DS684
Cloud Computing
Week 08

Regarding Labs and Assignments

- Class participation means more than Zoom attendance. You must actively participate in the discussion and labs, and answer questions.

- Must hit Submit button, otherwise no grade
- If you need extension in time, must send written request (***email***). Otherwise no grade and no makeup. Requests sent over Zoom chat do not count.
- For any technical difficulty (installation, Azure access, etc), you must send written explanation (***email***) before the deadline. Otherwise no grade and no makeup.

Teaching Schedule

Week 7: Azure Synapse Analytics Part I: Data Warehouse

Week 8: Azure Synapse Analytics Part II: Data Engineering

Week 9: Visualization using Power BI

Week 10: Azure Machine Learning

Week 11: Final project presentation

Agenda

- ETL
  - Traditional Data Processing Flow
- ELT Data Processing Flow
  - Medallion Architecture
- Data Processing Services
  - Lab: Fabric Data Pipeline
- Introduction to Databricks
- Introduction to Azure Synapse Analytics

General Tasks of Data Engineering

- Extracting (reading) from a source
- Transforming
  - Filtering
  - Calculation
  - Joining
  - Aggregation
  - etc.
- Loading (saving) into a target

Extract, Transform, Load (ETL)

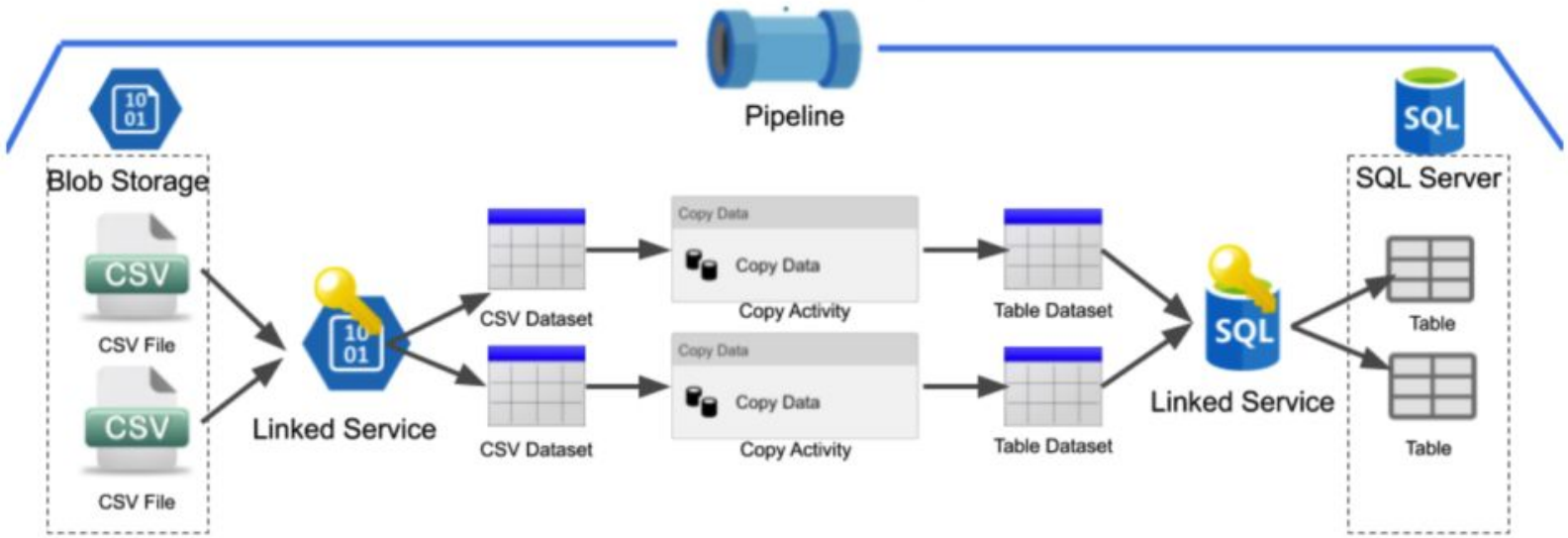
Different source systems will generate data that

- Comes in different format and frequency
  - Medicare history vs Medicaid history
- Is not joinable from different sources
  - Medical history vs medical device purchase history - different keys
- Is not clean
  - Rx usage history vs patient social network contents - lots of unrelated information

Extract, Transform, Load (ETL)

- Data from multiple sources (website, mobile, etc) are cleansed, consolidated, merged, and stored together
- Derived/aggregated values are calculated
- Loaded into more accessible schemas.

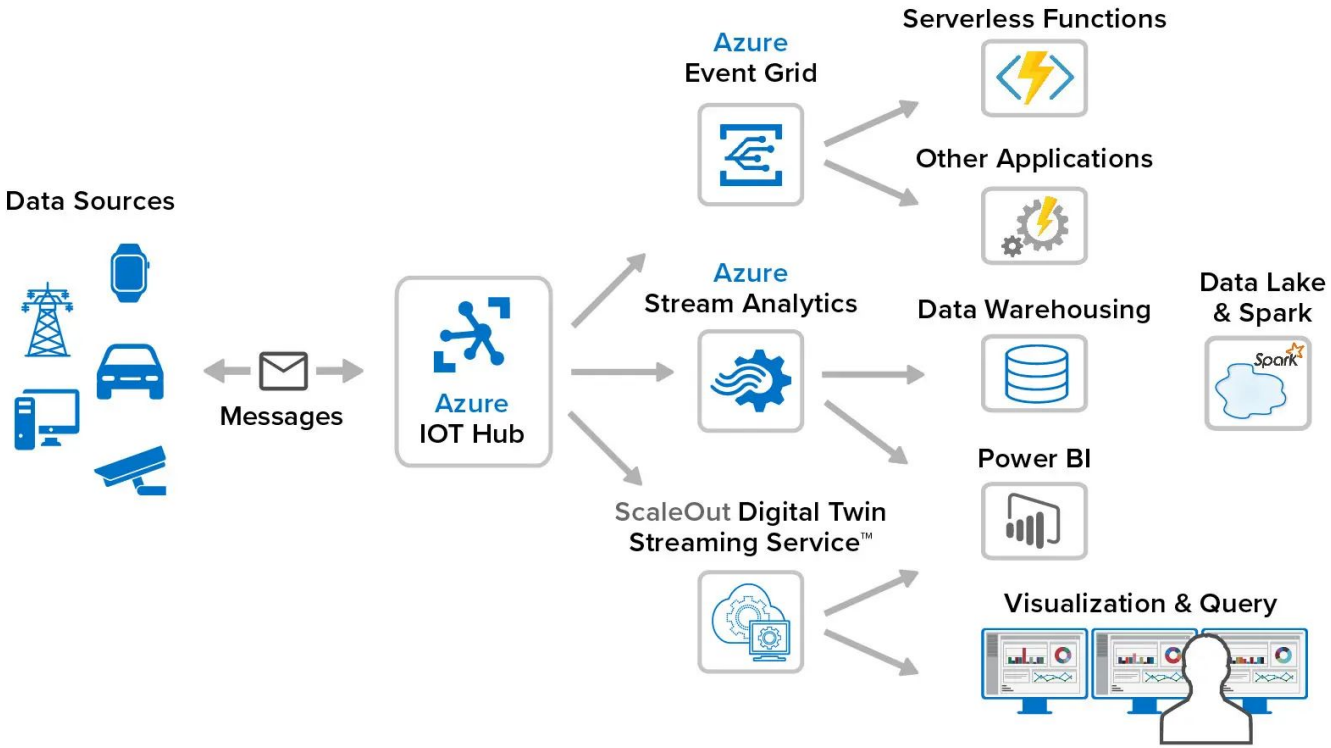
Extract, Transform, Load (ETL)



Batch vs Streaming

- Real world activities happen like a stream of events over time
  - Batch: Collect the events and process together
  - Streaming: Processing each event as it arrives
- Streaming is not a new concept/practice, but gets more attention as big data gains popularity
  - Velocity of data
  - Variety (source and format) of data
- Example: Internet of Things (IoT)

Azure IoT Streaming Example



Agenda

- ETL
  - Traditional Data Processing Flow
- ELT Data Processing Flow
  - Medallion Architecture
- Data Processing Services
  - Lab: Fabric Data Pipeline
- Introduction to Databricks
- Introduction to Azure Synapse Analytics

