SQL KONFERENZ 2025

Navigating Data Seas: Diving into Databricks with Orcas



Vitalija Bartusevičiūtė, Evidi

Vitalija Bartusevičiūtė

Senior Consultant, Data Engineer



- Evolutionary biologist at heart
- Data Engineer for a living
- ADF, Databricks, PBI, Synapse

Personal:

- Coffee snob
- Ceramics
- DogOps - 50%

in/vitalijabart/

vitalija.bartuseviciute@evidi.com

“Nobody rolls out of bed and is suddenly a Databricks expert”

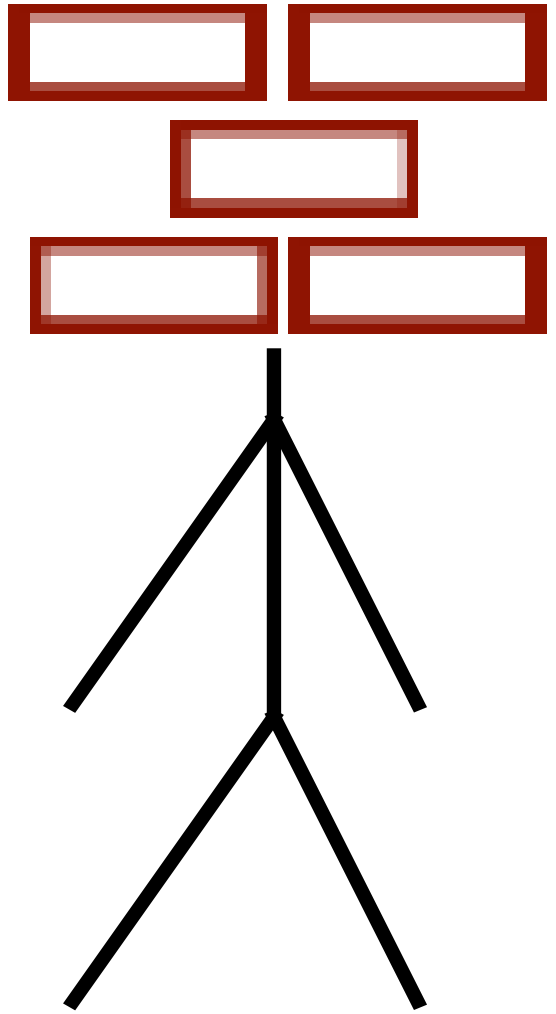
Hugo Lu

Snowflake vs. Databricks 2024 (actually useful) | by Hugo Lu | Medium



Why Brick-headed?

- ADF -> Databricks
- Solution for a major tech enterprise
- Took a certification
- Learning by doing
- Confusing at first
- AI - huge help



What is Databricks?



- Unified open analytics platform for building, deploying, sharing, and maintaining enterprise-grade data, analytics, and AI solutions.
- Strong commitment to open source
- Launched in 2013

Unity Catalog

Unity Catalog - universal catalog for data and AI assets.



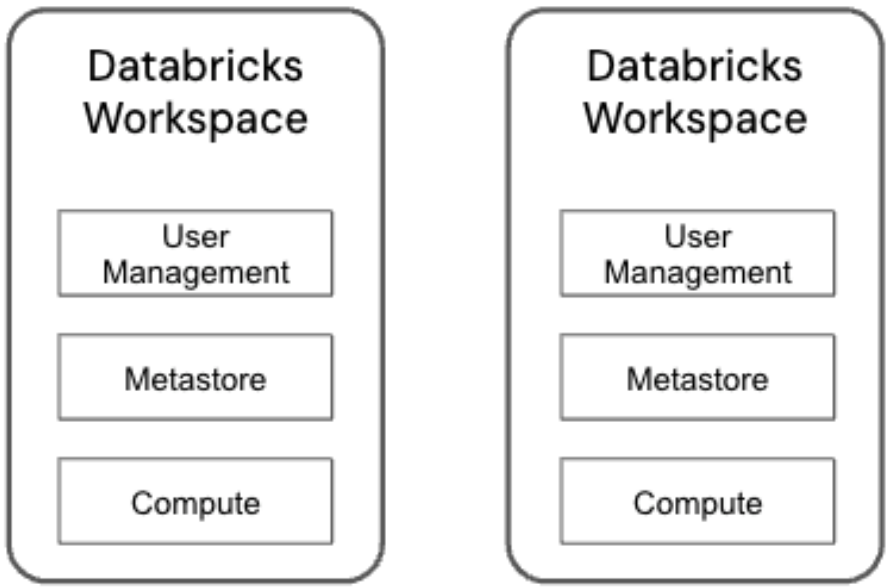
Unity Catalog

- **Governance** solution
- **Unified** across all workspaces
- **Data Discovery** – search interface
- **Lineage**
- **Access** to data (Connections, Clean rooms, Shares)

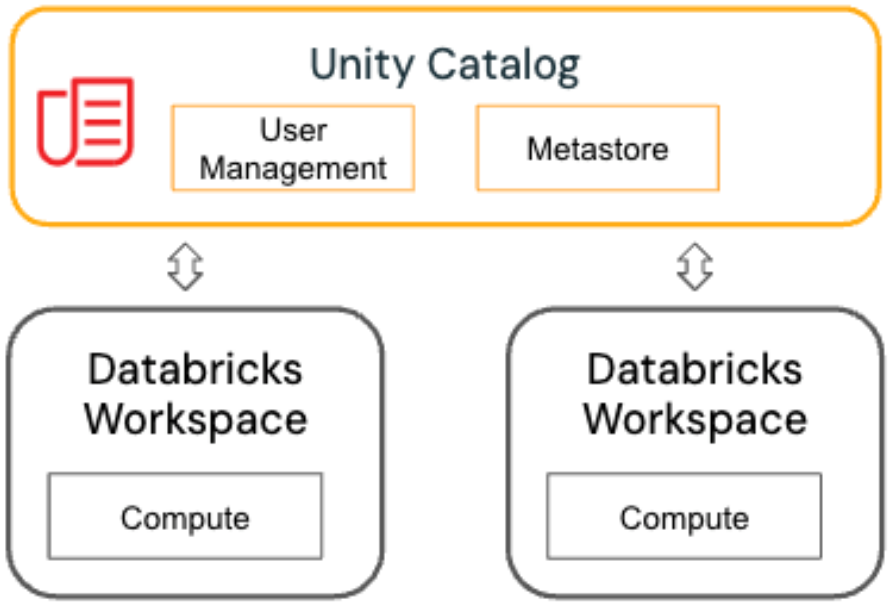


Unity Catalog

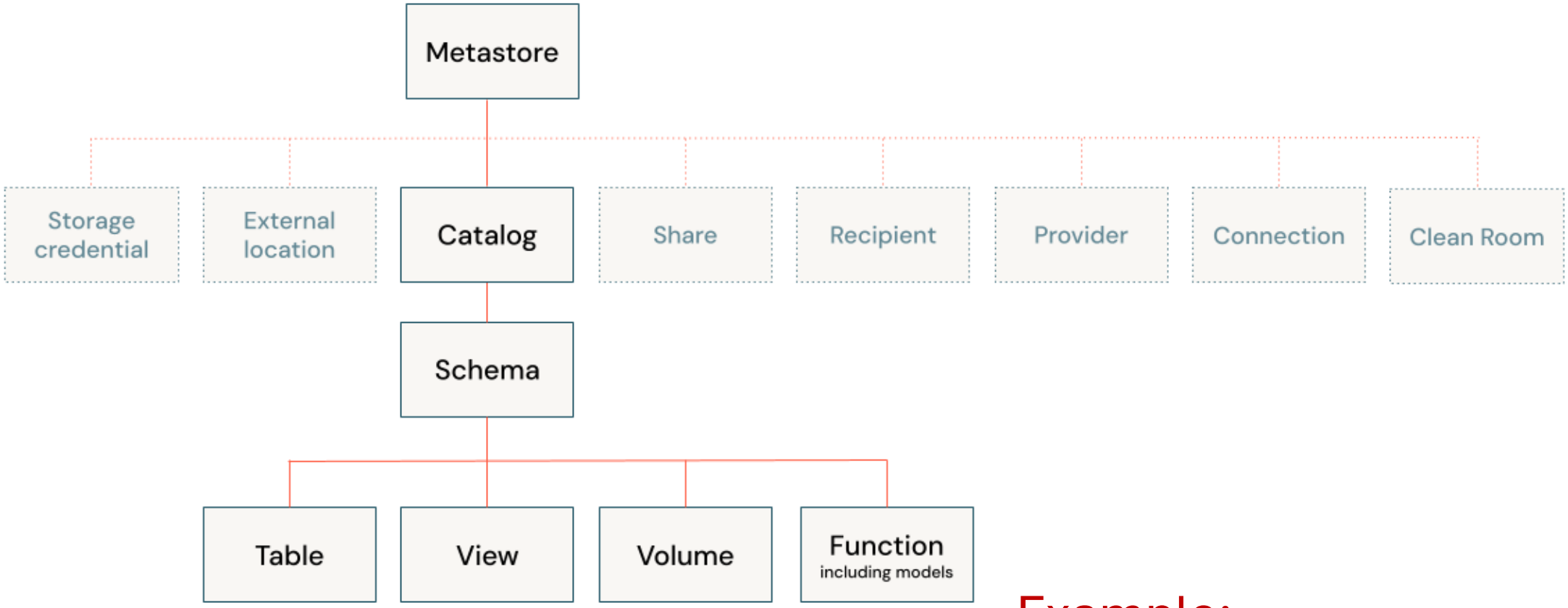
Without Unity Catalog



With Unity Catalog

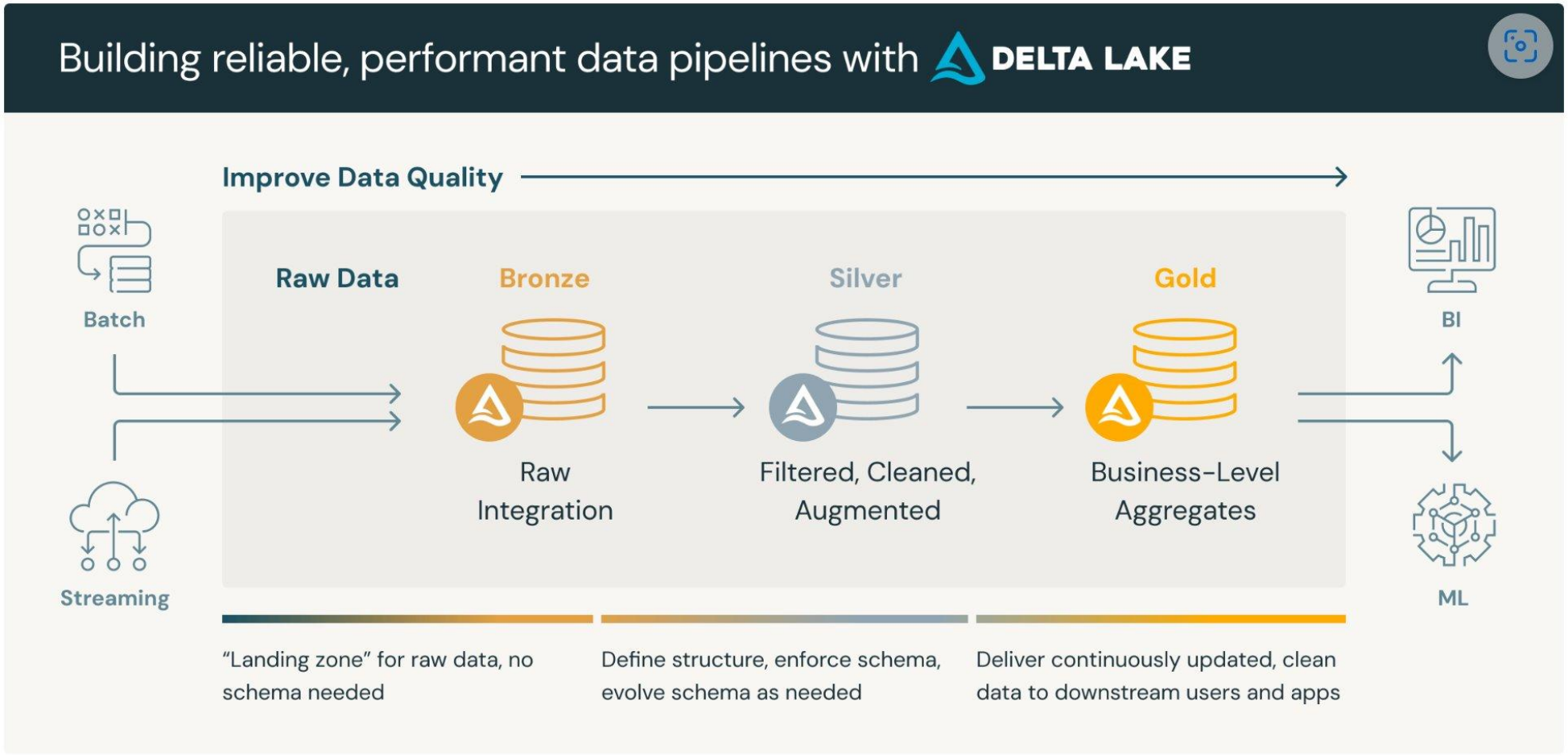


The Unity Catalog object model



Example:
Sandbox.Orcas.Orcas_bronze

The Medallion Architecture



Source: [What is a Medallion Architecture? (databricks.com)]

So...How do I start?





Databricks express setup

✦ Member-only story

DATA ENGINEERING

Databricks Express Setup

Set up your Databricks account in less than 2 minutes



Gaurav Thalpati · Follow

Published in CodeX · 3 min read · Nov 15, 2024

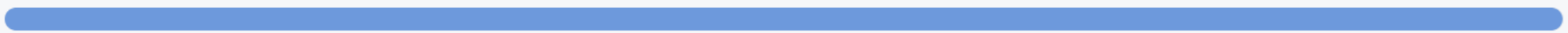
27

[Databricks Express Setup. Set up your Databricks account in less... | by Gaurav Thalpati | CodeX | Medium]

Upgrade to Databricks Premium

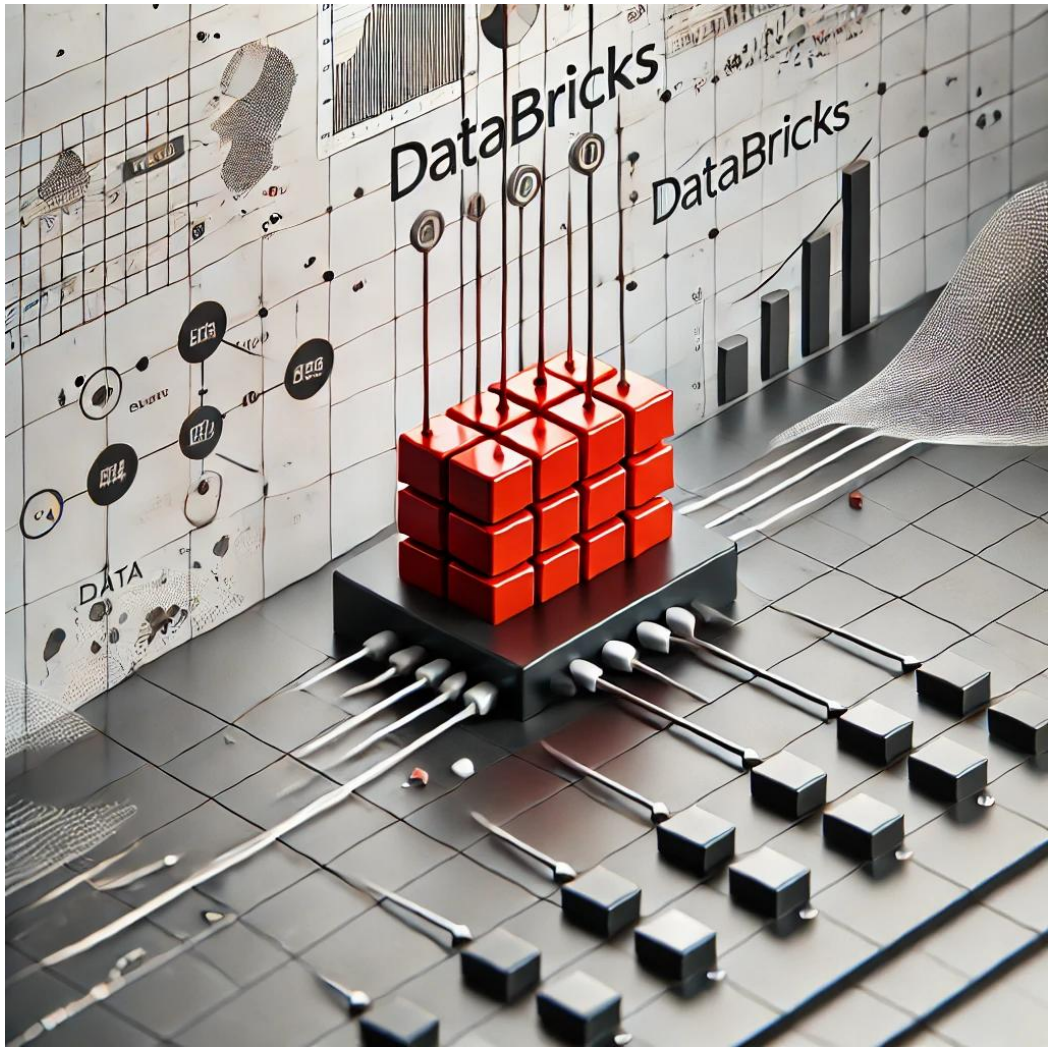
Add a payment method to remove product limits in trial, while keeping your credits. Keep using Databricks on a flexible, pay-as-you-go subscription.

