Lab 1: Upload Data to Databricks

Setting Up Your Data Foundation

Concepts Introduced

Unity Catalog - Databricks' unified governance solution for data and AI assets.

Key Concepts:

- **Catalog** - Top-level container (like a database server)
- **Schema** - Namespace within a catalog (like a database)
- **Volume** - Managed storage for files (CSV, HTML, images, etc.)

Why Unity Catalog?

- Centralized access control and auditing
- Files accessible from Spark, SQL, and AI agents
- Consistent path format: `/Volumes/<catalog>/<schema>/<volume>/`

What We're Doing

Upload source data to **Unity Catalog Volumes** as the foundation for the entire workshop.

Data Sources:

- **CSV Files** - Structured financial data (customers, accounts, transactions)
- **HTML Files** - Unstructured documents (customer profiles, research reports)

Destination:

```
/Volumes/<catalog>/<schema>/<volume>/
├── csv/    ← Structured data for graph import
└── html/   ← Documents for AI agent analysis
```

Two Upload Options

Option	Method	When to Use
Manual	Drag & drop in Databricks UI	Quick, visual, no setup
Programmatic	Python script with Databricks SDK	Automated, repeatable

Key Insight: Unity Catalog Volumes provide governed, secure storage that both Spark and AI agents can access.

What You'll Have

After this lab:

- **7 CSV files** ready for Neo4j import
  - Customers, Banks, Accounts, Companies, Stocks, Positions, Transactions
- **HTML documents** ready for Knowledge Agent indexing
  - Customer profiles, investment research, market analysis

Next: Import this data into Neo4j using the Spark Connector