Traditional Data Processing Flow

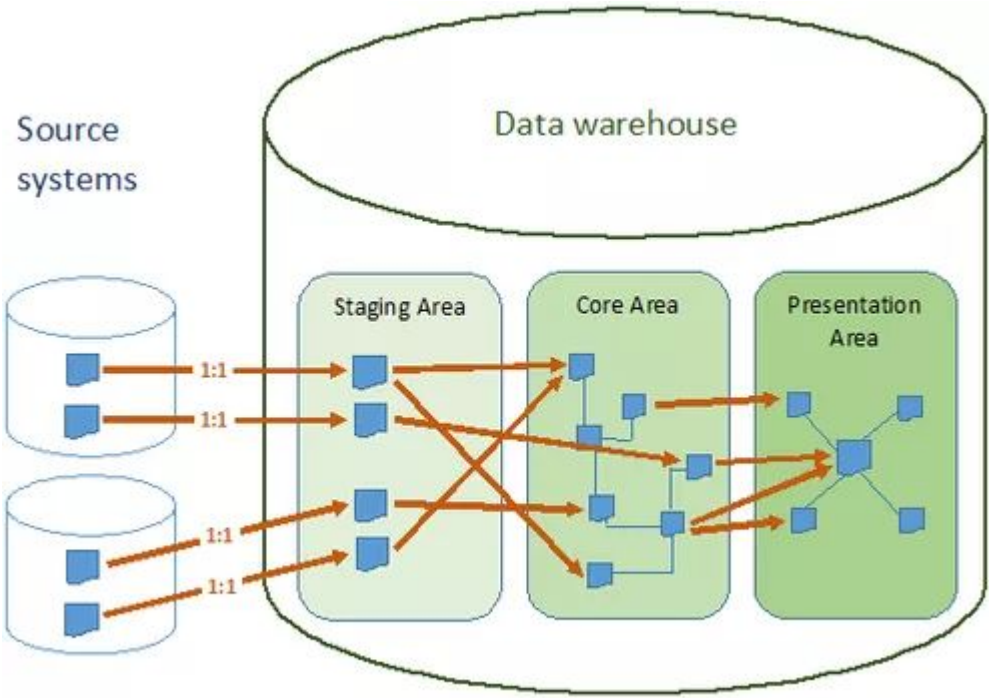
Source ->

Staging ->

Data Warehouse ->

Datamart

- Transformation happens between Staging area and Data Warehouse
- There might be multiple layers of staging



Traditional Data Architecture

Staging:

1. Load data as is
2. Data manipulations
3. Intermediate results

There might be multiple layers of staging

ETL workflow will move data between different layers of staging tables, to data warehouse, and finally to datamarts

Traditional Data Architecture

Datamart:

- Star schema
- Business oriented
- Organized around a particular business flow or activity

Look backwards, datamarts and data warehouses are usually designed first

Traditional Data Architecture

Central storage

- Star/Snowflake schema Data Warehouse, or
- 3NF complaint ODS: Since datamart has been designed as star schema, data warehouse can be 3NF, or
- Logical data warehouse: a layer of views on top of ODS

Agenda

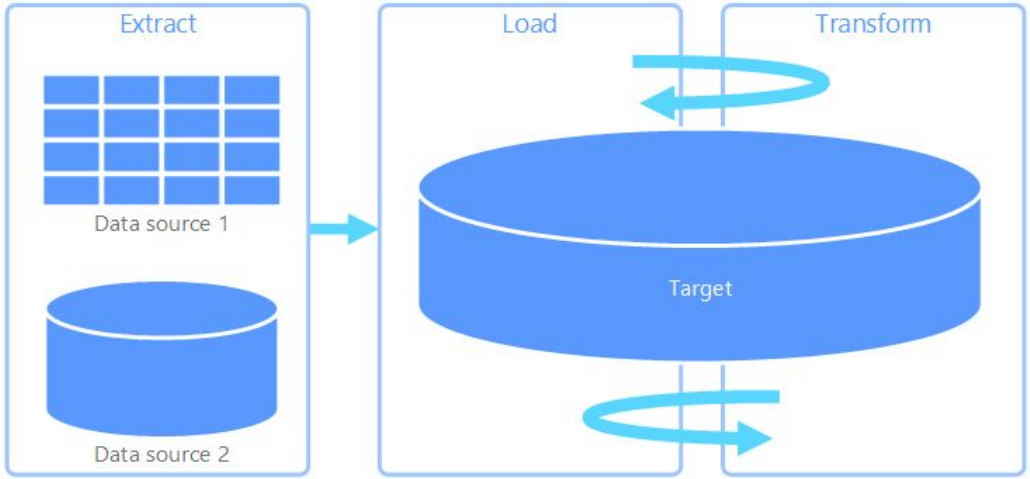
- ETL
  - Traditional Data Processing Flow
- ELT Data Processing Flow
  - Medallion Architecture
- Data Processing Services
  - Lab: Fabric Data Pipeline
- Introduction to Databricks
- Introduction to Azure Synapse Analytics

ELT Data Processing Flow

Load into data lake first. Process when read (reducing processing needs)

Raw data lake based Lakehouse

Challenge: Not all datasets are equal. Some requires more attention.



Agenda

- ETL
  - Traditional Data Processing Flow
- ELT Data Processing Flow
  - Medallion Architecture
- Data Processing Services
  - Lab: Fabric Data Pipeline
- Introduction to Databricks
- Introduction to Azure Synapse Analytics

Medallion Data Architecture

- Organize the data in a data lake

  Bronze -> Silver -> Gold

Medallion Data Architecture

Bronze: raw data, all as is

- Current practice is to collect data as detailed as possible

Medallion Data Architecture

Silver

- Cleansed: Handle missing, inconsistent, duplicated, and erroneous data
- Filtered: Remove unnecessary data
- Conformed: Make data type (date, string, integer) and data format (year, month, date e.g.) consistent
- Normalized/Denormalized: Convert between relational and non-relational schemas
- Feature engineering

Do you want to join (maintain foreign key) in silver stage?

- A matter of preference

Medallion Data Architecture

Gold: Ready for reporting

- Consumption-ready
- Project-specific
- Joined different datasets
- Denormalized into star schema
- Aggregated into statistics

Medallion Data Architecture

- Organize the data in a data lake

  Bronze -> Silver -> Gold

Medallion Architecture vs Traditional Staging

- Medallion Architecture and Staging are not exclusive
- Staging approach is still an important part of data warehouse ETL design
- You will see a mixture of both in your future jobs

Agenda

- ETL
  - Traditional Data Processing Flow
- ELT Data Processing Flow
  - Medallion Architecture
- Data Processing Services
  - Lab: Fabric Data Pipeline
- Introduction to Databricks
- Introduction to Azure Synapse Analytics

Medallion Architecture in Microsoft Fabric

In Microsoft Fabric, there is only one lake

Medallion architecture is achieved by using different lakehouses for each layer

- May be multiple lakehouses for same layer

Data Processing Services

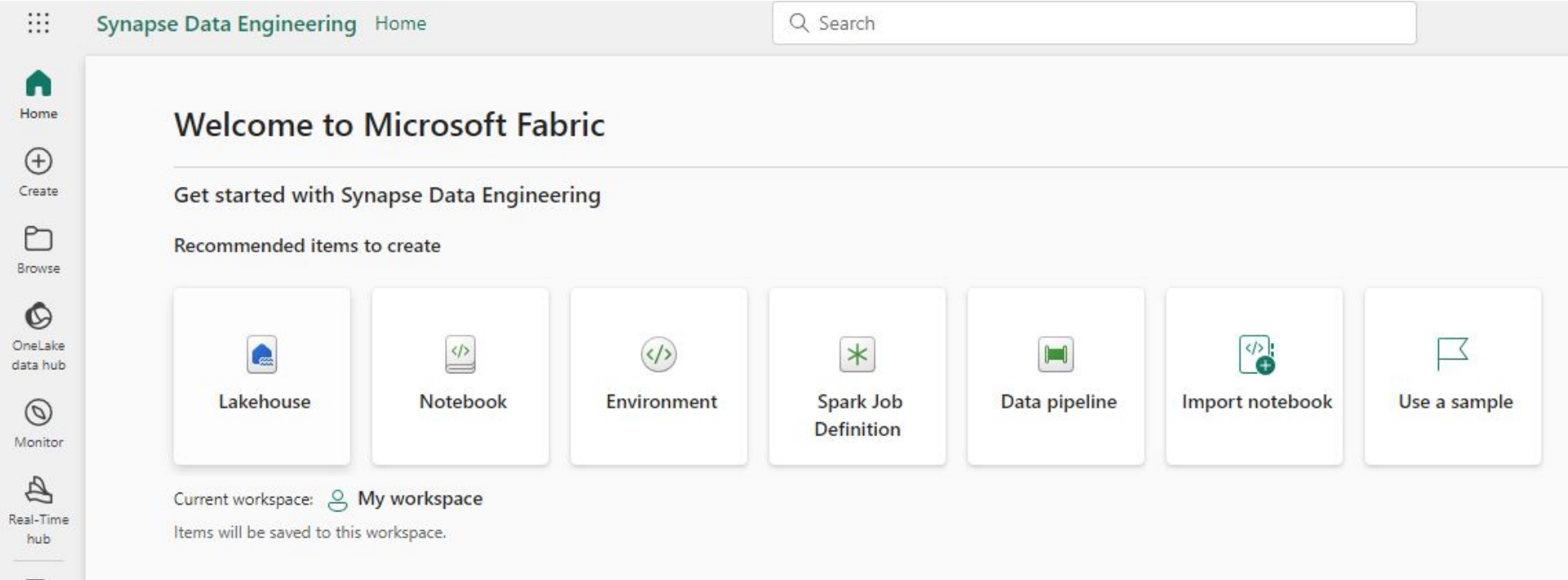
- General compute (VM, container, function)
- Analytical services offered by Microsoft
  - Fabric Data Factory/Data Engineering/Data Science
  - Synapse Data Pipeline (Azure Data Factory）

