Modern Data Mgmt: How to Set Up Your Data For Success

Alec Bialosky - Select Star

Data Council 2023

Fast-growing companies and F500s rely on Select Star for their Data Discovery needs

Alec Bialosky

Business Operations

pitney bowes

Fivetran

Opendoor

Square

BLOCK

FAIRE

Xometry

productboard

BOWERY

Handshake

LIVESPORT

Flock Freight

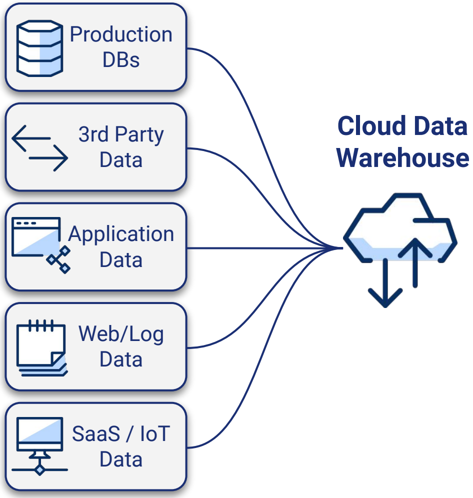
VTS

Agenda

I. Why Data Discovery?
II. Rollout Strategy
III. Best Practices

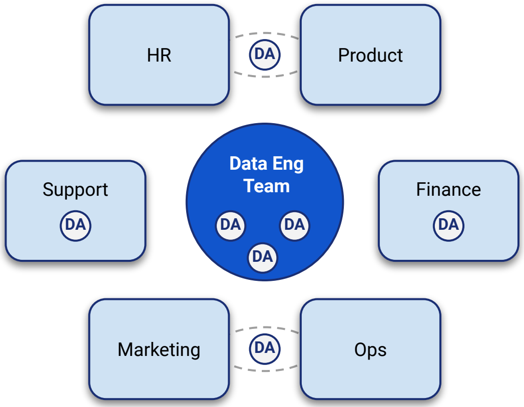
Changes to the modern data stack have led orgs to have more data, and more data users, than ever before…

Explosion in data volume & data sources going to CDWs

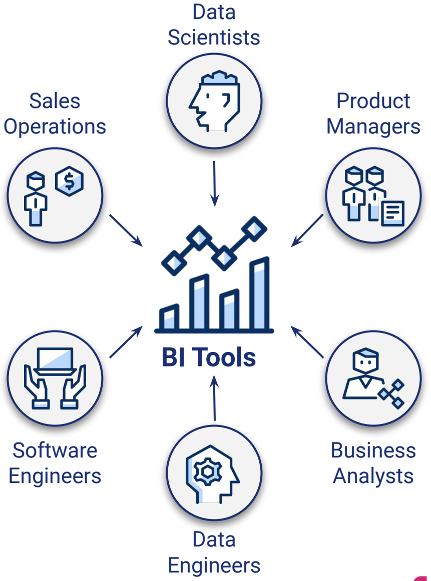


Decentralization of data ownership

Embedded Model



Democratization of data access



… Making finding and understanding data a big challenge in many organizations today

Marketing Analyst: Where can I find the customer engagement numbers?

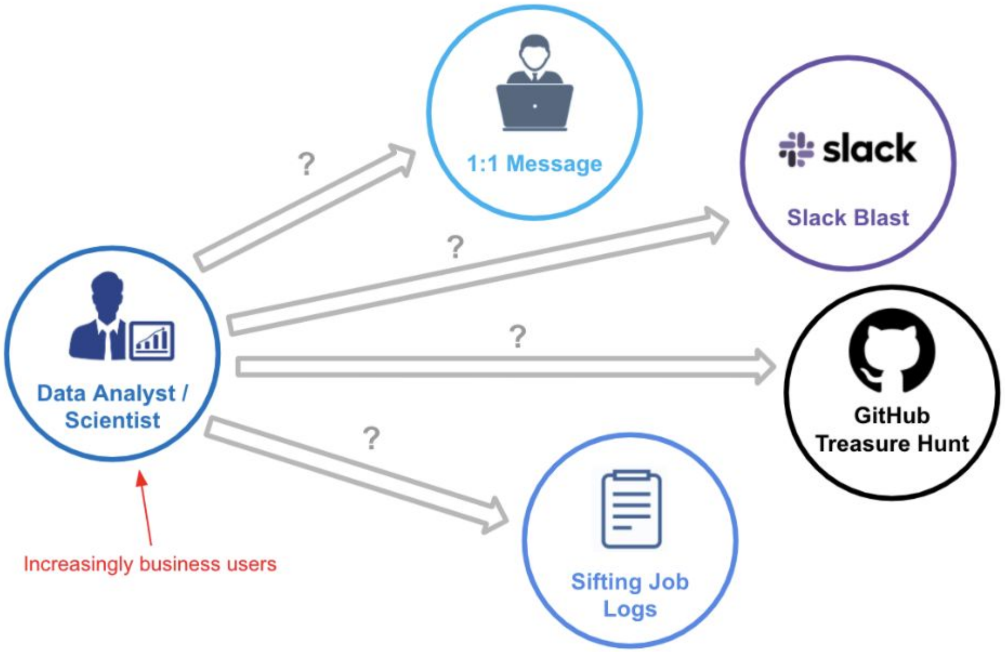
Product Manager: Where did the data come from?

