BUILD PIPELINE ON CLOUD

MENTOR: NGUYỄN HOÀNG VŨ THÂN
TĂNG THỊ THÚY VÂN

CONTEXT

01.

BASIC KNOWLEDGE

02.

CLOUD COMPUTING

03.

MICROSOFT AZURE

04.

DEMO

01. BASIC KNOWLEDGE

PYTHON

- Concept
- Compiler & Interpreter

SQL

- Concept
- OLTP & OLAP

PYTHON

DEFINITION

- High-level
- Open source
- Cross-platform

LIBRARIES

- Faker
- Pandas
- Datetime

card.py > ...

```python
# Khởi tạo đối tượng Faker
fake = Faker()

# Kết nối đến cơ sở dữ liệu SQL Server
sql_server_host = r'INTERN-NTNGHI-L\SQLSEVER'
sql_server_database = 'projectsale'
sql_server_user = 'sa'
sql_server_password = '1234567'

# Tạo SQLAlchemy engine
engine = create_engine(f'mssql+pyodbc://{sql_server_user}:{sql_server_password}@{sql_server_host}/{sql_server_database}?driver=ODBC

# Tạo DataFrame chứa dữ liệu
data = {'card_number': [], 'cc_name': [], 'pin': [], 'b_street': [], 'b_city': [], 'b_state': [], 'b_zip': []}

# Tạo một tập hợp để lưu trữ các card_number đã sử dụng
existing_card_numbers = set()

# Tạo 300 dòng dữ liệu
for _ in range(300):
    card_number = ''.join(random.choices('0123456789', k=16))

    # Kiểm tra xem card_number đã tồn tại chưa
    while card_number in existing_card_numbers:
        card_number = ''.join(random.choices('0123456789', k=16))

    # Thêm card_number vào tập hợp đã sử dụng
    existing_card_numbers.add(card_number)

    cc_name = fake.name()
    pin = ''.join(random.choices('0123456789', k=4))
    b_street = fake.street_address()
    b_city = fake.city()[:20]
    b_state = fake.state_abbr()
    b_zip = fake.zipcode()
```

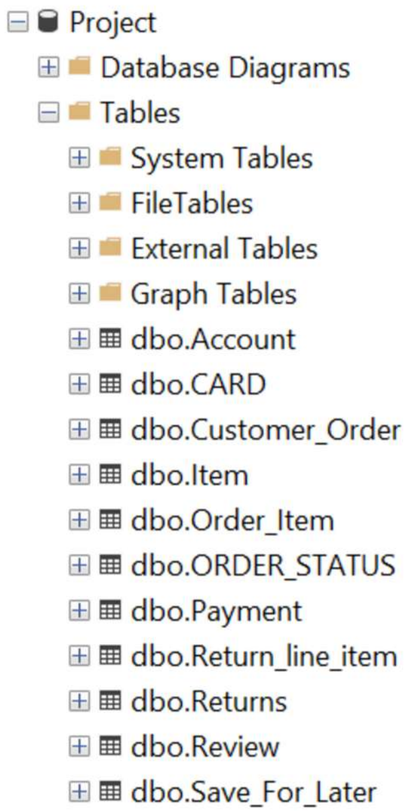
COMPILER & INTERPRETER

Criteria	Compiler	Interpreter
Input	Entire program	Single line of code
Output	Intermediate object code	No intermediate object code is generated
Mechanism	Compilation happens before execution	Compilation and execution happen simultaneously
Speed	Faster	Slower
Memory	Requires more memory due to intermediate object code generation	Requires less memory since it does not generate intermediate object code
Errors	Displays all errors after compilation, all at once	Displays errors one line at a time
Error Detection	Very difficult	Relatively easy
Programming Languages	C, C++, C#, Scala, typescript	PHP, Perl, Python, Ruby

