Incremental Ingestion

ScholarNest

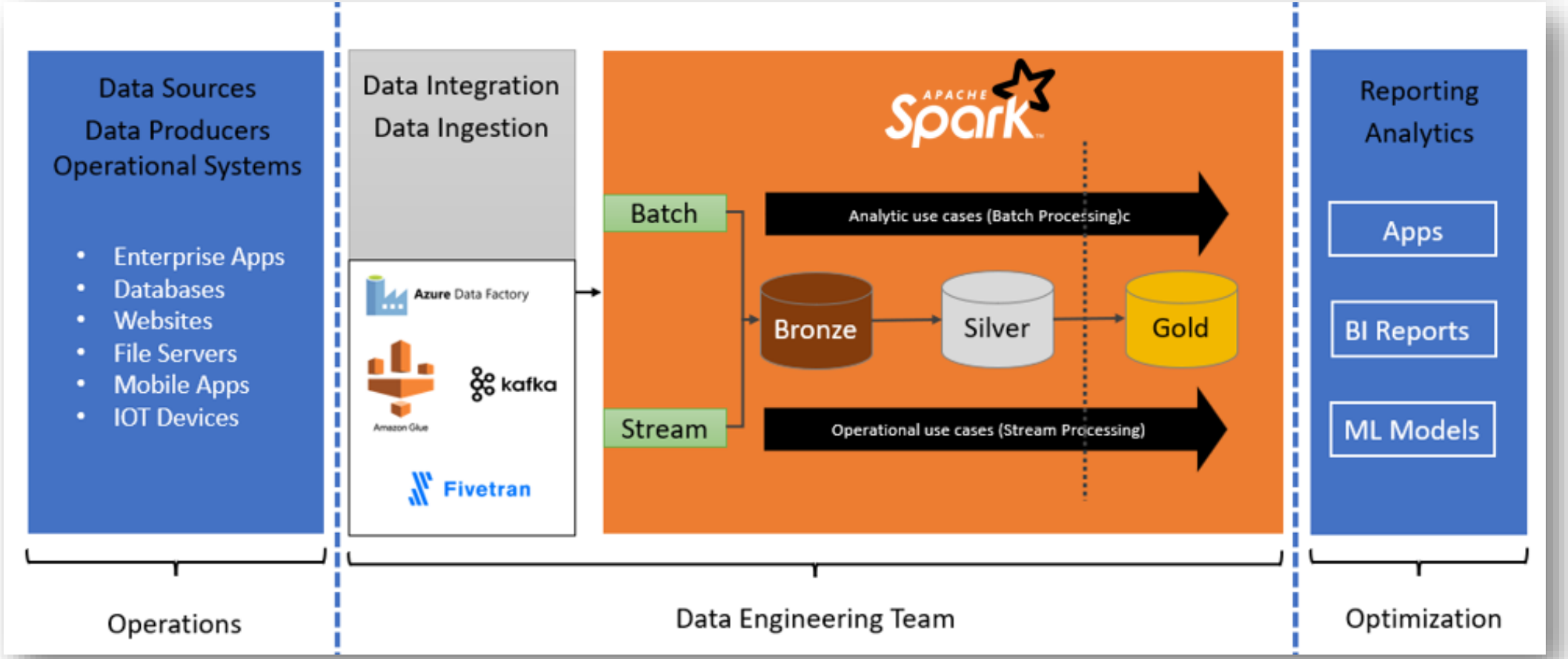
Incremental Data Ingestion

Databricks Cloud

- Architecture and Use case
- Copy Command
- Spark Streaming
- Autoloader
- Manual Schema Evolution
- Automatic Schema Evolution

