Starburst

Starburst Workshop:

Python data pipelines & products

Subtitle Goes Here

Starburst Academy
Data Universe 2024

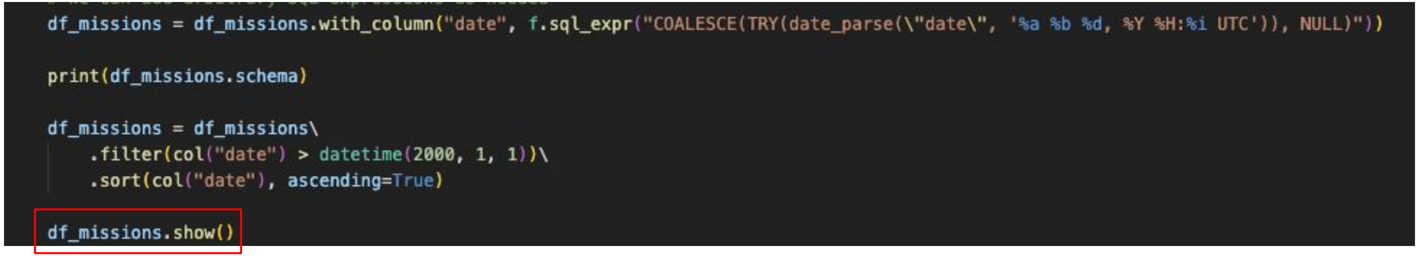
Workshop objectives

- Present the reference architecture spanning multiple zones
  - Land
  - Structure
  - Consume
- Build a Python-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
  - All code is in the Jupyter notebook – `Workshop4.ipynb`
- 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**

PyStarburst Overview

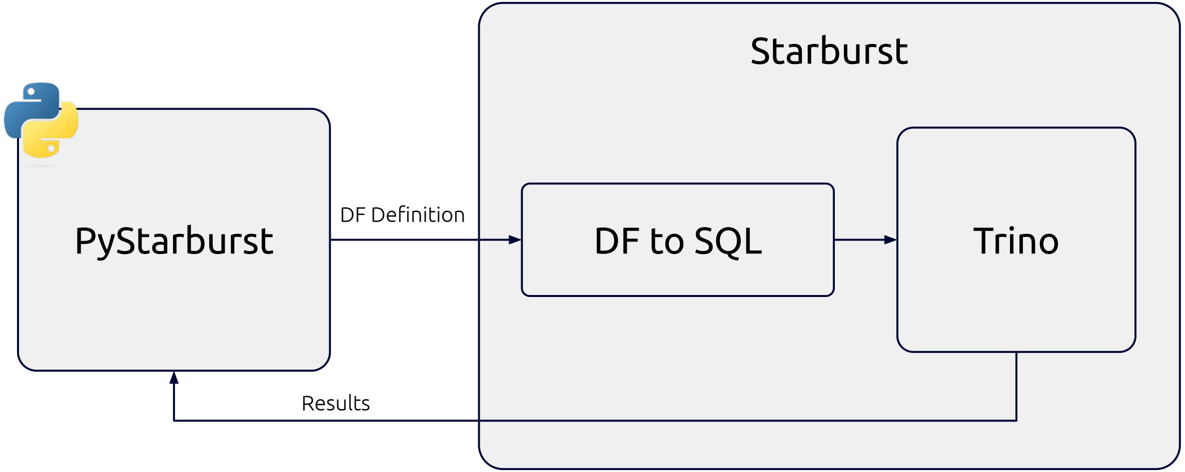


```
df_missions = df_missions.with_column("date", f.sql_expr("COALESCE(TRY(date_parse(\"date\", '%a %b %d, %Y %H:%i UTC')), NULL)"))

print(df_missions.schema)

df_missions = df_missions\
    .filter(col("date") > datetime(2000, 1, 1))\
    .sort(col("date"), ascending=True)

df_missions.show()
```



- PySpark-like Syntax
- Lazy execution
- Python gets converted to SQL
- Heavy lifting done by Trino

Links: API Docs & Example Code

Currently supported on Starburst Galaxy and Starburst Enterprise.

