



Starburst Workshop: SQL data pipelines & products

Starburst Academy
Data Universe 2024

Workshop objectives

- Present the reference architecture spanning multiple zones
 - Land
 - Structure
 - Consume
- Build a SQL-based pipeline that spans the reference architecture's zones
- Create and secure granular data products for your downstream consumers

For today...

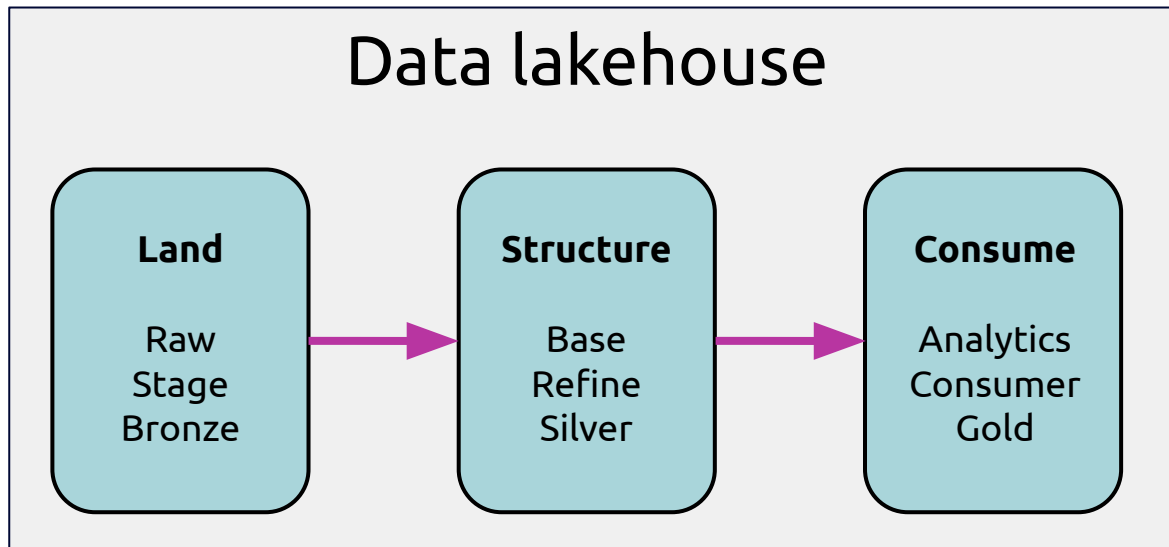
- POIs
 - The lab guide has it ALL!
 - The S3 credentials will expire after the weekend
 - Copy/n/paste in bulk with accompanying **SQL.txt** file
- Approach
 - I'm going to perform the labs myself & you can...
 - do them along with me
 - or just watch (and maybe do later)
 - We ALL come from DIFFERENT experiences & backgrounds, so...
 - **ASK QUESTIONS**

