How to make your data drive value, not risks?

VITALIS KAVALIAUSKAS, CTO

April 27th, 2022

baltic amadeus

Facts:

30+
Years in business

100+
Customers

200+
Employees

Technology partner in:

- ✓ Consultancy
- ✓ Ominichannel
- ✓ Data & Analytics
- ✓ Cloud

Banking & finance

Energy

Telecommunications

Healthcare & Pharmacy

Insurance

IT & consulting

Agenda for today

I. Why do we need data governance? (5 min.)

II. How to choose the best data governance model? (10 min.)

III. Data governance implementation: challenges and required capabilities (35 min.)

I. Why do we need data governance?

baltic amadeus

Data-driven strategy is now vital to succeed

With the power of data, you can effectively support decisions such as:

- revenue growth
- profitability
- customer satisfaction
- organisations

Three main challenges we need to solve

Huge volumes of data coming from everywhere

Speed

Trust

baltic amadeus

Why should we modernise our approach to data?

The story of one book

baltic amadeus

What if we could:

- ✓ organise data at scale,
- ✓ extract hidden data value,
- ✓ deliver data everyone can trust?

The essence of data governance

Collection of processes

Roles

Policies and standards

Metrics

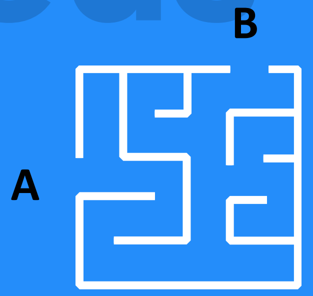
- ✓ maximise data’s value,
- ✓ manage its risks,
- ✓ reduce the cost of data management.

Compliance is important but not the only driver for data governance

Regulatory compliance needs to incorporate **multiple controls** and **foster accountability** for data protection.

It should be verifiable in practice, not just defined by legal guidelines written on paper.

Compliance **on paper**



Compliance **in practise**

Data governance provides multiple benefits

Common understanding

Higher data quality

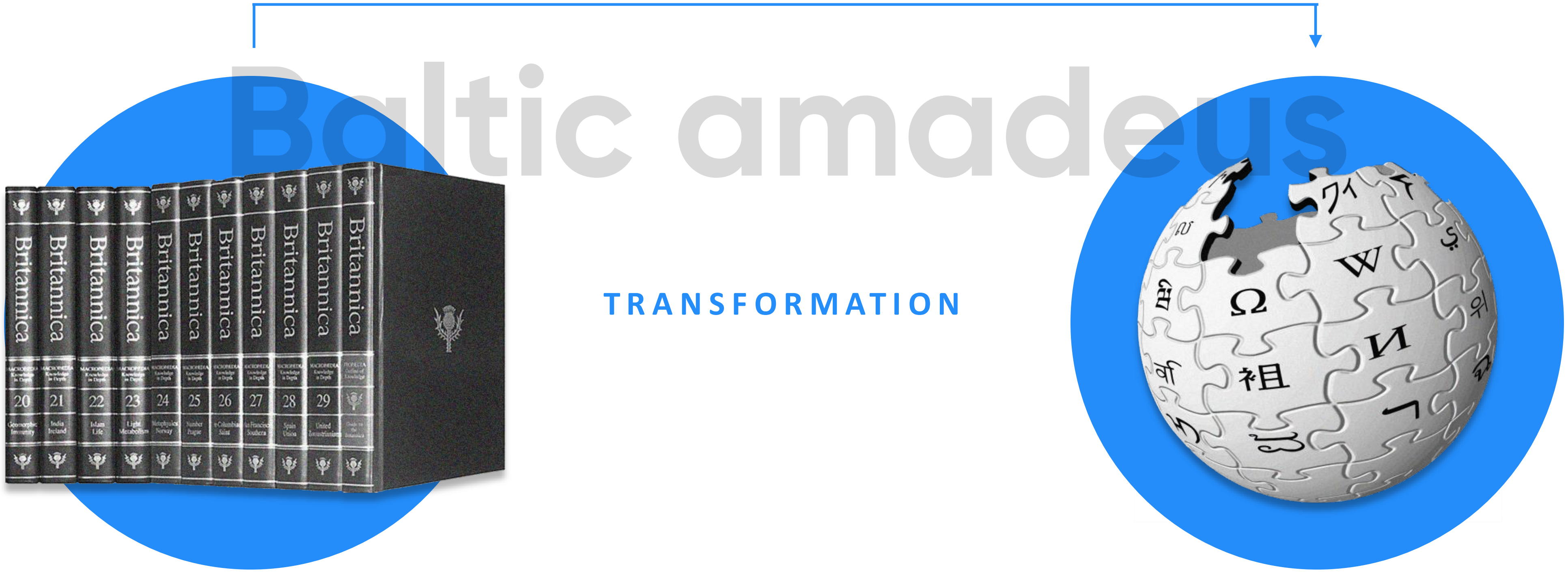
360-degree views

Improved data management

Easy access

II. How to choose the best data governance model?

Transformation of „knowledge industry“



TRANSFORMATION

Why do we need to review and modernise data governance models?

Multiplication of new data-driven roles

IT's budget and resources are relatively flat

Data comes from everywhere

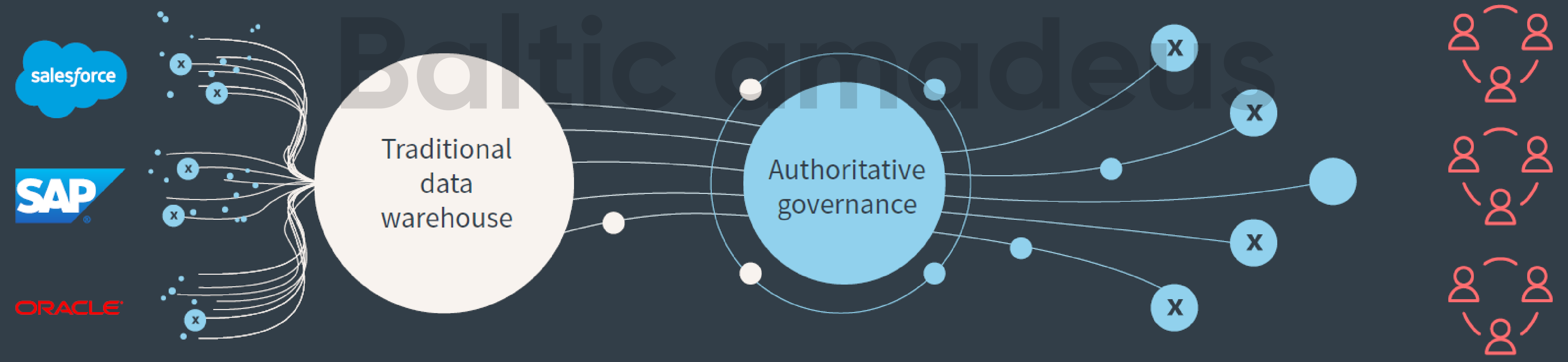
Established "governance with the NO"

Data is needed faster than ever

ECONOMICS OF DATA IS BROKEN

baltic amadeus

The traditional model



Restricted data access

Limited user reach

Pros

- ✓ Quality can be excellent with this model.
- ✓ Defined by central model to collect and reconcile data.
- ✓ Relies on a team of data professionals armed with well-defined methodologies and practices.

Cons

- ✓ Requires a lot of effort to bring data accurately and quickly as consumers want.
- ✓ Do not address the growing needs for new and various data types.
- ✓ People look for other ways, such as shadow IT to meet their data needs.

Organisations that cannot evolve from this model lose:

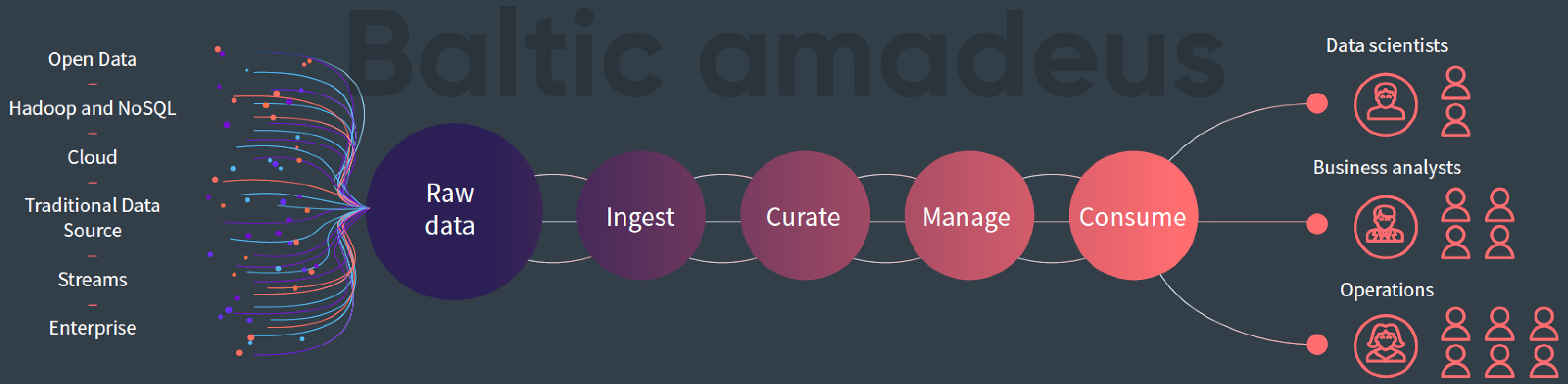
control

accuracy

speed

security

The data lake model



Any data

Costs, time to value, scalability, governance, risks

Any data worker

Pros

- ✓ Raw data can be ingested with minimal upfront implementation costs.
- ✓ Cloud infrastructure/services can drastically accelerate the data ingestion process.
- ✓ The model is more agile – it scales across data sources, use cases, and audiences.