General Manager: How did you calculate this metric?

Data Scientist: “Do we have this data?”

Finance Analyst: Why are my revenue numbers look different than yours?

Problem: Data teams waste time discovering data, or answering questions for others



Opendoor:
"Every day, we get these questions from occasional data users and there are 200 of them. It's a lot for our team to support everyone."

Handshake:
"Even for experienced SQL users, it takes 3-4 months before they are comfortable discovering data on their own"

Bowery Farming:
- 10-30 data questions on Slack every week
- *"When you get a question about something you didn't work on, I have to go find the query they are asking about, and look up the tables and columns it queries, which alone takes 10-20 minutes every single time"*

Existing solutions are resource intensive and/or don’t meet today’s needs

Wiki / Documentation

Confluence

Notion

Google Sheets

Outdated Information:

Manual documentation can’t keep up with the changes in the DWH / BI tool

Custom DIY / Open Source

Data Hub

Apache Atlas

Amundsen

Heavy Engineering Investment:

6-12 months to productionize a custom tool and requires continuous investment for upkeep

Enterprise Data Catalog

Informatica

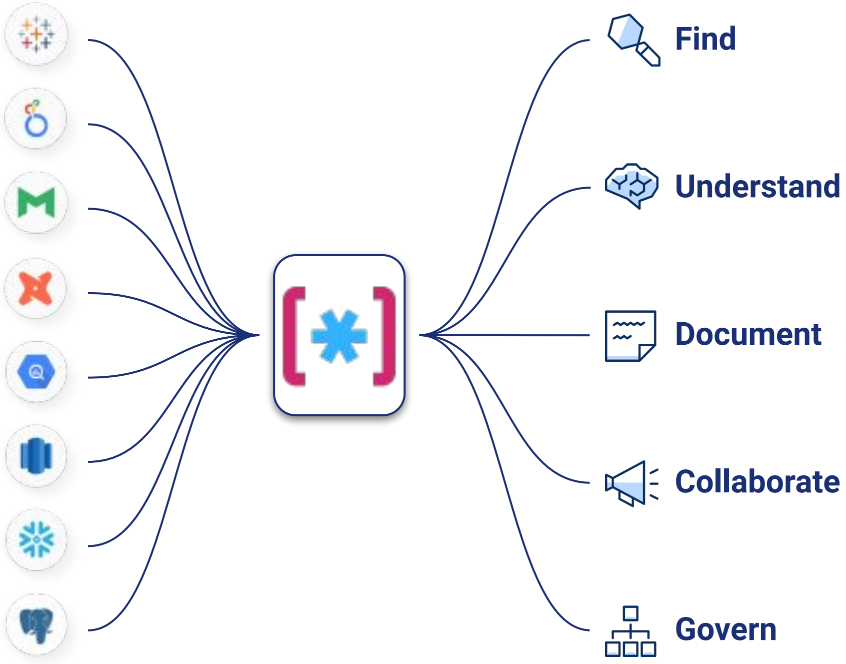
Alation

collibra

Poor UX and Manual Workflow:

Upfront financial investment, dedicated Data Governance team with training required

Data Discovery Platforms create a single source of truth for data within your organization, saving time and increasing trust



Find

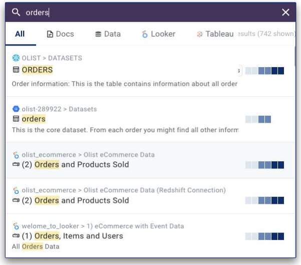
Understand

Document

Collaborate

Govern

Universal Search



orders

All · Docs · Data · Looker · Tableau · esults (742 shown)

OLIST > DATASETS
ORDERS
Order information: This is the table contains information about all order

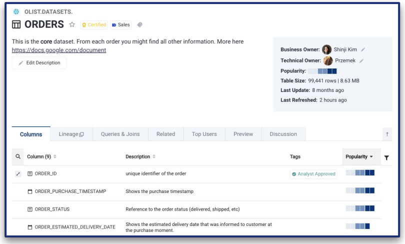
olist-289922 > Datasets
orders
This is the core dataset. From each order you might find all other inform

olist_ecommerce > Olist eCommerce Data
(2) Orders and Products Sold

olist_ecommerce > Olist eCommerce Data (Redshift Connection)
(2) Orders and Products Sold

welome_to_looker > 1) eCommerce with Event Data
(1) Orders, Items and Users
All Orders Data

Automated Insights & Documentation



OLIST.DATASETS.