Data pipelines

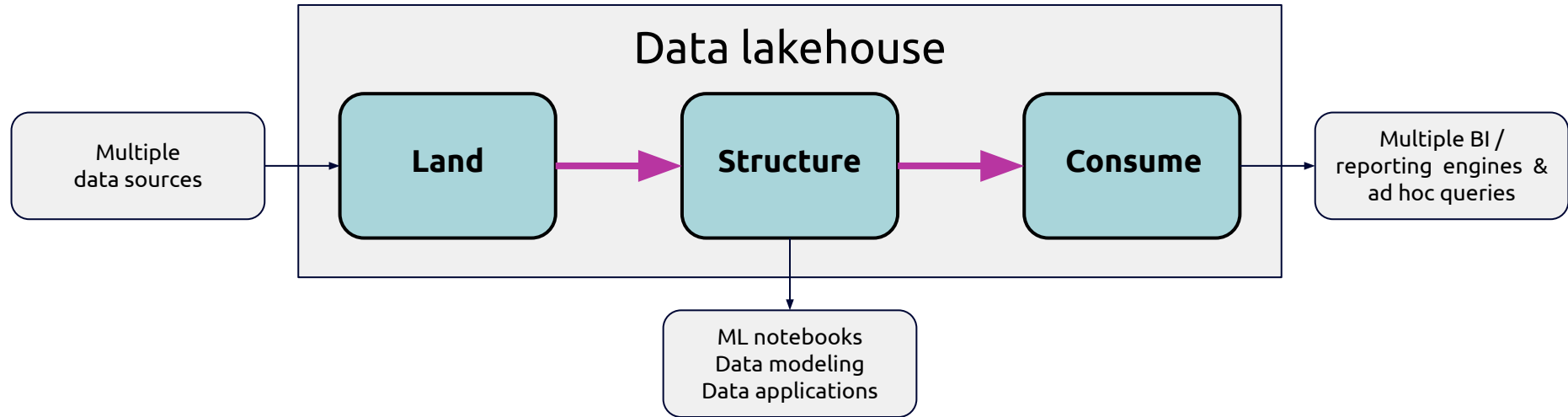
Reference architecture

Reference architecture

The reference architecture centers around the data lakehouse and how we classify our data assets into distinct zones. Data pipelines populate the zones.



Activities across the architecture

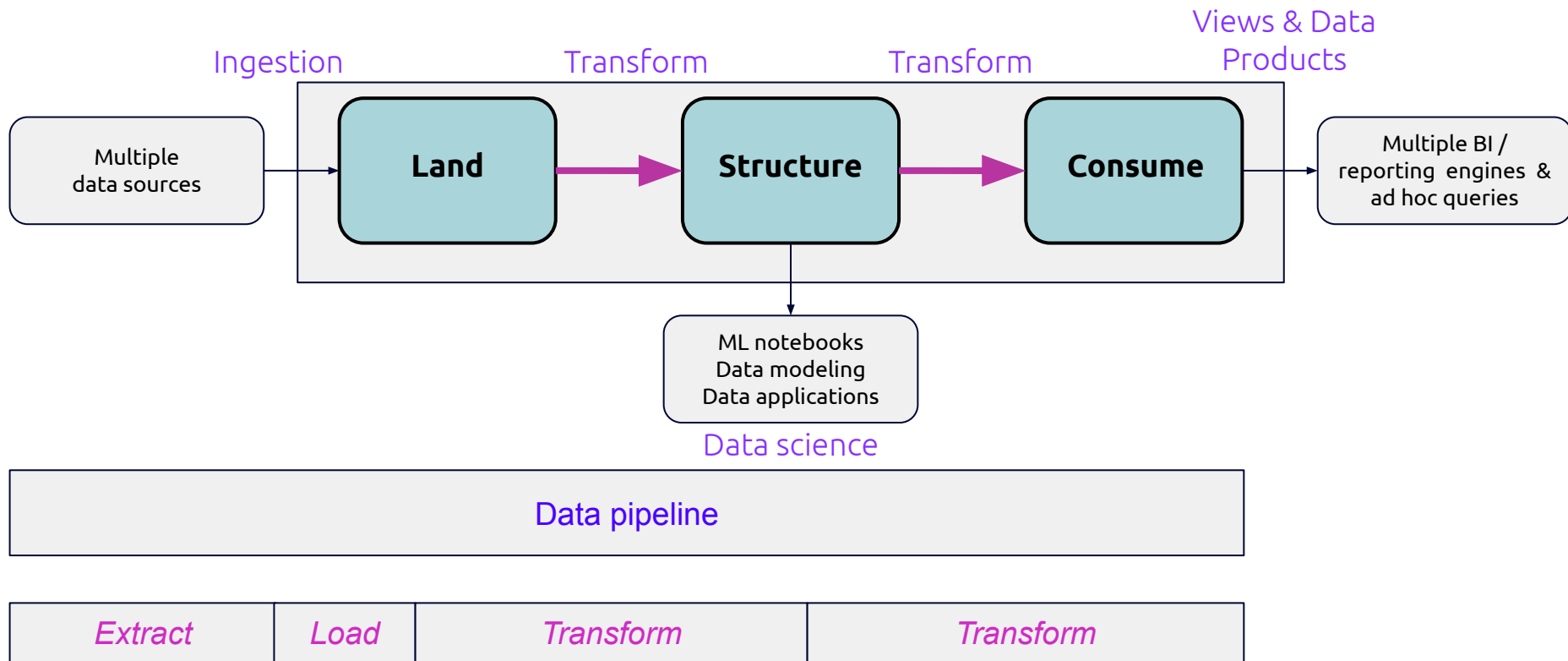


The diagram illustrates a data engineering process flow, structured into three main stages: **Land**, **Structure**, and **Consume**, all contained within a central box labeled **Data Products**.

