

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)

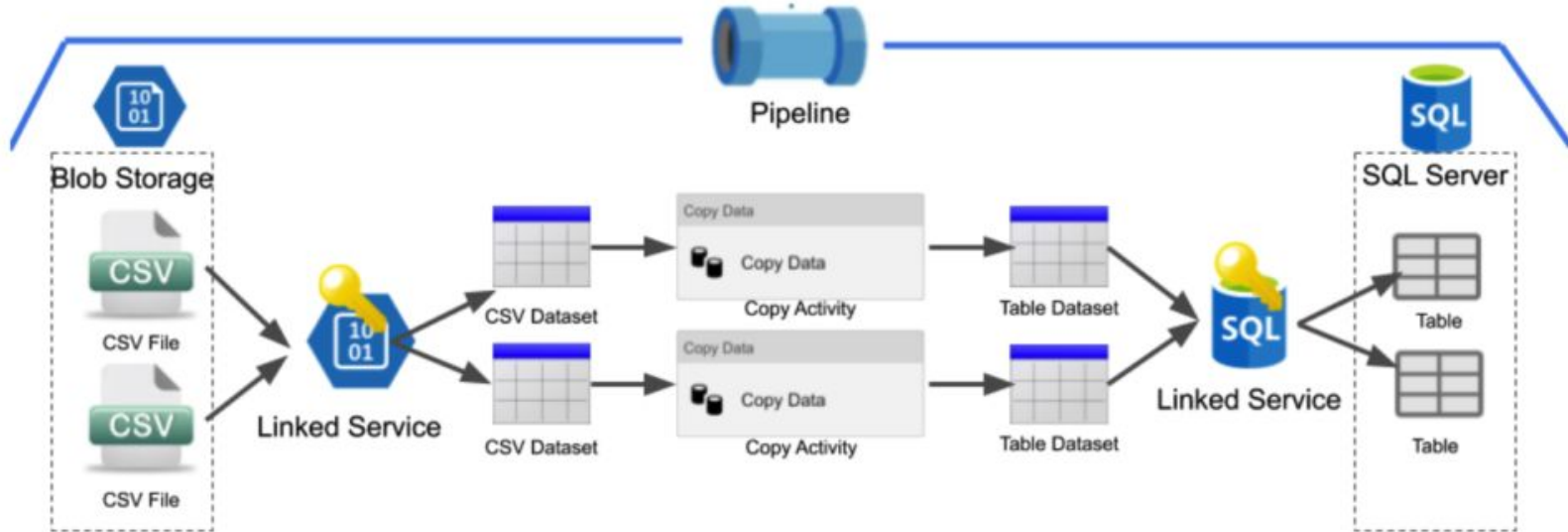
Different source systems will generate data that

- 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.

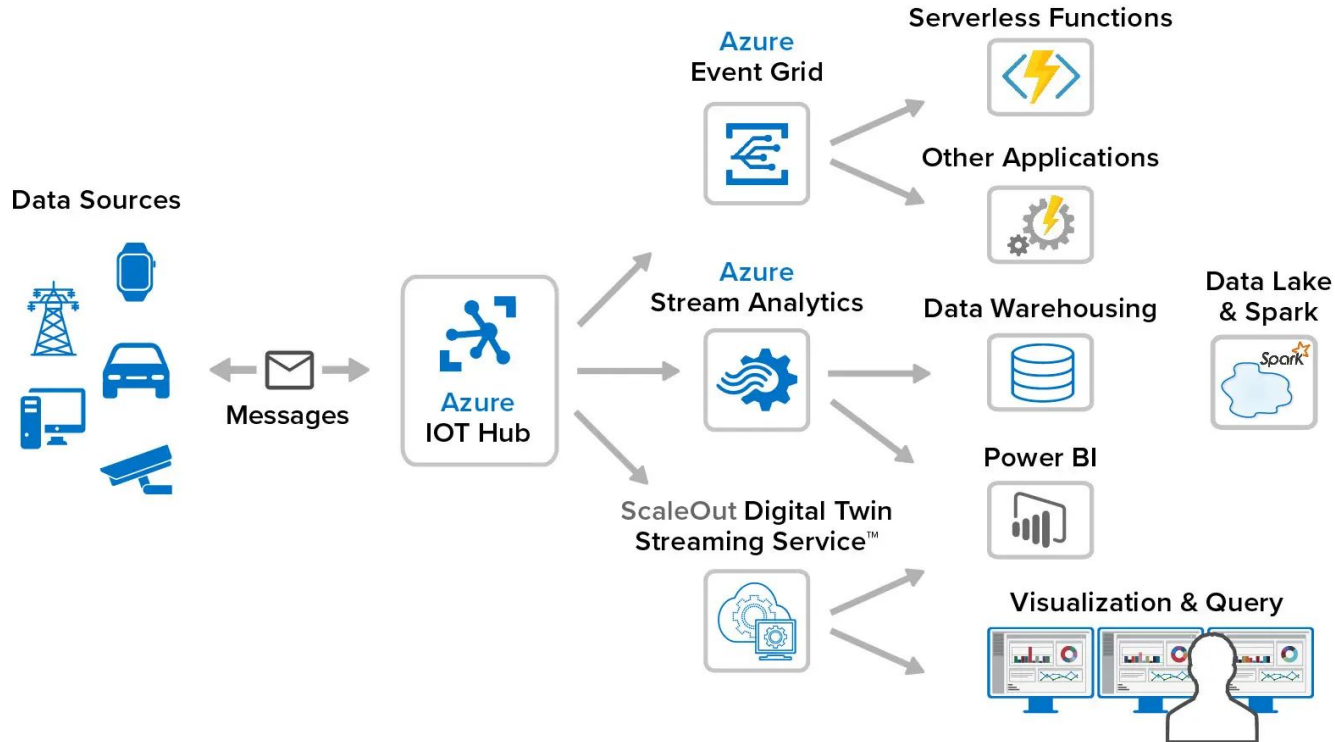
Extract, Transform, Load (ETL)