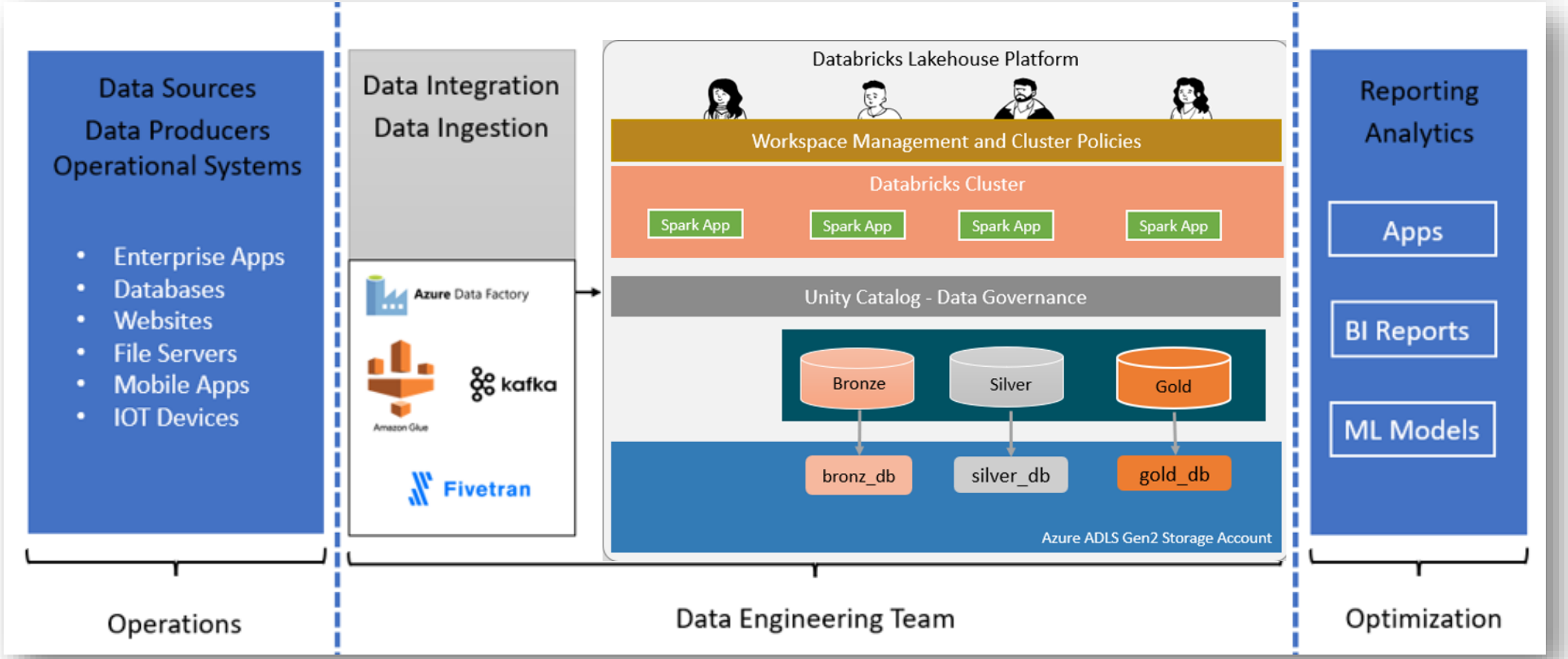
Lakehouse Medallion Architecture

ScholarNest



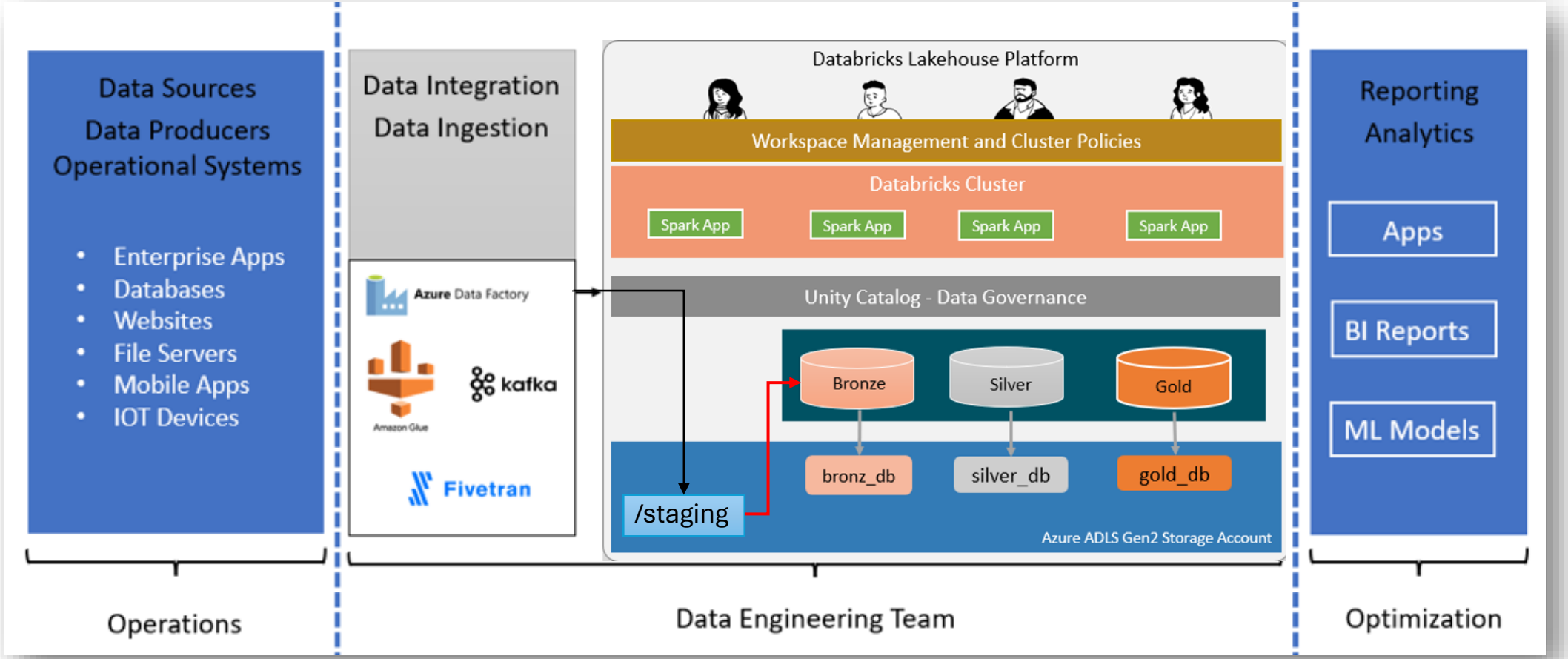
Lakehouse Medallion Architecture

ScholarNest



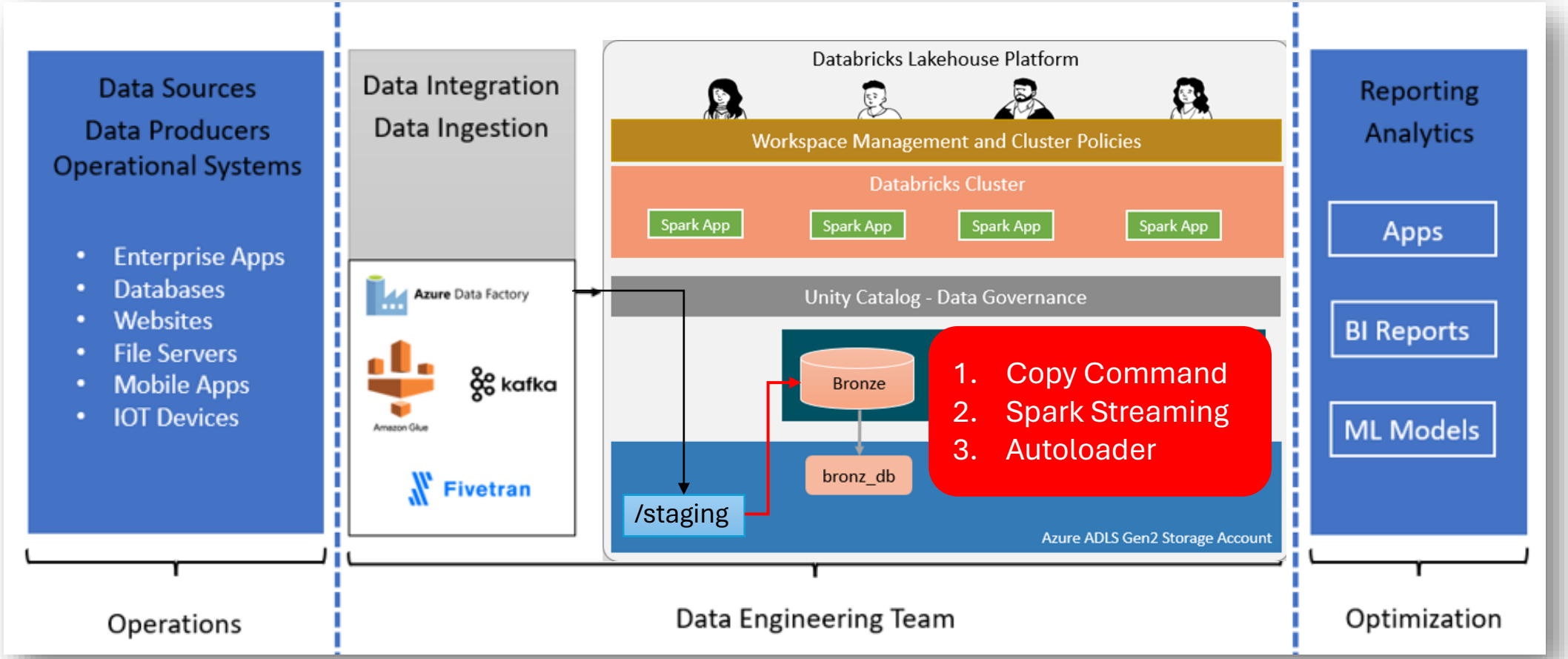
