

Microsoft - DAT278x

From Graph to Knowledge Graph

– Algorithms and Applications

Challenge Lab Environment Setup Guide

The challenge labs for DAT278x will run on Microsoft Azure Data Lake Analytics environment. You can sign-up with an Azure free account, upload the data and finish the challenge labs assignment.

Lab assignments are running on U-SQL – a data processing language that unifies a declarative SQL-like syntax with C# programming. U-SQL can be used to process both structured and unstructured data in big data environments.

Prior basic knowledge and experience with SQL, U-SQL, Azure Data Lake Storage (ADLS), and Azure Data Lake Analytics (ADLA) would be required to complete the challenge labs.

You can refer to below resources to fill the knowledge gaps if any:

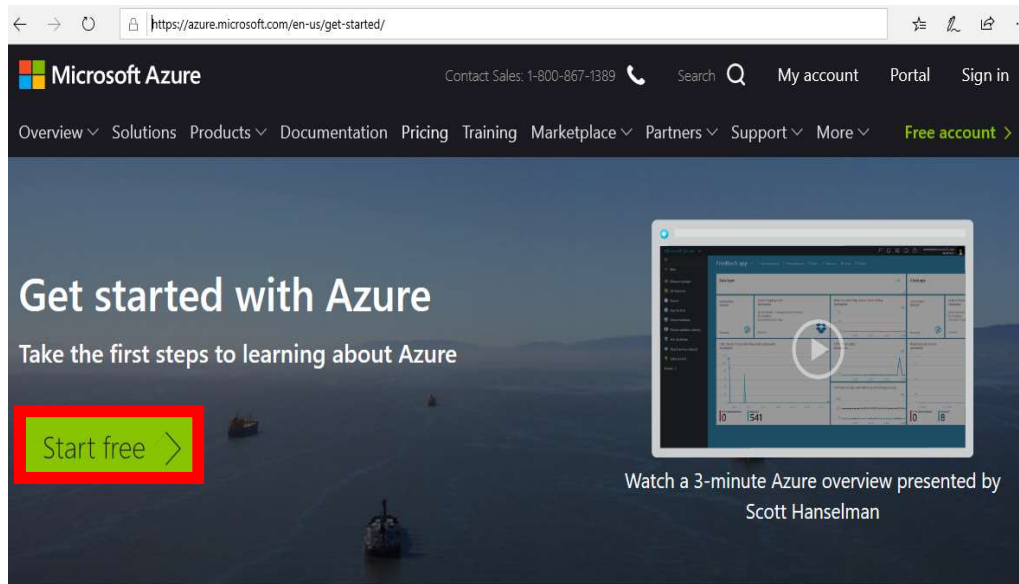
- EdX course on [Processing Big Data with Azure Data Lake Analytics](#)
- U-SQL resource site: usql.io

In the remaining of this document, we include step-by-step instructions to:

1. create an Azure free account;
2. create an Azure Data Lake Analytics (ADLA) resource and an Azure Data Lake Storage (ADLS) resource;
3. create folder structure in Azure Data Lake Storage (ADLS) to prepare for challenge labs runs;
4. upload data files to Azure Data Lake Storage (ADLS);
5. run a smoke test USQL script to verify the environment is setup properly.

1. Create a free Azure account:

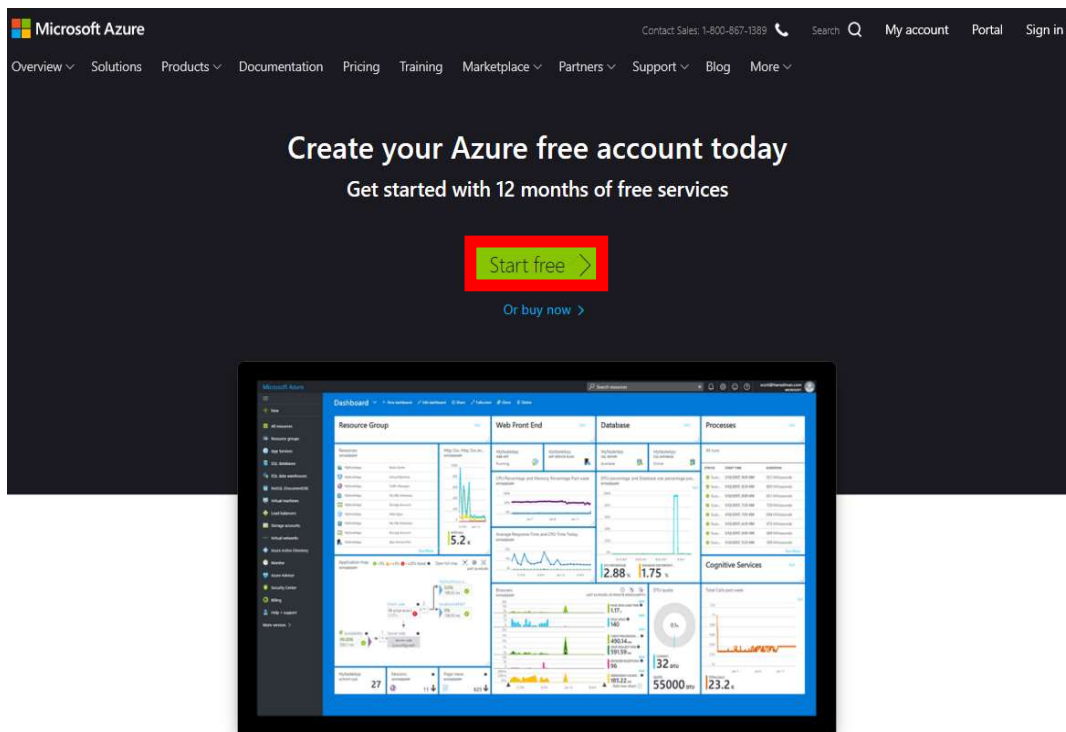
1.1 Go to this page: <https://azure.microsoft.com/en-us/get-started/> ; click on “start-free”.



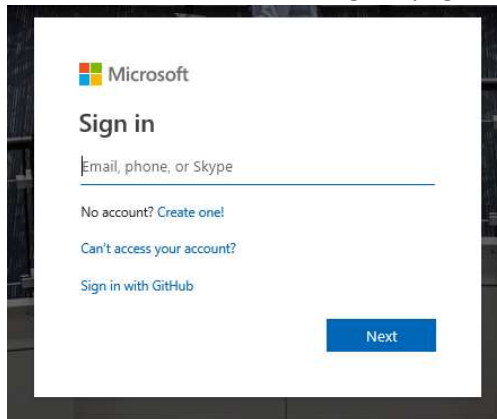
Deploy your first solution in 10 minutes or less

Try out these short tutorials on how to use Azure and start building projects right away.

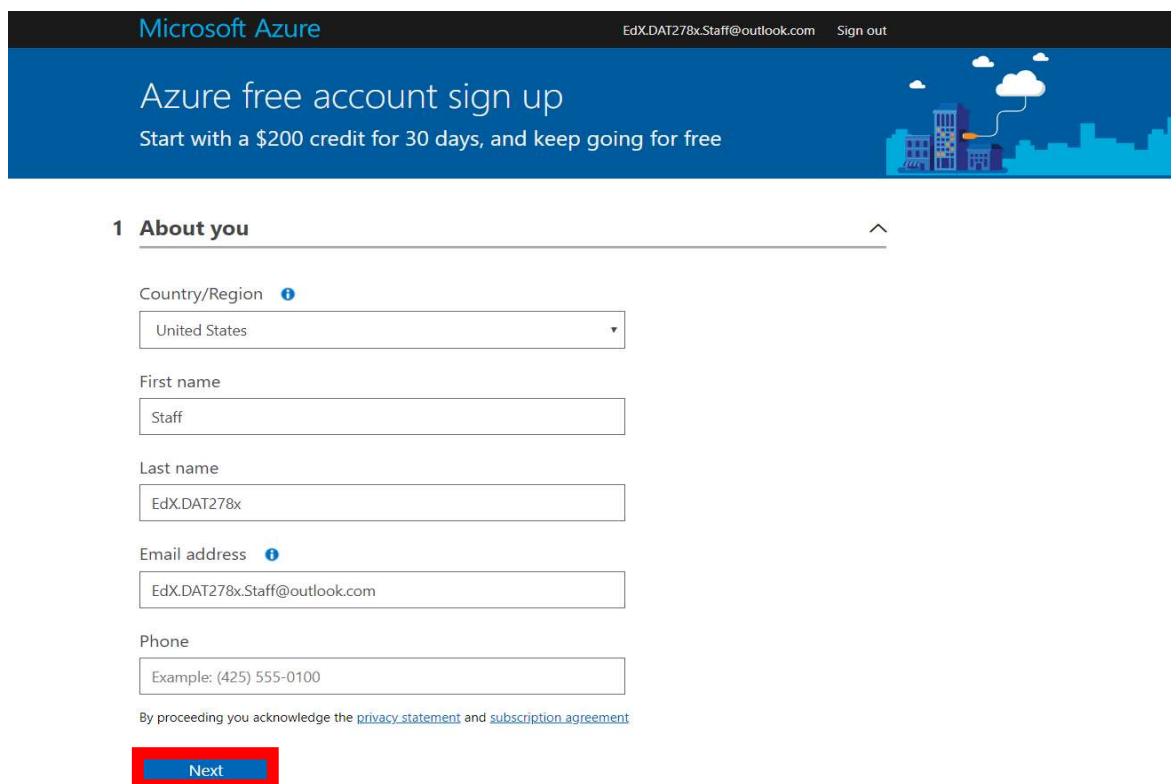
1.2 Land on below page, click on “start free”



