

# Data Engineering in the Cloud

## Building Data Mapping Flows with ADF

Xuemao Zhang  
East Stroudsburg University

January 18, 2025

# Outline

- Building Data Mapping Flows
- Prerequisites
- Lab

# Building Data Mapping Flows

- Graphical Definition: In Azure Data Factory (ADF), mapping data flows allow for the graphical definition of data transformations.
- Code-Free Transformation: These data flows enable data engineers to design data transformation algorithms without writing any code.
- Execution within Pipelines: Data flows are executed as activities within ADF pipelines, using scalable Apache Spark clusters.
  - ▶ These clusters are managed and operated by ADF.
- Operationalization: Data flow activities can be scheduled, monitored, and regulated using ADF's built-in features.

# Prerequisites

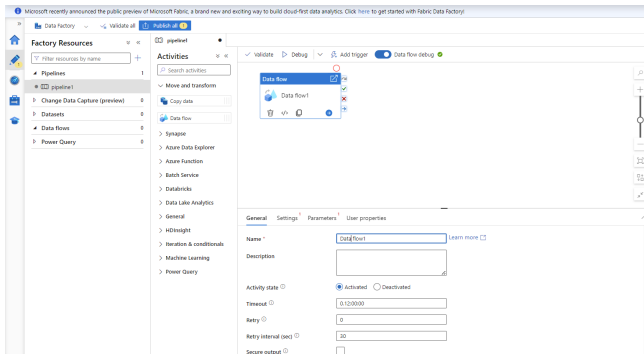
- Objective: You have the containers in the Azure Storage Gen-2 for input and output. We will fetch the length of the course names and have a mapping to its data based on that.
- The lab is similar as that from Lecture 17.

# Prerequisites

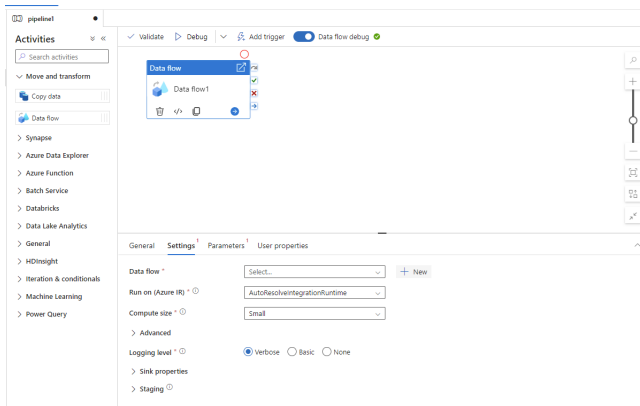
- Prerequisites (5 minutes):
  - ▶ Create a datalake (storage account with hierarchical namespace enabled)
    - ★ To lower the cost, you may choose Redundancy as LRS
  - ▶ Create two containers in the storage: `raw` and `output`
    - ★ Upload the data set `CoursesData.csv` to the container `raw`
  - ▶ Create an Azure Data Factory and launch Azure Data Factory Studio
- In the lab, we transform the data and then copy the transformed data to the container `output`

# Lab

- 1 In the Azure Data Factory you just created, go to the “Author tab”
- 2 Click on the “Pipelines” to create a pipeline
  - And drag and drop the “Data Flow” activity onto the pipeline canvas



- 3 Click on + New to add a data flow



# Lab

- 4 In Source settings, click + New to try to add the data set CoursesData.csv

The screenshot shows the Databricks Source settings configuration page. At the top, there's a header with 'pipeline1' and 'dataflow1'. Below this, there are tabs for 'Validate', 'Data flow debug', and 'Debug Settings'. The main area displays a source component named 'source1' with 'Columns: 0 total'. Below the source component is a dashed box labeled 'Add Source' with a downward arrow. The bottom section is titled 'Source settings' and contains several fields: 'Output stream name' (set to 'source1'), 'Description' (with a text area containing 'Add source dataset'), 'Source type' (with 'Dataset' and 'Inline' options), 'Dataset' (with a dropdown menu showing 'Select...' and a '+ New' button), and 'Options' (with checkboxes for 'Allow schema drift' and 'Infer drifted column types').