Cons

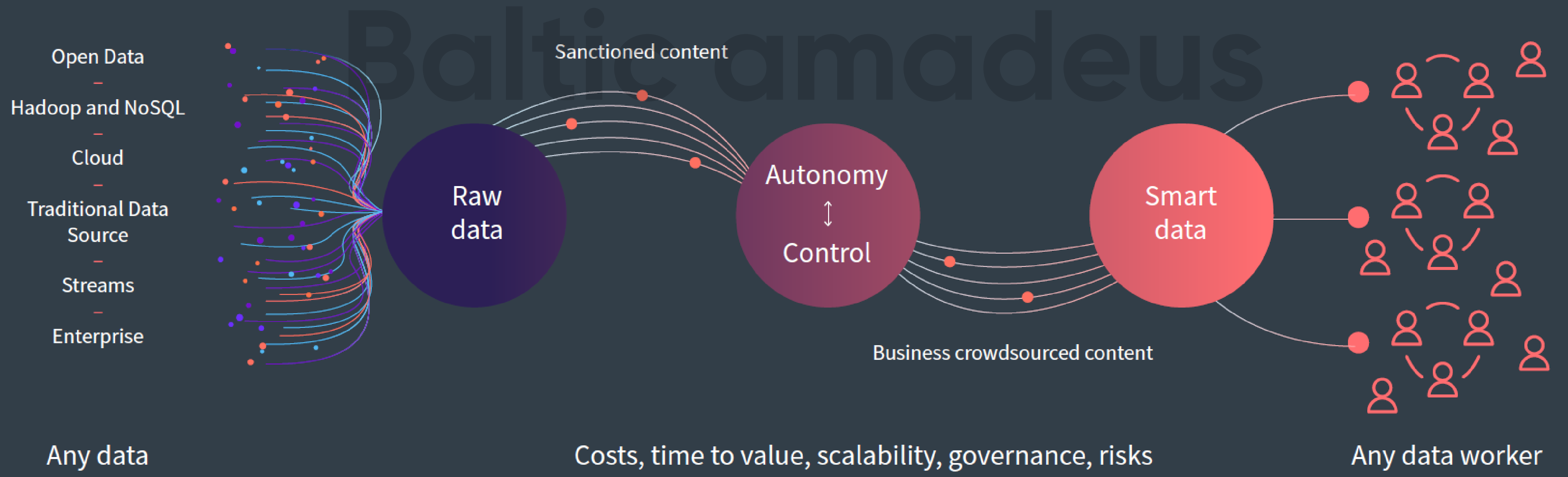
- ✓ Only the most data-savvy people can access raw data, while others still require structured data.
- ✓ Need to establish stronger control if you target a wider audience.

Forget the governance in your data lake, and it will become a data swamp

DATA **LAKE**

DATA **SWAMP**

The model of collaborative governance



Open Data
–
Hadoop and NoSQL
–
Cloud
–
Traditional Data Source
–
Streams
–
Enterprise

Sanctioned content

Raw data

Autonomy ↕ Control

Smart data

Business crowdsourced content

Any data

Costs, time to value, scalability, governance, risks

Any data worker

Scale trust and reach through collaboration

1200

admins

300 000

editors

55 000 000

articles

Pros

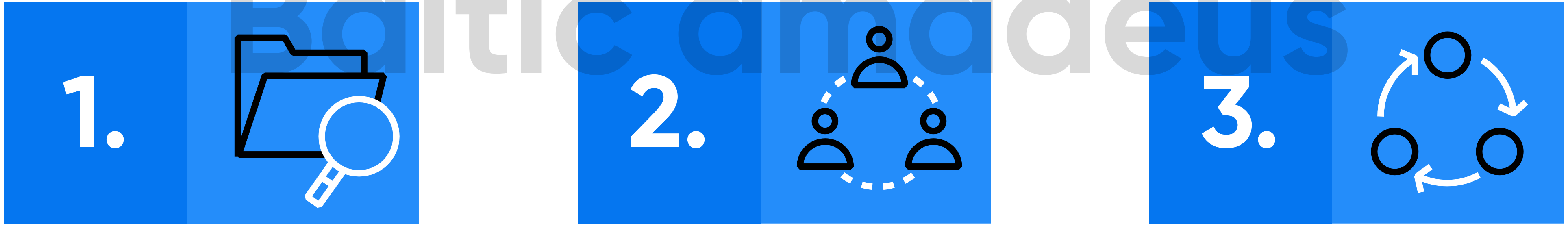
- ✓ Retains the best data lake properties – it scales across data sources, use cases, and audiences.
- ✓ Engage the entire business in turning raw data into trusted information.
- ✓ A system of trust can scale by leveraging smart and workflow-driven self-service tools with embedded data quality controls.

Cons

- ✓ Rather complement than replace the top-down approach.
- ✓ Heavily regulated processes, such as risk data aggregation in financial services, and some unique data, like consumer credit card information, specific particular attention.

III. Data governance implementation: challenges and required capabilities

Three steps to implement data governance



Discover and cleanse

Organise and empower

Automate and enable

1. Discover and cleanse

Do not go blind with your data

1. Discover and cleanse

- a. **Know and understand** your data
- b. **Involve business** users in data profiling
- c. **Establish trust** with advanced profiling
- d. **Delegate data cleansing** with self service
- e. **Orchestrate data** integrity across pipelines

a. Know and understand your data

Challenges:

- ✓ Manual data exploration does not work anymore
- ✓ Data sprawl demands a more automatic and systematic approach.

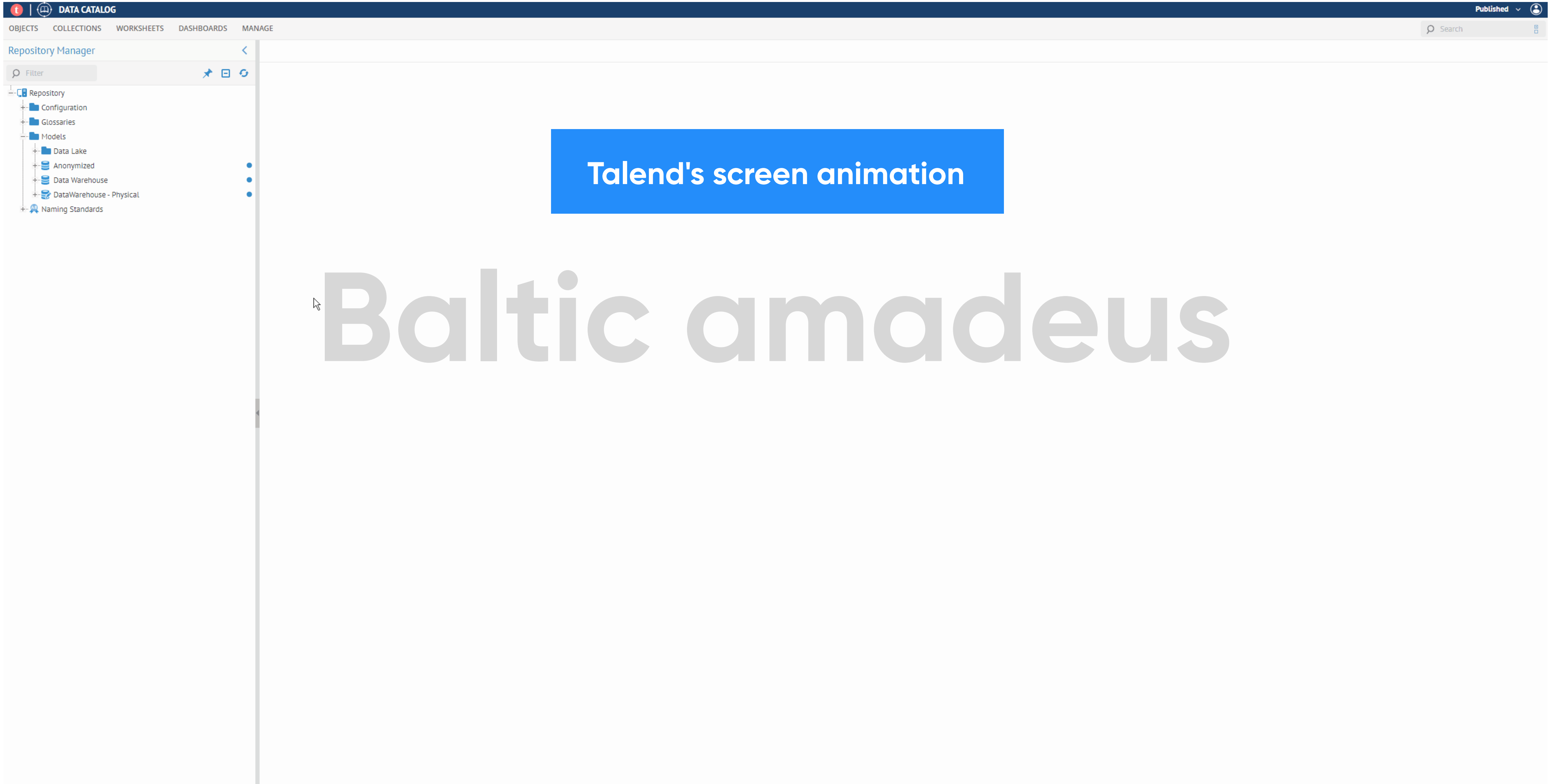
a. Know and understand your data

Capabilities:

- ✓ crawler for automated discovery of datasets;
- ✓ broad range of browse and search methods;
- ✓ easy-to-use sampling to assess data at a glance;
- ✓ automated relationship discovery between datasets;
- ✓ integrated business glossary, semantic;
- ✓ automated profiling and classification.

talend

1. 2. 3.

