AN ECKERSON GROUP TOP TIER REPORT

Next-Generation Data Catalogs

Five Products You Should Know

By Lyndsay Wise & Wayne W. Eckerson
July 2024

Eckerson Group

ABOUT THE AUTHORS

Lyndsay Wise is the data governance lead at the Royal College of Physicians and Surgeons of Canada. Wise has worked as an industry analyst and evangelist and served as a director of market intelligence for a top BI firm where she ran customer and competitive intelligence and worked with customers to design and implement BI strategies. In 2007, Wise founded WiseAnalytics, and for eight years was an independent analyst and consultant providing services to small and medium-sized businesses on BI, analytics, data warehousing, dashboard deployments. Wise is the author of Using Open Source Platforms for Business Intelligence: Avoid Pitfalls and Maximize ROI.

Wayne W. Eckerson is an international thought leader in the data and analytics field since the early 1990s. He is a noted speaker, sought-after consultant, and widely read author. Eckerson has conducted groundbreaking research studies, chaired numerous conferences, and consulted with organizations around the world during his 25 years in the field. Eckerson has also written two books: The Secrets of Analytical Leaders: Insights from Information Insiders (2012) and Performance Dashboards: Measuring, Monitoring, and Managing Your Business (2005/2010). He has degrees from Williams College and Wesleyan University.

ABOUT ECKERSON GROUP

Eckerson Group is a boutique consultancy that helps organizations get more value from data. We specialize in developing actionable data strategies and tailored data solutions for commercial and non-profit organizations. Our consultants each have 25+ years of hands-on experience and specialize in data architecture, data governance, data literacy, self-service, data products, and operating models. Our clients say we are hard-working, insightful, and humble. It stems from our love of data and desire to help organizations turn data into insights and action. **Explore what Eckerson Group can do for you!**

TABLE OF CONTENTS

Market Trends

MARKET TRENDS

Faster, better, cheaper

That's the mantra of a new breed of data catalogs that are easy to use, geared to business users, and built natively for the cloud. These next-generation data catalogs offer 80% of the functionality of incumbent data catalogs at half the price. They are gaining ground quickly, adding features faster than we can document for this report!

Their timing couldn't be better

The market for data catalogs is growing rapidly, moving from an early adopter to mainstream market. Organizations big and small now recognize the importance of a metadata repository to support data discovery and self-service as well as data governance, compliance, data migration, and cost optimization. Meanwhile, incumbent data catalogs, such as Collibra, Alation, and Informatica, are struggling to gain traction with non-technical users and justify their six-figure annual license fees, creating an opening for a new breed of data catalogs.

5
VENDORS

Five Vendors

This report profiles five next-generation data catalogs: Actian Zeenea, Castor Doc, Metaphor, Select Star, and ThinkData Works. This is not an exhaustive list. There are about a dozen vendors we could have profiled. But these five vendors are good representations of this emerging category. Each profile in this report examines the product's heritage, focus, differentiators, and challenges.

Startups

The vendors profiled in this report are startups. Most were founded just four years ago in 2020, the exceptions being Actian Zeenea (2017) and ThinkData Works (2014). All have under $7.5 million in annual recurring revenue. Right now, CastorDoc and Select Star have the most customers, revenues, employees, and funding. (All companies profiled in this report asked us to keep their revenues and sales confidential; some also requested confidentiality for the number of customers, employees, and partners.) Interestingly, two were founded in France, CastorDoc and Actian Zeenea. CastorDoc now has its headquarters in Brooklyn, NY, while Actian will be introducing Zeenea functionality to existing and prospective U.S. clients.

Use Cases

Most of the products were designed to facilitate data discovery by regular business users. All capture, index, classify, and tag metadata and provide a user interface that makes it easy for users to find and evaluate the data they want. Some go beyond indexing metadata and provide direct access to data. These include ThinkData Works via a data virtualization tool, Actian Zeenea through a data marketplace, and Metaphor through data products published to domain pages.

To ensure trustworthy data, most products now offer some data governance features. Most support business glossaries, usage reports, role-based access, and automatic classification and tagging of personal or sensitive information. But many are adding bulk metadata editing, automatic term and table descriptions, schema change alerts, impact analysis, and policy engines. Some vendors, such as Select Star, also support data migration and cost optimization use cases.

Pricing

All vendors in this report offer aggressive pricing, at least compared to incumbent data catalogs, which cost around $150,000 a year for an enterprise license. For example, Select Star, which publishes its pricing, offers a $10,800/year package for 25 users, 2 data sources, and 250 tables. In general, an annual enterprise license for next-generation data catalogs runs between $70,000 and $90,000 a year. Pricing models vary. Most have user pricing, although ThinkData Works offers unlimited users and Actian Zeenea and CastorDoc offer unlimited viewers but charge for administrators and stewards. Metaphor offers unlimited connectors, tables, and assets.

