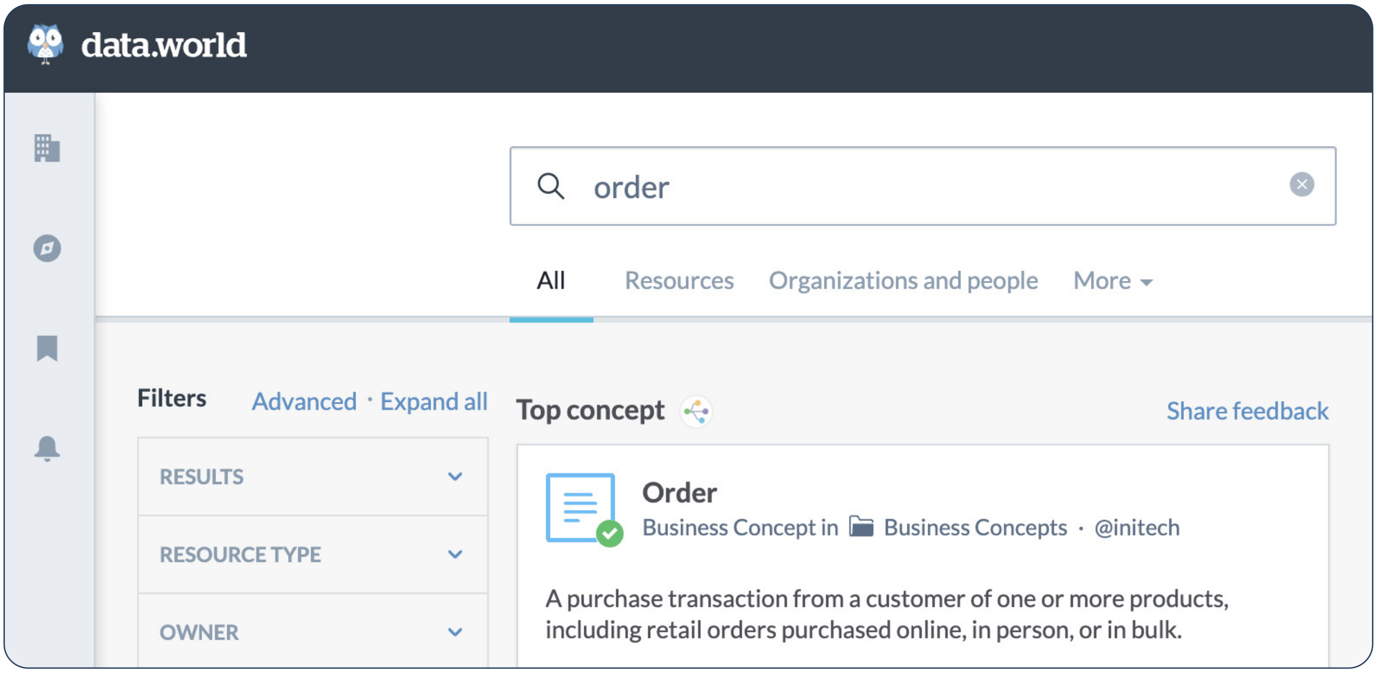
- **Simple Search:** Given a keyword, match it to a relevant concept and take into account the surrounding context.
- **Advanced Search:** Given a keyword, filter through customized metadata fields and apply operators such as AND, OR, NOT.
- **Graph Search:** Navigate the graph, exploring the relationships by writing a graph query language, such as SPARQL, to search and find anything.



A knowledge-graph-powered data catalog enables the entire spectrum of search. Let's take a look at the search experience gets more powerful and expressive the deeper you explore.

Simple search

Simple search is when you type keywords, e.g. "Order" in the search bar, on the entry page of the data catalog:



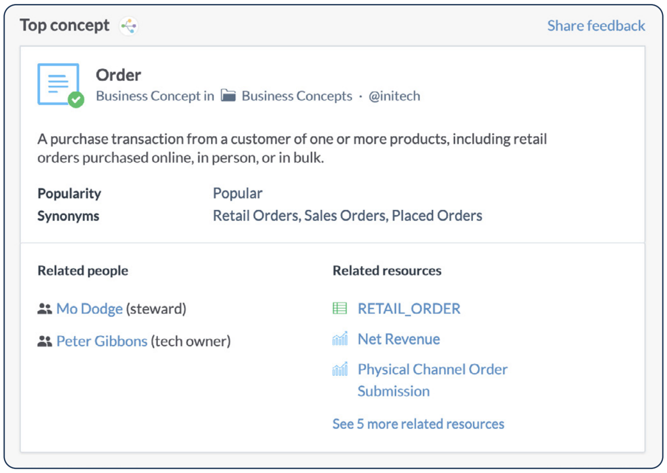
A purchase transaction from a customer of one or more products, including retail orders purchased online, in person, or in bulk.