SQL

DEFINITION

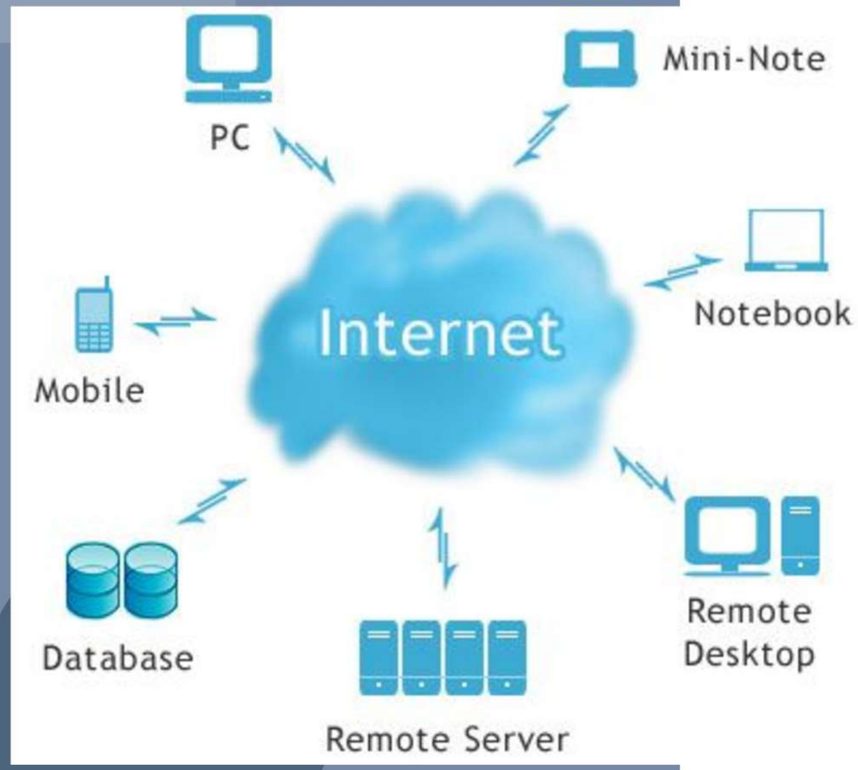
- Programming language
- Storing and processing
- Relational databases



OLTP & OLAP

Feature	OLAP	OLTP
Purpose	OLAP helps you analyze large volumes of data to support decision-making	OLTP helps you manage and process real-time transactions
Data source	OLAP uses historical and aggregated data from multiple sources	Uses real-time and transactional data from a single source
Data structure	OLAP uses multidimensional (cube) or relational databases	Uses relational databases
Data model	Uses star schema, snowflake schema, or other analytical models	Uses normalized or denormalized models
Data volume	Requires large storage capacity. Think terabytes (TB) and petabytes (PB).	Has relatively small storage requirements. Think gigabytes (GB).
Response time	Has longer response times, typically in seconds or minutes	Has shorter response times, typically in milliseconds
Example applications	Analyze trends, predict customer behavior, and identify profitability	Process payments, manage customer data, and fulfill orders