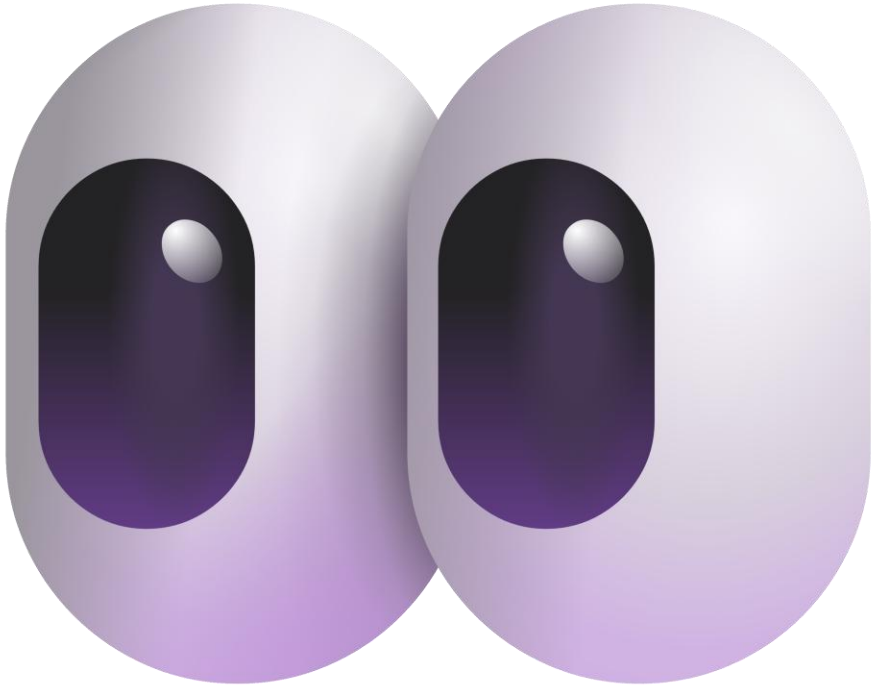
Your credits



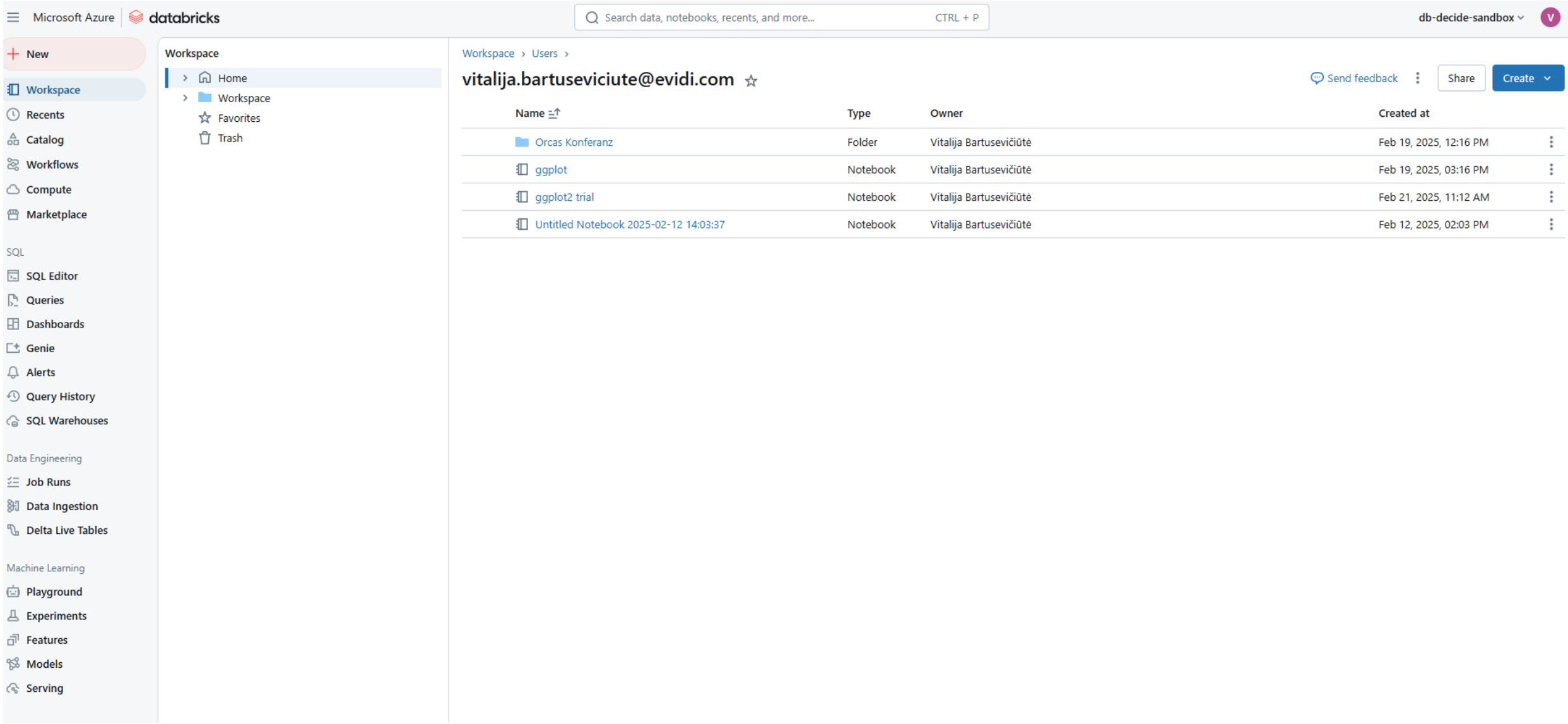
$400 remaining out of $400 Updated about 2 hours ago

Databricks workspace

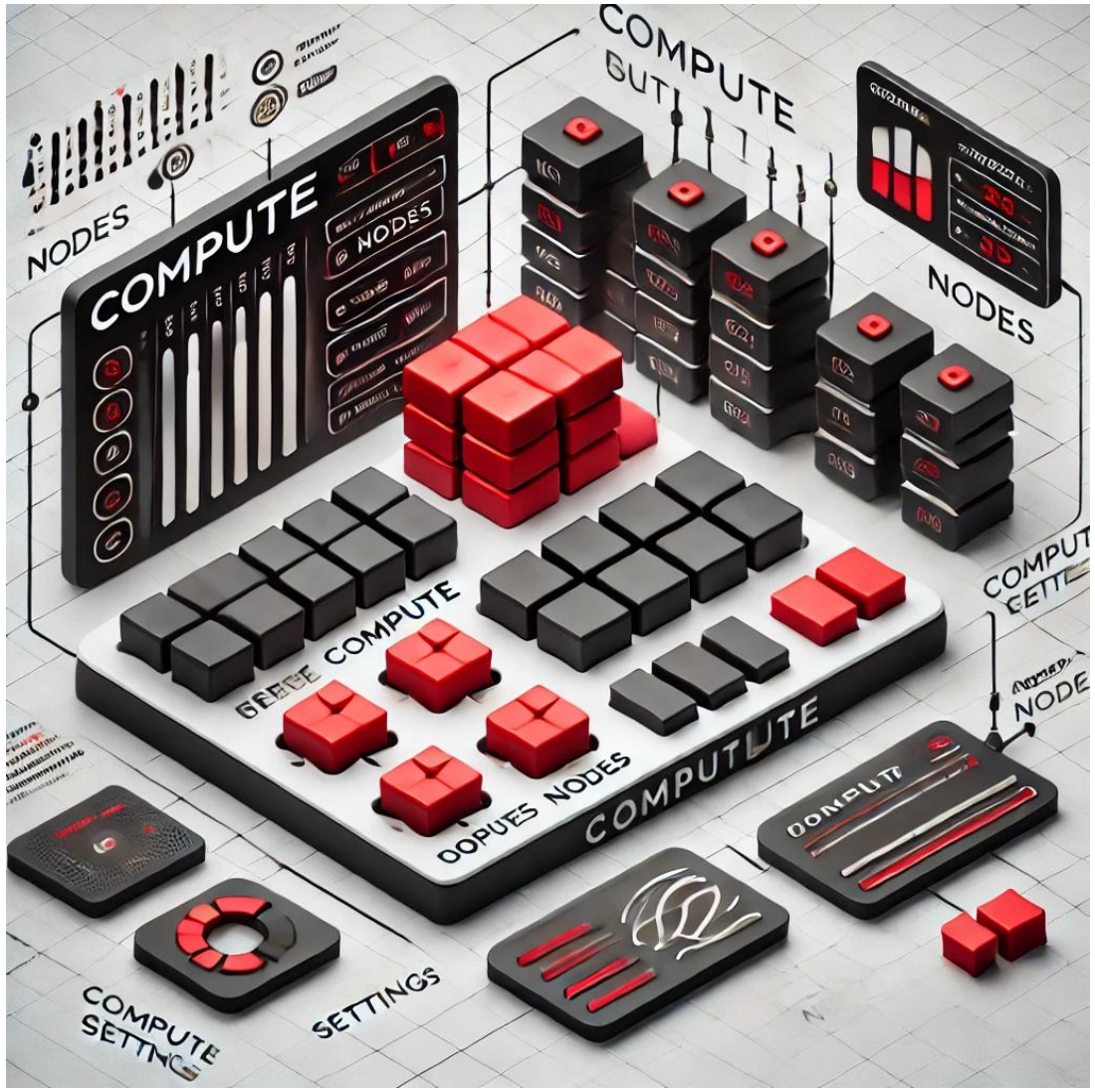




DEMO let’s have a look



So I have a workspace… How do I Compute?





Types of compute

- Serverless compute for notebook
- Serverless compute for jobs
- **All-purpose compute**
- **Jobs compute**
- Instance pools
- **Serverless SQL warehouses**
- Classic SQL warehouses

What do you plan to do?

- Generally trying things in notebooks – Python, SQL, R, Scala – All purpose compute
- SQL Editor & SQL – SQL Serverless
- Run a job – job compute

You don’t know what you want – All purpose compute.

All-purpose compute

Vitalija Bartusevičiūtė's Cluster

Policy

Unrestricted

Multi node Single node

Access mode Single user access

Dedicated (formerly: Single user) Vitalija Bart...

Performance

Databricks runtime version

Runtime: 15.4 LTS (Scala 2.12, Spark 3.5.0)

Use Photon Acceleration

Worker type Min workers Max workers

Standard_D4ds_v5 16 GB Memory, 4 Cores 2 8 Spot instances

Driver type

Same as worker 16 GB Memory, 4 Cores

Enable autoscaling

Terminate after 120 minutes of inactivity

Multi or single node?



Multi node

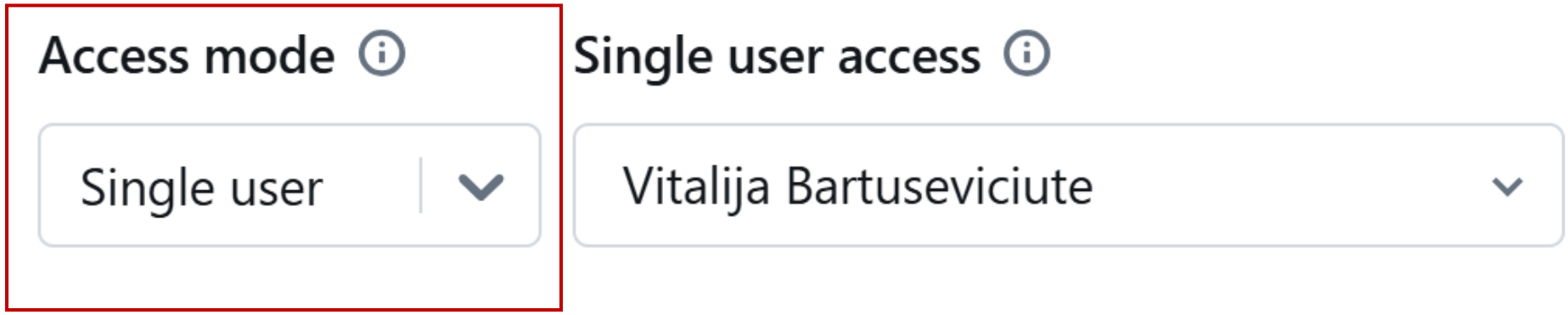
Single node

- Large-scale data processing
- Sharing is caring
- Can’t be scaled to 0 workers
- GPU scheduling

- Large-scale data will exhaust the resources
- Not for sharing
- Can be scaled to 0 workers
- GPU scheduling not enabled

Depends on the use case

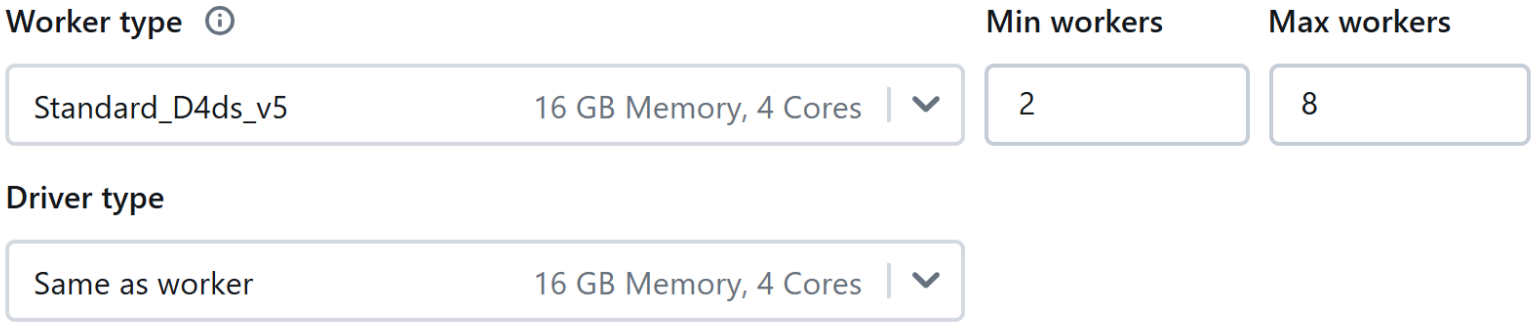
Access mode



WHO can use the cluster?

