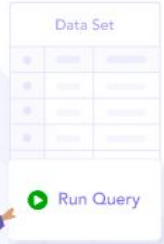
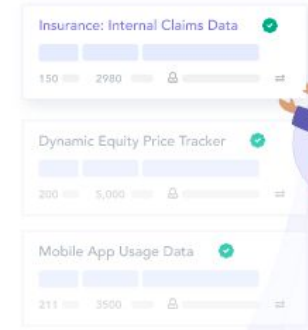
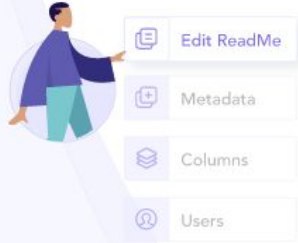


atlan

# The Ultimate Guide to Evaluating a data catalog



# The 5 steps to finding the right data catalog for your organization



## STEP 1

# Define organizational needs for a data catalog

1

Start by identifying the **TOP THREE CHALLENGES** that are causing data initiatives to fail at your organization

2

**MAP ORGANIZATIONAL NEEDS** or challenges with the core functionalities of data catalog offerings

3

Evaluate **NON-FUNCTIONAL CONSIDERATIONS** related to data catalogs and how relevant they are to the needs of your organization

# Conduct surveys and interviews across the organization to identify top reasons for failure of data projects

Data data everywhere, **just not when you need it.**



**Sam the Business Manager** 1:55 PM

Hey @ian\_IT I made a request for the data 14 days ago. Any ETA when you'll share it?

Dependencies... **Live with it.**



**Ian the IT Admin**

2:01 PM

Your request for report automation will be prioritised next quarter, and will go live in 6 months.

Murphy's Law strikes data.  
**At play, every day.**



**Derek the DevOps Engineer** 1:33 AM

The last 3 runs failed. @sam\_biz check with the client that the web service URL is the same?

Data is a **black box**



**Larry the Data Lake Owner** 7:22 PM

@Chief Data Officer I am frustrated. The business keeps demanding insane outcome and don't help at all in fixing foundational data quality. We have made so much progress but no one in business even understands.



**Dalia the Data Scientist** 5:01 PM

@alex\_analyst what does variable `column_xy881` stands for in the data set `sales_mm_blr_2919.csv`?

Human tribal knowledge;  
**siloed and often, lost.**

Human grunt work **max.**



**Sam the Business Manager** 1:55

Hey @alex\_analyst can you please rerun the excel report so that we can send it to boss?

# Sample Survey Questionnaire (Part 1)

CATEGORY	QUESTION	OPTIONS (TO BE CUSTOMIZED)
<b>ROLE &amp; USER PERSONA</b>	Which one of the following personas best fit your role at the organization?	<ul style="list-style-type: none"><li>• Data Scientist</li><li>• Data Analyst</li><li>• Data Engineer</li><li>• Business Analyst</li><li>• Business Manager</li><li>• Data Steward</li><li>• IT Support (Data Provisioning)</li><li>• Cloud Team</li><li>• Other (If yes, then what persona and how is it different from the ones listed above?)</li></ul>
	What are your top three priorities this year?	<i>[Insert text]</i>
	<ol style="list-style-type: none"><li>1. What are your top three challenges that affect productivity?</li><li>2. How many hours would you save by solving each of the three challenges every week?</li></ol>	<i>[Insert text]</i>
<b>DATA-RELATED ACTIVITIES</b>	<ol style="list-style-type: none"><li>1. Which of the following data-related activities do you perform?</li><li>2. How much time do you spend on each activity every week?</li></ol>	<ul style="list-style-type: none"><li>• Getting access to data</li><li>• Finding the right dataset</li><li>• Understanding context associated with data (e.g. understanding of column names)</li><li>• Running quality checks on datasets</li><li>• Exploring data &amp; running queries</li></ul>