02. CLOUD COMPUTING



- Providing services
- Through the Internet

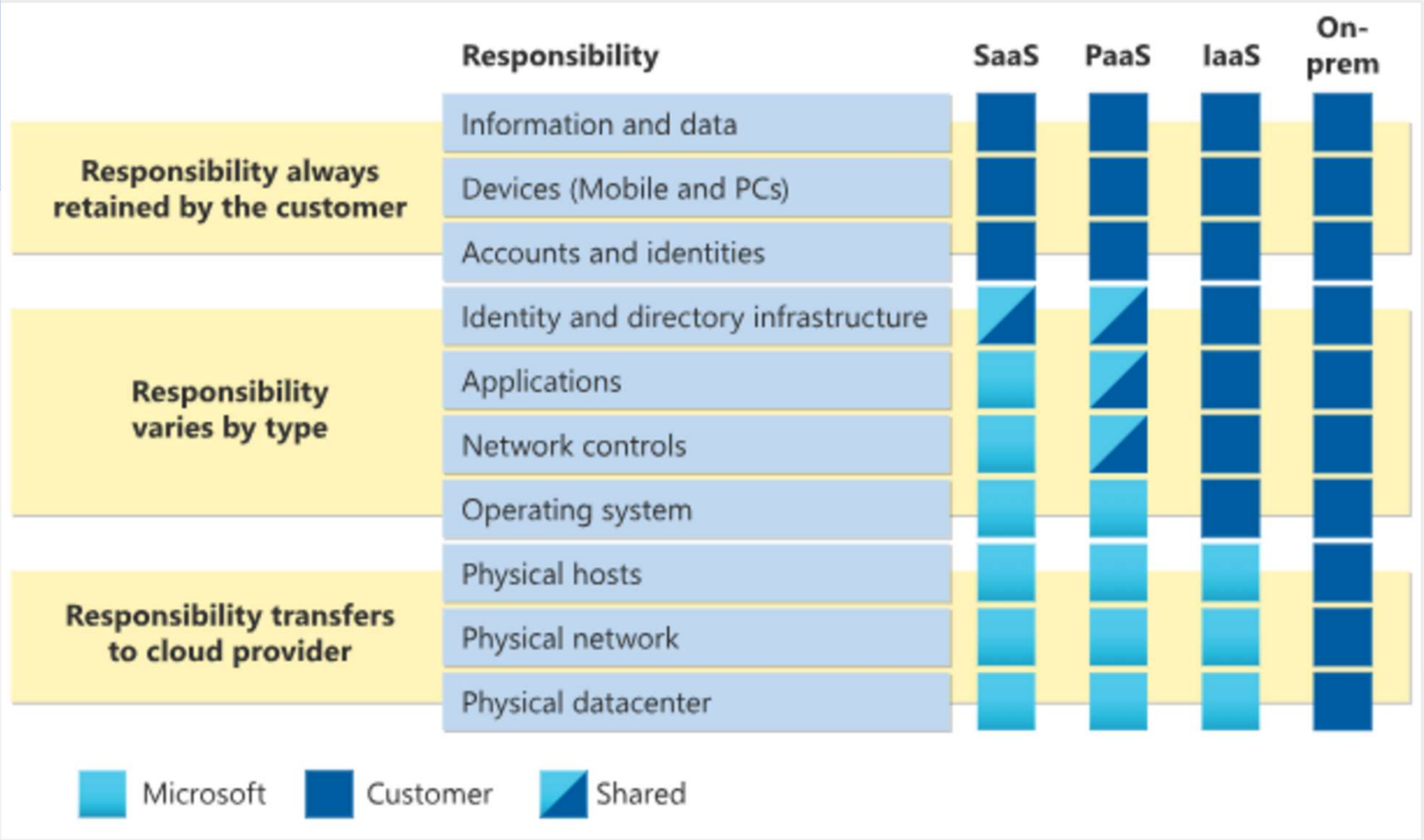
CLOUD SERVICE TYPES

- Iaas
- Paas
- Saas

CLOUD MODELS

- Private Cloud
- Public Cloud
- Hybrid Cloud

RESPONSIBILITIES OF EACH CLOUD SERVICE TYPE



CLOUD MODELS

Public Cloud	Private Cloud	Hybrid Cloud
No capital expenditures to scale up	Complete control	Provides the most flexibility
Quickly provisioned and deprovisioned	Data is not collocated with other organizations' data	Determine where to run their applications
Pay only for what they use	Hardware must be purchased for startup and maintenance	Control security, compliance, or legal requirements
Insufficient control over resources and security	Responsible for hardware maintenance and updates	No mention