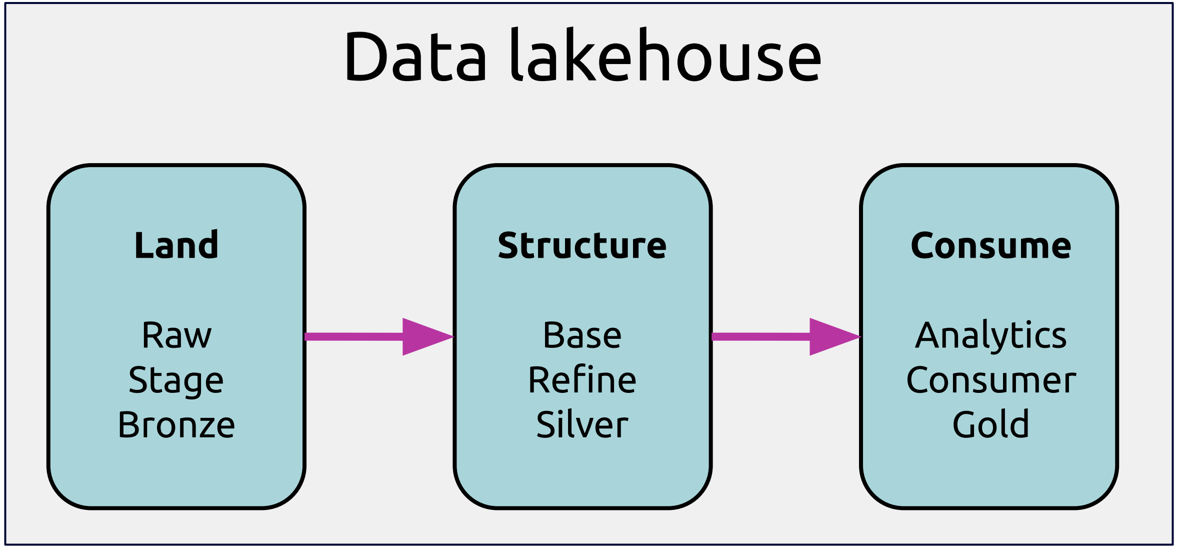
Data pipelines

Reference architecture

Starburst

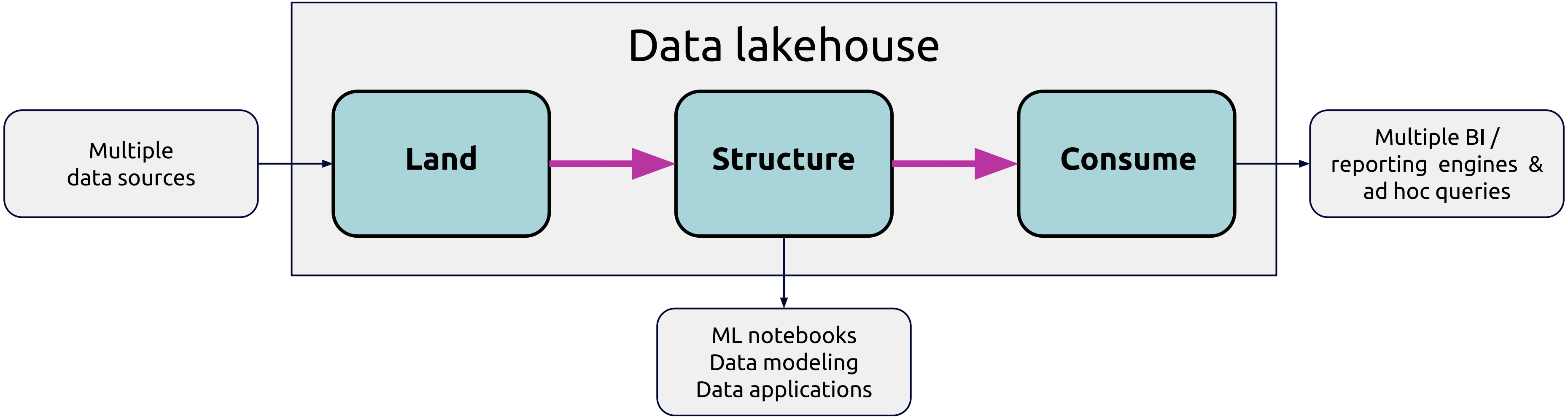
Reference architecture

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



Starburst

Activities across the architecture



Data lakehouse

Multiple data sources

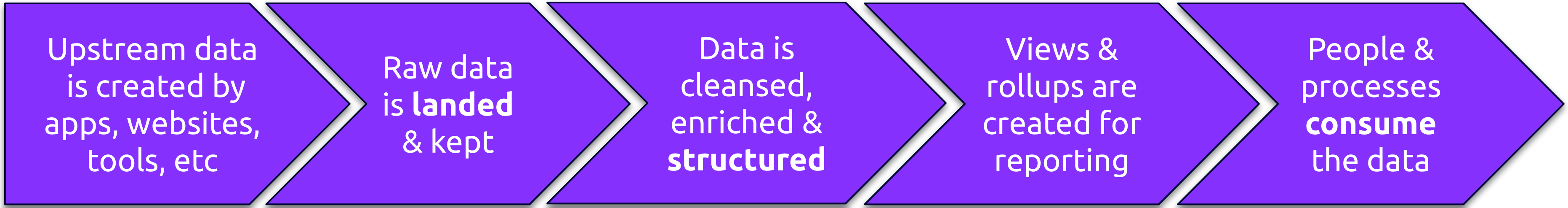
Land

Structure

Consume