Data Processing Tools

- SQL
- Spark (python, scala, Java)
- Low-code/no-code ETL mapping

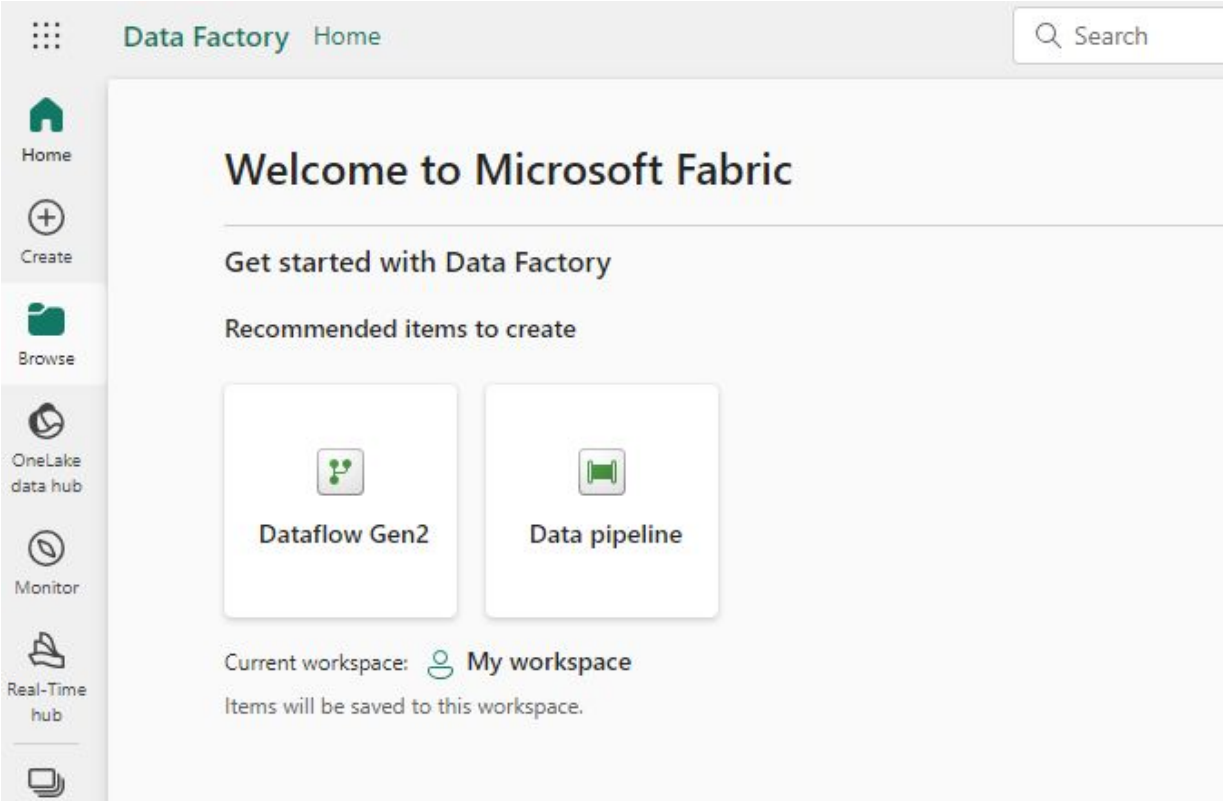
Fabric Data Engineering/Data Science

Spark notebooks, including support for python and SQL

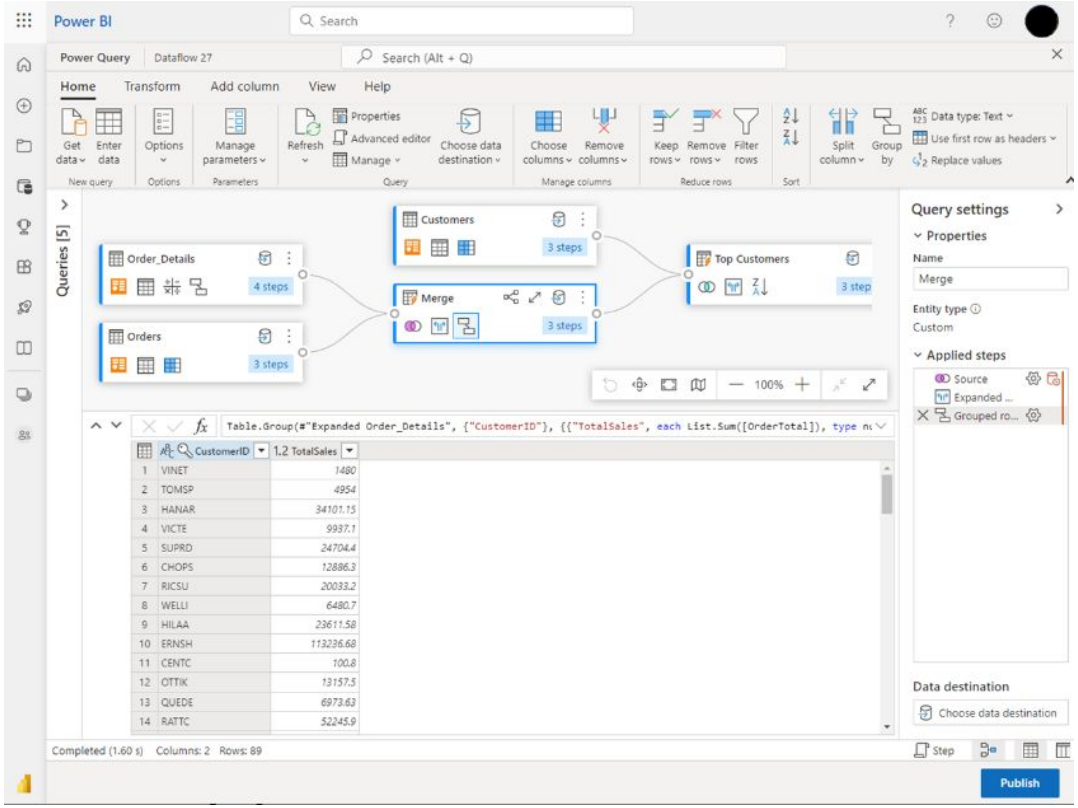


Fabric Data Factory

UI Based data transformation



Fabric Data Factory



Agenda

- ETL
  - Traditional Data Processing Flow
- ELT Data Processing Flow
  - Medallion Architecture
- Data Processing Services
  - Lab: Fabric Data Pipeline
- Introduction to Databricks
- Introduction to Azure Synapse Analytics

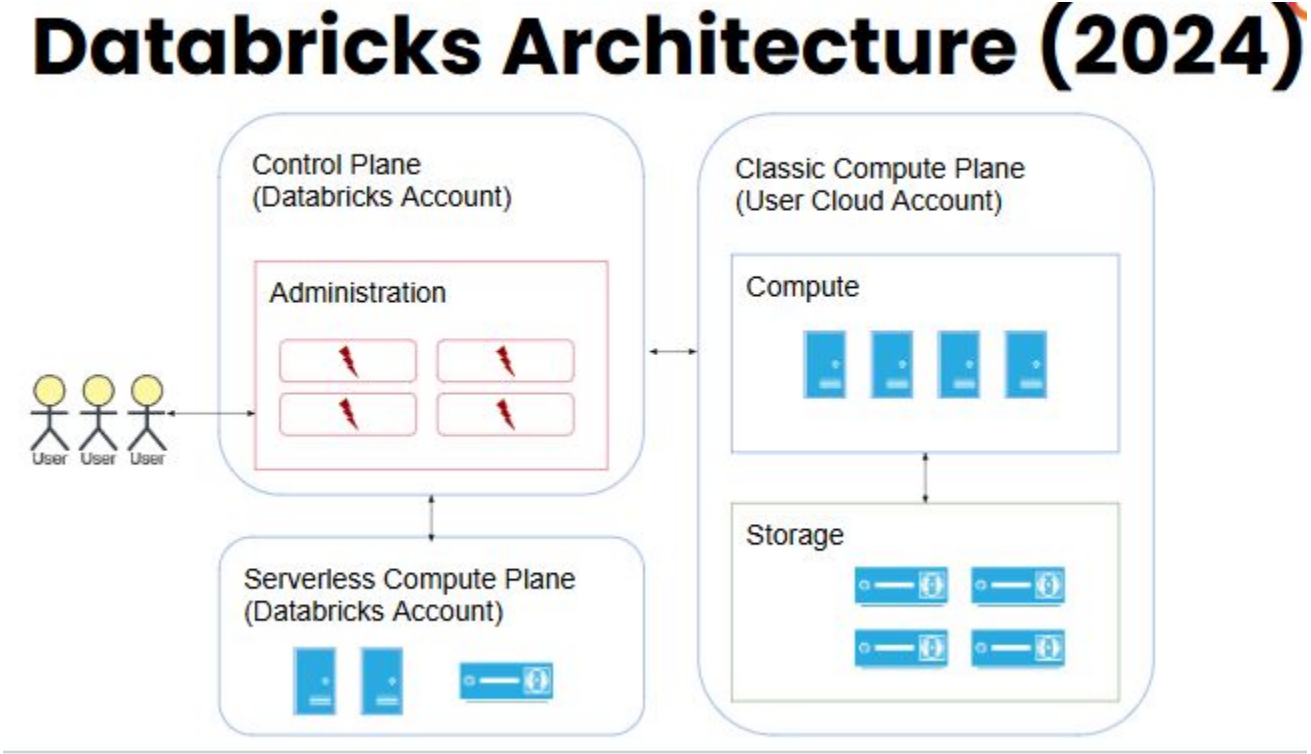
Fabric Data Engineering Demo and Lab

- Create lakehouses
- Import data
- Create notebook
  - Access files
- Create data transformation between layers
  - Create filter
  - Create new derived column: Concatenate, Substring
  - Create data type conversion
  - Create join
  - Save to target