- Dedicated (formerly: Single user) – only one user
- Standard (formerly: Shared Mode) – multiple users with data isolation among them. (Requires Premium);
- No Isolation shared – multiple users with no data isolation among them

Workers & Drivers?



- Worker nodes – run Spark executors required for properly functioning compute resource.
- Driver node - maintains state information of all notebooks attached to compute resource.
- Enabled autoscaling – min and max workers for the compute resource. Databricks chooses appropriate number of workers required to run your job.
- Autoscaling not enabled – must enter fixed amount of workers.

Databricks Runtime



Set of core components that run on your compute

Runtime versioning

- Long Term Support

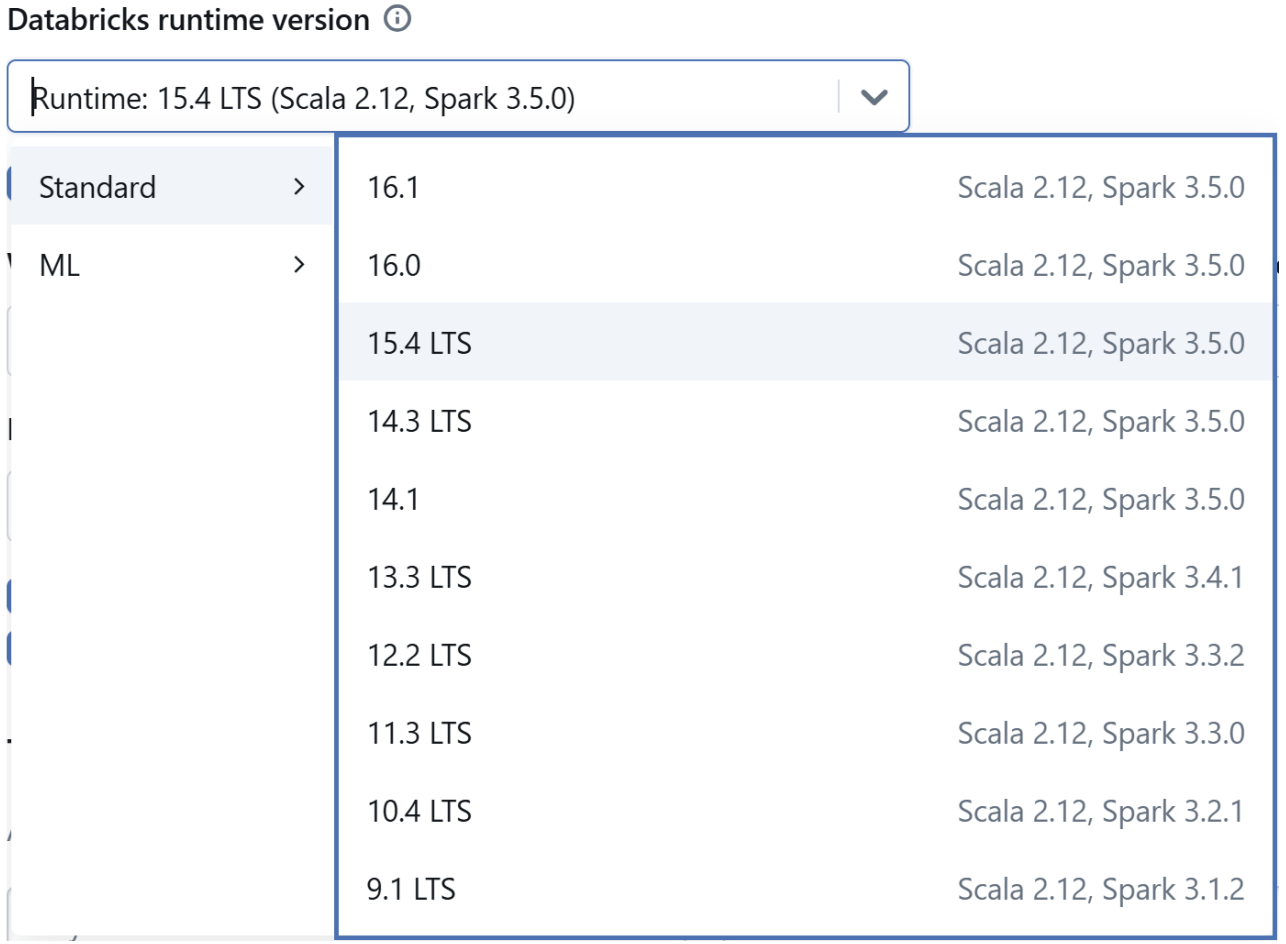
(15.4 LTS)

- Major

(15.5; 16.0)

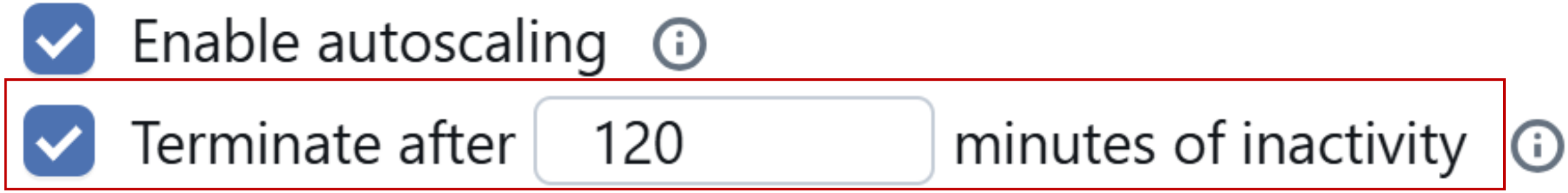
- Feature

(Increment to the version number by decimal i.e. 16.0 to 16.1)



All-purpose compute - use most current version to ensure you have the latest optimizations and most up-to-date computability

Autoscaling

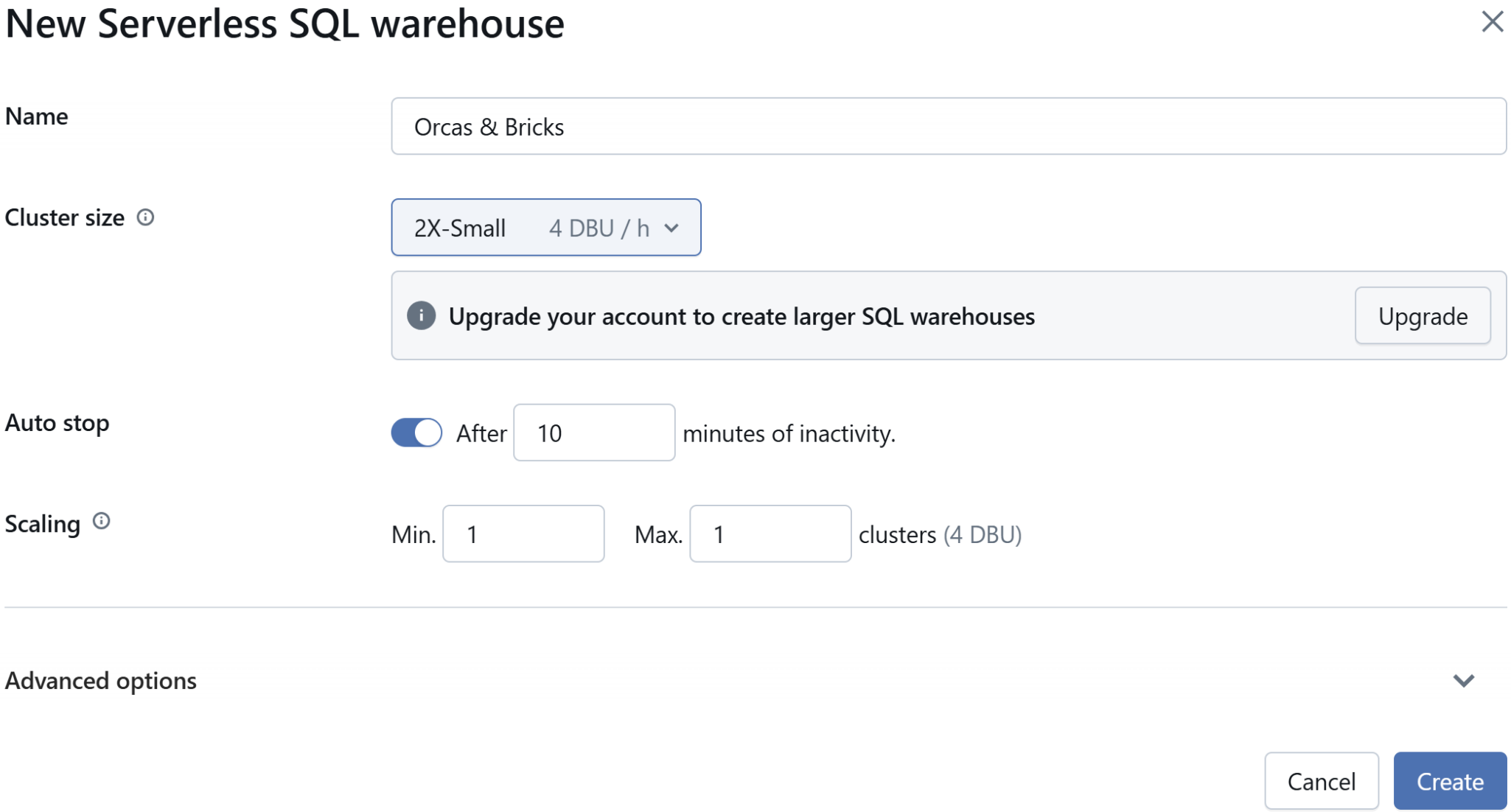


What is optimal for your team?

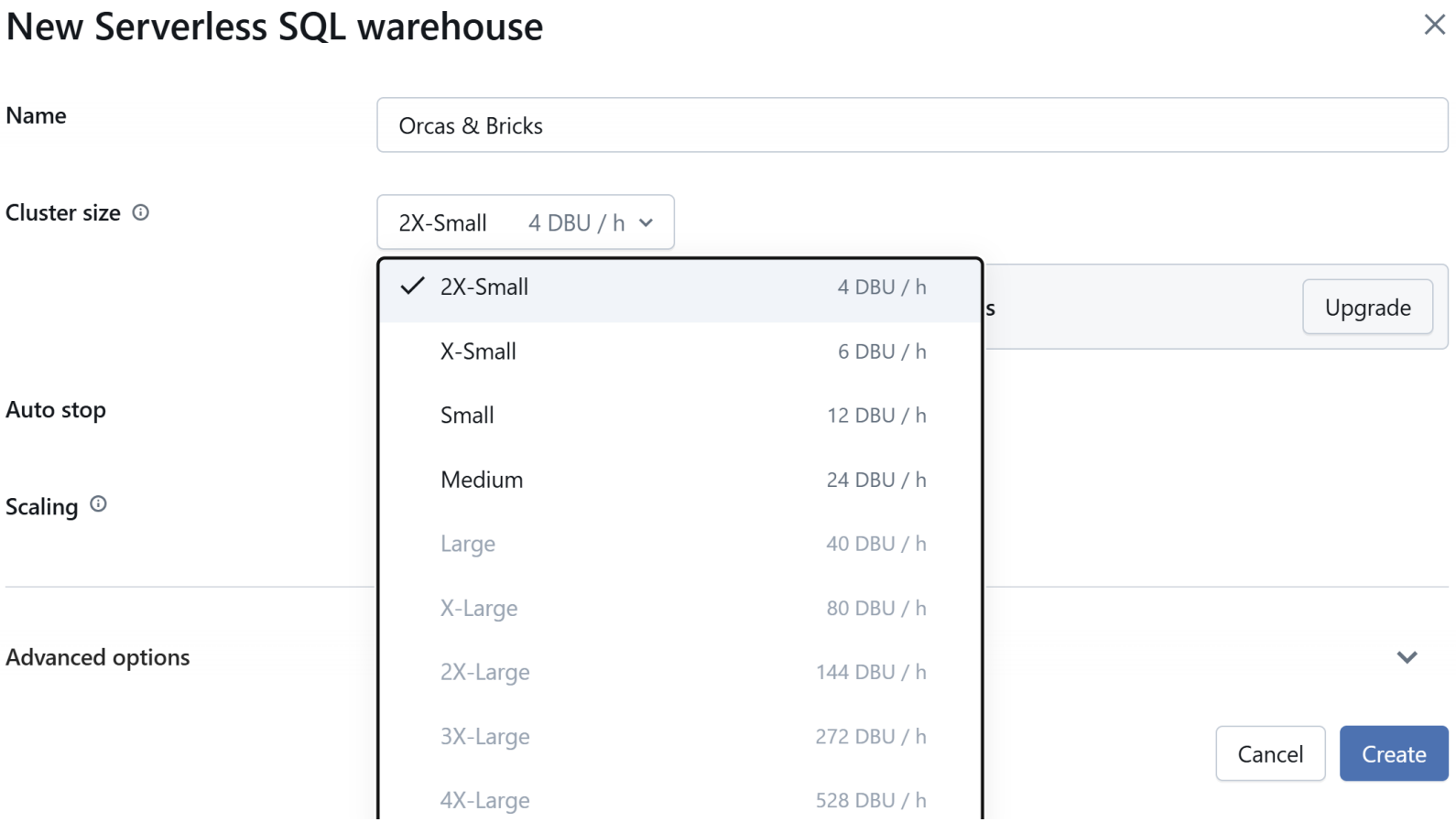
- Optimizes cluster utilization by automatically allocating cluster resources based on workload volume.

- Shuts down under-utilized nodes while guaranteeing there are no failed tasks during shutdown.

Serverless SQL Warehouses



Cluster size

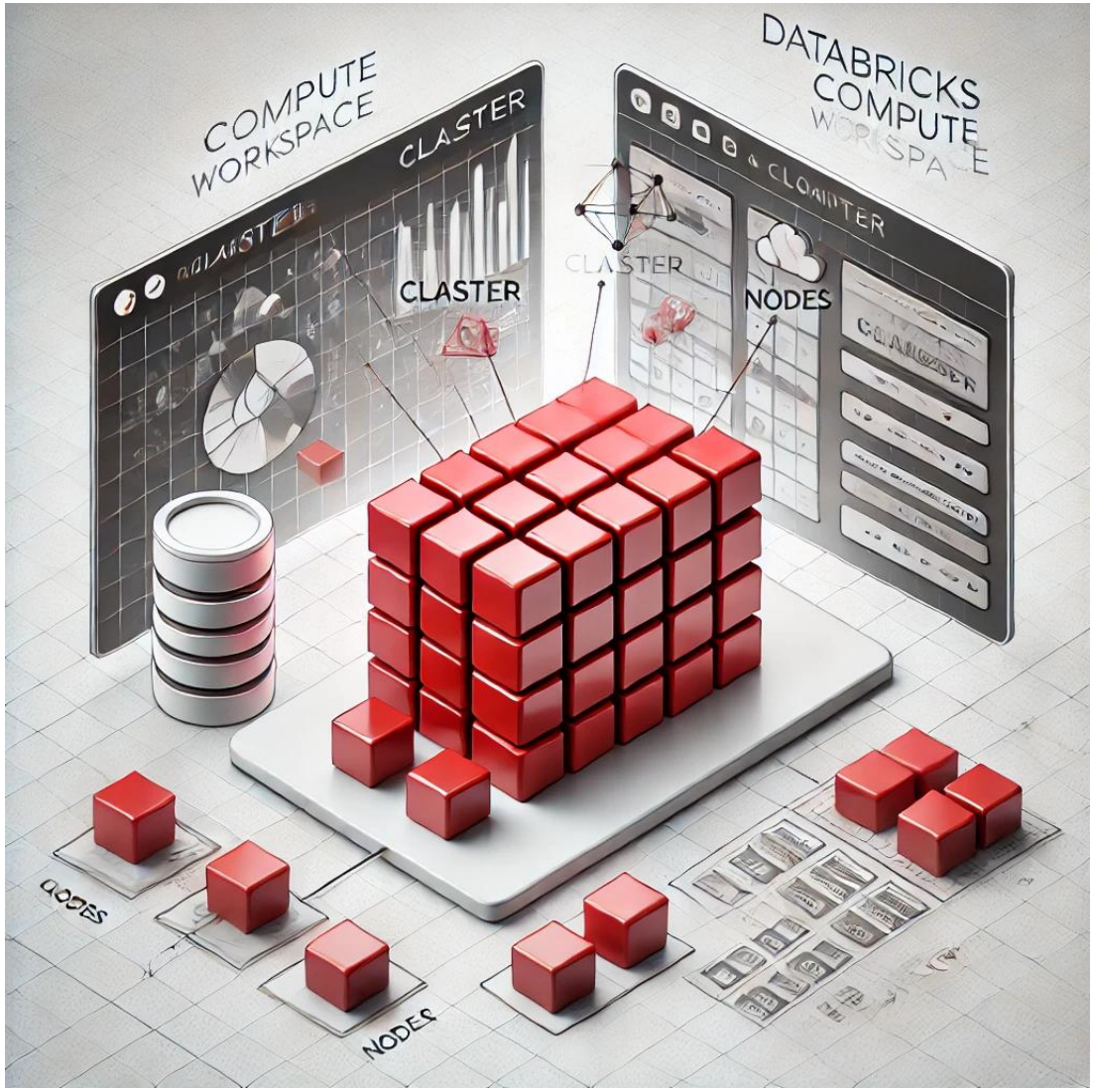


Auto-stop



- Auto stop – stops if its idle for specifies number of minutes
- Scaling – sets minimum and maximum number of clusters to be used in a query
- Increasing max clusters if you want to handle more concurrent users for a given query
- Recommended a cluster for every 10 concurrent queries