03. MICROSOFT AZURE

- Azure Blob Storage
- Azure Data Factory
- Azure Databricks

03. MICROSOFT AZURE

AZURE BLOB STORAGE

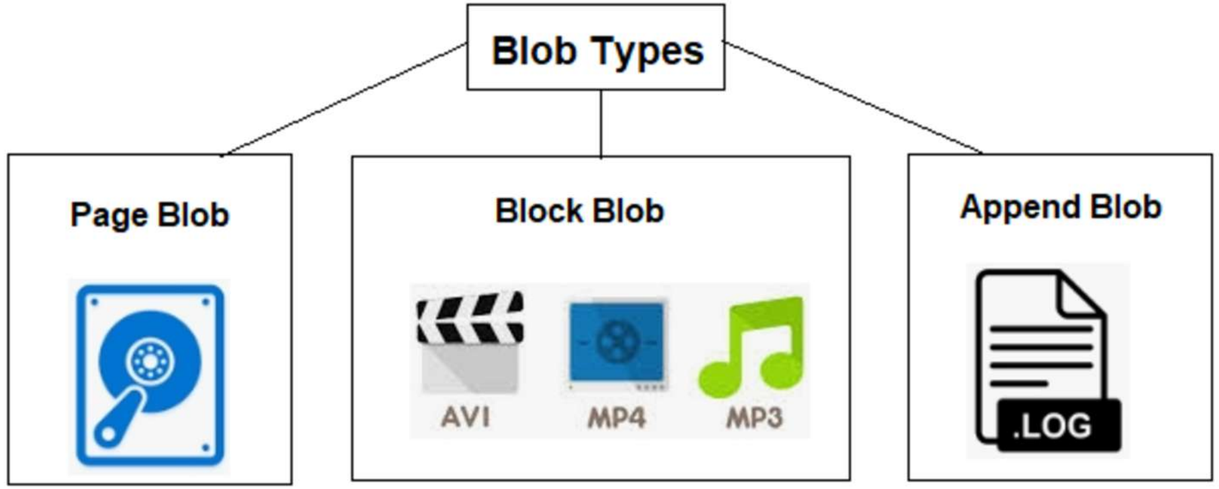
- Microsoft's object storage solution
- Store large amounts of unstructured data

Main components of Azure Blob Storage

Storage Account

Container

Blob



03. MICROSOFT AZURE

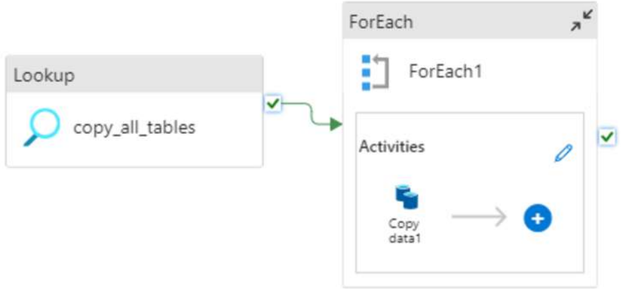
AZURE DATA FACTORY

Key features

- Data Integration Service
- Data Transformation
- Data Movement
- Automation and Scheduling

