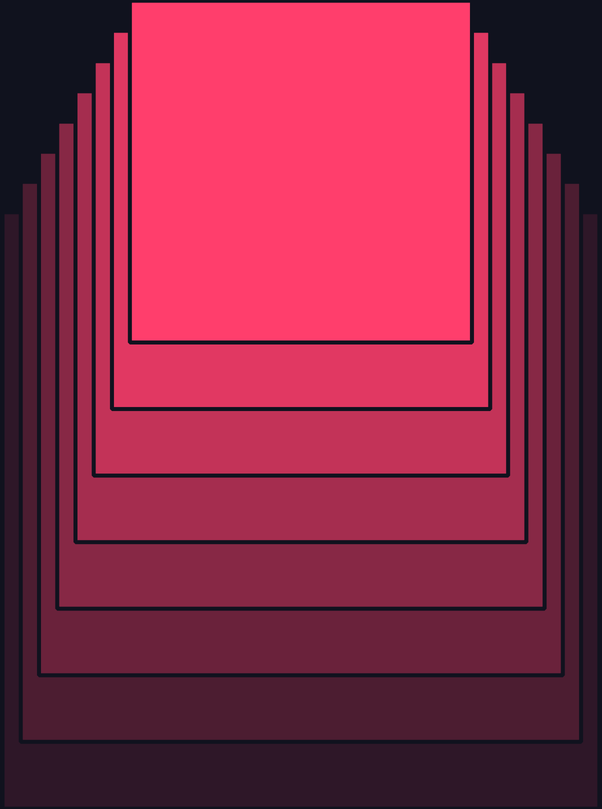
Product safe harbor statement

This information is provided to outline Databricks’ general product direction and is for **informational purposes only**. Customers who purchase Databricks services should make their purchase decisions relying solely upon services, features, and functions that are currently available. Unreleased features or functionality described in forward-looking statements are subject to change at Databricks discretion and may not be delivered as planned or at all

DATA+AI SUMMIT
BY databricks

FILE-BASED APPLICATIONS WITH VOLUMES IN UNITY CATALOG



Adriana Ispas, Databricks (R&D)
Adrian Ionescu, Databricks (R&D)

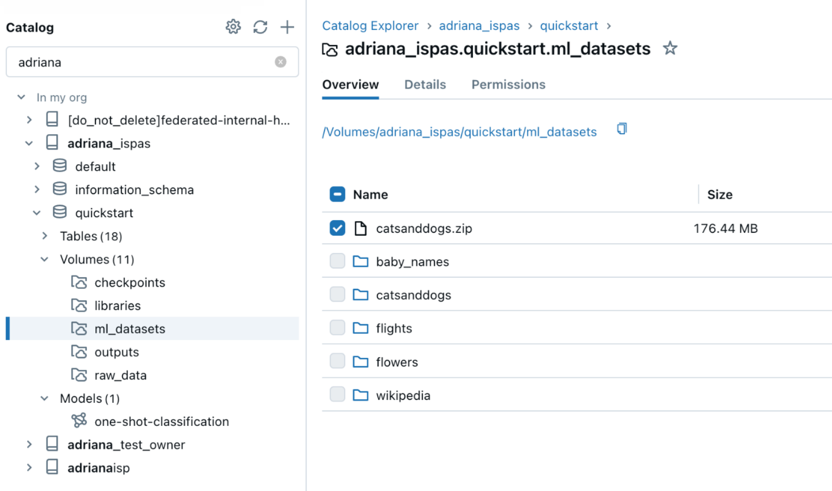
©2024 Databricks Inc. — All rights reserved

Volumes in Unity Catalog

Governance and management ofnon-tabular data

Collections of directories and files of arbitrary formats, structured, semi-structured & unstructured

…



Catalog

adriana

- In my org
  - [do_not_delete]federated-internal-h...
  - adriana_ispas
    - default
    - information_schema
    - quickstart
      - Tables (18)
      - Volumes (11)
        - checkpoints
        - libraries
        - ml_datasets
        - outputs
        - raw_data
      - Models (1)
        - one-shot-classification
  - adriana_test_owner
  - adrianaisp

Catalog Explorer › adriana_ispas › quickstart ›

adriana_ispas.quickstart.ml_datasets ☆

Overview | Details | Permissions

/Volumes/adriana_ispas/quickstart/ml_datasets

☐ Name	Size
☑ catsanddogs.zip	176.44 MB
☐ baby_names	
☐ catsanddogs	
☐ flights	
☐ flowers	
☐ wikipedia	

… cataloged inside schemas `<catalog>.<schema>.<volume>`, following the core principles of the UC object model

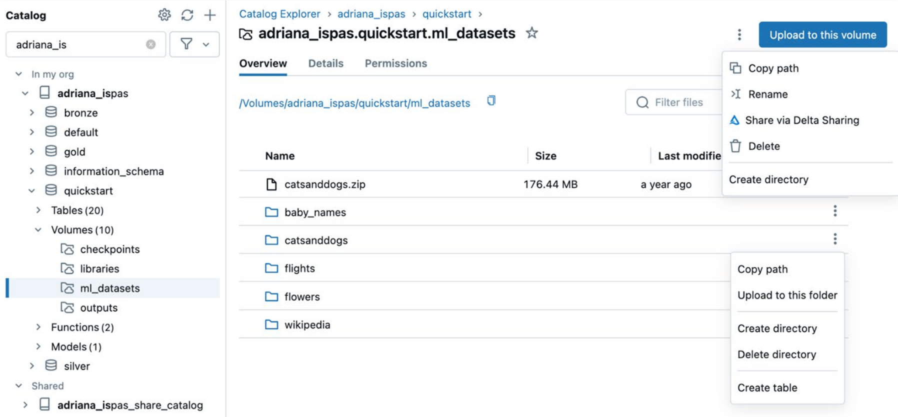
```sql
GRANT CREATE VOLUME ON SCHEMA my_schema
TO `domain admins`;

-- Create a managed volume
CREATE VOLUME my_catalog.my_schema.my_volume;

-- … or an external volume
CREATE VOLUME my_catalog.my_schema.my_external_volume;
LOCATION 's3://<el_bucket_path>/<sub_path>';

GRANT READ VOLUME, WRITE VOLUME
ON VOLUME my_volume
TO `user group`;
```

©2024 Databricks Inc. — All rights reserved

Volumes in Unity Catalog

Capabilities to access, store, process, and share files of any format

Manage or share files with the convenience of a user interface ...