1.3 Land on the Microsoft Sign in page, sign-in with your Microsoft Account.



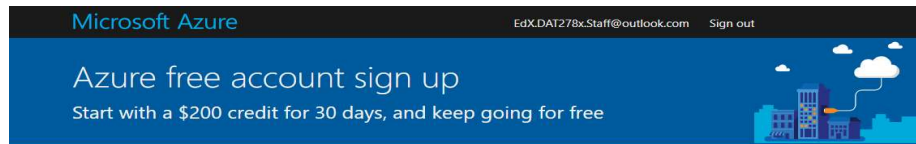
1.3 Land on Microsoft Azure sign-up page, “about you” section, fill in your information and click “Next”.



1.4 Provide phone information and credit card information for identity verification purpose.
(PLEASE NOTE: Your credit card won't be charged unless you upgrade your subscription. For EdX DAT278x course, the “free trial” credit would be enough to finish all labs.)

For students only: if you are a student with a valid school email address, you can access Azure without providing credit card information. For more info, please refer below links.

<https://azure.microsoft.com/en-us/free/free-account-students-faq/>



1 About you ⌵

2 Identity verification by phone ⌵

A text or phone call helps us make sure this is you.

Country code

United States (+1) ⌵

Phone number

Text me

Call me

We delivered a code to your phone.

Verification code

Verify code

I did not receive a code

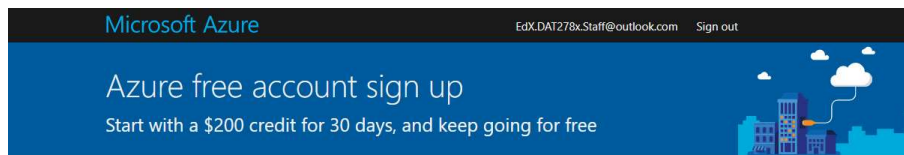
3 Identity verification by card ⌵

We ask for your credit card number to verify your identity and to keep out spam and bots.

You won't be charged unless you upgrade.



1.5 To accept the "Agreement" and click the "sign up" to proceed.



1 About you ⌵

2 Identity verification by phone ⌵

3 Identity verification by card ⌵

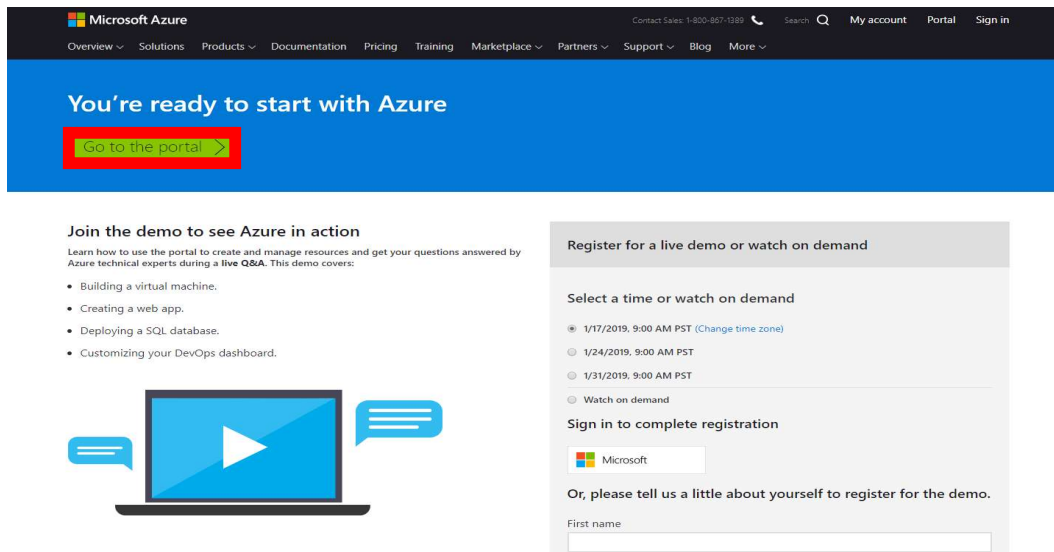
4 Agreement ⌵

☒ I agree to the [subscription agreement](#), [offer details](#), and [privacy statement](#)

I will receive information, tips, and offers from Microsoft or selected partners about Azure, including Azure Newsletter, Pricing updates, and other Microsoft products and services.