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

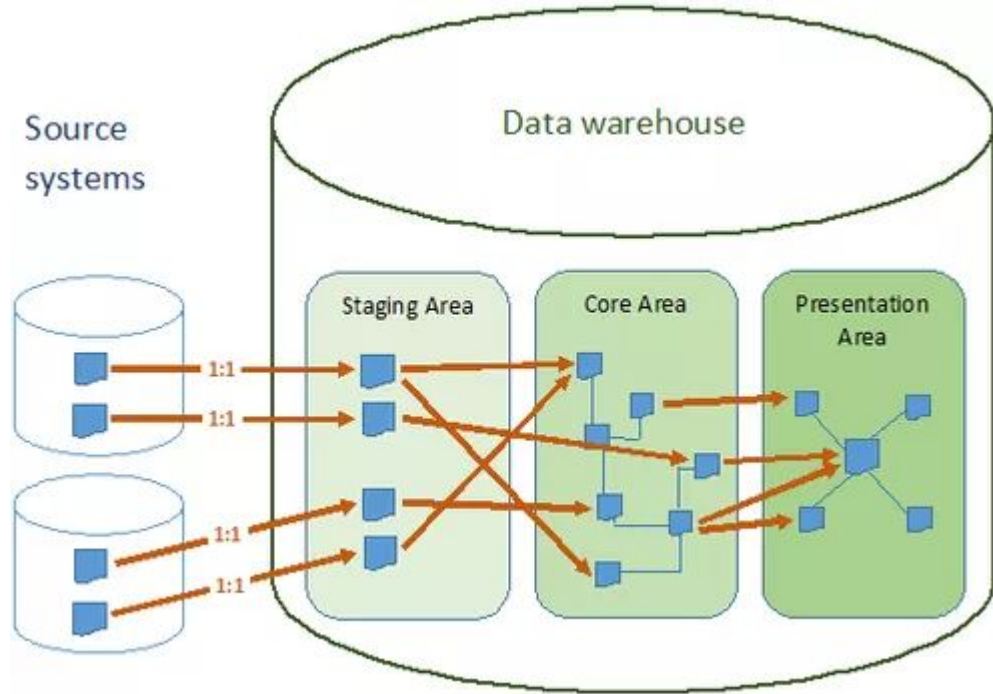
Source ->

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

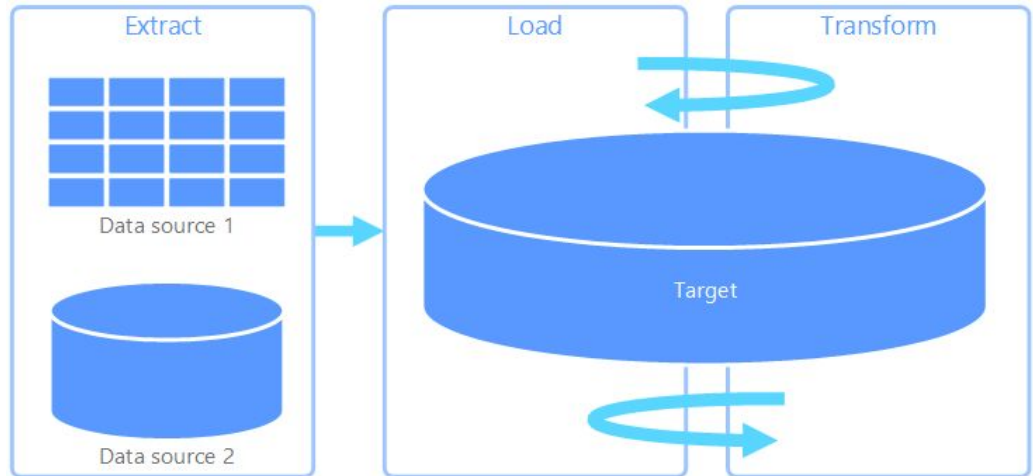
- ETL
 - Traditional Data Processing Flow
- **ELT Data Processing Flow**
 - Medallion Architecture
- Data Processing Services
 - Lab: Fabric Data Pipeline
- Introduction to Databricks
- Introduction to Azure Synapse Analytics

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

The screenshot shows the Microsoft Fabric Synapse Data Engineering interface. At the top, there is a header bar with the text "Synapse Data Engineering" and "Home" on the left, and a search bar on the right. Below the header, the main content area is titled "Welcome to Microsoft Fabric". Underneath this title, it says "Get started with Synapse Data Engineering". A section titled "Recommended items to create" displays seven cards: "Lakehouse", "Notebook", "Environment", "Spark Job Definition", "Data pipeline", "Import notebook", and "Use a sample". Each card has a corresponding icon. On the left side of the interface, there is a vertical sidebar with icons and labels for "Home", "Create", "Browse", "OneLake data hub", "Monitor", and "Real-Time hub". At the bottom of the main content area, it says "Current workspace: My workspace" and "Items will be saved to this workspace."

Synapse Data Engineering Home

Search

Welcome to Microsoft Fabric

Get started with Synapse Data Engineering

Recommended items to create

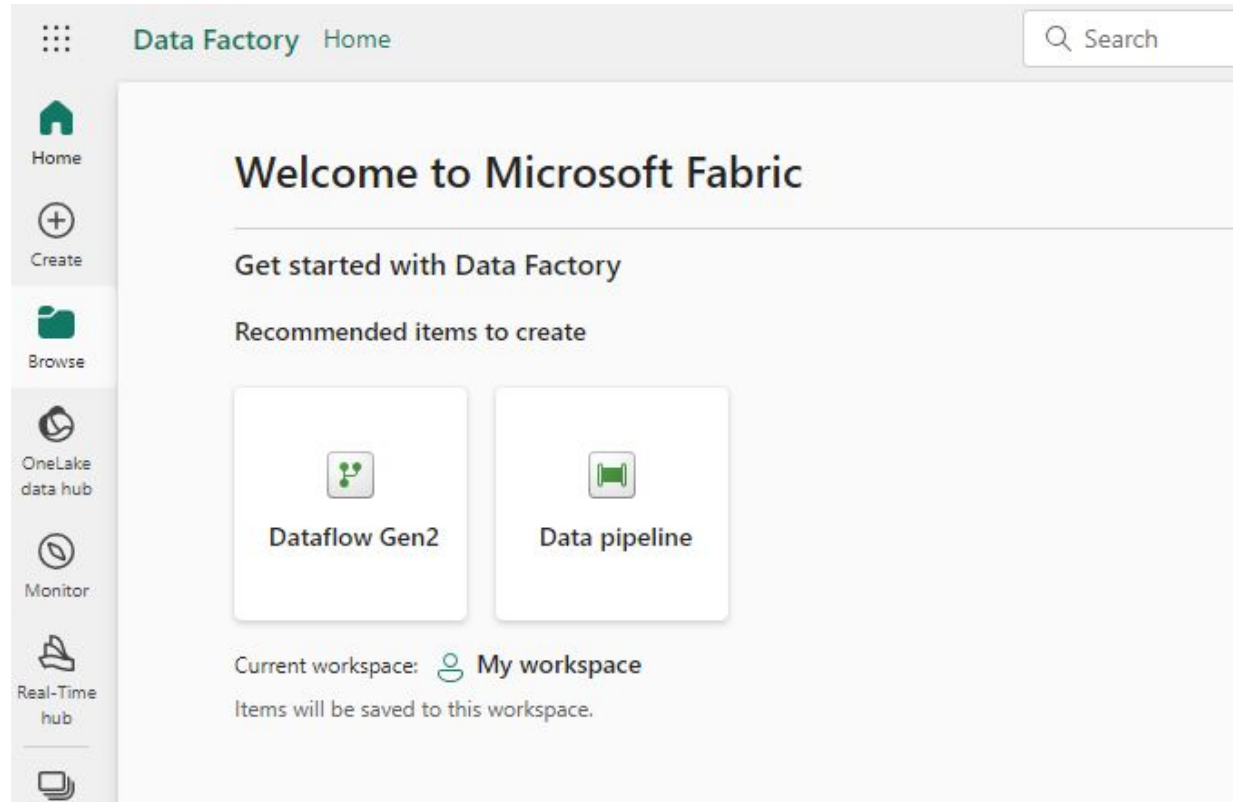
- Lakehouse
- Notebook
- Environment
- Spark Job Definition
- Data pipeline
- Import notebook
- Use a sample

Current workspace: My workspace

Items will be saved to this workspace.