Multiple BI / reporting engines & ad hoc queries

ML notebooks
Data modeling
Data applications



Upstream data is created by apps, websites, tools, etc

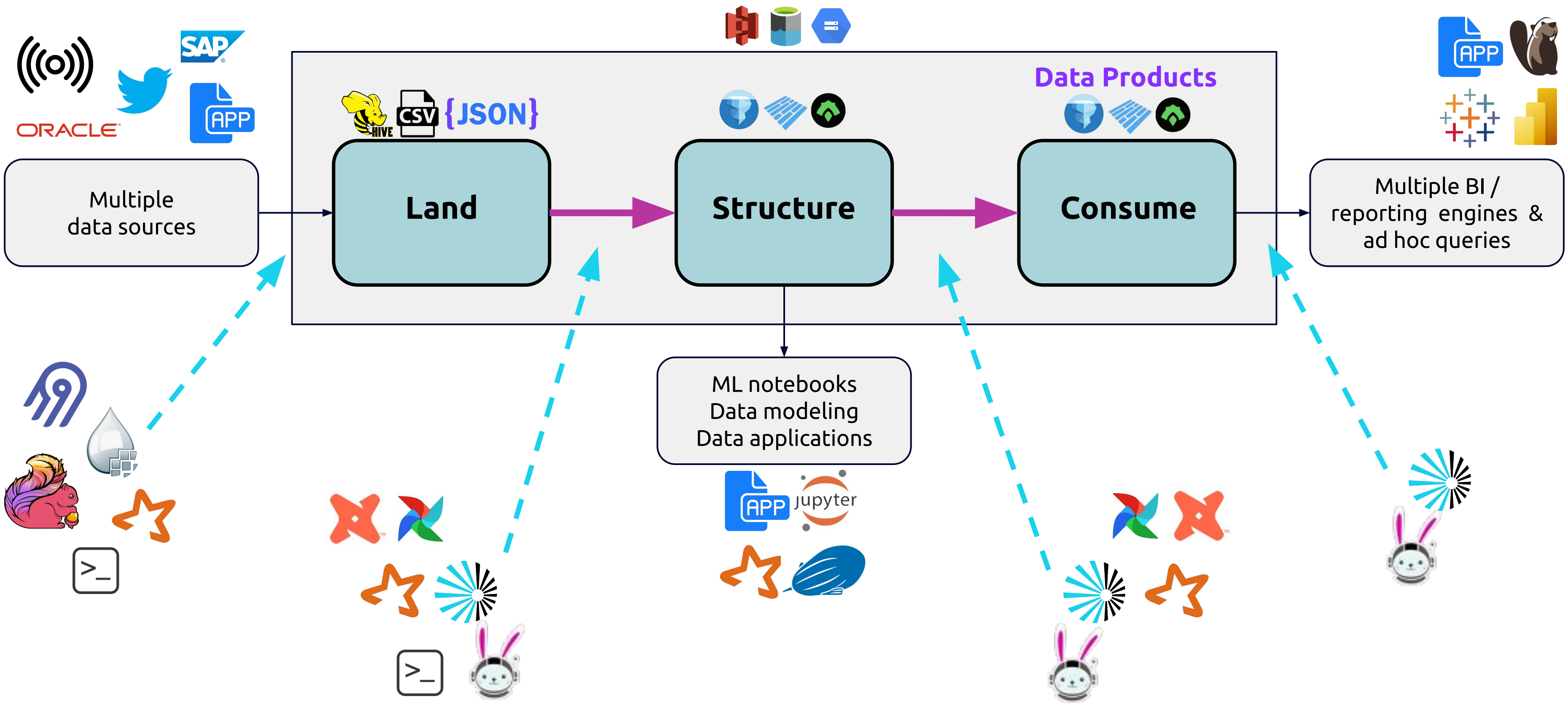
Raw data is **landed** & kept

Data is cleansed, enriched & **structured**

Views & rollups are created for reporting

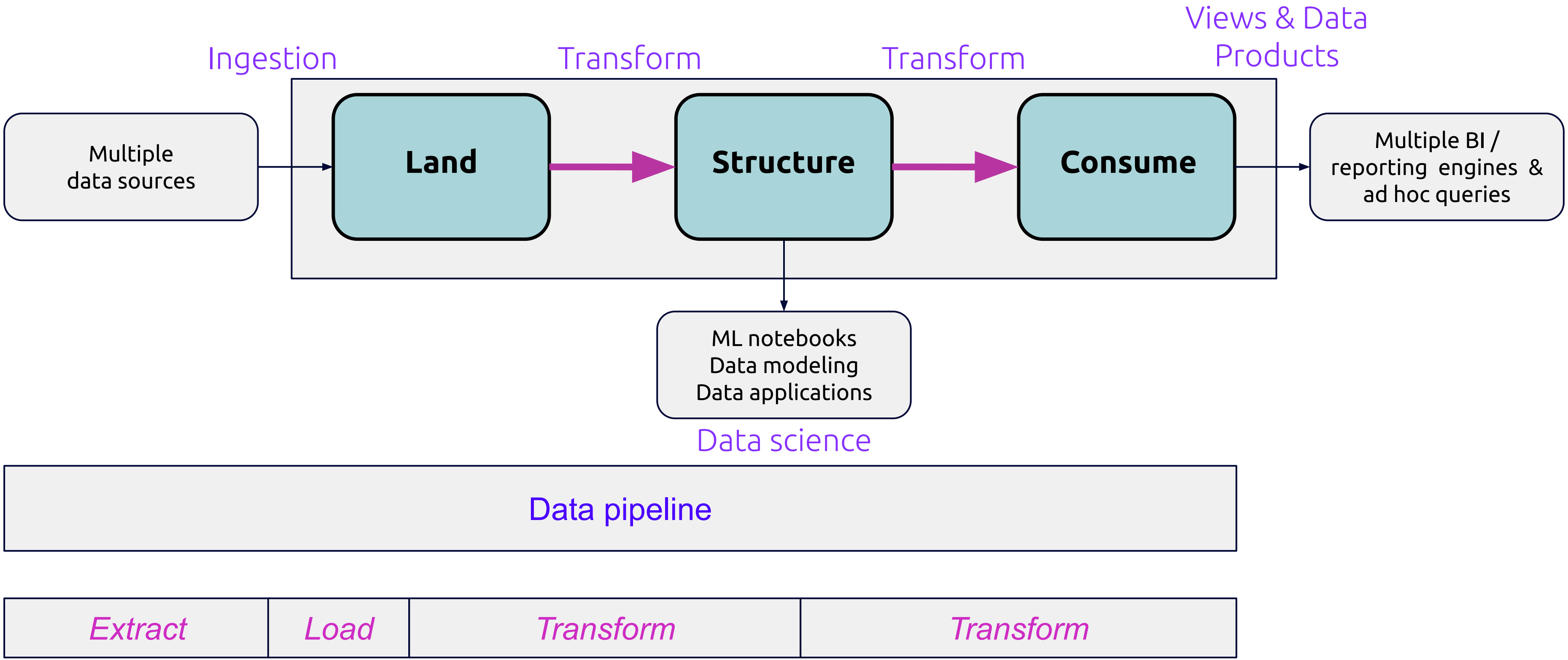
People & processes **consume** the data

Tools & technologies