pipeline1 dataflow1

✓ Validate Data flow debug Debug Settings

source1  
Columns:  
0 total

Add Source

Source settings Source options Projection Optimize Inspect Data preview

Output stream name \* source1 [Learn more](#)

Description Add source dataset [Reset](#)

Source type \* Dataset Inline

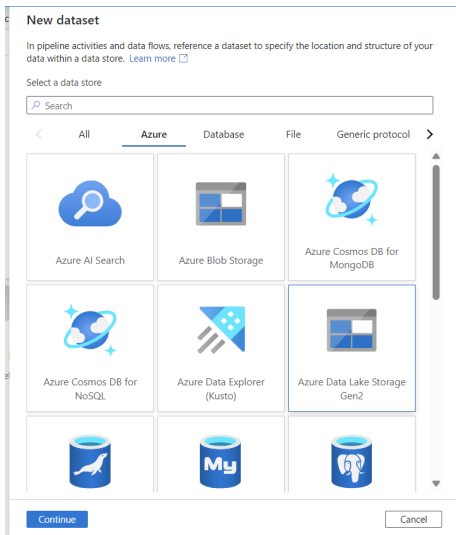
Dataset \* Select... [+ New](#)

Options ☒ Allow schema drift [?](#) ☐ Infer drifted column types [?](#)



# Lab

- 5 Select “Azure Data Lake Storage Gen-2” account and click on “Continue”











# Lab

- 6 Select “DelimitedText” and click on “Continue”

Select format

Choose the format type of your data

 Avro	 Binary	 DelimitedText
 Excel	 JSON	 ORC
 Parquet	 XML	

# Lab

- 7 Set the properties and click on “OK”
- A Linked service needs to be created

The screenshot shows a 'Set properties' dialog box with the following fields and options:

- Name:** A text input field containing 'CoursesCSV'.
- Linked service \*:** A dropdown menu showing 'AzureDataLakeStorage1' with an edit icon to its right.
- File path:** A path field with three segments: 'raw' (selected), 'Directory', and 'CoursesData.csv'. There are folder and file icons to the right of the last segment.
- First row as header:** A checkbox that is checked.
- Import schema:** Three radio buttons: 'From connection/store' (selected), 'From sample file', and 'None'.
- Advanced:** A link with a right-pointing chevron to expand more options.
- Buttons:** 'OK' (blue), 'Back', and 'Cancel' at the bottom.

- 8 Turn on the dataflow debug so we can Test connection

The screenshot displays the Databricks Dataflow interface. At the top, there's a 'Publish all' button and a 'dataflow1' tab. Below this, a 'Validate' button is active, and a 'Data flow debug' toggle is turned on. A 'Debug Settings' link is also visible. The main workspace shows a pipeline diagram with a single node labeled 'source1' with a 'Columns: 5 total' label. Below the diagram is an 'Add Source' button. The bottom section is the 'Source settings' configuration panel. It includes fields for 'Output stream name' (set to 'source1'), 'Description' (set to 'Import data from CoursesCSV'), and 'Source type' (set to 'Dataset'). The 'Dataset' type is selected, showing a 'CoursesCSV' dataset. The 'Options' section has 'Allow schema drift' checked and 'Infer drifted column types' unchecked. A 'Connection successful' status is shown with a green checkmark. A 'Test connection' button is available, along with 'Open' and '+ New' buttons.