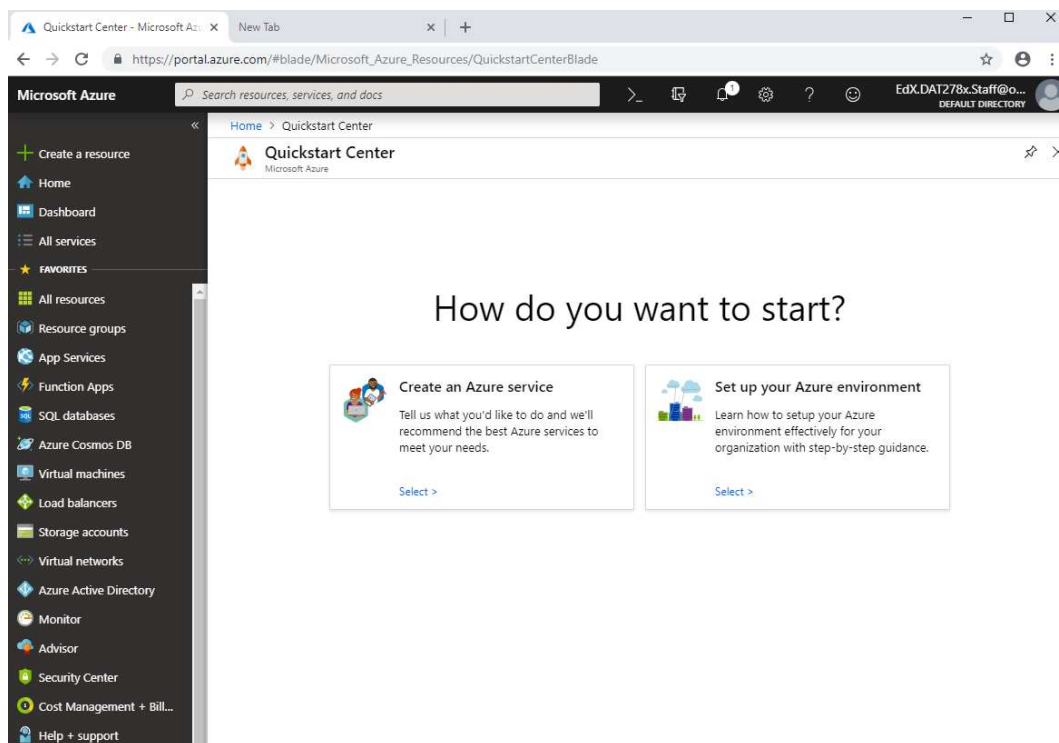
Sign up

1.6 You can see below page after successfully sign-up with Azure account, click on “Go to the portal”

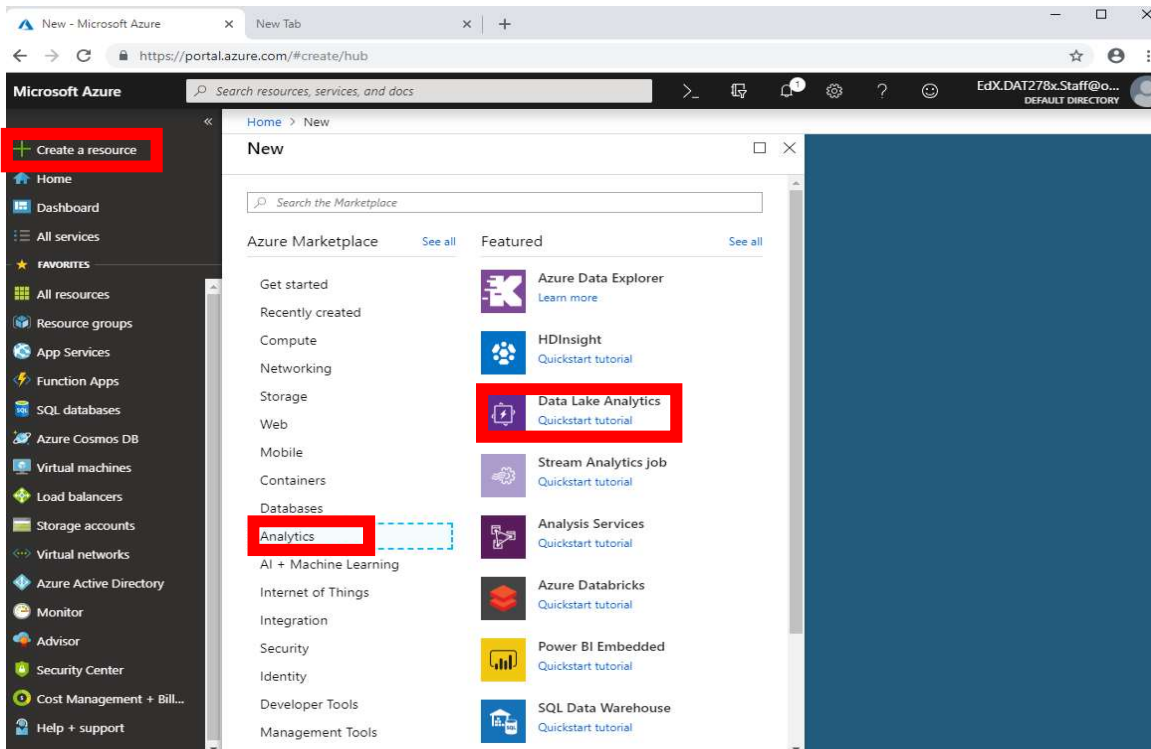


2. On Azure Portal, to create an Azure Data Lake Analytics (ADLA) resource and an Azure Data Lake Storage (ADLS) resource

2.1 Sign on to the Azure Portal to see below page:



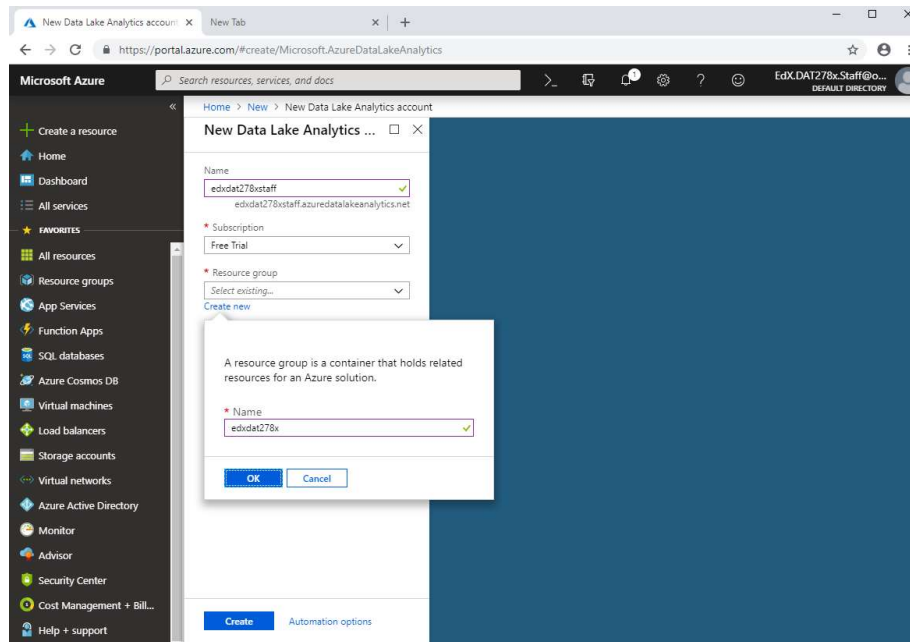
2.2 Click on “Create a resource” -> “Analytics” -> “Data Lake Analytics”



2.3 Fill in values for following items:

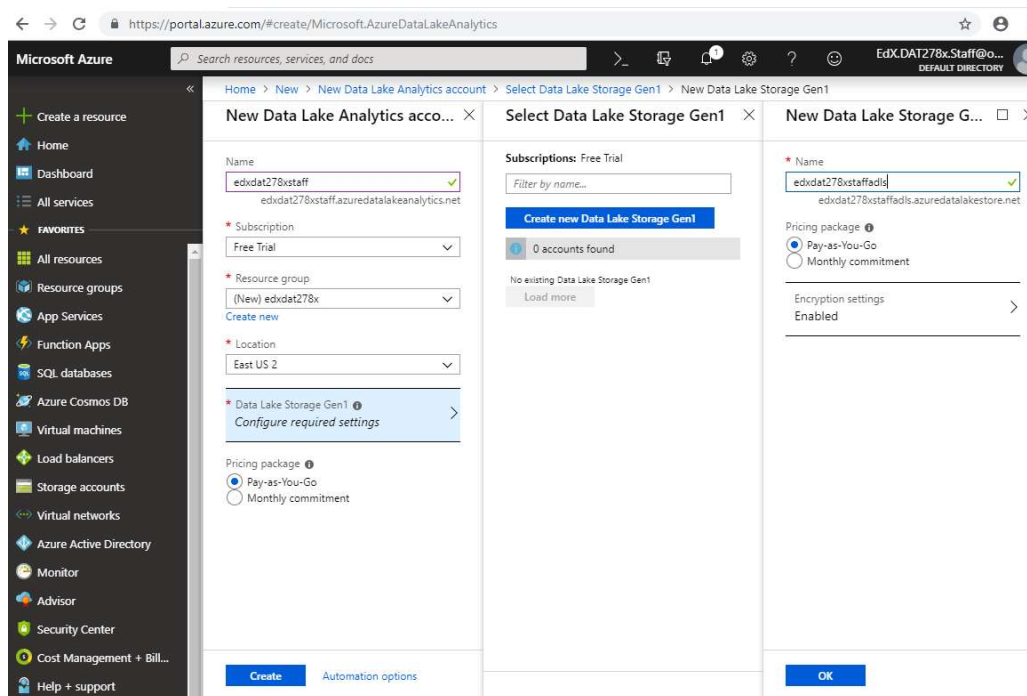
- Name:** Name your Data Lake Analytics account (Only lower-case letters and numbers allowed).
- Subscription:** Choose the Azure subscription (select "free trial" for DAT278x course unless you have other subscriptions) used for the Analytics account.
- Resource Group:** Select an existing Azure Resource Group or create a new one (if this is the first time you use Azure or you want a separate Resource Group for DAT278x course).
- Location:** Select an Azure data center for the Data Lake Analytics account (using default value is fine).
- Data Lake Store:** Follow the instruction to create a new Data Lake Store account, or select an existing one.
- Optionally,** select a **pricing tier** for your Data Lake Analytics account.

2.3.1 Pick a new name for Resource Group

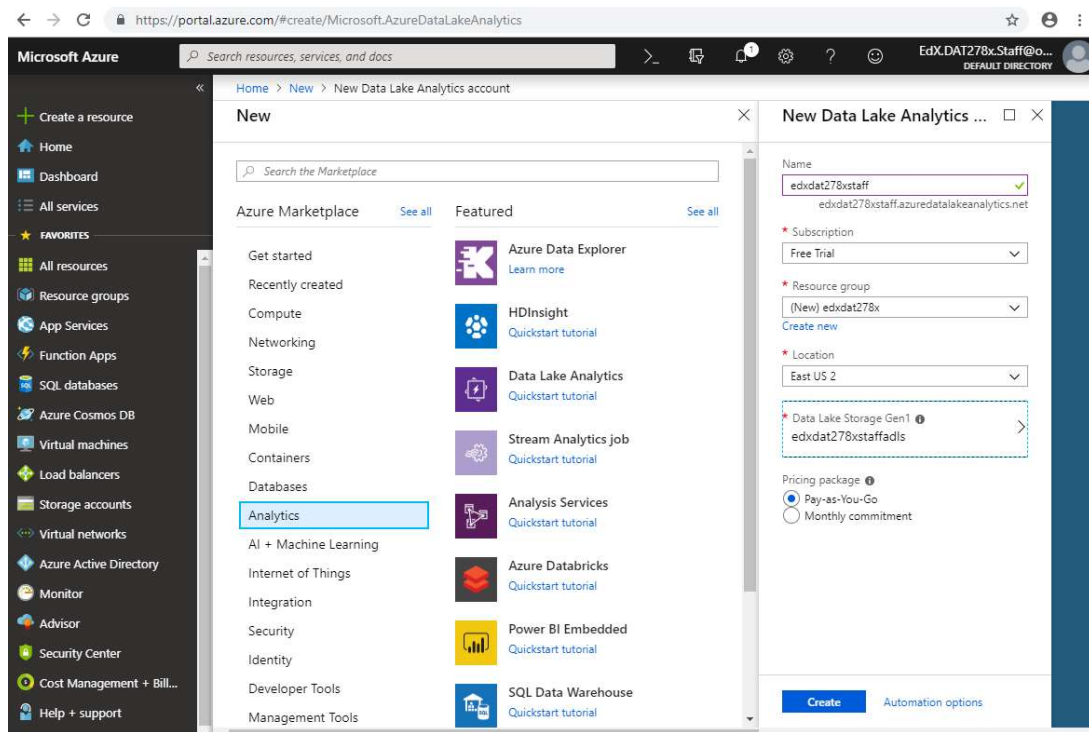


2.3.2 Create a new Azure Data Lake Storage (ADLS) account to associate with the Azure Data Lake Analytics (ADLA) account.