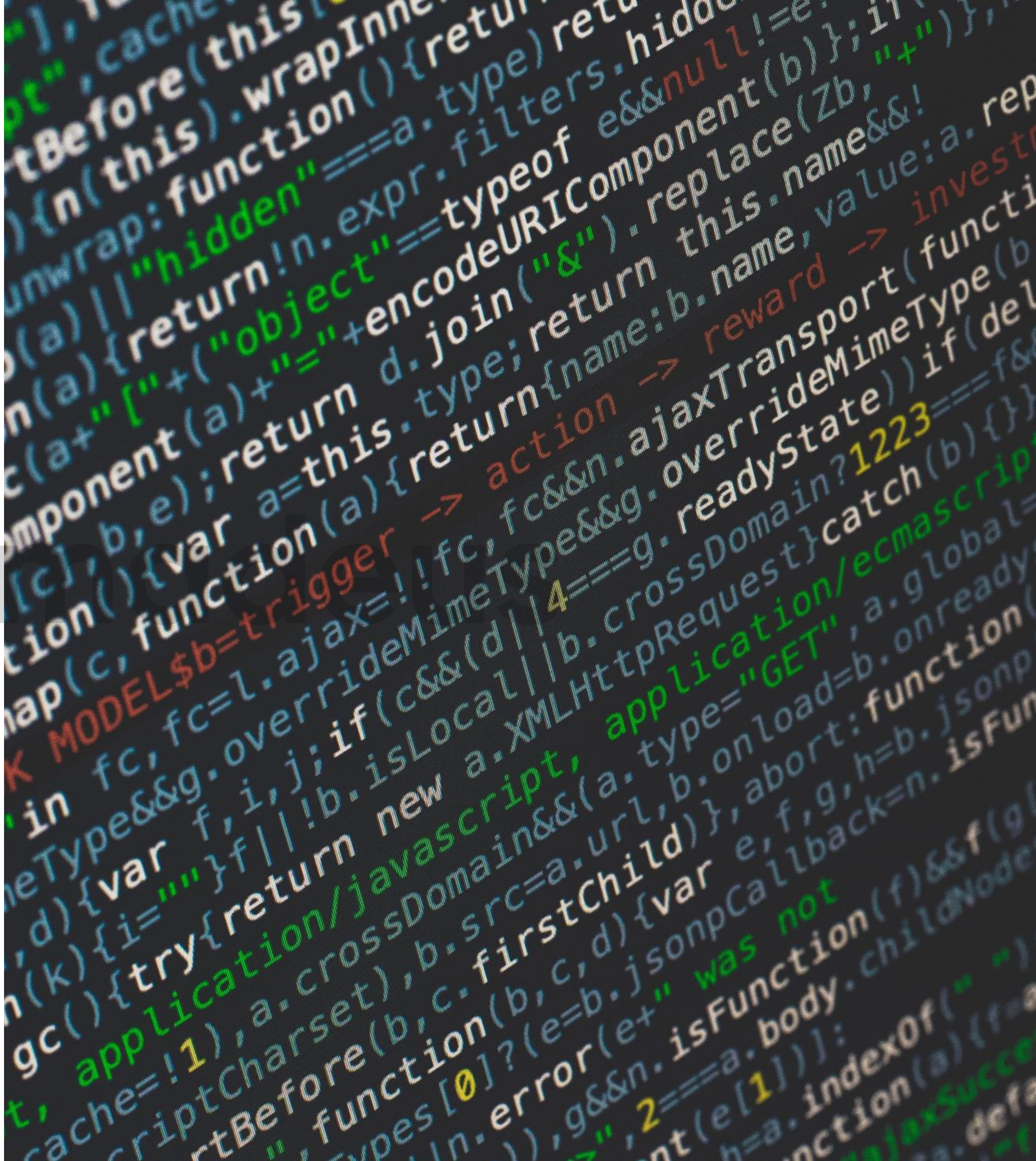


Talend's screen animation

b. Involve business users in data profiling

Challenges:

- ✓ We need to understand the data before we can fix it.
- ✓ Accurate diagnosis is required since data often comes in hidden formats, inoperable, or unstructured.
- ✓ People who know the data best are not technology experts. They need tools that can hide the technical complexity.



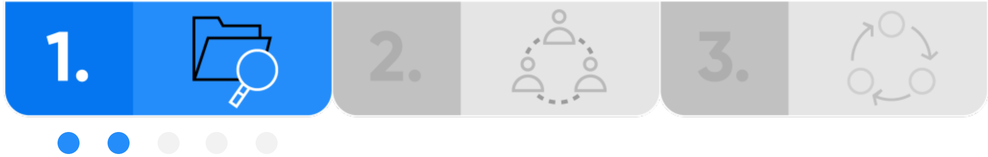
1. 2. 3.

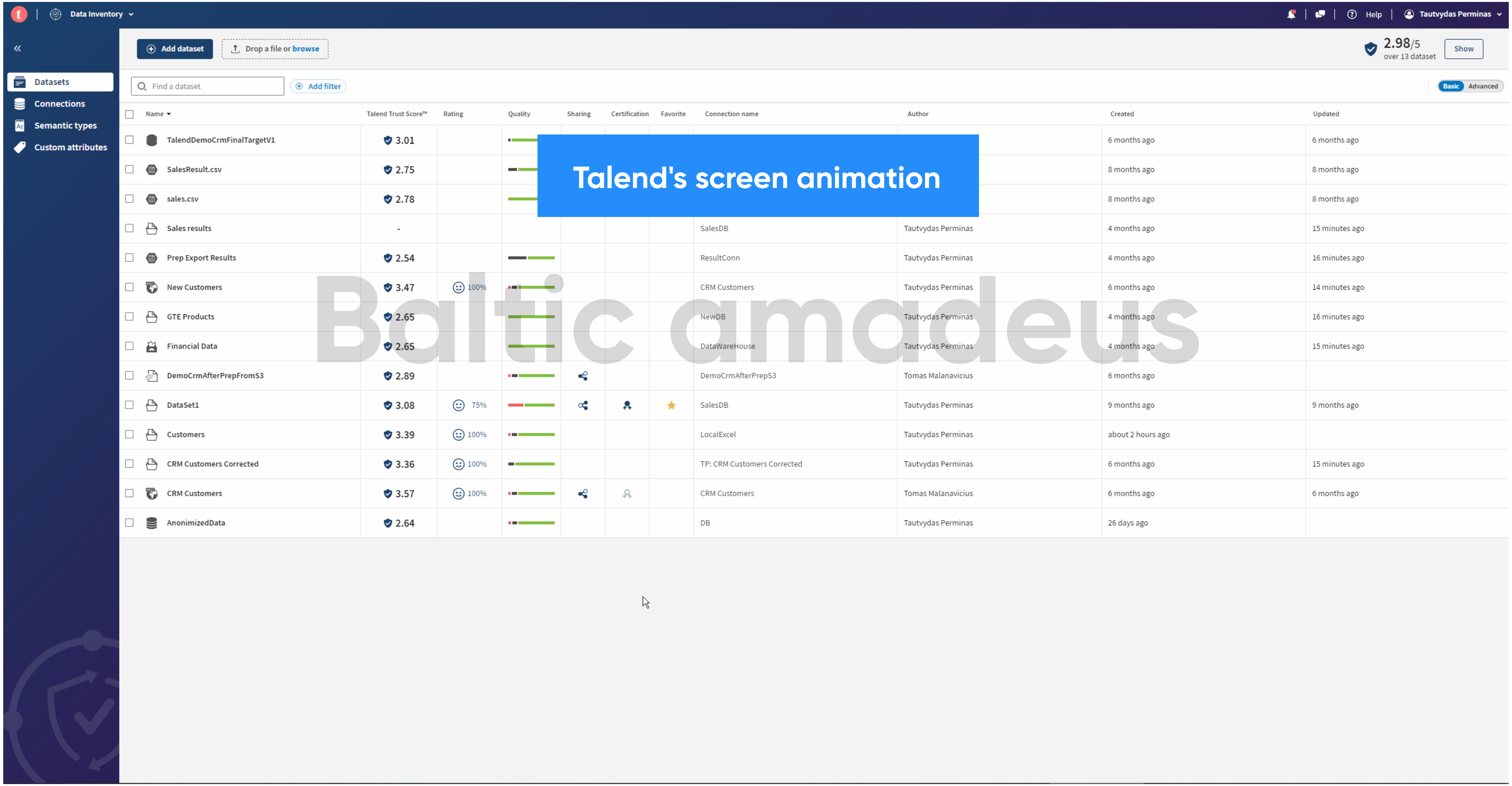
b. Involve business users in data profiling

Capabilities:

- ✓ simple, fast, and visual user experience for data exploration;
- ✓ automated data quality assessment with the help of indicators, trends, and patterns;
- ✓ easy identification of inaccurate, inconsistent, and incomplete data.

talend





	Name	Talend Trust Score™	Rating	Quality	Sharing	Certification	Favorite	Connection name	Author	Created	Updated
□	TalendDemoCrmFinalTargetV1	3.01								6 months ago	6 months ago
□	SalesResult.csv	2.75								8 months ago	8 months ago
□	sales.csv	2.78								8 months ago	8 months ago
□	Sales results	-						SalesDB	Tautvydas Perminas	4 months ago	15 minutes ago
□	Prep Export Results	2.54						ResultConn	Tautvydas Perminas	4 months ago	16 minutes ago
□	New Customers	3.47	☺ 100%					CRM Customers	Tautvydas Perminas	6 months ago	14 minutes ago
□	GTE Products	2.65						NewDB	Tautvydas Perminas	4 months ago	16 minutes ago
□	Financial Data	2.65						DataWareHouse	Tautvydas Perminas	4 months ago	15 minutes ago
□	DemoCrmAfterPrepFromS3	2.89			✓			DemoCrmAfterPrepS3	Tomas Malanavicius	6 months ago	
□	DataSet1	3.08	☺ 75%		✓	✓	★	SalesDB	Tautvydas Perminas	9 months ago	9 months ago
□	Customers	3.39	☺ 100%					LocalExcel	Tautvydas Perminas	about 2 hours ago	
□	CRM Customers Corrected	3.36	☺ 100%					TP: CRM Customers Corrected	Tautvydas Perminas	6 months ago	15 minutes ago
□	CRM Customers	3.57	☺ 100%		✓	✓		CRM Customers	Tomas Malanavicius	6 months ago	6 months ago
□	AnonimizedData	2.64						DB	Tautvydas Perminas	26 days ago	

Talend's screen animation

c. Establish trust with advanced profiling

Challenges:

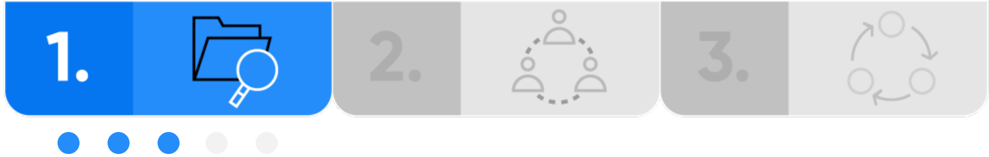
- ✓ Top-down approach needs a deeper look into the data.
- ✓ E.g., risk data aggregation/reporting – defined by formal principles and related regulations.
- ✓ Working with complex data structures requires the involvement of IT specialists and comprehensive tools.