ORDERS ☆ Certified Sales

This is the **core** dataset. From each order you might find all other information. More here https://docs.google.com/document

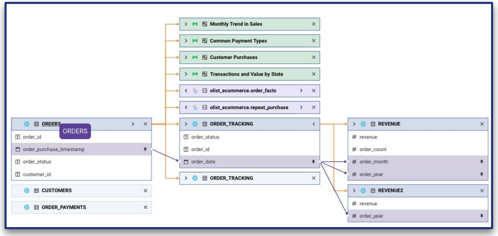
Edit Description

Business Owner: Shinji Kim
Technical Owner: Przemek
Popularity:
Table Size: 99,441 rows | 8.63 MB
Last Update: 8 months ago
Last Refreshed: 2 hours ago

Columns · Lineage · Queries & Joins · Related · Top Users · Preview · Discussion

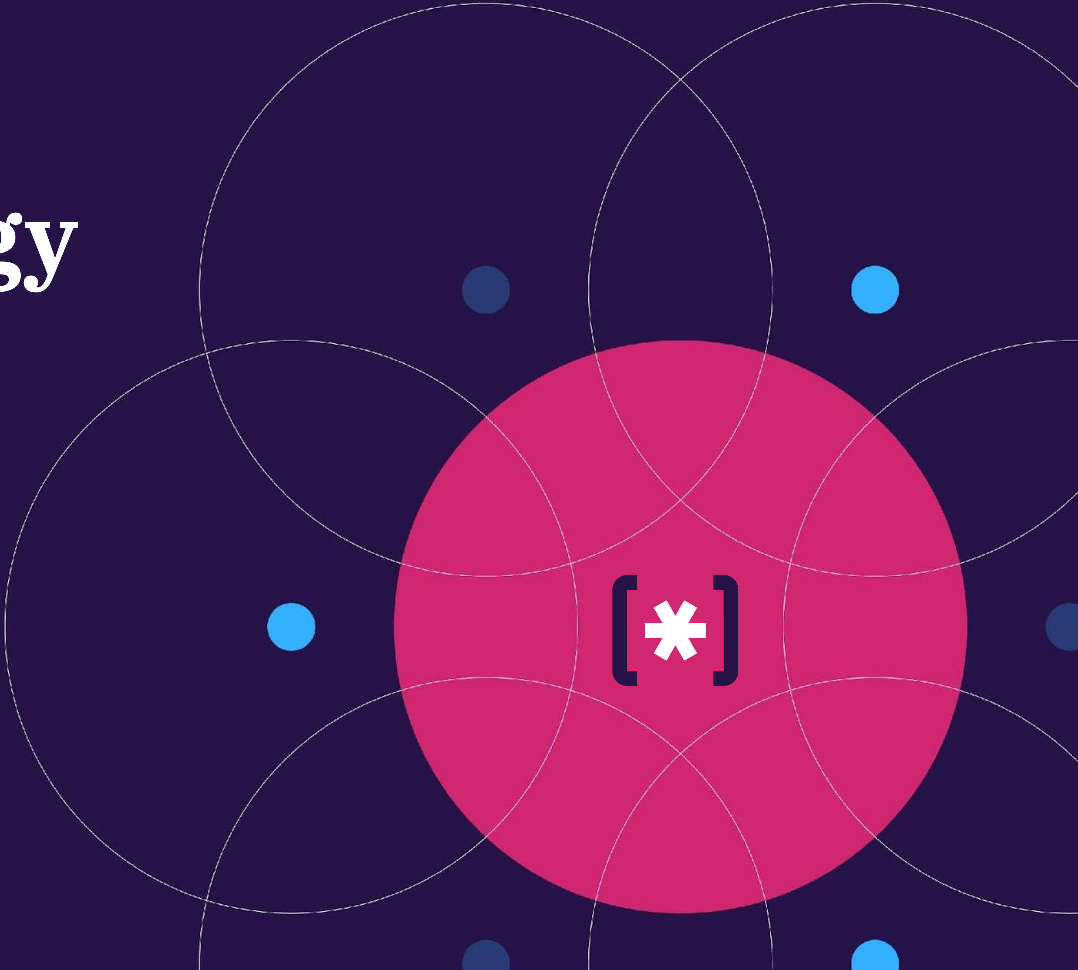
Column (9)	Description	Tags	Popularity
ORDER_ID	unique identifier of the order	Analyst Approved	
ORDER_PURCHASE_TIMESTAMP	Shows the purchase timestamp		
ORDER_STATUS	Reference to the order status (delivered, shipped, etc)		
ORDER_ESTIMATED_DELIVERY_DATE	Shows the estimated delivery date that was informed to customer at the purchase moment.		