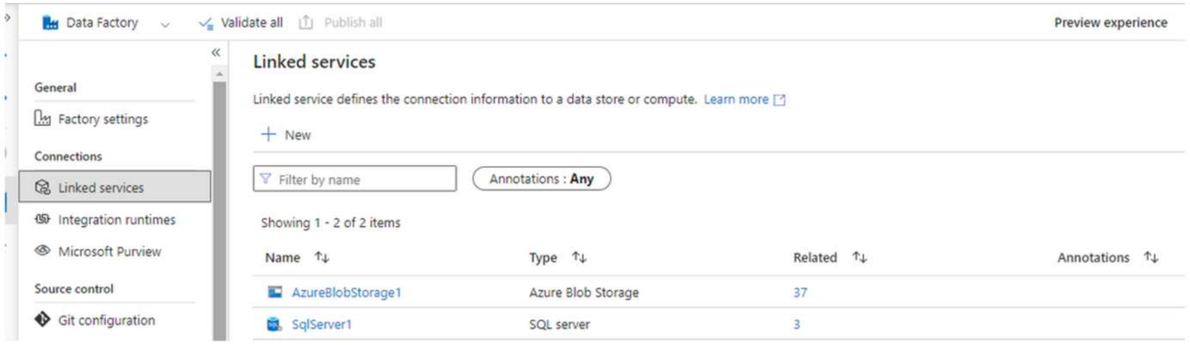
Key components

- Pipeline



- Activity
- Dataset
- Trigger

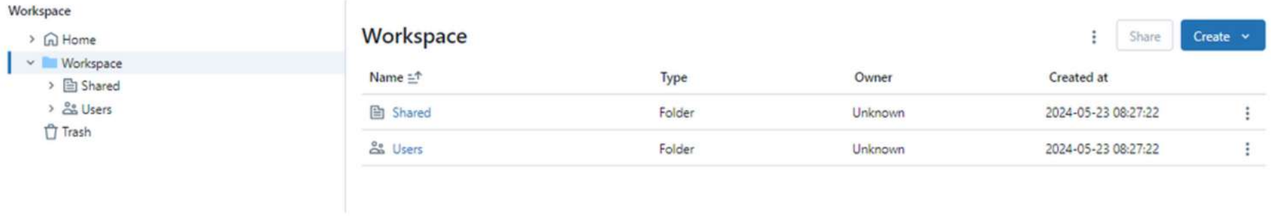
- Linked Service



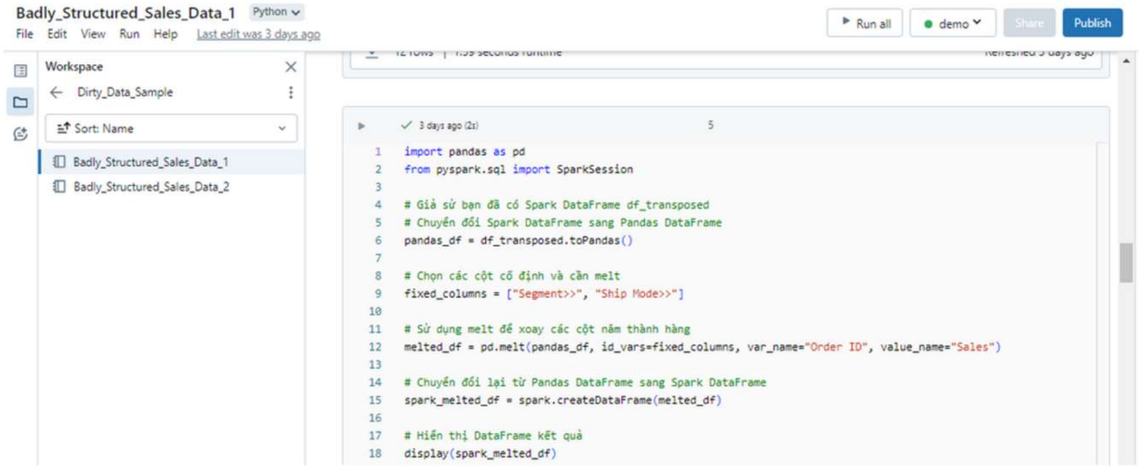
03. MICROSOFT AZURE

DATABRICKS

- Workspace



- Notebook



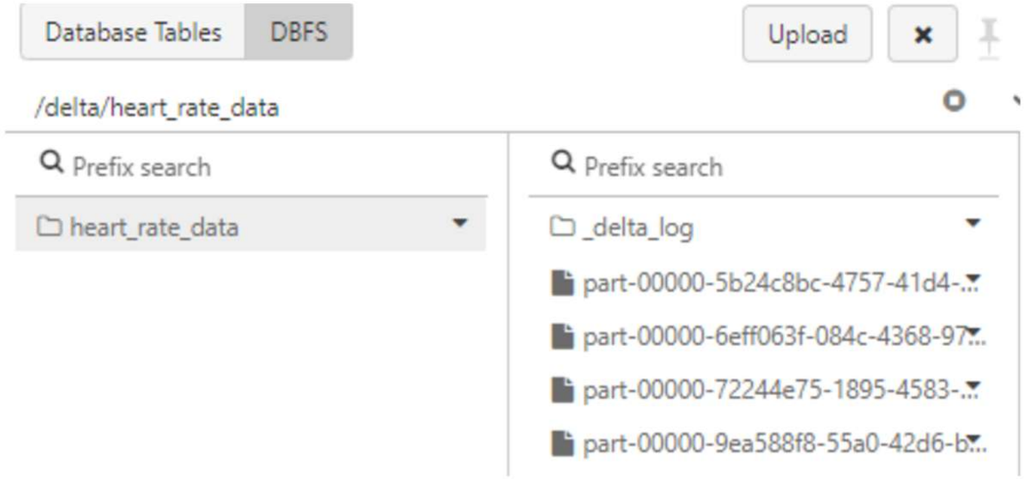
03. MICROSOFT AZURE

DATABRICKS

- Clusters

State	Name	Runtime	Active memory	Active cores	Active DBU / h	Source	Creator	Notebooks
✓	demo	12.2	15 GB	2 cores	1	UI	dlkhanh-batch16bd@sd...	-

