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 lab assignments.

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:

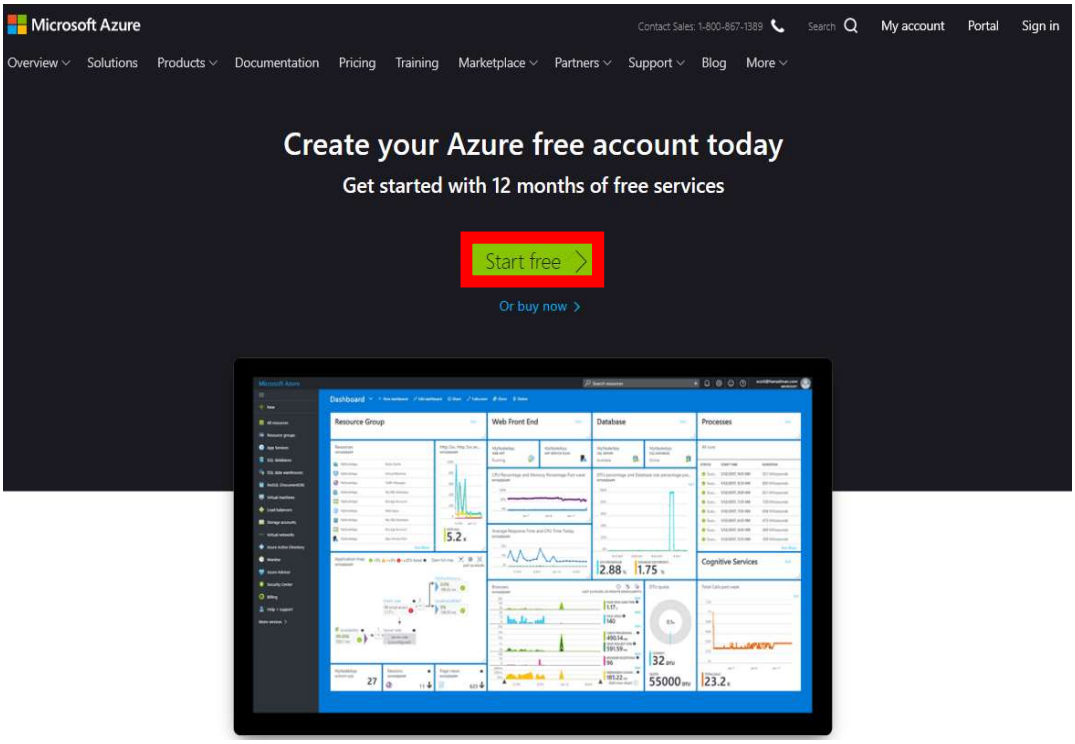
- EdX course on Processing Big Data with Azure Data Lake Analytics
- U-SQL resource site: usql.io

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

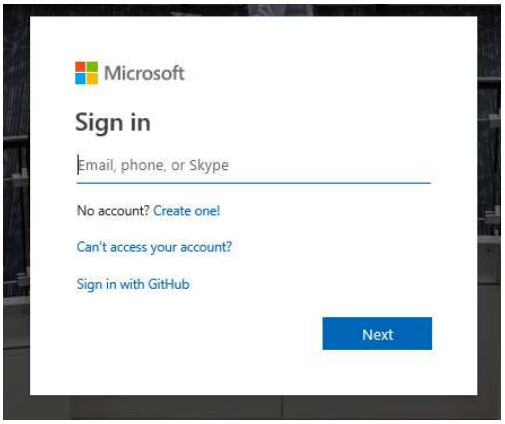
1. Create an Azure free account.
2. Create an Azure Data Lake Analytics (ADLA) resource and an Azure Data Lake Storage (ADLS) resource.
3. Create folder structure in Azure Data Lake Storage (ADLS) to prepare for smoke testing and challenge lab runs.
4. Upload data files to Azure Data Lake Storage (ADLS).
5. Run a smoke test U-SQL script to verify that the environment is setup properly.

1. Create a free Azure account:

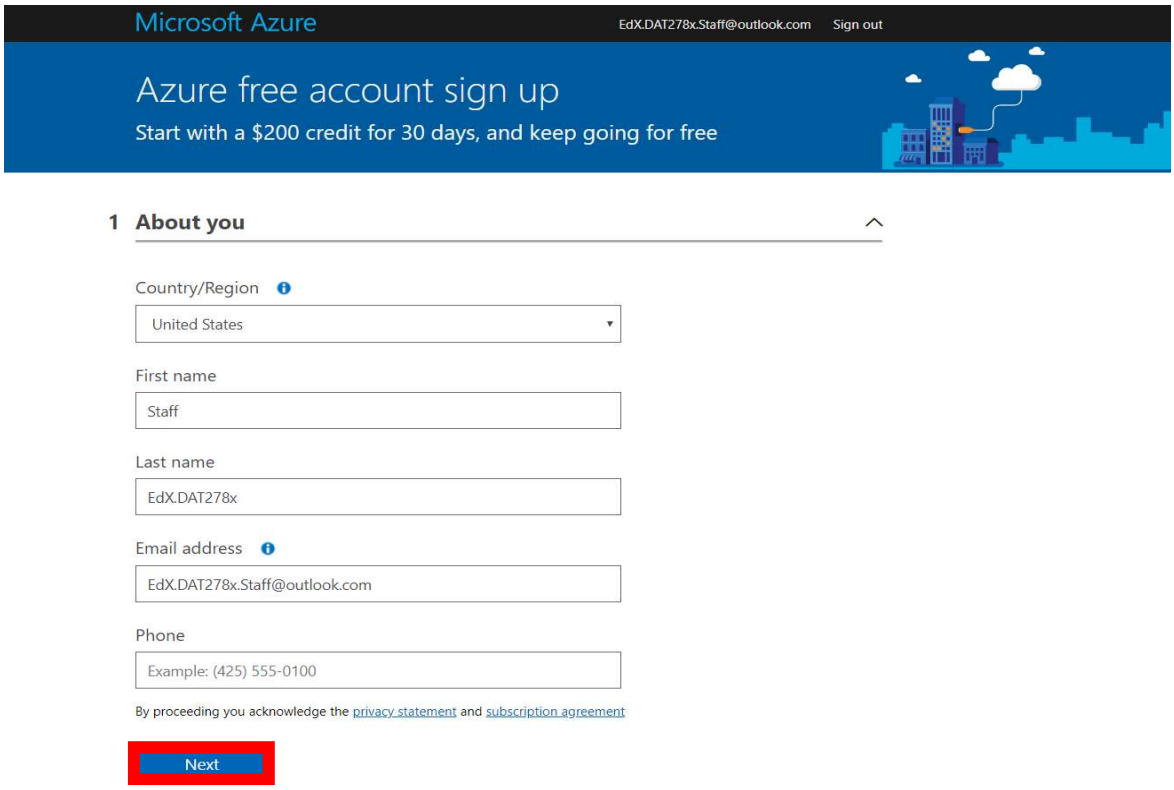
1.1 Go to this page: https://azure.microsoft.com/free ; click on “start-free”.