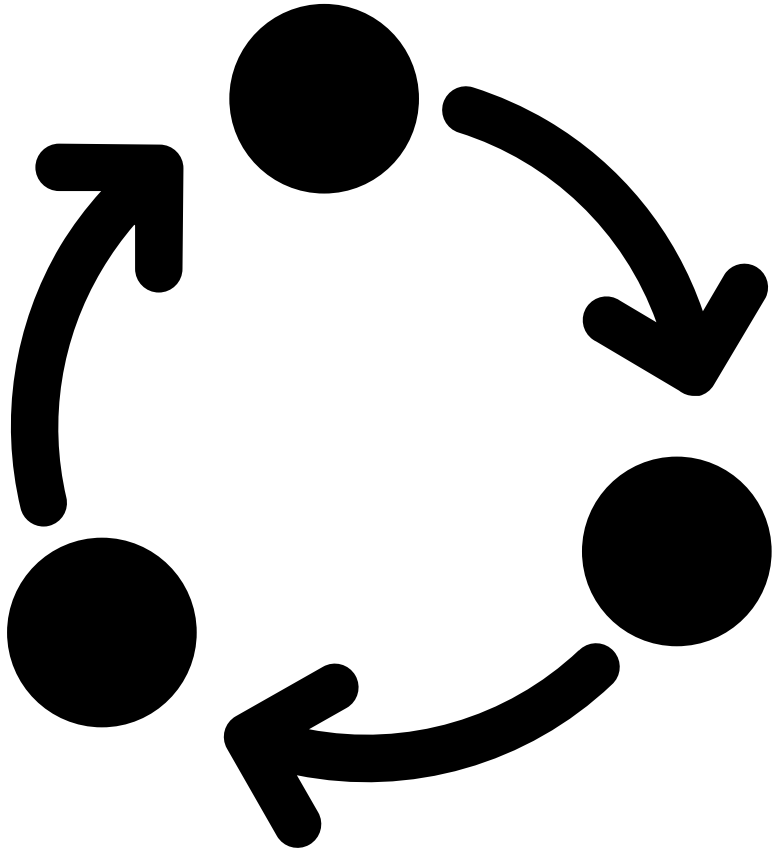
Creating SQL Serverless Warehouse





Job Compute

- Meant for scheduled jobs and batch processing
- Automatically created when a job is scheduled and automatically terminated once job is completed
- Smart to use LTS Databricks Runtime versions
- Cost effective



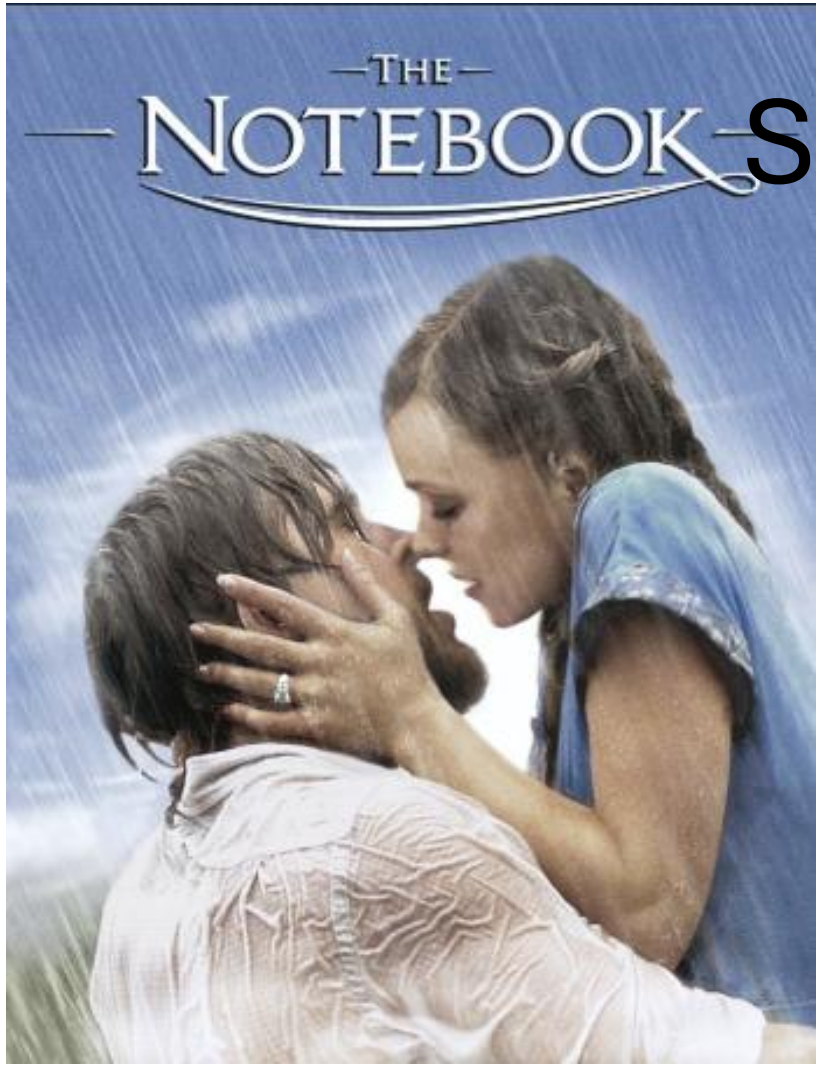
To sum up - simplify

- All-purpose compute - notebooks
- Job compute – jobs
- SQL Serverless – SQL

- Good to keep latest version of DBFS
- Remember to set up auto-terminate



So… Where is a good place to play with my data?



Source: The Notebook (2004) - Posters — The Movie Database (TMDB)

Notebooks

- Common tool for developing code and presenting results
- Offer real-time coauthoring in multiple languages
- Python, SQL, Scala, R
- Magic command %SQL %R %Python

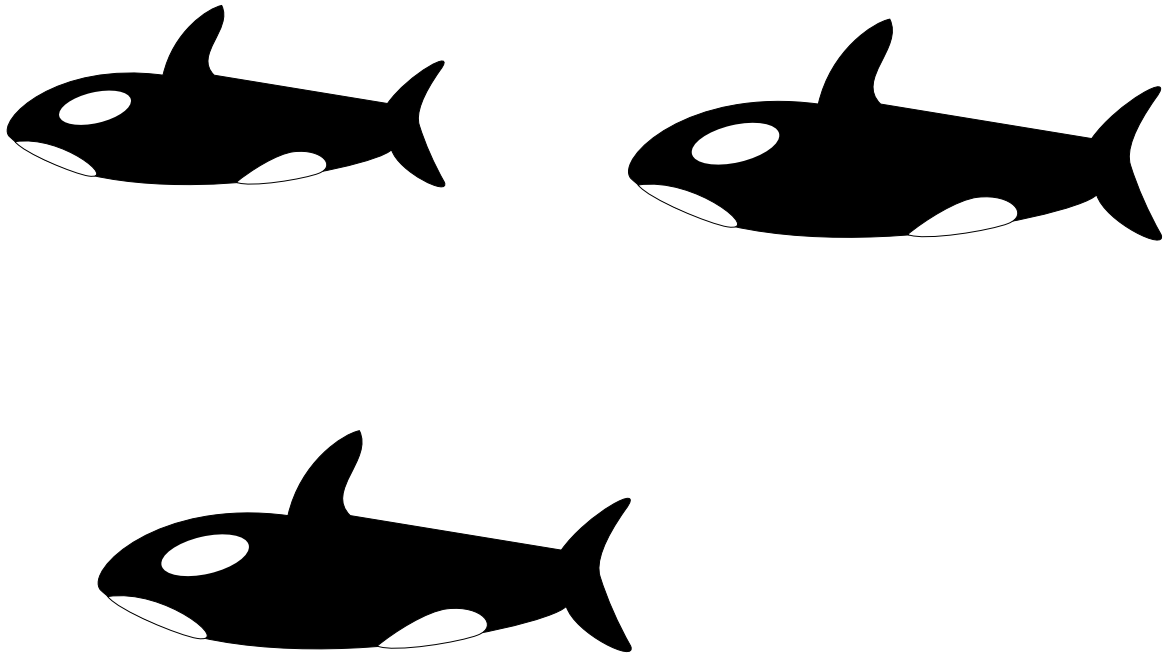
Nice place to start

Developing code in Databricks notebooks

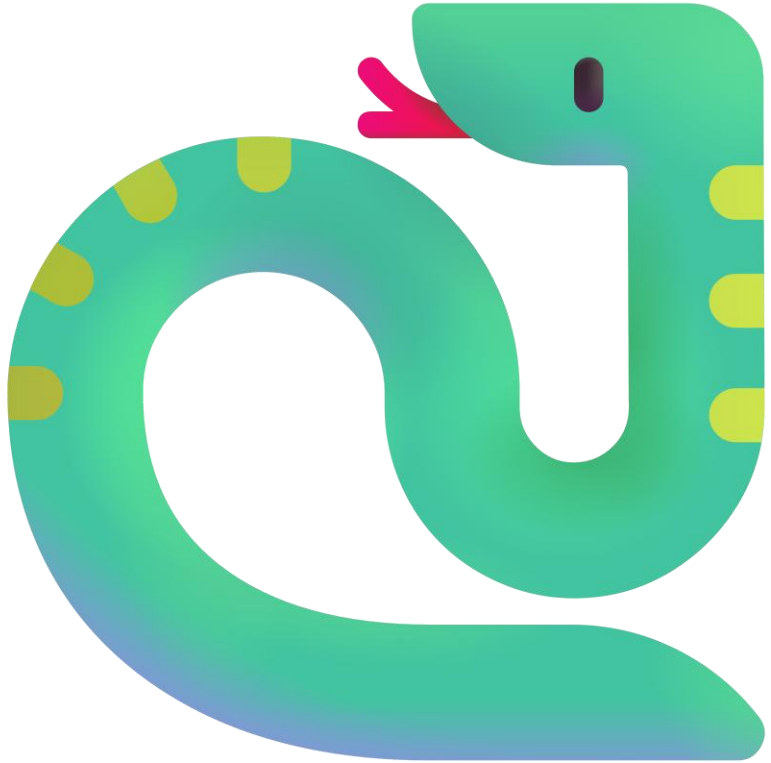
- Autocomplete
- Automatic formatting
- Possibility to combine Python & SQL in a notebook
- Databricks assistance
- Possible to export/import
- Possibility to schedule in workflows

Collaboration in Notebooks

- Collaborative notebooks
- Nice, but I like my colleagues to stay away from coding at the same time
- Can attend other workspaces



Notebooks. Python



Notebooks. Python

11:05 PM (33s) 1 Python

```python
from pyspark.sql.functions import rand, round
from pyspark.sql.types import IntegerType, StringType, DateType
import random
from datetime import datetime, timedelta

# Generate random data
def generate_random_date(start_date, end_date):
    return start_date + timedelta(days=random.randint(0, (end_date - start_date).days))

data = []
start_date = datetime(2000, 1, 1)
end_date = datetime(2025, 1, 19)
genders = ['Male', 'Female']
pods = ['Pod A', 'Pod B', 'Pod C', 'Pod D']
locations = ['Pacific Ocean', 'Atlantic Ocean', 'Indian Ocean', 'Arctic Ocean']
diets = ['Fish', 'Seals', 'Squid', 'Penguins']

for _ in range(200):
    age = random.randint(0, 100)
    date_spotted = generate_random_date(start_date, end_date)
    gender = random.choice(genders)
    pod = random.choice(pods)
    location = random.choice(locations)
    diet = random.choice(diets)
    data.append((age, date_spotted, gender, pod, location, diet))
```

Notebooks. SparkSQL



Notebooks. SparkSQL

15 hours ago (9s) 2 SQL

```
%sql
Select * from whales.orcas.orcas
```

See performance (1) Analyze

Table +

	age	date_spotted	gender	pod	location	diet
3	60	2024-03-09T00:00:00.000+00:...	Male	Pod B	Arctic Ocean	Penguins
4	49	2001-03-15T00:00:00.000+00:...	Female	Pod A	Pacific Ocean	Seals
5	41	2015-10-02T00:00:00.000+00:...	Female	Pod B	Indian Ocean	Fish
6	3	2021-01-05T00:00:00.000+00:...	Female	Pod A	Arctic Ocean	Seals
7	19	2003-02-22T00:00:00.000+00:...	Female	Pod A	Indian Ocean	Fish
8	64	2010-08-02T00:00:00.000+00:...	Male	Pod A	Arctic Ocean	Fish
9	90	2007-07-17T00:00:00.000+00:...	Female	Pod D	Arctic Ocean	Squid
10	65	2000-08-08T00:00:00.000+00:...	Male	Pod D	Arctic Ocean	Fish
11	61	2007-06-12T00:00:00.000+00:...	Female	Pod D	Atlantic Ocean	Fish
12	17	2000-07-22T00:00:00.000+00:...	Male	Pod C	Atlantic Ocean	Fish
13	58	2009-08-17T00:00:00.000+00:...	Male	Pod B	Pacific Ocean	Squid