Figure 2. Simple search

When it comes to search, simplicity is in the eyes of the beholder. But, the simplicity that you are experiencing when you search for a single term actually requires a complex technological setup behind the scenes. Simple search is normally powered by indexing that allows for the ranking of search results based on perceived relevance. With indexed search, the word, "Order" will be returned as the top hits. This is the baseline case for any data catalog.

Knowledge graphs, however, increase the expressivity of simple search. With graph-powered search, the strength of potential hits are measured against a complete graph of your data, pointing you to most connected — and thereby strongest — search results. In other words, a knowledge graph doesn't just deliver an indexed list of items; it also infers context from the graph to provide more accurate responses or recommendations. This is why search engines like Google and Bing use knowledge graphs.

In the example to the right (Figure 3), you can see that a search for the term “Order” actually reveals a 360 view of the term by providing a definition, popularity score, additional resources, the name of the data steward, and other terms users have searched for. This is the power of simple search via a knowledge graph.


Top concept

Share feedback

Order
Business Concept in Business Concepts · @initech

A purchase transaction from a customer of one or more products, including retail orders purchased online, in person, or in bulk.

Popularity	Popular
Synonyms	Retail Orders, Sales Orders, Placed Orders

Related people

- Mo Dodge (steward)
- Peter Gibbons (tech owner)

Related resources

- RETAIL_ORDER
- Net Revenue
- Physical Channel Order Submission

See 5 more related resources

Figure 3. Unveiling the top concept via data.world Eureka Answers™

Advanced search

There are moments when simple search is not enough. For example, you may need to find all dashboards related to “Orders” that have an Approved Status, come from either the Sales or Marketing domains, and were updated in the last month. In order to accomplish that search, you need to combine keywords “Order”, metadata fields (dashboards have a status attribute that can have a value “Approved”) with operators like AND, OR, or NOT. Advanced search can seem meticulous, but mastering it can be a gamechanger for effectively finding data.

Sometimes, when you add more elements to your advanced search, you need to put in a little more thought to get your search just right. You can find guidance on how to do this here.

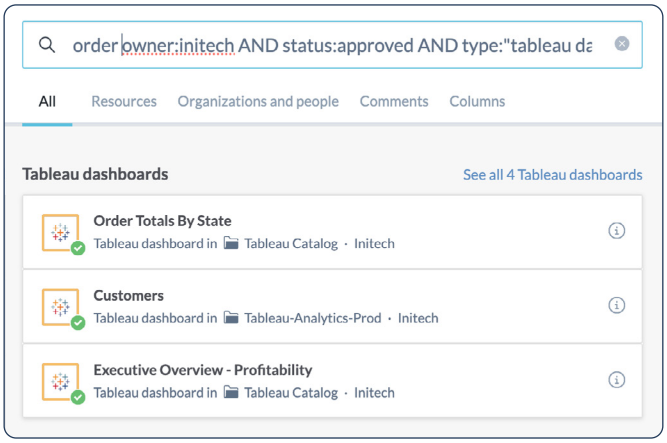

```
order owner:initech AND status:approved AND type:"tableau da
```

Tableau dashboards

See all 4 Tableau dashboards

Order Totals By State
Tableau dashboard in Tableau Catalog · Initech

Customers
Tableau dashboard in Tableau-Analytics-Prod · Initech

Executive Overview - Profitability
Tableau dashboard in Tableau Catalog · Initech