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

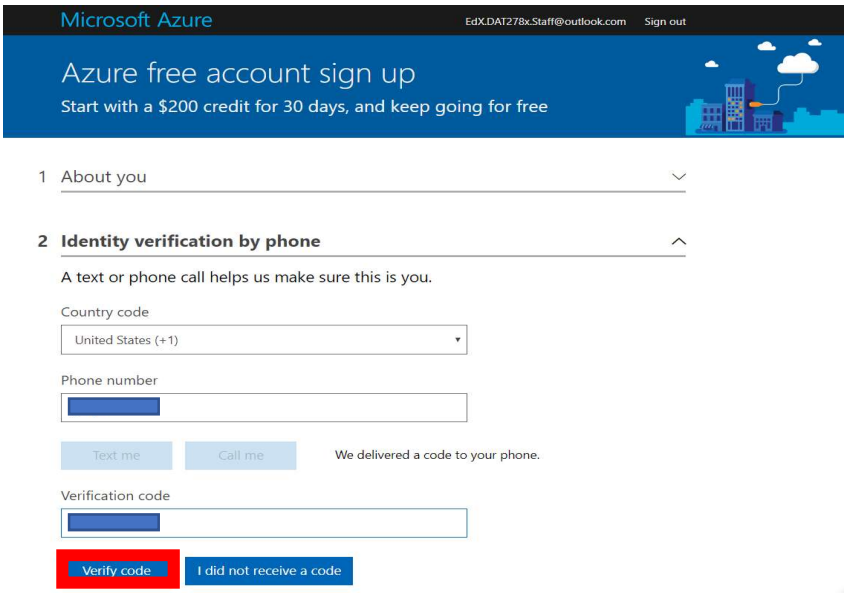


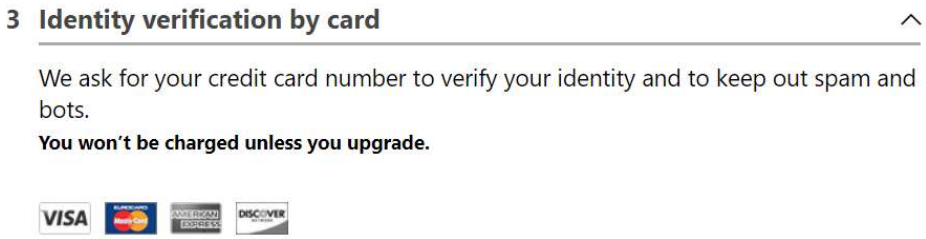
1.3 Land on Microsoft Azure sign-up page, “about you” section, fill in your information and click “Next”.



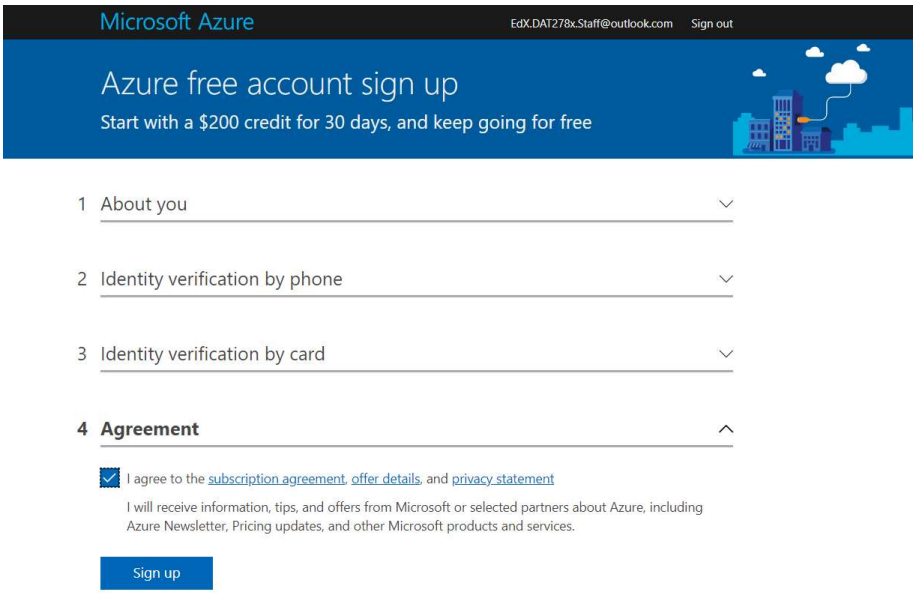
1.4 Provide phone information and credit card information for identity verification purpose. (**PLEASE NOTE:** Your credit card will **NOT** be charged unless you upgrade your subscription. For EdX DAT278x course, the “free trial” credit would be enough to finish all labs.)

For students only: if you are a student with a valid school email address, you can sign up Azure without providing credit card information. For more info, please refer below links. https://azure.microsoft.com/free/free-account-students-faq/