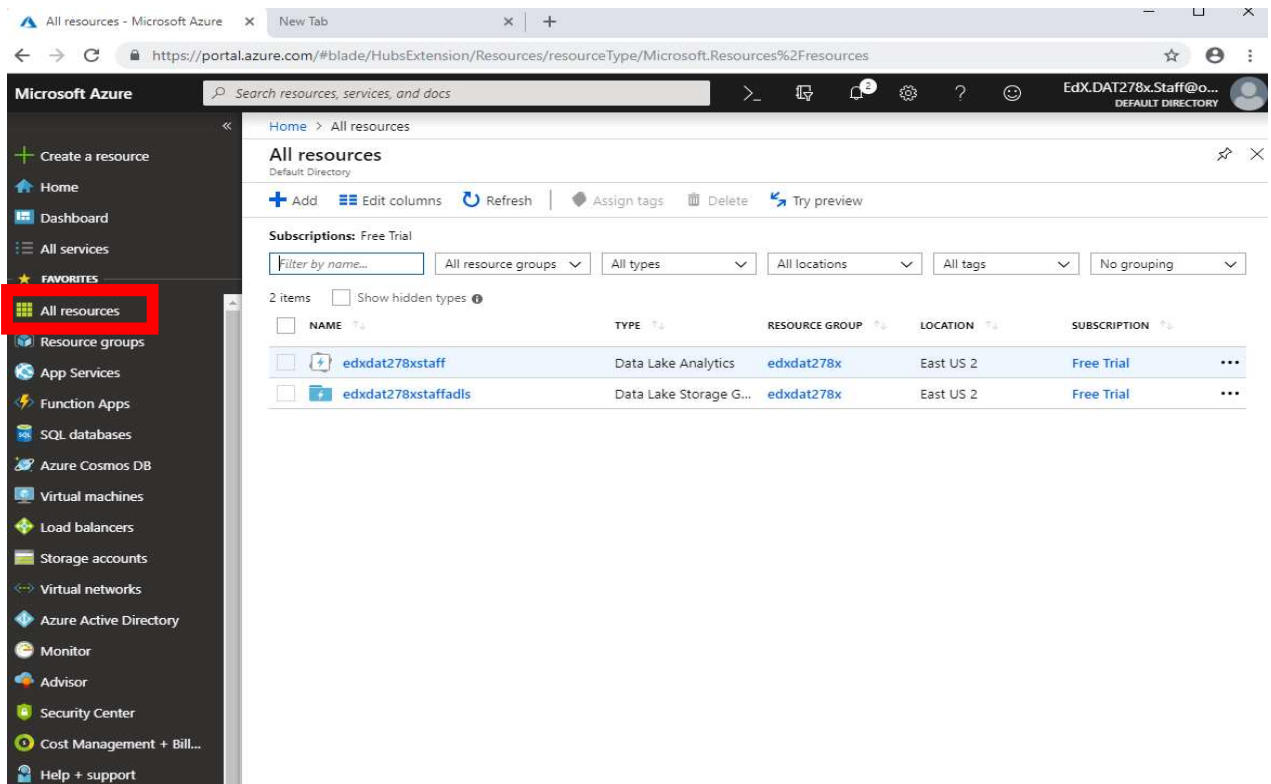
Use the default “Pay-as-You-Go” option for “free trial” subscription, pick a different package if you have other subscriptions.)



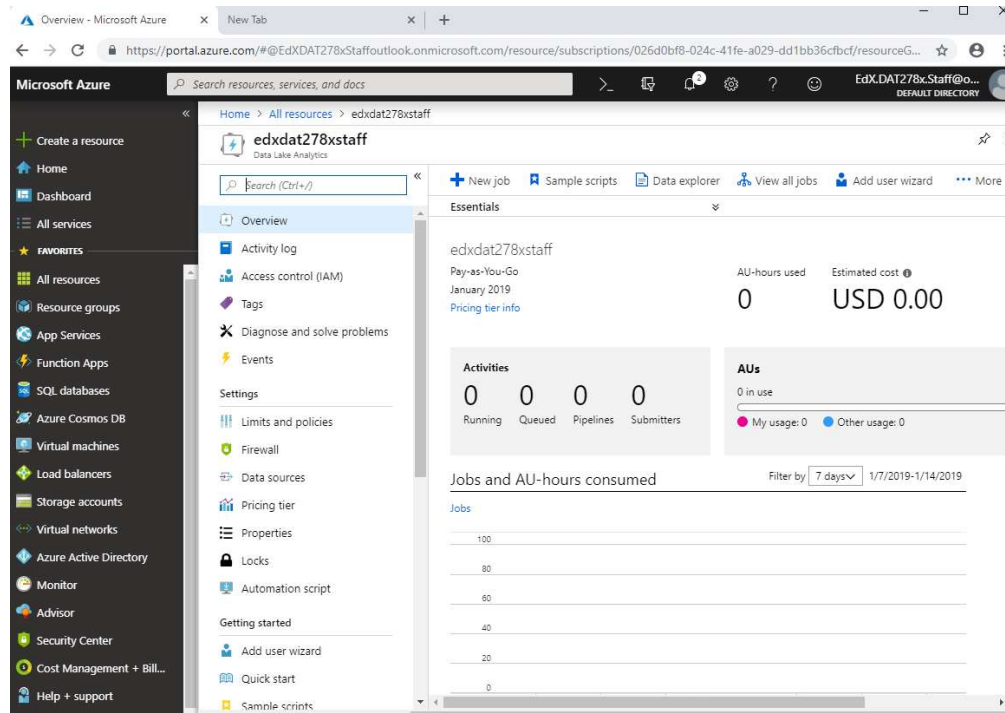
2.3.3 Validate all the fill-in values are correct and then click “Create”.



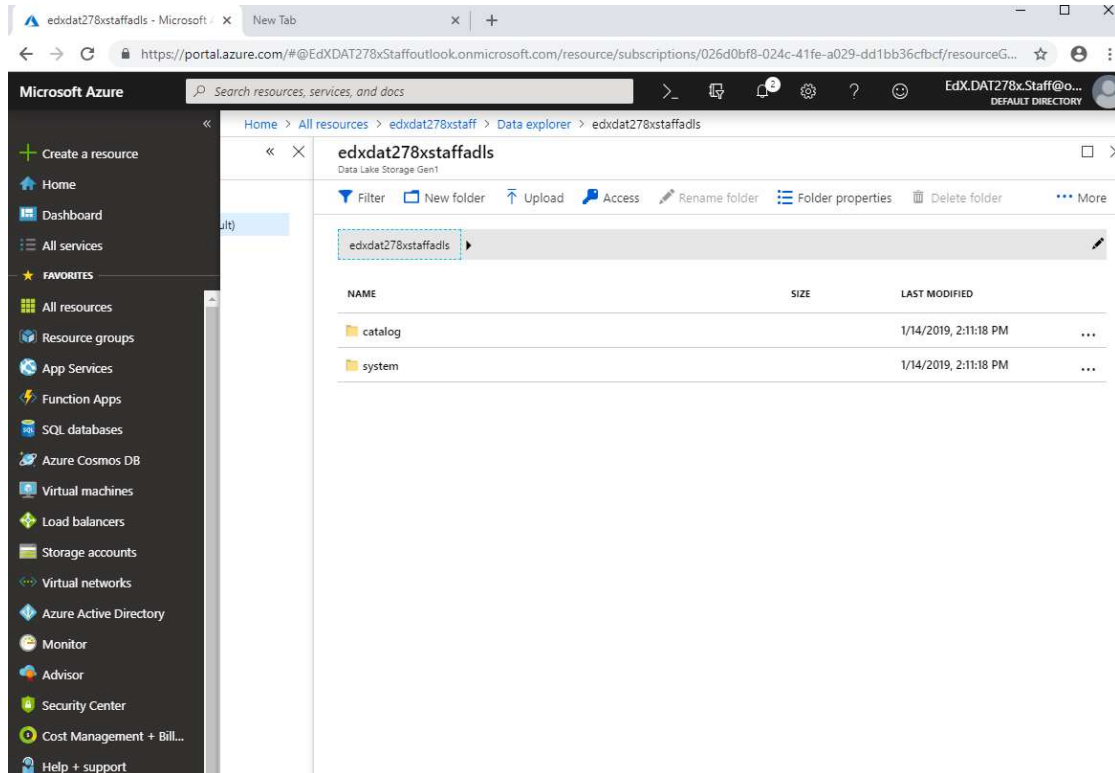
2.4 View the created Data Lake Analytics and Data Lake Store accounts in “All resources”.