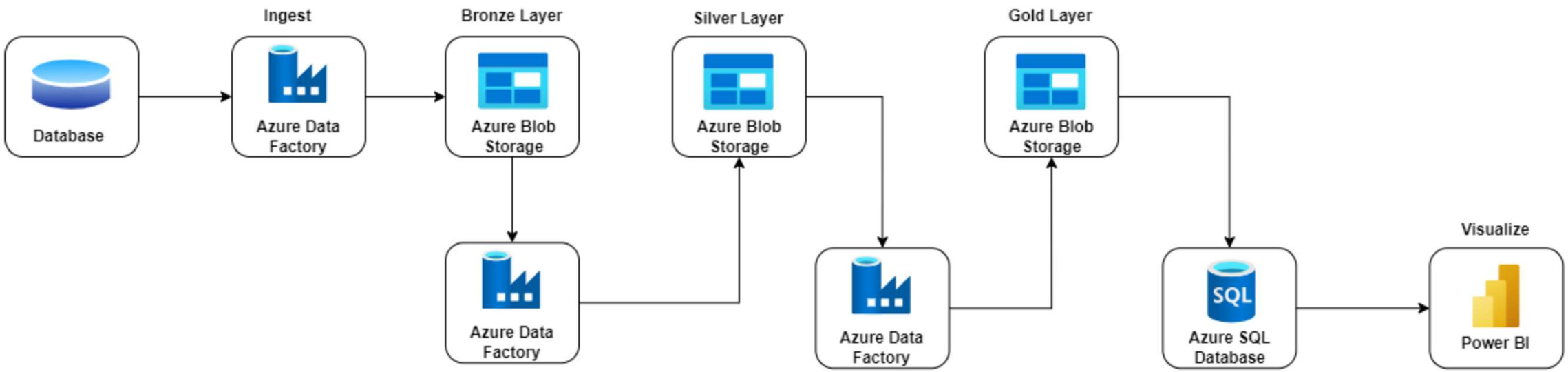
- Delta Lake



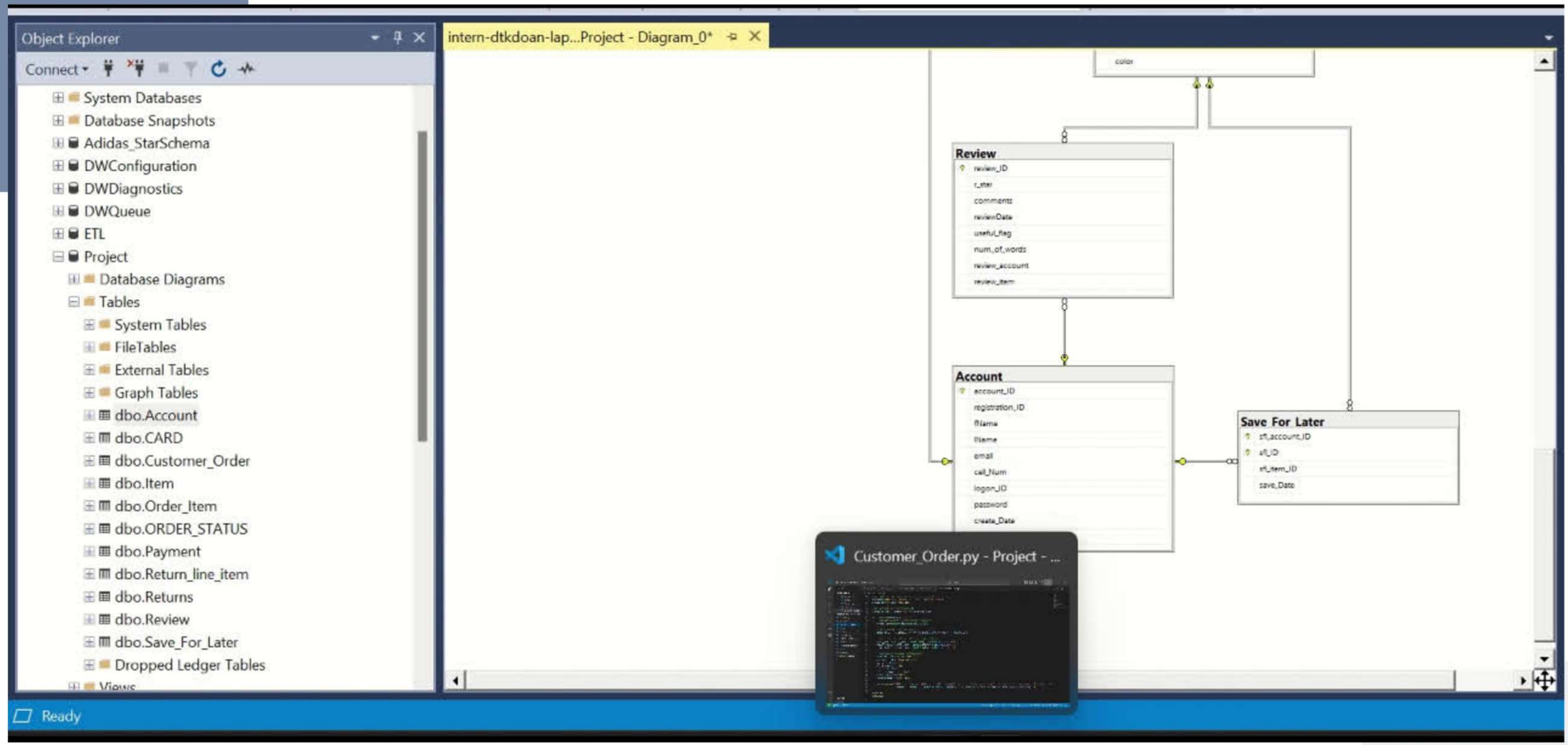
04. DEMO

Azure
Data Factory

DATA FLOW


Ingest
Bronze Layer
Silver Layer
Gold Layer
Visualize
Database
Azure Data Factory
Azure Blob Storage
Azure Blob Storage
Azure Blob Storage
Azure Data Factory
Azure Data Factory
SQL
Azure SQL Database
Power BI

04. DEMO


Object Explorer
Connect
System Databases
Database Snapshots
Adidas_StarSchema
DWConfiguration
DWDiagnostics
DWQueue
ETL
Project
Database Diagrams
Tables
System Tables
FileTables
External Tables
Graph Tables
dbo.Account
dbo.CARD
dbo.Customer_Order
dbo.Item
dbo.Order_Item
dbo.ORDER_STATUS
dbo.Payment
dbo.Return_line_item
dbo.Returns
dbo.Review
dbo.Save_For_Later
Dropped Ledger Tables
Views
intern-dtkdoan-lap...Project - Diagram_0*
color
Review
review_ID
r_star
comments
reviewDate
useful_flag
num_of_words
review_account
review_item
Account
account_ID
registration_ID
fName
lName
email
cell_Num
logon_ID
password
create_Date
Save_For_Later
sfl_account_ID
sfl_ID
sfl_item_ID
save_Date
Customer_Order.py - Project - ...
Ready

04. DEMO

databricks

Thank You