Fabric Data Factory

UI Based data
transformation



Fabric Data Factory

Power BI Dataflow 27

Search (Alt + Q)

Home Transform Add column View Help

Get data Enter data Options Manage parameters Refresh Advanced editor Choose data destination Choose columns Remove columns Keep rows Remove rows Filter rows Sort Split column Group by

Queries [5]

Order_Details (4 steps)

Orders (3 steps)

Merge (3 steps)

Customers (3 steps)

Top Customers (3 steps)

Query settings

Properties

Name: Merge

Entity type: Custom

Applied steps

Source, Expanded, Grouped rows

Data destination

Choose data destination

Table.Group(#"Expanded Order_Details", {"CustomerID"}, {"TotalSales", each List.Sum([OrderTotal]), type nullable number)

CustomerID	TotalSales
1 VINET	1480
2 TOMSP	4954
3 HANAR	34101.15
4 VICTE	9937.1
5 SUPRD	24704.4
6 CHOPS	12886.3
7 RICSU	20033.2
8 WELU	6480.7
9 HILAA	23611.58
10 ERNSH	113236.68
11 CENTC	100.8
12 OTTIK	13157.5
13 QUEDE	6973.63
14 RATTC	52245.9

Completed (1.60 s) Columns: 2 Rows: 89

Publish

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

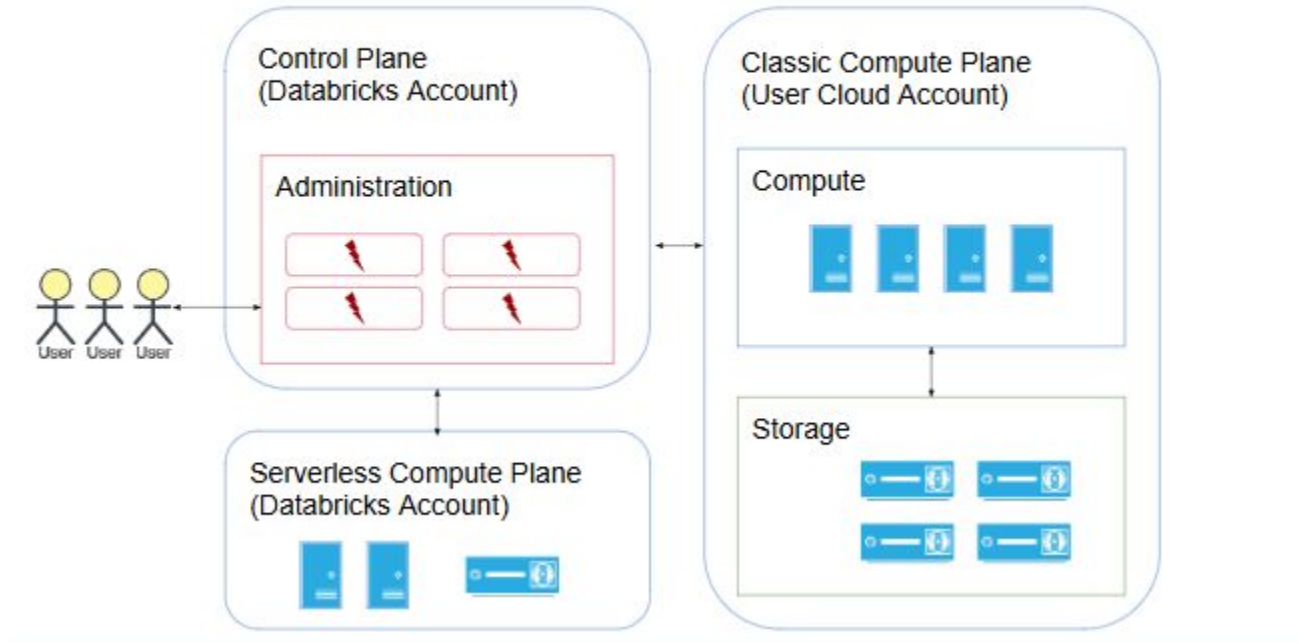
- Create lakehouses
- Import data
- Create notebook
 - Access files
- Create data transformation between layers
 - Create filter
 - Create new derived column: Concatenate, Substring
 - Create data type conversion
 - Create join
 - Save to target

Agenda

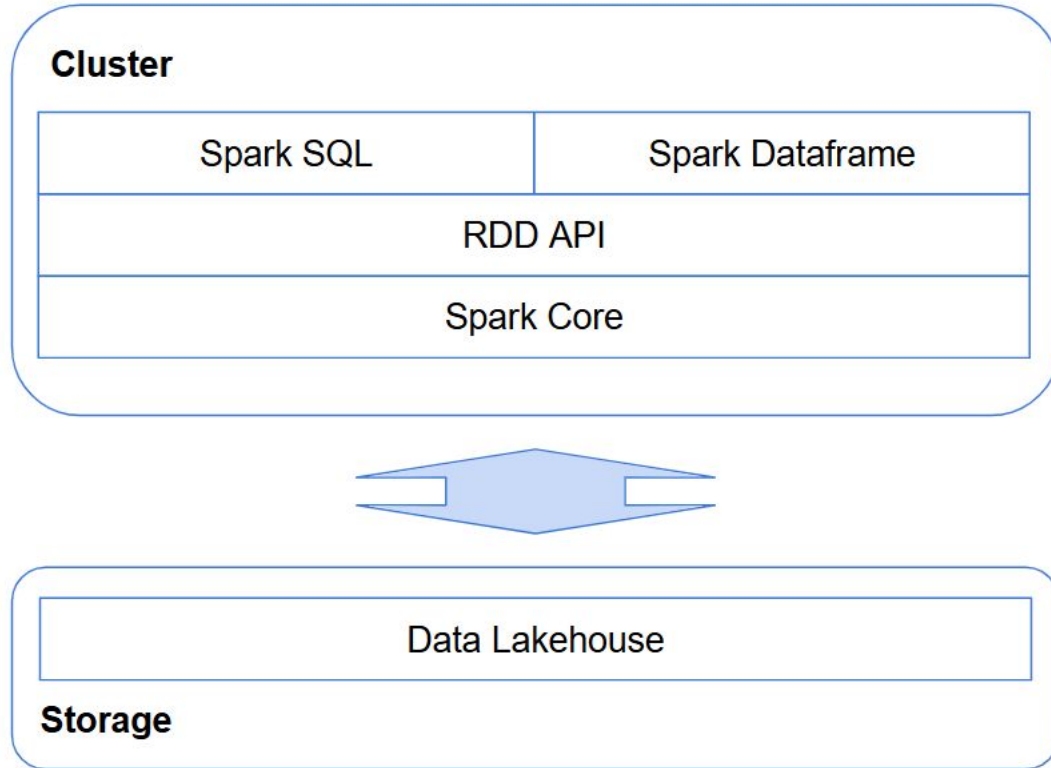
- ETL
 - Traditional Data Processing Flow
- ELT Data Processing Flow
 - Medallion Architecture
- Data Processing Services
 - Lab: Fabric Data Pipeline
- **Introduction to Databricks**
- Introduction to Azure Synapse Analytics