Figure 4. String-based advanced search in a data catalog

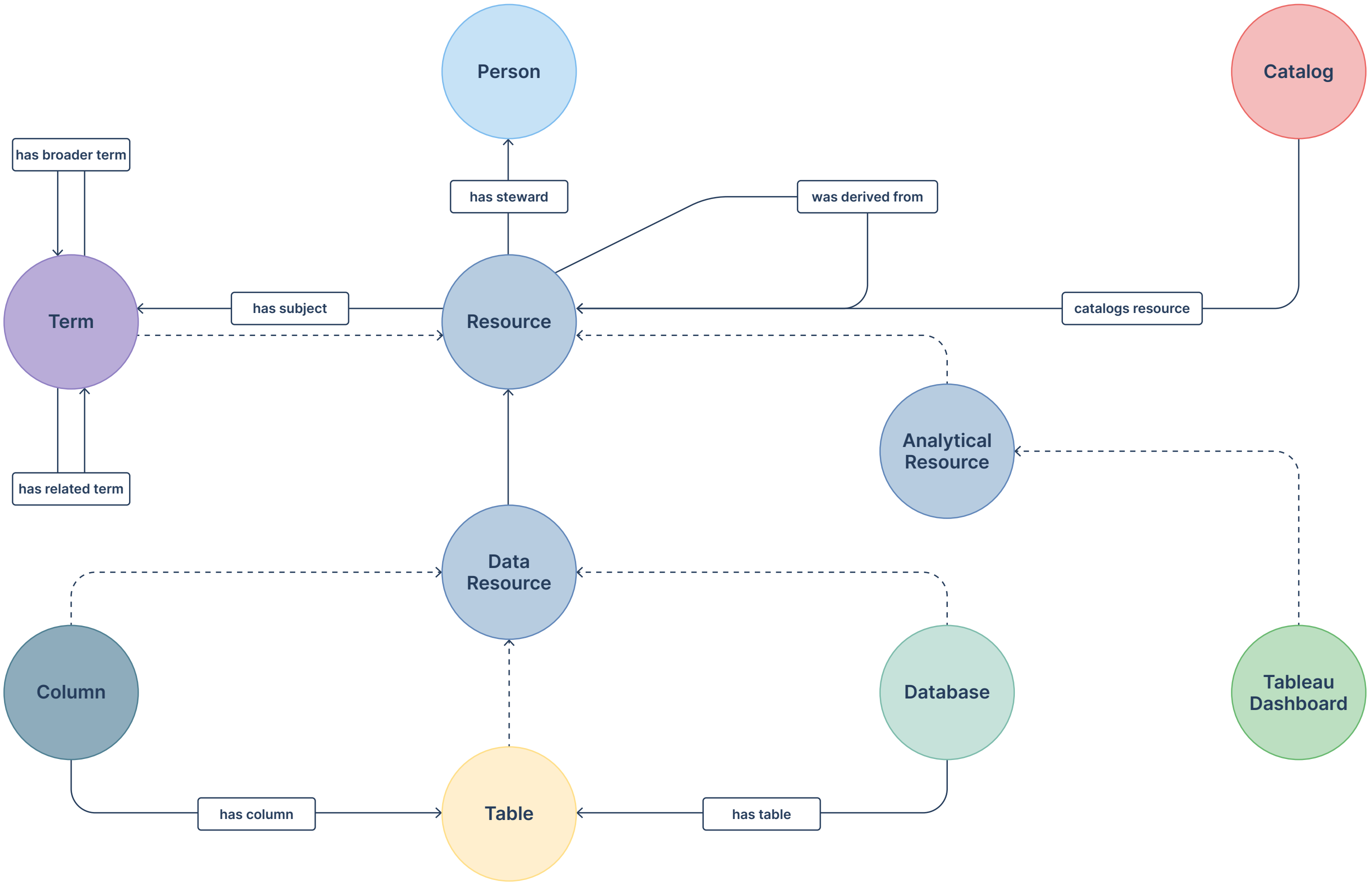
A knowledge graph powered data catalog provides full flexibility and extensibility on how to model metadata through an ontology. A data catalog ontology defines the concepts, attributes and relationships of how metadata resources should be organized.

A core data catalog ontology should consist of the following:

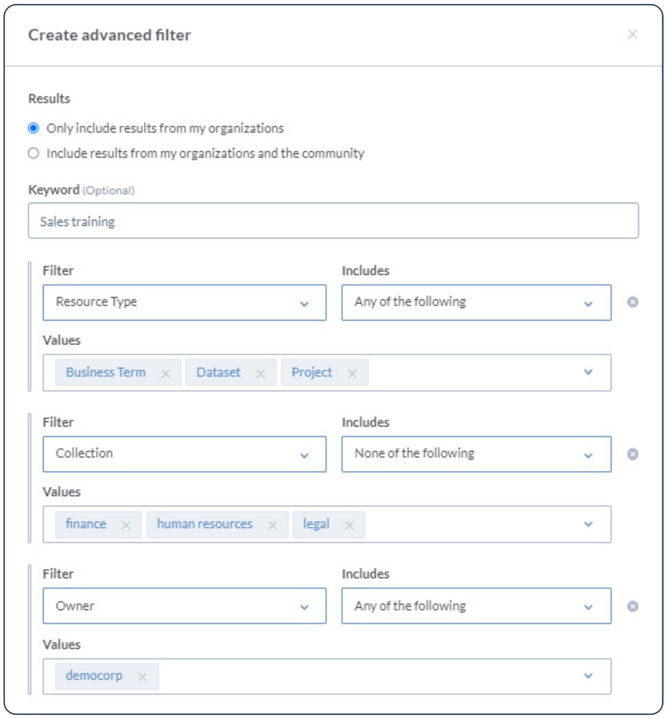
- A metadata resource can be either a Data, Analytics, or a Term resource.
- Data resources are Databases, Tables, and Columns. A database has tables. A table has columns.
- Analytical resources are Tableau Dashboard, etc.
- Term resources are business glossary terms and they can have relationships between them: related, broader, etc.
- A Resource can be related to a Term.
- A Resource has a steward which is a Person.
- A Resource can be derived from another resource.
- A catalog is a collection of Resources.

These elements can be used in the Advanced Search. Additionally, all of these elements can be fully extended upon.

If we put this all together, this is what the data catalog ontology visually looks like:



Advanced search can also be done using search builder tools in order to avoid writing long search strings. For example, a filter template (Figure 5) is easy to use and reduces the uncertainty that can come with not knowing whether a long search string is formulated correctly. Filtering provides a logical, operational overview of the advanced search.



Create advanced filter

Results

● Only include results from my organizations
○ Include results from my organizations and the community

Keyword (Optional)

Sales training

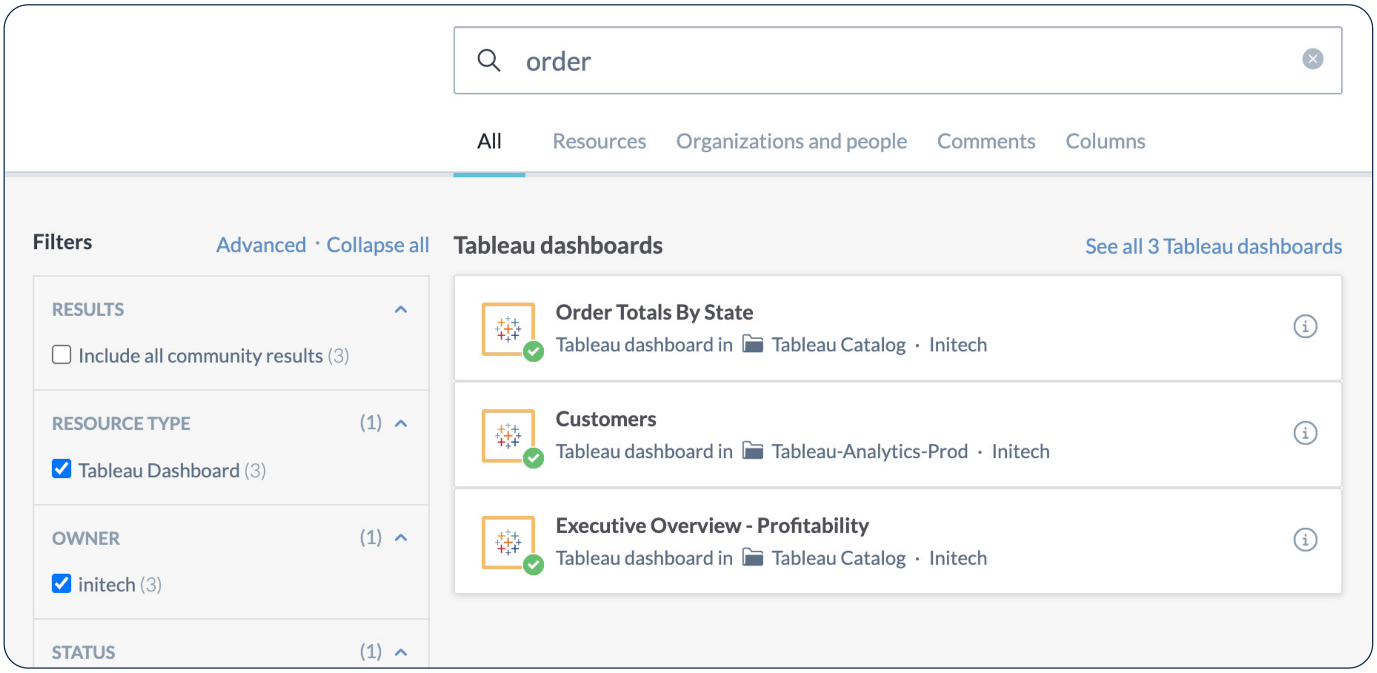
Filter: Resource Type — Includes: Any of the following
Values: Business Term × Dataset × Project ×

