Delta Sharing

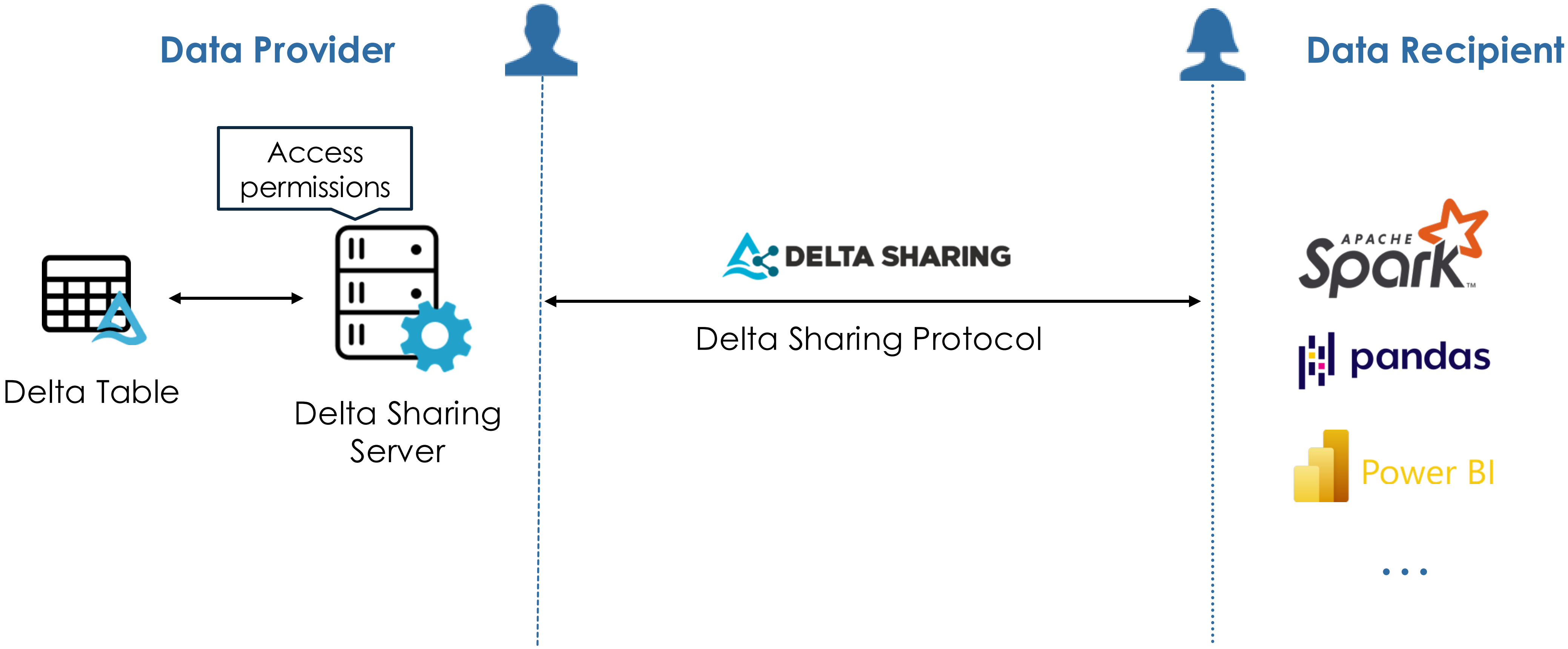
Learning Objectives

- What is Delta Sharing
- How it works under the hood
- Costs and limitations

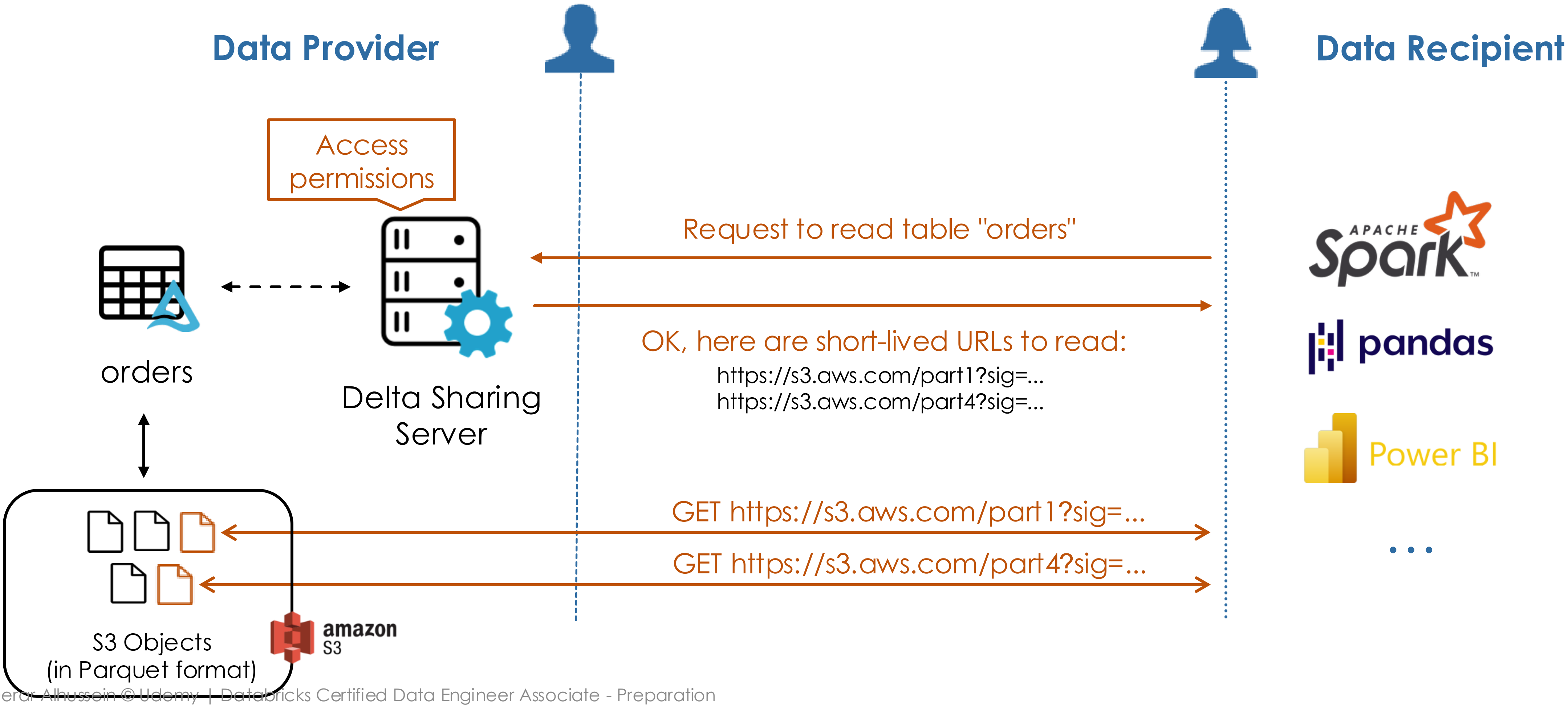
Delta Sharing

- Open protocol for secure data sharing
- Share live data with organizations regardless of computing platforms

How it Works


Data Provider
Data Recipient
Access permissions