Source settings Source options Projection Optimize Inspect Data preview

Output stream name \* source1 [Learn more](#)

Description Import data from CoursesCSV [Reset](#)

Source type \* Dataset Inline

Dataset \* CoursesCSV

Options ☒ Allow schema drift [?](#) ☐ Infer drifted column types [?](#)

Connection successful [Test connection](#) [Open](#) [+ New](#)

# Lab

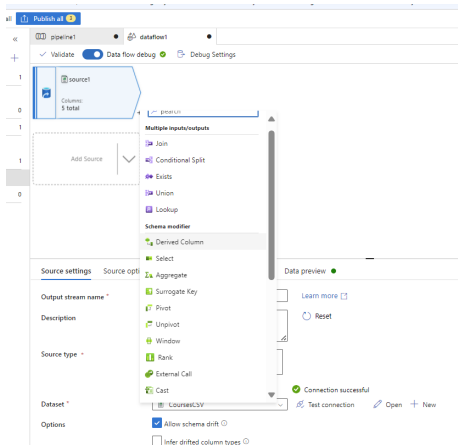
- 9 Let's click on Detect data type in the Projection tab
  - ▶ Nothing changes

The screenshot shows the Databricks Dataflow interface. At the top, there's a 'Publish all' button and a 'dataflow1' tab. Below this, there are tabs for 'Validate', 'Data flow debug', and 'Debug Settings'. The main area shows a source named 'source1' with 'Columns: 5 total'. Below the source, there's an 'Add Source' button. At the bottom, there's a tabbed interface with 'Source settings', 'Source options', 'Projection', 'Optimize', 'Inspect', and 'Data preview'. The 'Projection' tab is selected, showing options to 'Define default format', 'Detect data type', 'Import projection', and 'Reset schema'. Below these options is a table with columns 'Column name', 'Type', and 'Format'.

Column name	Type	Format
CourseTitle	string	Specify format
Description	string	Specify format
Availability	string	Specify format
Cost	string	Specify format
Date	string	Specify format

# Lab

- 10 Now you will fetch a few “desired columns”. To do so, click on “+” and select the “derived columns” option



# Lab

- 11 Give a name to the output and select the column below. Click on the “expression builder”

The screenshot displays the Databricks Dataflow interface. At the top, there's a 'Publish all' button and tabs for 'pipeline1' and 'dataflow1'. Below these are tabs for 'Validate', 'Data flow debug', and 'Debug Settings'. The main workspace shows a dataflow diagram with a 'source1' node (labeled 'Import data from CoursesCSV') and a 'derivedColumn1' node (labeled 'Columns: 5 total'). A dashed box labeled 'Add Source' is visible below the diagram. On the right side, there are zoom controls (search, +, -, and a vertical slider).

The 'Derived column's settings' panel is open at the bottom, showing the following configuration:

- Output stream name \***: `derivedColumn1` [Learn more](#)
- Description**: `Creating/updating the columns 'CourseTitle, Description, Availability, Cost, Date'` [Reset](#)
- Incoming stream \***: `source1`
- Columns \***: [+ Add](#) [Clone](#) [Delete](#) [Open expression builder](#)

Column	Expression
<input checked="" type="checkbox"/> CourseTitle	<input type="text" value="Enter expression..."/> <a href="#">Open expression builder</a>

- 12 In the expression builder, write the code
  - Then click on Save and finish

**Dataflow expression builder**

derivedColumn1

Derived Columns

+ Create new

LengthCourseTitle

Column name \*

LengthCourseTitle

Expression

length(CourseTitle)

+ - \* / || && ! ^ ~ <> [ ] > <

Expression elements

All

Functions

Input schema

Parameters

Cached lookup

Data flow library functions

Locals

Expression values

Filter by keyword

+ Create new

CourseTitle

Description

Availability

Cost

Date

abs(as\_numeric\_value)

acos(as\_numeric\_value)

add(as\_first\_expression, as\_second\_expression)

Data preview Refresh

Save and finish Cancel Clear contents



# Lab

- 13 You can select any expression that you want to fetch. I have chosen to understand the length of the course title

The screenshot displays the Databricks Dataflow interface. At the top, there's a 'Publish all' button and a 'dataflow1' tab. Below this, a visual pipeline shows a source node 'source1' (Import data from CoursesCSV) connected to a derived column node 'derivedColumn1' (Columns: 6 total). A dashed box labeled 'Add Source' is visible below the pipeline.