Filter: Collection — Includes: None of the following
Values: finance × human resources × legal ×

Filter: Owner — Includes: Any of the following
Values: democorp ×

Figure 5. Advanced search through filtering

Faceted search provides yet another method of advanced search that eliminates the need for lengthy, difficult to read query strings.



order

Filters — Advanced · Collapse all

RESULTS
□ Include all community results (3)

RESOURCE TYPE (1)
☑ Tableau Dashboard (3)

OWNER (1)
☑ initech (3)

STATUS (1)

Tableau dashboards — See all 3 Tableau dashboards

Order Totals By State
Tableau dashboard in Tableau Catalog · Initech

Customers
Tableau dashboard in Tableau-Analytics-Prod · Initech

Executive Overview - Profitability
Tableau dashboard in Tableau Catalog · Initech

Figure 6. Advanced search through faceted search

Graph search

Just like simple search, advanced search has its limitations. Consider a situation where your organization is audited for GDPR. You need to find all analytics that consume data from tables that do not have Personal Identifiable Information (PII) but that are derived from tables that originally did have PII. Additionally, you need to find the transformations that were involved.

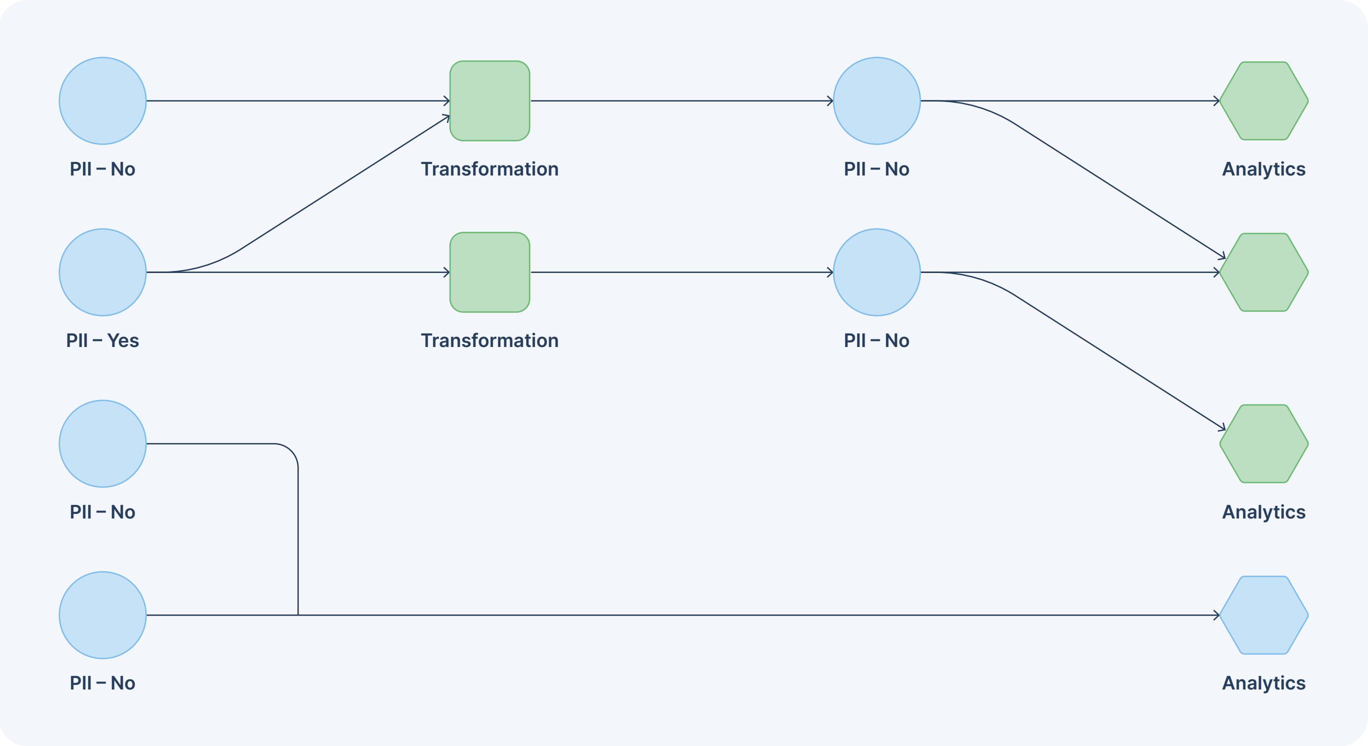
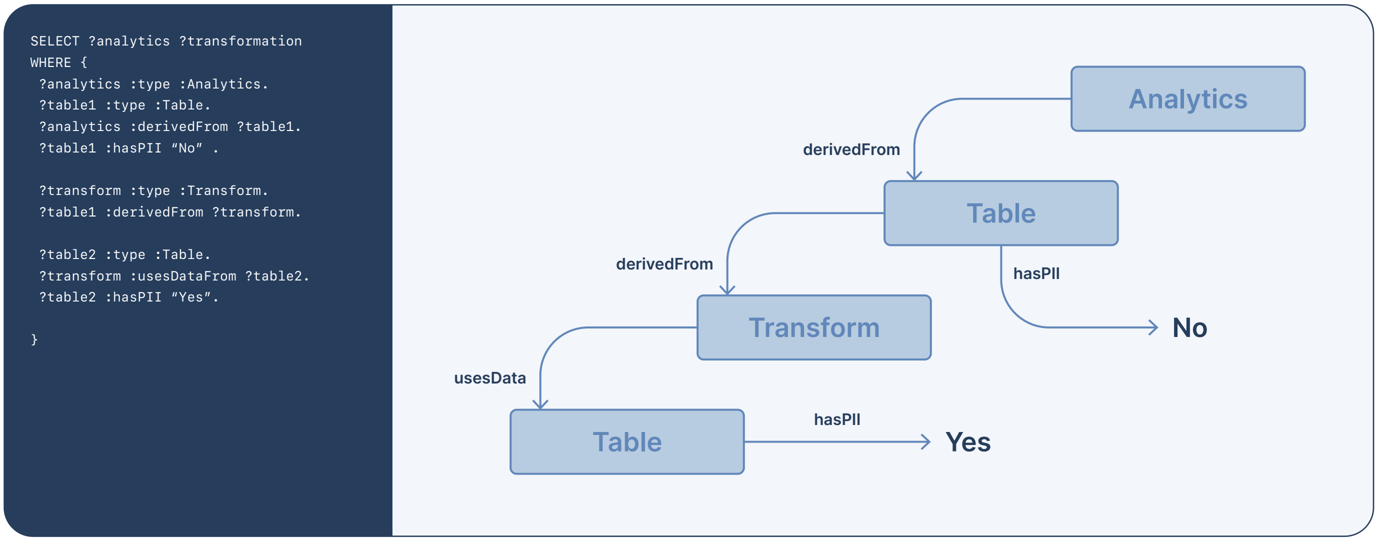