Column Level Lineage



II.

Rollout Strategy



Make sure there is alignment across the organization on the Why, the Who, and the How for Data Discovery

Objectives

Why are we deploying a Data Discovery tool? What are the specific pain points we are trying to solve?

Stakeholders

Who will be responsible for the rollout? Who will the end users be?

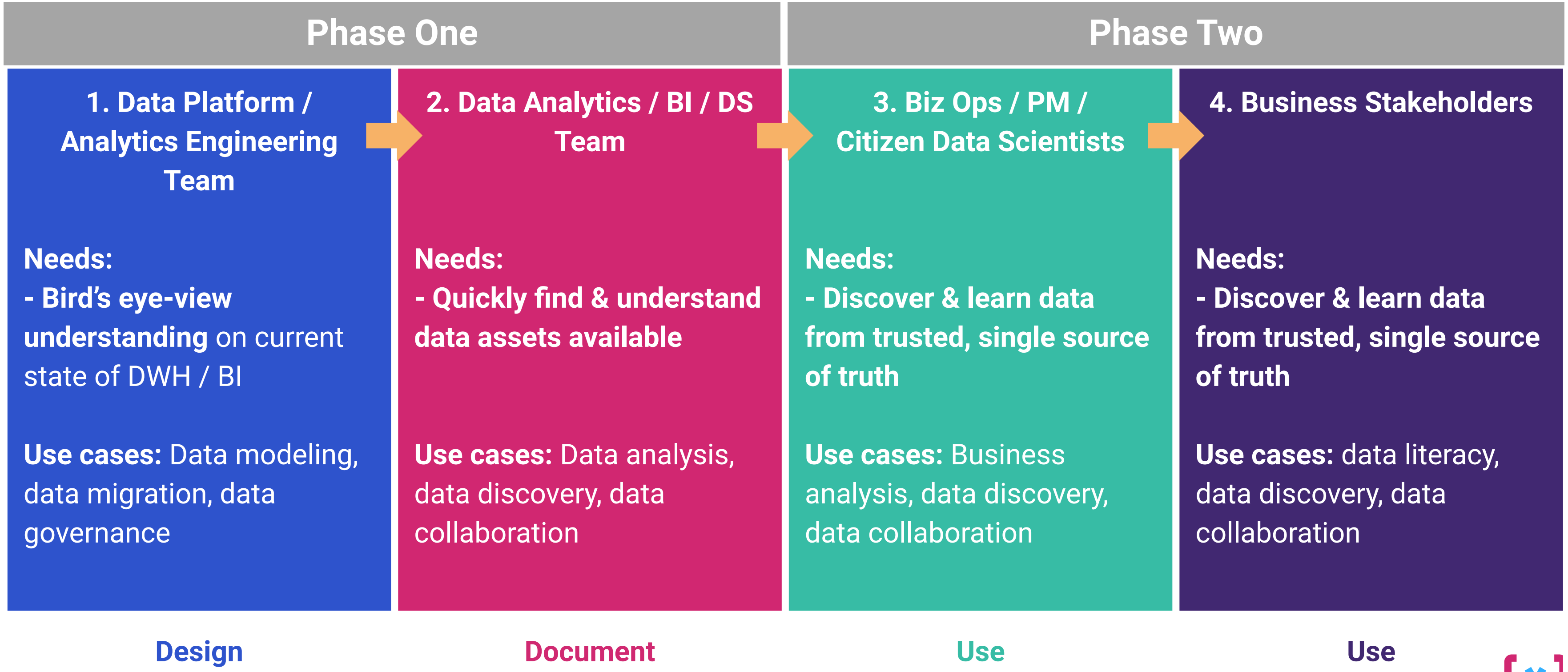
Milestones

How do we measure success?