Starburst

Summary



Starburst

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

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

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

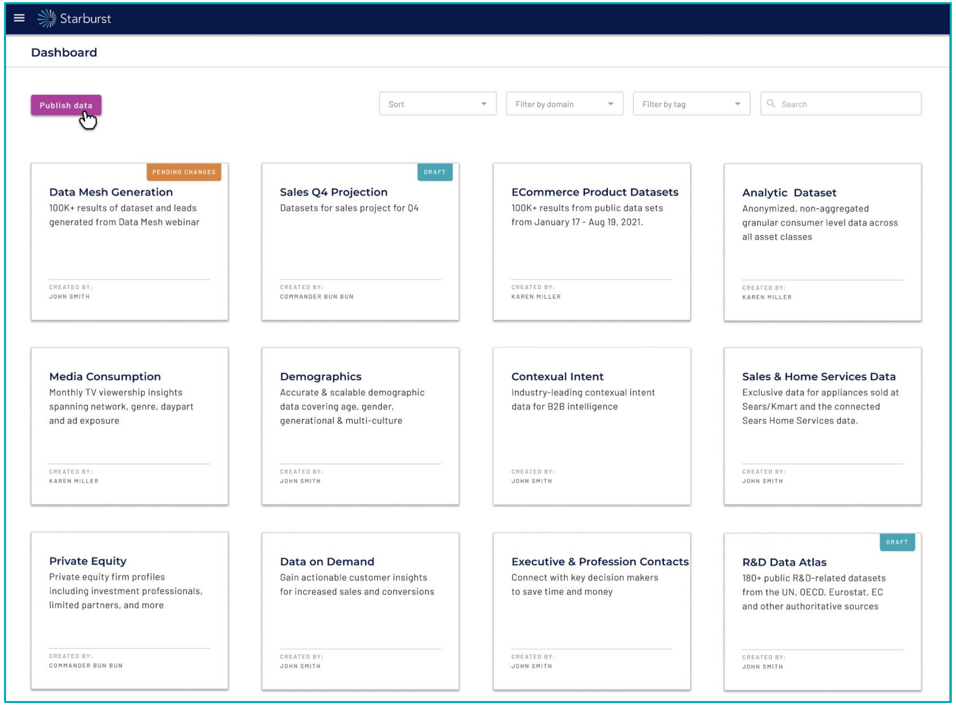
Self-describing: any end user outside the domain which produces the data product should have all of the information they require to use the data