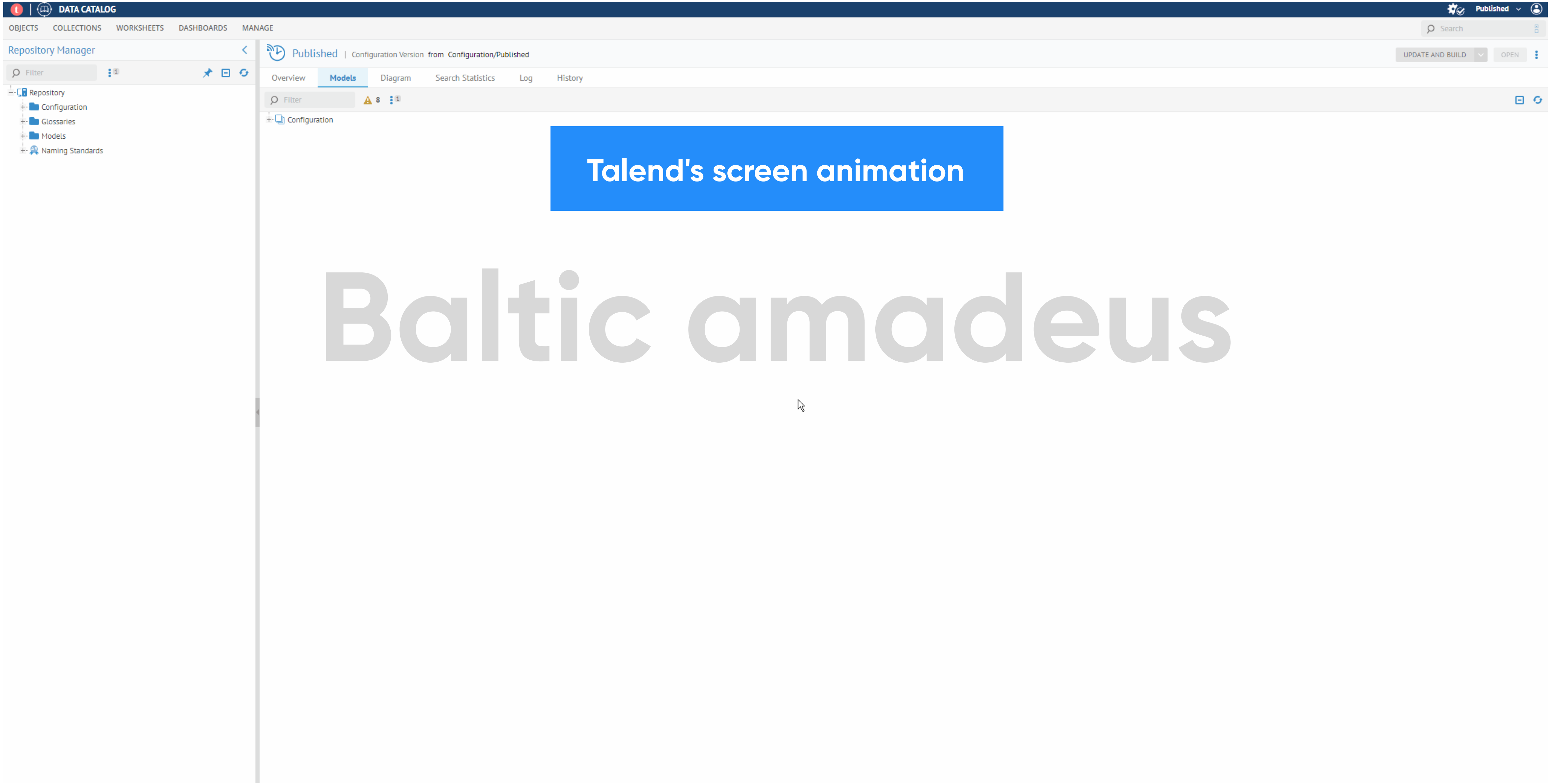
c. Establish trust with advanced profiling

Capabilities:

- ✓ connect to virtually any data sources to analyse data structure;
- ✓ define/analyse data using metadata repository;
- ✓ visualise the enterprise architecture;
- ✓ assess data quality and integrity at various levels using diverse analysis methods: database, table, column, content, redundancy, and correlation analysis.

talend



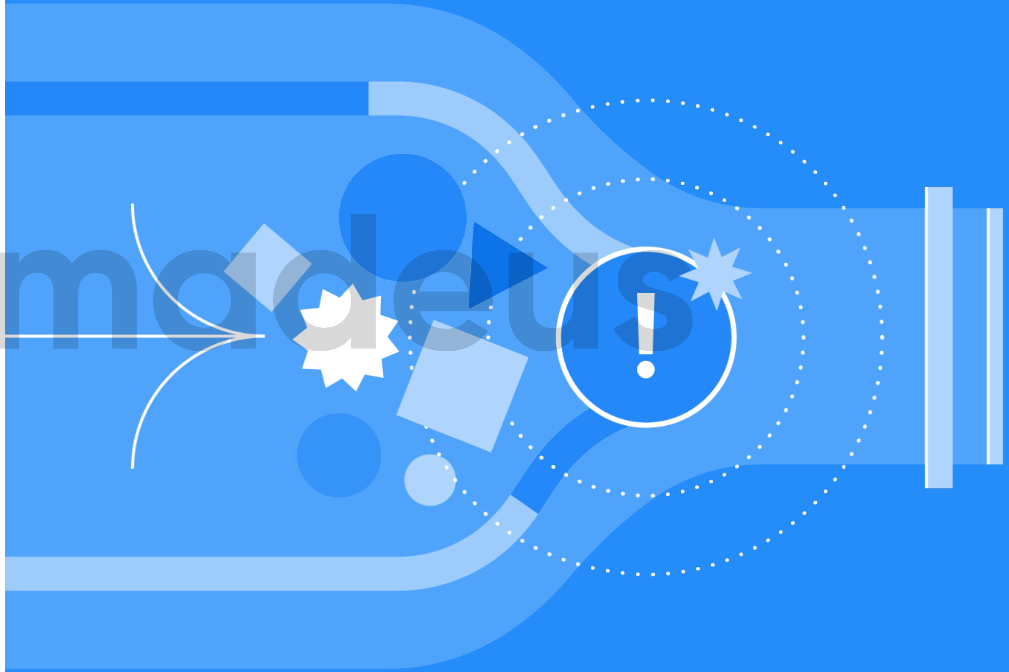


Talend's screen animation

d. Delegate data cleansing with self-service

Challenges:

- ✓ Data is not the responsibility of a single central organisation.
- ✓ Centralised data governance creates bottlenecks.
- ✓ As more non-technical people are involved in data preparation, we need smart tools to reduce complexity and minimise repetitive, manual work.



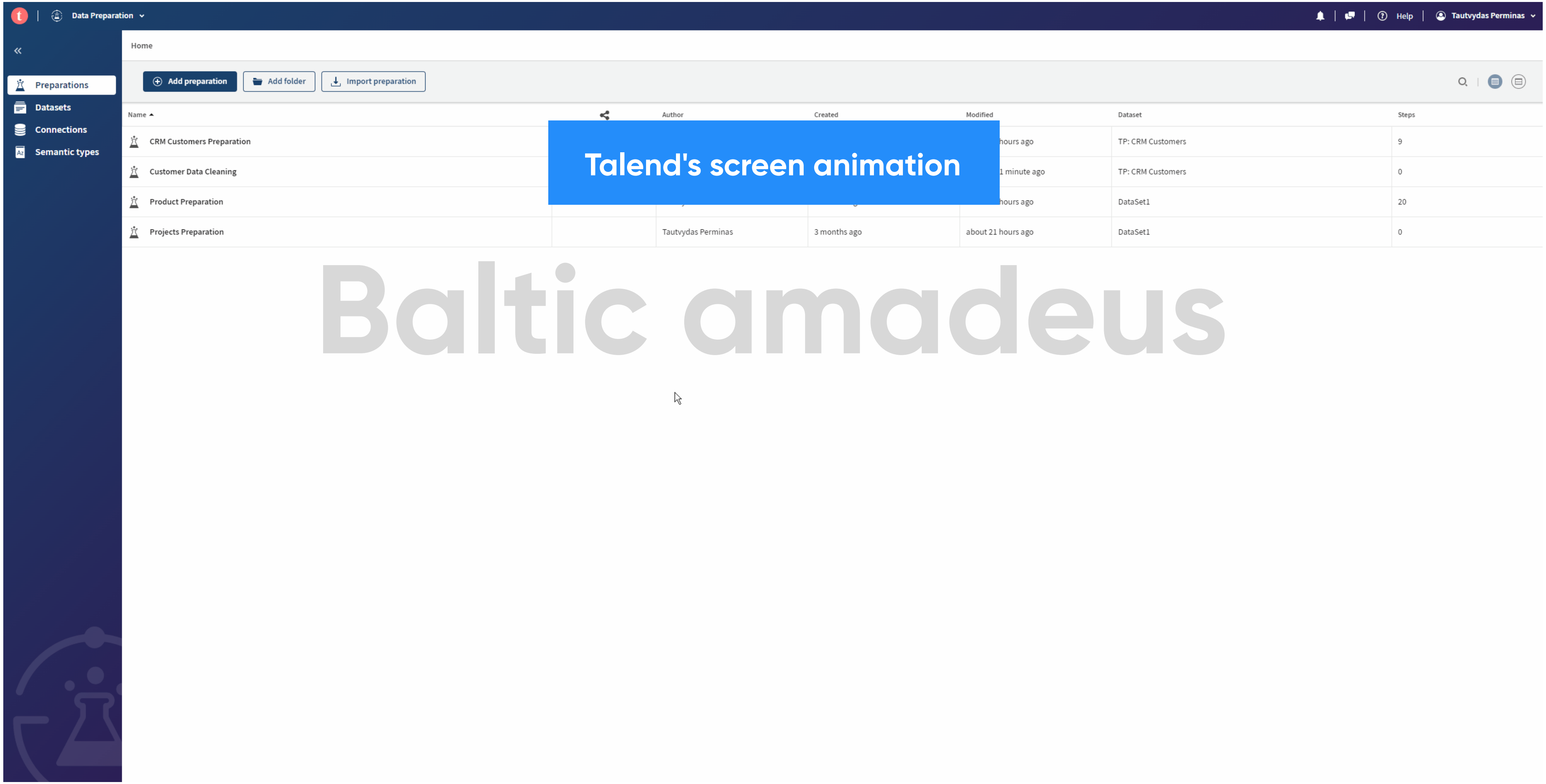
d. Delegate data cleansing with self-service

Capabilities:

- ✓ built-in, automatic data visualisation and statistics to understand data briefly;
- ✓ intuitive and smart functions for data cleansing and standardising;
- ✓ predefined and custom libraries of semantic types and regular expression-based rules.

talend

1. 2. 3.