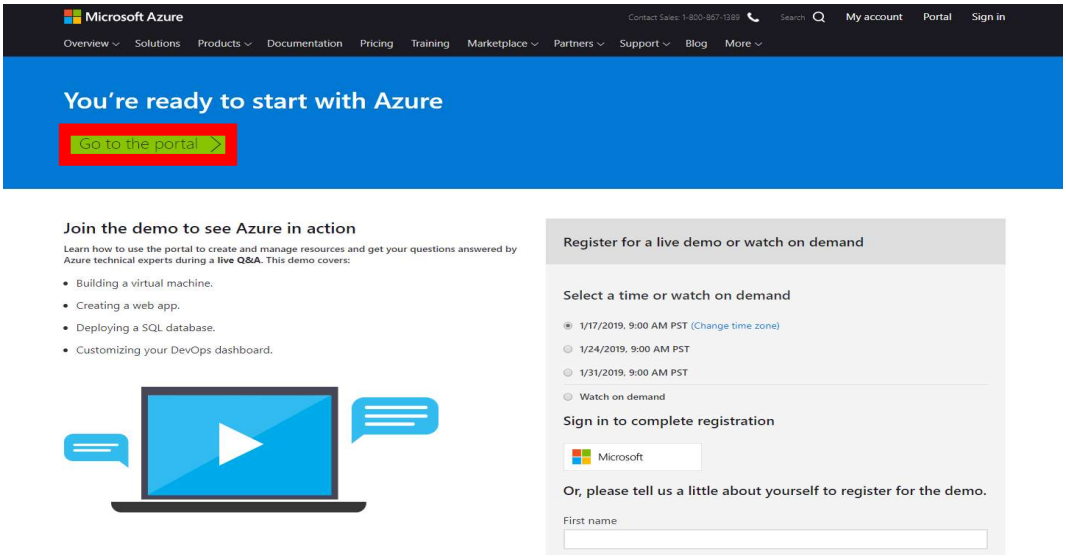




1.5 To accept the "Agreement" and click the "Sign up" to proceed.

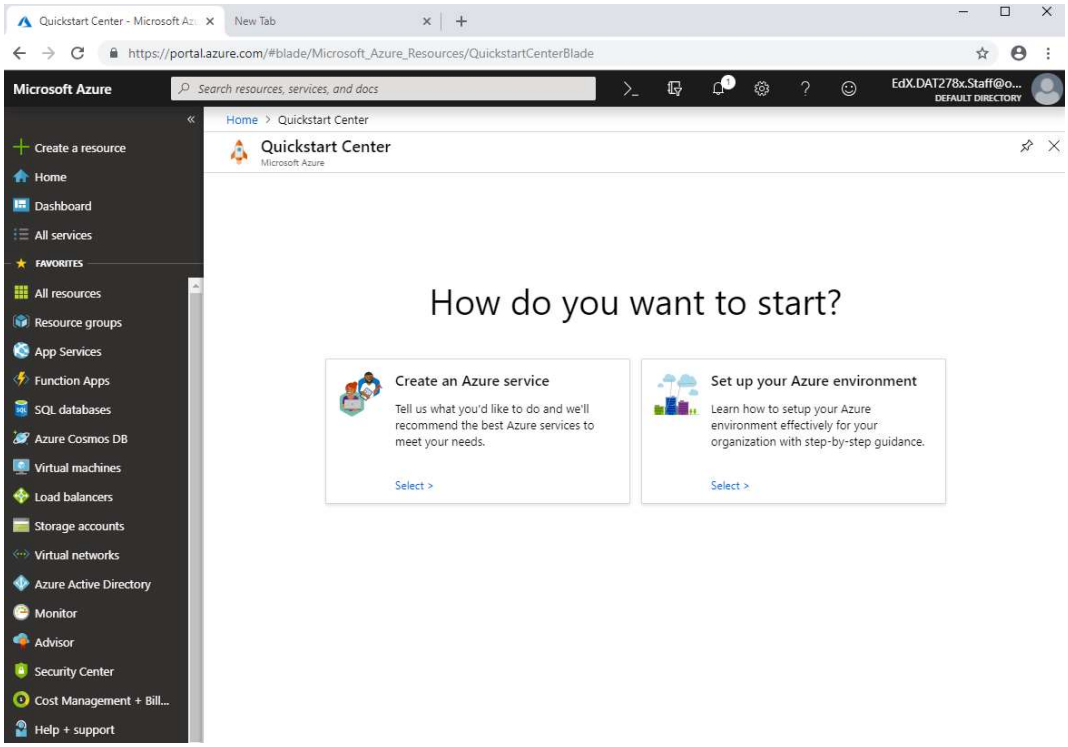


1.6 You can see below page after successfully sign up with Azure account, click on "Go to the portal"

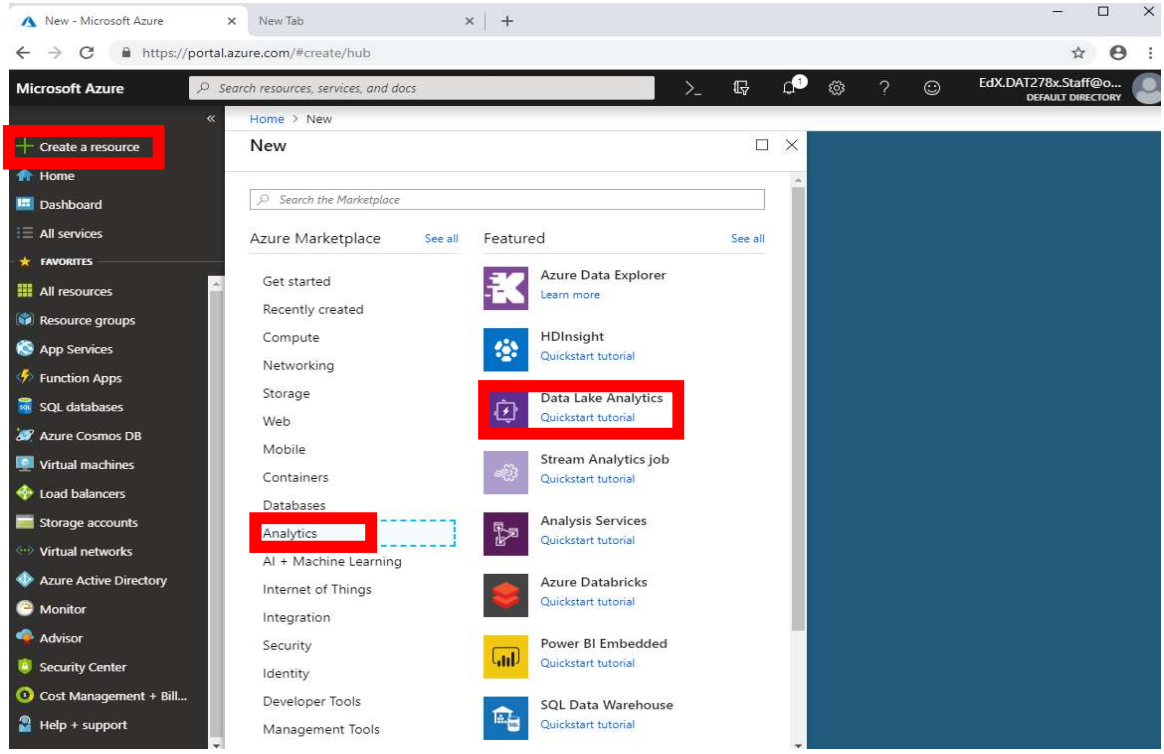


2 On Azure Portal, create an Azure Data Lake Analytics (ADLA) resource and an Azure Data Lake Storage (ADLS) resource

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



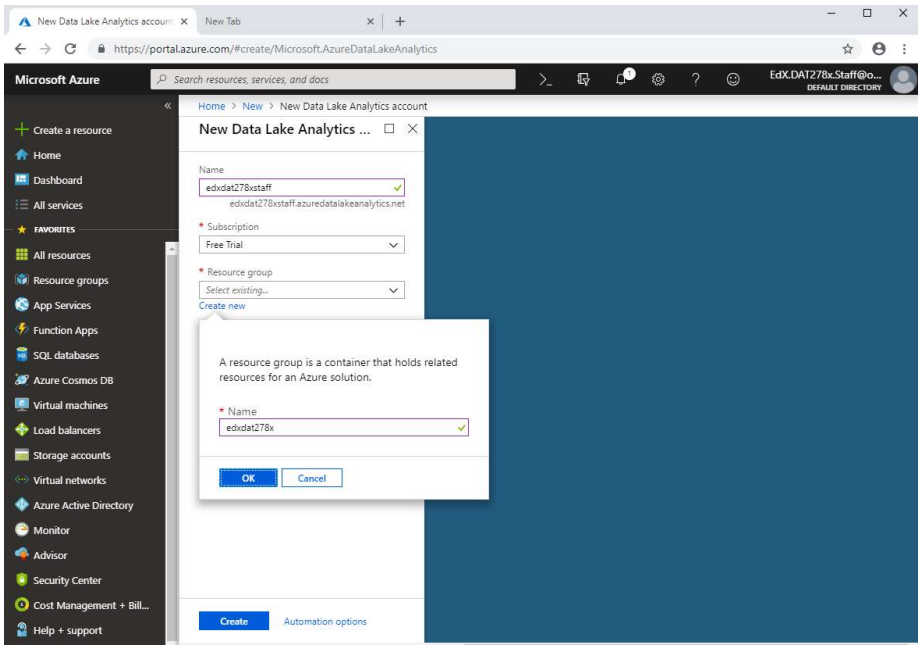
2.2 Click on “Create a resource” -> “Analytics” -> “Data Lake Analytics”