Databricks Architecture

Databricks Architecture (2024)



Databricks Data Architecture

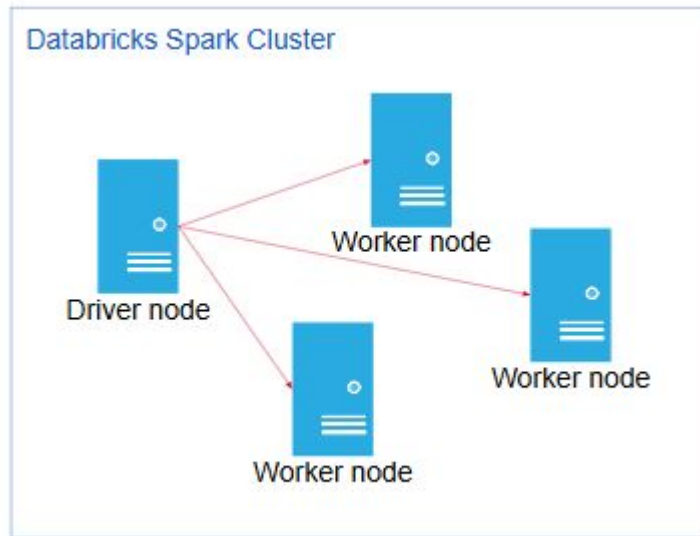


Databricks Cluster

Databricks compute is built on top of Apache Spark.

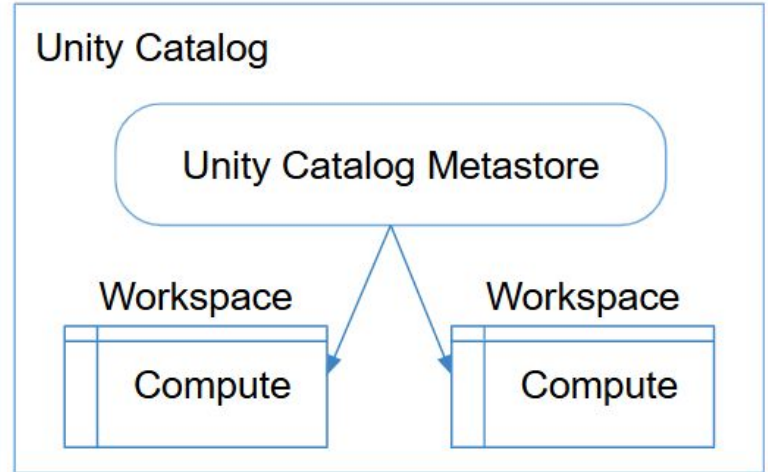
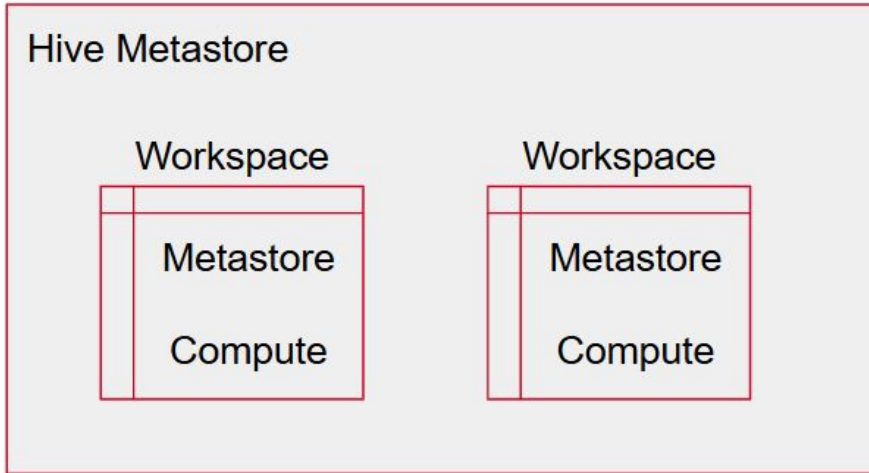
The primary unit of operation in Spark is cluster, a set of virtual machines (nodes) organized together for workload processing.