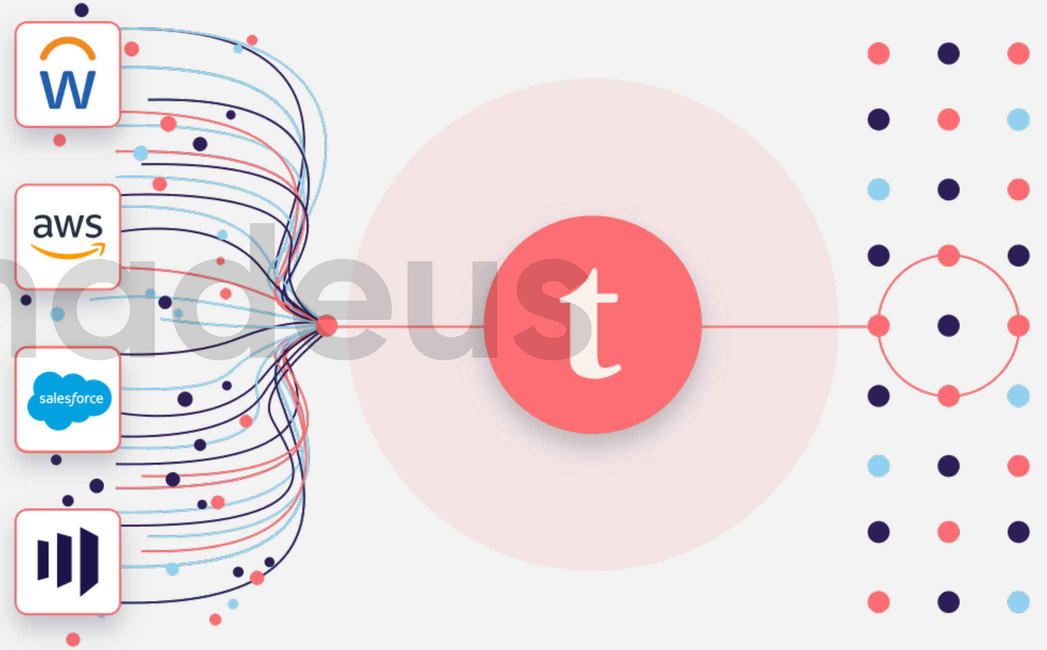
Name		Author	Created	Modified	Dataset	Steps
CRM Customers Preparation				hours ago	TP: CRM Customers	9
Customer Data Cleaning				1 minute ago	TP: CRM Customers	0
Product Preparation				hours ago	DataSet1	20
Projects Preparation		Tautvydas Perminas	3 months ago	about 21 hours ago	DataSet1	0

Talend's screen animation

e. Orchestrate data integrity across pipelines

Challenges:

- ✓ Data quality is not a stand-alone operation.
- ✓ It is crucial to run data quality operations upfront, natively from the data sources and the data lifecycle to deliver trusted data.
- ✓ It ensures that any data user or app could consume trusted data at the end.



e. Orchestrate data integrity across pipelines

Capabilities:

- ✓ apply data quality controls and remediations to the ingested data sources;
- ✓ run controls at any place (on-premises, in cloud, Big Data cluster) and at any time (on data at rest or streaming data);
- ✓ profile, cleanse, and standardise in any format or size.

talend

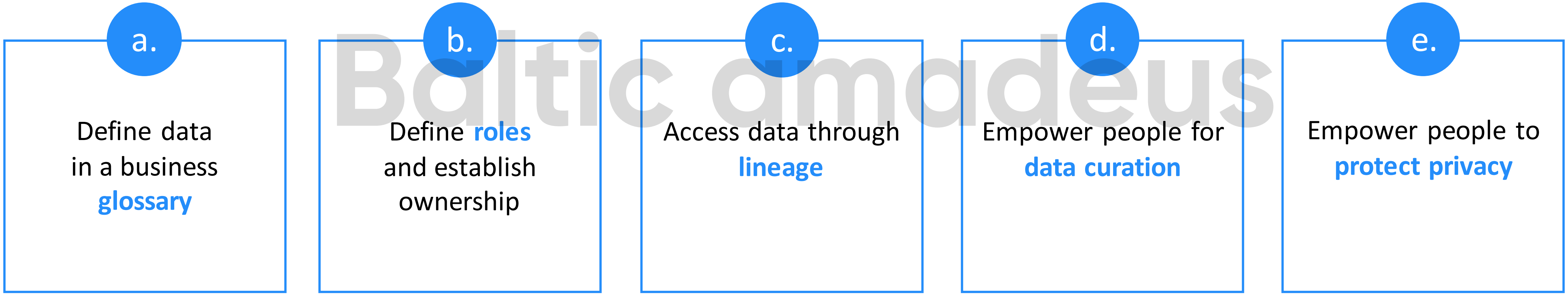
Key takeaway

Delegate data quality operations to business users
in a self-service mode while keeping control

2. Organise and empower

It is time to organise data assets for massive consumption

2. Organise and empower



a.

Define data in a business **glossary**

b.

Define **roles** and establish ownership

c.

Access data through **lineage**

d.

Empower people for **data curation**

e.

Empower people to **protect privacy**

a. Define data in a business glossary

Challenges:

- ✓ Without a clear definition, data can be very ambiguous.
- ✓ Incorrect interpretation causes misunderstandings.
- ✓ A business glossary is required to reach an agreement between all stakeholders.

customer buyer consumer

account THE SAME THING – MANY TITLES? client

purchaser user prospect

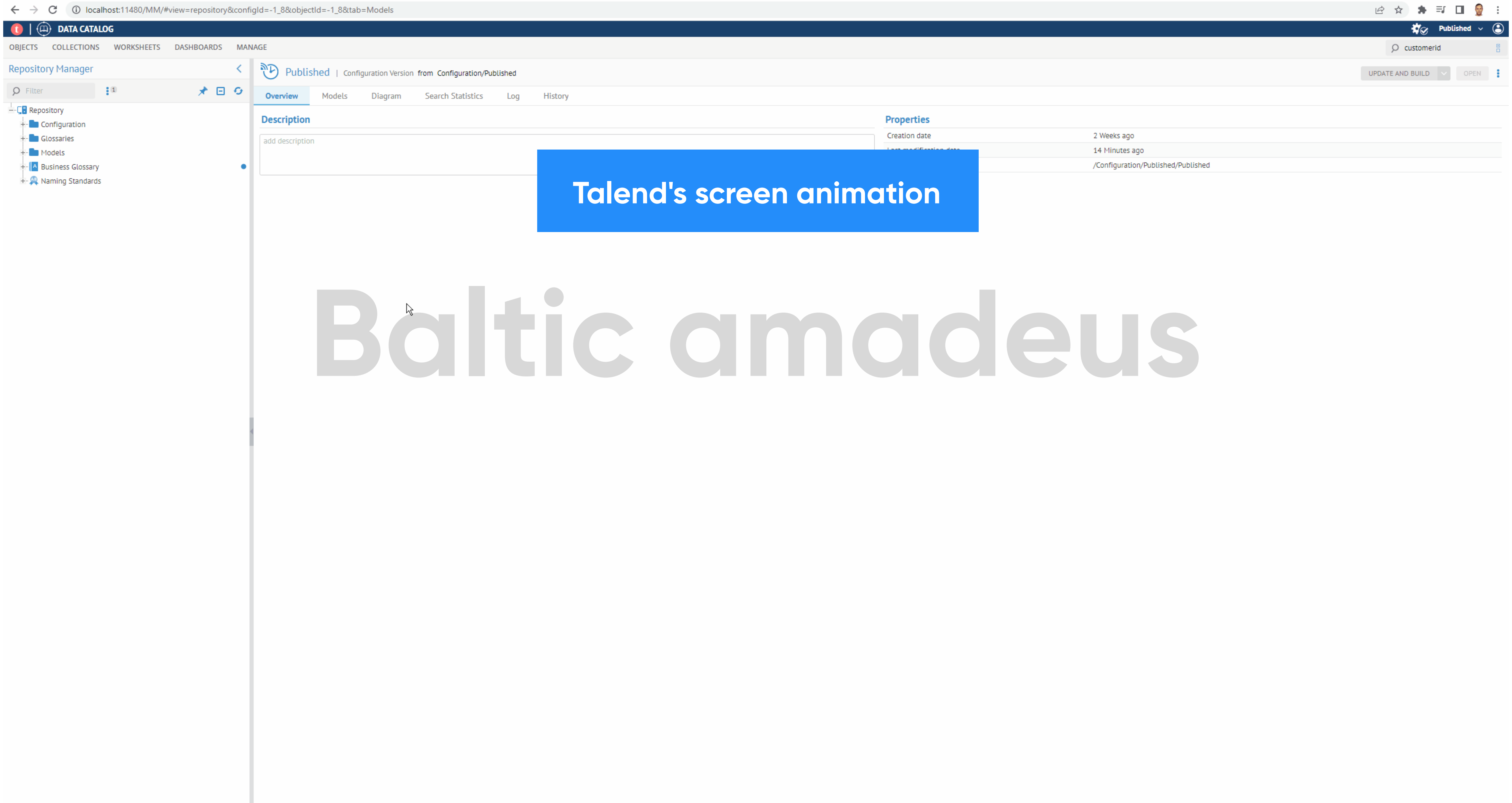
a. Define data in a business glossary

Capabilities:

- ✓ maintain an enterprise business glossary of **terminology, definitions, codes, validation rules** etc.;
- ✓ **aggregate** business terms to sub/categories;
- ✓ use **semantic mappings** to describe how elements in a source model define elements in a destination model;
- ✓ **organise**: maintain versions, assign responsibilities, manage workflow.

talend

1. 2. 3.