The 'Derived column's settings' panel is open, showing the following details:

- Output stream name:** derivedColumn1
- Description:** Creating/updating the columns 'CourseTitle', 'Description', 'Availability', 'Cost', 'Date', 'lengthCourseTitle'
- Incoming stream:** source1
- Columns:** A table listing the columns and their expressions.

Column	Expression
lengthCourseTitle	length(CourseTitle)

- 14 Preview the data, you can see that one column is added to the data

Published all 1

pipeline1 • dataflow1 •

✓ Validate Data flow debug Debug Settings

source1 Import data from CoursesCSV

derivedColumn1 Columns: 6 total

Add Source ✓

Derived column's settings Optimize Inspect **Data preview**

Number of rows INSERT 100 UPDATE 0 DELETE 0 UPSERT 0 LOOKUP 0 ERROR 0 TOTAL

Refresh Typecast Modify Map drifted Statistics Remove Export to CSV

CourseTitle	Description	Availability	Cost	Date	LengthCourseTitle
C# Basics	This is a course for ...	On demand	\$25/mo	2021-06-15T04:44:3...	9
Java Basics	This is a course for ...	On demand	\$25/mo	2021-06-15T04:44:2...	11
Mindfulness for Wel...	This is a course for ...	Upcoming	Free	2021-06-15T04:44:2...	46
Introduction to Pro...	This is a course for ...	On demand	Free	2021-06-15T04:44:3...	56
Existential Well-bein...	This is a course for ...	On demand	Free	2021-06-15T04:44:1...	74
secedu - Security E...	This is a course for ...	On demand	Free	2021-06-15T04:44:1...	55
Foundations of Co...	This is a course for ...	On demand	Free	2021-06-14T17:57:0...	44

# Lab

- 15 You can add as many columns as you want.
- 16 Once done, click on “+” and click on “sink”

The screenshot shows the Apache Airflow Dataflow interface. At the top, there's a header with 'pipeline1' and 'dataflow1'. Below it, there are tabs for 'Validate', 'Data flow debug', and 'Debug Settings'. The main area displays a pipeline diagram with a 'source1' node (Import data from SourcesCSV) and a 'derivedColumn1' node (Columns: 6 total). A 'Sink' node is visible at the bottom right. A context menu is open over the 'Sink' node, showing options like 'Unpivot', 'Window', 'Rank', 'External Call', 'Cast', 'Formatters', 'Flatten', 'Parse', 'Stringify', 'Row modifier', 'Filter', 'Sort', 'Alter Row', 'Assert', 'Flowlets', 'Flowlet', 'Destination', and 'Sink'. The 'Sink' option is highlighted. Below the pipeline diagram, there's a table showing the 'Derived column's settings' with columns for 'Number of rows', 'INSERT', 'UPDATE', and 'DELETE'. The table has 10 rows of data, including course titles like 'CP Basics', 'Java Basics', 'Mindfulness for Wel...', 'Introduction to Pro...', 'Existential Well-bein...', and 'car arts - Sanskrit F'. The table also shows the 'Cost' and 'Free' status for each row. The 'Sink' node is highlighted in the table, showing a cost of 'On demand' and a free status.

Number of rows	INSERT	UPDATE	DELETE
100	0	0	0
Refresh	Typecast	Modify	Map
%	CourseTitle	Cost	Description
+	CP Basics	This is a course for ...	
+	Java Basics	This is a course for ...	
+	Mindfulness for Wel...	This is a course for ...	
+	Introduction to Pro...	This is a course for ...	
+	Existential Well-bein...	This is a course for ...	
+	car arts - Sanskrit F	This is a course for ...	