2.3 Fill in values for following items:

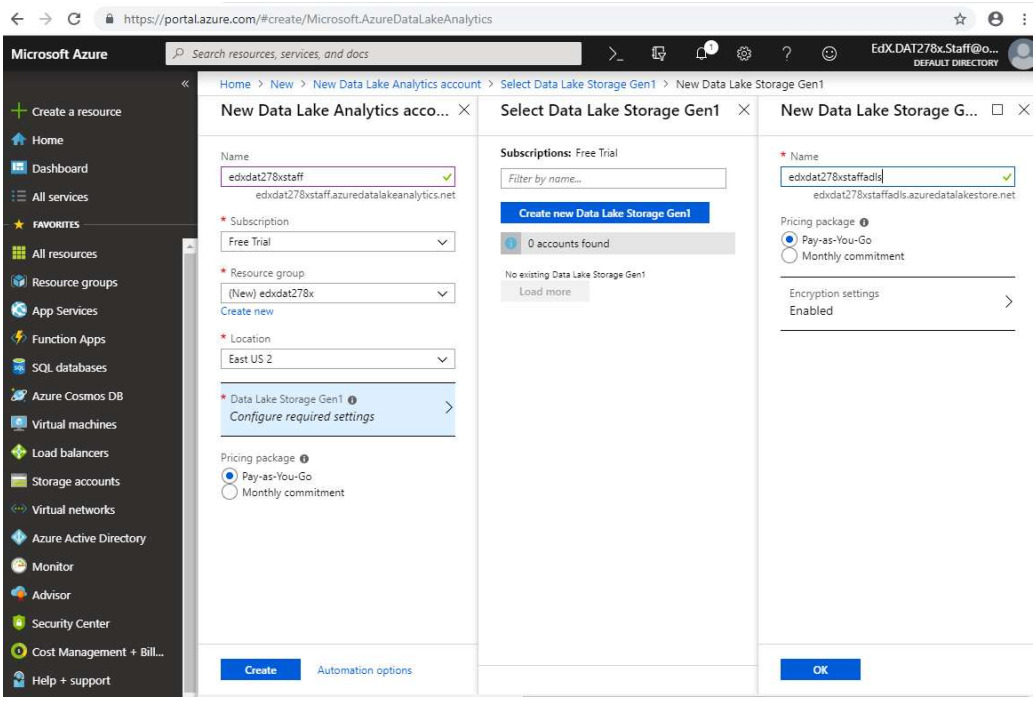
a. **Name**: Name your Data Lake Analytics account (Only lower-case letters and numbers allowed).
b. **Subscription**: Choose the Azure subscription (select "free trial" for DAT278x course unless you have other subscriptions) used for the Analytics account.
c. **Resource Group:** Select an existing Azure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

2.3.1 Pick a new name for Resource Group

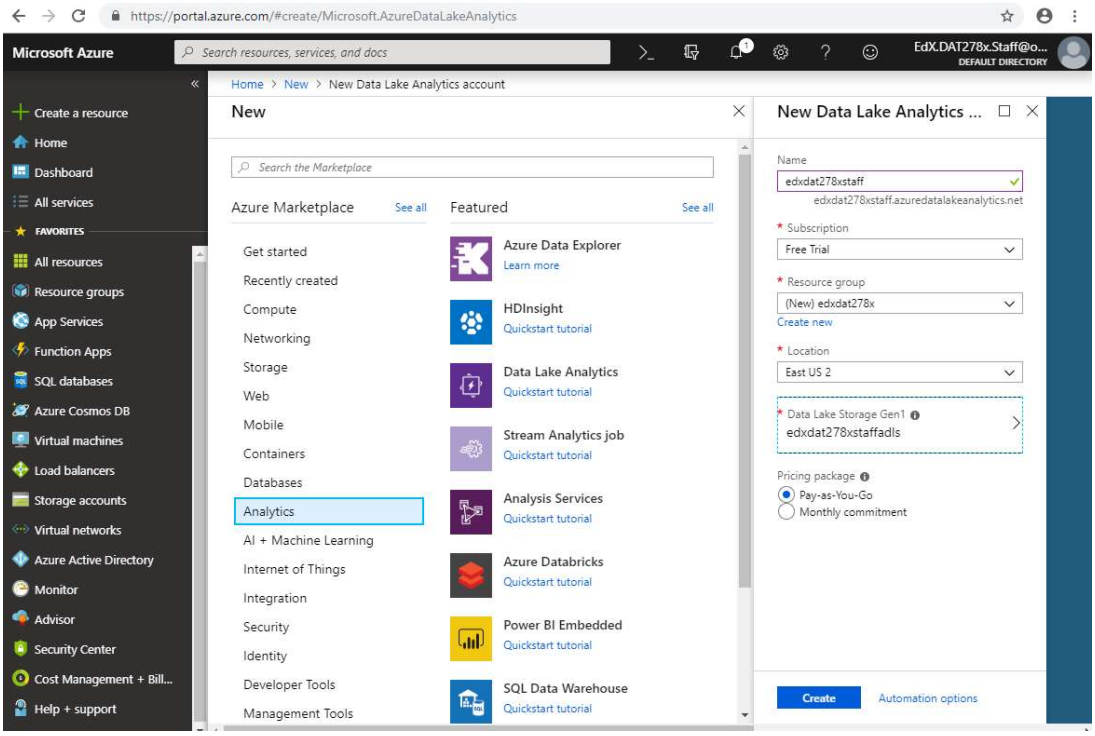


2.3.2 Create a new Azure Data Lake Storage (ADLS) account to associate with the Azure Data Lake Analytics (ADLA) account.