- Driver nodes
- Worker nodes
- Single-node cluster



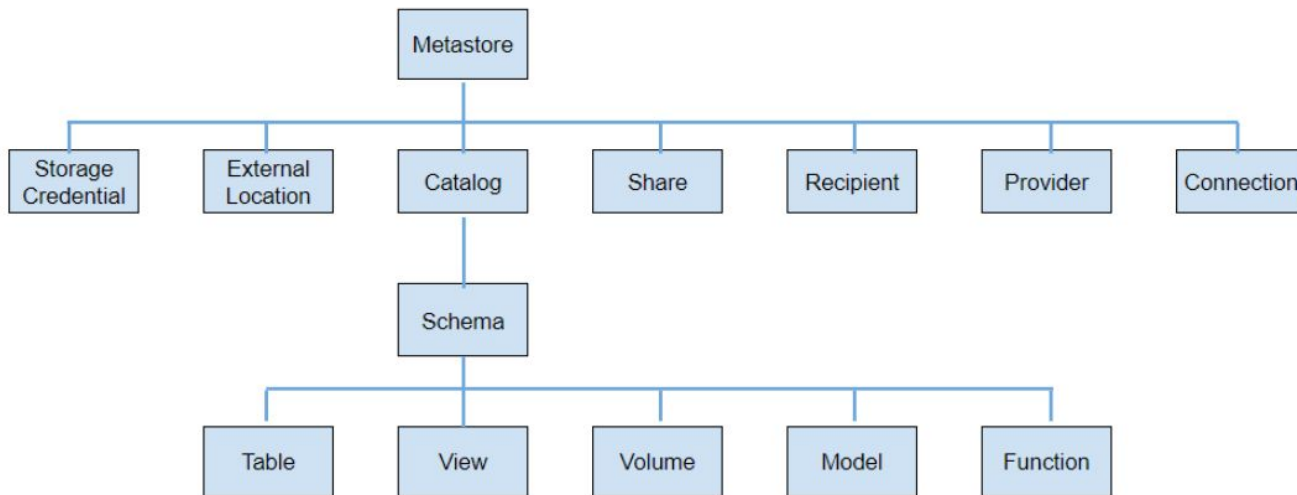
Databricks Unity Catalog

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



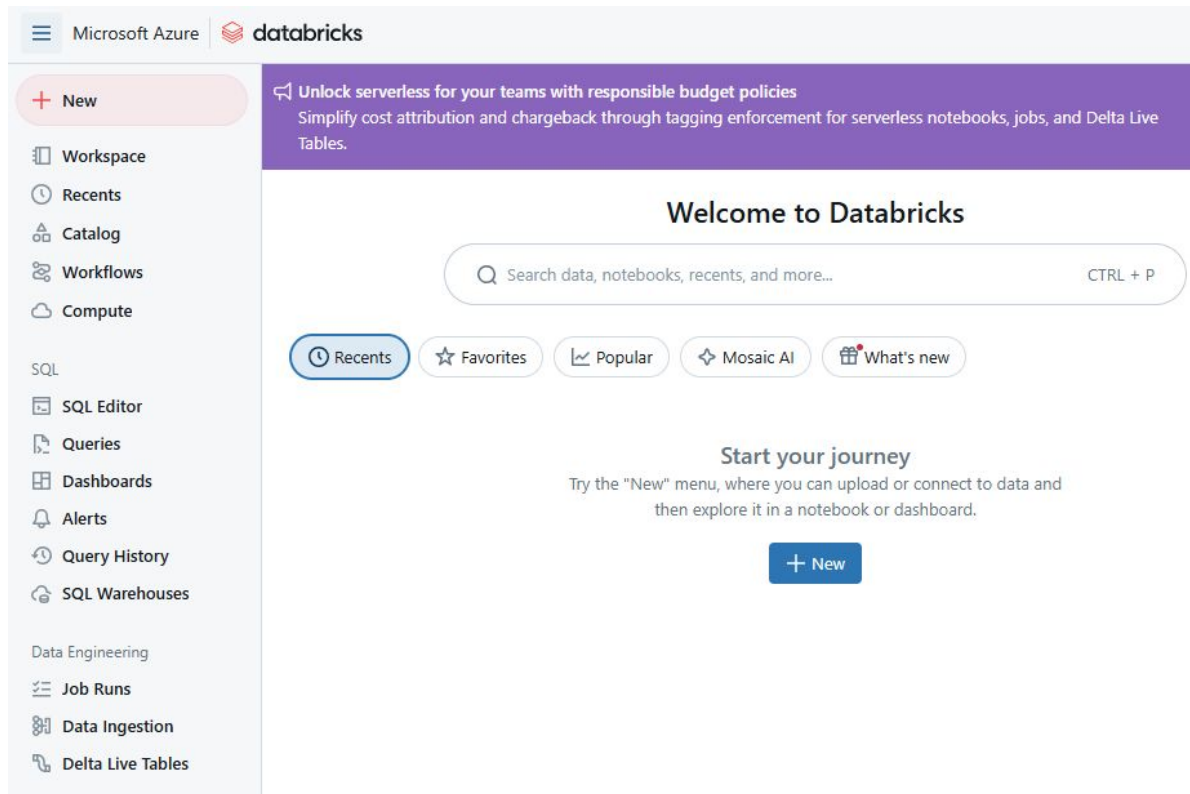
Databricks Objects

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



Databricks User Interface

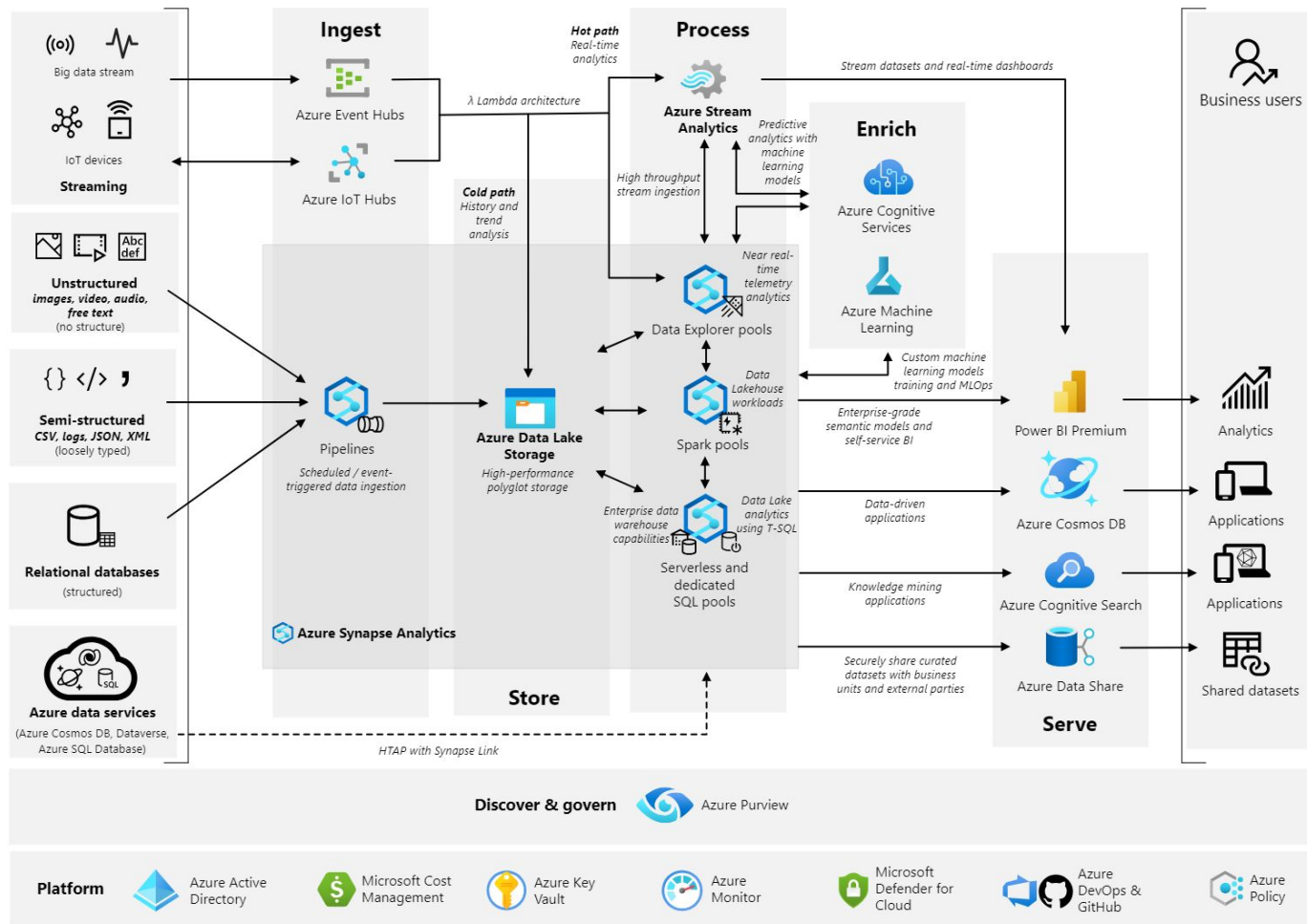
- Workspace
- Compute (cluster)
- Data Engineering
- SQL
- Catalog