Architecture

The products are all cloud-native, SaaS products. Some offer private virtual clouds, such as Metaphor and ThinkData Works. Most run on AWS, although CastorDoc also runs on Google Cloud Platform and ThinkData Works runs on all three major cloud platforms. Two vendors, Metaphor and Actian Zeenea, use a knowledge graph to support contextual relationships among assets, users, and interactions and facilitate expansive search. Most can pull data from on-premises systems: CastorDoc and Actian Zeenea supply agents, Metaphor uses Python scripts, and ThinkData Works uses data virtualization. Actian Zeenea offers data catalog federation where each business domain can have its own metamodel that synchronizes with an enterprise metamodel.

Integration

Most products offer 50+ connectors to a range of applications, including cloud data platforms, cloud applications, BI tools, data observability tools, and office productivity applications. In addition, most can pull data from on-premises systems: CastorDoc and Actian Zeenea supply agents, Metaphor uses Python scripts, and ThinkData Works uses data virtualization. Most tools display data quality scores and flags derived from integrations with data observability platforms, such as Monte Carlo, Soda, Metaplane, and BigEye. Uniquely, ThinkData Works has its own data quality detection technology. Some vendors offer two-way synchronization of tags, documentation, and access controls with dbt and/or Snowflake

GUI and Lineage

Given their focus on self-service and data discovery for regular business users, all vendors profiled in this report prize simplicity and ease of use. Actian Zeenea is perhaps the most relentless in its pursuit of a clean user interface; it regularly deletes functionality, such as threaded discussions, when it sees low user adoption. Metaphor, CastorDoc, and other products offer attractive, personalized home pages and a social experience which includes the ability to follow, like, rate, and comment on data entries. Most have reinvented data lineage and impact analysis displays, especially Metaphor, Select Star, and Castor Doc. Their goal is to make data lineage easier for non-technical people without compromising granularity and a comprehensive view from source to target, including BI tools.

Embedding

To foster adoption and ease of use, most vendors have embedded their catalogs in other applications, including Slack and Microsoft Teams. CastorDoc and Metaphor also support wikis, such as Confluence, and other applications, such as Sharepoint. These deep integrations enable users to search the catalog, submit requests, contact a steward, like an asset, and link to related terms and lineage displays. Most products, including Metaphor, CastorDoc, and Select Star, also integrate with Google Chrome, allowing users to see metadata related to the content displayed within their browser.

GenAI

Select Star, Castor Doc, and Metaphor are aggressively leveraging generative AI capabilities to improve ease of use and make their data catalogs more accessible to average users. They have adopted genAI to support natural language searches of data catalog contents and return natural text answers. They also use genAI to create labels

and textual descriptions of data assets and summarize conversations. Some are using AI to create glossary definitions, although others don't think this functionality is ready for prime time. Some are working on co-pilots to help users create SQL and Python code when working in tools integrated with the data catalog.

Our Methodology

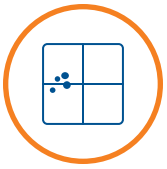
OUR METHODOLOGY

Qualifications

In the fall of 2023, we reached out to more than a dozen vendors that we felt represented this category. Some didn't respond to our request (Stemma, Amundsen, Infogix, Dataedo, Secoda), some declined (DataGalaxy), others had grown too pricey (Atlan, data.world), some were open source (Apache Atlas, OpenMetadata), some we didn't learn about until afterwards (Gravitino), and some were platforms that bundled a data catalog along with other functionality, such as data quality and data observability (Ataccama, OvalEdge).

Methodology

Rather than sending out a worksheet to vendors and getting "Yes/No" responses to questions, we scheduled multiple briefings with each vendor to drill into the details behind our evaluation criteria. We captured the details in note form, which we subsequently validated and scrubbed for legibility. This approach required us to adapt our taxonomy of criteria as we learned about new features and attributes we didn't think about initially. Since our research spanned nine months, vendors have likely added or changed features since we documented their capabilities. So please validate this research with your own.



Quadrant Chart

The report also contains a quadrant chart that plots each vendor on two axes, "company scale" and "functionality", with lower left area representing emerging players and upper right area representing established players. The five vendors are clustered in the middle of the chart, offering strong value for the functionality (green), a focus on data discovery (shape), with growing customer bases (size).

Scoring

To create the quadrant chart, we scored the product worksheets on two dimensions: company scale and functionality. For company scale, we used normalized counts. That is, for each section in the company scale section, we normalized the counts for each item. For example, we divided annual recurring revenues by $1 million. A vendor with

$980,000 in annual recurring revenues (ARR) received a score of .98, while a vendor with $5 million in ARR received a score of 5. For customer count, we divided by 10 to generate a normalized score. A vendor with 120 customers received a score of 12 and a vendor with 12 customers received a score of 1.2.