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

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



PENDING CHANGES
Data Mesh Generation
100K+ results of dataset and leads generated from Data Mesh webinar
CREATED BY:
JOHN SMITH

DRAFT
Sales Q4 Projection
Datasets for sales project for Q4
CREATED BY:
COMMANDER BUN 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

Contexual 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 BUN BUN

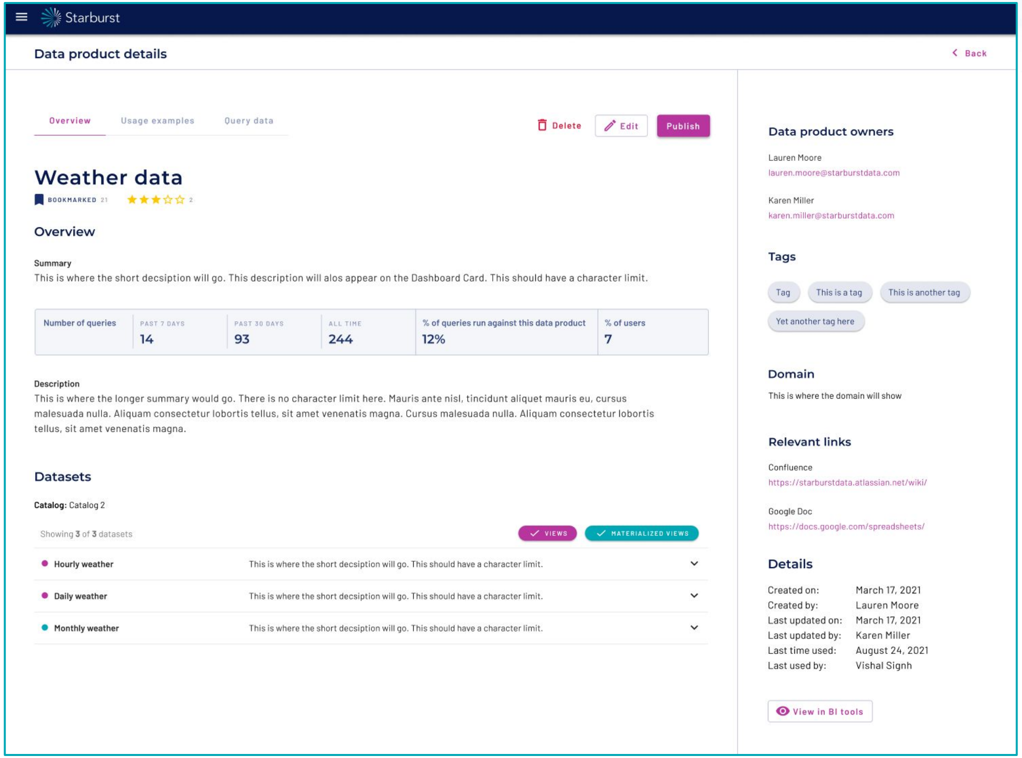
Data on Demand
Gain actionable customer insights for increased sales and conversions
CREATED BY:
JOHN SMITH

Executive & Profession Contacts
Connect with key decision makers to save time and money
CREATED BY:
JOHN SMITH

DRAFT
R&D Data Atlas
180+ public R&D-related datasets from the UN, OECD, Eurostat, EC and other authoritative sources
CREATED BY:
JOHN SMITH