Figure 7. Graph of Analytics, tables with PII, transformations, tables without PITT

This goes beyond the advanced search capabilities. You need to both navigate and interrogate the relationships between the resources. There's simply no way to answer this question via strings, filtered, or faceted search. You cannot reach the level of specificity necessary for your search engine to comprehend the question.

This complex question, however, can be answered in a graph search. Figure 7 displays the relationships between analytics that consume data from tables that do not have PII data but that are derived from tables that originally do have PII data. Additionally, the transformations involved are represented by the green nodes.

Unlike traditional data catalogs, knowledge graph-powered data catalogs can be queried using a graph query language, such as SPARQL. For those unfamiliar with SPARQL, here's a useful tutorial. Figure 8 depicts the SPARQL graph query and its corresponding visualization to answer the question at hand.



```
SELECT ?analytics ?transformation
WHERE {
 ?analytics :type :Analytics.
 ?table1 :type :Table.
 ?analytics :derivedFrom ?table1.
 ?table1 :hasPII “No” .

 ?transform :type :Transform.
 ?table1 :derivedFrom ?transform.

 ?table2 :type :Table.
 ?transform :usesDataFrom ?table2.
 ?table2 :hasPII “Yes”.

}
```

Figure 8. Example SPARQL Query

The following common searches that can be performed in a knowledge-graph-powered data catalog:

Lineage path queries: Common use cases for data lineage include impact analysis (e.g. what is the downstream impact of a column change?) and root cause analysis (e.g. what upstream event caused my dashboard to break?).