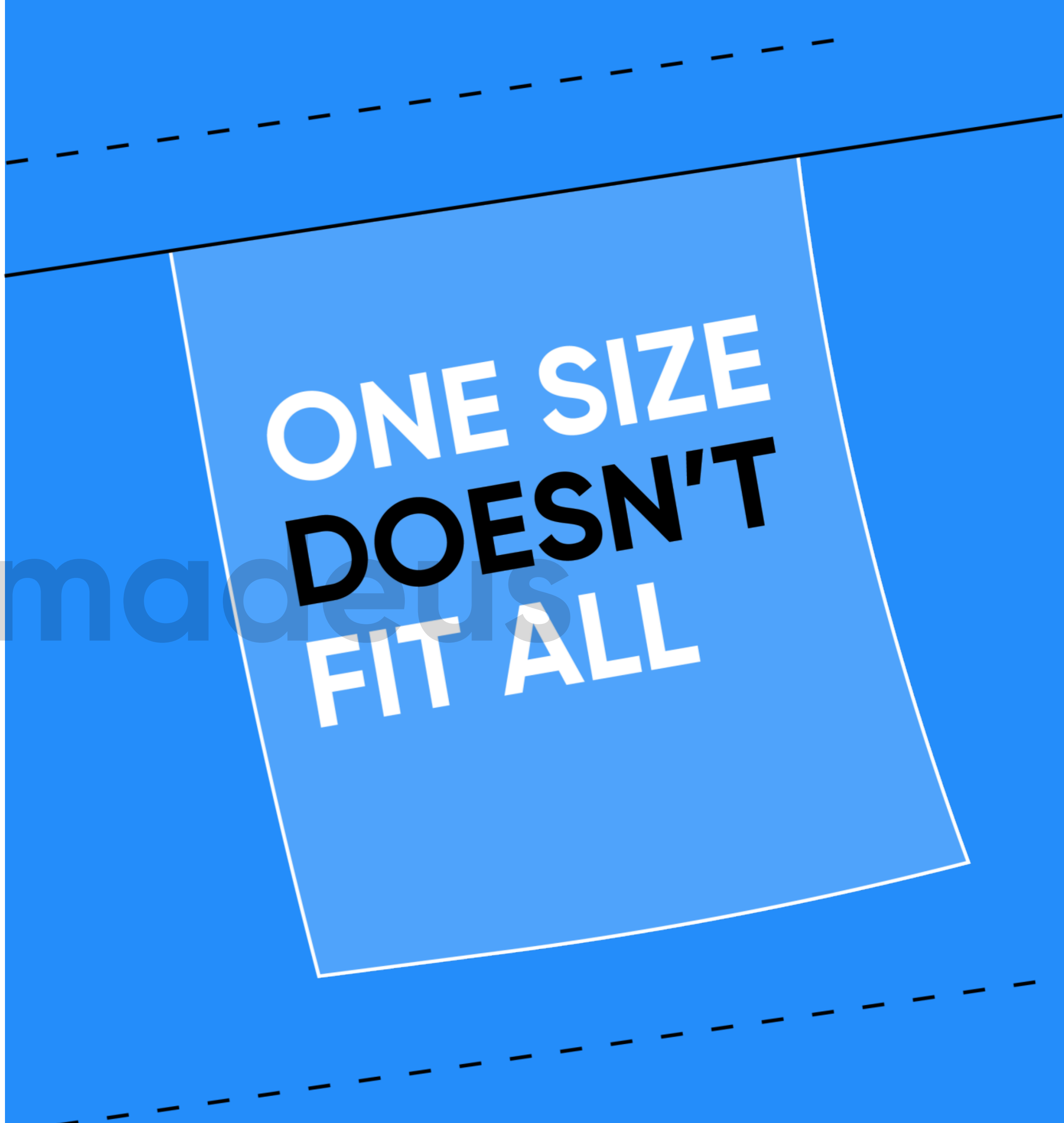
Creation date	2 Weeks ago
Last modification date	14 Minutes ago
	/Configuration/Published/Published

Talend's screen animation

b. Define roles and establish ownership

General trends:

- ✓ Roles depend on the organisation's structure, culture, risk management practices etc.
- ✓ Roles **shifting from centralised to decentralised positions** in the line of business departments.
- ✓ Effective governance **requires expertise** in compliance regulations and data management.



Critical roles in a data governance:

Chief data officers

Data protection officers

Data architects

Data engineers and developers

Data scientists

Data stewards

Business analysts

b. Define roles and establish ownership

Capabilities:

- ✓ **role-based access control** and work-flow roles;
- ✓ user and group **assignments to data assets**: categories and subcategories;
- ✓ flexible **modes of user authentication** (OAuth, SAML, etc.);
- ✓ usage **statistics** and **audit logs**.

talend

c. Access data through lineage

Challenges:

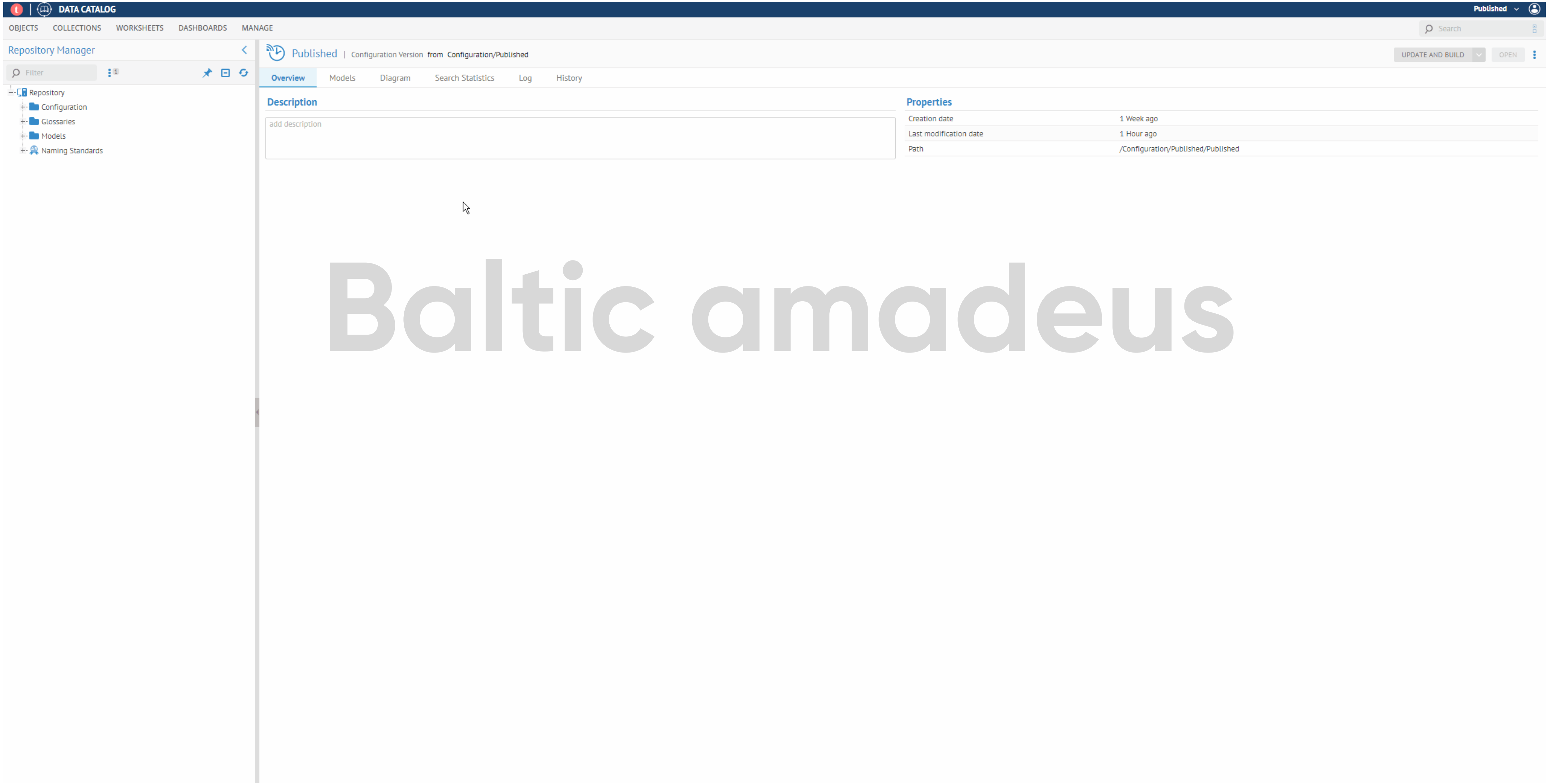
- ✓ How to **explain data** in your systems and analytics?
- ✓ How do quickly **answer audit trails** as requested by the competent authorities?
- ✓ How to identify **new data sources** in your data lake?
- ✓ How do we assess the **impact of IT change** on the data chain?

c. Access data through lineage

Capabilities:

- ✓ track data lineage to understand where the data comes from and how it was processed;
- ✓ trace data impact on understanding how changing some element can impact the whole data chain;
- ✓ trace semantic definition to discover the meaning of the report fields;
- ✓ track semantic usage to identify where data is held and potentially accessible.

talend



Published | Configuration Version from Configuration/Published

Description

add description

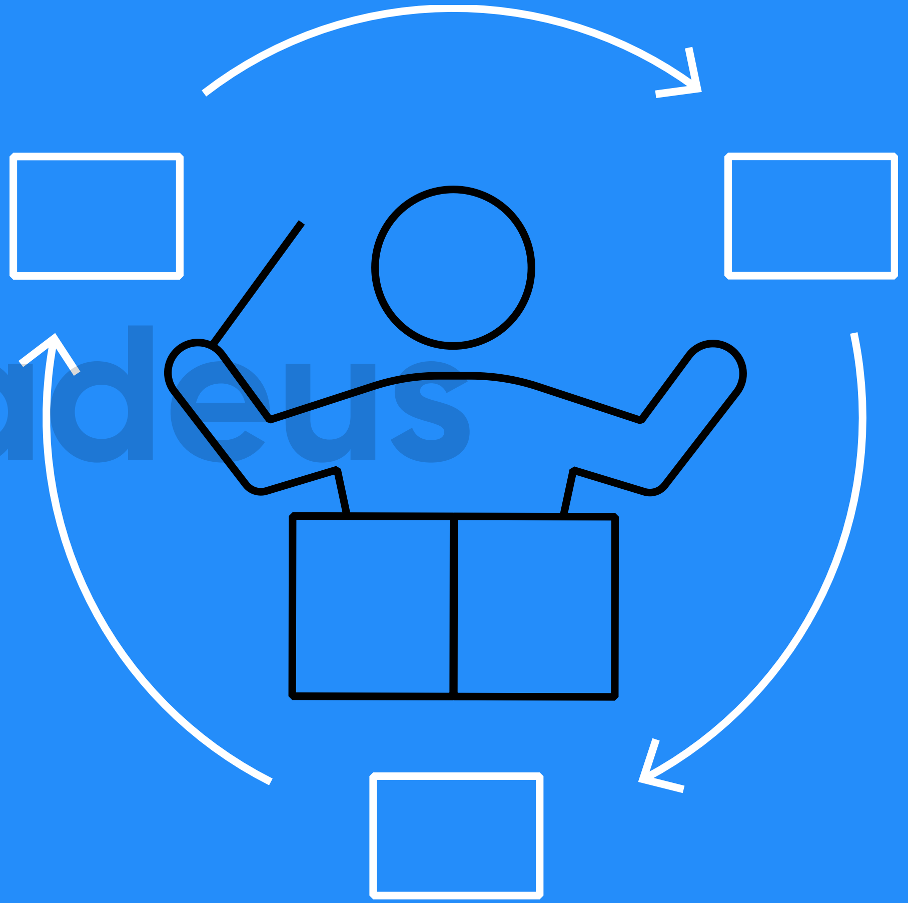
Properties

Creation date	1 Week ago
Last modification date	1 Hour ago
Path	/Configuration/Published/Published

d. Empower people for data curation

Challenges:

- ✓ It is not enough to give people tools to explore data.
- ✓ It is crucial to enable **data curation** and **remediation** by clearly defining who must do what.
- ✓ Data owners should manage everything by themselves and **act as orchestrators**.