## Sample Survey Questionnaire (Part 2)

CATEGORY	QUESTION	OPTIONS (TO BE CUSTOMIZED)
<b>DATA-RELATED ACTIVITIES</b>	<ol style="list-style-type: none"><li>1. Which data-related activities do you face productivity challenges with today?</li><li>2. How many hours would you save every week by solving these challenges?</li></ol>	<ul style="list-style-type: none"><li>• Getting access to data</li><li>• Finding the right dataset</li><li>• Understanding context associated with data (e.g. understanding of column names)</li><li>• Running quality checks on datasets</li><li>• Exploring data &amp; running queries</li></ul>
<b>COLLABORATION</b>	<ol style="list-style-type: none"><li>1. In a typical week, what are the different user personas that you collaborate with on data?</li><li>2. Do you face any challenges while collaborating with others on data? If yes, explain.</li><li>3. How much time would you save every week if these challenges were solved?</li></ol>	<ul style="list-style-type: none"><li>• Data Scientist</li><li>• Data Analyst</li><li>• Data Engineer</li><li>• Business Analyst</li><li>• Business Manager</li><li>• Data Steward</li><li>• IT Support (Data Provisioning)</li><li>• Cloud Team</li><li>• Other (If yes, then what persona and how is it different from the ones listed above?)</li></ul>

# Next, understand the 6 core capabilities of data catalogs



## DISCOVERY

Clear and comprehensive view of all data assets within the organization



## KNOWLEDGE

All context, information and business know-how around data assets



## TRUST

Information on data quality and coverage along with usage of data assets with the organization



## COLLABORATION

Intuitive UI for diverse team members to effectively work together on data assets



## GOVERNANCE

Manage access rights to data assets to ensure legal and regulatory compliance



## SECURITY

Ensure secure and compliant usage of data assets within the organization

# Map these capabilities to key organizational needs and define what you're looking for in a data catalog

PRIORITY	TOP ORG NEEDS/ PROBLEM STATEMENTS	PRIORITY DATA CATALOG CAPABILITY (S)
1	<b>Long waiting periods for business users</b> to access required data → leads to long lead times (>14-21 days) and productivity loss.	<b>Discovery:</b> Master search that enables users to easily find data assets using attached metadata, glossary terms, classifications, and more. <b>Governance:</b> Easy-to-use & customizable access policies for every data asset to manage data provisioning to various users.
2	<b>Lack of trust</b> of both data analysts and business users on the quality and accuracy of data assets	<b>Trust:</b> Add customizable data quality (DQ) rules to fix blank values, duplicated records, etc. and assign an auto-generated data quality score to each data asset. Set automated alerts and flags in case certain DQ checks do not succeed.
3	<b>No clear context around data assets:</b> multiple versions of same dataset exist in different locations leading to confusion and repeated back and forth	<b>Knowledge:</b> Comprehensive data dictionary (or profile) for every data asset i.e. bring together structured metadata for every data asset in an easy to consume manner. Provide ability to manually create logical views and tagging such as groups, type of data, use cases, etc.



# But it isn't just about functional capabilities – data catalogs will be the foundation for your data ecosystem. That's why the best catalogs...

1

Have an ability to integrate with other tools and technology

Standalone Catalogs that do not integrate into existing tool sets of data users often end up as *"just another tool"*

Ability to integrate into your team's favorite tools like Tableau and Excel help ensure adoption within the organization



2

Are built in a way that prevents technology lock-in

A data catalog is a foundational tool for data driven organization and will by-design get deeply embedded in your organization. As technology continues to evolve rapidly, it is important that **you do not get locked into the 2020 Oracle!**



3

## Interface optimized for business, not just IT

Gone are the days where metadata management was solely for IT teams. Non-technical, business team members are rapidly become active and regular users of data and need to be able to use the data catalog effectively. **Ease of use and intuitiveness of the UI is paramount.**

**LOW CODE ENVIRONMENT**

**DIY AUTOMATION AND ALERTS**

**VISUAL QUERY BUILDER**

**DESIGN-FIRST APPROACH**

4

## Ease of onboarding and setup

It is essential to consider the **hidden costs of engineering time and support.** Several traditional catalogs require a full time engineering team to setup, support and maintain — over and top of high cost of engaging consultants and experts.