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



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

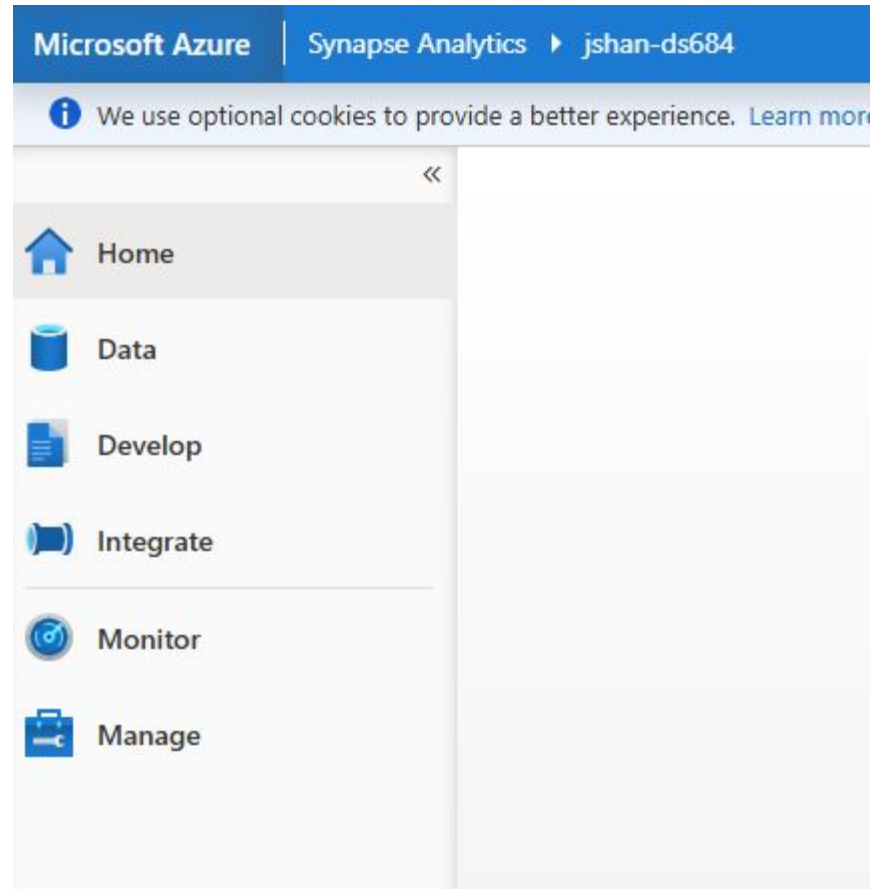
- Azure Data Lake
- Synapse Data Warehouse

Synapse User Interface

Data: storage

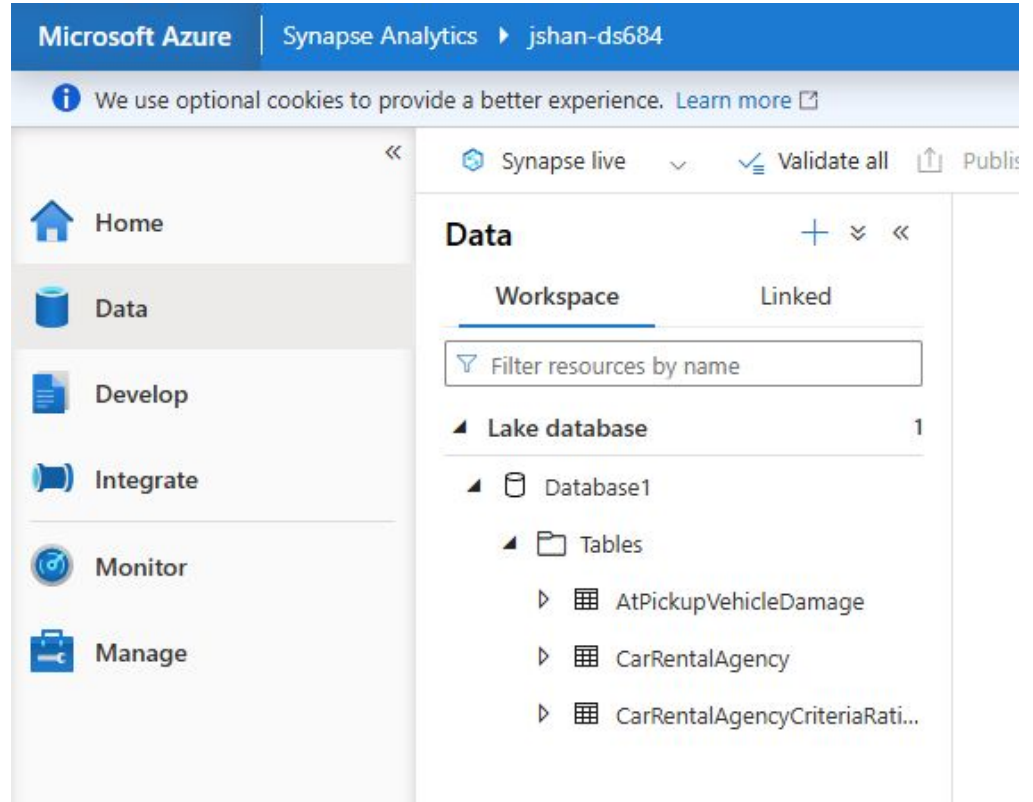
Develop: compute

Integrate: Orchestration



Synapse User Interface

Manually map files in data lake
into Synapse Lakehouse, or
Create managed files



Integration Dataset

The screenshot displays the Microsoft Azure Synapse Analytics user interface. At the top, the header shows 'Microsoft Azure | Synapse Analytics' followed by the workspace name 'jshan-ds684'. A cookie notice is visible below the header. The left sidebar contains navigation icons for Home, Data (selected), Develop, Integrate, Monitor, and Manage. The main content area is titled 'Data' and includes a 'Workspace' tab and a 'Link' tab. A search bar labeled 'Filter resources by name' is present. Below the search bar, a list of resources is shown under the heading 'Azure Data Lake Storage Gen2', including 'jshan-ds684 (Primary - d...' and '(Attached Containers)'. A dropdown menu is open, showing options for creating a new dataset: 'Workspace', 'SQL database', 'Lake database', 'Data Explorer database (preview)', 'Linked', 'Connect to external data', 'Integration dataset' (highlighted), and 'Browse gallery'.

Microsoft Azure | Synapse Analytics ▶ jshan-ds684

We use optional cookies to provide a better experience. [Learn more](#)

Home Data Develop Integrate Monitor Manage

Synapse live Validate all Publish all

Data

Workspace Link

Filter resources by name

Azure Data Lake Storage Gen2

- ▶ jshan-ds684 (Primary - d...
- ▶ (Attached Containers)

Integration Dataset Options:

- Workspace
- SQL database
- Lake database
- Data Explorer database (preview)
- Linked
- Connect to external data
- Integration dataset
- Browse gallery

Integration Dataset

The screenshot displays the Microsoft Azure Synapse Analytics user interface. At the top, the header shows 'Microsoft Azure | Synapse Analytics' followed by the workspace name 'jshan-ds684'. A notification banner below the header states: 'We use optional cookies to provide a better experience. [Learn more](#)'. The left sidebar contains navigation icons and labels for 'Home', 'Data' (which is selected), 'Develop', 'Integrate', 'Monitor', and 'Manage'. The main content area is titled 'Data' and includes a 'Workspace' tab and a 'Link' tab. Below these tabs is a search bar labeled 'Filter resources by name'. Under the 'Workspace' tab, there are two expandable sections: 'Azure Data Lake Storage Gen2' containing 'jshan-ds684 (Primary - d...)' and '(Attached Containers)'. A dropdown menu is open, showing options for creating or linking data sources: 'Workspace' (selected), 'SQL database', 'Lake database', 'Data Explorer database (preview)', 'Linked' (a sub-section header), 'Connect to external data', 'Integration dataset', and 'Browse gallery'.