/Volumes/adriana_ispas/quickstart/ml_datasets

Name	Size	Last modifie
catsanddogs.zip	176.44 MB	a year ago
baby_names		
catsanddogs		
flights		
flowers		
wikipedia		

... or using path-based access via APIs: Spark APIs, FUSE, dbutils, REST, SQL, CLI, Terraform, and more

```
spark.read.text("/Volumes/my_catalog/my_schema/my_volume/dat
a.txt").show()


SELECT * FROM
csv.`/Volumes/my_catalog/my_schema/my_volume/data.csv`


%sh curl http://<address>/text.zip >
/Volumes/my_catalog/my_schema/my_volume/tmp/text.zip

import os
os.listdir('/Volumes/my_catalog/my_schema/my_volume/my_dir')
```

Path: `[dbfs:]/Volumes/<catalog>/<schema>/<volume>/<path>`

©2024 Databricks Inc. — All rights reserved

Volumes in Unity Catalog

Seamlessly integrated across the platform

Track lineage for non-tabular data along with the ~~tabular data and~~ ML models

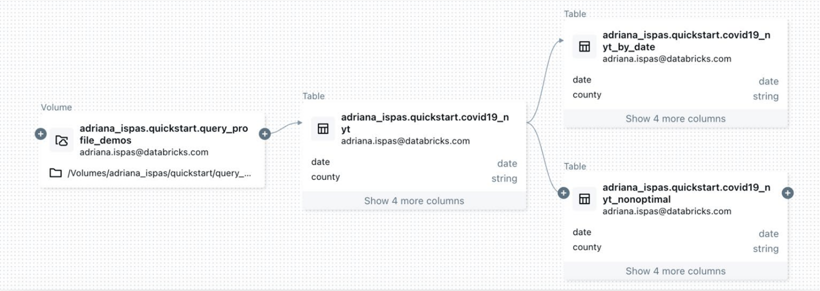
Share non-tabular data, along with tables, notebooks, ~~and AI models,~~ across clouds, regions, and accounts

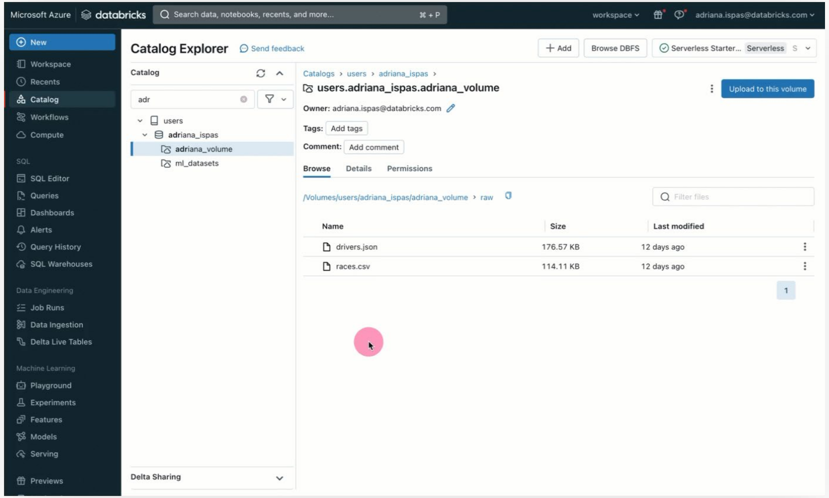
Configure Job file arrival triggers for storage ~~accessed through~~ Volumes

Store source files and configure cluster libraries, job ~~dependencies,~~ and init scripts

Create tables from files stored in Volumes using UI ~~wizards~~

Audit user actions with system tables and information schemas


Volume
adriana_ispas.quickstart.query_profile_demos
adriana.ispas@databricks.com
/Volumes/adriana_ispas/quickstart/query_...
Table
adriana_ispas.quickstart.covid19_nyt
adriana.ispas@databricks.com
date
county
date
string
Show 4 more columns
Table
adriana_ispas.quickstart.covid19_nyt_by_date
adriana.ispas@databricks.com
date
county
date
string
Show 4 more columns
Table
adriana_ispas.quickstart.covid19_nyt_nonoptimal
adriana.ispas@databricks.com
date
county
date
string
Show 4 more columns