- Land:** This stage receives input from **Multiple data sources** (represented by logos like ORACLE, SAP, APP, and a Twitter bird). It involves data formats like **CSV** and **JSON** (indicated by a document icon). A yellow bee icon labeled **HIVE** is also present.
- Structure:** This stage involves **ML notebooks**, **Data modeling**, and **Data applications** (represented by a box with these labels). It also includes logos for **Jupyter** and a blue document icon labeled **APP**.
- Consume:** This stage outputs to **Multiple BI / reporting engines & ad hoc queries** (represented by a box with these labels). It includes logos for various BI tools like Tableau, Power BI, and Qlik.

Arrows indicate the flow of data from **Land** to **Structure** to **Consume**. Dashed blue arrows point from the **Multiple data sources** box to the **Land** stage, and from the **Multiple BI / reporting engines & ad hoc queries** box to the **Consume** stage. A solid black arrow points from the **Structure** stage down to the **ML notebooks** box.

Summary



Data products

Data consumers & producers

Qualities of data products



Discoverable: end users and other domains need to be able to discover and access a given data product



Self-describing: any end user outside the domain which produces the data product should have all of the information they require to use the data



Addressable: the data should have a straightforward and documented way of being programmatically accessed, e.g. via SQL



Interoperable: governance should ensure that the data complies to any inter- or intra-domain standards or regulations, so the end user can confidently use the data without concern



Trustworthy: end users should be able to understand the level of data quality and ideally view the provenance (lineage) of the data so they can be confident in any analyses using the data product



Secure: data products should fold any authorization into the access control provided by the data mesh experience plane, which is where data product consumption occurs

Starburst Data Products

Data
producers



Starburst

Dashboard

Publish data

Sort Filter by domain Filter by tag Search

Data Mesh Generation 100K+ results of dataset and leads generated from Data Mesh webinar PENDING CHANGES CREATED BY: JOHN SMITH	Sales Q4 Projection Datasets for sales project for Q4 DRAFT CREATED BY: COMMANDER SUN BUN	ECommerce Product Datasets 100K+ results from public data sets from January 17 - Aug 19, 2021. CREATED BY: KAREN MILLER	Analytic Dataset Anonymized, non-aggregated granular consumer level data across all asset classes CREATED BY: KAREN MILLER
Media Consumption Monthly TV viewership insights spanning network, genre, daypart and ad exposure CREATED BY: KAREN MILLER	Demographics Accurate & scalable demographic data covering age, gender, generational & multi-culture CREATED BY: JOHN SMITH	Contextual Intent Industry-leading contextual intent data for B2B intelligence CREATED BY: JOHN SMITH	Sales & Home Services Data Exclusive data for appliances sold at Sears/Kmart and the connected Sears Home Services data. CREATED BY: JOHN SMITH
Private Equity Private equity firm profiles including investment professionals, limited partners, and more CREATED BY: COMMANDER SUN BUN	Data on Demand Gain actionable customer insights for increased sales and conversions CREATED BY: JOHN SMITH	Executive & Profession Contacts Connect with key decision makers to save time and money CREATED BY: JOHN SMITH	R&D Data Atlas 180+ public R&D-related datasets from the UN, OECD, Eurostat, EC and other authoritative sources DRAFT CREATED BY: JOHN SMITH

Data
consumers



Discover, create, publish, and manage data products based on multiple datasets

The screenshot shows the 'Data product details' page for 'Weather data' in the Starburst interface. The page is divided into several sections: Overview, Usage examples, and Query data. The Overview section includes a summary, a table of query statistics, a description, and a list of datasets. The right sidebar contains information about data product owners, tags, domain, relevant links, and details.

Data product details

Overview Usage examples Query data

Weather data

BOOKMARKED 21 ★★★★★ 2

Overview

Summary
This is where the short description will go. This description will also appear on the Dashboard Card. This should have a character limit.

Number of queries	PAST 7 DAYS	PAST 30 DAYS	ALL TIME	% of queries run against this data product	% of users
	14	93	244	12%	7

Description
This is where the longer summary would go. There is no character limit here. Mauris ante nisl, tincidunt aliquet mauris eu, cursus malesuada nulla. Aliquam consectetur lobortis tellus, sit amet venenatis magna. Cursum malesuada nulla. Aliquam consectetur lobortis tellus, sit amet venenatis magna.