Delta Table
Delta Sharing Server
DELTA SHARING
Delta Sharing Protocol
APACHE Spark
pandas
Power BI
. . .

Under the Hood



Data Provider

Data Recipient

Access permissions

orders

Delta Sharing Server

Request to read table "orders"

OK, here are short-lived URLs to read:

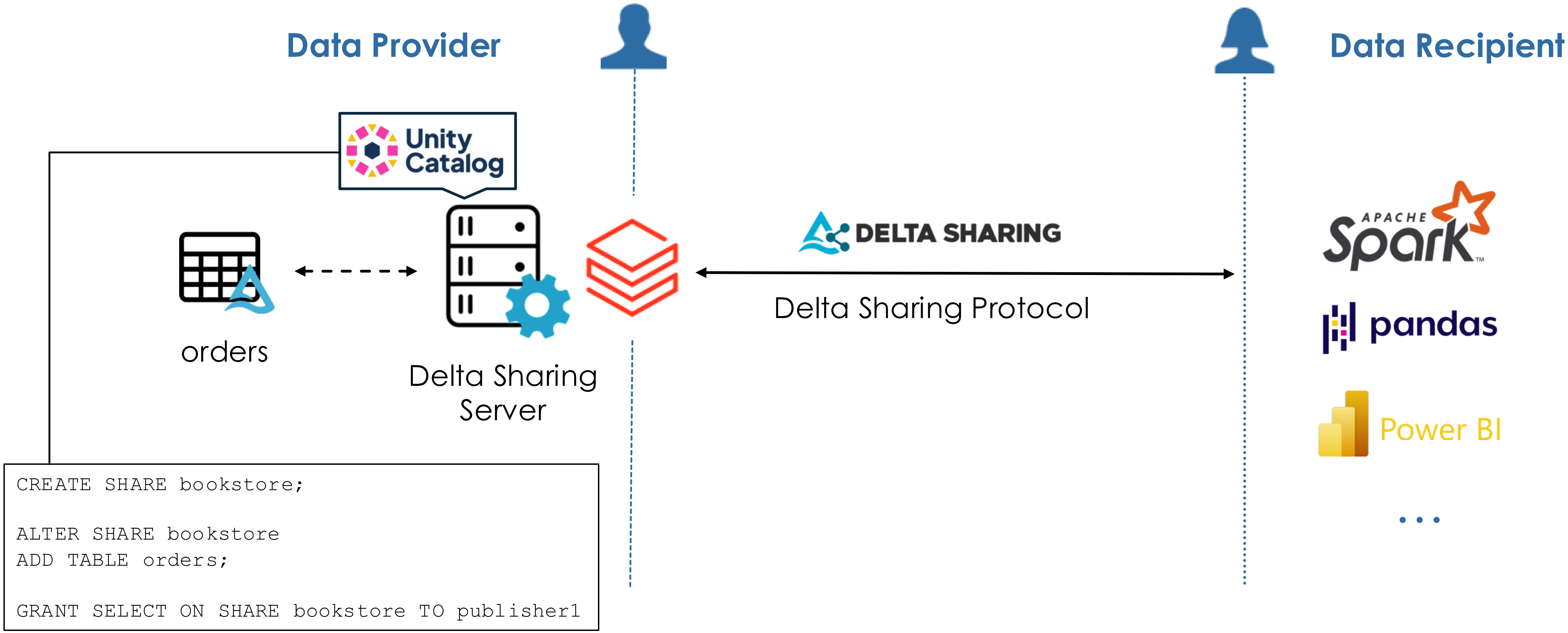
https://s3.aws.com/part1?sig=...
https://s3.aws.com/part4?sig=...