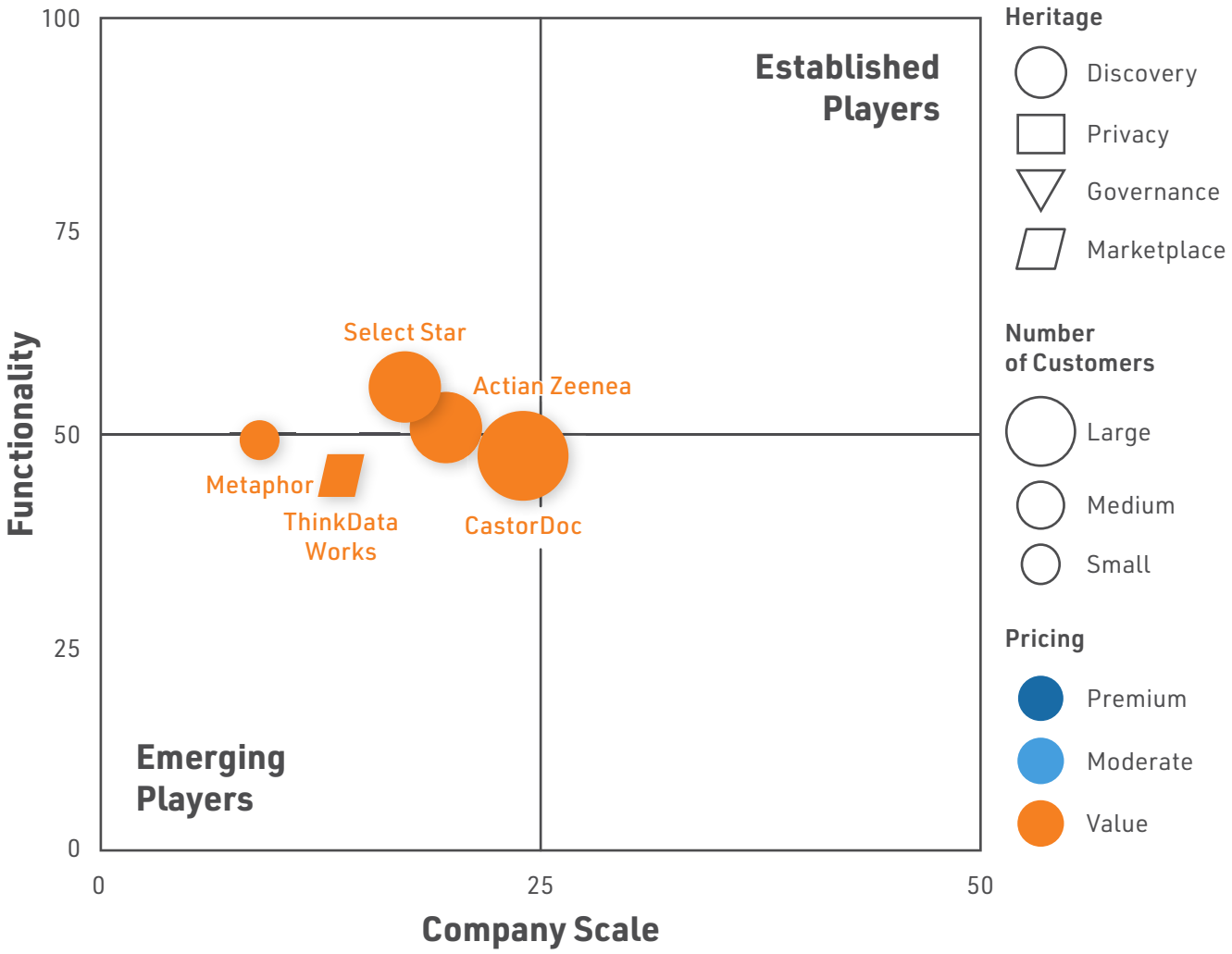
We judged functionality subjectively on a 4-point scale with 4= Excellent, 3 = Good, 2 = Average, and 1= Poor or non-existent. To generate scores, we compared written descriptions of vendor functionality in each of the 19 categories in the worksheet. In some cases, we reviewed vendor videos to validate functionality before entering a score. We sent the worksheets (but not scores) to vendors to validate our inputs.

Companion Worksheet

The heart of the report is the companion worksheet that we compiled for each product. Each worksheet contains our synthesized notes for 40+ evaluation criteria divided into four categories: 1) company background 2) company scale 3) product background and 4) product functionality. We scored the items in the second and fourth categories to create the quadrant chart described above. Feel free to add your own notes to the worksheet and score the products as you see fit.

Vendor Profiles

VENDOR COMPARISON CHART



ACTIAN ZEENEA

FOUNDED
2017

€15M
FUNDING

Focus

Actian Zeenea was founded in 2017 with the mission to facilitate data democratization, data access, and data stewardship through metadata management. Based in France, Actian Zeenea has 60+ customers today, based largely in Europe, but the company is aggressively moving to North America, targeting large companies ($5 billion+) that want to replace incumbent data catalogs, such as Collibra, Informatica, and IBM. The company's current customers are Microsoft shops that span multiple industries include giants, such as Maersk, Renault, Stellantis, ACC Automotive, Groupe BPCE, and Lufthansa Cargo. Its largest installation indexes 2+ million assets sourced from 20+ systems and supports almost 10,000 users, according to Actian Zeenea.