Datasets

Catalog: Catalog 2

Showing 3 of 3 datasets

Hourly weather	This is where the short description will go. This should have a character limit.	✓ VIEWS
Daily weather	This is where the short description will go. This should have a character limit.	✓ MATERIALIZED VIEWS
Monthly weather	This is where the short description will go. This should have a character limit.	

Data product owners

Lauren Moore
lauren.moore@starburstdata.com

Karen Miller
karen.miller@starburstdata.com

Tags

Tag This is a tag This is another tag

Yet another tag here

Domain

This is where the domain will show

Relevant links

Confluence
<https://starburstdata.atlassian.net/wiki/>

Google Doc
<https://docs.google.com/spreadsheets/>

Details

Created on: March 17, 2021
Created by: Lauren Moore
Last updated on: March 17, 2021
Last updated by: Karen Miller
Last time used: August 24, 2021
Last used by: Vishal Singh

[View in BI tools](#)

Allow data owners
& data engineers to
**define relevant
metadata** to data
consumers

Starburst Enterprise

?

Vishal

<>

Provide column description for Dataset detail: customer_informations

You can fill in Column description below to provide better context. This step is optional.

Column name	Data type	Column description(optional)
custkey	bigint	Customer key
name	varchar(25)	Name
address	varchar(40)	Address
nationkey	bigint	Country the customer belong to
phone	varchar(15)	Phone number
acctbal	double	Account
mktsegment	varchar(10)	Segment
comment	varchar(117)	Comment