GET https://s3.aws.com/part1?sig=...

GET https://s3.aws.com/part4?sig=...

S3 Objects
(in Parquet format)

Delta Sharing on Databricks



```
CREATE SHARE bookstore;

ALTER SHARE bookstore
ADD TABLE orders;

GRANT SELECT ON SHARE bookstore TO publisher1
```

Delta Sharing on Databricks

- The Databricks-to-Databricks sharing protocol
  - Share data between Databricks clients
  - Collection of tables, views, volumes, and notebooks
- The Databricks open sharing protocol
  - Share data with users on any computing platform

Costs

- Delta Sharing does not require data replication

- Egress cost
  - Within a region: no egress cost
  - Cross-clouds or cross-regions: cloud vendor charges data egress fees, instead:
    - Clone the shared data to local regions
    - Share data from Cloudflare R2

Limitations

- Read-Only Access Model
- Data Format Constraints
  - Only Delta tables are supported for sharing

Lakehouse Federation

Learning Objectives

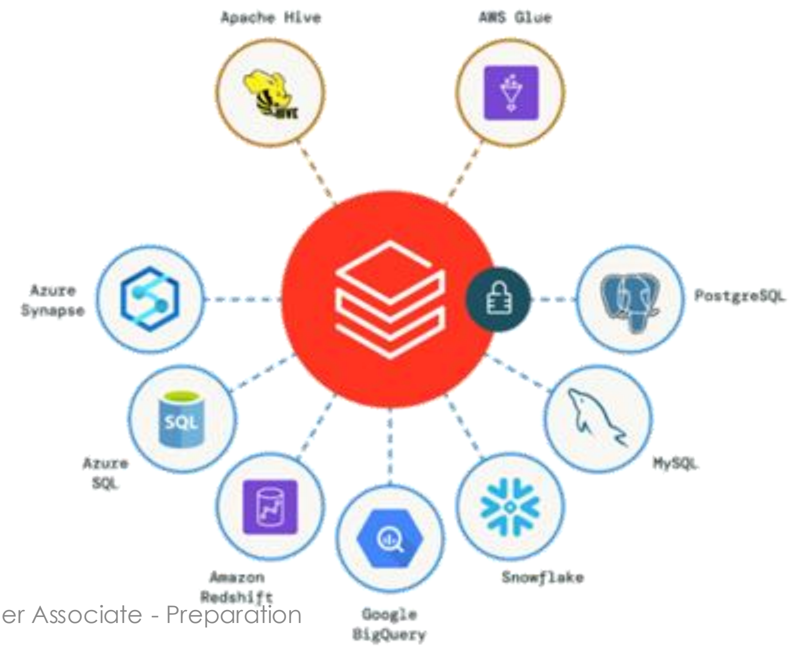
- Data ingestion challenges
- Lakehouse Federation

Data Ingestion Challenges

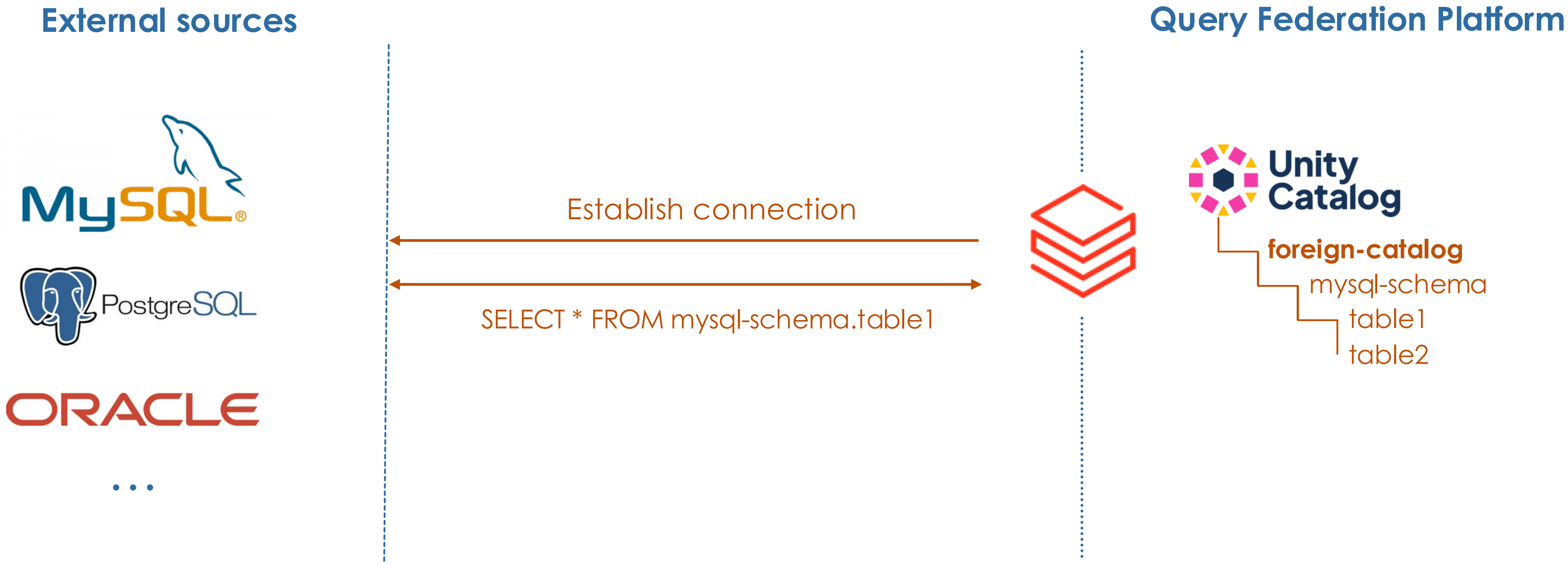
- Data ingestion is recommended for most use cases
  - high data volumes
  - low-latency querying
  - third-party API limits.
- Ingestion results in duplicate data that may become stale
  - Use Delta Sharing to receive live data from the source