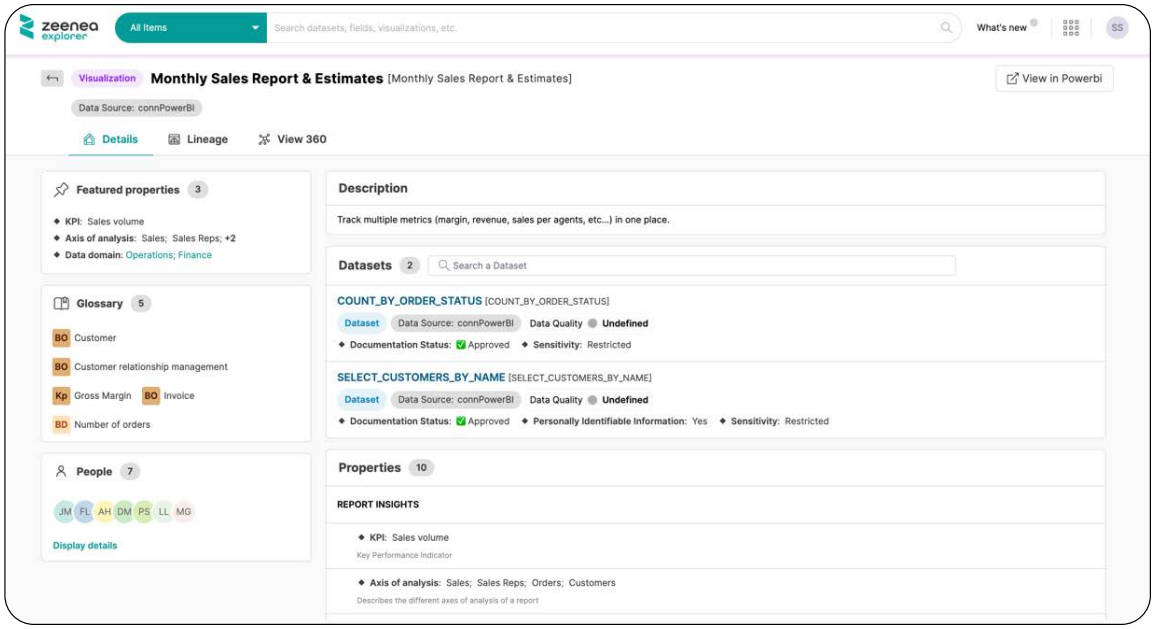
Shortly after this report published, HCLTech acquired Zeenea in September, merging it with its Actian business division which offers holistic data solutions to companies worldwide.

Actian Zeenea's guiding principle for product development is simplicity. Its goal is to provide a sophisticated product that is simple and easy to use. The company recently redesigned its user GUI, user home pages, and data steward dashboard to improve usability. Counter to software orthodoxy, it deleted functions, such as threaded discussions, because customers weren't using the feature, company officials said. It also refuses to jump on the industry bandwagon and implement "hot" new features, such as Gen AI, until they're proven to deliver customer value.

Differentiators

The Actian Zeenea Data Discovery Platform is more than just a data catalog: it bundles at no cost a data marketplace for publishing and consuming data products, putting Actian Zeenea on the cutting-edge of all data catalogs. Actian Zeenea offers 70+ proprietary data connectors that capture and manage metadata to support discovery, lineage, and classification. The data catalog comes with two user interfaces: Actian Zeenea Explorer for business users to discover data and Actian Zeenea Studio for data stewards to govern data.

Zeenea Explorer – Asset View



Actian Zeenea Explorer displays asset definitions, contacts, curators, descriptions, properties, lineage, data model, data sets, data fields, visualizations, and suggests related items. (See below.) With Actian Zeenea Studio, data stewards can create rules and alerts and define custom properties, item types and roles and add them to the metamodel. Actian Zeenea Studio also contains a portal or dashboard that provides everything stewards need to govern data: it displays watch lists, action items, and completion levels and provides reports on the health and performance of the catalog.

Challenges

As a young company, Actian Zeenea doesn't yet offer all the bells & whistles for incumbent competitors. For instance, it doesn't support policy management or built-in workflows, relying on third party tools instead. The product only supports one classifier out-of-the-box (personal information), but customers can create their own classifiers and use its unique "fingerprinting" function that turns data into vectors, making classification more accurate. Although Actian Zeenea has not embraced GenAI like some of its competitors, it does use it to summarize long descriptions. In addition, it has not embraced social features to enhance the user experience and it does not yet integrate with Slack or Chrome.

CASTORDOC

FOUNDED
2020

$25.5M
RAISED

Focus

Founded in 2020, CastorDoc aims to simplify data discovery for business users and data scientists. Among the vendors profiled in this report, it has raised the most money ($25.5 million) and has the most customers, revenues, partners, and employees. The company was founded in France but is headquartered in Brooklyn, NY. The company targets mid-market as well as enterprise companies where employees struggle to find, access, and trust data to do their jobs. Recently, customers have begun turning to the platform to support data governance. Its marquee customers include Deliveroo, Veolia, Panasonic, Freeport-McMoran, Taylormade, and the ACLU. Among vendors profiled in this report, it has one of the largest customer implementations with 1+ million assets and 1,000+ users.