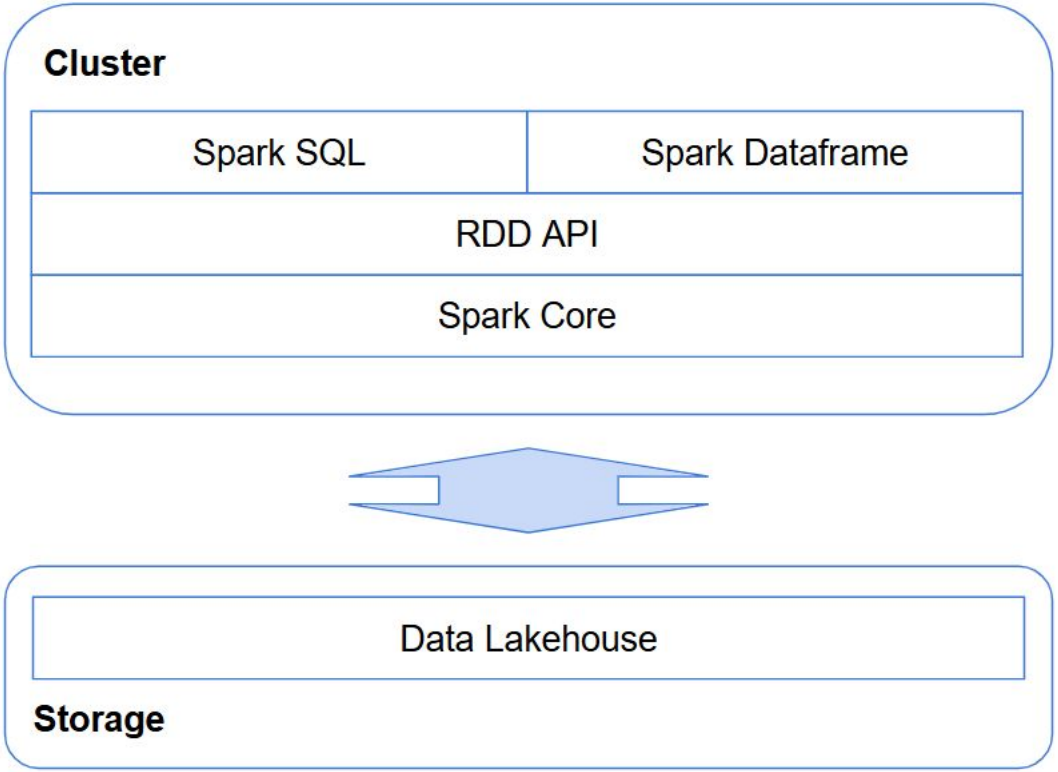
Agenda

- ETL
  - Traditional Data Processing Flow
- ELT Data Processing Flow
  - Medallion Architecture
- Data Processing Services
  - Lab: Fabric Data Pipeline
- Introduction to Databricks
- Introduction to Azure Synapse Analytics

Databricks Architecture



Databricks Data Architecture

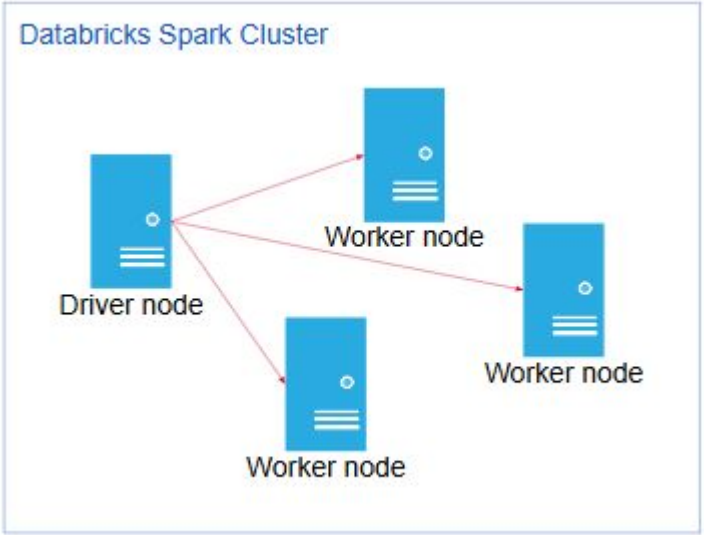


Databricks Cluster

Databricks compute is built on top of Apache Spark.

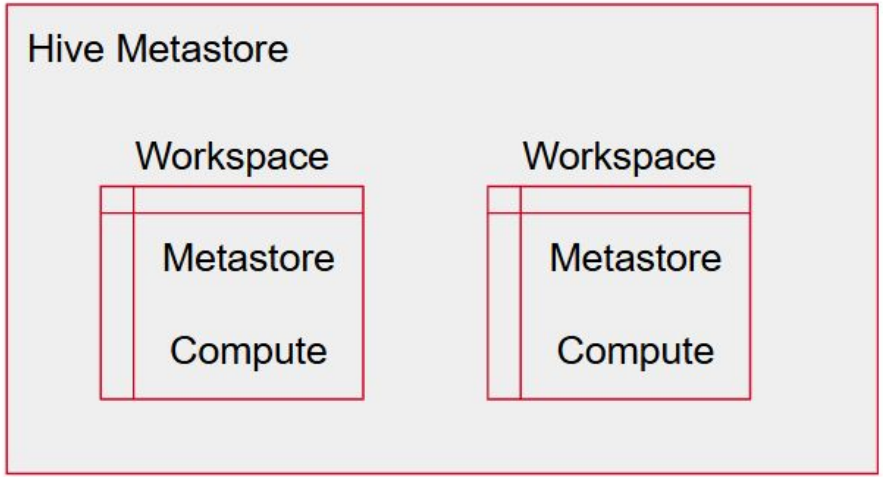
The primary unit of operation in Spark is cluster, a set of virtual machines (nodes) organized together for workload processing.

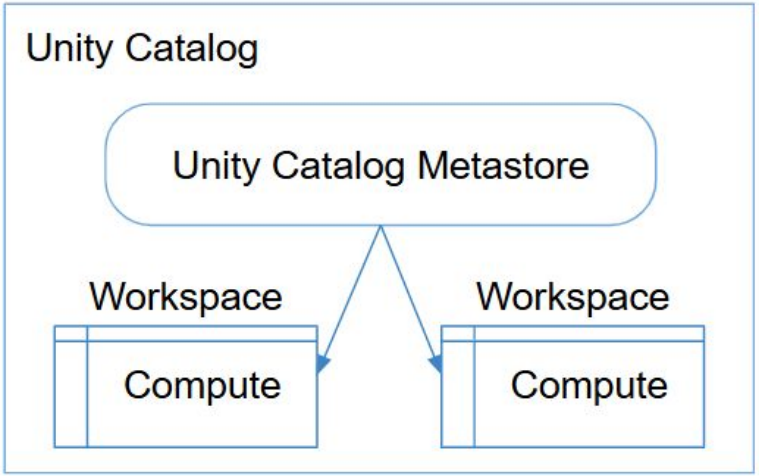
- Driver nodes
- Worker nodes
- Single-node cluster



Databricks Unity Catalog

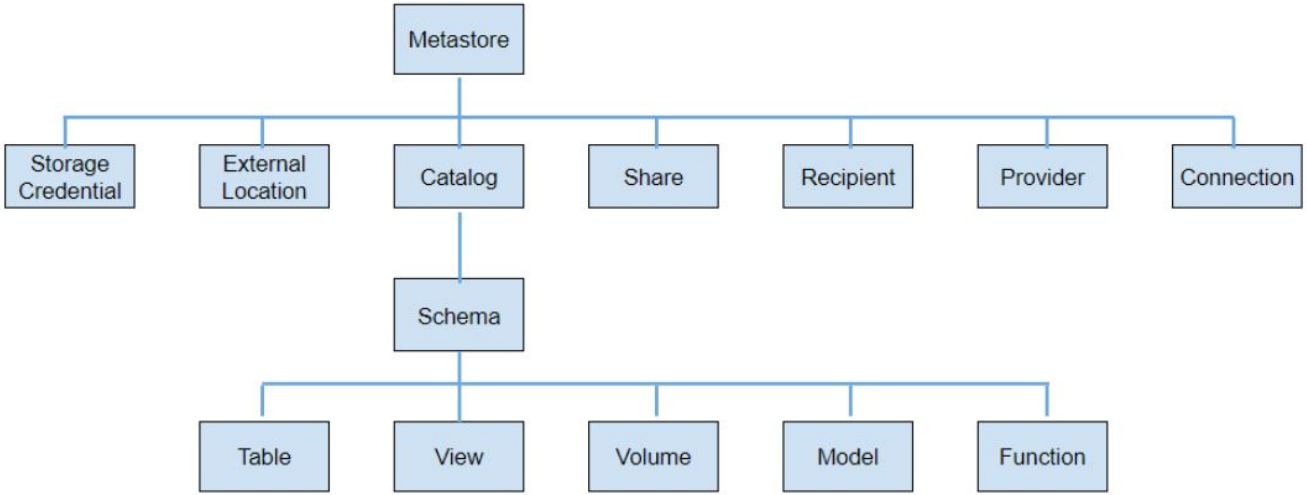
Provides centralized, cross-workspace control over access control, auditing, lineage, and data discovery