Cost	Free	Date	Length	CourseTitle
125/mo	2021-06-15T04:44:3...	9		
125/mo	2021-06-15T04:44:2...	11		
Free	2021-06-15T04:44:2...	46		
Free	2021-06-15T04:44:3...	56		
On demand	2021-06-15T04:44:1...	74		
On demand	2021-06-15T04:44:1...	95		

# Lab

- 17 Again, we need to click + New to define the output file

The screenshot displays the Databricks Dataflow configuration interface. At the top, there are tabs for 'pipeline1' and 'dataflow1'. Below these, there are buttons for 'Validate', 'Data flow debug', and 'Debug Settings'. The main workspace area is currently empty, showing a dashed box with the text 'Add Source' and a downward arrow. At the bottom, there is a navigation bar with tabs: 'Sink', 'Settings', 'Errors', 'Mapping', 'Optimize', 'Inspect', and 'Data preview'. The 'Sink' tab is selected. The configuration for the sink is as follows:

- Output stream name \***: A text input field containing 'sink1'.
- Description**: A text input field containing 'Add sink dataset'.
- Incoming stream \***: A dropdown menu showing 'derivedColumn1'.
- Sink type \***: Three buttons labeled 'Dataset', 'Inline', and 'Cache'. The 'Dataset' button is selected.
- Dataset \***: A dropdown menu showing 'Select...'.
- Options**: Two checkboxes: 'Allow schema drift' (checked) and 'Validate schema' (unchecked).

There is a '+ New' button next to the 'Dataset' dropdown menu.

# Lab

- 16 Enter a name to it and click OK

## Set properties

Name

CoursesOutput

Linked service \*

AzureDataLakeStorage1



File path

output

/

Directory

/

File name



First row as header



Import schema

☒ From connection/store ☐ From sample file ☐ None

> Advanced

## 19 Test connection

**Publish all**

pipeline1 • dataflow1

✓ Validate ☒ Data flow debug Debug Settings

Add Source

**Sink** Settings Errors Mapping Optimize Inspect Data preview

Output stream name \*  [Learn more](#)

Description  Reset

Incoming stream \*

Sink type \*

☒ Dataset
 ☐ Inline
 ☐ Cache

Dataset \*  Test connection Open New

Skip line count

Options

☒ Allow schema drift

☐ Validate schema

Connection successful

# Lab

- 20 In the settings, select the option as default

The screenshot displays the Databricks Dataflow interface. At the top, there's a 'Publish all' button and a 'pipeline1' tab. Below this, a 'Data flow debug' section shows a visual pipeline: 'source1' (Import data from CoursesCSV) connects to 'derivedColumn1' (Creating/updating the columns 'CourseTitle', 'Description', 'Availability', 'Cost', 'Date', 'LengthCourseTitle'), which then connects to 'sink1' (Columns: 6 total). The 'sink1' component is highlighted with a blue border.

Below the pipeline diagram is an 'Add Source' button. The main interface is divided into tabs: 'Sink', 'Settings' (selected), 'Errors', 'Mapping', 'Optimize', 'Inspect', and 'Data preview'. The 'Settings' tab contains the following configuration options:

- Clear the folder**: ☐
- File name option \***:
- Quote All**: ☐
- Headers**:  ANY
- Umask**:
  - Owner: ☐ R ☐ W ☐ X
  - Group: ☐ R ☒ W ☐ X
  - Others: ☐ R ☒ W ☐ X
  - Octal**:
- Pre/post commands**:
  - File pre command**:  + 🗑️
  - File post command**:  + 🗑️

- 21 Again, you may preview the output data

Pipeline1

Validate ☒ Data flow debug ☒ Debug Settings

source1 Import data from CoursesCSV + derivedColumn1 Creating/updating the columns CourseTitle, Description, Availability, Cost, Date, LengthCourseTitle + sink1 Columns: 6 total