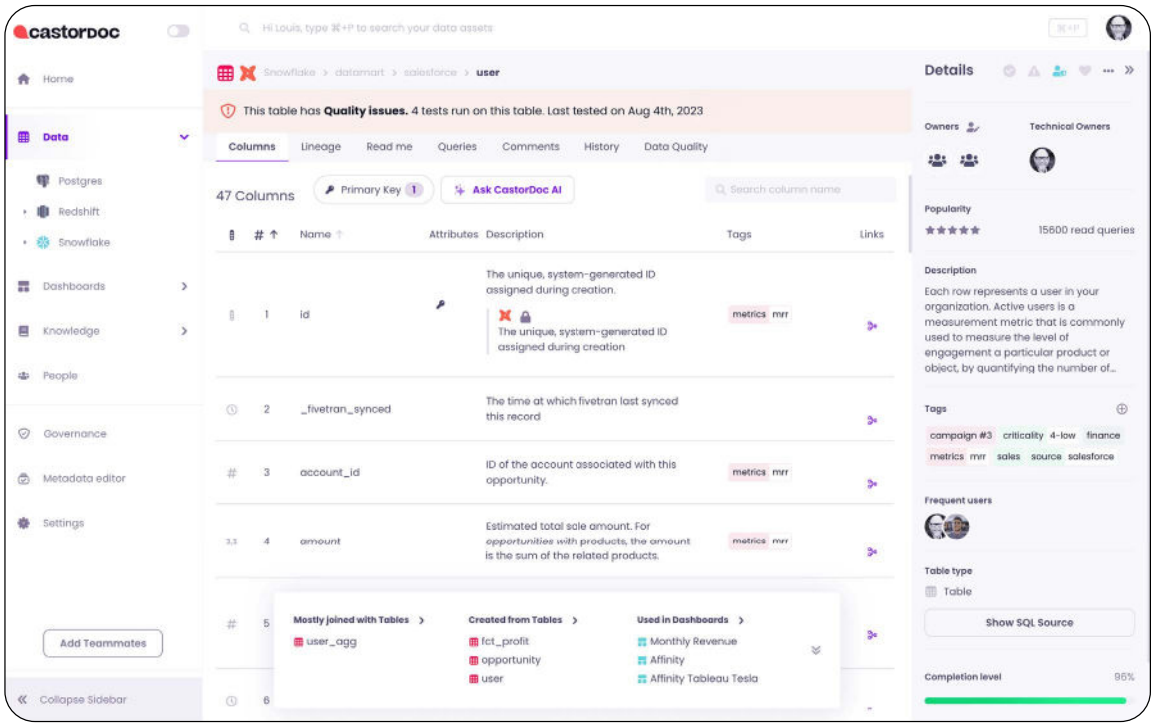
Differentiators

CastorDoc offers a clean user interface designed to provide business users with all the information they need to find and evaluate data assets for analytics purposes. For any asset, users can view table and field descriptions, tags, frequent users, popularity scores, lineage, related queries, data quality scores, flags, and alerts, comments by other users, and dashboards where the asset is used. (See below.) CastorDoc offers 10+ out-of-the-box classifiers, including those for personal information, software provenance, and deprecated, popular, and new assets. Uses can create their own classifiers.

CastorDoc runs on both AWS and Google Cloud platform. It has connectors to most of the popular cloud data platforms, BI tools, and data quality tools, and it can reach into on-premises systems via an agent architecture. It can push data into Snowflake and dbt and is working on similar pushback connectors to Looker, Tableau, and BigQuery. Customers can embed CastorDoc inside Chrome, Microsoft Teams, Slack, Notion, and Confluence. These connectors enable table-, column- and software-based lineage and tagging to related assets. CastorDoc has above average capabilities for automation, business glossary, data lineage, and customizable data assets and attributes.

CastorDoc also has an AI-based assistant that enables users to search for data assets using a conversational interface. It also uses AI to generate business glossary terms,

CastorDoc Data Catalog



create table and column descriptions, and generate SQL queries in response to text prompts. The company also offers a Dashboard Q&A that lets users ask questions in natural language about a dashboard they are viewing.