Different objectives require involvement of different stakeholders - goals must be clear

Prevent Data Outages	Self-Service Data Analytics	Data Model Cleansing	Data Governance
Why? ● *Reduce downtime for key dashboards and reports* ● *Prevent data bottlenecks by proactively ensuring pipelines are up-to-date*	**Why?** ● *Empower more users to rely on data-driven decision making* ● *Reduce burden on data engineering teams*	**Why?** ● *Reduce storage and compute costs* ● *Ensure alignment across different business silos* ● *Track PII across your data model*	**Why?** ● *Improve data governance by understanding data consumption* ● *Implement data ownership for consistency and management*
How? ● *Quickly find dependencies if dashboards look "off" with automated lineage and impact analyses* ● *Integrate lineage into CI/CD pipeline via API to proactively prevent outages*	**How?** ● *Learn the data model quickly* ● *Find popular data assets, example queries, and how power users interact with data* ● *Curate your data with descriptions and tags to provide context to data consumers*	**How?** ● *Find unused tables and dashboards, or your most expensive queries* ● *Propagate tags based on data lineage* ● *Build docs and metrics tied directly to your data model*	**How?** ● *Understand who's accessed what and when* ● *Assign business and technical owners* ● *Compare lineage across multiple reports or analyses*
Who? ● *Data Engineers* ● *Data Analysts*	**Who?** ● *Data Engineers* ● *Data Analysts* ● *Business Stakeholders*	**Who?** ● *Data Engineers* ● *Data Analysts*	**Who?** ● *Data Engineers* ● *Data Analysts* ● *Business Stakeholders*

Data Discovery Starts with the Data Platform & BI Analytics Team working together



Phase One		Phase Two	
1. Data Platform / Analytics Engineering Team	2. Data Analytics / BI / DS Team	3. Biz Ops / PM / Citizen Data Scientists	4. Business Stakeholders
Needs: **- Bird's eye-view understanding** on current state of DWH / BI	**Needs:** **- Quickly find & understand data assets available**	**Needs:** **- Discover & learn data from trusted, single source of truth**	**Needs:** **- Discover & learn data from trusted, single source of truth**
Use cases: Data modeling, data migration, data governance	**Use cases:** Data analysis, data discovery, data collaboration	**Use cases:** Business analysis, data discovery, data collaboration	**Use cases:** data literacy, data discovery, data collaboration
Design	**Document**	**Use**	**Use**

Data Platform Team - Clean up data model and implement data catalog design

Tagging Structure

Define the tagging structure for the organization

- How will data be classified and grouped for end users to enable easy discovery and quick understanding?
- Unique to each organization
- Category vs Status

Data Deprecation

Remove obsolete or duplicate data to increase signal vs noise

- Check popularity and usage to understand what is no longer active or needed
- Review downstream dependencies to ensure deprecation will not have unintended consequences

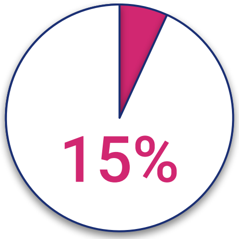
Assign Data Owners

Determine who is responsible for owning different data assets

- Owners can verify data documentation and validity
- Start with most popular data assets and look at top users as potential data owners

Documenting all of your data can be a herculean effort, so start with the most important / popular data assets

Data Usage



Across all tables synced to Select Star, only ~15% have been queried in the past 90 days