Owner: adriana.ispas@databricks.com

Tags: Add tags

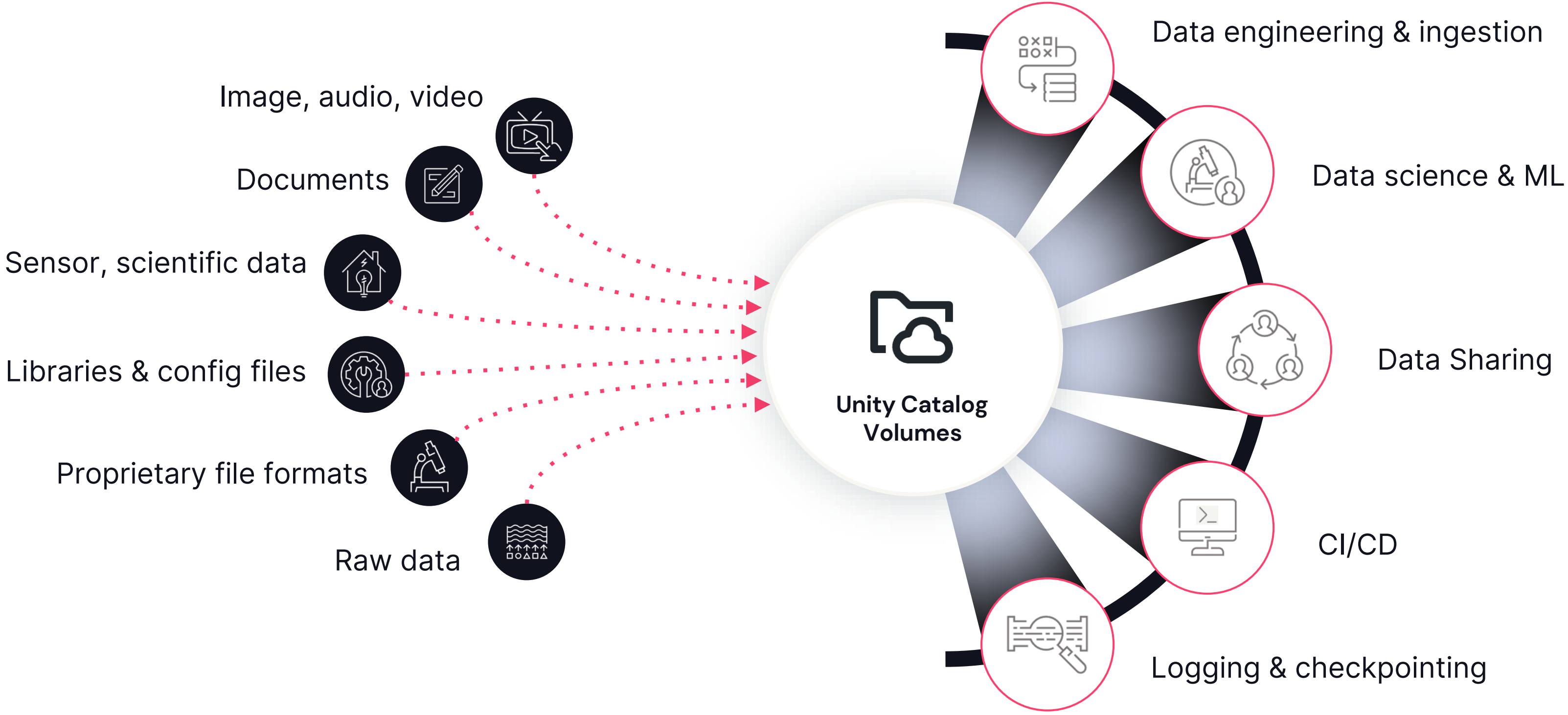
Comment: Add comment

/Volumes/users/adriana_ispas/adriana_volume › raw

Name	Size	Last modified
drivers.json	176.57 KB	12 days ago
races.csv	114.11 KB	12 days ago

Volumes use cases

Unlock new processing capabilities for arbitrary files



Non-tabular data

File-based workloads

VOLUMES IN ACTION:

1/ AI
2/ INGESTION

©2024 Databricks Inc. — All rights reserved

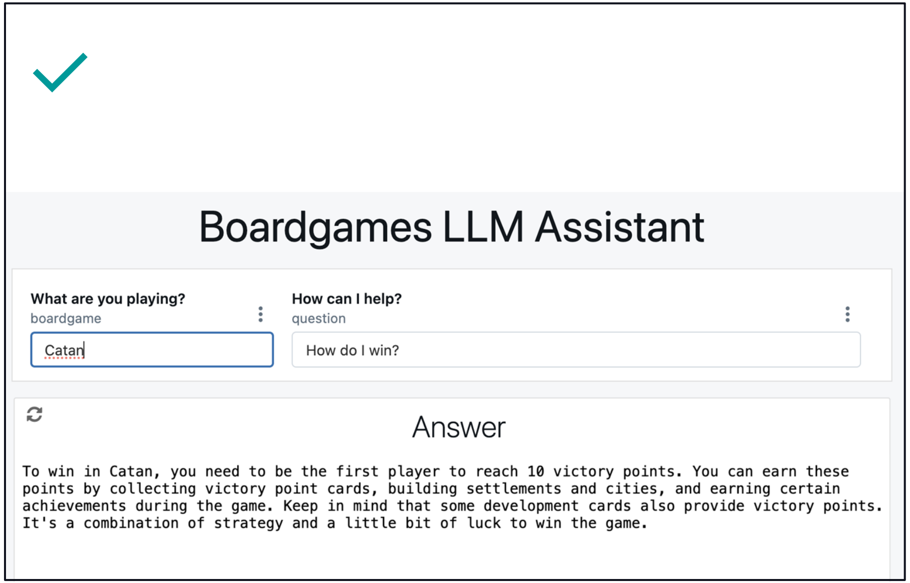
1/

USING VOLUMES FOR AI APPLICATIONS

©2024 Databricks Inc. — All rights reserved

Scenario/ Existing LLM

Knows about popular games ...



What are you playing?
boardgame
Catan

How can I help?
question
How do I win?

Answer

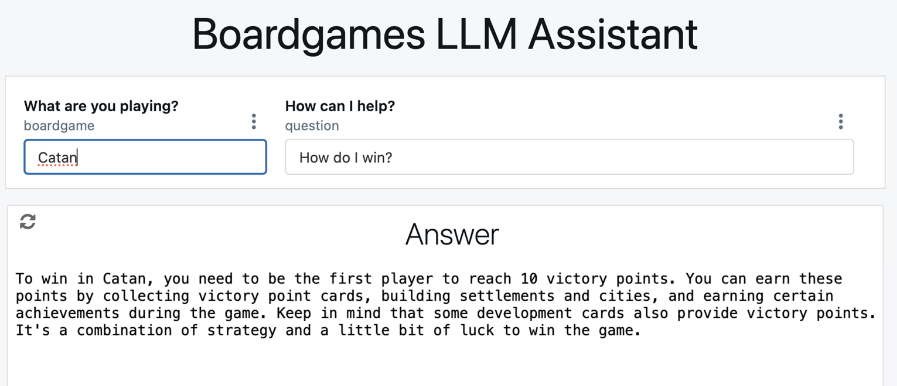
```
To win in Catan, you need to be the first player to reach 10 victory points. You can earn these
points by collecting victory point cards, building settlements and cities, and earning certain
achievements during the game. Keep in mind that some development cards also provide victory points.
It's a combination of strategy and a little bit of luck to win the game.
```