Add Source

Sink Settings Errors Mapping Optimize Inspect **Data preview**

Number of rows + INSERT N/A UPDATE N/A DELETE N/A UPSERT N/A LOOKUP N/A ERROR N/A TO

Refresh Statistics Export to CSV

CourseTitle	abc ↑↓	Description	abc ↑↓	Availability	abc ↑↓	Cost	abc ↑↓	Date	abc ↑↓	LengthCourseTitle	abc ↑
C# Basics		This is a course for p...		On demand		\$25/mo		2021-06-15T04:44:3...		9	
Java Basics		This is a course for p...		On demand		\$25/mo		2021-06-15T04:44:2...		11	
Mindfulness for Well...		This is a course for p...		Upcoming		Free		2021-06-15T04:44:2...		46	
Introduction to Prob...		This is a course for p...		On demand		Free		2021-06-15T04:44:3...		56	
Existential Well-bein...		This is a course for p...		On demand		Free		2021-06-15T04:44:1...		74	
sec.edu - Security En...		This is a course for p...		On demand		Free		2021-06-15T04:44:1...		55	
Foundations of Com...		This is a course for p...		On demand		Free		2021-06-14T17:57:0...		44	
The Ancient Greek H...		This is a course for p...		On demand		Free		2021-06-14T17:55:1...		34	
The Science of Gastr...		This is a course for p...		Upcoming		Free		2021-06-14T17:55:1...		25	
Chinese Thought: A...		This is a course for p...		Upcoming		Free		2021-06-14T17:56:5...		61	
4G Network Essentials		This is a course for p...		On demand		Free		2021-06-14T17:55:1...		21	
Computation Struct...		This is a course for p...		On demand		Free		2021-06-14T17:55:1...		49	



# Lab

- 22 Now click on “Publish all” button
- 23 Now go to the “Pipelines” tab and you should be able to see the complete pipeline

The screenshot displays the Microsoft Fabric Data Factory user interface. At the top, a banner mentions the public preview of Microsoft Fabric. The left sidebar, titled 'Factory Resources', shows a tree view with 'Pipelines' selected, containing 'pipeline1'. Below it are 'Datasets' (CoursesCSV, CoursesOutput) and 'Data flows' (dataflow1). The main workspace is divided into three panes. The 'Activities' pane on the left lists 'Move and transform' activities: 'Copy data' and 'Data flow'. The central canvas shows a 'Data flow' activity named 'Data flow1'. The right pane, titled 'Settings', has tabs for 'General', 'Settings', 'Parameters', and 'User properties'. The 'Settings' tab is active, showing configuration for 'Data flow1', including 'Run on (Azure IR)' set to 'AutoResolveIntegrationRuntime' and 'Compute size' set to 'Small'. The 'Logging level' is set to 'Verbose'.

- 24 Click on “add trigger” and “trigger now” button to trigger the pipeline. This will take some time to trigger.

All pipeline runs > pipeline1 - Activity runs

Rerun Cancel Refresh Update pipeline List Gantt

Data flow

Data flow1

Activity runs

Pipeline run ID 15266a0c-18aa-46fd-ae9d-8659d4e63f50

All status Monitor in Azure Metrics Export to CSV

Showing 1 - 1 items

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Activity run ID
Data flow1	In progress	Data flow	7/25/2024, 11:19:11 AM	4s			9a13e34b-faa7-494a-b504

- 25 Click on the Monitor tab

The screenshot shows the 'Pipeline runs' interface. The left sidebar contains navigation options: Dashboards, Runs, Pipeline runs (selected), Trigger runs, Change Data Capture (prev...), Runtimes & sessions, Integration runtimes, Data flow debug, Notifications, and Alerts & metrics. The main panel has tabs for 'Triggered' and 'Debug'. Below the tabs are filters: 'Filter by run ID or name', 'Local time: Last 24 hours', 'Pipeline name: pipeline1', 'Status: All', and 'Runs: Latest runs'. There are also buttons for 'Copy filters', 'Export to CSV', and 'Add filter'. The table shows 1 item, with the following data:

Pipeline name %s	Run start %s	Run end %s	Duration	Triggered by	Status %s	Run	Parameters
pipeline1	7/25/2024, 11:19:09 AM	--	1m 33s	Manual trigger	In progress	Original	

The screenshot shows the 'Pipeline runs' interface after the run has completed. The table now shows the run with a status of 'Succeeded'.

Pipeline name %s	Run start %s	Run end %s	Duration	Triggered by	Status %s	Run	Parameters
pipeline1	7/25/2024, 11:19:09 AM	7/25/2024, 11:22:23 AM	3m 15s	Manual trigger	Succeeded	Original	

- 26 You can go to the output container to preview the output data.