Challenges

CastorDoc has recently added features to support data stewards and curators, including the ability to bulk edit fields, view dozens of usage reports, and create tags and glossary terms. But it still lacks governance workflows and policy management. In addition, to create custom reports, customers must download data in CSV or XLS or ask CastorDoc professional services to create a custom report for them. Finally, CastorDoc does not currently offer an internal data marketplace for publishing and consuming data products or a data provisioning system for users to request access to data. Metaphor

METAPHOR

FOUNDED
2020

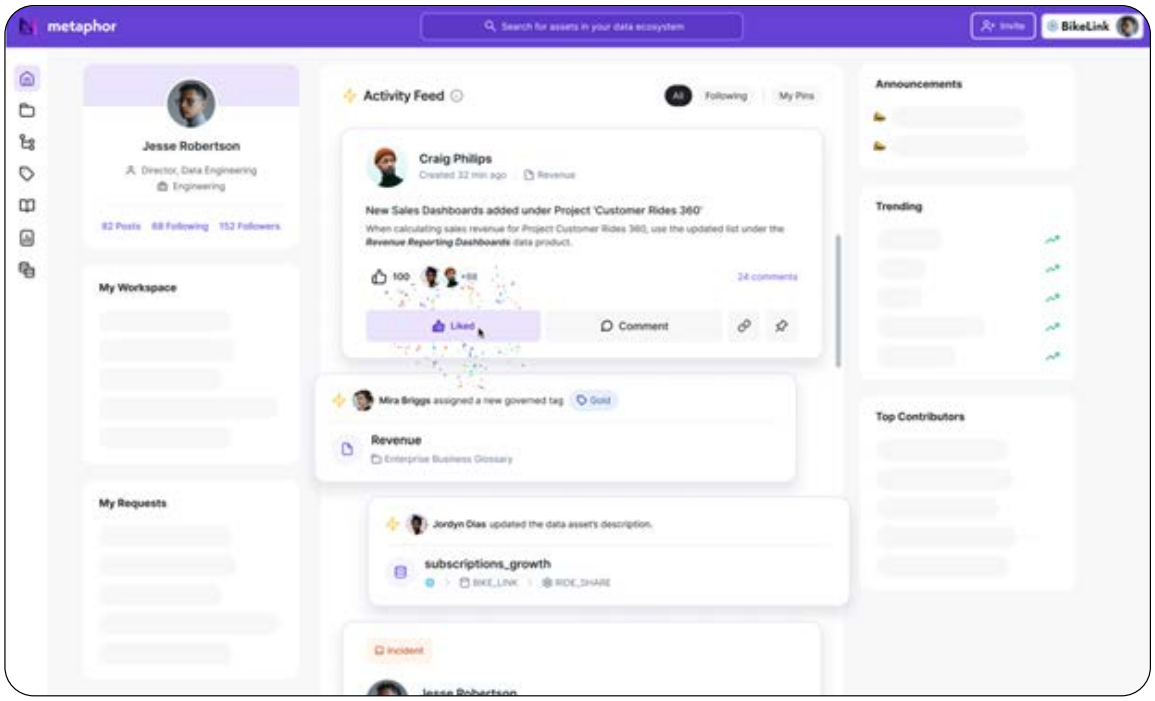
$10.3M
RAISED

Focus

Metaphor is a social platform for data with a user-friendly interface that is designed to ensure that everyone in an organization can get value from a data catalog with little to no training. Metaphor also believes that the best data catalog is no catalog at all – that is, users can use their native applications, such as Slack and Teams, to ask questions about data, submit requests, and view relevant metadata, including lineage and incidents, among other things.

Founded in 2020, Metaphor is one of the smallest data catalog vendors we profiled in terms of revenues, customers, and employees, but it has some very large brand name customers, including Wolf Subzero Cove, AAA Life Insurance, First American, Monday.com, and Fiverr. Its largest installation supports 5,000+ users, 100,000+ assets, and integrations with numerous sources. Metaphor has raised $10.3 million in two seed rounds in 2020 and 2023.

Custom Home Page for a Metaphor User



Differentiators

Besides its embedded capabilities, Metaphor creates a social experience in its SaaS-based interface, fostering collaboration by allowing users to like, follow, and comment on assets. Its homepage offers a social activity feed, custom workspaces with links to a user's recently viewed assets and requests, and sections for announcements, trending assets, and top contributors. The goal is to offer a simple, personalized GUI that hides complexity until it's needed. (See image below.)

Underneath the covers, Metaphor uses a knowledge graph to relate assets, people, queries, comments, and other attributes, which powers its search and lineage functions. It recently added GenAI capabilities to support natural language search, generate glossary definitions, automate documentation, and summarize conversations. It offers end-to-end, column-level lineage from popular cloud data platforms, such as Snowflake and Databricks, to leading BI tools, such as PowerBI and Thoughtspot.

Challenges

Although Metaphor doesn't support a formal data marketplace, authorized users can publish data products to a domain page that users can access. It also doesn't classify metadata or data, but can import classifications from third-party applications, such as Snowflake, which it exposes as tags. Designed to support discovery and self-service, Metaphor, like other upstart catalog vendors, is moving into data governance. Although it does not have a data governance portal, it offers packaged data governance reports and customers can export data via its SQL data warehouse and build their own reports.

SELECT STAR

FOUNDED
2020

$20M
FUNDING

Focus

Select Star was founded in 2020 as a SaaS-based data discovery and governance platform. The product is easy to use, fast to deploy, embeds within office productivity tools, and is half the cost of incumbent data catalogs, making it a big hit among mid-market firms. But larger enterprises with 10,000+ employees, such as Fivetran, Pitney Bowes, and nib, an Australian health insurance company, have also discovered Select Star and validated its scalability. Today, Select Star's largest customer installations support 2,000+ users and 1+ million assets.