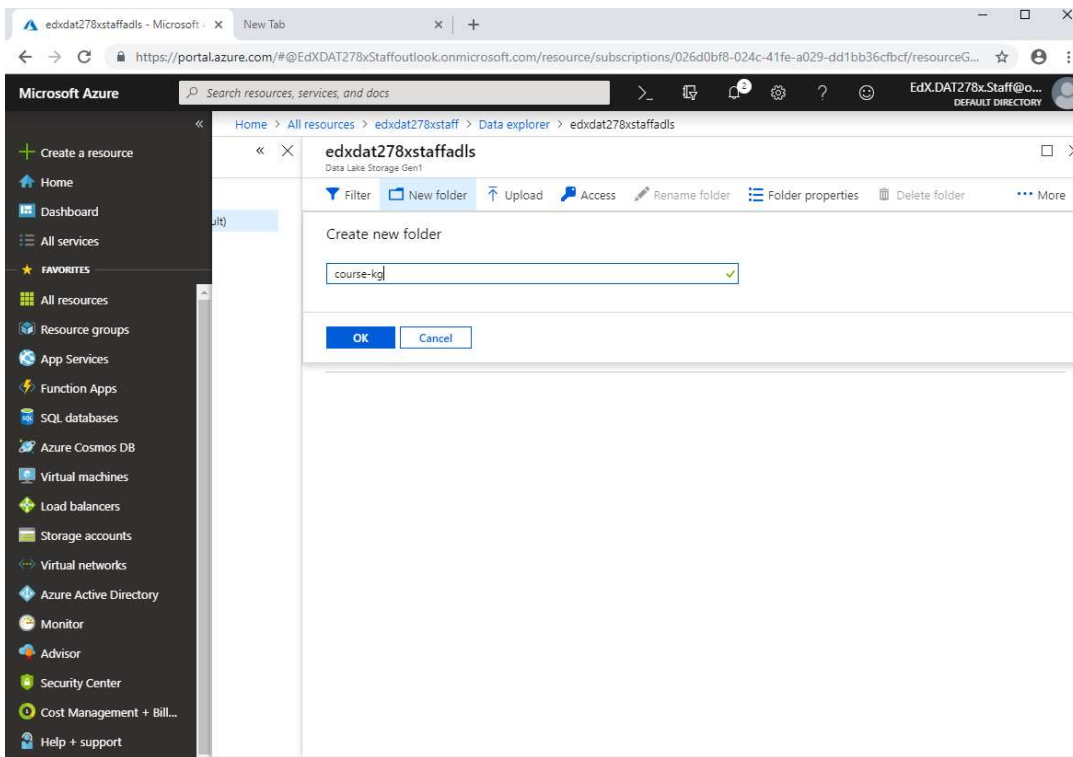
3. Create folder structure in Azure Data Lake Store (ADLS) to prepare for challenge labs runs.
 - 3.1 Go to Azure Data Lake Analytics account and click **Data explorer**.



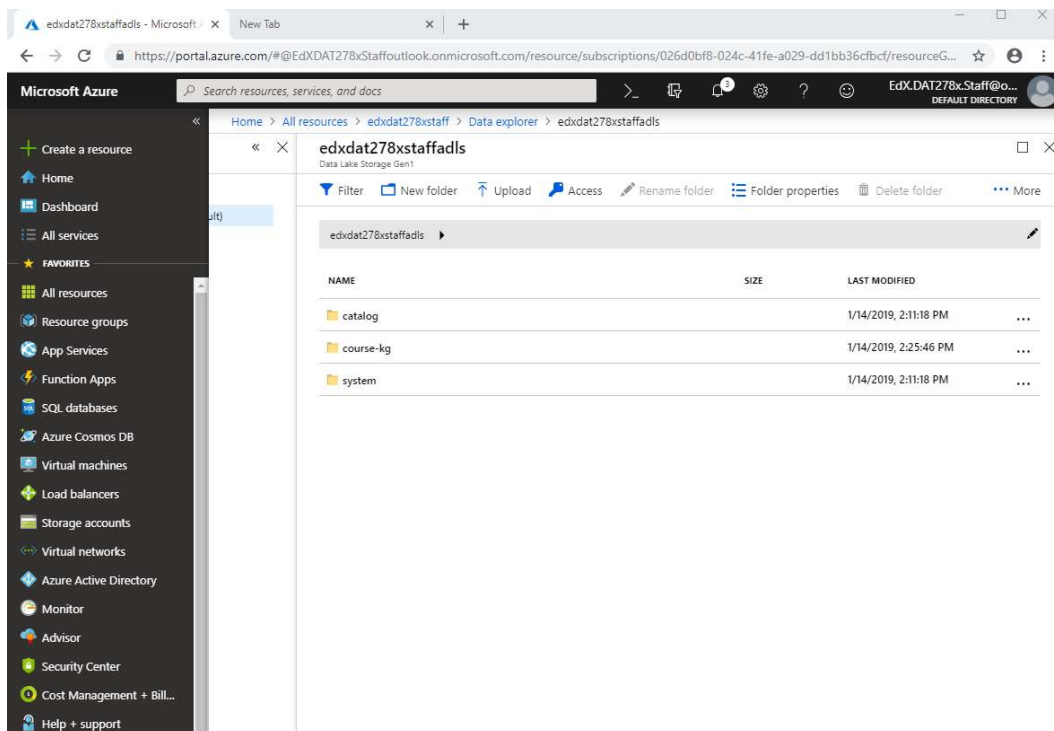
- 3.2 Click **New folder** to create a new folder under Azure Data Lake Store account.