Preview

< Back

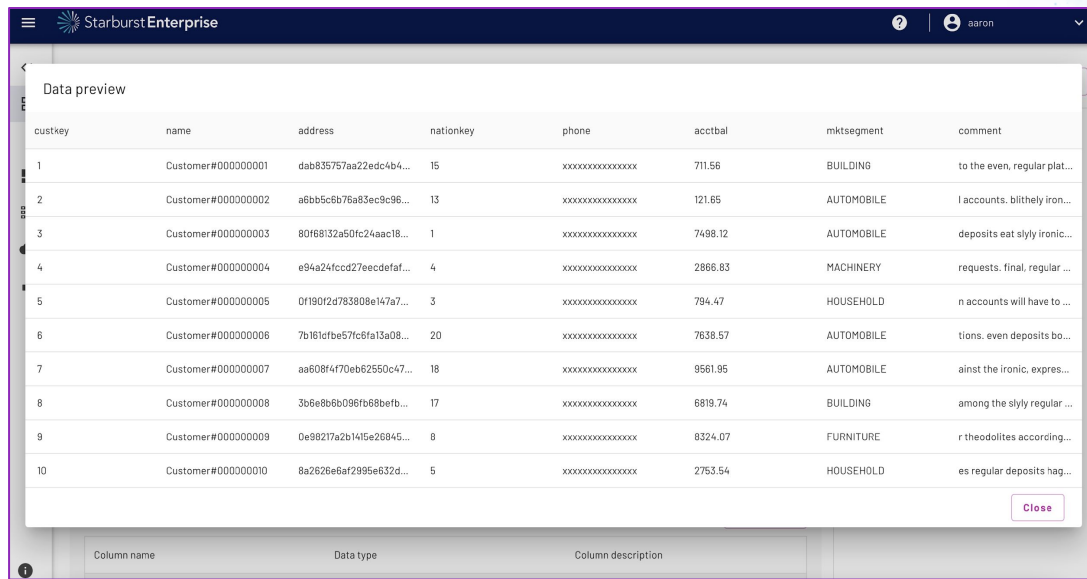
+ Add another dataset

Cancel

Save as draft

Save and continue

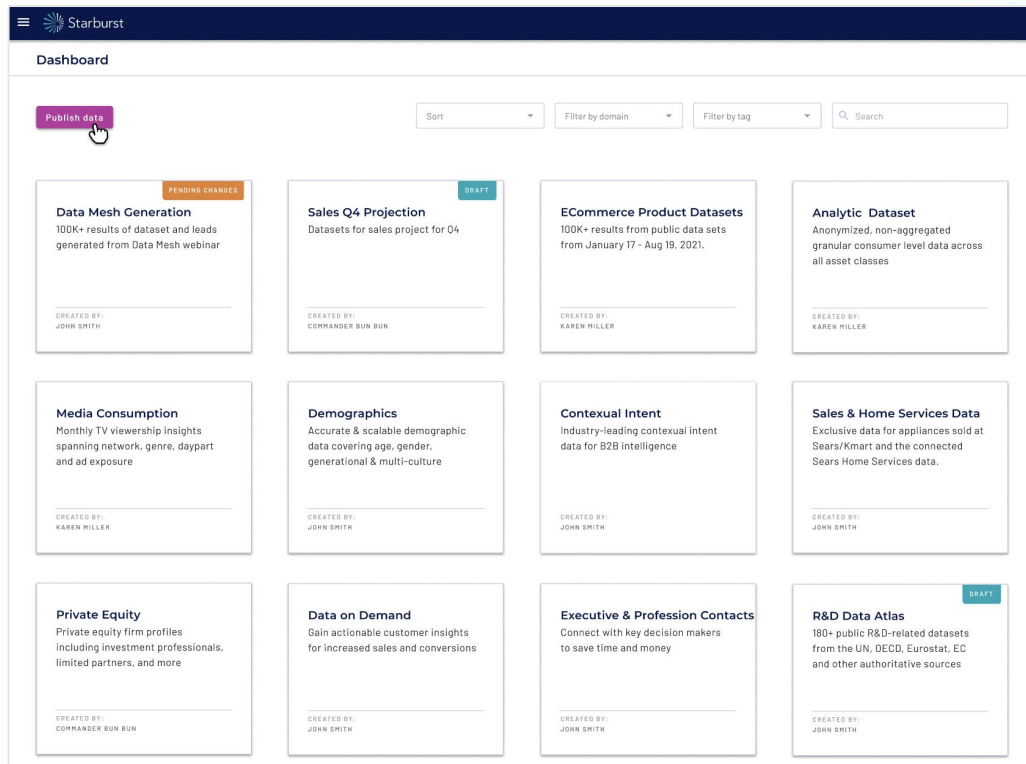
Secure your data products with **access control**, ensuring consistent **governance** from the source level



The image shows a screenshot of the Starburst Enterprise interface. At the top, there is a dark blue header with the Starburst logo and the text "StarburstEnterprise". On the right side of the header, there is a user profile icon labeled "aaron" and a question mark icon. Below the header, a "Data preview" window is open, displaying a table with 10 rows and 8 columns. The columns are labeled: custkey, name, address, nationkey, phone, acctbal, mktsegment, and comment. The data is presented in a clean, white table with alternating row colors. A "Close" button is located at the bottom right of the preview window. Below the preview window, a table structure is visible with columns for "Column name", "Data type", and "Column description".

custkey	name	address	nationkey	phone	acctbal	mktsegment	comment
1	Customer#000300001	dab835757aa22edc4b4...	15	xxxxxxxxxxxxxxxx	711.56	BUILDING	to the even, regular plat...
2	Customer#000300002	a6bb5c6b7ba83ec9c96...	13	xxxxxxxxxxxxxxxx	121.65	AUTOMOBILE	I accounts, blithely iron...
3	Customer#000300003	80f68132a50fc24aac18...	1	xxxxxxxxxxxxxxxx	7498.12	AUTOMOBILE	deposits eat slyly ironic...
4	Customer#000300004	e94a24fccd27eecdefaf...	4	xxxxxxxxxxxxxxxx	2866.83	MACHINERY	requests, final, regular ...
5	Customer#000300005	0f190f2d783808e147a7...	3	xxxxxxxxxxxxxxxx	794.47	HOUSEHOLD	n accounts will have to ...
6	Customer#000300006	7b161cfbe57fc6fa13a08...	20	xxxxxxxxxxxxxxxx	7638.57	AUTOMOBILE	tions, even deposits bo...
7	Customer#000300007	aa608f470eb62550c47...	18	xxxxxxxxxxxxxxxx	9561.95	AUTOMOBILE	ainst the ironic, expres...
8	Customer#000300008	3b6e8b6b096f68b6fb...	17	xxxxxxxxxxxxxxxx	6819.74	BUILDING	among the slyly regular ...
9	Customer#000300009	0e98217a2b1415e26845...	8	xxxxxxxxxxxxxxxx	8324.07	FURNITURE	r theodolites according...
10	Customer#000300010	8a2626e6af2995e632d...	5	xxxxxxxxxxxxxxxx	2753.54	HOUSEHOLD	es regular deposits hag...