©2024 Databricks Inc. — All rights reserved

Scenario/ Existing LLM

… but does not know about the latest board games

✓



Boardgames LLM Assistant

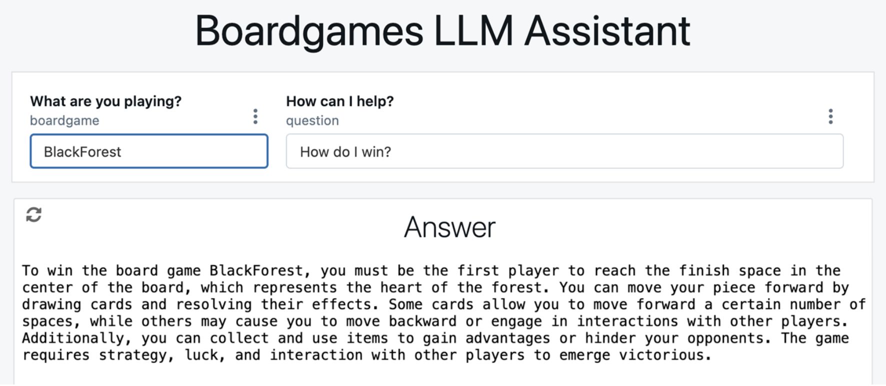
What are you playing?
boardgame
Catan

How can I help?
question
How do I win?

Answer

To win in Catan, you need to be the first player to reach 10 victory points. You can earn these points by collecting victory point cards, building settlements and cities, and earning certain achievements during the game. Keep in mind that some development cards also provide victory points. It's a combination of strategy and a little bit of luck to win the game.

✗



Boardgames LLM Assistant

What are you playing?
boardgame
BlackForest

How can I help?
question
How do I win?

Answer

To win the board game BlackForest, you must be the first player to reach the finish space in the center of the board, which represents the heart of the forest. You can move your piece forward by drawing cards and resolving their effects. Some cards allow you to move forward a certain number of spaces, while others may cause you to move backward or engage in interactions with other players. Additionally, you can collect and use items to gain advantages or hinder your opponents. The game requires strategy, luck, and interaction with other players to emerge victorious.

©2024 Databricks Inc. — All rights reserved

Scenario/ Board Games LLM

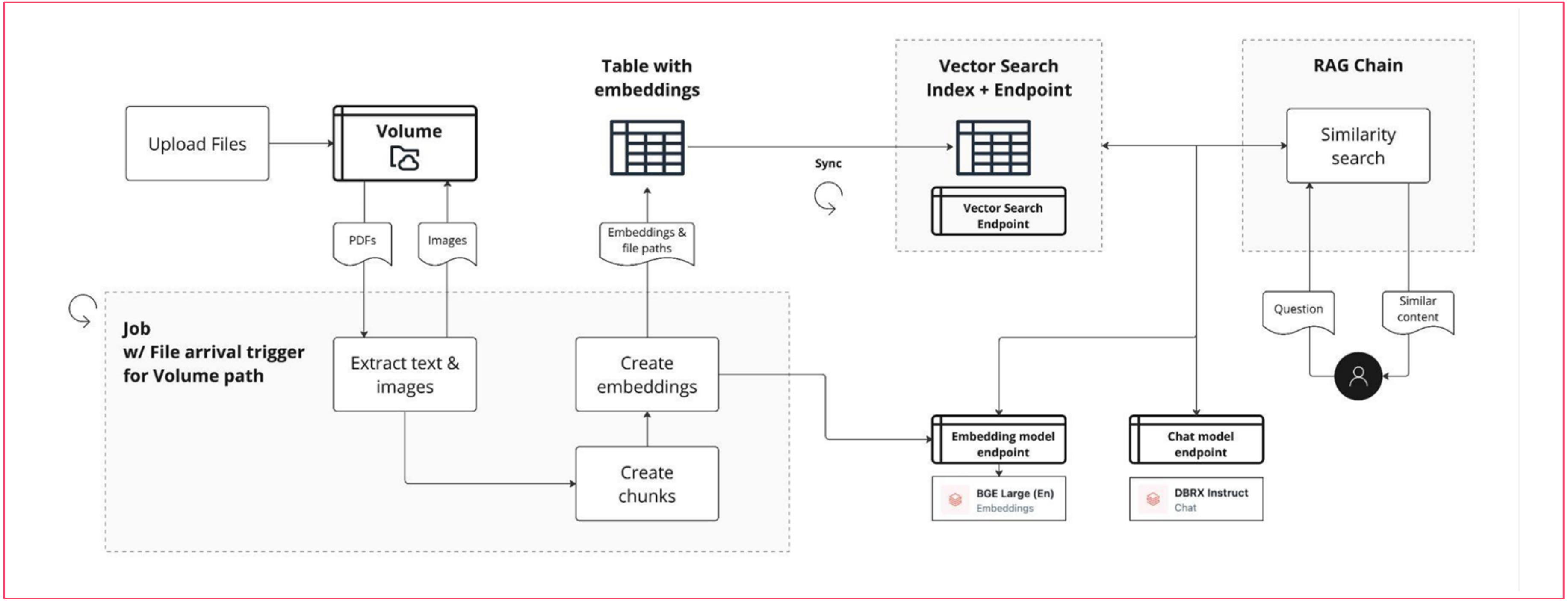
Specialize a model with information on latest Board Games

1. Upload and process unstructured data
2. Specialize an LLM with the new data
3. Search & get updated answers in real-time