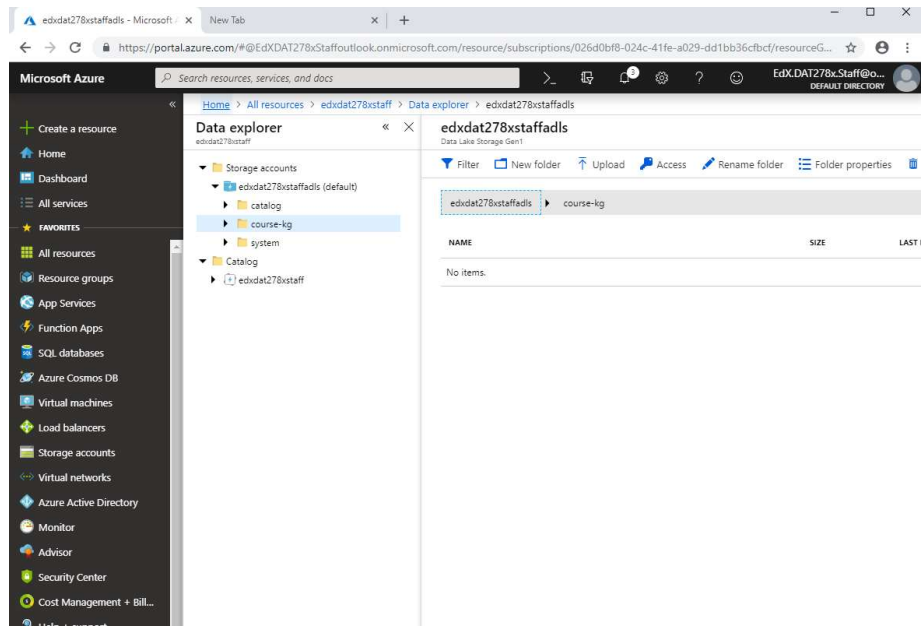


3.3 Use “course-kg” as the folder name and click **OK**.

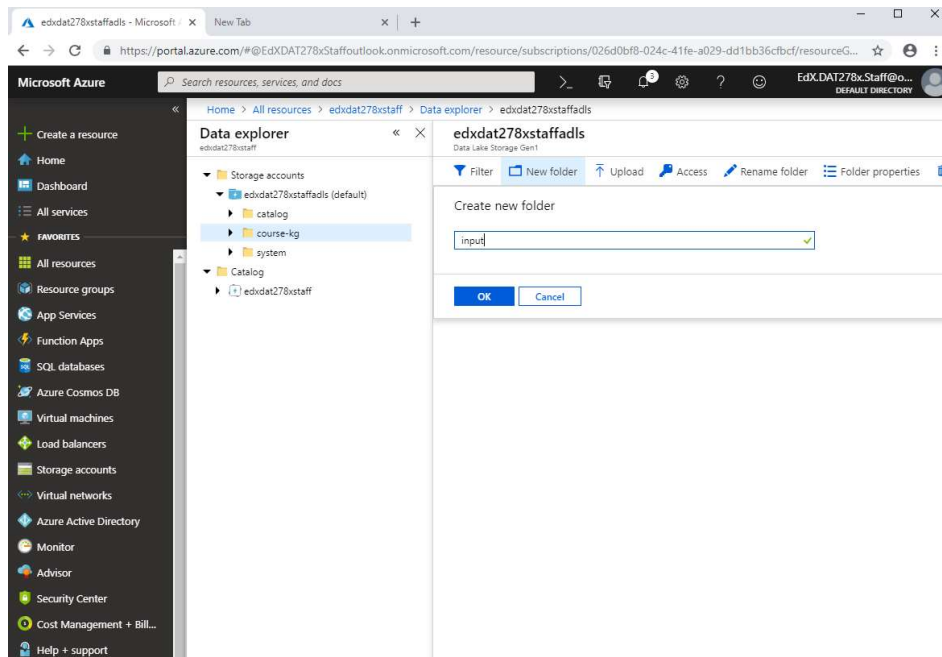


3.4 Click **course-kg** to go into this folder

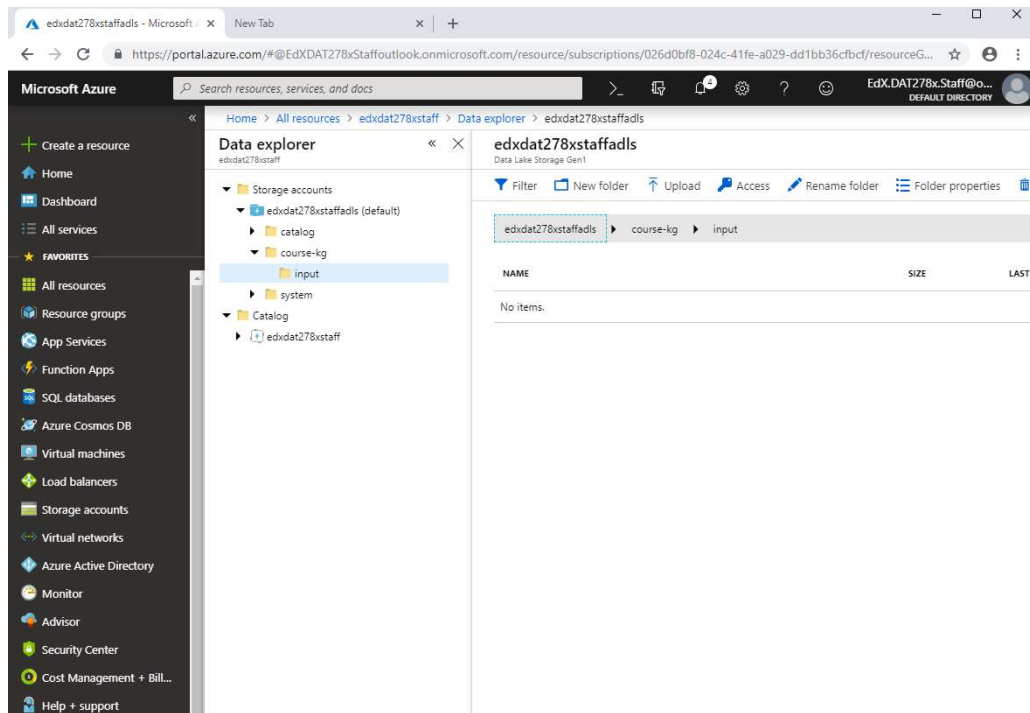




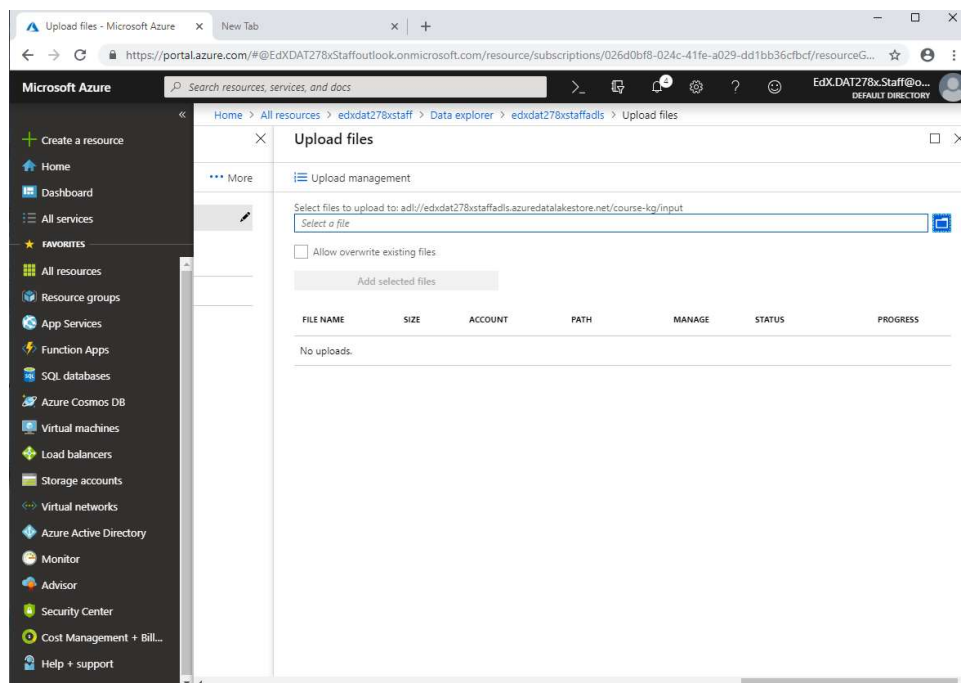
3.5 Create a new folder “input” under “course-k9” folder.



3.6 Click **input** to go into this folder.



4. Upload 2 saved data files to Azure Data Lake Store (ADLS) to prepare for Challenge Labs:
 - 4.1 Upload Paper_authors.tsv and Paper_venue.tsv one by one.



Upload files - Microsoft Azure

https://portal.azure.com/#@EdXDAT278xStaffoutlook.onmicrosoft.com/resource/subscriptions/026d0bf8-024c-41fe-a029-dd1bb36cfbct/resourceG...

Microsoft Azure

Home > All resources > edxdat278xstaff > Data explorer > edxdat278xstaffadls > Upload files

Upload files

Upload management

Select files to upload to: adl://edxdat278xstaffadls.azuredatalakestore.net/course-kg/input

Paper_authors.tsv

☐ Allow overwrite existing files

Add selected files

FILE NAME	SIZE	ACCOUNT	PATH	MANAGE	STATUS	PROGRESS
No uploads.						

Upload files - Microsoft Azure

https://portal.azure.com/#@EdXDAT278xStaffoutlook.onmicrosoft.com/resource/subscriptions/026d0bf8-024c-41fe-a029-dd1bb36cfbct/resourceG...

Microsoft Azure

Home > All resources > edxdat278xstaff > Data explorer > edxdat278xstaffadls > Upload files