Notebooks. R



R in notebooks

Bricks & Whales, lets go Python

File Edit View Run Help Last edit was now

Run all | Orcas and Bricks | Schedule | Share

Just now (3s) 8 R

```r
%r
library(tidyverse)
library(sparklyr)

# Connect to your Spark cluster
sc <- spark_connect(method = "databricks")

# Read the multi-part table name directly into an R dataframe using sparklyr
orcatable_df <- tbl(sc, "orcas.orcasandbricks.orcatable") %>%
  collect()

# View the first few rows of the dataframe
head(orcatable_df)

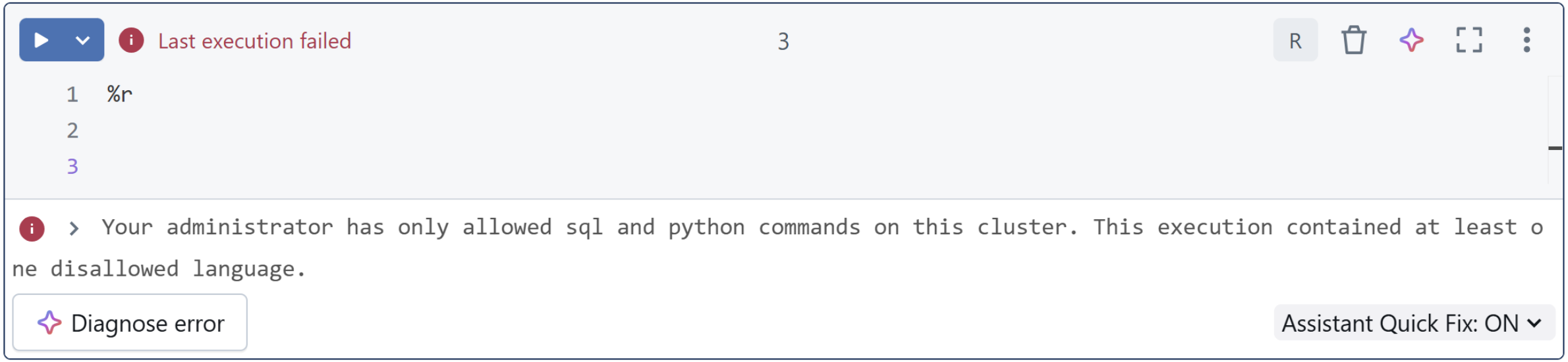
# Summary statistics for numeric columns
orcatable_df %>%
  summarise(
    avg_age = mean(age, na.rm = TRUE),
    avg_size = mean(size, na.rm = TRUE),
    avg_weight = mean(weight, na.rm = TRUE)
  )

# Count of orcas by location
orcatable_df %>%
  group_by(location) %>%
  summarise(count = n()) %>%
  arrange(desc(count))

# Plot size vs. weight
orcaplot <-ggplot(orcatable_df, aes(x = weight, y = size)) +
  geom_point() +
  labs(title = "Orcas Size vs. Weight", x = "Weight", y = "Size")

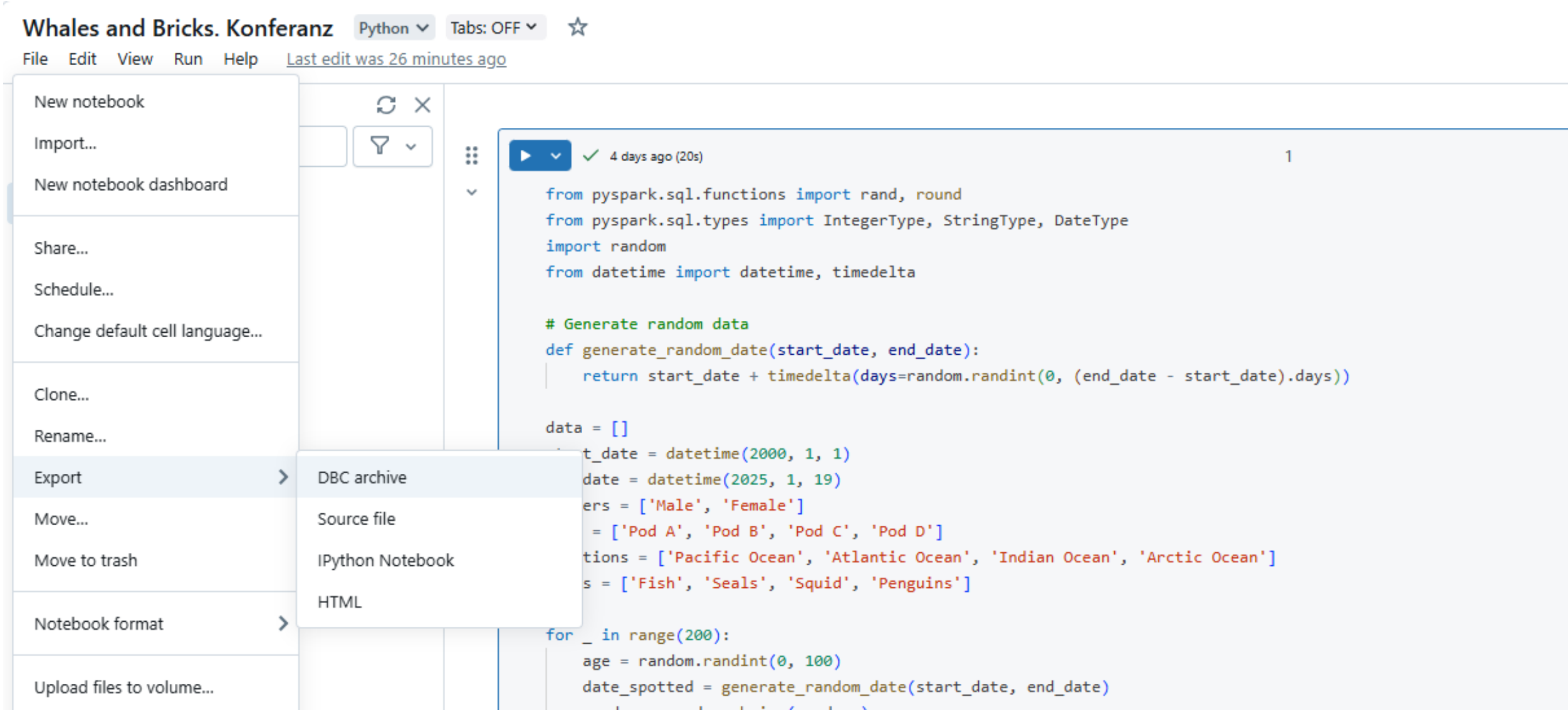
orcaplot
```

Note – on serverless free trial



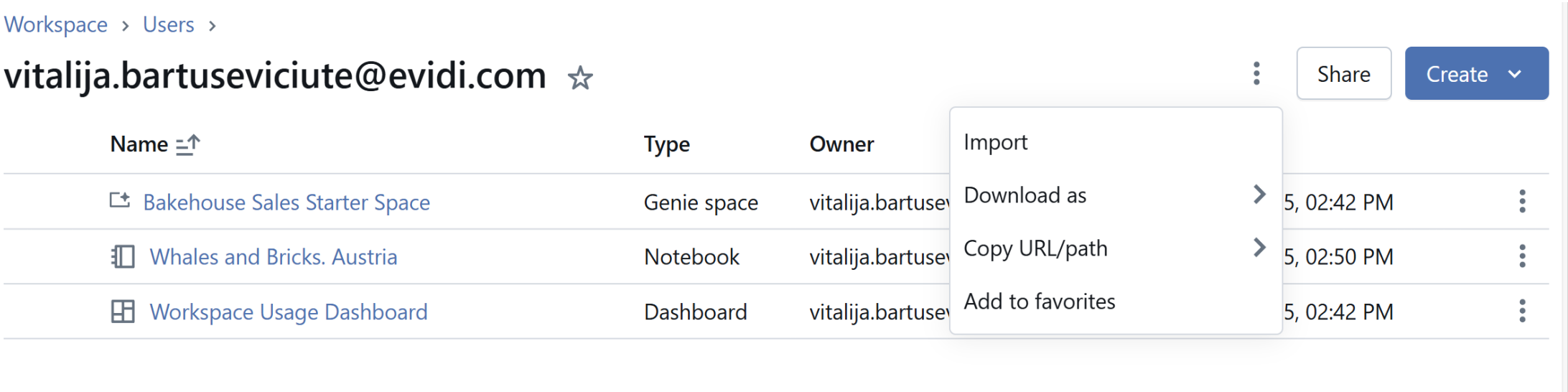
I see what I like… how do I

Export a notebook?

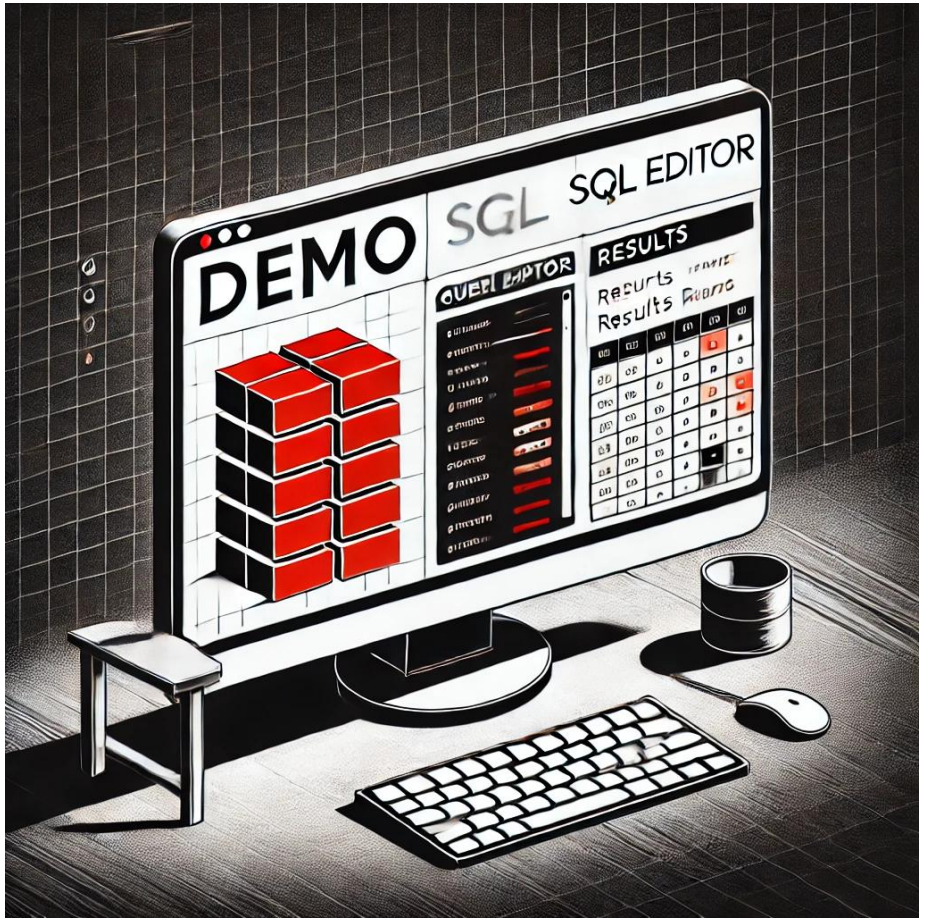

Whales and Bricks. Konferanz
Python
Tabs: OFF
File
Edit
View
Run
Help
Last edit was 26 minutes ago
New notebook
Import...
New notebook dashboard
Share...
Schedule...
Change default cell language...
Clone...
Rename...
Export
Move...
Move to trash
Notebook format
Upload files to volume...
DBC archive
Source file
IPython Notebook
HTML
4 days ago (20s)
1
from pyspark.sql.functions import rand, round
from pyspark.sql.types import IntegerType, StringType, DateType
import random
from datetime import datetime, timedelta
# Generate random data
def generate_random_date(start_date, end_date):
return start_date + timedelta(days=random.randint(0, (end_date - start_date).days))
data = []
t_date = datetime(2000, 1, 1)
date = datetime(2025, 1, 19)
ers = ['Male', 'Female']
= ['Pod A', 'Pod B', 'Pod C', 'Pod D']
tions = ['Pacific Ocean', 'Atlantic Ocean', 'Indian Ocean', 'Arctic Ocean']
s = ['Fish', 'Seals', 'Squid', 'Penguins']
for _ in range(200):
age = random.randint(0, 100)
date_spotted = generate_random_date(start_date, end_date)

I exported some notebook from somewhere… how do I

Import?


Workspace › Users ›
vitalija.bartuseviciute@evidi.com
Share
Create
Import
Download as
Copy URL/path
Add to favorites
Name
Type
Owner
Bakehouse Sales Starter Space
Genie space
vitalija.bartusev
5, 02:42 PM
Whales and Bricks. Austria
Notebook
vitalija.bartusev
5, 02:50 PM
Workspace Usage Dashboard
Dashboard
vitalija.bartusev
5, 02:42 PM

SQL Editor





SQL Editor

- To SQL editor generally one is coming with data already
- Author queries
- Browse data
- Create visualizations

Catalog

Type to search...

For you | All

- My organization
  - workspace
  - system
  - whales
    - default
    - information_schema
    - orcas
      - orcas
        - age
        - date_spotted
        - gender
        - pod
        - location
        - diet
- Shared
  - samples

New Query 2025-01-17 2:49pm

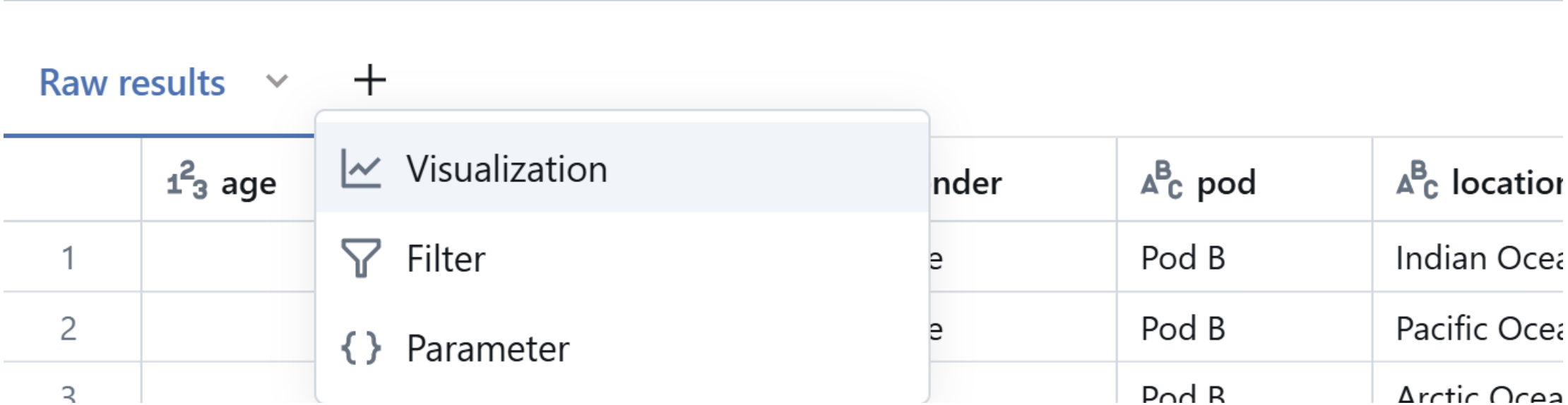
Run (1000) | whales . orcas | Serverless Starte... Serverless M | Save*

```
1 select * from orcas
```

Raw results

	age	date_spotted	gender	pod	location	diet
1	40	2004-02-15T00:00:00.0...	Female	Pod B	Indian Ocean	Squid
2	80	2014-12-04T00:00:00.0...	Female	Pod B	Pacific Ocean	Fish
3	60	2024-03-09T00:00:00.0...	Male	Pod B	Arctic Ocean	Penguins
4	49	2001-03-15T00:00:00.0...	Female	Pod A	Pacific Ocean	Seals
5	41	2015-10-02T00:00:00.0...	Female	Pod B	Indian Ocean	Fish
6	3	2021-01-05T00:00:00.0...	Female	Pod A	Arctic Ocean	Seals
7	19	2003-02-22T00:00:00.0...	Female	Pod A	Indian Ocean	Fish
8	64	2010-08-02T00:00:00.0...	Male	Pod A	Arctic Ocean	Fish
9	90	2007-07-17T00:00:00.0...	Female	Pod D	Arctic Ocean	Squid

Visualization in SQL editor



Orcas Diet

Visualization type

Pie

General | Series | Colors | Data labels

X column

diet

Y columns

diet | Count | -

Add column

Group by

Choose column...