1. 2. 3.

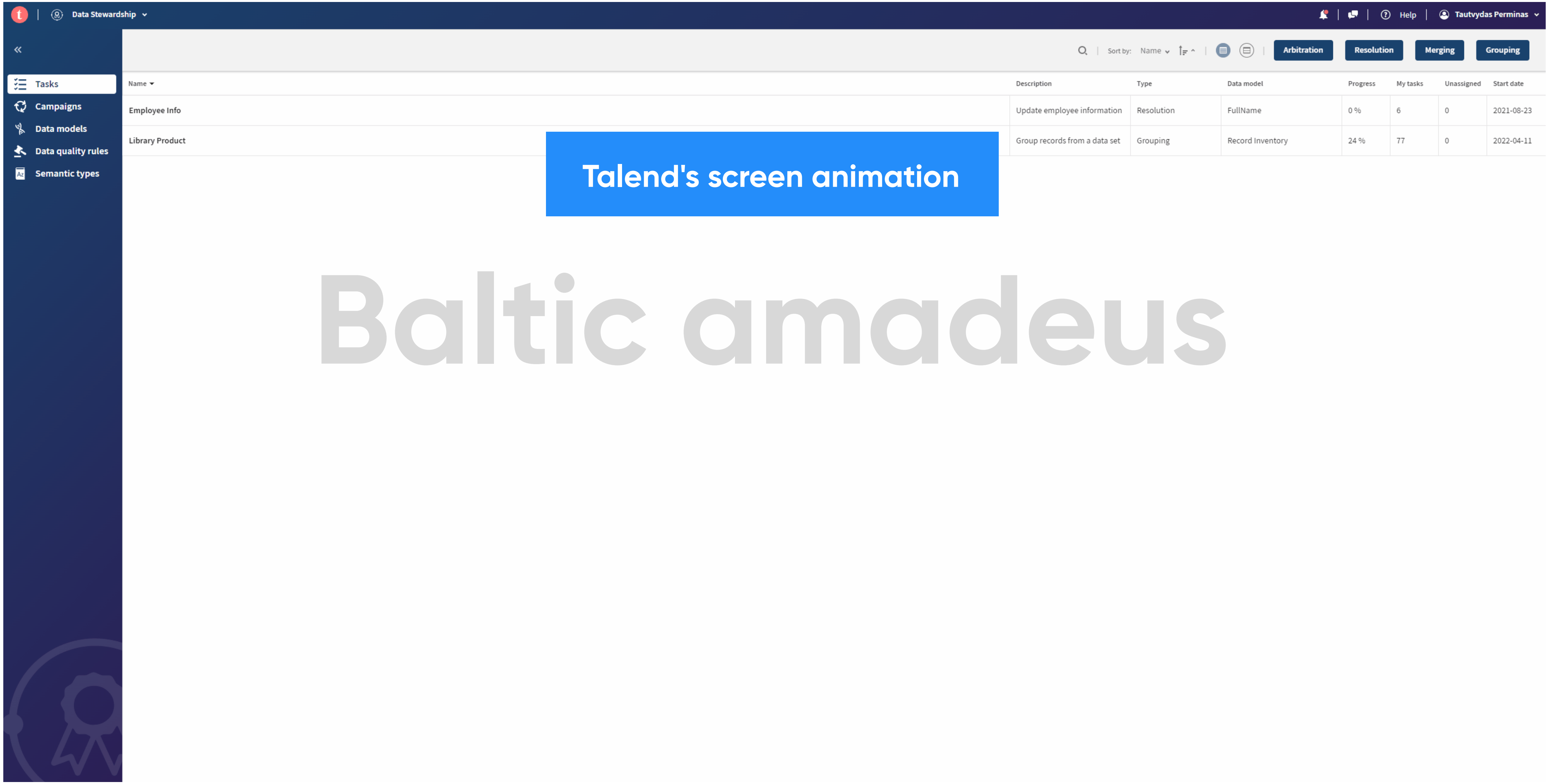
● ● ● ● ○

d. Empower people for data curation

Capabilities:

- ✓ design, orchestrate **data stewardship campaigns**;
- ✓ **delegate** data curation tasks to appropriate roles and **control** progress;
- ✓ **resolve, enrich and validate inconsistent data** in a user-friendly interface;
- ✓ **track/audit history** of curation/remediation.

talend



Name	Description	Type	Data model	Progress	My tasks	Unassigned	Start date
Employee Info	Update employee information	Resolution	FullName	0 %	6	0	2021-08-23
Library Product	Group records from a data set	Grouping	Record Inventory	24 %	77	0	2022-04-11

Talend's screen animation

e. Empower people to protect privacy

Challenges:

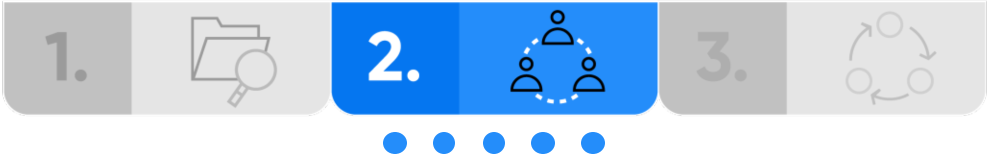
- ✓ Data security and data privacy is shared responsibility.
- ✓ A large audience should protect the data on their own.
- ✓ Data protection task delegation to people who might not be technical experts.

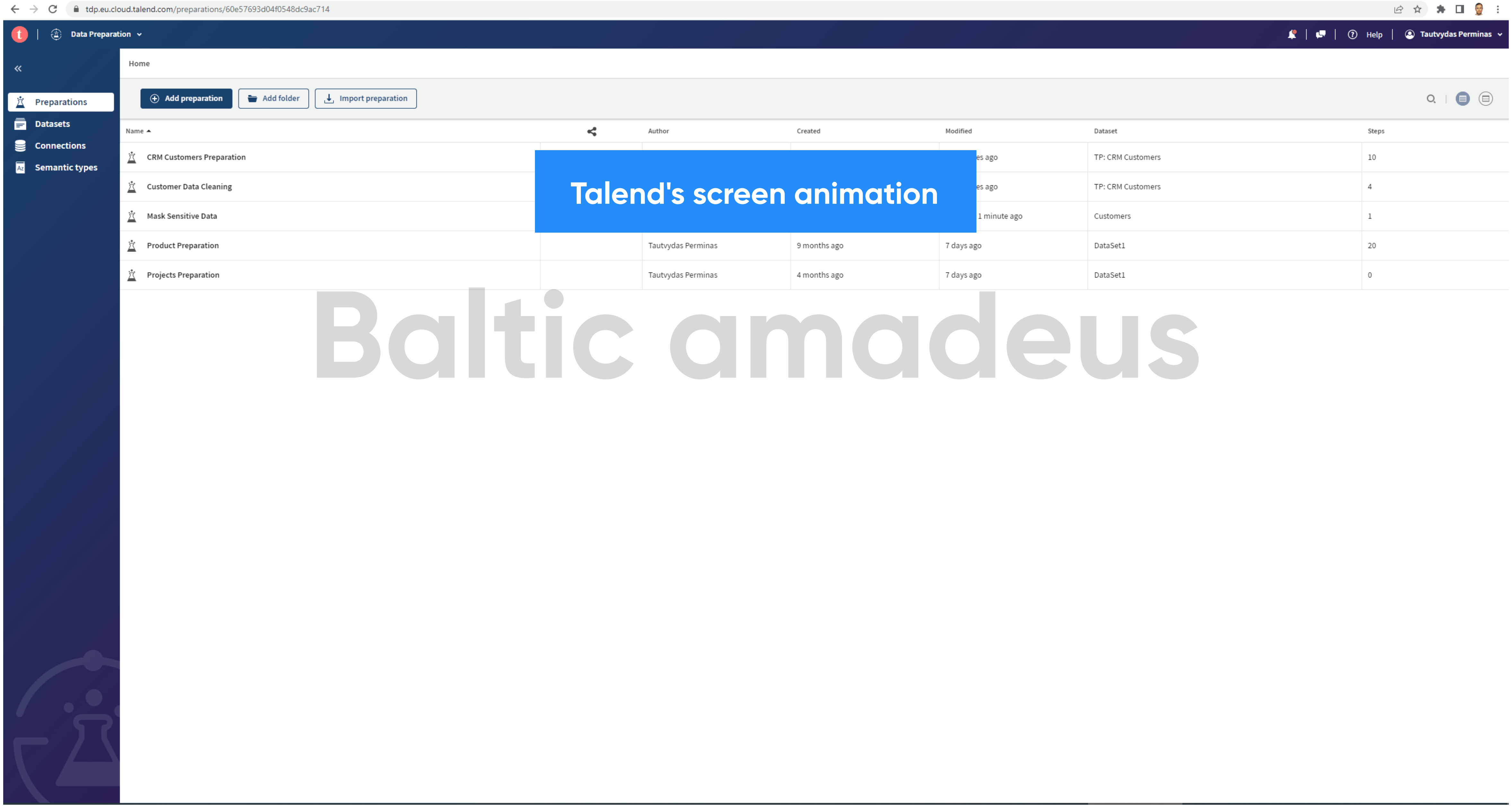
e. Empower people to protect privacy

Capabilities:

- ✓ data masking capabilities integrated across all applications;
- ✓ various functions for data masking, e.g.:
  - ✓ semantic masking by maintain data pattern;
  - ✓ random characters;
  - ✓ replacement;
  - ✓ etc.

talend





Home

Name ▲	Author	Created	Modified	Dataset	Steps
CRM Customers Preparation			es ago	TP: CRM Customers	10
Customer Data Cleaning			es ago	TP: CRM Customers	4
Mask Sensitive Data			1 minute ago	Customers	1
Product Preparation	Tautvydas Perminas	9 months ago	7 days ago	DataSet1	20
Projects Preparation	Tautvydas Perminas	4 months ago	7 days ago	DataSet1	0

Talend's screen animation

Key takeaway

Centralising data into a shareable environment
will save time and resources once operationalised

3. Automate and enable

Let’s extract all data values by delivering at scale

3. Automate and enable

a. Scale up data governance with **machine learning**

b. **Automate privacy** protection

c. **Embed controls** into the data chain

d. Publish trusted data **into applications**

a. Scale up data governance with ML

Capabilities:

ML-based features integrated into multiple applications:

- ✓ pattern recognition, best next action suggestion, smart data cleaning;
- ✓ smart data error resolution, matching, and deduplication;
- ✓ out-of-the-box and comprehensive algorithms for data mining/classification, cluster analysis, prediction, recommendation, regression.

b. Automate privacy protection

Capabilities:

- ✓ automatically spot sensitive/personal data against new data sources based on patterns, dictionaries or ontologies;
- ✓ automatically apply data masking or encryption on those elements;
- ✓ implement and automate other regulations such *as right of access, right of rectification, right to be forgotten.*

c. Embed controls into the data chain

Capabilities:

- ✓ control/orchestrate all your data pipelines in one place;
- ✓ rich set (>2 000) of data connectors and functions;
- ✓ operationalise and automate any jobs or flows to keep on structuring and cleaning your data along the data lifecycle.