Upload files

Upload management

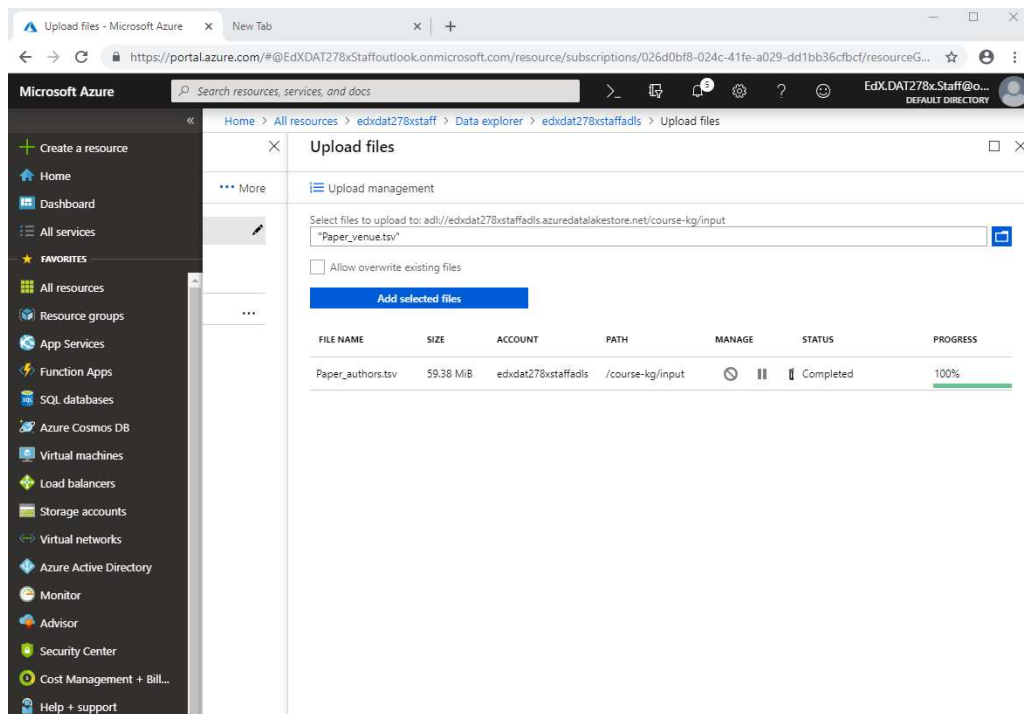
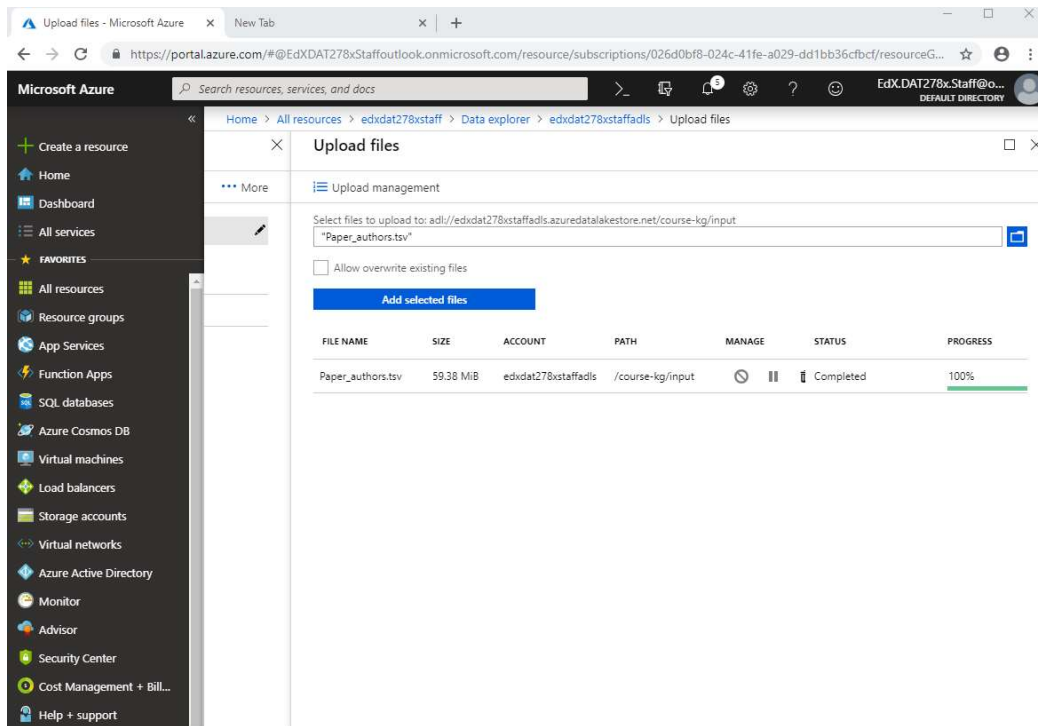
Select files to upload to: adl://edxdat278xstaffadls.azuredatalakestore.net/course-kg/input

Paper_authors.tsv

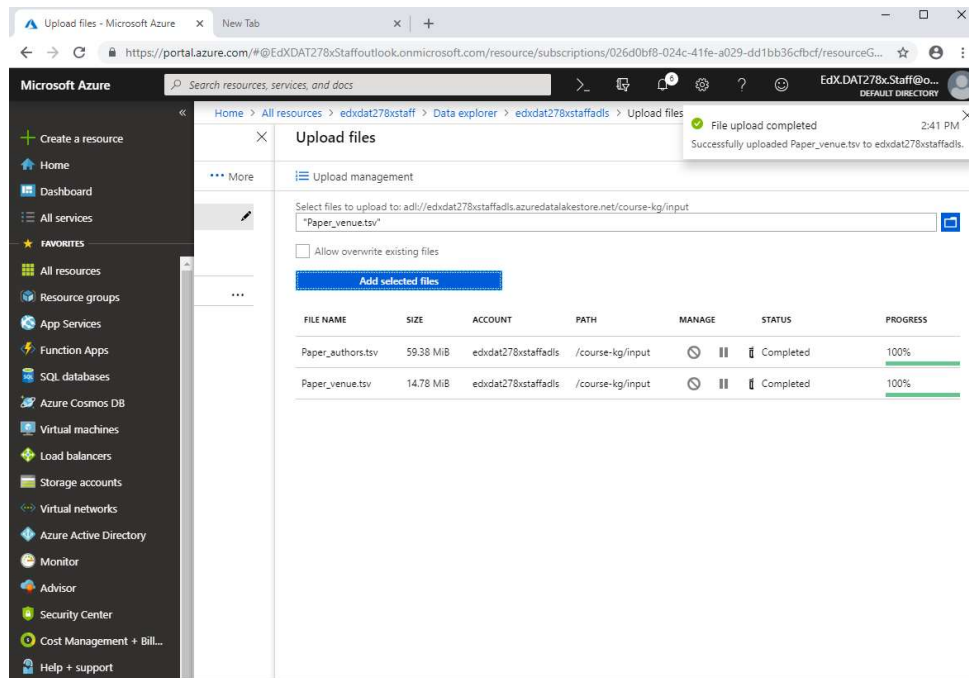
☐ Allow overwrite existing files

Add selected files

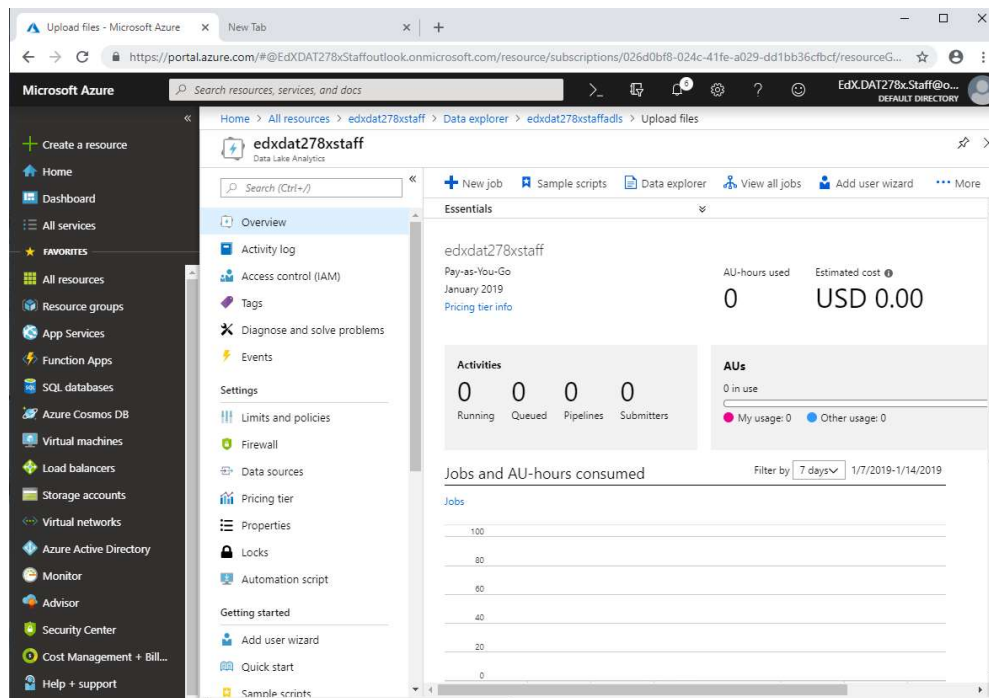
FILE NAME	SIZE	ACCOUNT	PATH	MANAGE	STATUS	PROGRESS
Paper_authors.tsv	59.38 MiB	edxdat278xstaffadls	/course-kg/input		Uploading	58%