While this visual feature will look appealing on a small amount of metadata when you need to see how data travels over tens or hundreds of columns and rows, the power of a knowledge-graph-based data catalog really flexes its muscle.

Example Lineage Path Queries	Business Value
Return a list of all downstream/ upstream resources for a given node	If there is a problem to be diagnosed in the data pipeline, a user does not want to go through a set of visualizations. Instead, they will want a list of resources that they can quickly check, thus reducing the time to identify the problem.
Does a path exist between two nodes	Understand how different resources are related to each other. For example, does a certain user have access to a sensitive table? If so, through what access controls? The resulting context helps ensure that compliance requirements are being met and reduces risk exposure.
Find the paths that connect to a node	Understand all the dependencies that a resource has in order to understand where to invest to manage and reduce the complexity.

360 Node queries: A knowledge graph enables you to understand everything connected to a node – not just one hop away but any number of hops away – providing a full overview of a given resource.

Example 360 Node Queries	Business Value
Return all the resources connected to a node	Provide a full overview of a resource, thus enabling visibility of what is known and unknown.
Identify the popularity of a node	Improve recommendations in order to reduce the time users spend looking for data.
Identify similarity between nodes	Create relationships between nodes to improve discoverability and reduce the time needed to find data.

Structural queries: Metadata is natively represented as a network of resources and the relationships between them. Just like a social network, metadata can have important nodes, cliques, and different communities. Understanding the metadata network provides insights on the IT landscape that can be used to identify and reduce complexity.

Example Structural Queries	Business Value
Identify central nodes that have a high node degree (many nodes connected to it)	A central resource implies that the node is highly dependent and can thus can be seen as a bottleneck. The right measures need to be put in place to keep it from affecting other processes. Identifying complex nodes that could be simplified helps reduce maintenance costs.
Similarity between different graph structures	The similarity may imply that there is duplicate data. Eliminating duplicate data helps reduce cost and risk exposure.
Identify isolated graph communities	The isolated graph community implies that it is not highly connected with the rest of the IT infrastructure and may not be business critical. Those resources could potentially be turned off, reducing cost.

Use cases

Now you understand the spectrum of search, let's look at how it can benefit your company. In this paper, we will only cover the three aforementioned use cases – data discovery, data governance, and cloud migration – but rest assured, knowledge-graph-powered search covers a far broader range of use cases.

Data Discovery

As a data analyst, you are tasked with running a shipping profitability analysis. To do so, you need to find and use data about Orders and Shipping Costs.

In a simple search, you would search for Order and Shipping Cost and sift through the list of results. In contrast, a knowledge graph-powered search would return a concept card describing the context of Order and Shipping with the related datasets – reducing the time you need to find data to complete your analysis.

In an advanced search, you would likely want to reduce the amount of results by filtering by specific fields. For example, the analysis is interested in datasets that have been updated in the last year, that are approved for use, and have a high popularity score. In addition to saving on the time needed to find the data, you are also finding the right data faster.

In a graph search, you may want to find the people and data stewards who have knowledge about the datasets pertaining to Orders and Shipping Costs. For this, you would navigate the graph to leverage the relationships between topics, datasets, and people. Now, you can have direct conversations with the people who can best provide guidance about the data. As a result, you avoid any potentially costly mistakes.

Data Governance

As a data steward, you are tasked with understanding the state of PII data and the risk exposure that the company faces in the wake of a data breach.

In a simple search, you would search for columns that are named with name, email, etc. that are likely candidates to have PII data. That search would return a list of columns and tables that you could then investigate to determine whether the data is properly masked and governed.

In an advanced search, you can filter by fields that already define whether the datasets have PII. Your notes show that the datasets have statuses, such as Approved and Pending. By filtering for the Pending status, you get a smaller, more accurate list of tables and columns to review.

In a graph search, you want to find the people who could potentially be responsible for the breach due to unauthorized access. You can query the graph to find all the users who have access to tables that are tagged as PII and join those users with the list of authorized users who can have access to PII data. As a result, you can more effectively identify any mismatches.