Data consumers

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



Data product details

Weather data

Overview

Summary
This is where the short decsiption will go. This description will alos 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. Cursus 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 decsiption will go. This should have a character limit.
● **Daily weather**	This is where the short decsiption will go. This should have a character limit.
● **Monthly weather**	This is where the short decsiption 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

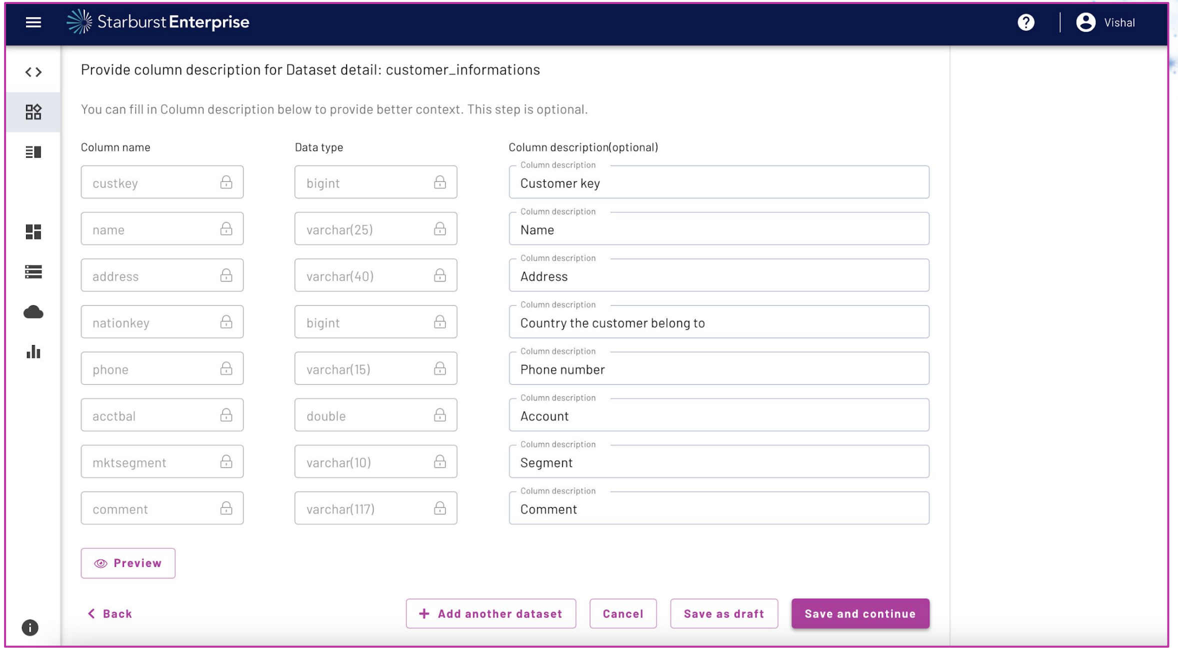
Confluence
https://starburstdata.atlassian.net/wiki/

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 Signh

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



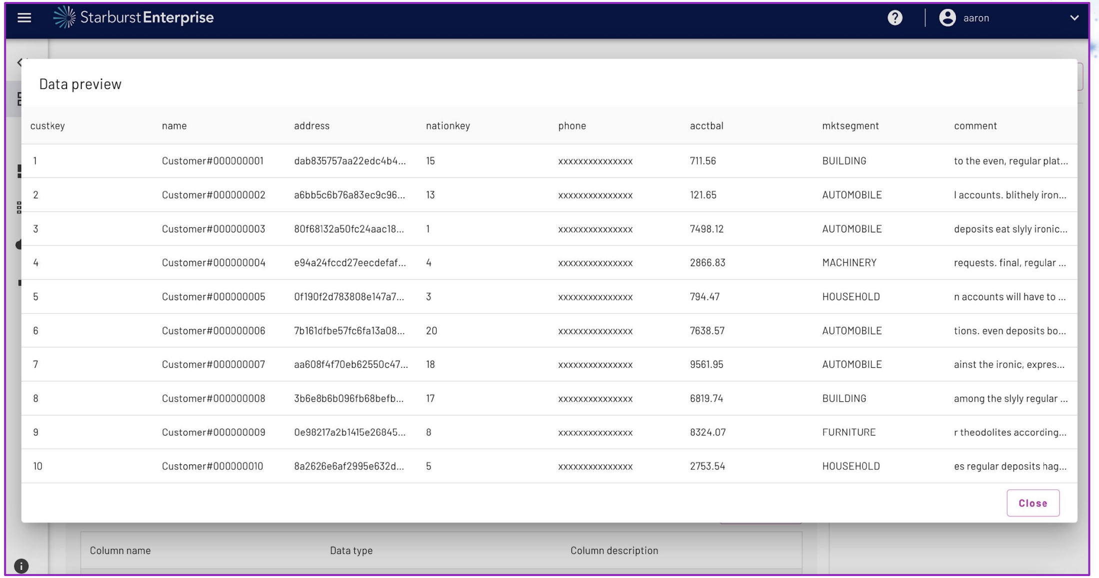
Provide column description for Dataset detail: customer_informations