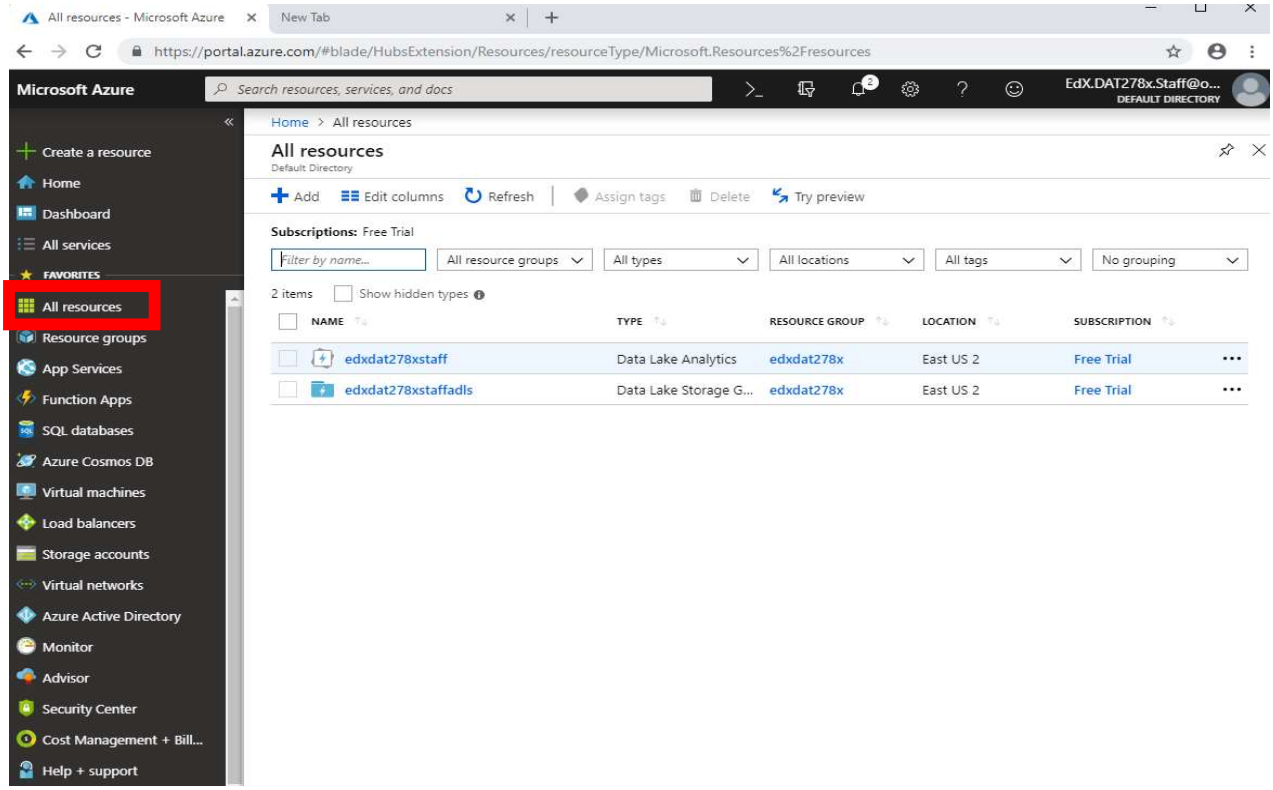
Use the default "Pay-as-You-Go" option for "free trial" subscription, pick a different package if you have other subscriptions.