**DIY SELF-SERVE SET UP**

**HANDS ON ONBOARDING**

**CLEAR STEPS OUTLINED**

**PERSONA-BASED USER FLOWS**

5

## Pay as you go models

The pricing model must be flexible and entirely aligned with the success of your data initiative.



### **Low initial investment:**

Products with high initial investments will require you to mobilize significant budgets — making it harder to kickstart



### **Scale costs as adoption occurs:**

The best kind of pricing models have costs mapped to internal adoption -- as more users adopt product, costs ↑

## STEP 2

# Create a customized evaluation criteria



Finalize a clear set of objective criteria that will guide the evaluation process

- ✓ Core capabilities mapped to organizational needs
- ✓ Other high priority considerations and how to evaluate them



# Sample Evaluation Criteria – Core Functionalities (Part 1)

CATEGORY	FEATURE REQUIREMENT	PRIORITY
Discovery	Master search that enables users to easily find data assets using attached metadata, glossary terms, classifications, and more.	1
Knowledge	Ability to inventory business glossary terms and link it to data assets at a table and column level.	1
	Create a data dictionary (or profile) for every data asset i.e. bring together structured metadata for every data asset in an easy to consume manner. Provide ability to manually create logical views and tagging such as groups, type of data, use cases, etc.	1
	If a data rule or structure changes, identify the impact of the change on all endpoints. Maintain versions and capture change history of data assets and context surrounding it in ReadMe and/ or its meta data.	2
Trust	Add customizable data quality rules (to fix blank values, duplicated records, etc) and assign an auto-generated data quality score to each data asset to help identify quality. Set automated alerts and flags in case certain DQ checks do not succeed.	3
Collaboration	Various users should be able to tag each other to flag discrepancies, make comments and ask questions on the data assets. How can the users work in a collaborative way? E.g. can a user make a comment about an entity or a KPI if he's not the owner?	2
	Owner of each data asset can approve recommendations made by other users. E.g. recommended edit of a classification or glossary term, etc.	2
	Save queries so that other users are able to reuse the common queries without having to rewrite the code every time	2
Governance	Customizable access policies for every data asset (table > row/ column > attribute) to manage access provided to various users.	1
	Real-time monitoring of who is accessing which data assets at any given time — provide and revoke access within a few clicks.	2
Security	Integrate with the Active Directory system for authentication and security purposes.	1
	Identify and automatically classify personal and private data (DNI, PAN, address, phone numbers, etc.	1

## Sample Evaluation Criteria – Other Considerations (Part 2)

CATEGORY	KEY CONSIDERATIONS	HOW TO EVALUATE?
Ability to integrate with every data user's favorite tools	<p>Need to test integrations with the following tools commonly used in organization:</p> <ul style="list-style-type: none"> <li>• Jupyter</li> <li>• Microsoft Excel</li> <li>• Power BI</li> <li>• Tableau</li> </ul>	<p>Level 1: Ask vendor if native integrations are available</p> <p>Level 2: Ask vendor to showcase the integrations</p> <p>Level 3: Test the integration deeply in the trial / Proof of Concept (POC)</p>
Avoid Technology Lock-In through Open API architecture	No special considerations	Understand the backend infrastructure of the tool — and its reliance on open API architecture
UI optimized for business, not just IT	Very essential to ensure that the catalog is easy for not just technical users but also the power business users.	<p>Level 1: Include business users in the evaluation process and rate easy of use (or intuitiveness) on a scale of 1 to 5</p> <p>Level 2: Include business users in the trial / POC to observe how easily they use the catalog without needing support</p>
Onboarding effort and time: what does it really take to get started?	Evaluate how many man hours will you need to invest in getting the data catalog set up on your environment.	<p>Level 1: Ask for detailed documentation around set up process — include technical requirements and time/ involvement required of different user personas</p> <p>Level 2: Evaluate how many ad-hoc requests came in (beyond scope) and how much additional time was spent</p>
Tenets of product pricing should support success	Evaluate how flexible and scalable the pricing model is — for both the current and future usage scenario	Under the pricing model — upfront cost, cost of running a full POC on your data environment, scaled-up costs, additional licenses or infrastructure costs