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

Starburst

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



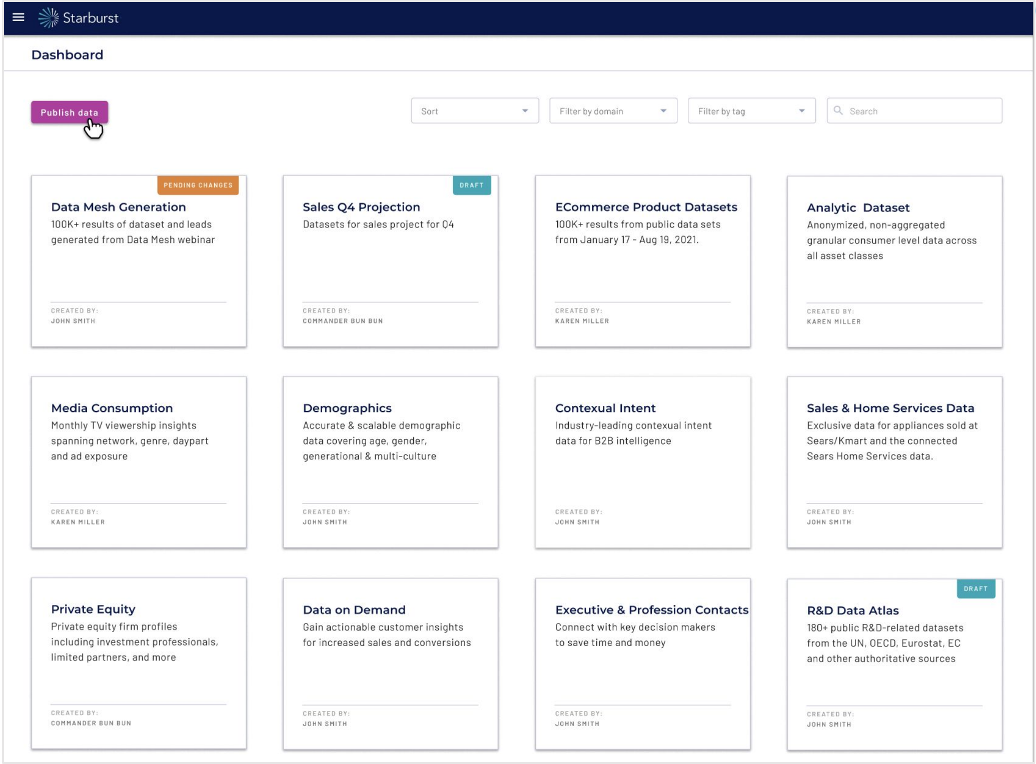
Data preview

custkey	name	address	nationkey	phone	acctbal	mktsegment	comment
1	Customer#000000001	dab835757aa22edc4b4...	15	xxxxxxxxxxxxxxx	711.56	BUILDING	to the even, regular plat...
2	Customer#000000002	a6bb5c6b76a83ec9c96...	13	xxxxxxxxxxxxxxx	121.65	AUTOMOBILE	l accounts. blithely iron...
3	Customer#000000003	80f68132a50fc24aac18...	1	xxxxxxxxxxxxxxx	7498.12	AUTOMOBILE	deposits eat slyly ironic...
4	Customer#000000004	e94a24fccd27eecdefaf...	4	xxxxxxxxxxxxxxx	2866.83	MACHINERY	requests. final, regular ...
5	Customer#000000005	0f190f2d783808e147a7...	3	xxxxxxxxxxxxxxx	794.47	HOUSEHOLD	n accounts will have to ...
6	Customer#000000006	7b161dfbe57fc6fa13a08...	20	xxxxxxxxxxxxxxx	7638.57	AUTOMOBILE	tions. even deposits bo...
7	Customer#000000007	aa608f4f70eb62550c47...	18	xxxxxxxxxxxxxxx	9561.95	AUTOMOBILE	ainst the ironic, expres...
8	Customer#000000008	3b6e8b6b096fb68befb...	17	xxxxxxxxxxxxxxx	6819.74	BUILDING	among the slyly regular ...
9	Customer#000000009	0e98217a2b1415e26845...	8	xxxxxxxxxxxxxxx	8324.07	FURNITURE	r theodolites according...
10	Customer#000000010	8a2626e6af2995e632d...	5	xxxxxxxxxxxxxxx	2753.54	HOUSEHOLD	es regular deposits hag...