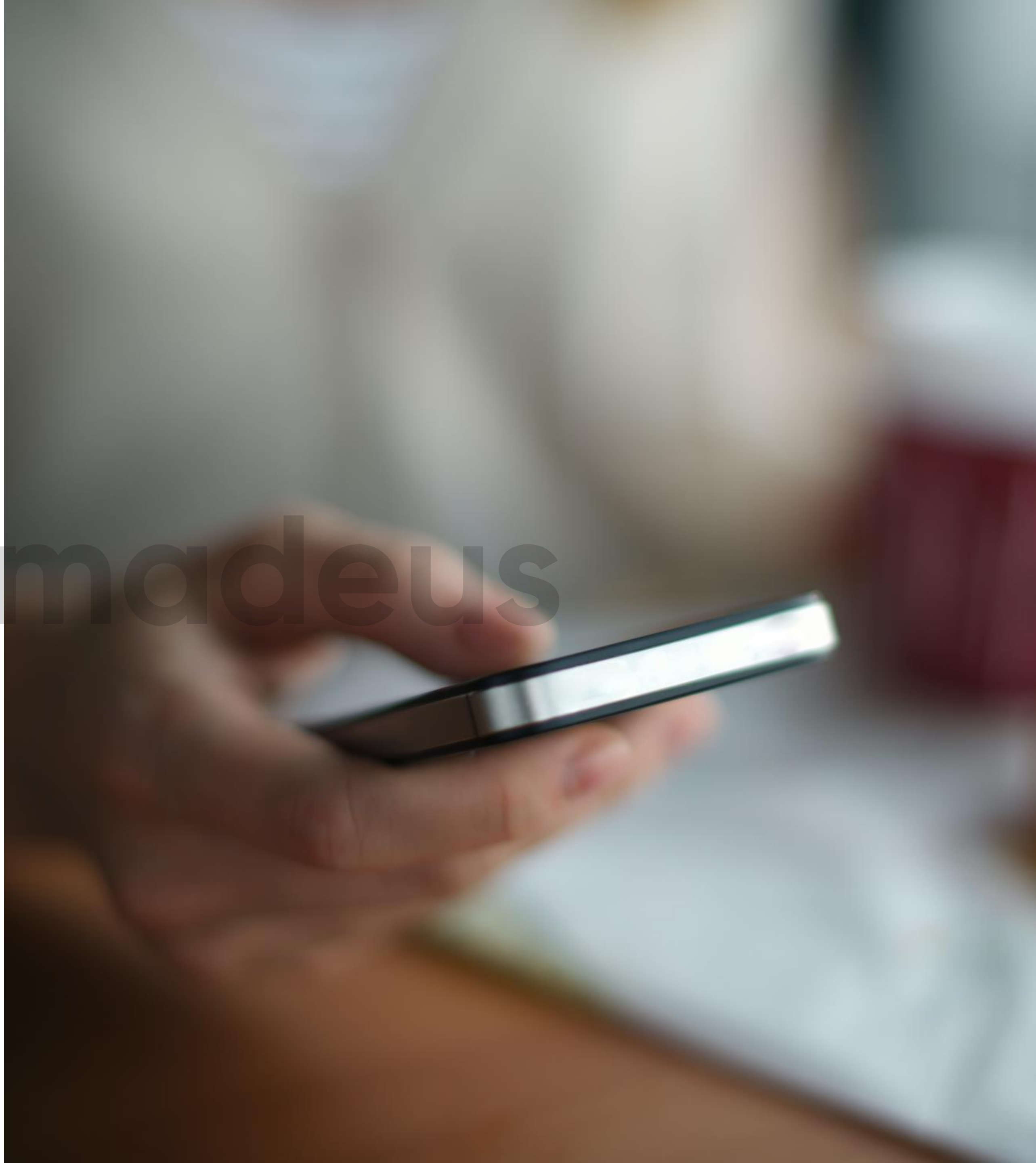
talend

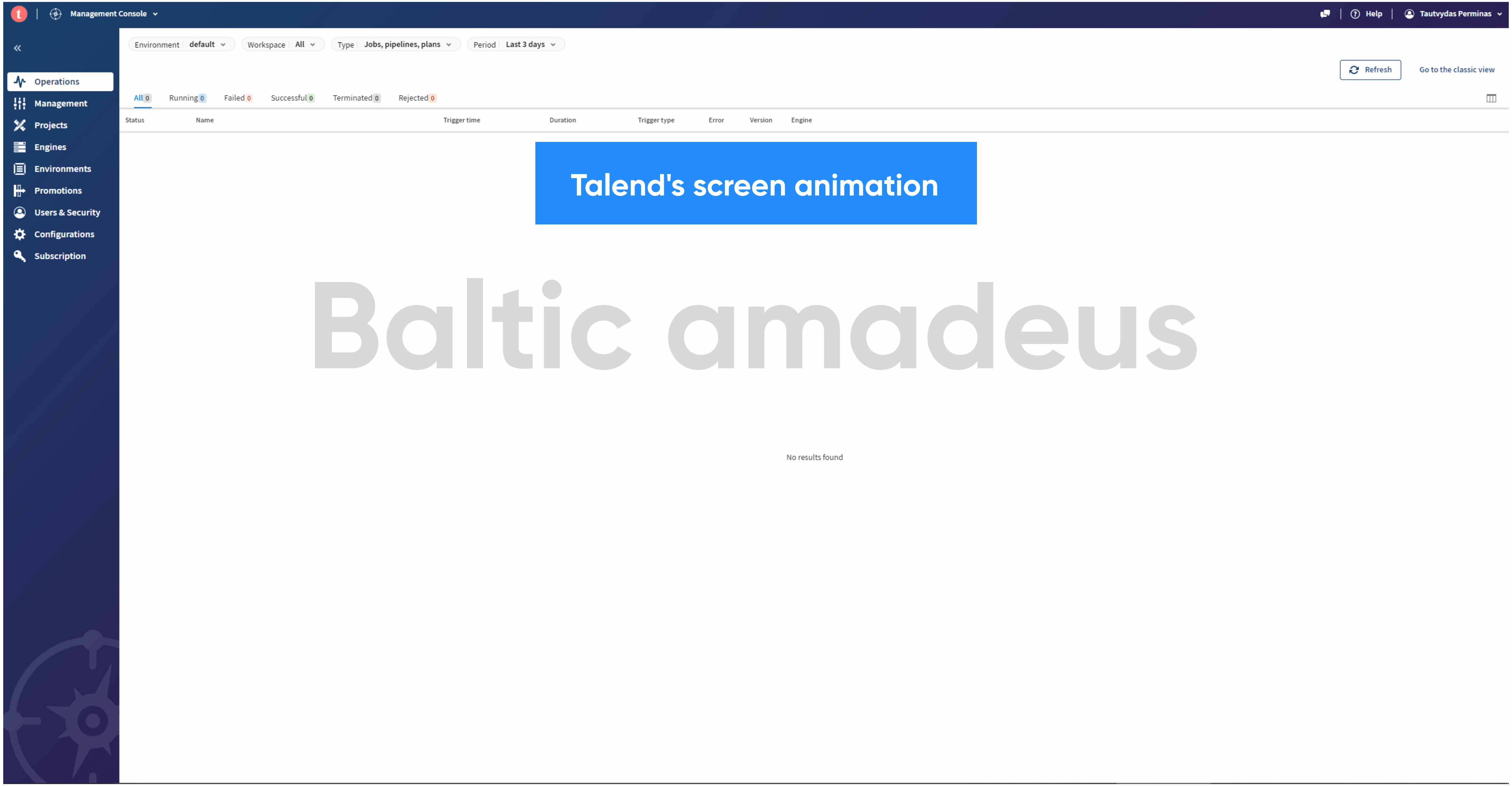
1. 2. 3.

d. Publish trusted data into applications

Capabilities:

- ✓ API/application integration in the same platform as data governance;
- ✓ easy to use, contract-based API designer;
- ✓ visual API tester to test, debug, and simulate real-life usage;
- ✓ auto-generated API reference documentation;
- ✓ automatic API mocking.





Environment: default | Workspace: All | Type: Jobs, pipelines, plans | Period: Last 3 days

All 0 | Running 0 | Failed 0 | Successful 0 | Terminated 0 | Rejected 0

Status	Name	Trigger time	Duration	Trigger type	Error	Version	Engine

No results found

Talend's screen animation

Key takeaway

Leverage the power of automation to streamline
your dataflows and use machine learning to scale faster

baltic amadeus

Successful governance requires people and processes

Find the right

people that cares about the data quality:

- ✓ start with data management strategy;
- ✓ establish processes;
- ✓ define and assign roles;
- ✓ find right data governance tool for your organisation.

Successful governance requires a modern data platform

Find the right

data governance solution for your organisation by looking for:

- ✓ scalable software that is easy to integrate with the organisation's existing environment;
- ✓ robust plug-and-play capabilities that are cost-efficient and easy to use;
- ✓ cloud-based applications to avoid the overhead required for on-premise systems.

Modern data platform

Should help you to:

- ✓ capture and understand data through discovery, profiling, and benchmarking;
- ✓ improve the quality of your data with validation, data cleansing, and data enrichment;
- ✓ integrate data with metadata-driven ETL and ELT;
- ✓ track and trace your data with end-to-end data lineage;
- ✓ control your data with tools that actively review and monitor;
- ✓ document your data to augment it with metadata;
- ✓ empower people who know data the best to contribute with self-service tools.

Time for your questions

Q&A SESSION