Lakehouse Federation

- Run queries against multiple data sources
- No data migration



Query Federation



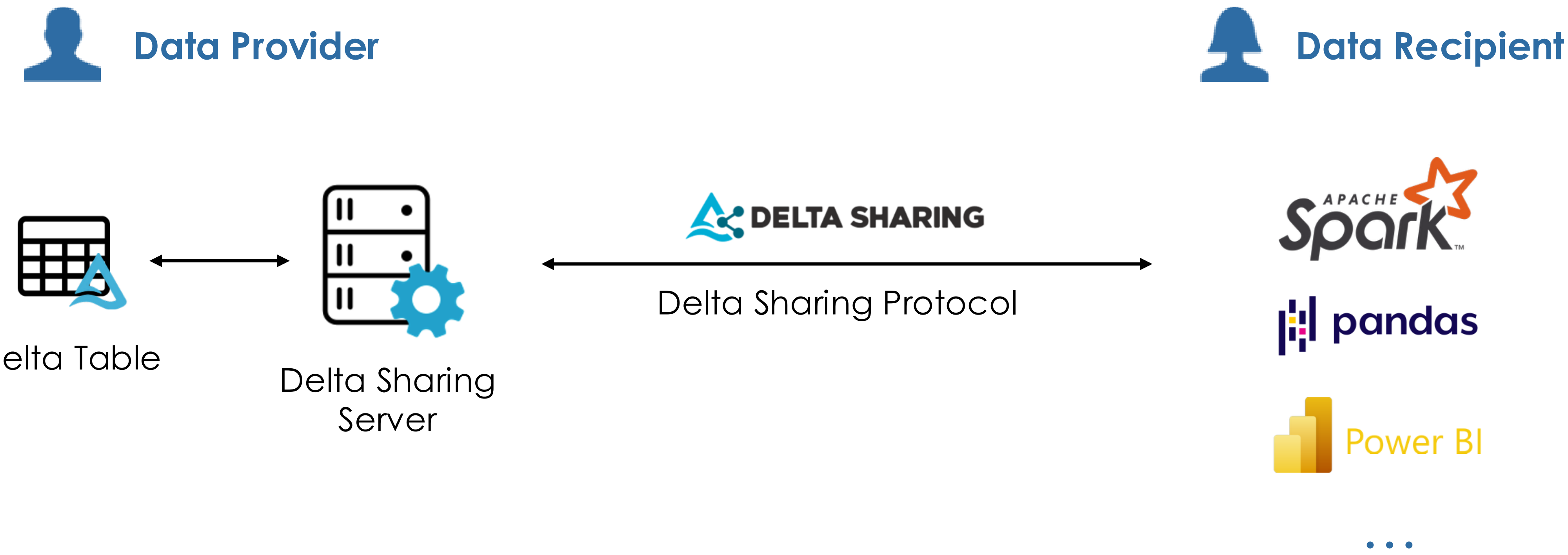
Use case

- Maintain live access to external systems.
- Ad hoc reporting or proof-of-concept access to operational data stored in external databases.