Column name	Data type	Column description

Starburst

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



Dashboard

PENDING CHANGES
Data Mesh Generation
100K+ results of dataset and leads generated from Data Mesh webinar
CREATED BY:
JOHN SMITH

DRAFT
Sales Q4 Projection
Datasets for sales project for Q4
CREATED BY:
COMMANDER BUN 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

Contexual 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

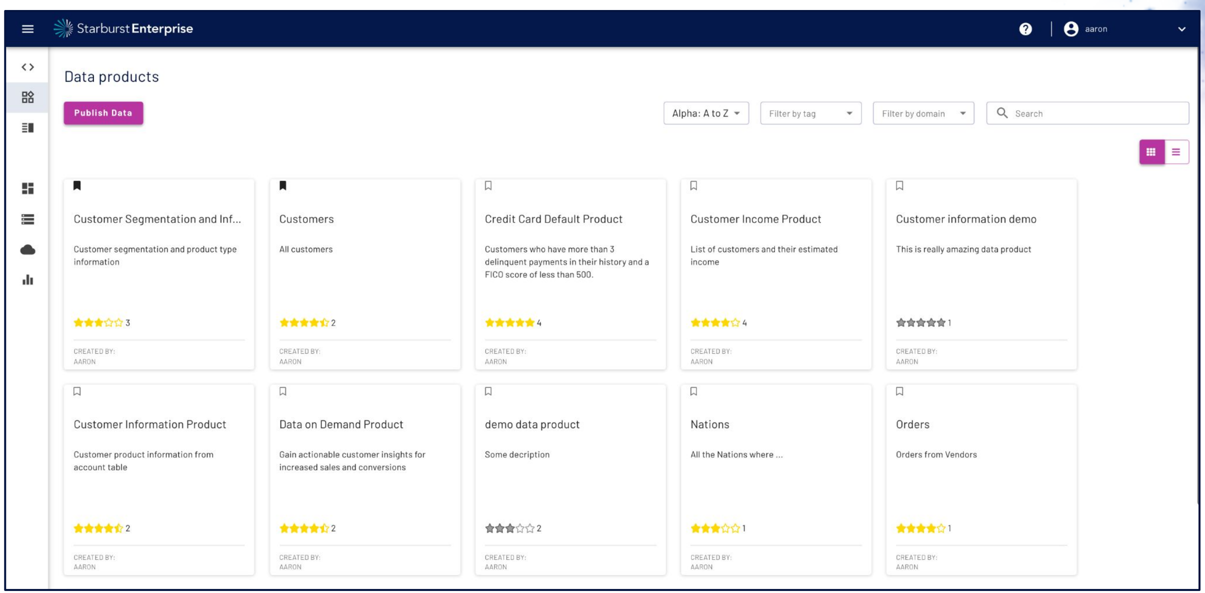
Private Equity
Private equity firm profiles including investment professionals, limited partners, and more
CREATED BY:
COMMANDER BUN BUN

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

DRAFT
R&D Data Atlas
180+ public R&D-related datasets from the UN, OECD, Eurostat, EC and other authoritative sources
CREATED BY:
JOHN SMITH

Share and rate your data products internally and track usage



Data products

Customer Segmentation and Inf...

Customer segmentation and product type information

★★★☆☆ 3

CREATED BY: AARON

Customers

All customers

★★★★⯪ 2

CREATED BY: AARON

Credit Card Default Product

Customers who have more than 3 delinquent payments in their history and a FICO score of less than 500.

★★★★★ 4

CREATED BY: AARON

Customer Income Product

List of customers and their estimated income

★★★★☆ 4

CREATED BY: AARON

Customer information demo

This is really amazing data product

★★★★★ 1

CREATED BY: AARON

Customer Information Product

Customer product information from account table

★★★★⯪ 2

CREATED BY: AARON

Data on Demand Product

Gain actionable customer insights for increased sales and conversions

★★★★⯪ 2

CREATED BY: AARON

demo data product

Some decription

★★★☆☆ 2

CREATED BY: AARON

Nations

All the Nations where ...

★★★☆☆ 1

CREATED BY: AARON

Orders

Orders from Vendors

★★★★☆ 1

CREATED BY: AARON

Starburst

HANDS-ON LABS!!

Starburst Academy