

# Data Mesh Workshop

---

GO  
DATA  
DRIVEN

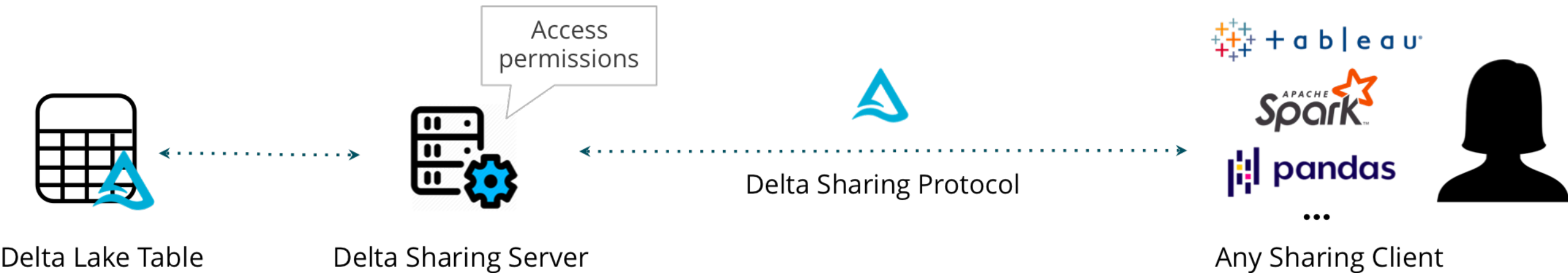
Kris Geusebroek  
Robert Rodger

# Delta Sharing

# Delta Sharing

Data Provider

Data Recipient



GO   
DATA  
DRIVEN

# A Quick word on Delta Lake

- Delta Sharing works with tables in the Delta Lake format

# The Delta Lake format

- Open Source storage layer for your data lake
- Based upon the parquet file format
- Adds the `\_delta\_log` metadata

# The Delta Lake format

- Brings you ACID transactions on your data (just like a database)
- Checkpoints
- Transaction log and Time travel

# Delta Sharing server

- Delta Sharing server is not a bottleneck in sharing data
- You don't share security details to your data

# No Bottleneck

- Data does not pass your Delta sharing server



# Security details

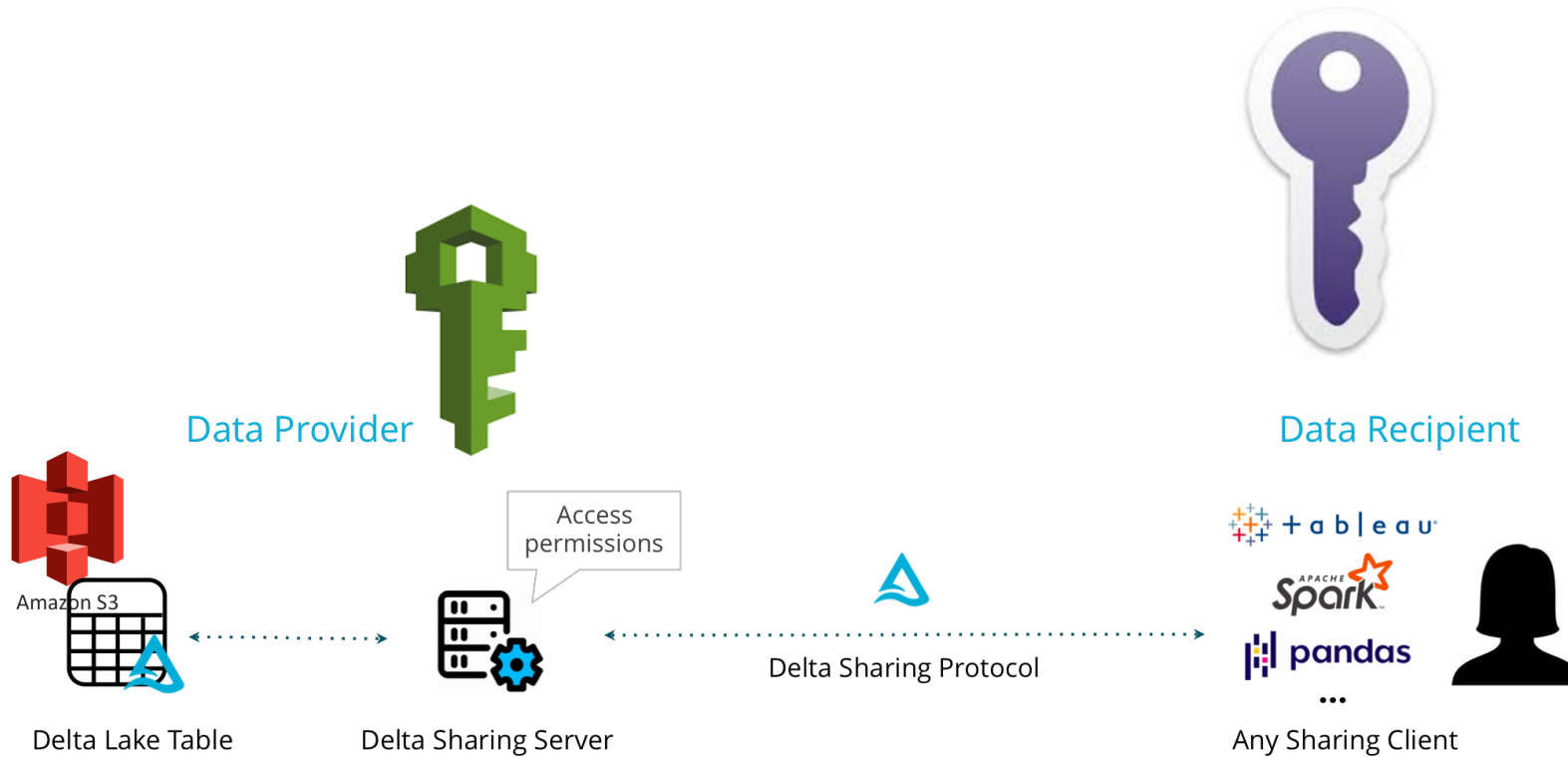
- Delta Sharing gives you signed url's to the dataset