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

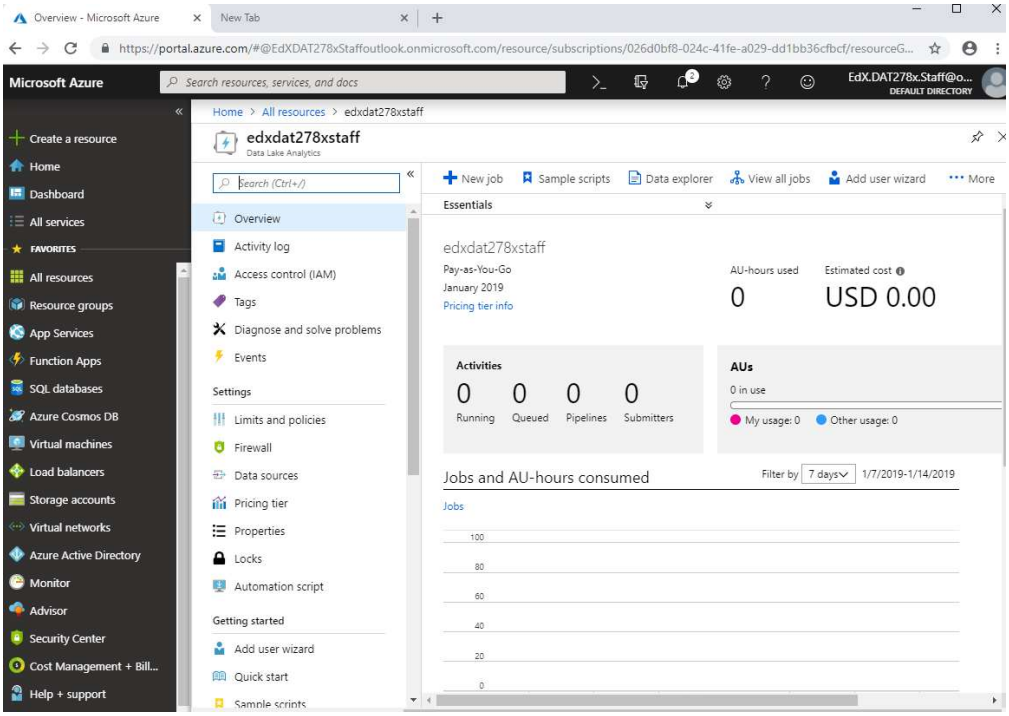


2.4 View the created Data Lake Analytics and Data Lake Storage resources in “All resources”.

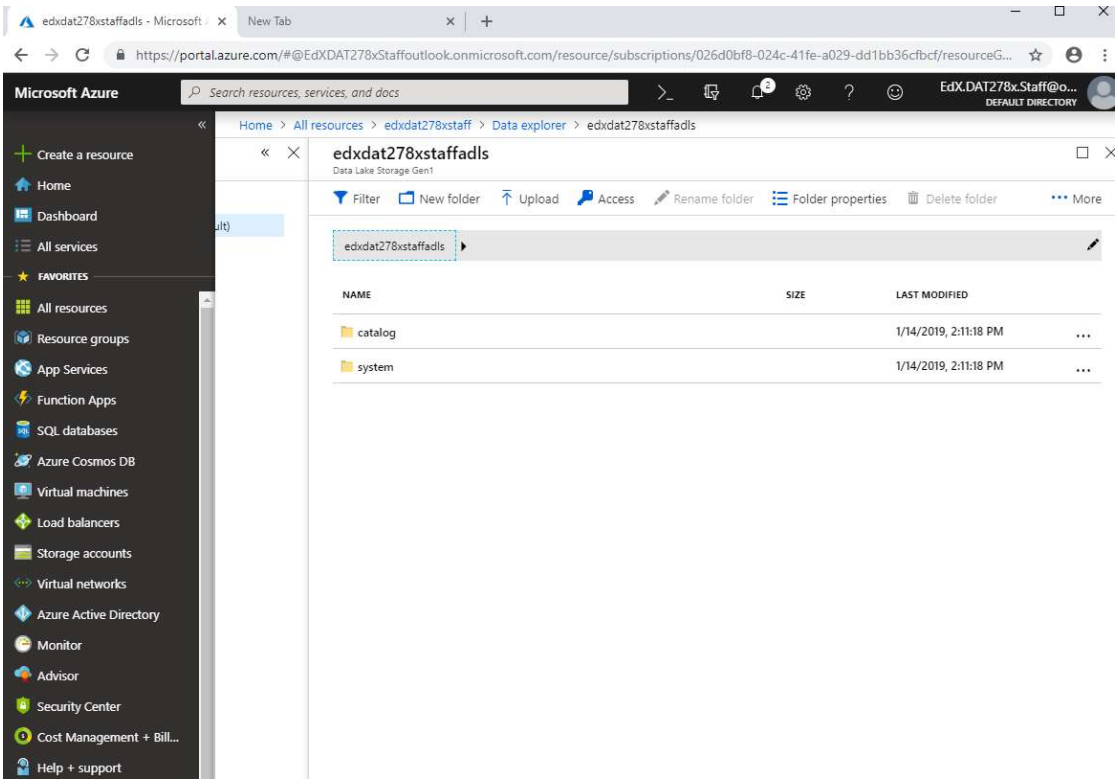


3 Create folder structure in Azure Data Lake Storage (ADLS) to prepare for challenge lab runs.

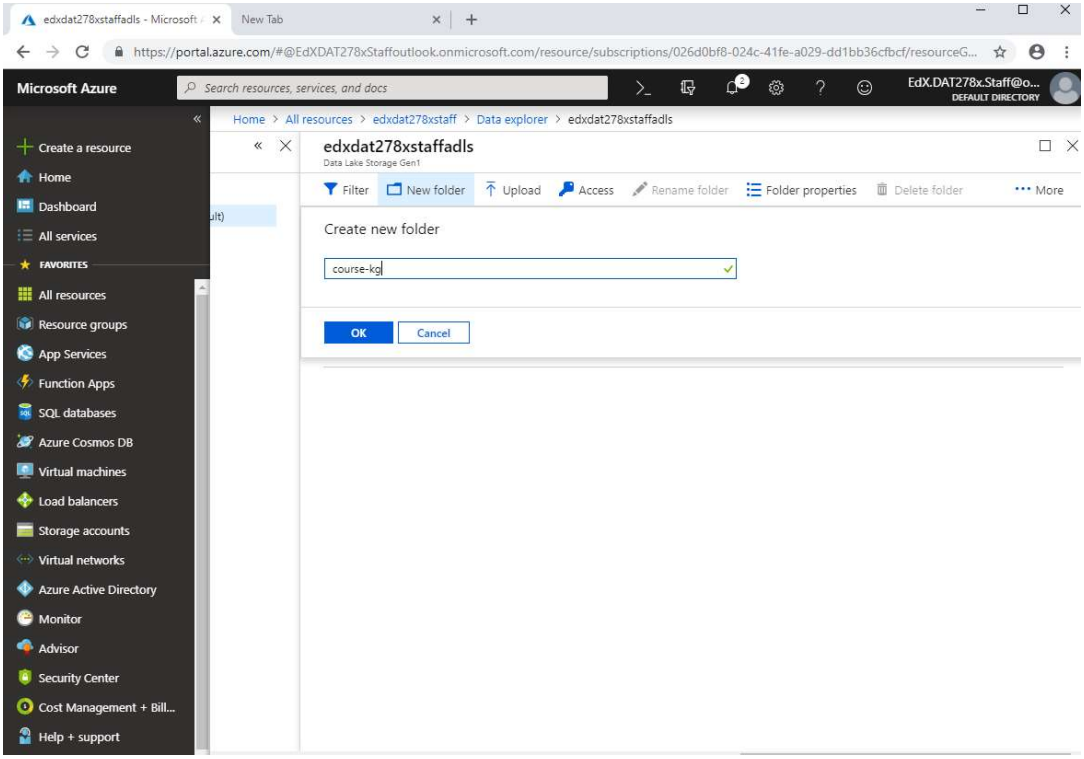
3.1 Go to Azure Data Lake Analytics account and click **Data explorer**.



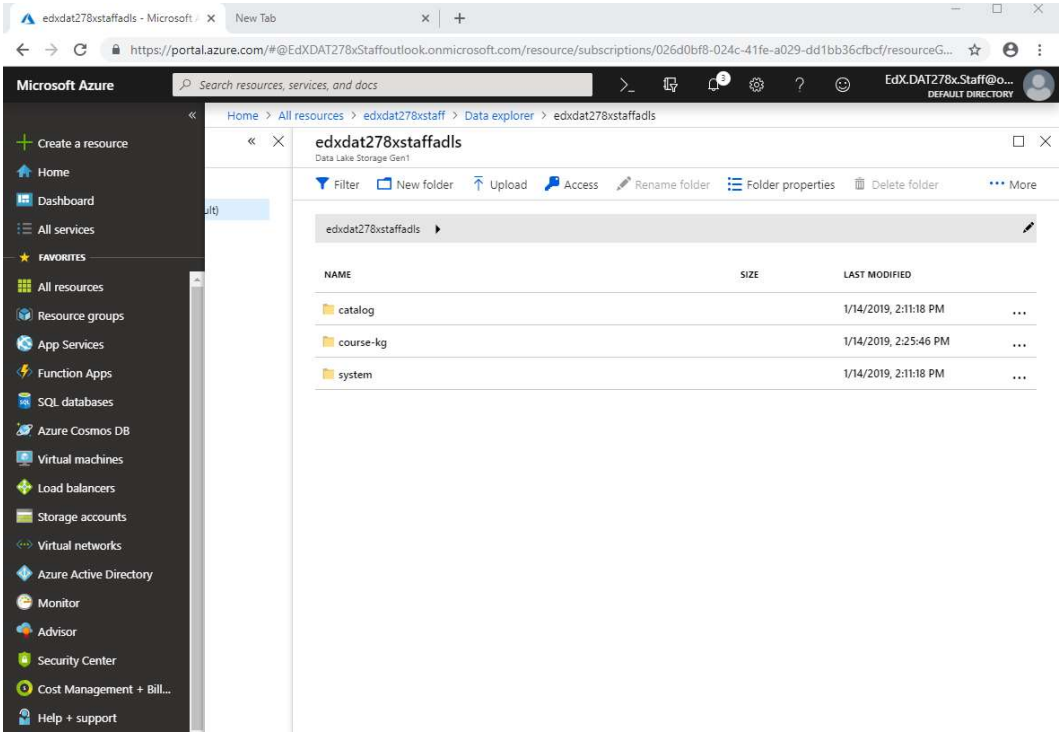
3.2 Click **New folder** to create a new folder under Azure Data Lake Storage account.