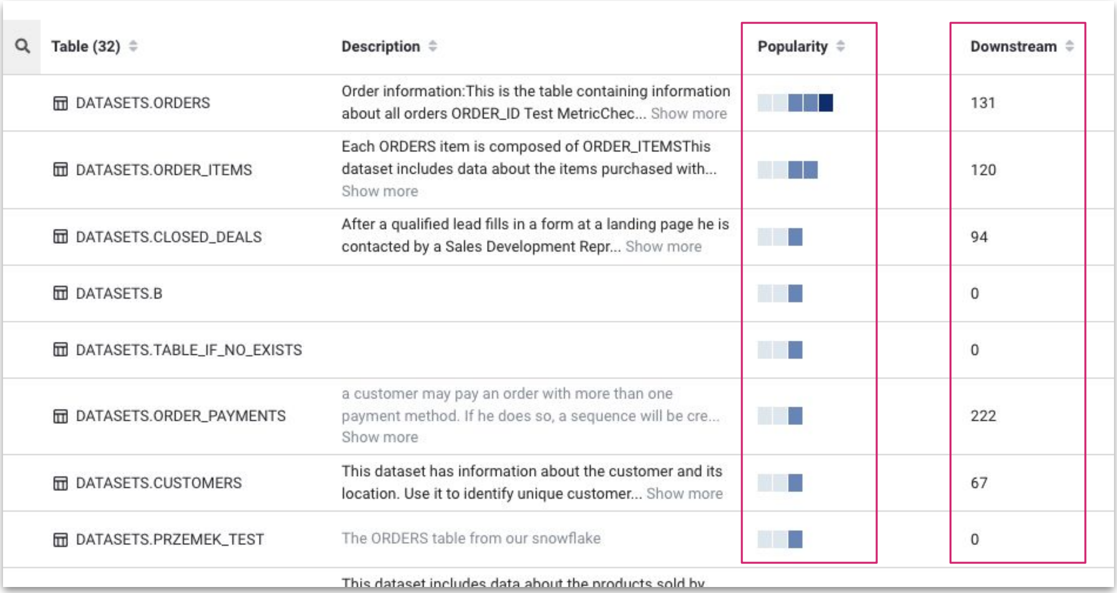


Table (32)	Description	Popularity	Downstream
DATASETS.ORDERS	Order information:This is the table containing information about all orders ORDER_ID Test MetricChec... Show more	▮▮▮	131
DATASETS.ORDER_ITEMS	Each ORDERS item is composed of ORDER_ITEMSThis dataset includes data about the items purchased with... Show more	▮▮	120
DATASETS.CLOSED_DEALS	After a qualified lead fills in a form at a landing page he is contacted by a Sales Development Repr... Show more	▮	94
DATASETS.B		▮	0
DATASETS.TABLE_IF_NO_EXISTS		▮	0
DATASETS.ORDER_PAYMENTS	a customer may pay an order with more than one payment method. If he does so, a sequence will be cre... Show more	▮	222
DATASETS.CUSTOMERS	This dataset has information about the customer and its location. Use it to identify unique customer... Show more	▮	67
DATASETS.PRZEMEK_TEST	The ORDERS table from our snowflake	▮	0
	This dataset includes data about the products sold by		

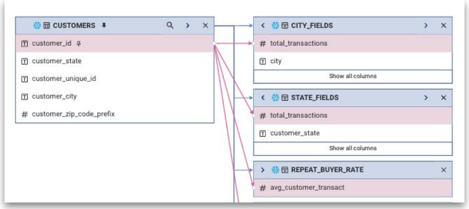
Analytics & Data Science - leverage power users to build out documentation

Data Discovery Platforms reduce the time required to curate data through automation and crowdsourcing

Automated Insights

Lineage, Top Users, Popularity, Entity Relationship Diagrams, Popular Queries & Joins

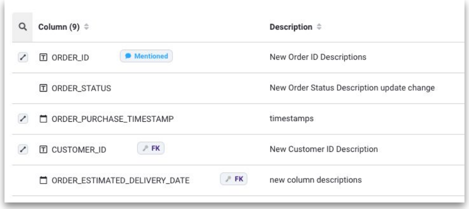
Data Discovery Platforms should do this automatically upon setup



Additional Curation

Add-in missing descriptions your data dictionary, assign tags to create curated data sets

Data Discovery Platforms should help speed this process up (i.e. tag / description propagation)

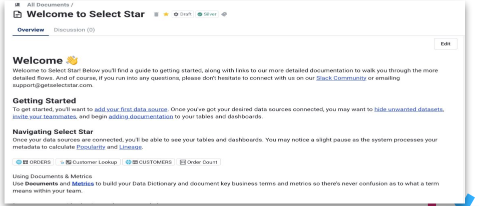


Column (9)		Description
ORDER_ID	Mentioned	New Order ID Descriptions
ORDER_STATUS		New Order Status Description update change
ORDER_PURCHASE_TIMESTAMP		timestamps
CUSTOMER_ID	FK	New Customer ID Description
ORDER_ESTIMATED_DELIVERY_DATE	FK	new column descriptions

Semantic Layer

Business glossaries, FAQs, or other additional documentation that helps users grasp the data model

Data Discovery Platforms should link this information directly to your data assets



All Documents /

Welcome to Select Star

Draft Silver

Overview Discussion (0)

Edit

Welcome 👋

Welcome to Select Star! Below you'll find a guide to getting started, along with links to our more detailed documentation to walk you through the more detailed flows. And of course, if you run into any questions, please don't hesitate to connect with us on our Slack Community or emailing support@getselectstar.com.

Getting Started

To get started, you'll want to add your first data source. Once you've got your desired data sources connected, you may want to hide unwanted datasets, invite your teammates, and begin adding documentation to your tables and dashboards.

Navigating Select Star

Once your data sources are connected, you'll be able to see your tables and dashboards. You may notice a slight pause as the system processes your metadata to calculate Popularity and Lineage.

ORDERS Customer Lookup CUSTOMERS Order Count

Using Documents & Metrics

Use **Documents** and **Metrics** to build your Data Dictionary and document key business terms and metrics so there's never confusion as to what a term means within your team.