> output >

Upload + Add Directory ...

Authentication method: Access key (Switch to Microsoft Entra user account)

Location: output

Search blobs by prefix (case-...)

Show deleted objects

Name

☐ \_SUCCESS ...

☐ part-00000-279b7edd-2ccc-4... ..

☒ part-00000-d5aebd62-731f-42... ..

part-00000-d5aebd62-731f-4285-845f-266d123318b4-c000.csv ...

Save Discard Download Refresh Delete

Overview Versions Edit Generate SAS

CourseTitle	Description	Availability	Cost	Date	Length	CourseTitle
C# Basics	This is a course for professionals	On demand	\$25/mo	2021-06-15T04:44:38.223Z	9	
Java Basics	This is a course for professionals	On demand	\$25/mo	2021-06-15T04:44:21.547Z	11	
Mindfulness for Wellbeing and Peak Performance	This is a course for professionals	Upcoming	Free	2021-06-15T04:44:31.702Z	46	
Introduction to Probability - The Science of Uncertainty	This is a course for professionals	On demand	Free	2021-06-15T04:44:31.332Z	56	
Existential Well-being Counseling: A Person-centered Experiential Approach	This is a course for professionals	On demand	Free	2021-06-15T04:44:12.533Z	74	
secedu - Security Engineering - Applied Cyber Security	This is a course for professionals	On demand	Free	2021-06-15T04:44:16.036Z	55	
Foundations of Computer Science for Teachers	This is a course for professionals	On demand	Free	2021-06-14T17:57:02.240Z	44	
The Ancient Greek Hero in 24 Hours	This is a course for professionals	On demand	Free	2021-06-14T17:55:17.612Z	34	
The Science of Gastronomy	This is a course for professionals	Upcoming	Free	2021-06-14T17:55:18.577Z	25	
Chinese Thought: Ancient Wisdom Meets Modern Science - Part 2	This is a course for professionals	Upcoming	Free	2021-06-14T17:56:51.946Z	61	
4G Network Essentials	This is a course for professionals	On demand	Free	2021-06-14T17:55:18.393Z	21	
Computation Structures - Part 1: Digital Circuits	This is a course for professionals	On demand	Free	2021-06-14T17:55:18.476Z	49	
Ecodeign for Cities and Suburbs	This is a course for professionals	In Session	Free	2021-06-14T14:24:48.367Z	32	
Build a Modern Computer from First Principles: From Nand to Tetriz (Project-Centered Course)	This is a course for professionals	Upcoming	Free	2021-06-14T14:24:45.734Z	92	
Reconciliation Through Indigenous Education	This is a course for professionals	Upcoming	Free	2021-06-14T14:24:45.429Z	43	
Lessons from Ebola: Preventing the Next Pandemic	This is a course for professionals	On demand	Free	2021-06-14T14:24:44.499Z	48	
Quantum Mechanics for Everyone	This is a course for professionals	In Session	Free	2021-06-14T14:24:44.209Z	30	
Justice	This is a course for professionals	In Session	Free	2021-06-14T14:24:43.264Z	7	
Bioethics: The Law, Medicine, and Ethics of Reproductive Technologies and Genetics	This is a course for professionals	In Session	Free	2021-06-14T14:24:44.189Z	82	
Communicating Corporate Social Responsibility (CSR)	This is a course for professionals	On demand	Free	2021-06-14T14:24:43.264Z	51	
Astrophysics: The Violent Universe	This is a course for professionals	In Session	Free	2021-06-14T14:22:10.094Z	34	
Humanity and Nature in Chinese Thought   中国哲学思想中的人类与自然观	This is a course for professionals	On demand	Free	2021-06-14T14:22:11.789Z	55	
Introduction to Philosophy: God, Knowledge and Consciousness	This is a course for professionals	On demand	Free	2021-06-14T14:22:10.909Z	60	

# License



This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/).