©2024 Databricks Inc. — All rights reserved

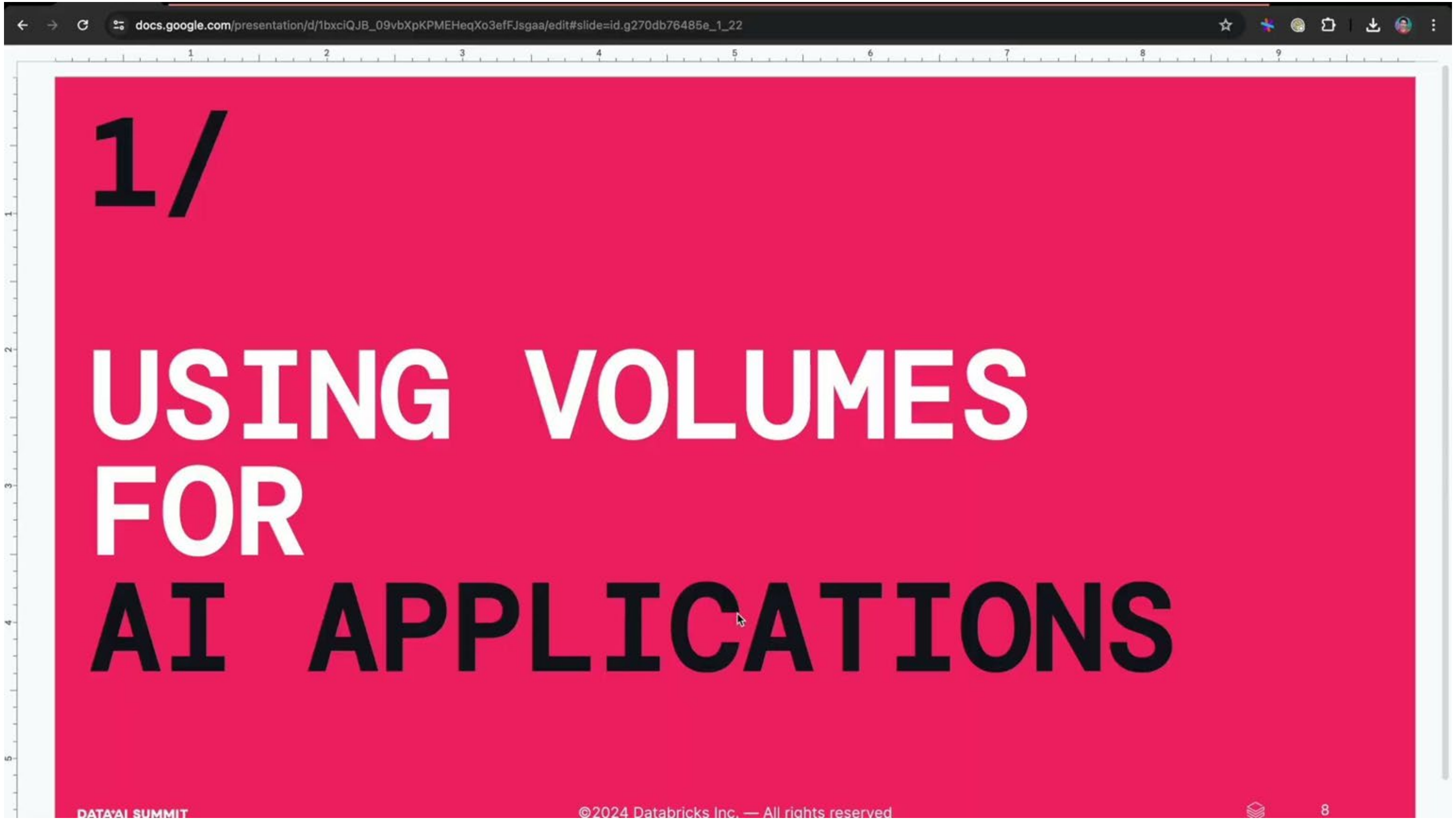
Specialize LLM with new PDF data

Components used



©2024 Databricks Inc. — All rights reserved

Live Demo



docs.google.com/presentation/d/1bxciQJB_09vbXpKPMEHeqXo3efFJsgaa/edit#slide=id.g270db76485e_1_22