Limitations

- Complex queries don’t benefit from the power of Databricks

How it Works



Data Provider

Data Recipient

Delta Table

Delta Sharing Server

DELTA SHARING

Delta Sharing Protocol

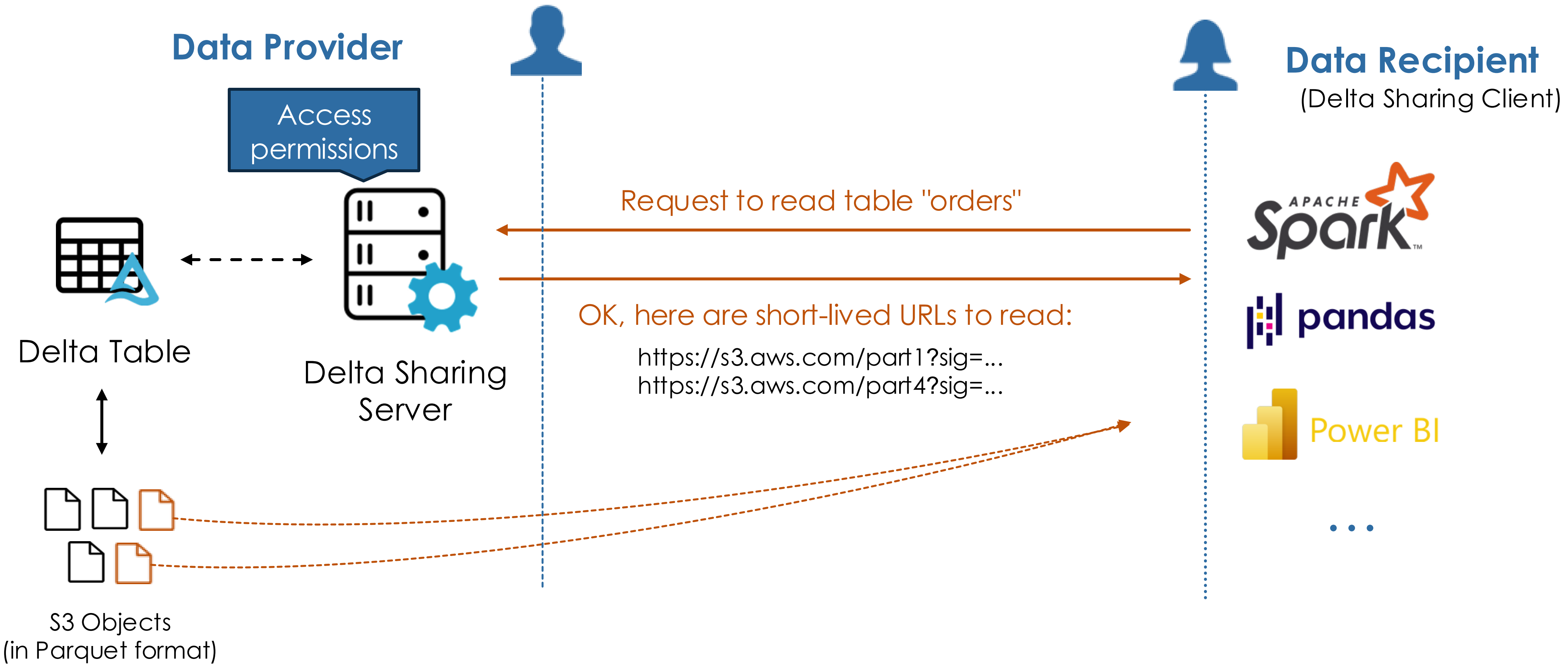
APACHE Spark

pandas

Power BI

. . .

How it Works



Data Provider

Access permissions

Delta Table

Delta Sharing Server

S3 Objects
(in Parquet format)

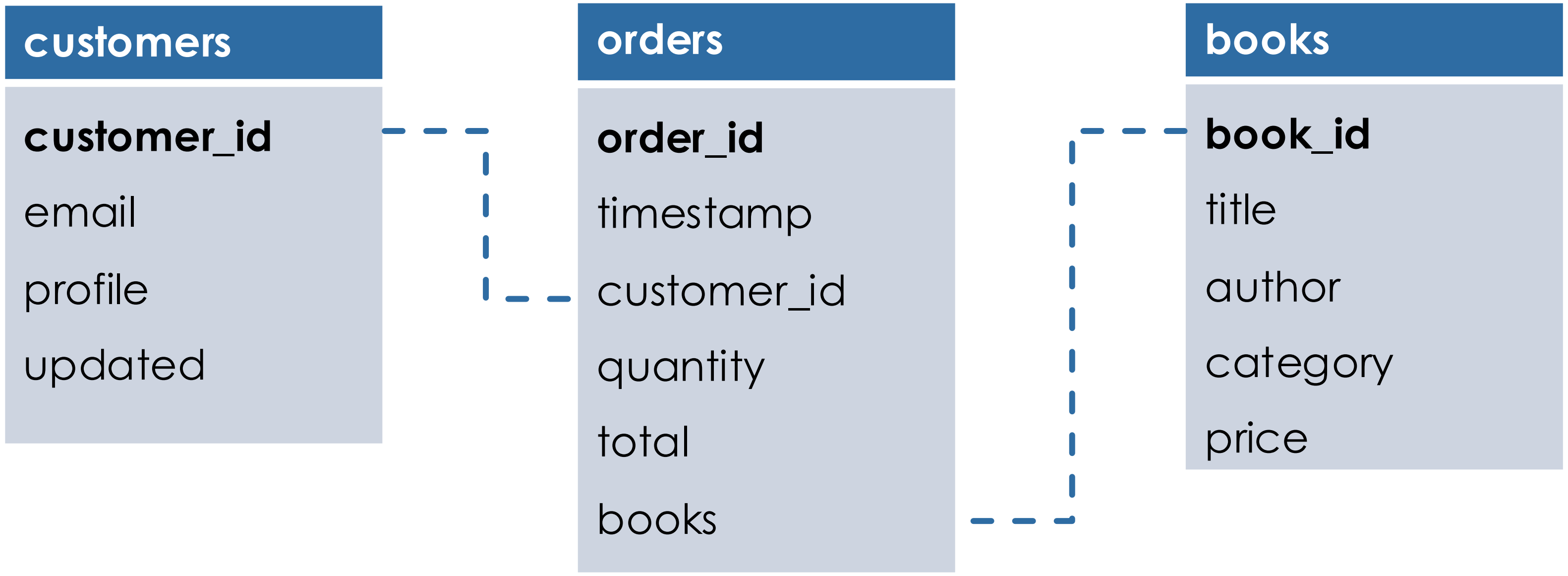
Data Recipient
(Delta Sharing Client)

Request to read table "orders"

OK, here are short-lived URLs to read:

https://s3.aws.com/part1?sig=...
https://s3.aws.com/part4?sig=...