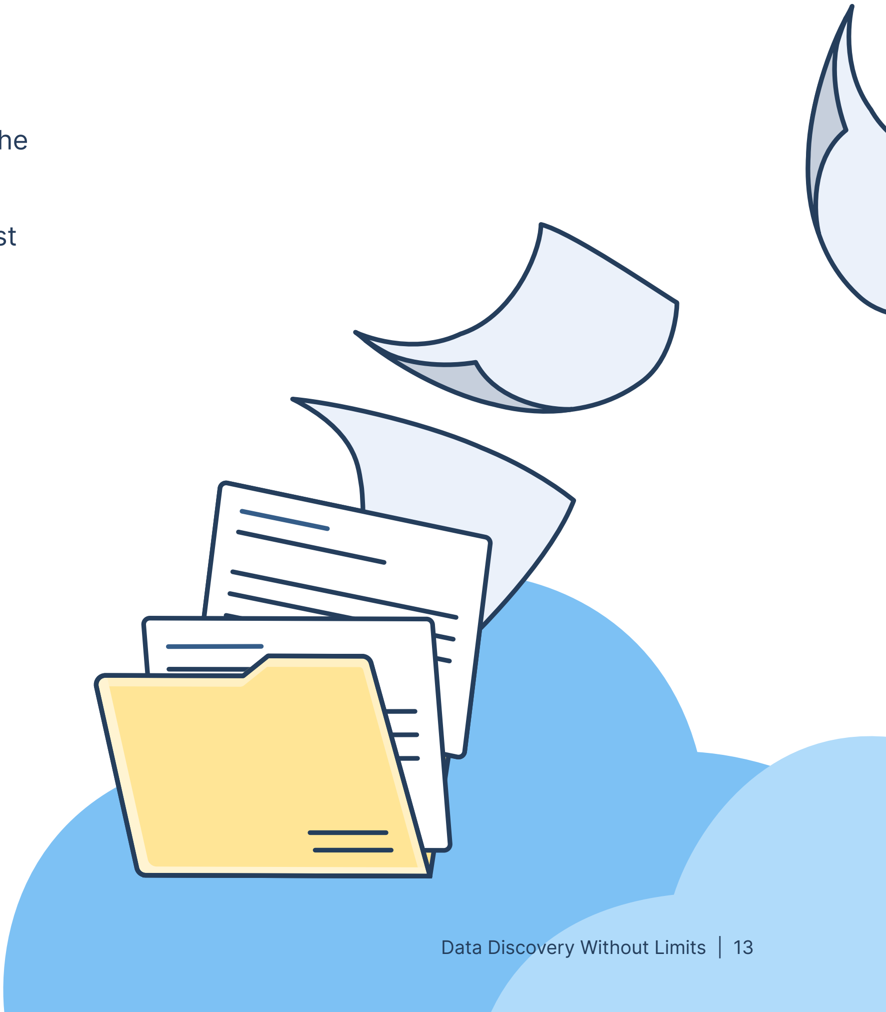
Cloud Migration

As a data engineer, you are tasked with creating a prioritized backlog of the on-prem data warehouse tables that need to be migrated to the cloud data warehouse.

In a simple search, you can find all the tables in the on-prem data warehouse. However, this requires combing through a long list of tables and as a first entry, the list is not all that useful.

In an advanced search, you can filter the results by selecting the tables that have a high number of columns and high popularity. With these filters, you'll get a better indication of the important and complex tables that need to be analyzed prior to a migration.

In a graph search, you want to focus on the tables that are critical for executive dashboards. For this particular use case, you would query the graph to find all the tables in the on-prem data warehouse that are upstream of the most popular executive facing dashboards and ordered by the number of joins that are used in views to create the tables. The ordering by the number of joins provides the level of complexity, thus creating a prioritized backlog (start with migrating the tables that are not that complicated and are critical to executive dashboards).



data.world

Conclusion

A data catalog is the front office for the data and analytics work that happens within an organization. Search is the starting point for everyone working with data and analytics. Thus, search is a foundational data catalog capability. Search represents a spectrum of expressivity from simple, through advanced, and all the way to graph search.

Not all data catalogs are created equal. We have described how knowledge graph-powered data catalogs support the entire spectrum of search:

- **Simple Search:** Find the relevant concept and provide the surrounding context.
- **Advanced Search:** Filter through customized metadata fields in the knowledge graph and apply operators such as AND, OR, NOT.
- **Graph Search:** Navigate the graph, exploring the relationships through graph query language, such as SPARQL, to search and find anything.

The spectrum of search provides richer experience for users to address use cases as:

- **Data discovery**, including finding relevant datasets all the way to finding the people who know the most about data.
- **Data governance**, including finding data that has PII all the way to finding the users who have access but are not supposed to have access to PII data.
- **Cloud migration**, including finding the tables of the on-prem database and creating a prioritized backlog of on-prem database tables that are critical to executive dashboards.

To see the complete spectrum of search in action, request a data.world demo.

Schedule a demo →