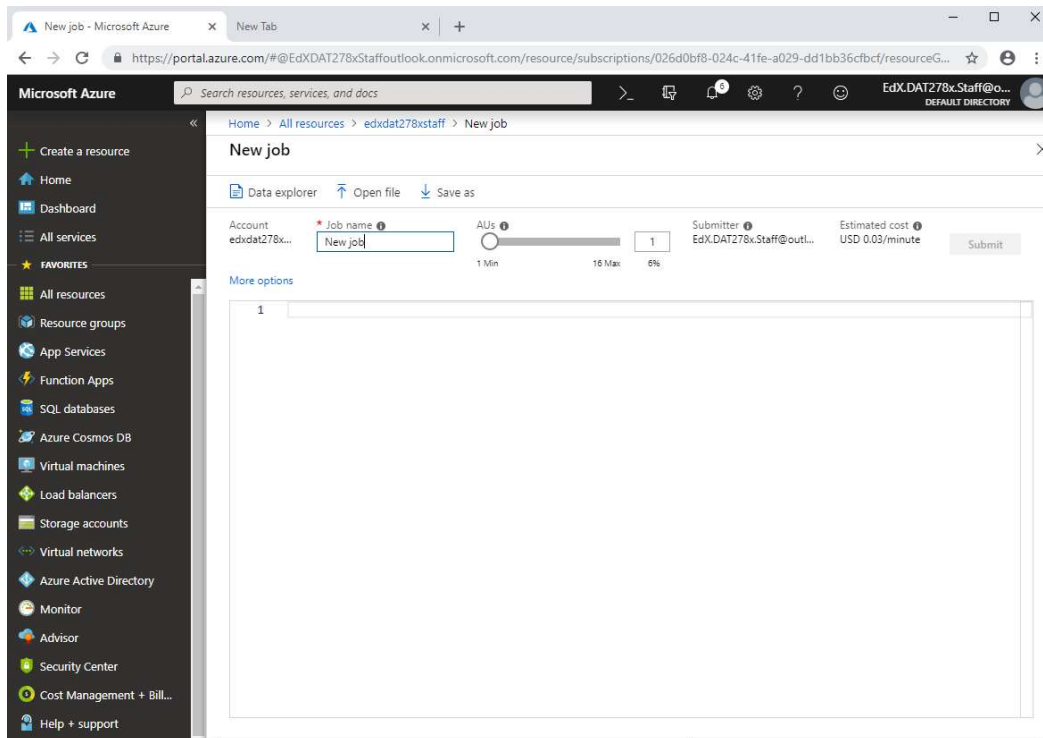
4.2 View uploaded files under “course-kg/input” folder.



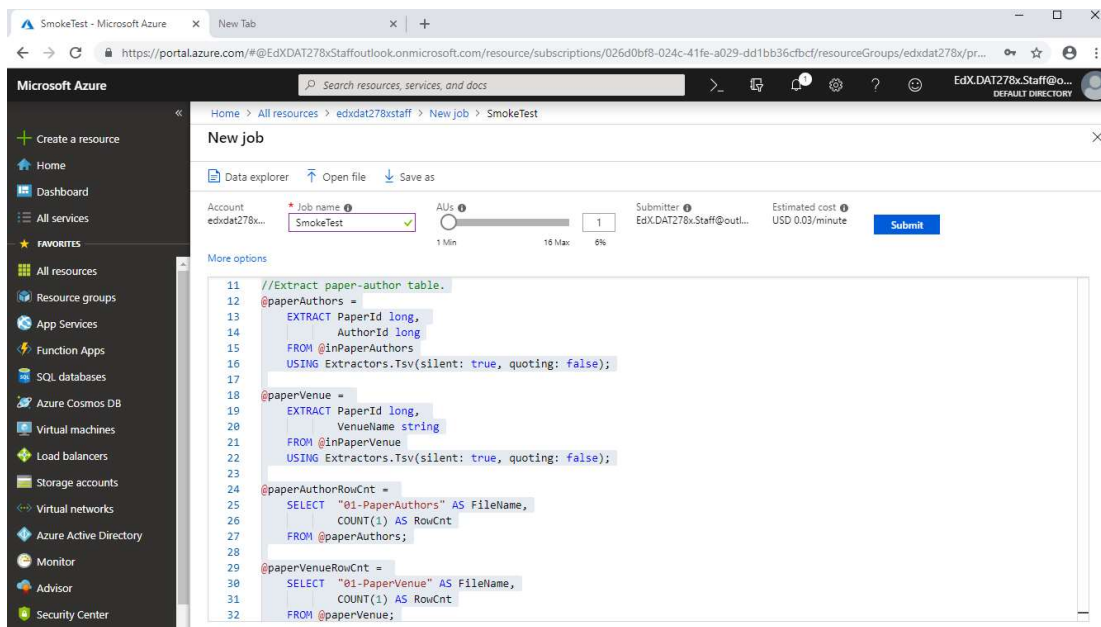
5. Run a smoke test U-SQL script to verify the environment is setup properly
 - 5.1 Go to Azure Data Lake Analytics account and click **New job** on the top menu.



- 5.2 Click **Open file** to open the downloaded U-SQL script “DAT278x_SmokeTest.usql”.



5.3 Click **Submit** to submit the job. The computing cost will be covered by [the Azure credit](#) that you received when you registered for the free trial account.



5.4 View the job progress as follow, and after running the script successfully, the results will be outputted to the `"/course-kg/output/"` folder.

SmokeTest
Job details

Refresh in 9 sec Resubmit Reuse script Cancel job

Status: Running

Progress: 100%
 AUs: 1
 Consumed AU-hours: 0
 Estimated cost: USD 0.01
 Efficiency: N/A
 Issues: 0 issues

Type: U-SQL
 Runtime version: release_20180820_adl_1780989
 Submitter: EdX.DAT278x.Staff@outlook.com
 Account: edxdat278xstaff
 Priority: 1000

Preparing: 18s
 Queued: N/A
 Running: N/A
 Duration: 24s

Job graph

Display Progress Playback 0s Zoom to fit

The job graph shows the following stages:

- Paper_authors.tsv** (1.4 MB) - SV1 Extract (1 vertex, R: 59.4 MB, W: 8 bytes, Stage progress: 100%)
- Paper_venue.tsv** (1.4 MB) - SV3 Extract (1 vertex, R: 14.8 MB, W: 8 bytes, Stage progress: 100%)
- SV2 Aggregate** (1 vertex, R: 8 bytes, W: 25 bytes, Stage progress: 100%)
- SV4 Aggregate** (1 vertex, R: 8 bytes, W: 23 bytes, Stage progress: 100%)
- SV5 Combine** (1 vertex, R: 48 bytes, W: 72 bytes, Stage progress: 100%)

The final output is **Results DAT278x Smoke...**

Data explorer
edxdat278xstaffadls

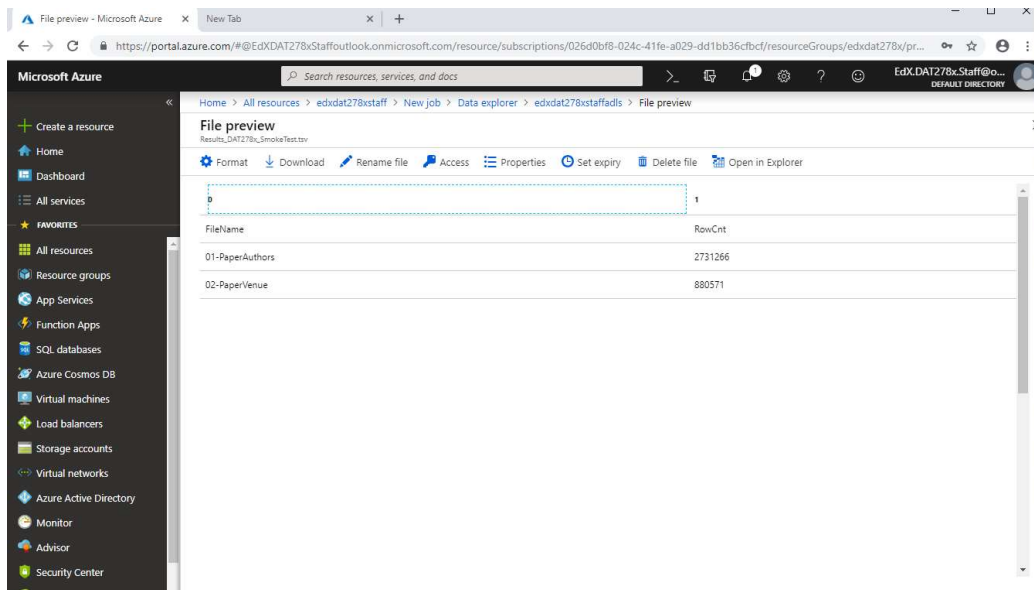
Filter New folder Upload Access Rename folder Folder properties Delete folder

Storage accounts
 ▼ edxdat278xstaffadls (default)
 catalog
 course-kg
 system
 Catalog
 ▼ edxdat278xstaff
 master

edxdat278xstaffadls course-kg output

NAME	SIZE	LAST MODIFIED
Results_DAT278x_SmokeTest.tsv	72 bytes	1/14/2019, 2:56:54 PM

5.5 Verify the result file contains following two lines of contents to pass the smoke test.



If you can follow above steps without issues and generate the exact contents for smoke testing scripts, congratulations! You have set up the lab environment successfully and you are ready for the challenge labs!

If you have any issues or questions on above steps, please use the discussion forum of DAT278x on the EdX platform for feedbacks, our course staff would monitor and answer questions you raise.