Microsoft Azure | Synapse Analytics ▶ jshan-ds684

We use optional cookies to provide a better experience. [Learn more](#)

Home Data Develop Integrate Monitor Manage

Synapse live Validate all Publish all

Data

Workspace Link

Filter resources by name

▲ Azure Data Lake Storage Gen2

- ▶ jshan-ds684 (Primary - d...)
- ▶ (Attached Containers)

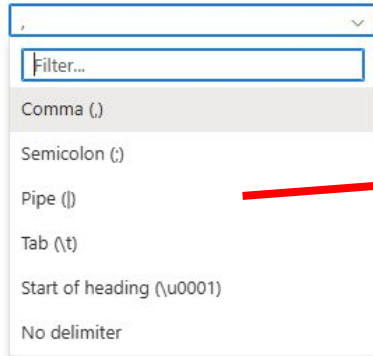
Workspace

- SQL database
- Lake database
- Data Explorer database (preview)
- Linked
 - Connect to external data
 - Integration dataset
- Browse gallery

Integration Dataset

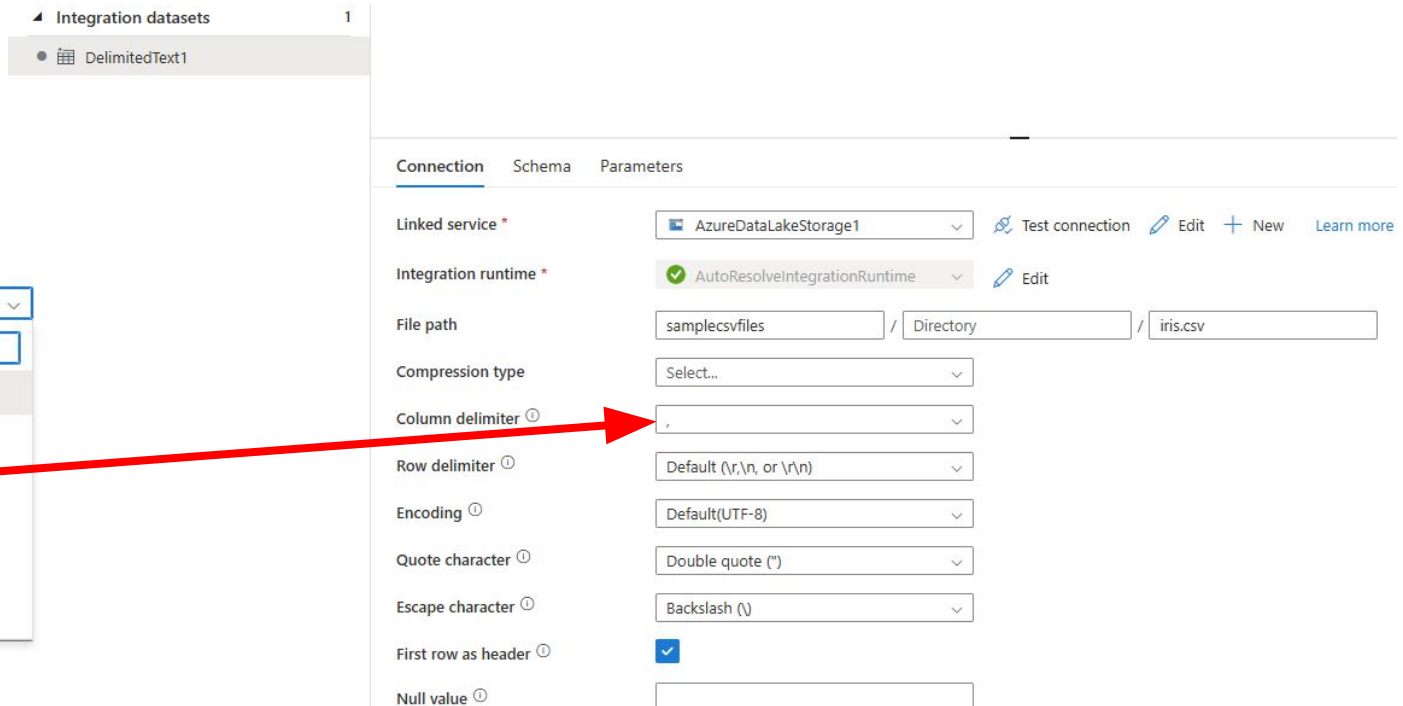
Flexible but manual: Fabric's options are much limited.

- Delimiter
- Header
- etc.



Filter...

- Comma (,)
- Semicolon (;)
- Pipe (|)
- Tab (\t)
- Start of heading (\u0001)
- No delimiter



Integration datasets 1

- DelimitedText1

Connection Schema Parameters

Linked service * AzureDataLakeStorage1 [Test connection](#) [Edit](#) [New](#) [Learn more](#)

Integration runtime * AutoResolveIntegrationRuntime [Edit](#)

File path samplescsvfiles / Directory / iris.csv

Compression type Select...

Column delimiter ,

Row delimiter Default (\r,\n, or \r\n)

Encoding Default(UTF-8)

Quote character Double quote (")

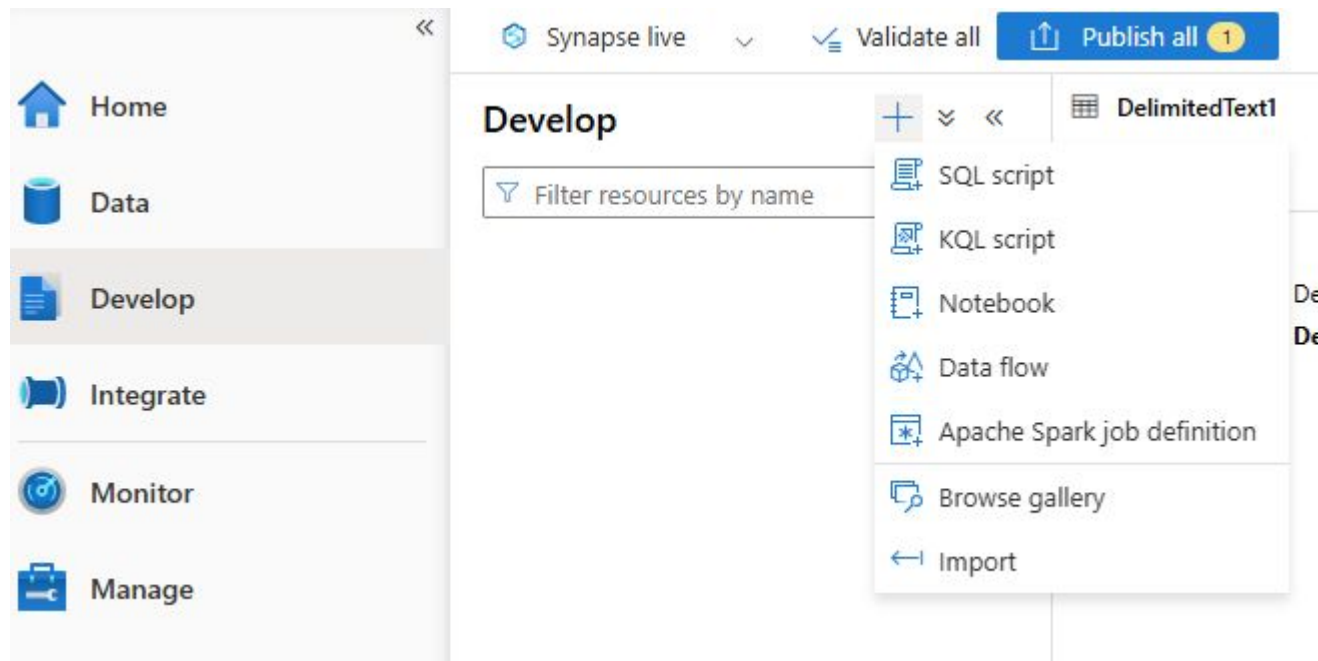
Escape character Backslash (\)