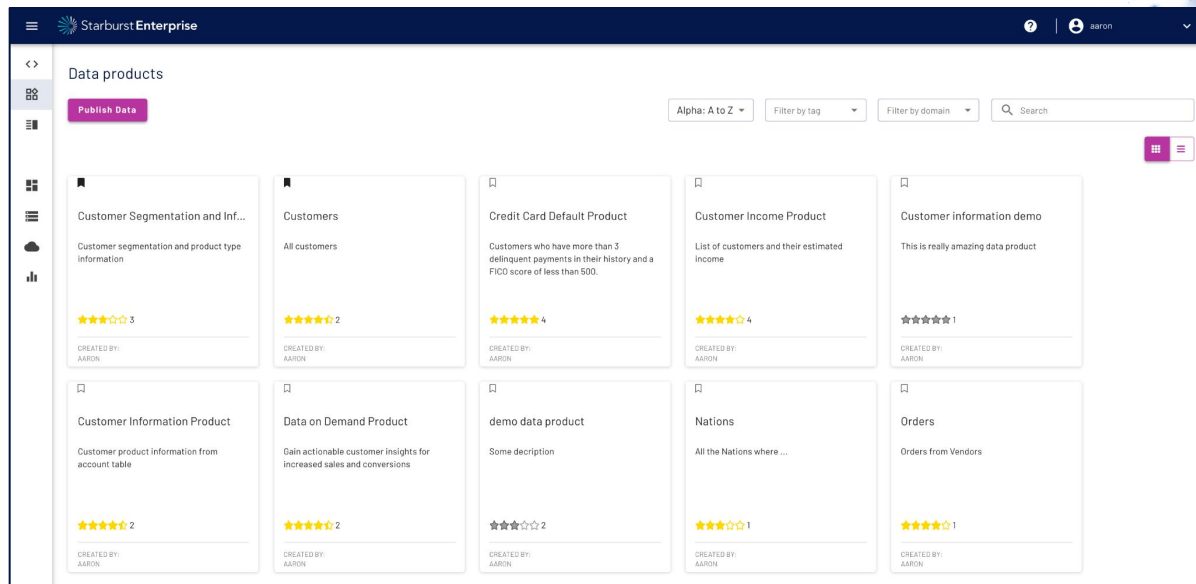
Query data products
that are trusted and
approved for frequent
business use



The screenshot displays the Starburst Dashboard interface. At the top, there is a dark blue header with the Starburst logo and a hamburger menu icon. Below the header, the word "Dashboard" is centered. A navigation bar contains a "Publish data" button with a cursor icon, and three filters: "Sort", "Filter by domain", and "Filter by tag", each with a dropdown arrow. A search bar is located on the right. The main content area is a grid of 12 data product cards, arranged in 3 rows and 4 columns. Each card has a title, a brief description, and a "CREATED BY:" field with the creator's name. Some cards have status tags: "PENDING CHANGES" (orange) for "Data Mesh Generation", "DRAFT" (teal) for "Sales Q4 Projection", "ECommerce Product Datasets", and "R&D Data Atlas".

Product Name	Description	Status	Created By
Data Mesh Generation	100K+ results of dataset and leads generated from Data Mesh webinar	PENDING CHANGES	JOHN SMITH
Sales Q4 Projection	Datasets for sales project for Q4	DRAFT	COMMANDER SUN SUN
ECommerce Product Datasets	100K+ results from public data sets from January 17 - Aug 19, 2021.	DRAFT	KAREN MILLER
Analytic Dataset	Anonymized, non-aggregated granular consumer level data across all asset classes		KAREN MILLER
Media Consumption	Monthly TV viewership insights spanning network, genre, daypart and ad exposure		KAREN MILLER
Demographics	Accurate & scalable demographic data covering age, gender, generational & multi-culture		JOHN SMITH
Contextual Intent	Industry-leading contextual intent data for B2B intelligence		JOHN SMITH
Sales & Home Services Data	Exclusive data for appliances sold at Sears/Kmart and the connected Sears Home Services data.		JOHN SMITH
Private Equity	Private equity firm profiles including investment professionals, limited partners, and more		COMMANDER SUN SUN
Data on Demand	Gain actionable customer insights for increased sales and conversions		JOHN SMITH
Executive & Profession Contacts	Connect with key decision makers to save time and money		JOHN SMITH
R&D Data Atlas	180+ public R&D-related datasets from the UN, OECD, Eurostat, EC and other authoritative sources	DRAFT	JOHN SMITH

**Share and rate your
data products**
internally and track
usage





HANDS-ON LABS!!

Starburst Academy