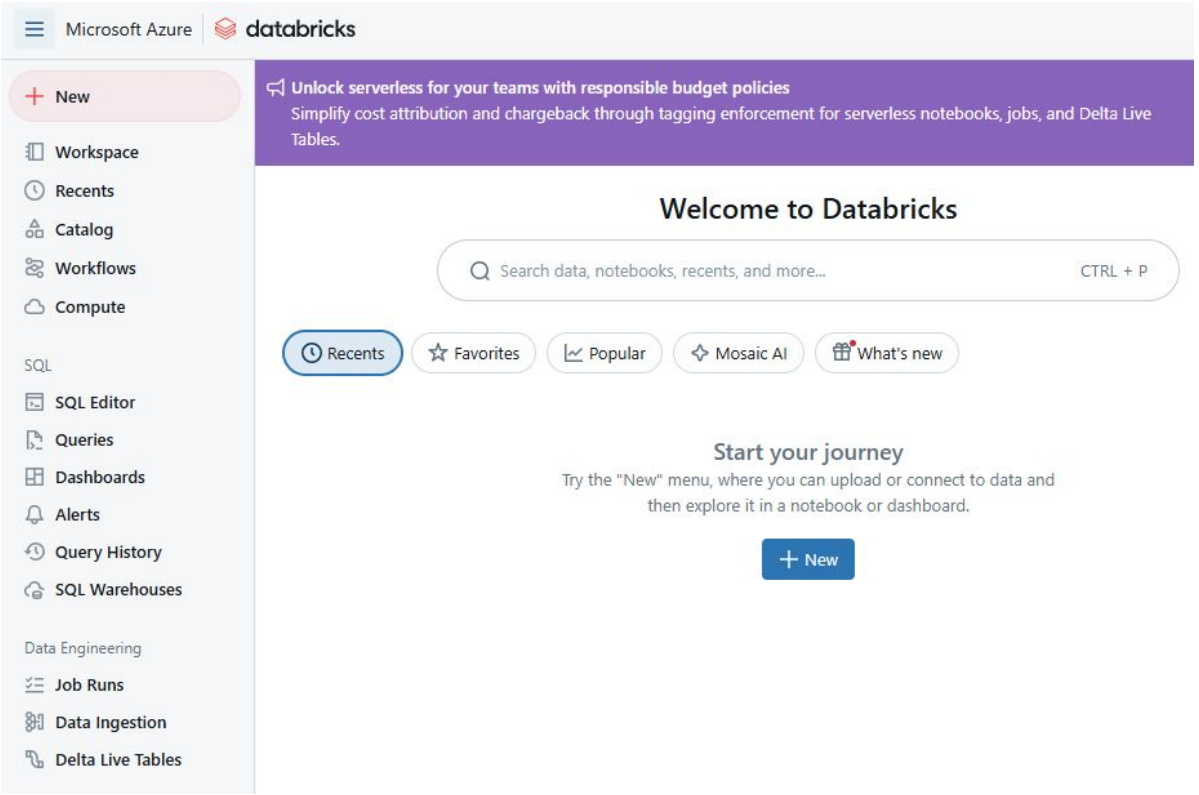
Databricks Objects

- Data engineering jobs require the manipulation of relational datasets
  - Table, view, etc.
  - Schema and higher hierarchies



Databricks User Interface

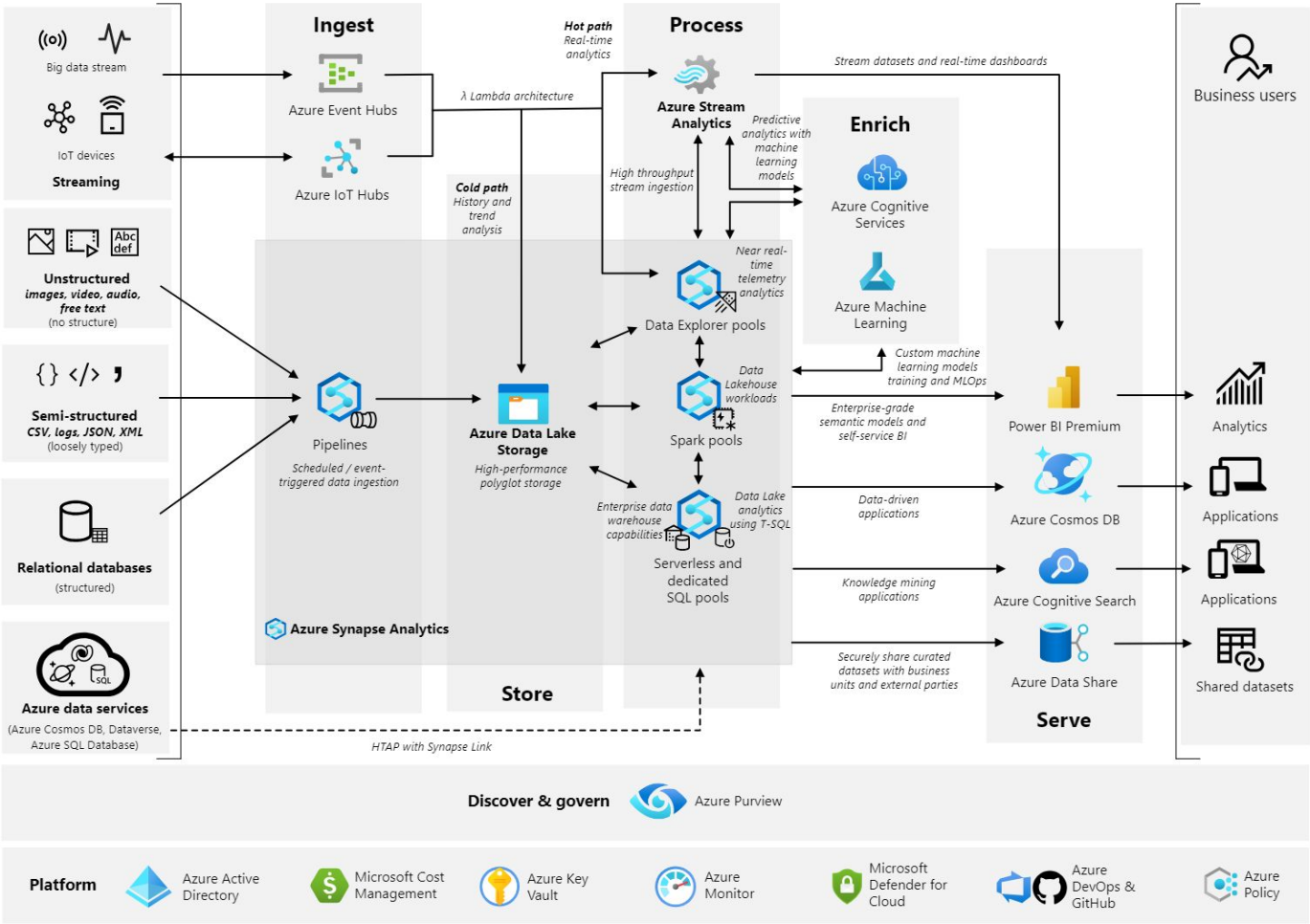
- Workspace
- Compute (cluster)
- Data Engineering
- SQL
- Catalog



Agenda

- ETL
  - Traditional Data Processing Flow
- ELT Data Processing Flow
  - Medallion Architecture
- Data Processing Services
  - Lab: Fabric Data Pipeline
- Introduction to Databricks
- Introduction to Azure Synapse Analytics

Azure Synapse Analytics



Microsoft Azure

Azure Synapse Analytics Architecture

Data Engineering

- Dedicated SQL Pool: Formally SQL Data Warehouse. A full scale data warehouse product
- Serverless SQL Pool: Lakehouse solution
- Spark Pool: Running Spark ecosystem

Storage

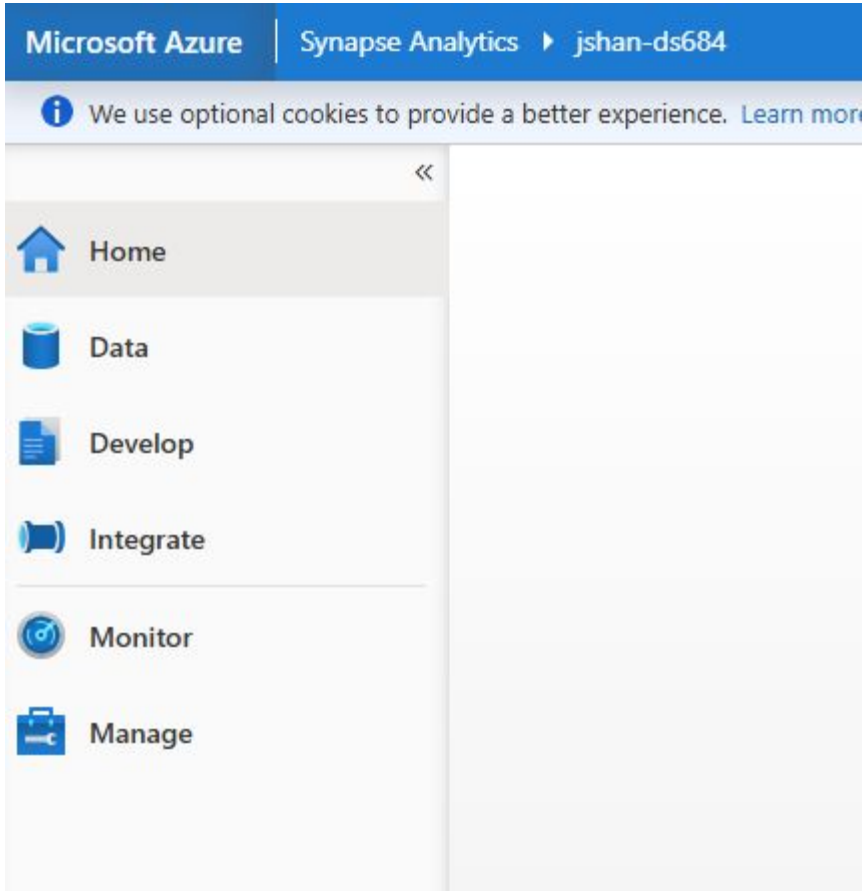
- Azure Data Lake
- Synapse Data Warehouse