Lakehouse Medallion Architecture

ScholarNest



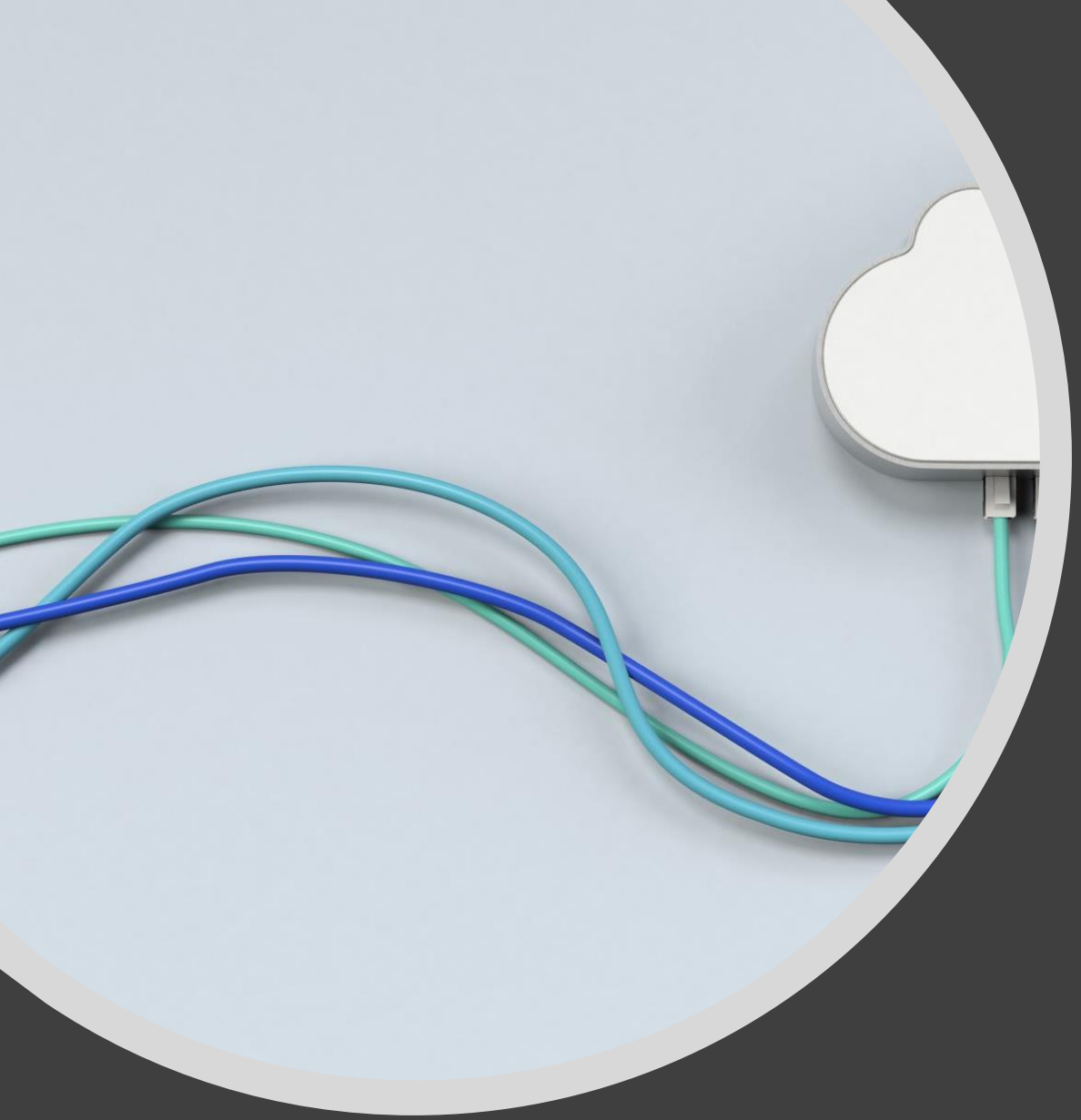
Lakehouse Medallion Architecture

ScholarNest



ScholarNest

Databricks Autoloader



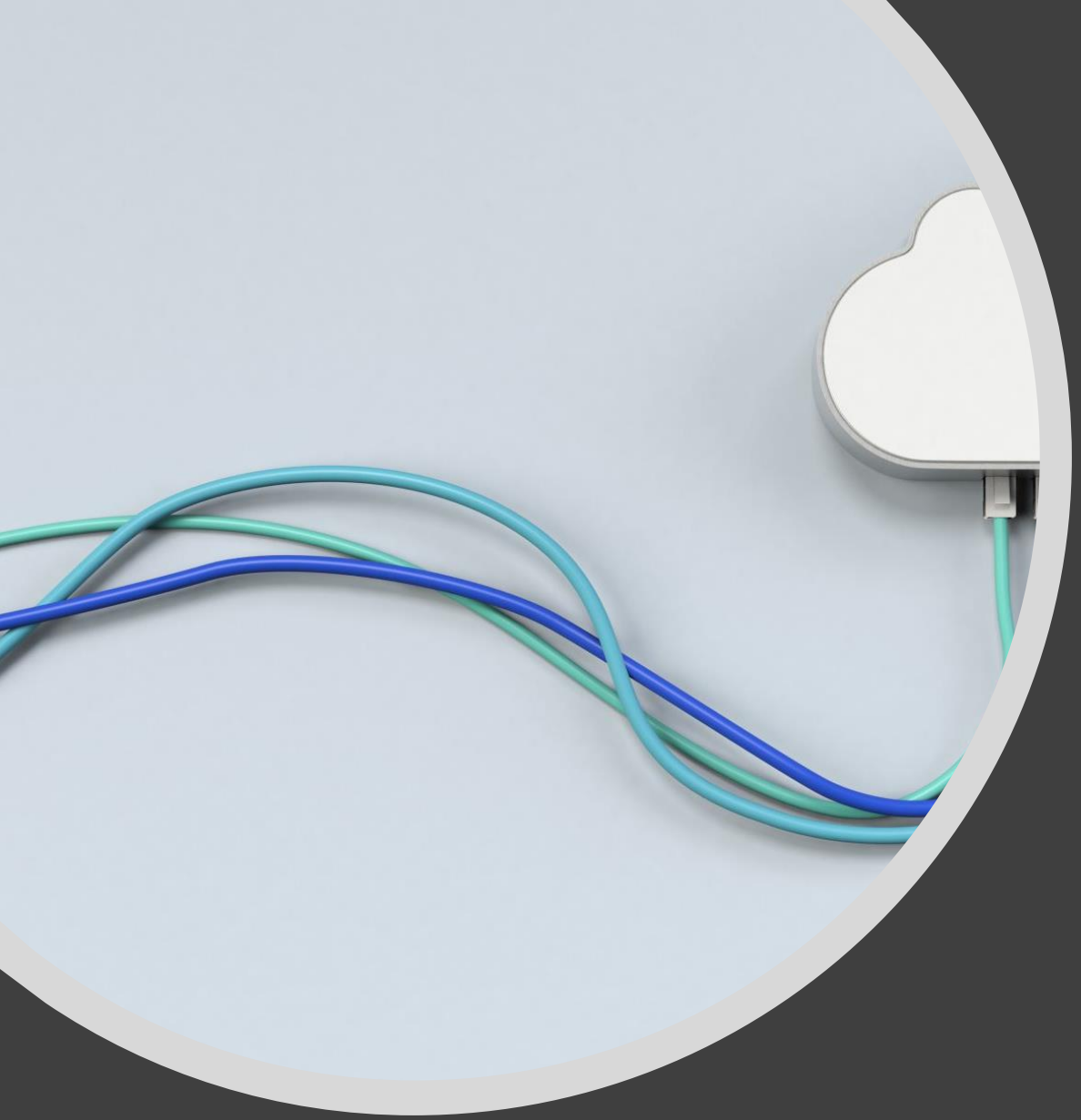
ScholarNest

What is Databricks Autoloader?

A framework to efficiently process new data files from cloud storage.

1. Amazon S3
2. Azure Data Lake Storage Gen2
3. Google Cloud Storage
4. Databricks File System

JSON, CSV, PARQUET, AVRO, ORC, TEXT, BINARY FILE



ScholarNest

Why Autoloader?

1. Optimized file listing
    1. Cloud-native APIs
    2. Fewer API calls
    3. Incremental listing
    4. Optional file notification service
2. Simplified schema evolution
3. Simplified data rescue