Prepare for rollout to data consumers and broader organization

1 Develop “Rules of Engagement”

Establish best practices for data documentation and common questions

Example Rules

- Where are descriptions updated, in dbt or directly in the Data Discovery Platform?
- What do I do if I disagree with a description for a particular table or column?
- When can new tags be created?
- How do I add/change/remove a tag on a data asset if I am not data manager?
- Who do I contact if I don’t have access to the data I need?

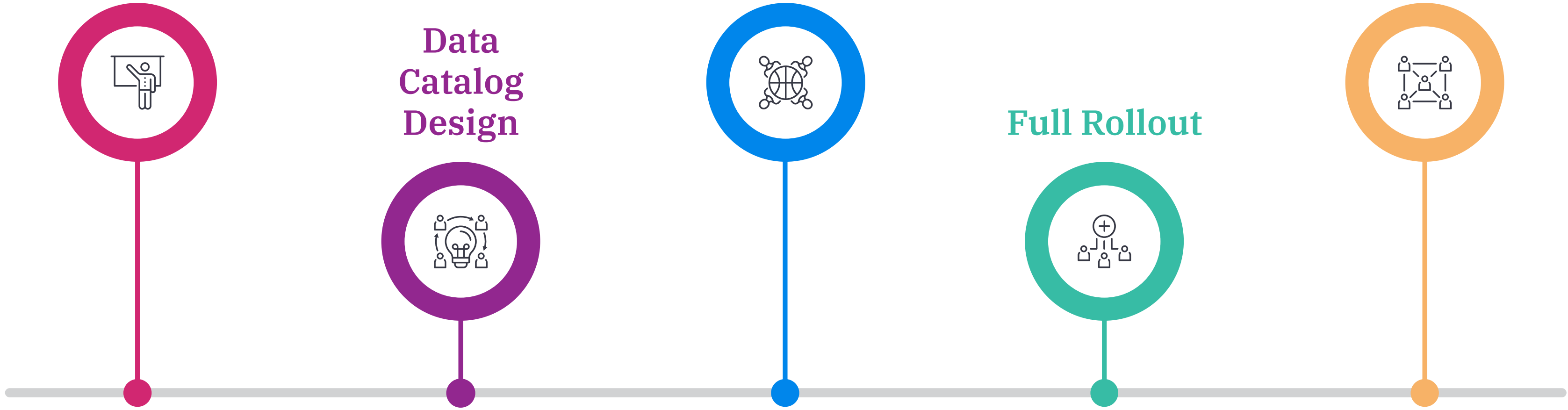
2 Prepare Onboarding Materials

Orient users to the platform and how it can be leveraged for their work

Example Materials

- 30 minute product demo
- How-to guide highlighting common workflows
- Data Trivia
  - Example questions that can be answered via the Data Discovery platform

Set dates for achieving key milestones and track progress



Platform Setup

1-2 Days

Data sources connected

Insights generated (e.g. Lineage, Popularity, Etc.)

Data Catalog Design

1-2 Weeks

Data team onboarded

Tagging structure defined

Owners Assigned

Data Curation

Varies

Documentation for popular / important data assets

Curated Data Sets Created

Training curriculum developed

Full Rollout

Data Consumers onboarded

Data Discovery platform worked into daily workflows

Usage Tracking

% of data assets documented

% of data assets with owners

Data Discovery Platform Usage

III.

Data Discovery Best Practices

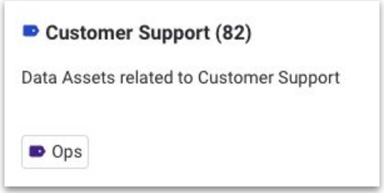
Tips and Tricks

Best Practices - Tags

Use Tags to build collections of curated data assets for end users to understand who it's for and how it should be used

Category Tags

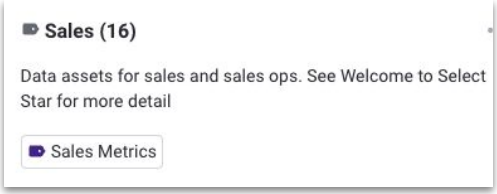
- 1st level - key domains (e.g. teams or products)
    - 2nd level - sub domains
    - Can have different sub domains for each tag



Customer Support (82)

Data Assets related to Customer Support

Ops



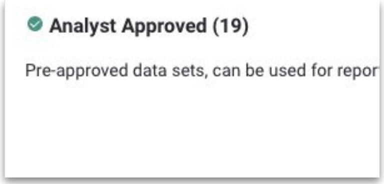
Sales (16)

Data assets for sales and sales ops. See Welcome to Select Star for more detail

Sales Metrics

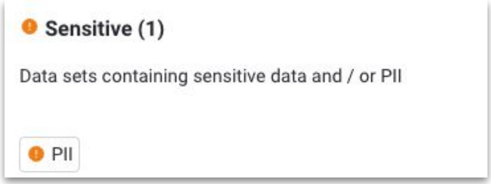
Status Tags

- PII / Sensitive
- To Be Deprecated
- Gold / Silver / Bronze Status
- Approved for Reporting
- ....