Synapse User Interface

Data: storage

Develop: compute

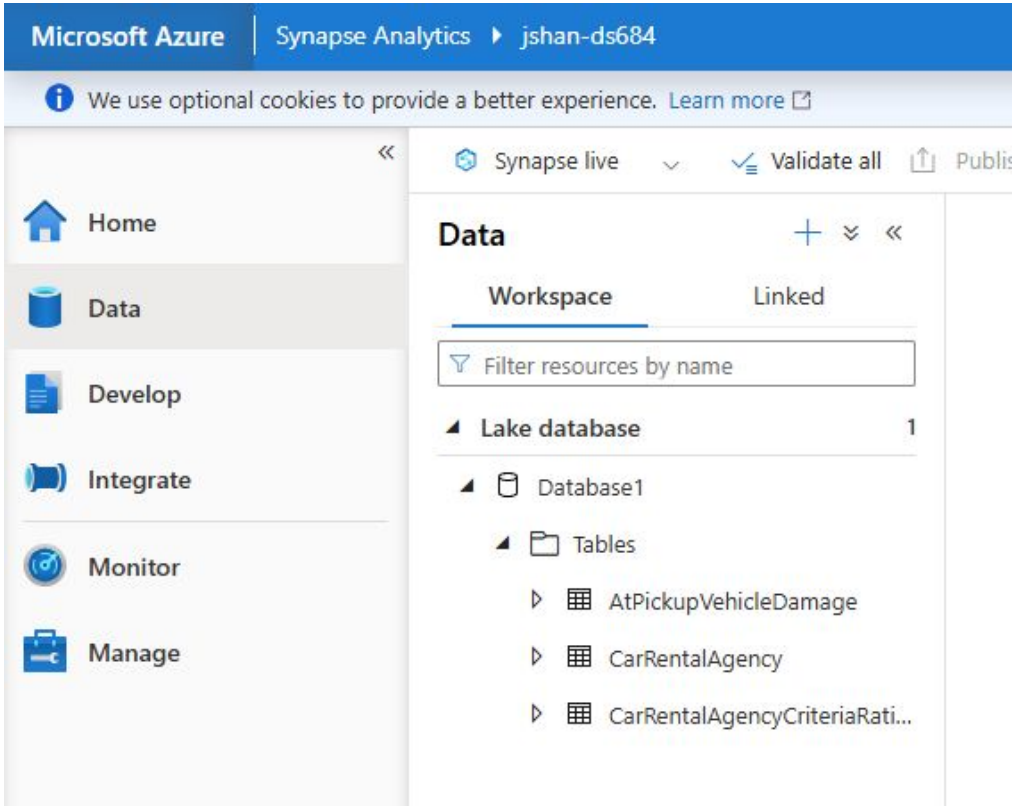
Integrate: Orchestration



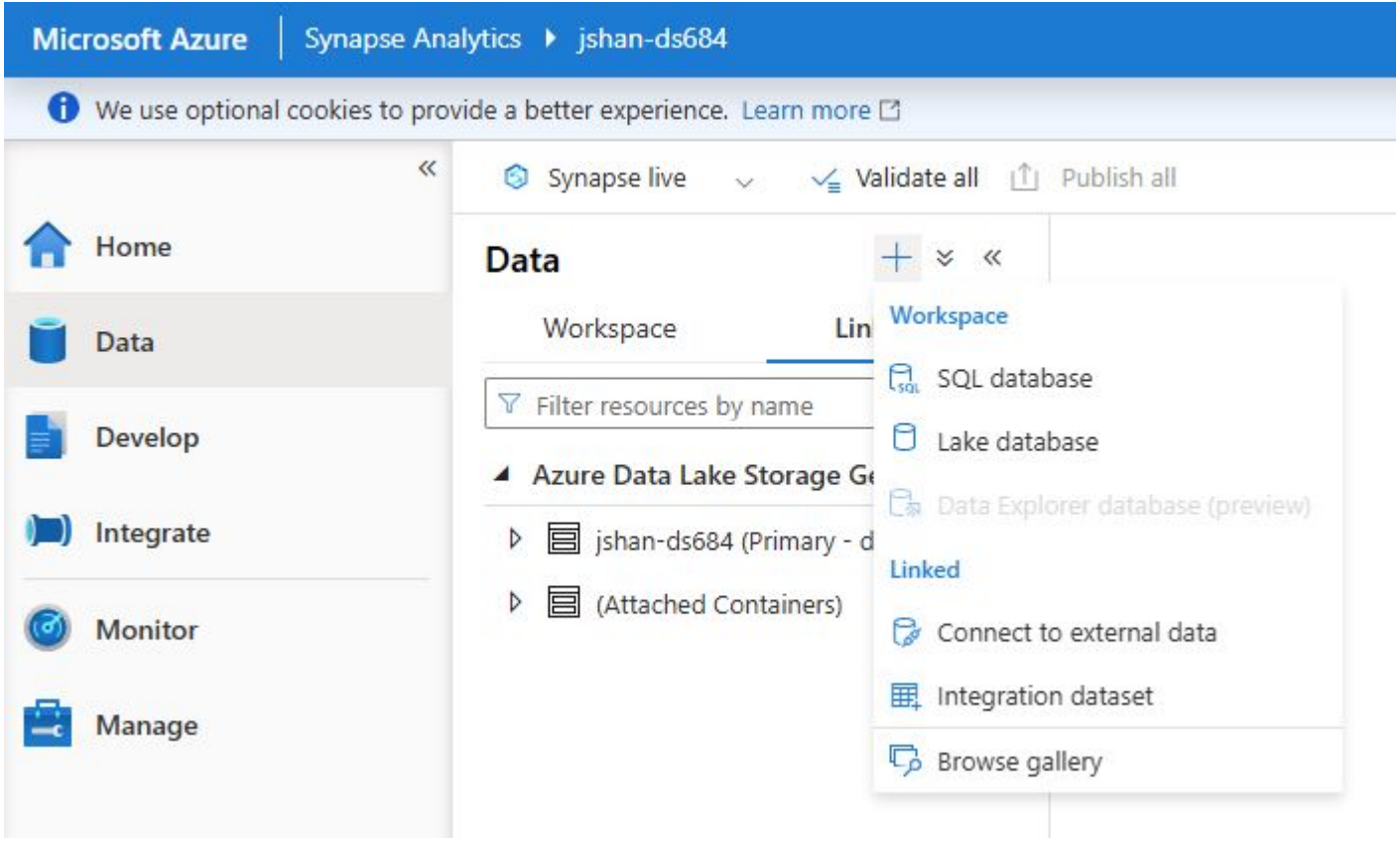
Synapse User Interface

Manually map files in data lake into Synapse Lakehouse, or

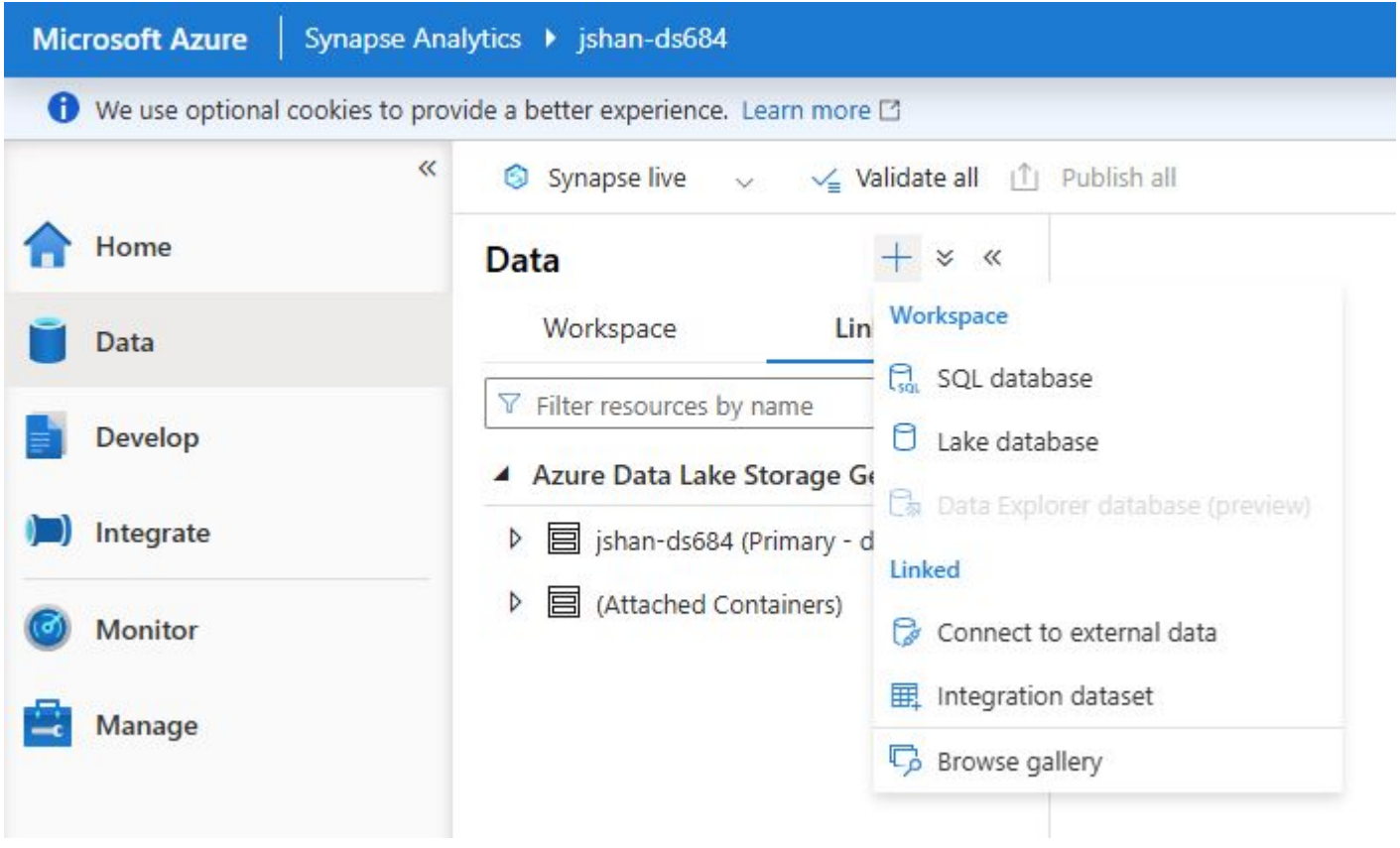