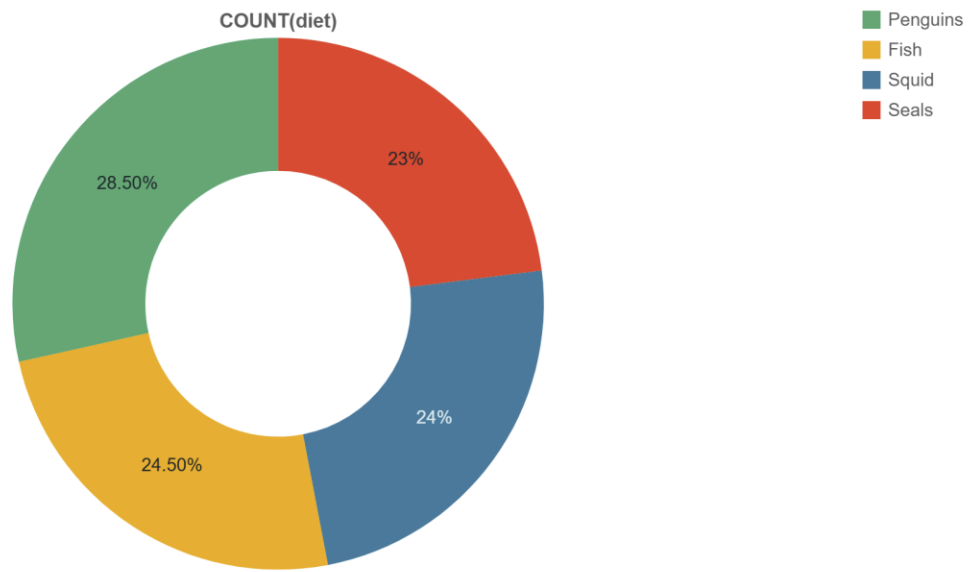
Direction

Counterclockwise

Missing and NULL values

Convert to 0 and display in chart

New charts: ON



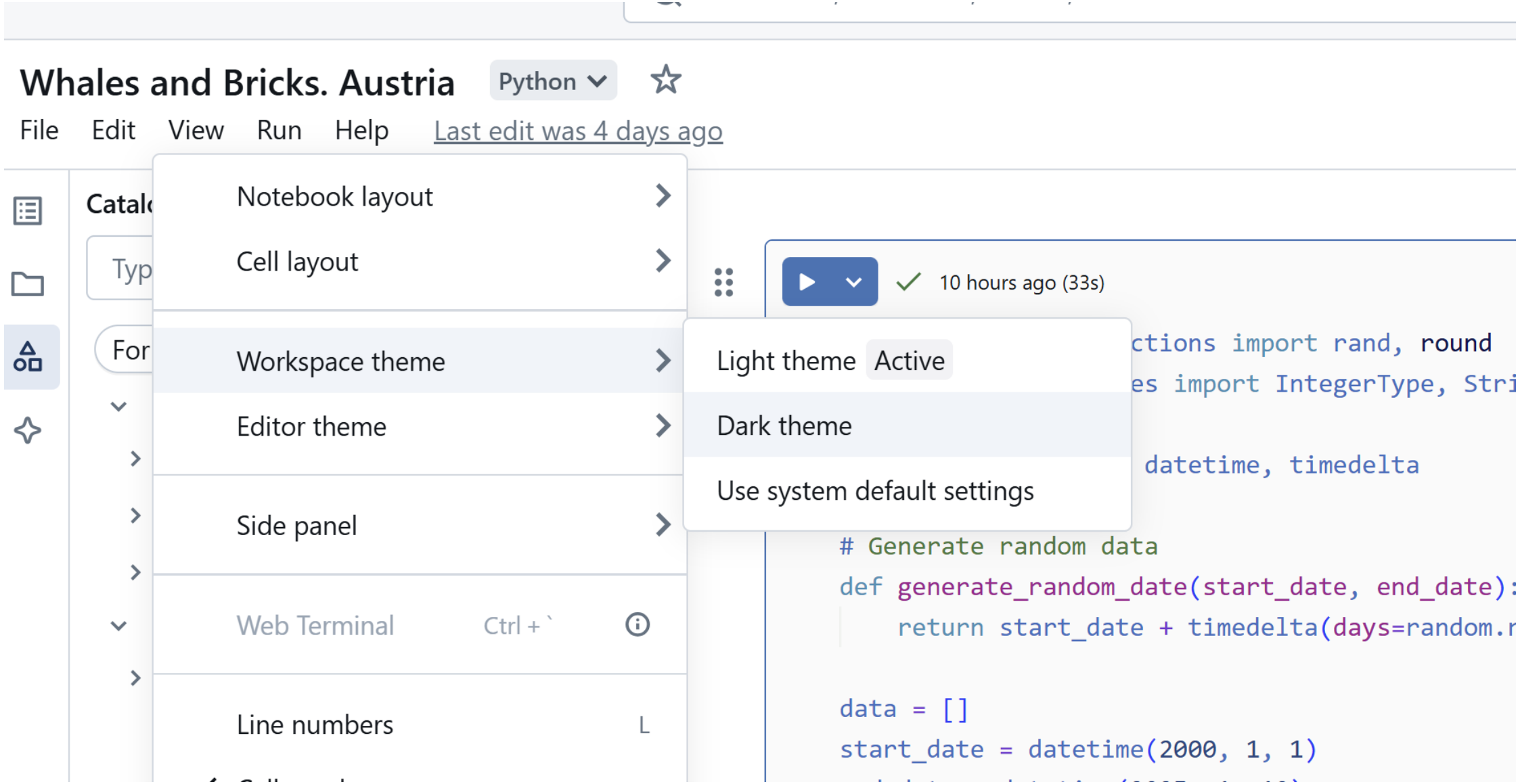
Catalog explorer



Dark mode



For the dark souls out there ;)


Whales and Bricks. Austria
Python
File Edit View Run Help
Last edit was 4 days ago
Notebook layout
Cell layout
Workspace theme
Editor theme
Side panel
Web Terminal Ctrl + `
Line numbers L
Light theme Active
Dark theme
Use system default settings
10 hours ago (33s)
ctions import rand, round
es import IntegerType, Stri
datetime, timedelta
# Generate random data
def generate_random_date(start_date, end_date):
return start_date + timedelta(days=random.r
data = []
start_date = datetime(2000, 1, 1)

I have an Excel, how do I query?



The classic…

The medallion of Orcas





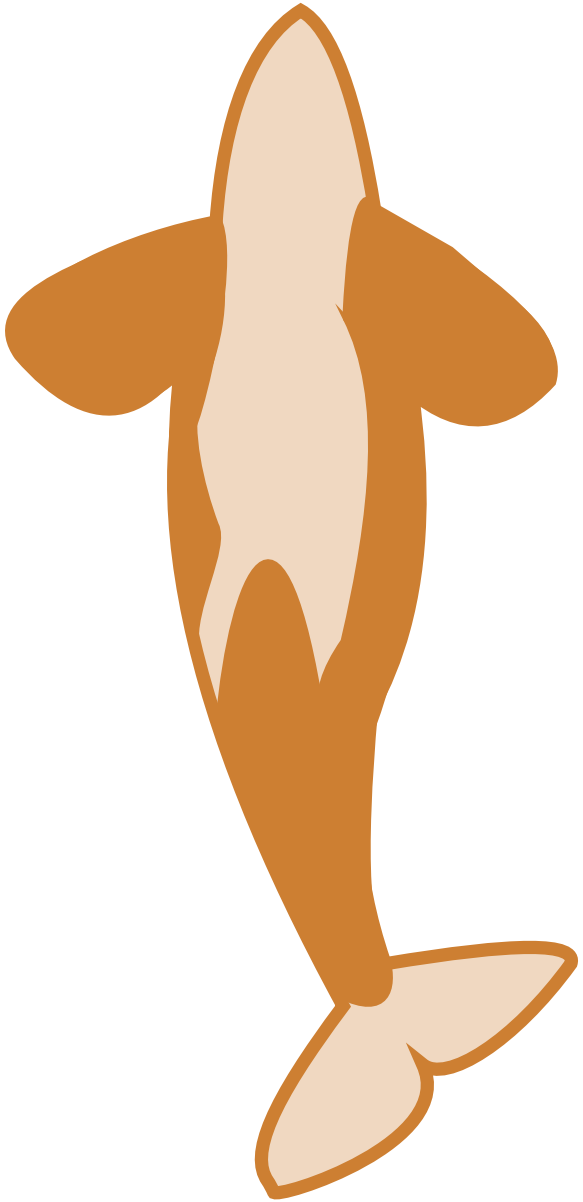
Bronze layer

- Land data from external source systems “as-is"
- Additional metadata columns (when loaded & so on)
- The focus in is quick Change Data Capture and the ability to provide an historical archive of source (cold storage)

Orcas bronze

- Picking just the column datatypes as they are

- In client solutions:
- historization
- technical columns
- column names as is



Silver layer
cleansed and conformed data

- Brings the data from different sources into an Enterprise view
- Only minimal or "just-enough" transformations and data cleansing
- Speed and agility to ingest and deliver the data prioritized.

Silver

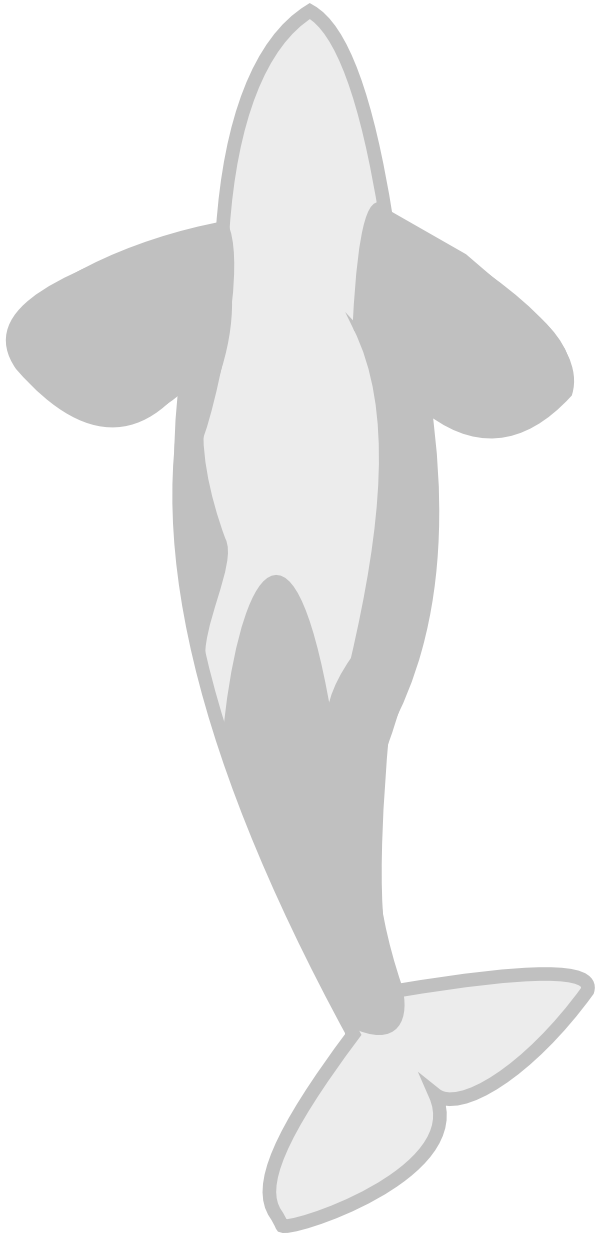
- Adjusting names:

ORC_Name_Name

- Adjusting datatypes

- In our solutions:

- foreign keys
- business logic



Gold

- Typically organized in consumption-ready "project-specific" databases.

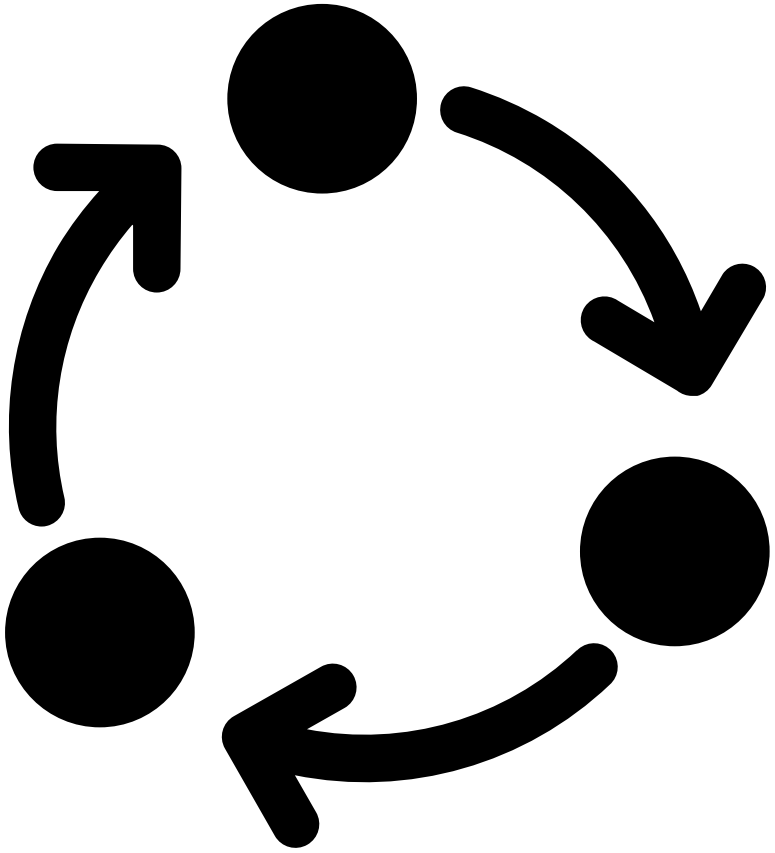
- The Gold layer is for reporting and uses more de-normalized and read-optimized data models with fewer joins.

- The final layer of data transformations and data quality rules are applied here.

Gold

- Getting ready for Power BI
- Client is very particular on naming
- Adding a new aggregation

Automating a job



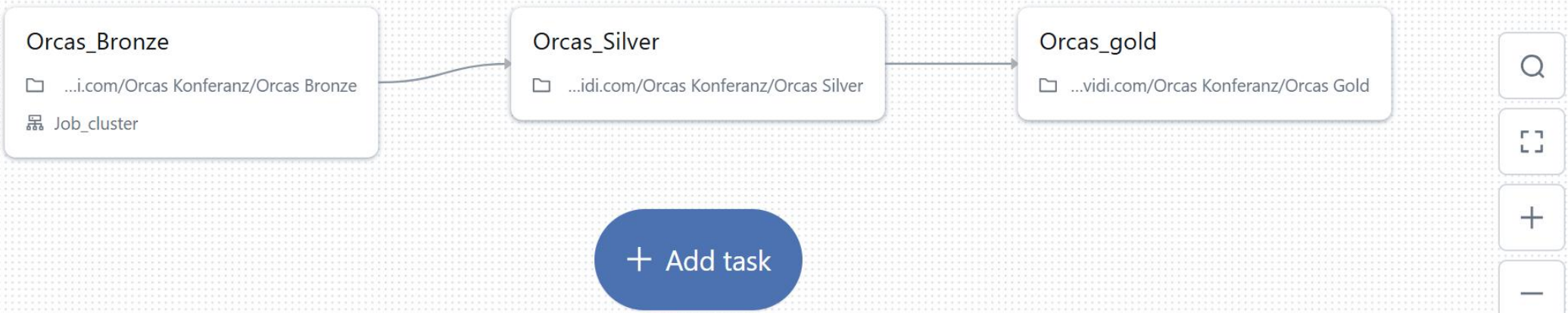
Workflows

- Data orchestration tool.
- Can create a workflow to automate running the data ingestion, processing, and analysis steps using a Databricks job.
- Can parametrize.
- Email/teams/slack notifications if something failed.