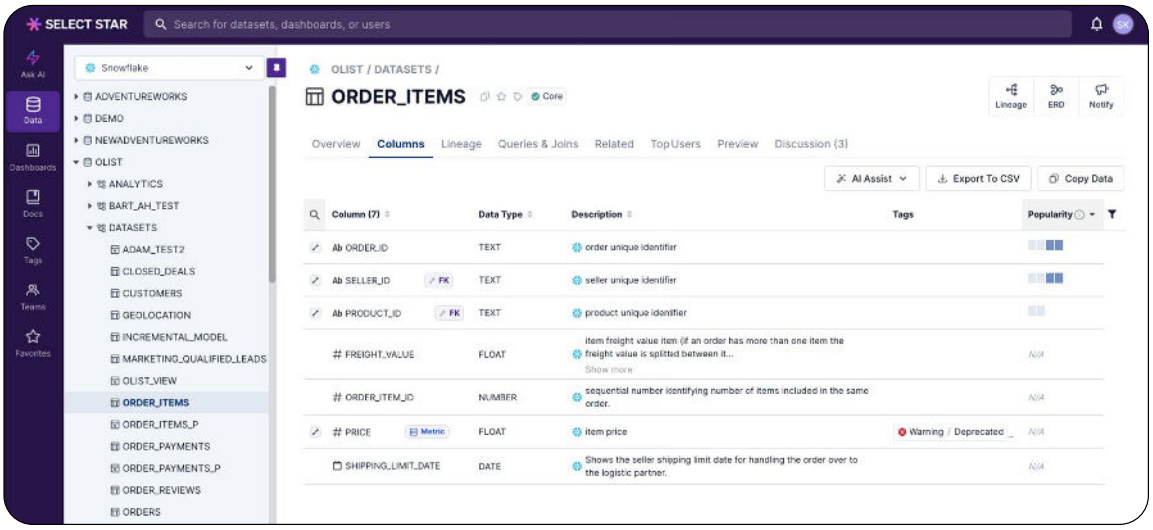
Select Star has received $20 million in funding, boasts 50 customers. The product is ideal for power users and data developers who need to find data spread across multiple cloud applications and environments. It's also designed to help data stewards and data curators manage data and comply with privacy and security regulations. Select Star has simplified GUI that puts relevant information for each type of user on a single pane. (See image below.)

Differentiators

For a startup, Select Star packs a lot of advanced functionality. Besides extracting and loading metadata automatically, Select Star analyzes query logs to compute column-level lineage, which it displays as tiles, graphs or hierarchies, making it easy for different types of users to understand the origins of data and its impact on downstream applications. It also mines query logs to track usage, calculate asset popularity, identify recent queries, suggest tags and descriptions, and create entity-relationship diagrams that make it easier for users to form SQL queries and joins.

The product also has a popular AI assistant based on OpenAI's ChatGPT that is now embedded in Slack so users can ask questions about data and generate SQL without logging into a data catalog. The product can also detect source system changes and notify users of downstream impacts. From a compliance perspective, when Select Star ingests data, it auto-detects, tags, and masks personal information, such as name, address, and social security number, automating discovery and governance tasks.

Select Star Data Consumption Viewer



Select Star also has a Chrome extension, displaying metadata for objects displayed within a browser. It also supports two-way synchronization of tags and documentation with dbt and Snowflake, and it imports data quality metrics and tests from Monte Carlo and dbt. Finally, it tracks Snowflake consumption via a dashboard that enables customers to better manage data processing costs.

Challenges

Select Star primarily connects to cloud sources and has fewer metadata connectors than some data catalog products profiled in this report. Although it has a business glossary, it doesn't auto-generate terms based on indexed data or provide a workflow for data stewards and owners to identify, prioritize, and define terms. Although Select Star embeds in Slack and Chrome, it does not yet support Microsoft Teams. It also does not have a policy engine although it does support policy-based access control. Although Select Star support nine out-of-the box reports (e.g., Active Users, Top Contributors), it does not support custom ad hoc reporting. Finally, Select Star indexes metadata but does not provide direct access to data via either a provisioning mechanism or a data marketplace.