Bookstore



customers

- **customer_id**
- email
- profile
- updated

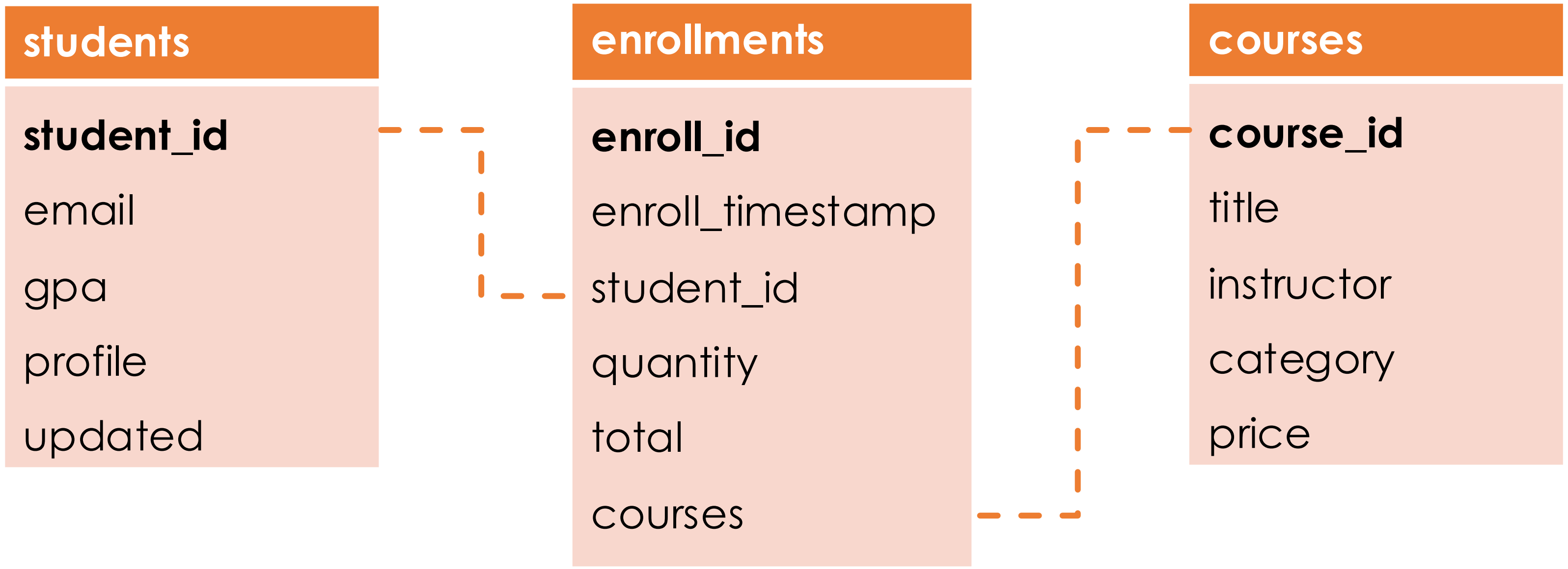
orders

- **order_id**
- timestamp
- customer_id
- quantity
- total
- books

books

- **book_id**
- title
- author
- category
- price

School



students

student_id
email
gpa
profile
updated

enrollments

enroll_id
enroll_timestamp
student_id
quantity
total
courses

courses

course_id
title
instructor
category
price

Certification Overview

Learning Objectives

- Format and structure of the exam.
- The topics covered in the exam.
- Types of questions provided on the exam