## STEP 3

# Understand the data catalog vendors and offerings in the market

There are broadly three types of Data Catalog offerings in the market

1

Traditional Data Catalogs



Collibra

\$\$\$



Alation

\$\$\$

2

Open Source Data Catalogs

**amundsen**

Open Source

3

Modern Data Catalogs

**atlan**

\$

## Traditional Data Catalogs



Heavily optimized to be deployed on  
**On-Premise Storage**

**Legacy organizations** — have a smooth and seasoned implementation process

**Limitations around data movement and scalability:** data in motion (streaming) is not supported very well

Very **limited support for cloud-based storage/ warehouses**

Built with the technical user in mind, optimized for IT, **challenging for business users**

Heavily optimized to be deployed on  
**on-premise storage**

**High effort and time investment** in maintenance and support — requires several full time engineering resources

## Open Source Catalogs



**Cloud native** but has varying documentation across cloud providers

**Open source project** — free of any license cost

**Feature light but UI is simple to use and intuitive.** Easy to understand for various users

With **limited features currently** — features like data preview and SQL query is in beta.

**Nascent open source project:** few contributors.

**Difficult to set-up** — for data ingestion need to create airflow dags or run python scripts. Templates available. an entire data engineering team required to help with the initial set-up and manage infrastructure and daily issues.

## Modern Data Catalogs



Cloud Native to all 3 cloud providers.

Feature rich with an easy to use and intuitive UI — **easy for business users** to use