Workflows › Jobs & pipelines ›

Medallion of Orcas_Germany ☆

Runs Tasks



Job_cluster

Cluster name

Job_cluster

Policy

Unrestricted

Multi node Single node

Access mode

Single user

UI | JSON

Summary

2 Workers — 32 GB Memory, 8 Cores

1 Driver — 16 GB Memory, 4 Cores

Runtime — 15.4.x-scala2.12

Unity Catalog Photon

Standard_D4ds_v5 7.5 DBU/h

Performance

Databricks runtime version

Runtime: 15.4 LTS (Scala 2.12, Spark 3.5.0)

Use Photon Acceleration

Worker type

Standard_D4ds_v5 — 16 GB Memory, 4 Cores

Workers

2

Spot instances

Driver type

Same as worker — 16 GB Memory, 4 Cores

Enable autoscaling

Tags

Cancel Confirm

Notifications to MS Teams

https://www.vitalijalovesdata.com/



Vitalija B · Feb 14 · 2 min read

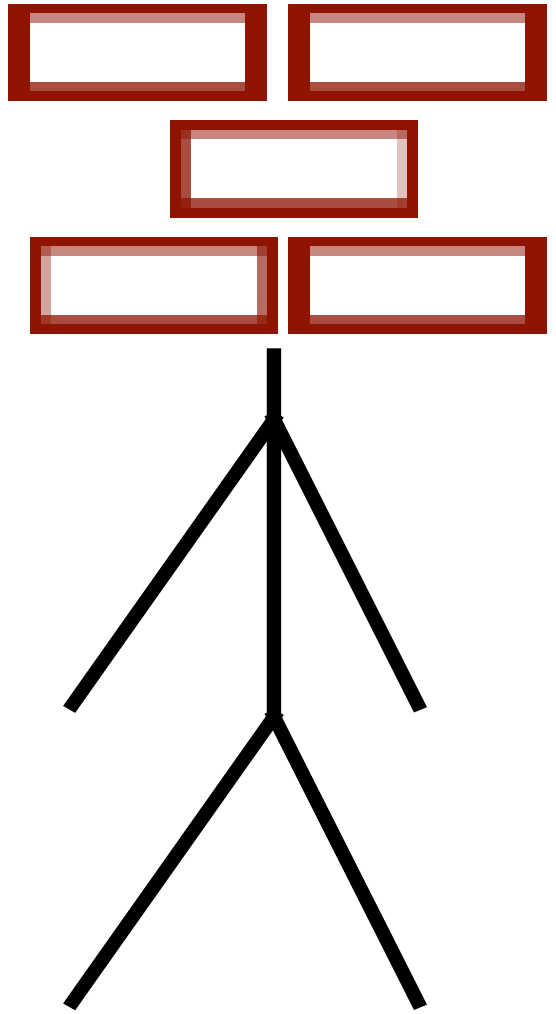
Databricks to Teams notifications via workflows

Summary

- Bricks are great, but not perfect
- Lots of functionalities
- Mature PaaS
- Requires a bit of googling,

once you are onto it, it’s great

- Good to have standard procedures in place, so it’s not a full mess



KUDOS



Brage Rønning Tukkensæter



Ben Weissman

NEW STARS OF DATA



André Kamman



Emilie Rønning



SQL KONFERENZ 2025

Feedback

SQL KONFERENZ 2025



Questions, inquiries, chitchat?



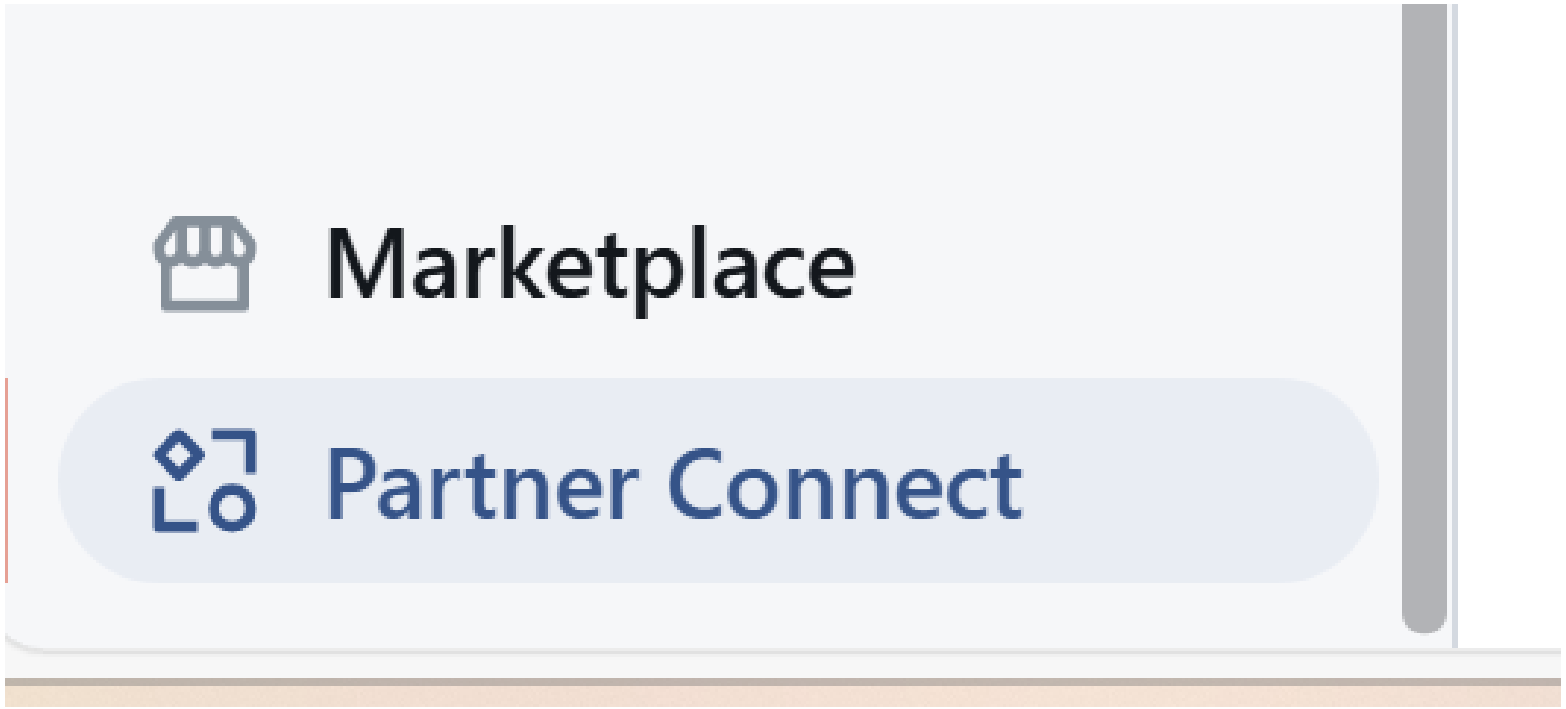
Vitalija Bartusevičiūtė

Senior Consultant - Data Engineer | PhD Candidate in Evolutionary Biology

Oslo, Oslo, Norway · Contact info

Connect to PBI




Marketplace
Partner Connect

Connect to partner



Microsoft Power BI

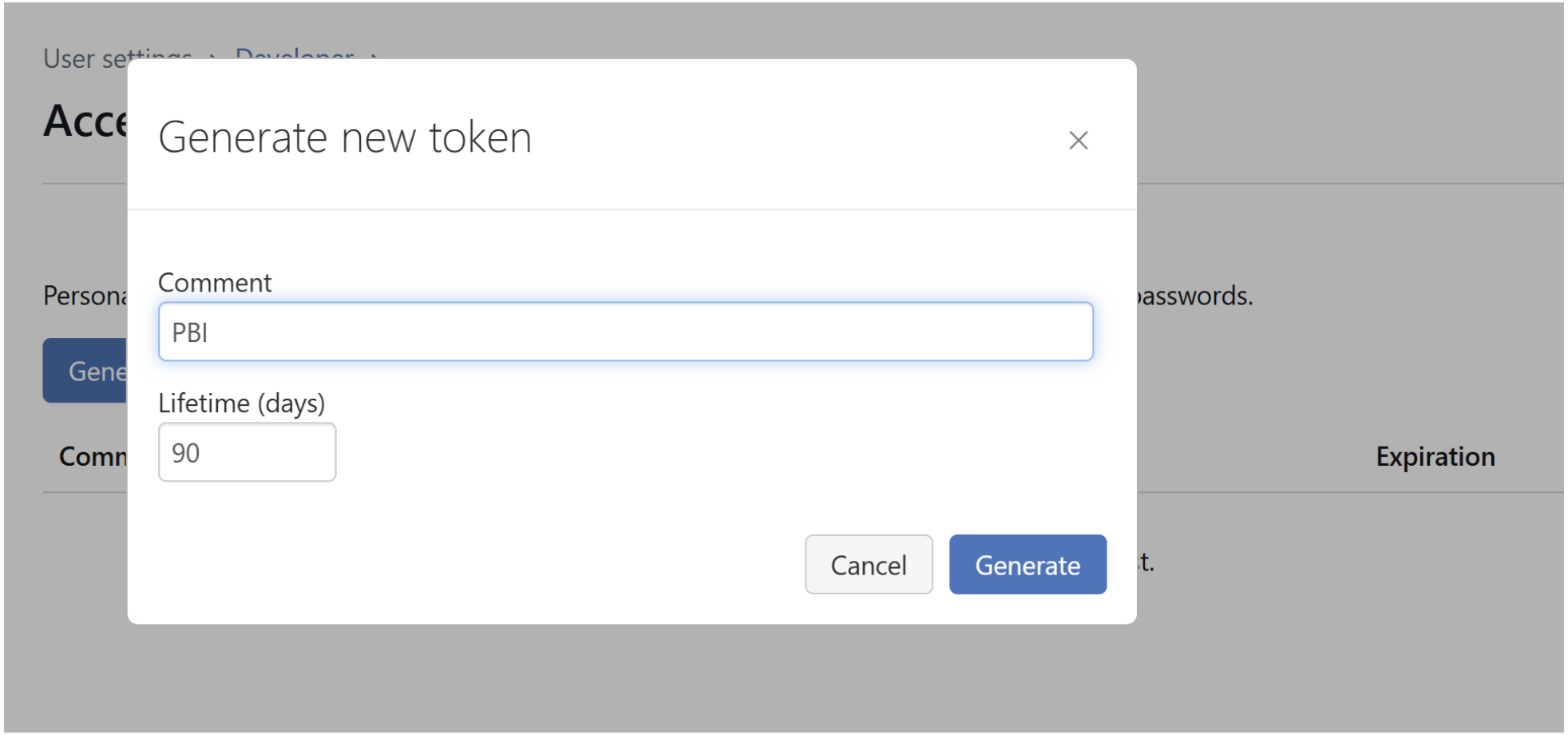
Quickly find meaningful insights within your data and easily build rich, visual analytic reports.

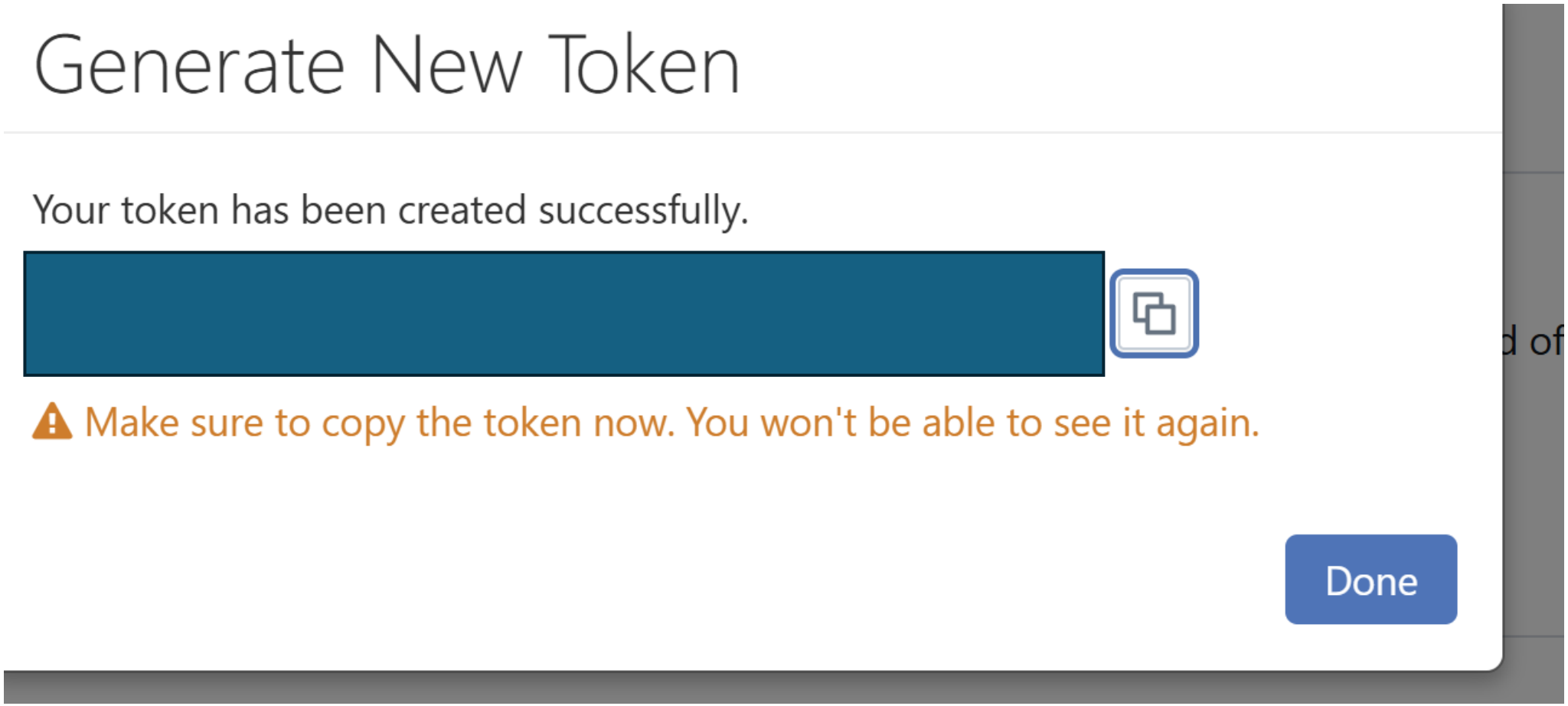
You can use Partner Connect to connect Power BI Desktop to a Databricks cluster or SQL warehouse. Select the target cluster or SQL warehouse, and then download and open the connection file (.pbids) to start Power BI Desktop. You must have Power BI Desktop version 2.99.563.0 or above installed.