<s3://demodata/silver/world/cities/part-00000-file1.snappy.parquet>

->

[http://s3server:4563/demodata/silver/world/cities/part-00000-file1.snappy.parquet?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20211101T200043Z&X-Amz-SignedHeaders=host&X-Amz-Expires=900&X-Amz-Credential=test%2F20211101%2Fus-east-1%2Fs3%2Faws4\\_request&X-Amz-Signature=e2290dcf59172649362a4e982889ab2e34e633ca9ae3b34a2e3bf79ce7ae386a](http://s3server:4563/demodata/silver/world/cities/part-00000-file1.snappy.parquet?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20211101T200043Z&X-Amz-SignedHeaders=host&X-Amz-Expires=900&X-Amz-Credential=test%2F20211101%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=e2290dcf59172649362a4e982889ab2e34e633ca9ae3b34a2e3bf79ce7ae386a)

# Delta Sharing



# Delta Sharing protocol

- Delta Sharing concepts:
  - Share: Combination of different datasets to share
  - Schema: Subdomain inside a share
  - Table: Delta table storage location

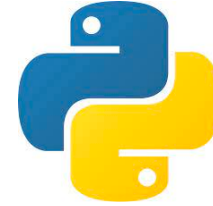
# Demo Environment



Amazon S3



Delta sharing server



GO  
DATA  
DRIVEN

# Demo: Delta Sharing Protocol

# Reading data with Delta Sharing

- The profile file
  - Endpoint
  - Authorization token
- Built your table url
  - Profile file + `#share.schema.table`
  - `load_as_pandas`
  - `format("deltaSharing").load`

# Demo: Read data with python

# Demo: Read data with spark



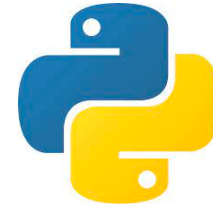
# Delta Sharing the future

- Support Azure filesystem (next version 0.3.0)
- Support adding shares/schemas/tables dynamically (Available in the rust based implementation called riverbank)
- Support GFS

# Delta Sharing workshop

- Instruqt
  - <https://play.instruqt.com/>
  - Each participant get's a separate environment

# Recap: Demo Environment



Amazon S3



Delta sharing server



GO  
DATA  
DRIVEN

# Workshop

## Challenges:

- Verify running services
- Environment in depth
- Protocol notebook
- Protocol terminal
- Read data python notebook
- Read data python terminal
- Read data spark notebook
- Read data spark terminal
- Playtime

# Instruqt: Workshop environment

<https://play.instruqt.com/godatadriven/tracks/datamesh-workshop-2021>

Start the track and wait for about 2 minutes to let the virtual machine boot.

When the machine is ready hit **Start** to get started

# Challenge 1

The first challenge is just a simple docker command to verify that all 6 services are running correctly

# Challenge 2

This next challenge is to get you a bit more familiar with how this workshop environment is setup.

We open a Jupiter notebook where a explanation of the environment is available.

# Challenge 3

This next challenge is to guide you through the Delta Sharing Protocol.

There is a Jupiter notebook (same as the one demoed) to guide you.



# Challenge 4

This next challenge is to guide you through the Delta Sharing Protocol again.

This time without a notebook but through plain old curl requests.

A few hints are available in the challenge sidebar. Feel free to experiment a bit further.

# Challenge 5

This next challenge is to guide you through reading data with python with Delta Sharing.

There is a Jupiter notebook (same as the one demoed) to guide you.

# Challenge 6

This next challenge is to guide you through reading data with python with Delta Sharing.

This time without a notebook but through starting a python interpreter and executing code yourself.

# Challenge 7

This next challenge is to guide you through reading data with spark with Delta Sharing.

There is a Jupiter notebook (same as the one demoed) to guide you.

# Challenge 8

This next challenge is to guide you through reading data with spark with Delta Sharing.

This time without a notebook but through starting a spark shell and executing code yourself.

# Challenge 9

This is a free format challenge.

Extend your knowledge by extending the previous challenges. Or explore all files available in the Editor tab to get a better understanding how this environment works together.

# Instruqt: Workshop environment (Recap)

<https://play.instruqt.com/godatadriven/tracks/datamesh-workshop-2021>

Start the track and wait for about 2 minutes to let the virtual machine boot.

When the machine is ready hit **Start** to get started

# Questions?

GO   
DATA  
DRIVEN



# Material

<https://github.com/godatadriven/datamesh>