Create managed files



Integration Dataset



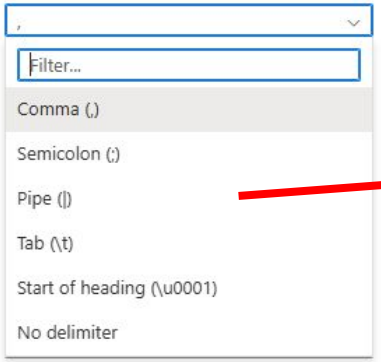
Integration Dataset

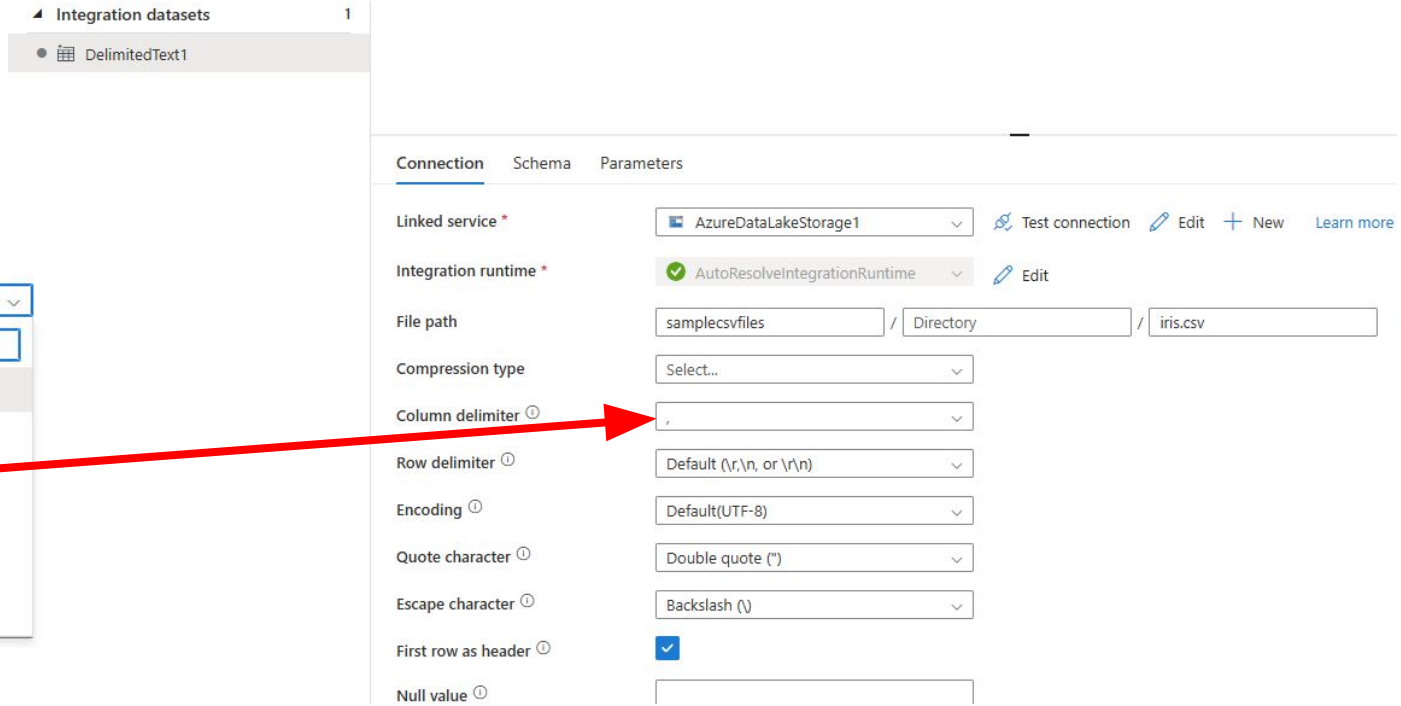


Integration Dataset

Flexible but manual: Fabric's options are much limited.

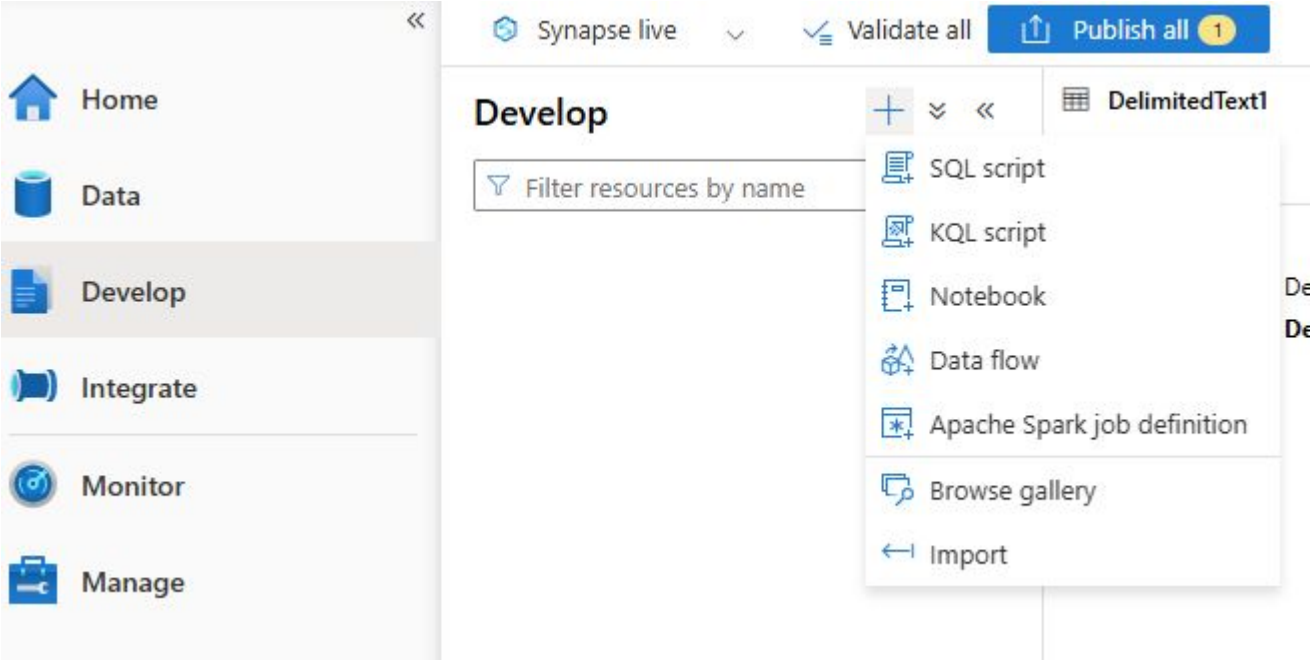
- Delimiter
- Header
- etc.