Analyst Approved (19)

Pre-approved data sets, can be used for repor



Sensitive (1)

Data sets containing sensitive data and / or PII

PII

Best Practices - Docs & Metrics

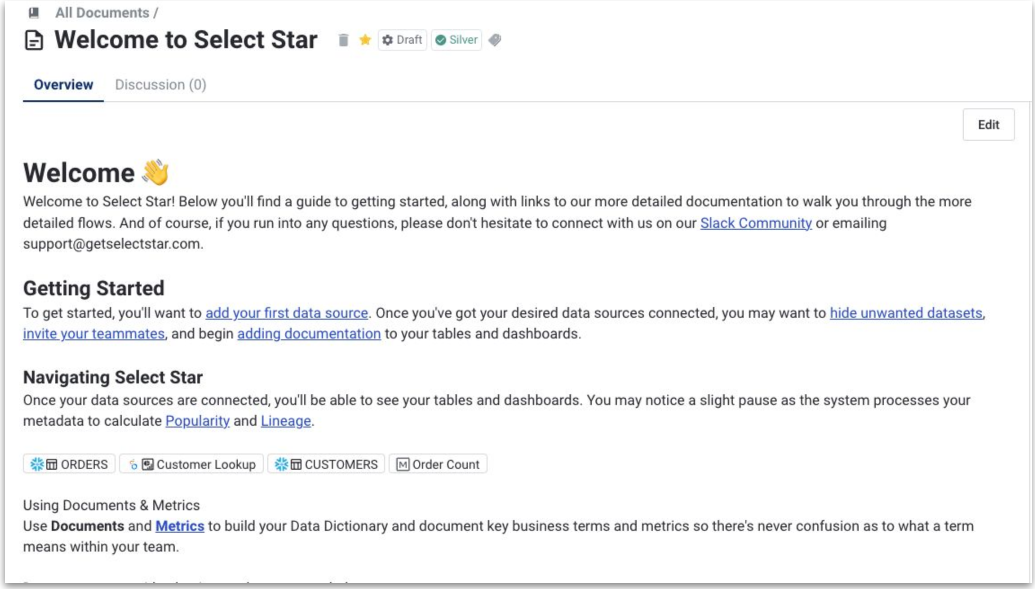
Docs and Metrics should be created to provide additional context and increase alignment on north start KPIs

Example Docs

- Onboarding guide
- General Business Glossary
- Change logs
- Additional info about specific data assets

Metrics

- Definition - *What is this metrics?*
- Business Question - *Why is it important?*
- Calculation - *How do we calculate it?*
  - Include SQL where relevant
- Represented As - *Where do I find this metric?*
- Dimensions - *How should I analyze it?*



Welcome 👋

Welcome to Select Star! Below you'll find a guide to getting started, along with links to our more detailed documentation to walk you through the more detailed flows. And of course, if you run into any questions, please don't hesitate to connect with us on our Slack Community or emailing support@getselectstar.com.

Getting Started

To get started, you'll want to add your first data source. Once you've got your desired data sources connected, you may want to hide unwanted datasets, invite your teammates, and begin adding documentation to your tables and dashboards.

Navigating Select Star

Once your data sources are connected, you'll be able to see your tables and dashboards. You may notice a slight pause as the system processes your metadata to calculate Popularity and Lineage.

Using Documents & Metrics

Use **Documents** and **Metrics** to build your Data Dictionary and document key business terms and metrics so there's never confusion as to what a term means within your team.

Best Practices - User Engagement

Provide training and processes to drive engagement and get the most out of your investment in Data Discovery

Training

Train users to ensure they can navigate the platform and understand when & how it should be used

- Onboarding sessions
- Data Trivia
- Rules of Engagement
- Q&A Sessions

Using

Build Data Discovery into daily workflows

- Share links to relevant pages in the platform when answering questions
- Facilitate Q&A within the platform
- Leverage integrations with existing platforms (e.g. Slack, Chrome extensions)
- Highlight success stories (e.g. I solved this question on my own with the platform)

Tracking

Track usage to measure adoption and impact

- # of active users
- Documentation fill rate
- Discussion items

Thank You

To learn more about Select Star and Data Discovery, check us out at **www.selectstar.com** and or reach out to sales@getselectstar.com for a demo