1/

USING VOLUMES FOR AI APPLICATIONS

DATA+AI SUMMIT

©2024 Databricks Inc. — All rights reserved

8

©2024 Databricks Inc. — All rights reserved

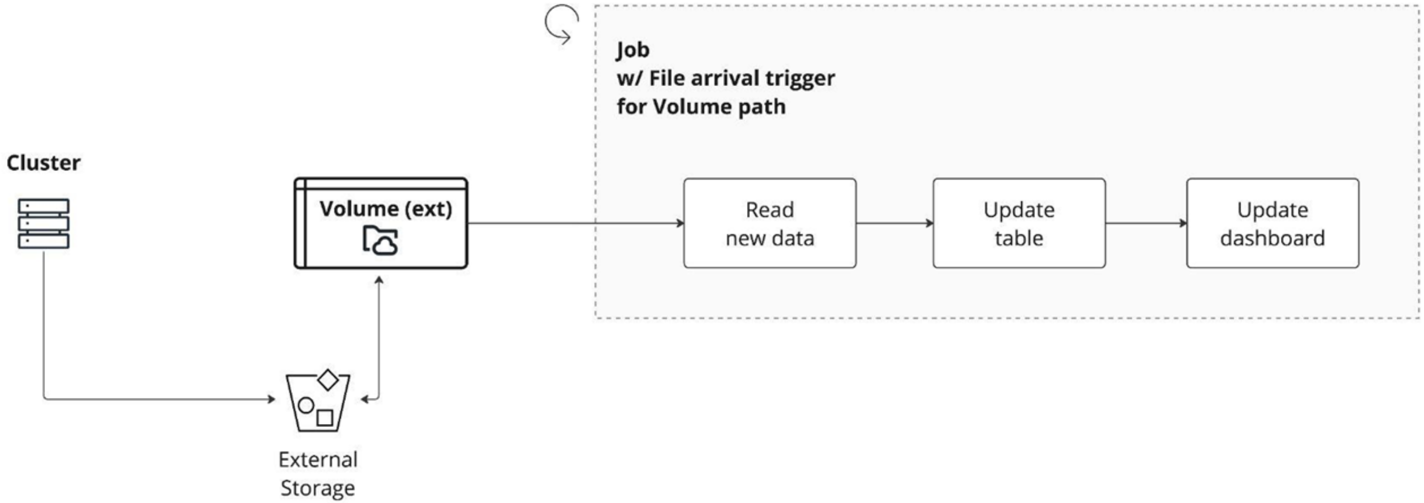
2/

USING VOLUMES FOR INGESTION

©2024 Databricks Inc. — All rights reserved

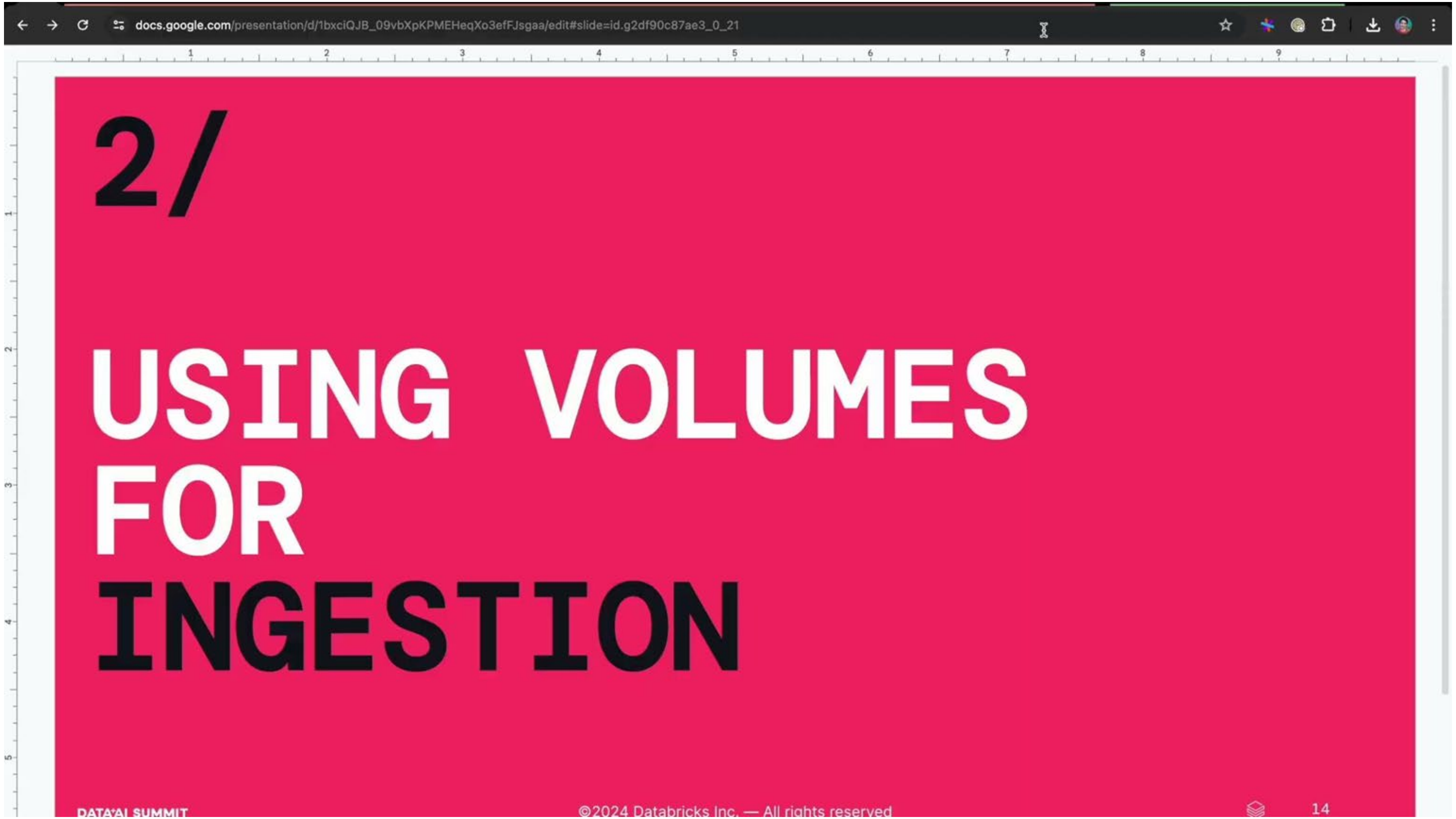
Ingest logs to a tabular format

Components used



©2024 Databricks Inc. — All rights reserved

Live Demo


docs.google.com/presentation/d/1bxciQJB_09vbXpKPMEHeqXo3efFJsgaa/edit#slide=id.g2df90c87ae3_0_21

2/

USING VOLUMES FOR INGESTION

DATA+AI SUMMIT

©2024 Databricks Inc. — All rights reserved

14

©2024 Databricks Inc. — All rights reserved

Ingestion using SQL

File operations for your data warehousing applications

SQL

```sql
-- 1. Upload local files to a Volume
PUT '/home/raw/data.json INTO '/Volumes/demo/ingestion/raw_data/nyc_taxi_trips_2024_05.json'

-- 2. List to assess the list of files
LIST '/Volumes/demo/ingestion/raw_data/'

-- 3. Add the files to an existing table
COPY INTO demo.ingestion.nyc_taxi_trips
FROM '/Volumes/demo/ingestion/raw_data/nyc_taxi_trips_2024_05.json'
FILEFORMAT = json
FORMAT_OPTIONS ('multiline' = 'true')

-- 4. Remove the files if no longer needed
REMOVE '/Volumes/demo/ingestion/raw_data/nyc_taxi_trips_2024_05.json'
```

©2024 Databricks Inc. — All rights reserved

WRAP UP

©2024 Databricks Inc. — All rights reserved

Volumes for file-based applications

Recap and recommendations

- → Use Volumes to kick start your AI, ingestion and CI/CD pipelines
  - ✓ Unified governance for tabular and non-tabular data sets
  - ✓ Streamlined developer experience, across languages and platform capabilities
  - ✓ Automatic syncing of application logic to your source data sets
  - ✓ End-to-end lineage across applications
  - ✓ Audit of user actions
- → Use managed Volumes for convenience and ease of use; use external Volumes for existing data and interoperability with other systems and tools
- → Use Volume Sharing for collaboration across regions, Clouds and accounts; implicit cross-workspace sharing within the same metastore

©2024 Databricks Inc. — All rights reserved

DATA+AI SUMMIT

Questions?

©2024 Databricks Inc. — All rights reserved

Learn more at the summit!

Databricks
Events App

Tells us what you think

- We kindly request your valuable feedback on this session.
- Please take a moment to rate and share your thoughts about it.
- You can conveniently provide your feedback and rating through the **Mobile App**.

What to do next?

- Discover more related sessions in the mobile app!
- Visit the Demo Booth: Experience innovation firsthand!
- More Activities: Engage and connect further at the Databricks Zone!