THINKDATA WORKS

FOUNDED
2014
$8.2M
FUNDING

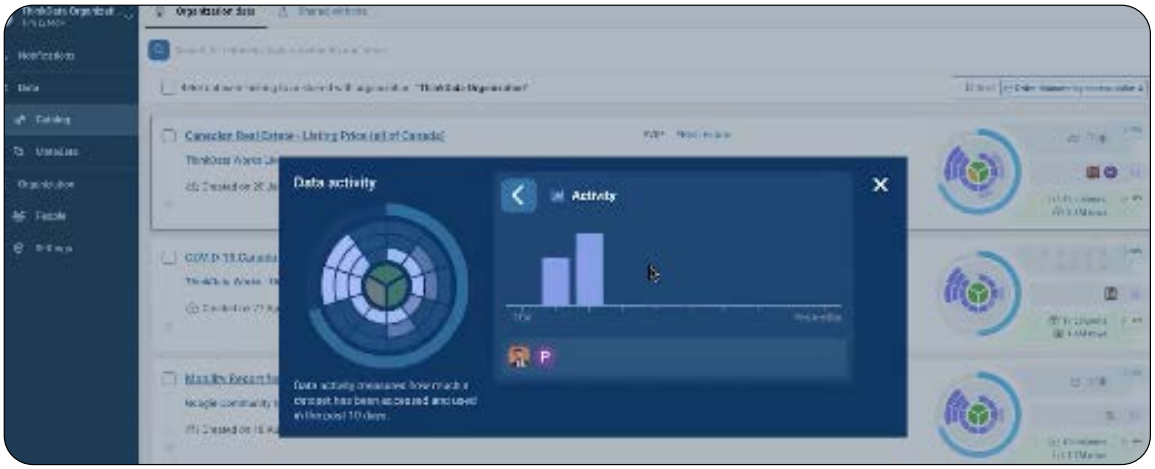
Focus

ThinkData Works was founded in 2014 to help organizations address the challenges of managing, accessing, and deriving value from scattered data sources. Their original product was designed as a data marketplace for open data and used a purpose-built ETL to connect to thousands of open data sets. In 2020, the company pivoted its strategy to focus its technology on data democratization, metadata management, and data migration. Although the company is one of the smaller vendors profiled in this report, it has two patents and some brand name clients, including RBC, Scotiabank, and Roche Pharmaceuticals.

Differentiators

Unique among data catalog vendors, ThinkData Works has data virtualization technology that gives users direct access to data, including on-premises and legacy systems or cloud data platforms and applications. Like most data catalogs, the ThinkData Works Platform harvests and indexes metadata from a variety of sources, displaying descriptive, administrative, and structural metadata, including provenance, schema, tags, etc.
But unlike its competitors, ThinkData Works enables customers to query data they've discovered in the data catalog, no matter where it's located, and share these views as data products. This ability to query and package data assets turns ThinkData Works from a data catalog to an internal data marketplace.

ThinkDataWorks Data Quality Monitoring Graph



ThinkData Works is also one of the few data catalog vendors that does its own data health monitoring, saving customers the cost of leveraging a third-party data observability tool. The product runs statistical analysis on datasets at user-configured intervals and adds scores to the catalog page to enable better data quality controls and analysis. The product displays flags and notifies users about data inconsistencies that may affect downstream processes. (See below.) ThinkData Works supports column-level data lineage for data products created in the platform.

Thanks to its data virtualization technology, ThinkData Works can manage governance rules for an entire data ecosystem. Data catalog administrators can set access roles and propagate them to downstream systems. Once users have been granted access to a system via the data catalog, they can access that system directly without going through the data catalog. This saves a lot of churn and wasted time.

So, a catalog administrator can grant a user access to a database, a DBA could grant me access using the catalog, but I don't need to access the data from the catalog in order to use it. I can connect to PowerBI (or any other analytics tool) and pull whatever data I have access to. The Catalog is the governance middleware that simplifies downstream access. In a system WITHOUT virtualized governance, you'd have to create governance rules at the data source level (ie give analysts access to the raw data in the DB) or set access rules in each downstream tool, which becomes difficult to maintain.

Challenges

Given its heritage and technology foundation, ThinkData Works is not a typical data catalog. Given its repositioning four years ago, the vendor is now gaining a foothold in the metadata management market. Its use of data virtualization gives its data catalog a unique spin by giving users direct access to data and the ability to publish data products. However, not all companies want both capabilities in a single toolset. In addition, ThinkData Works does not integrate tightly with some cloud platforms, which limits its ability to extract ETL metadata needed to build lineage views. Additionally, ThinkData Works does not have a policy management system and lacks full support for active metadata.

Conclusion

CONCLUSION: A NEW BREED

If your organization or department is looking for a data catalog or to replace an existing one, consider the data catalogs profiled in this report. These products represent the next generation of data catalogs that focus on ease of use, speed of deployment, and affordability. Although they may not have all the bells and whistles of incumbent products, they are catching up quickly. Those that rise to the top will soon compete with incumbents for enterprise deployments that require data discovery, data governance, cost optimization, and data migration. If you are new to metadata management, these products will evolve with your organization's needs.

ABOUT ECKERSON GROUP

Eckerson Group is a boutique consultancy that helps organizations get more value from data. We specialize in developing actionable data strategies and tailored data solutions for commercial and non-profit organizations. Our consultants each have 25+ years of hands-on experience and specialize in data architecture, data governance, data literacy, self-service, data products, and operating models. Our clients say we are hard-working, insightful, and humble. It stems from our love of data and desire to help organizations turn data into insights and action. **Explore what Eckerson Group can do for you!**