No engineering time or resources needed to maintain. **Completely DIY.**

**Easy Set-Up** — Jumpstart to full functionality on day 1

**Modular and scalable pricing** — no separate set up cost, pricing scales with adoption

**Newer in the market** compared to legacy data catalogs like Alation and Collibra

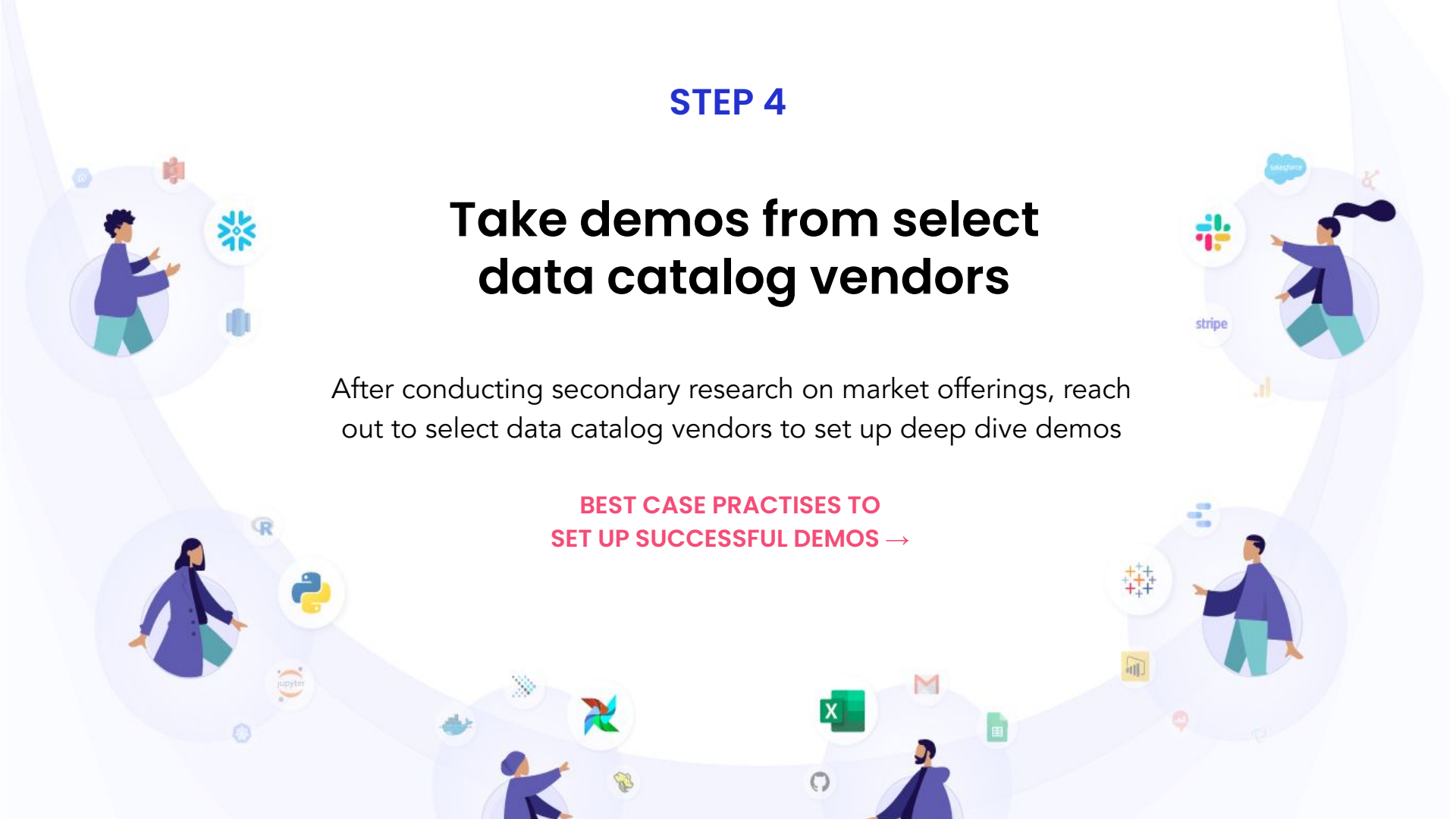
**Not optimized for on-premise** data sources — very limited support

## STEP 4

# Take demos from select data catalog vendors

After conducting secondary research on market offerings, reach out to select data catalog vendors to set up deep dive demos

**BEST CASE PRACTISES TO  
SET UP SUCCESSFUL DEMOS →**





# Best case practices to set up successful data catalog demos

1

## Share finalized evaluation criteria document with vendors in advance

- ✓ Help vendors to understand your problem statements and organizational context upfront
- ✓ Serves as a great alignment document to make sure all questions and priorities are discussed during the demo

2

## Ensure stakeholders from different teams attend the product demos

- ✓ Bring together different user personas for the demo to gather the most comprehensive feedback possible

3

## Conduct a data architecture compatibility check

- ✓ It is critical that the data catalog is compatible with both the current data architecture of the organization — as well as the vision and roadmap for the next 1 - 3 years

## STEP 5

# Execute hands-on proof of concept (POC)

After taking demos, reach out to selected data catalog vendors to set up hands on POC

**BEST CASE PRACTISES TO  
SET UP SUCCESSFUL POC →**



## L1: PRE-POC ACTIONS

### ✓ CONDUCT A DESIGN SPRINT

Identify top use cases and user flows to test during the POC

Ensure that the key capabilities in the evaluation criteria are included

### ✓ SET UP POC ORG STRUCTURE

Prepare ideal tech architecture: sources, connectors, etc.