Get trained and certified

- Visit the Learning Hub Experience at Moscone West, 2nd Floor!
- Take complimentary certification at the event; come by the Certified Lounge
- Visit our Databricks Learning website for more training, courses and workshops! databricks.com/learn

Volumes in Unity Catalog

Governance and management ofnon-tabular data

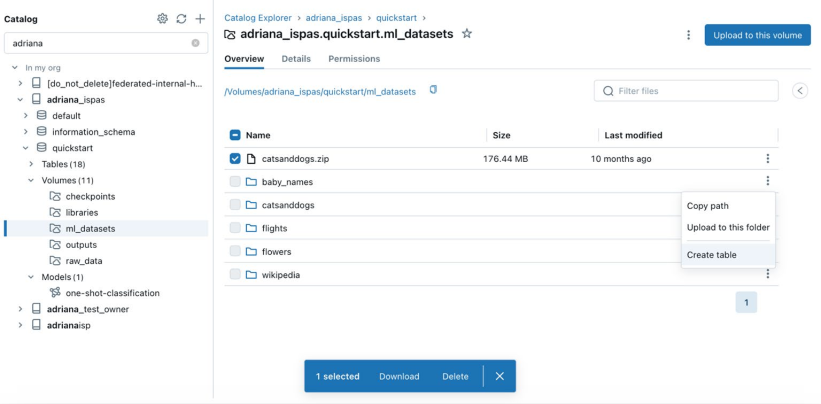
Govern collections of directories and files of arbitrary ~~formats; structured~~, semi-structured & unstructured

Acting as a logical storage unit; abstraction layer over ~~Cloud object storage~~

Provide capabilities to access, store, process, and share ~~files in any format~~

Work with files via UI, Spark APIs, FUSE, dbutils, REST, ~~SQL, Databricks~~ CLI, Terraform, and more

Track lineage for non-tabular data along with the tabular data and ML models



Catalog Explorer › adriana_ispas › quickstart ›

adriana_ispas.quickstart.ml_datasets ☆

/Volumes/adriana_ispas/quickstart/ml_datasets

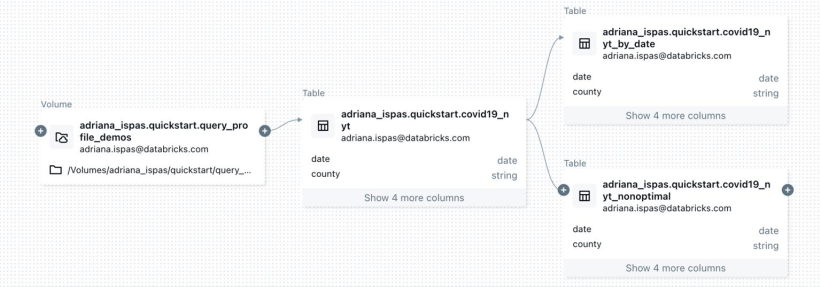
☐ Name	Size	Last modified
☑ catsanddogs.zip	176.44 MB	10 months ago
☐ baby_names		
☐ catsanddogs		
☐ flights		
☐ flowers		
☐ wikipedia		

Catalog

adriana

- In my org
  - [do_not_delete]federated-internal-h...
  - adriana_ispas
    - default
    - information_schema
    - quickstart
      - Tables (18)
      - Volumes (11)
        - checkpoints
        - libraries
        - ml_datasets
        - outputs
        - raw_data
      - Models (1)
        - one-shot-classification
  - adriana_test_owner
  - adrianaisp

Upload to this volume
Overview
Details
Permissions
Filter files
Copy path
Upload to this folder
Create table
1
1 selected
Download
Delete



Volume
adriana_ispas.quickstart.query_profile_demos
adriana.ispas@databricks.com
/Volumes/adriana_ispas/quickstart/query_...
Table
adriana_ispas.quickstart.covid19_nyt
adriana.ispas@databricks.com
date
county
date
string
Show 4 more columns
Table
adriana_ispas.quickstart.covid19_nyt_by_date
adriana.ispas@databricks.com
date
county
date
string
Show 4 more columns
Table
adriana_ispas.quickstart.covid19_nyt_nonoptimal
adriana.ispas@databricks.com
date
county
date
string
Show 4 more columns

©2024 Databricks Inc. — All rights reserved

Managed & external volume types

Both represent *volumes* of storage in a Cloud object storage location

Managed Volume

```
CREATE VOLUME <catalog>.<schema>.<volume_name>
```

- Store files in the default storage location for the parent schema
- Fast to set up – no need to configure Cloud storage access & credentials
- Convenient for adding new data, e.g., upload files for quick data explorations.

External Volume

```
CREATE EXTERNAL VOLUME <catalog>.<schema>.<volume_name>
LOCATION 's3://<external_location_path>/<directory>';
```

- Store files in an external cloud storage location referenced at Volume creation
- Stage access to pre-existing data w/o copying it, e.g., image or video data deposited by IoT or medical devices
- Provide direct file access from outside Databricks.

©2024 Databricks Inc. — All rights reserved

Working with files

Familiar file system tools and utilities for working with non-tabular data

Usage	Example

Databricks file system utilities	`dbutils.fs.ls("/Volumes/my_catalog/my_schema/my_volume/")`
	`%fs ls /Volumes/my_catalog/my_schema/my_volume/`

Apache Spark™ APIs	`spark.read.text("/Volumes/my_catalog/my_schema/my_volume/data.txt").show()`
Apache Spark™ SQL / DBSQL	``SELECT * FROM csv.`/Volumes/my_catalog/my_schema/my_volume/data.csv` ``

Pandas	`import pandas as pd`
	`df = pd.read_csv('/Volumes/my_catalog/my_schema/my_volume/data.csv')`

Shell commands via %sh	`%sh curl http://<address>/text.zip > /Volumes/my_catalog/my_schema/my_volume/tmp/text.zip`
Library installs using %pip	`%pip install /Volumes/my_catalog/my_schema/my_volume/my_library.whl`

Operating system file utilities	`import os`
	`os.listdir('/Volumes/my_catalog/my_schema/my_volume/path/to/directory')`
	
	`from zipfile import ZipFile`
	`zf = ZipFile('/Volumes/my_catalog/my_schema/my_volume/text.zip', 'r')`

©2024 Databricks Inc. — All rights reserved

System tables and information schema

Answer questions about the usage of Volumes

Who has access to this volume?

```sql
SELECT grantee, volume_name, privilege_type
FROM system.information_schema.volume_privileges
WHERE volume_name = "my_volume";
```

Who accessed this volume in the past 7 days?

```sql
SELECT user_identity.email, request_params.operation
FROM system.access.audit
WHERE event_date >= current_date() - INTERVAL 7 DAYS
AND action_name LIKE "%generateTemporaryVolumeCredential%"
AND request_params.volume_full_name =
"my_catalog.my_schema.my_volume";
```

Is this volume used to create any tables?

```sql
SELECT target_table_full_name
FROM  system.lineage.table_lineage
WHERE source_type = "PATH" AND target_type = "TABLE" AND
source_path LIKE
"%Volumes/my_catalog/my_schema/my_volume%";
```

Who has last downloaded this particular file?