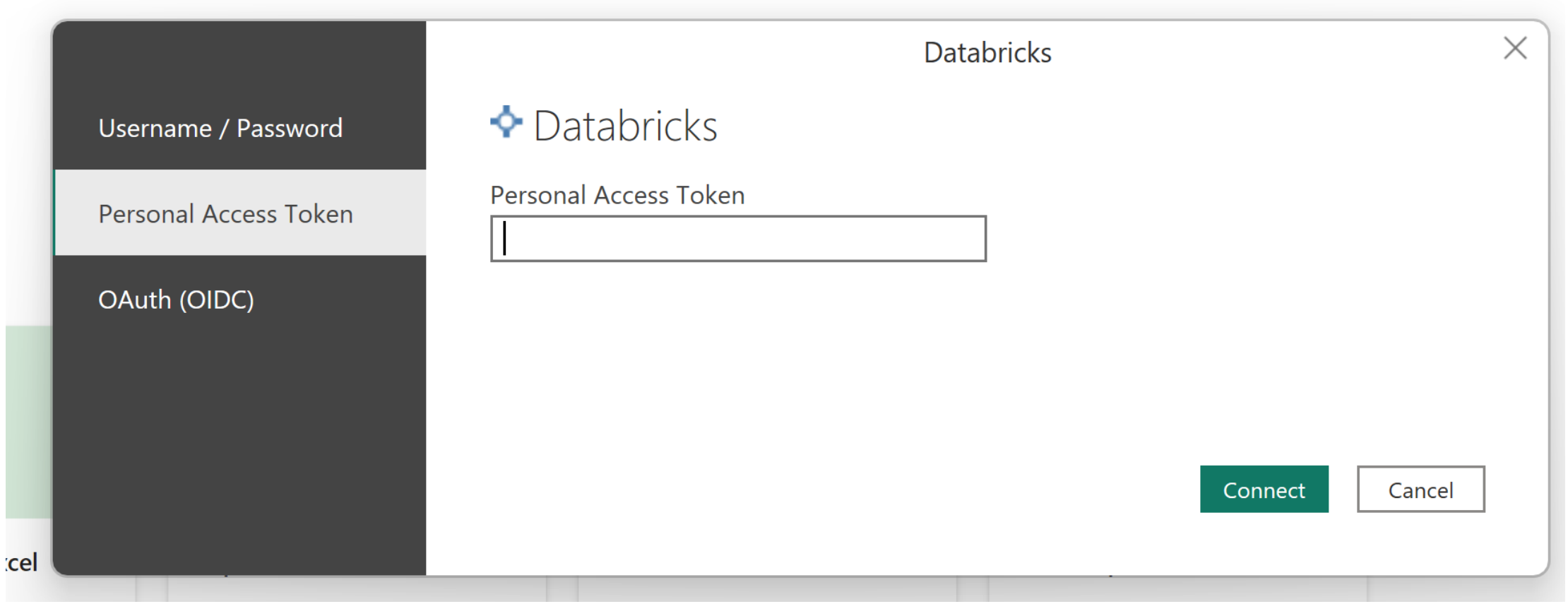
Learn more

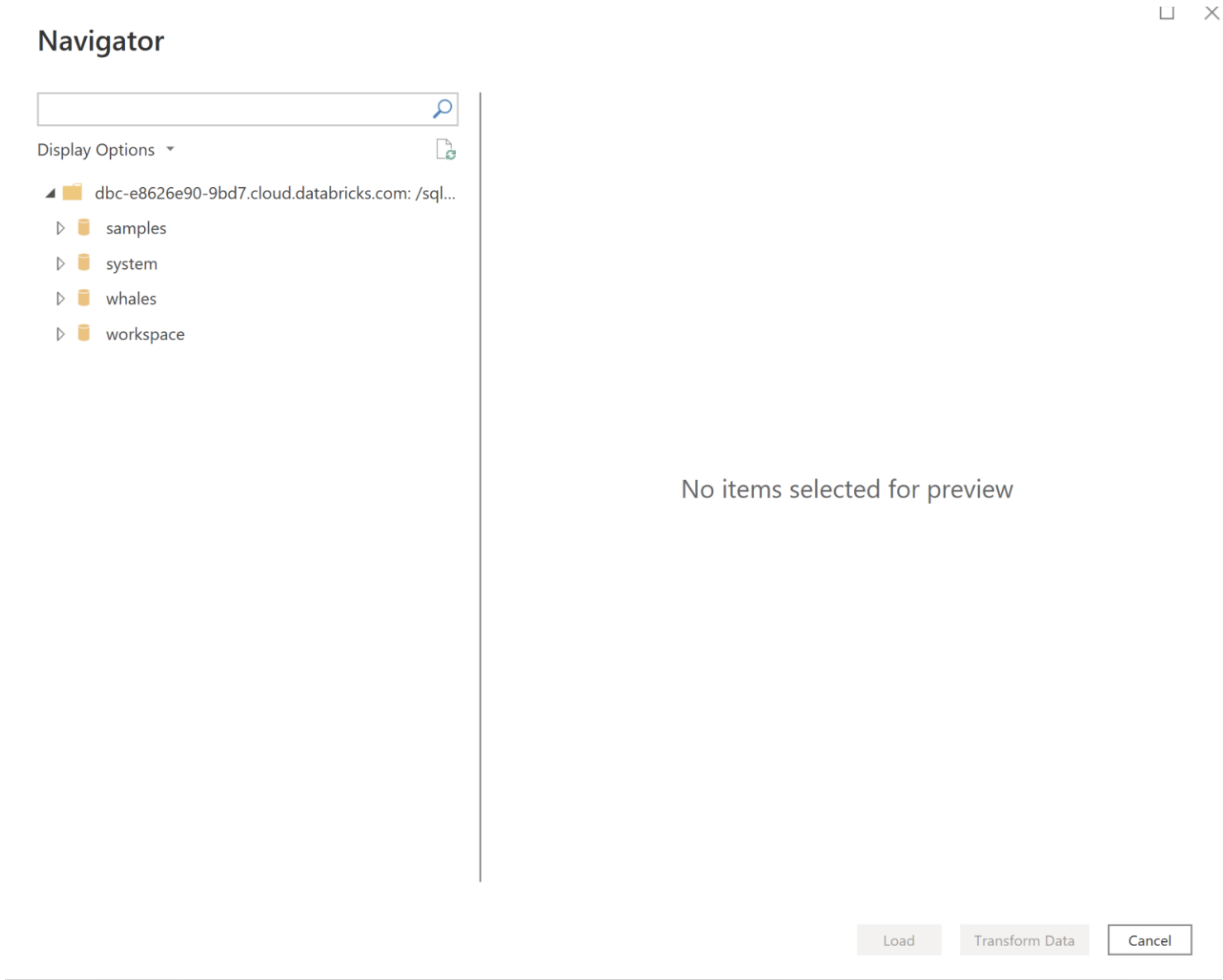
Compute Serverless Starter Warehouse

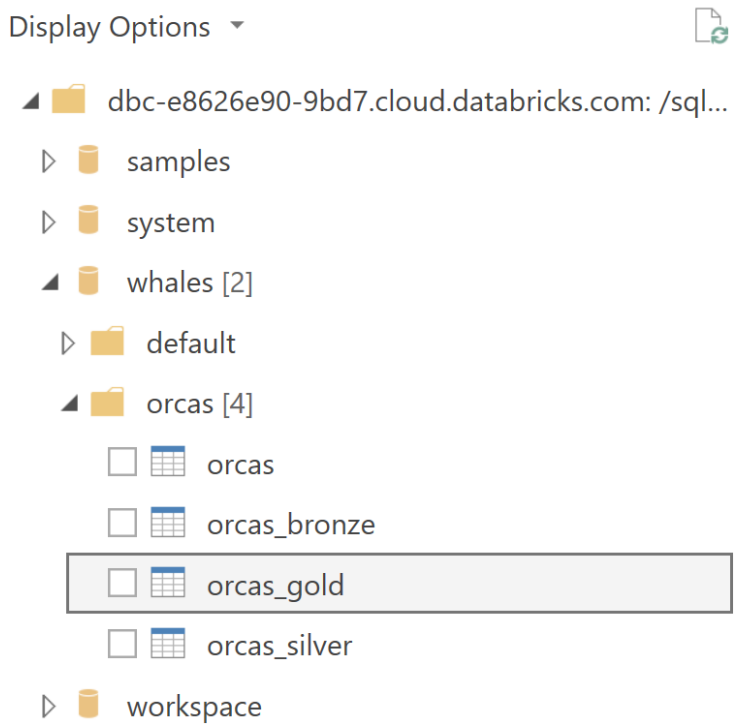
Download connection file | Cancel

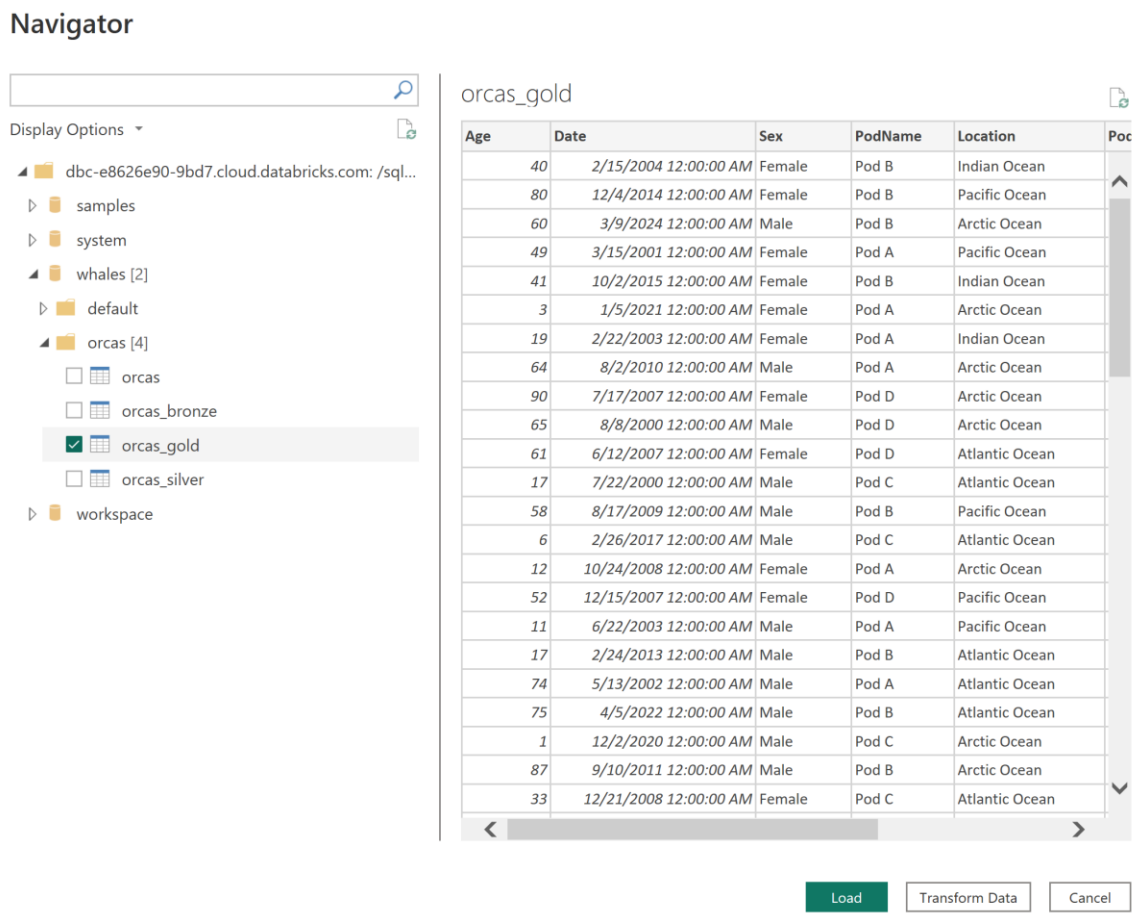

Generate new token
Comment
PBI
Lifetime (days)
90
Cancel
Generate
Expiration

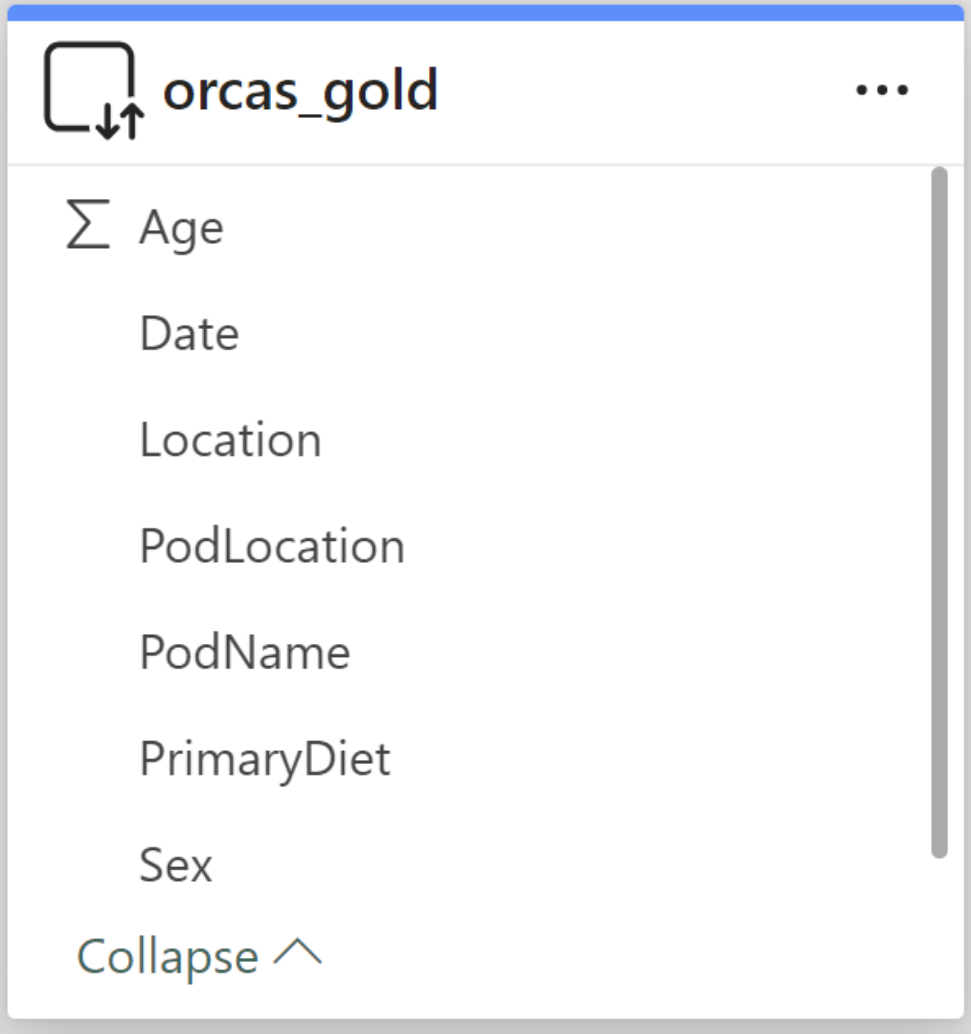

Generate New Token
Your token has been created successfully.
Make sure to copy the token now. You won't be able to see it again.
Done


Databricks
Username / Password
Personal Access Token
OAuth (OIDC)
Databricks
Personal Access Token
Connect
Cancel


Navigator
Display Options
dbc-e8626e90-9bd7.cloud.databricks.com: /sql...
samples
system
whales
workspace
No items selected for preview
Load
Transform Data
Cancel


Display Options
dbc-e8626e90-9bd7.cloud.databricks.com: /sql...
samples
system
whales [2]
default
orcas [4]
orcas
orcas_bronze
orcas_gold
orcas_silver
workspace


Navigator
Display Options
dbc-e8626e90-9bd7.cloud.databricks.com: /sql...
samples
system
whales [2]
default
orcas [4]
orcas
orcas_bronze
orcas_gold
orcas_silver
workspace
orcas_gold
Age Date Sex PodName Location Poc
40 2/15/2004 12:00:00 AM Female Pod B Indian Ocean
80 12/4/2014 12:00:00 AM Female Pod B Pacific Ocean
60 3/9/2024 12:00:00 AM Male Pod B Arctic Ocean
49 3/15/2001 12:00:00 AM Female Pod A Pacific Ocean
41 10/2/2015 12:00:00 AM Female Pod B Indian Ocean
3 1/5/2021 12:00:00 AM Female Pod A Arctic Ocean
19 2/22/2003 12:00:00 AM Female Pod A Indian Ocean
64 8/2/2010 12:00:00 AM Male Pod A Arctic Ocean
90 7/17/2007 12:00:00 AM Female Pod D Arctic Ocean
65 8/8/2000 12:00:00 AM Male Pod D Arctic Ocean
61 6/12/2007 12:00:00 AM Female Pod D Atlantic Ocean
17 7/22/2000 12:00:00 AM Male Pod C Atlantic Ocean
58 8/17/2009 12:00:00 AM Male Pod B Pacific Ocean
6 2/26/2017 12:00:00 AM Male Pod C Atlantic Ocean
12 10/24/2008 12:00:00 AM Female Pod A Arctic Ocean
52 12/15/2007 12:00:00 AM Female Pod D Pacific Ocean
11 6/22/2003 12:00:00 AM Male Pod A Pacific Ocean
17 2/24/2013 12:00:00 AM Male Pod B Atlantic Ocean
74 5/13/2002 12:00:00 AM Male Pod A Atlantic Ocean
75 4/5/2022 12:00:00 AM Male Pod B Atlantic Ocean
1 12/2/2020 12:00:00 AM Male Pod C Arctic Ocean
87 9/10/2011 12:00:00 AM Male Pod B Arctic Ocean
33 12/21/2008 12:00:00 AM Female Pod C Atlantic Ocean
Load
Transform Data
Cancel


orcas_gold
Age
Date
Location
PodLocation
PodName
PrimaryDiet
Sex
Collapse