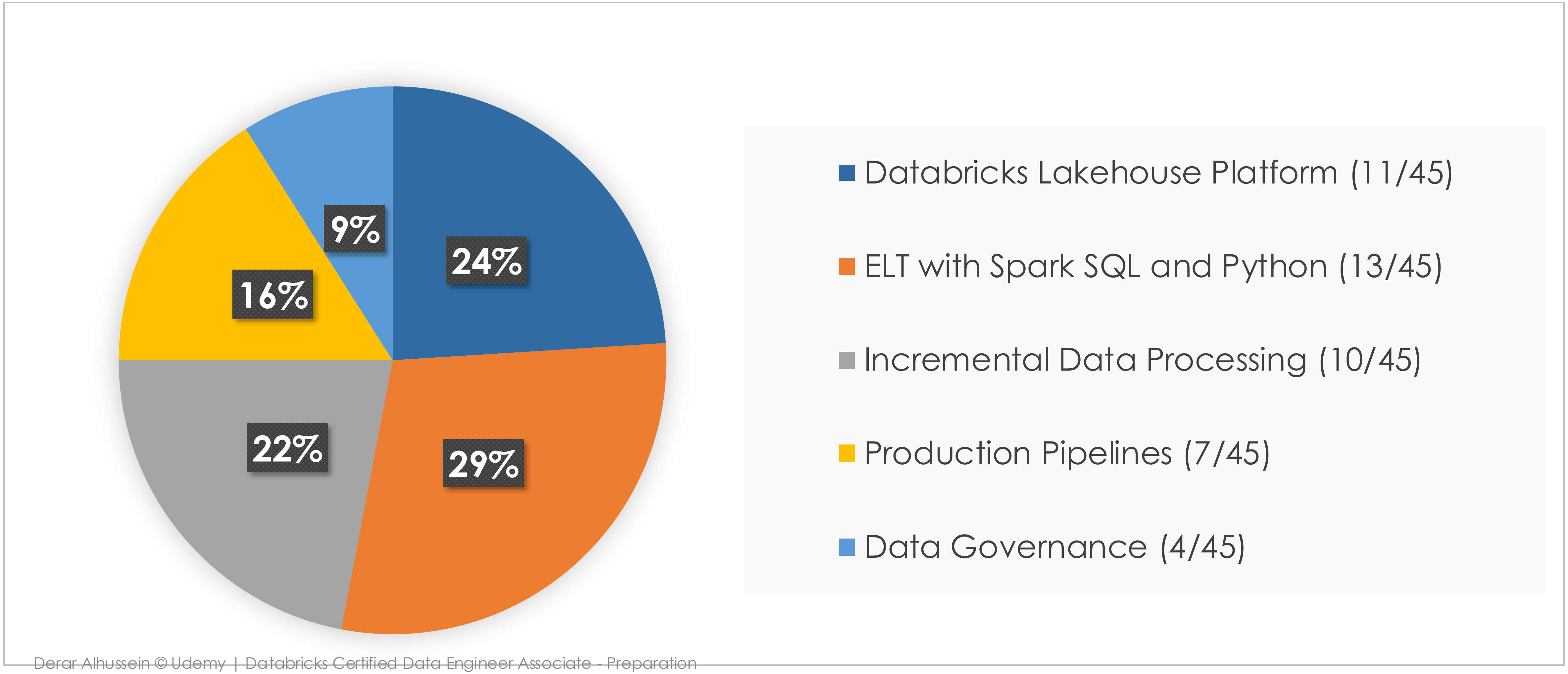
Exam Details

- Time allotted to complete exam = 1.5 hours

- Number of Questions = 45 (+5 Experimental)
- Passing scores = At least 70% (32/45)

- Exam fee = $200
- Exam retake fee = $200

Exam Questions



Out-of-scope

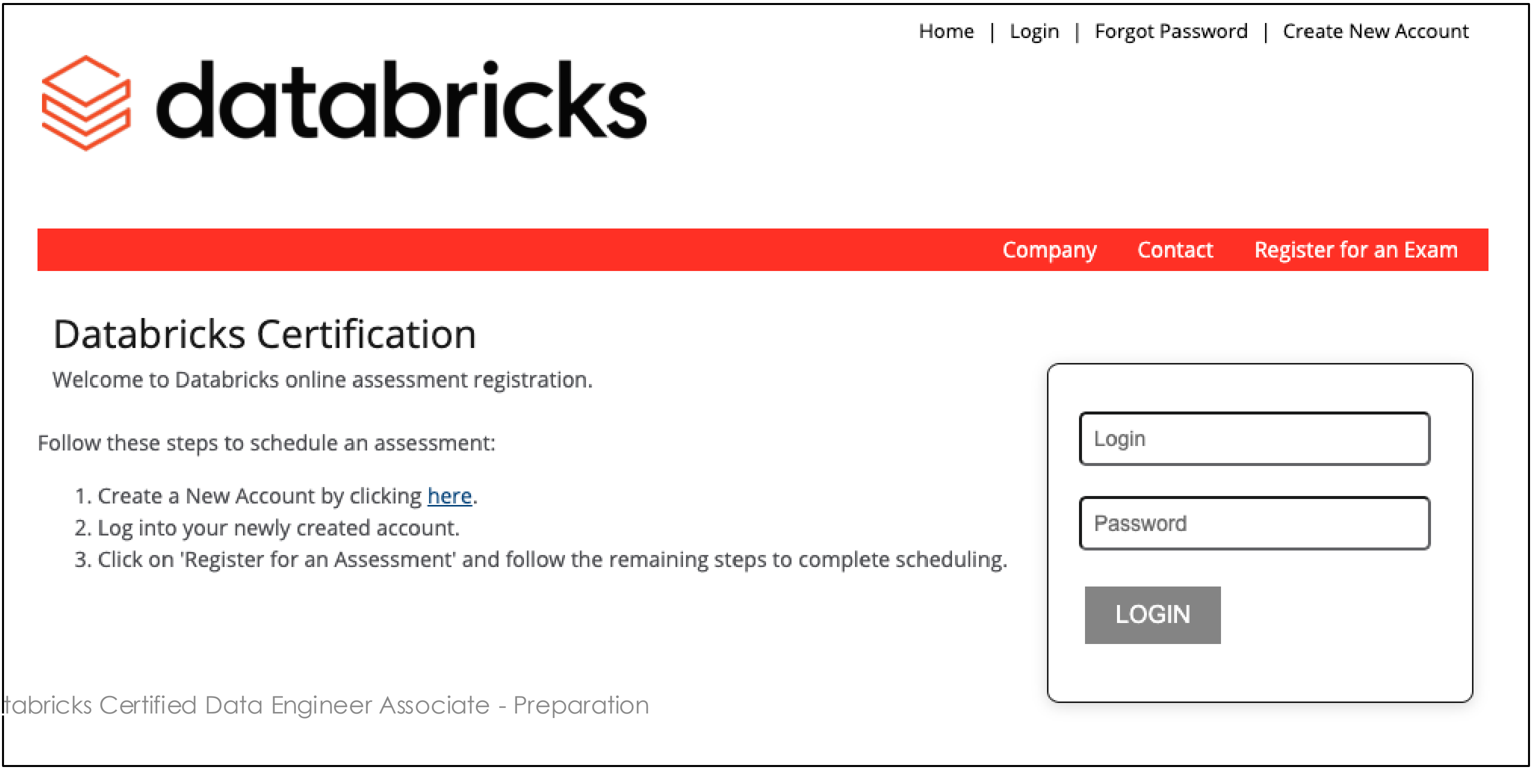
- Apache Spark internals
- Databricks CLI and REST API
- Change Data Capture CDC/CDF
- Data modeling concepts
- GDPR/CCPA
- Monitoring and logging production jobs
- Dependency management
- Testing

Code Examples

- Code will be in mainly in SQL
- Otherwise, Python

Exam Platform

https://www.webassessor.com/databricks



Databricks Certification

Welcome to Databricks online assessment registration.