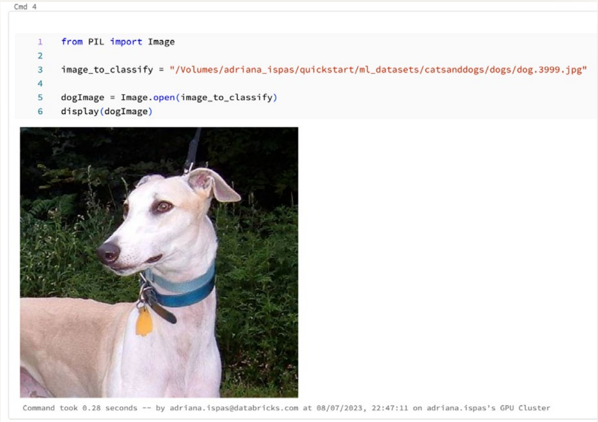
```sql
SELECT user_identity.email, max(event_time) as last_access
FROM system.access.audit
WHERE service_name LIKE "filesystem"
AND action_name = "filesGet"
AND request_params.path LIKE
"/Volumes/my_catalog/my_schema/my_volume/data.csv"
GROUP BY 1
ORDER BY last_access DESC;
```

©2024 Databricks Inc. — All rights reserved

Volumes in Unity Catalog

Example use cases

- ML on large collections of unstructured data such as image, audio, video, text or PDF files
- Upload and query non-tabular data files for data exploration
- Work with tools w/o native support for Cloud object storage APIs, that expect files in the local file system on cluster machines
- Stage and pre-process raw data files in early stages of ingestion before loading into tables, e.g., using Autoloader or COPY INTO
- Store libraries, certificates, or config files of arbitrary formats, e.g., .whl or .txt, for cluster, notebook-scoped, or job libs
- Share large collections of files with other users within or across workspaces
- … many more.



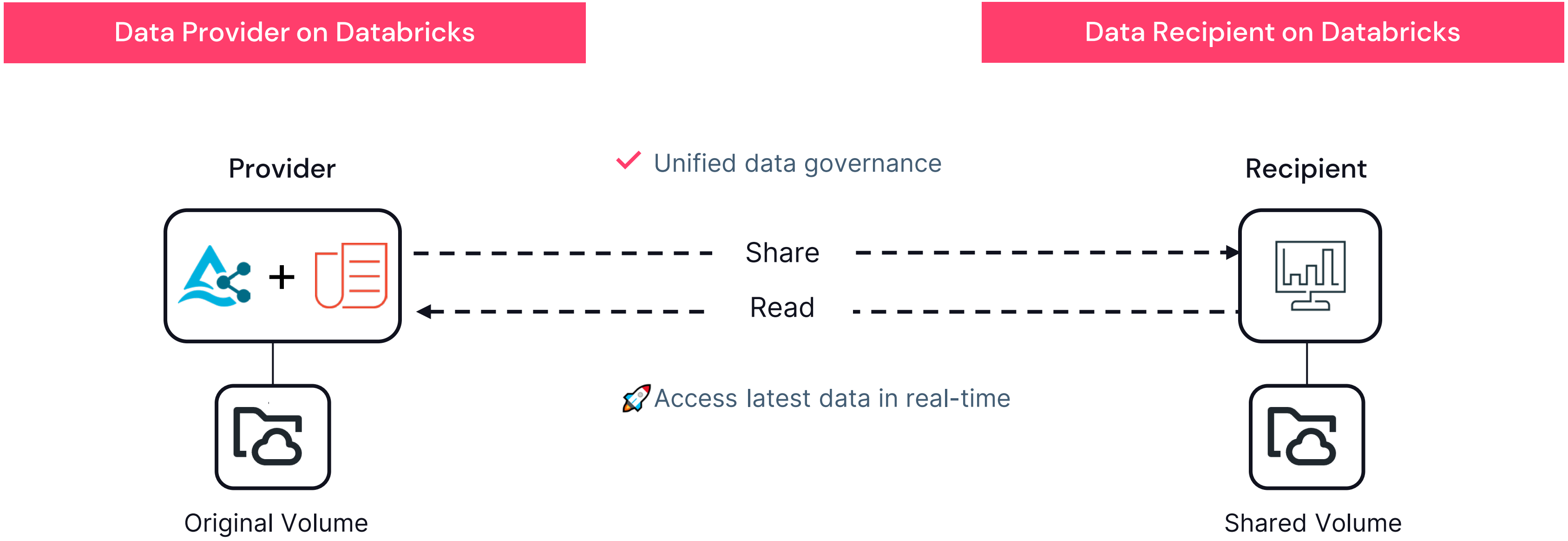
```
from PIL import Image

image_to_classify = "/Volumes/adriana_ispas/quickstart/ml_datasets/catsanddogs/dogs/dog.3999.jpg"

dogImage = Image.open(image_to_classify)
display(dogImage)
```

```
Command took 0.28 seconds -- by adriana.ispas@databricks.com at 08/07/2023, 22:47:11 on adriana.ispas's GPU Cluster
```

©2024 Databricks Inc. — All rights reserved

Volume Sharing

Delta Sharing protocol for Volumes


Data Provider on Databricks
Data Recipient on Databricks
Provider
✓ Unified data governance
Recipient
+
Share
Read
🚀Access latest data in real-time
Original Volume
Shared Volume

©2024 Databricks Inc. — All rights reserved