First row as header ☒

Null value

Development Tools

Many compute options, but not conveniently integrated



Integrate and Orchestration

Microsoft Azure | Synapse Analytics | jshan-ds684

We use optional cookies to provide a better experience. [Learn more](#)

Home
Data
Develop
Integrate
Monitor
Manage

Synapse live Validate all Publish all

Integrate

Filter resources by name

- Pipeline
- Link connection
- Copy Data tool
- Browse gallery
- Import from pipeline template

Integrate and Orchestration

The screenshot displays the Synapse Studio user interface, specifically the 'Integrate' tab. On the left is a vertical navigation pane with icons and labels for 'Home', 'Data', 'Develop', 'Integrate' (which is highlighted), 'Monitor', and 'Manage'. The main workspace is divided into three sections. The top section contains a header with 'Synapse live', a 'Validate all' button, and a 'Publish all' button with a yellow badge showing the number '1'. Below this is the 'Integrate' section, which includes a search bar labeled 'Filter resources by name' and a list of 'Pipelines' with a count of '1'. The 'Pipeline 1' item is selected and highlighted. To the right of the main workspace is a sidebar titled 'Activities' with a search bar labeled 'Search activities'. Below the search bar is a list of activity categories, each preceded by a right-pointing chevron: 'Synapse', 'Move and transform', 'Azure Data Explorer', 'Azure Function', 'Batch Service', 'Databricks', 'Data Lake Analytics', 'General', 'HDInsight', 'Iteration & conditionals', and 'Machine Learning'.

Home
Data
Develop
Integrate
Monitor
Manage

Synapse live Validate all Publish all 1

Integrate

Filter resources by name

Pipelines 1

Pipeline 1

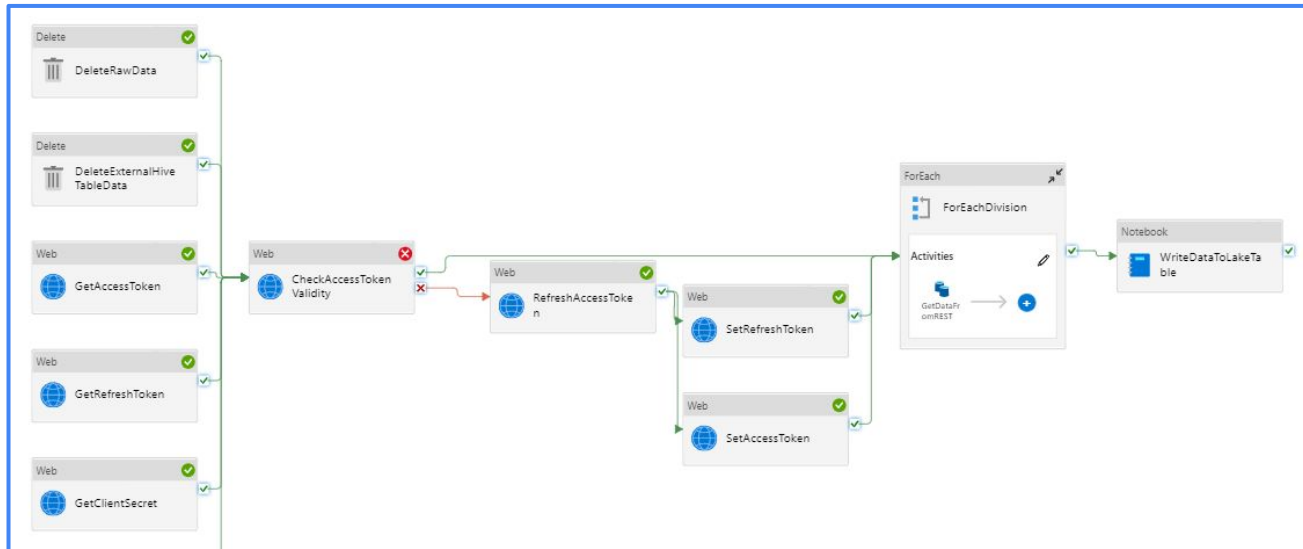
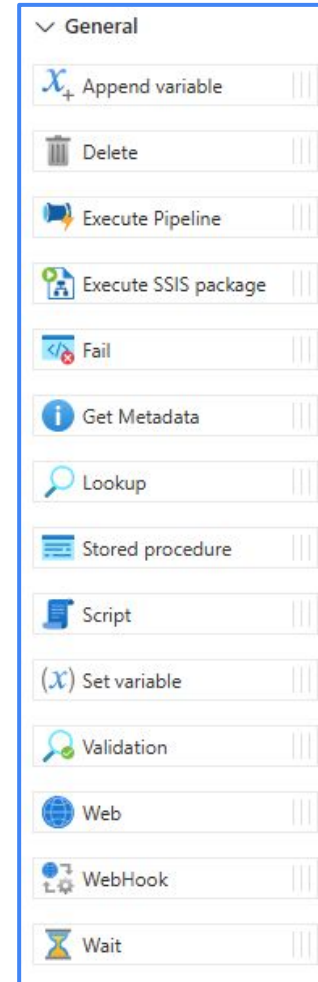
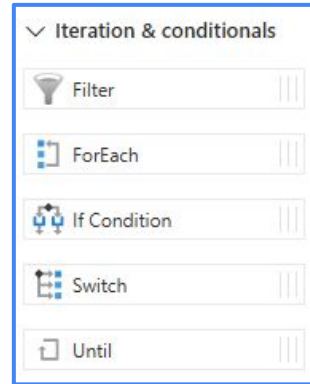
Activities

Search activities

- > Synapse
- > Move and transform
- > Azure Data Explorer
- > Azure Function
- > Batch Service
- > Databricks
- > Data Lake Analytics
- > General
- > HDInsight
- > Iteration & conditionals
- > Machine Learning

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