Follow these steps to schedule an assessment:

1. Create a New Account by clicking here.
2. Log into your newly created account.
3. Click on 'Register for an Assessment' and follow the remaining steps to complete scheduling.

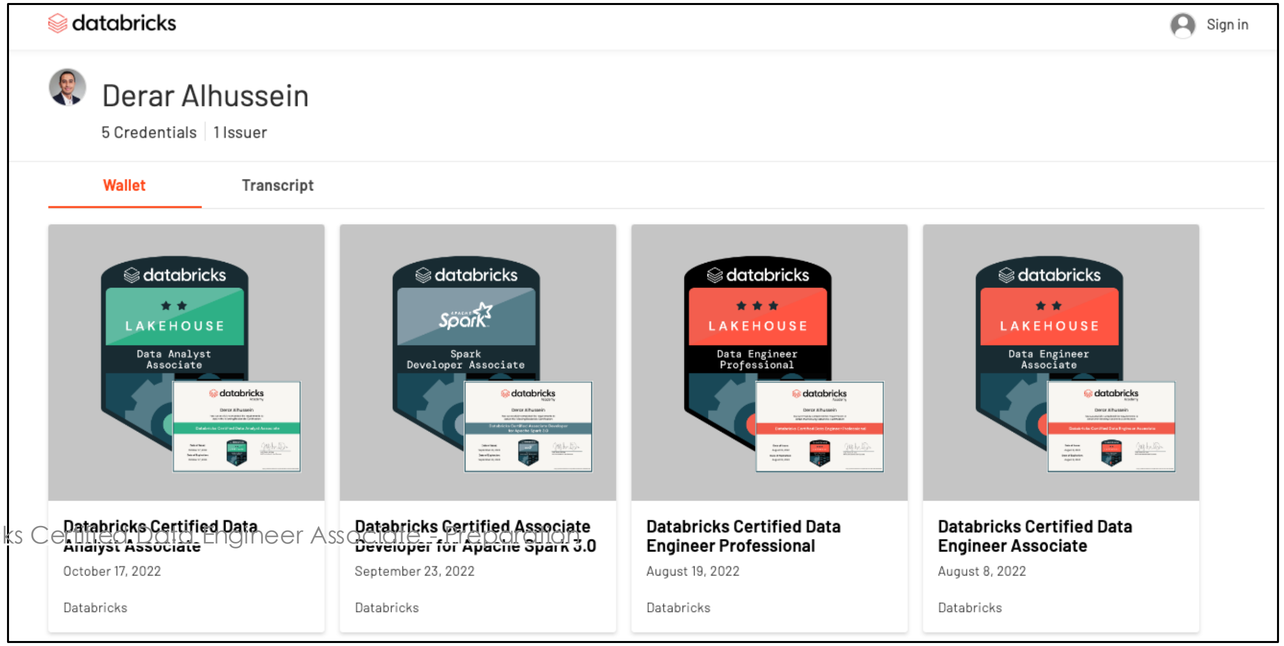
Exam Proctoring

- Monitored via webcam by a Webassessor proctor.
- Asked to provide valid, photo-based identification.
- Proctor does not provide assistance on the content of the exam
- No test aids will be available during the exam

Exam Result

- Certification exams are automatically graded.
- 24 hours to receive the badge and Certificate

https://credentials.databricks.com



Derar Alhussein

5 Credentials | 1 Issuer

Databricks Certified Data Analyst Associate
October 17, 2022
Databricks

Databricks Certified Associate Developer for Apache Spark 3.0
September 23, 2022
Databricks

Databricks Certified Data Engineer Professional
August 19, 2022
Databricks

Databricks Certified Data Engineer Associate
August 8, 2022
Databricks

Questions Types

- Multiple-choice questions – Only one correct answer

- Questions Types:
  - Conceptual Questions
  - Code-Based Questions

Conceptual Questions

- Which part of the Databricks Lakehouse Platform can data engineers use to orchestrate jobs ?

A. Repos

B. Workflows

C. Data Explorer

D. Databricks SQL

E. Cluster

Conceptual Questions

- Which part of the Databricks Lakehouse Platform can data engineers use to orchestrate jobs ?

A. Repos

B. Workflows

C. Data Explorer

D. Databricks SQL

E. Cluster

Code-Based Questions

```
spark.table("sales")
.writeStream
.option("checkpointLocation", checkpointPath)
._____

.table("new_sales")
```

If you want the query to execute a single micro-batch to process all of the available data, which of the following lines of code should you use to fill in the blank ?

A. trigger(once=True)

B. trigger(continuous="once")

C. processingTime("once")

D. trigger(processingTime="once")

E. processingTime(1)

Code-Based Questions

```
spark.table("sales")
    .writeStream
    .option("checkpointLocation", checkpointPath)
    ._____
    .table("new_sales")
```

If you want the query to execute a single micro-batch to process all of the available data, which of the following lines of code should you use to fill in the blank ?

A. trigger(once=True)

B. trigger(continuous="once")

C. processingTime("once")

D. trigger(processingTime="once")

E. processingTime(1)

Practice Exam

- Databricks official Practice test

- PDF file

Derar Alhussein

www.linkedin.com/in/DerarAlhussein