Properly onboard business users that will be involved in the POC execution

## L2: DURING POC ACTIONS

### ✓ ENSURE DEDICATED TIME SPENT

>80% of the time of the core team should be spent on execution

Capture detailed feedback at each step of POC

### ✓ STAGGERED IMPLEMENTATION

Do not aim to conduct POCs with more than 2 vendors

Do not run multiple POCs simultaneously

## L3: POST POC ACTIONS

### ✓ DEEP DIVE FEEDBACK

Set up calls with the vendor team to share detailed feedback on the POC experience and gauge response to feedback

### ✓ QUICK SPRINT TO DECISION

Bring together the core team and other decision makers to make a quick decision on the data catalog to prevent lag and leverage current momentum

## Create a foundation for a successful POC



High touch communications through a channel on Slack or Microsoft Teams



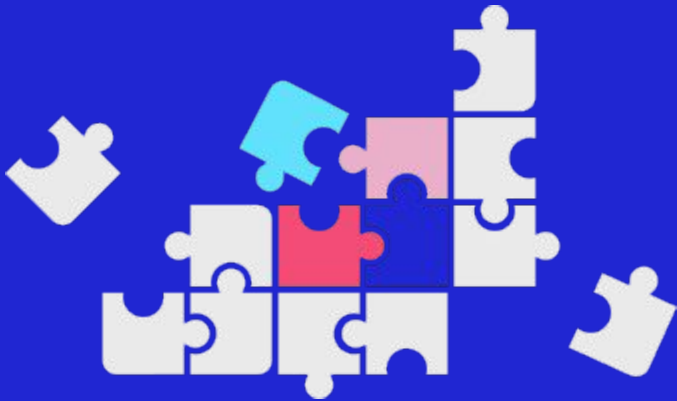
Adopt the partner mindset: create continuous feedback loops with vendor team



Success depends on the people as much as on technology — manage POC team well

Finally —

# Choose a partner, not a vendor



How do they respond to feedback? Are they listening or only trying to sell?

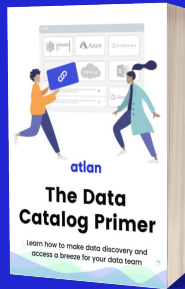
Are they upfront and honest?

Do they share their limitations with you while acknowledging positives in other competitors?

THANK YOU FOR READING THE ULTIMATE GUIDE TO

# Evaluating a Data Catalog

Check out our other resources for data teams



EBOOK

**The Data Catalog Primer:**  
Everything you need to  
know to setup one

[Download →](#)



EBOOK

**The Ultimate Guide to  
Building a Business Case  
for a Data Catalog**

[Download →](#)

Compiled with ❤️ by Atlan

# Explore **atlan**

The first data catalog built for the future



atlan | Hope you are having a good day, Adam! | What's New

Catalog

+ NEW TABLE

Search through your tables...

Filters

Table Owners

Search...

Sarah

Sandra

James

Darwin

Geeta

Status

Verified

Motor Insurance: Internal Claims Data

Motor insurance claims data for fraud analytics and customer exp

Fraud Analytics Customer Experience

98,088 rows 39 columns Ujwal ++ 9,576 rows 9 hours

AAA Rated Insured

This is a dataset of highest rated customers of the company

customer ratings Health

21,600 rows 39 columns Ujwal ++ 6,048 rows 9 hours

Outlet sales raw 2018


Click to add description.


Bootcamp

27,751,391 rows 7 columns Ankita 27,751,391 rows

Outlet sales raw 2019

Click to add description.

 **Cloud-native**  
data catalog

 **24 hours**  
to get up and running

 **Democratization**  
for business

 **Governance**  
for IT

 [Watch Guided Demo](#)

 [Take Guided Tour](#)



Trusted by data teams around the world



MINDSHARE



INMOBI™



Mahindra



OLA



milkbasket



Kotak  
Kotak Mahindra Bank



UNITED NATIONS

We are proud to be supported by



Rajan Anandan

Former MD Google India



WATERBRIDGE  
VENTURES



Ratan Tata

Chairman Emeritus Tata Sons



Manoj Menon

Partner & MD Frost & Sullivan (APAC)