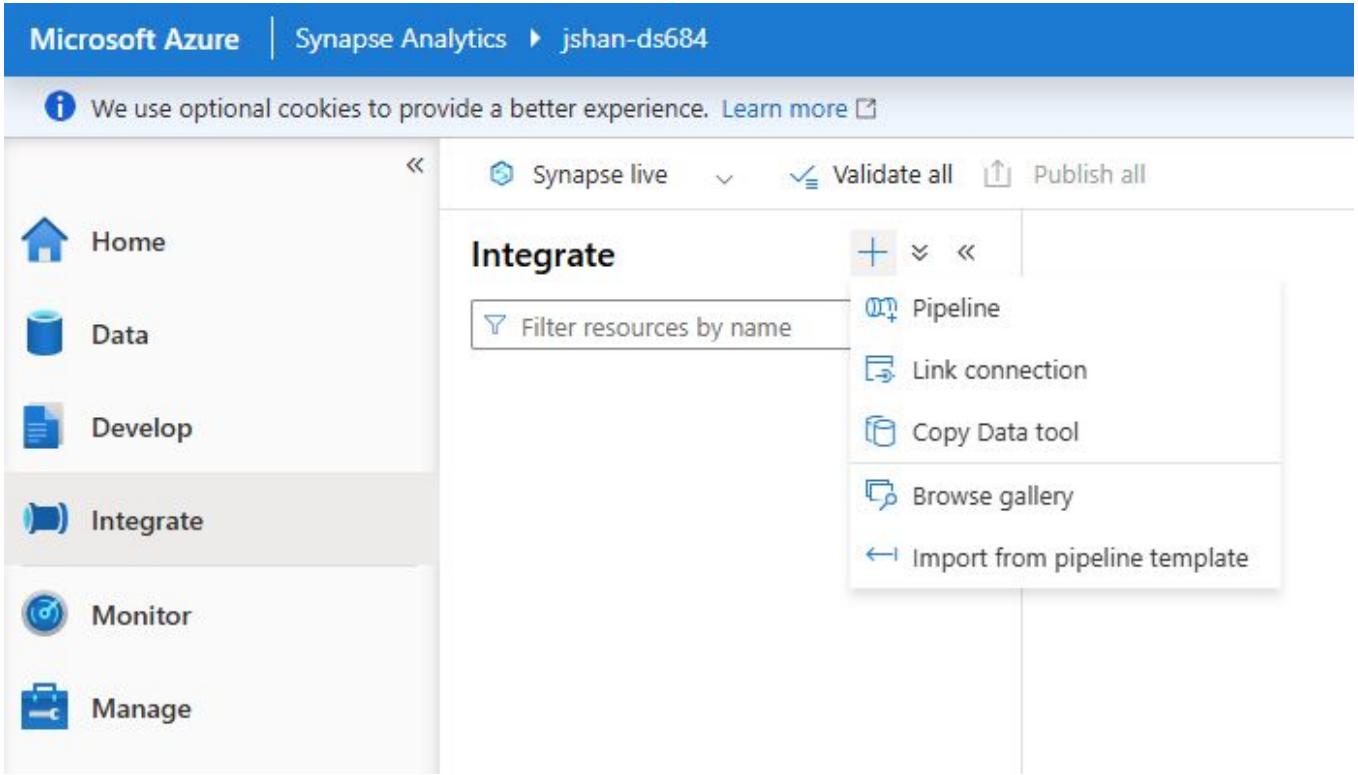


Development Tools

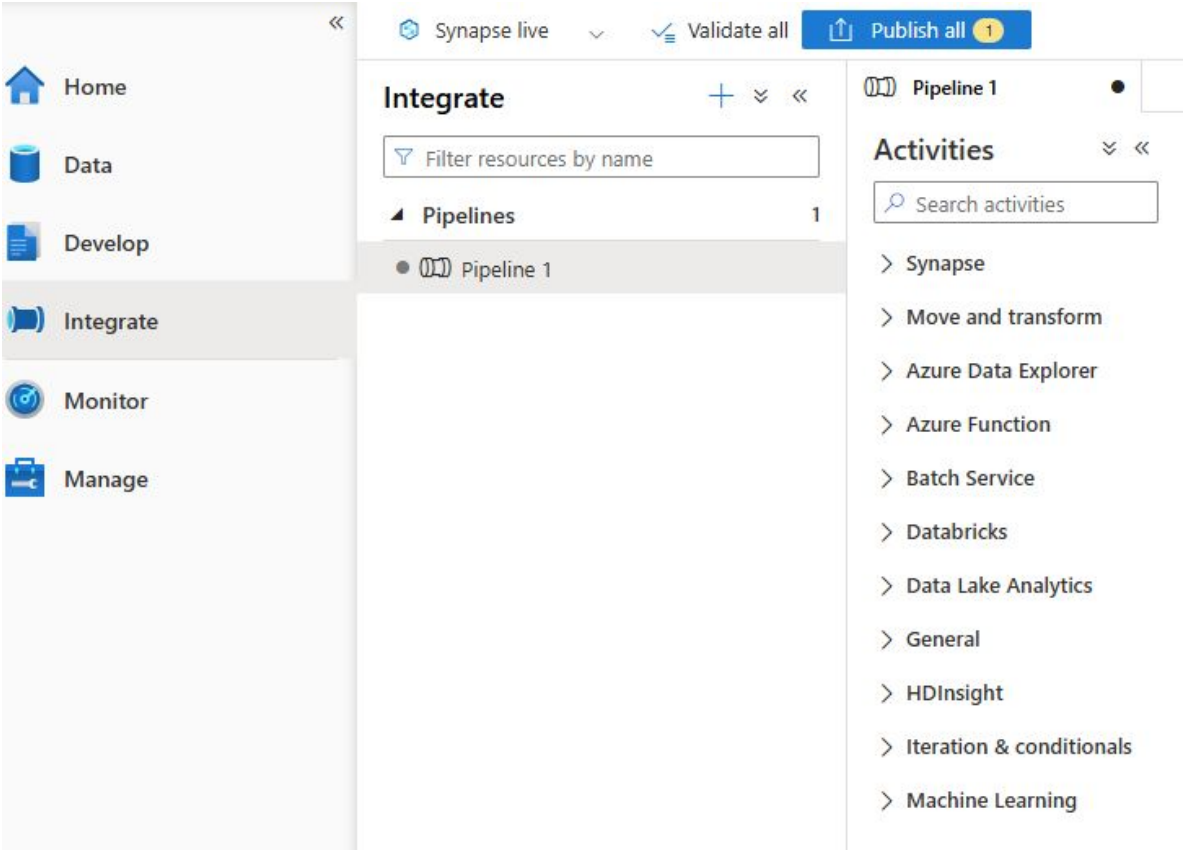
Many compute options, but not conveniently integrated



Integrate and Orchestration

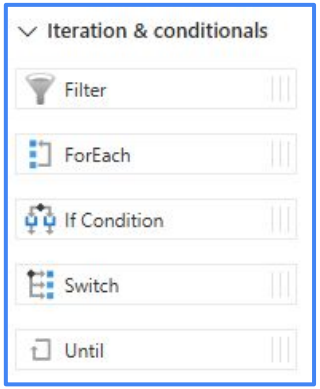


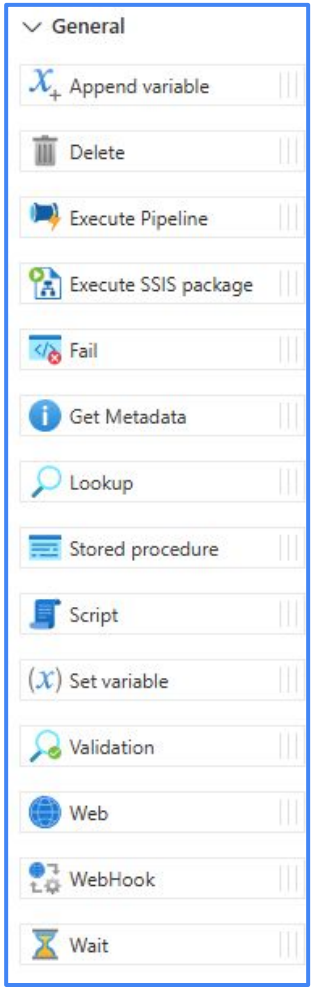
Integrate and Orchestration

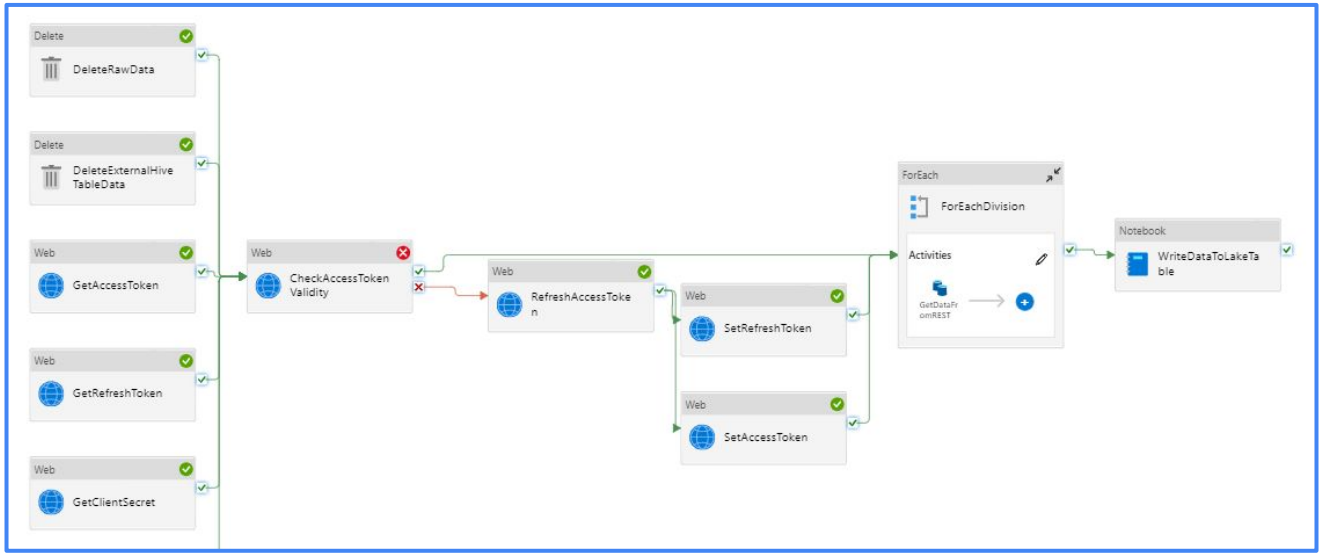


Integrate and Orchestration

Many available tools







Designing the Final Project

How would you design your final project database?

Final Project

Review Assignment 07

- Table creation

Assignment 08

- Data processing
  - End result: a consolidated dataset