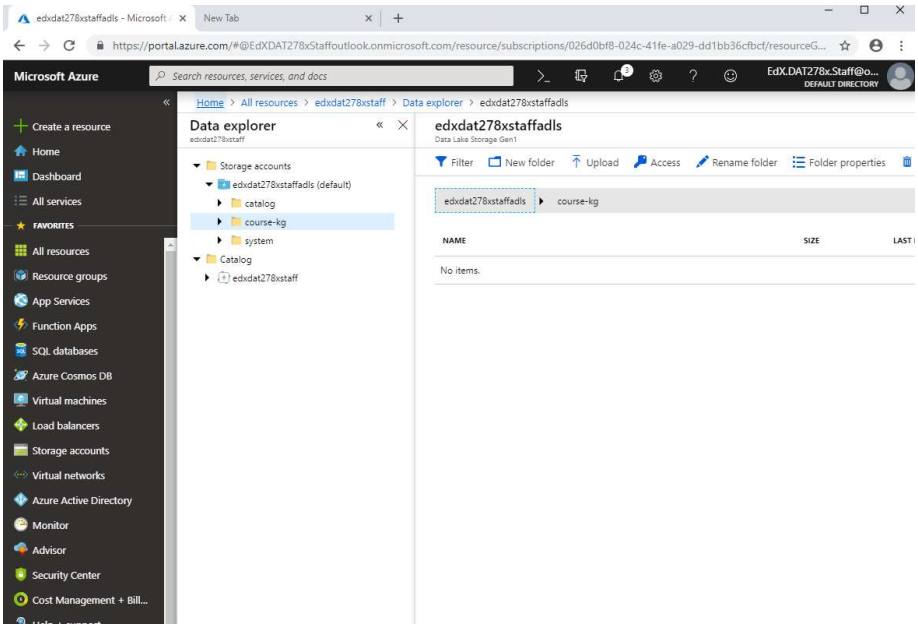


3.3 Use “**course-kg**” as the folder name and click **OK**.

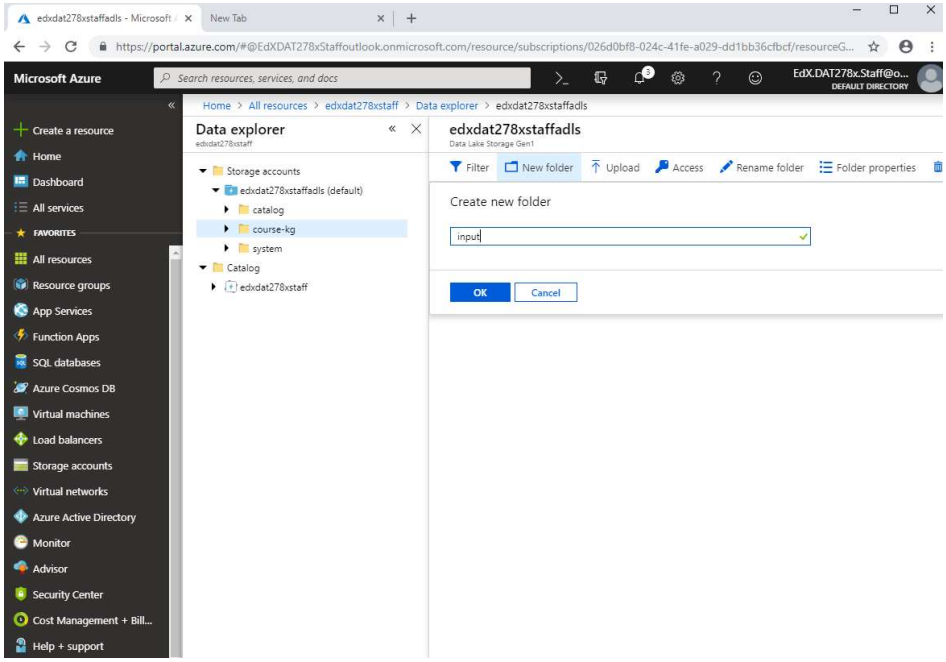


3.4 Click **course-kg** to go into this folder

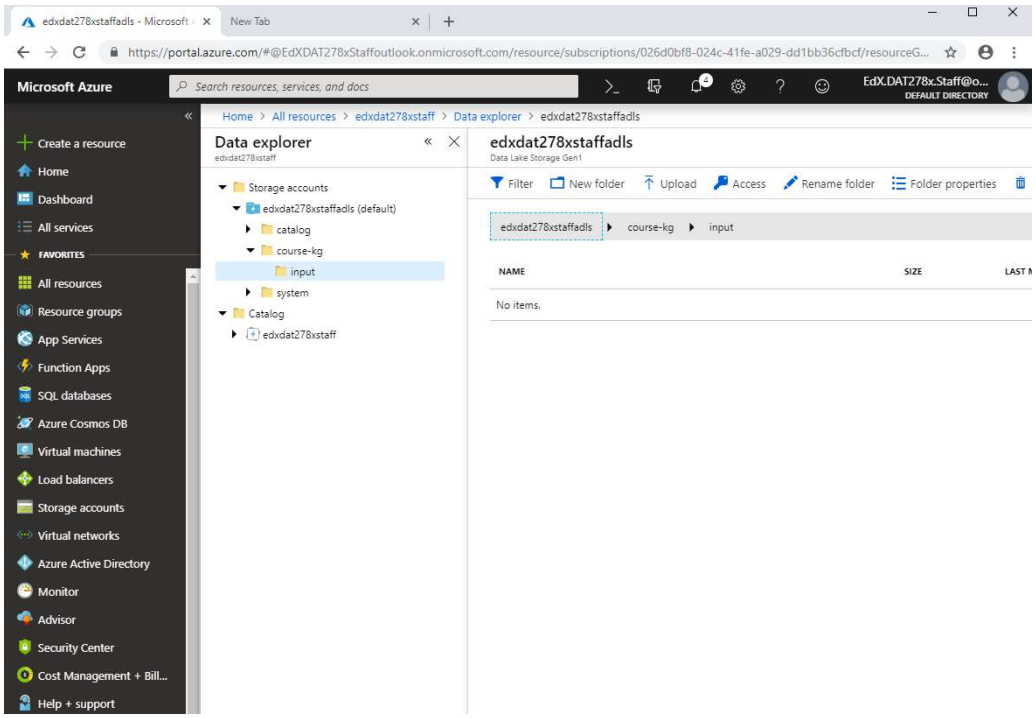




3.5 Create a new folder "**input**" under "course-kg" folder.

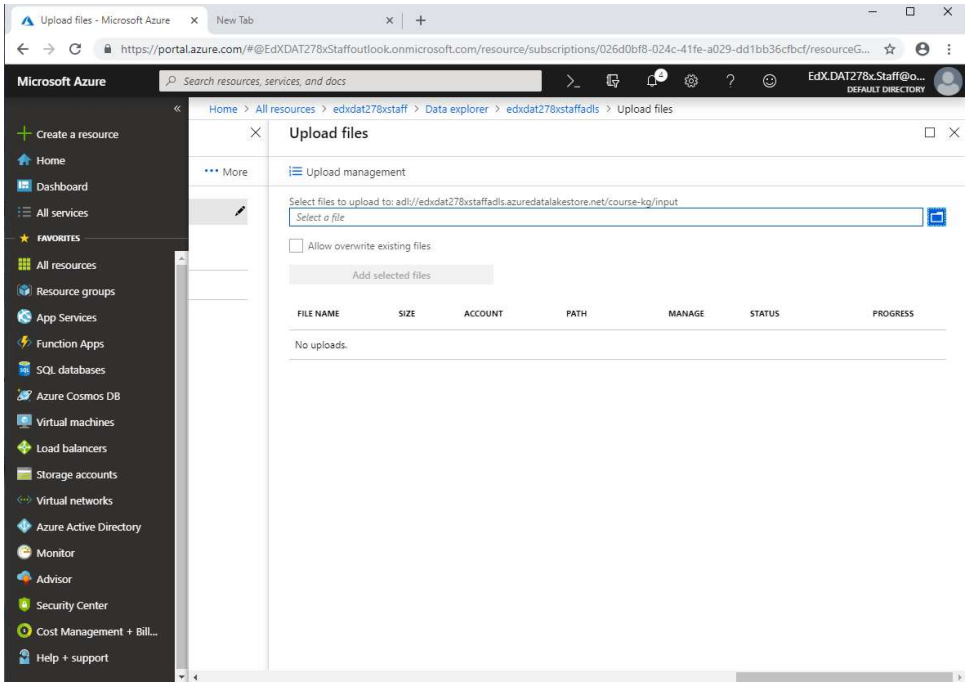


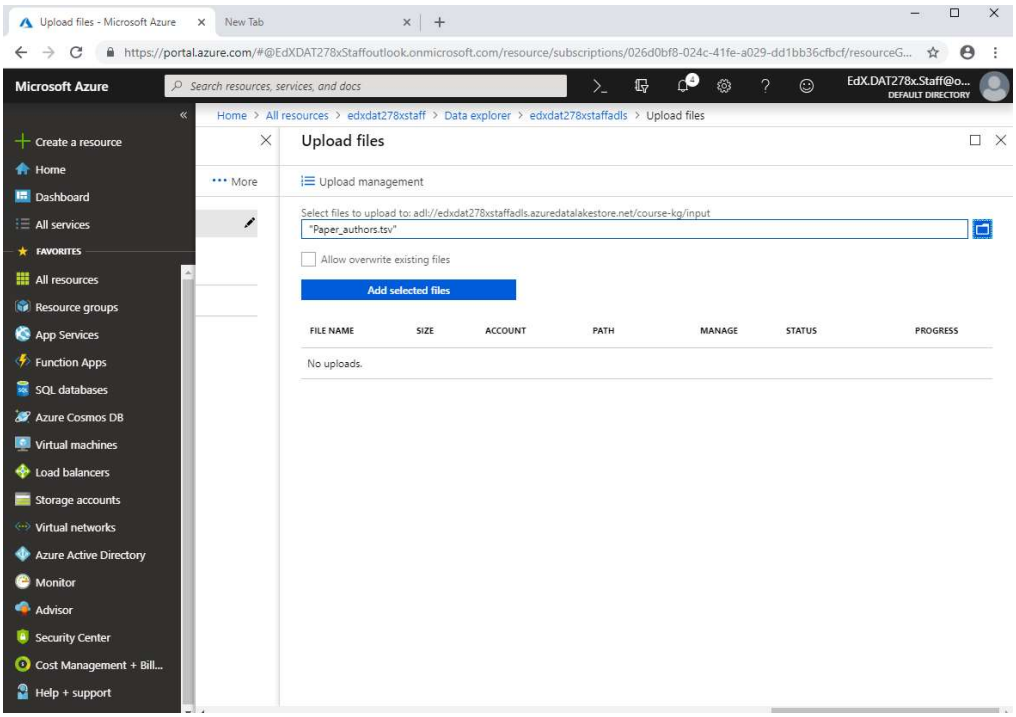
3.6 Click **input** to go into this folder.

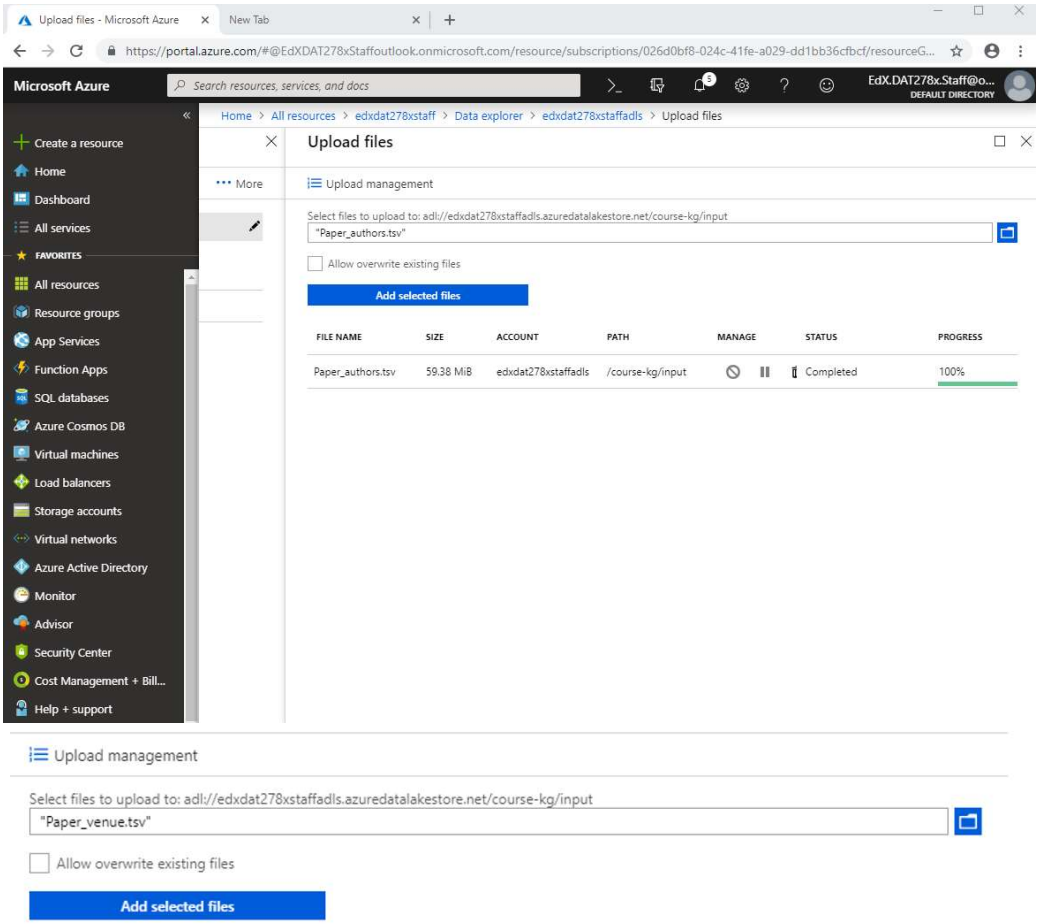

edxdat278xstaffadls - Microsoft
New Tab
https://portal.azure.com/#@EdXDAT278xStaffoutlook.onmicrosoft.com/resource/subscriptions/026d0bf8-024c-41fe-a029-dd1bb36cfbcf/resourceG...
Microsoft Azure
Search resources, services, and docs
EdX.DAT278x.Staff@o...
DEFAULT DIRECTORY
Home > All resources > edxdat278xstaff > Data explorer > edxdat278xstaffadls
Create a resource
Home
Dashboard
All services
FAVORITES
All resources
Resource groups
App Services
Function Apps
SQL databases
Azure Cosmos DB
Virtual machines
Load balancers
Storage accounts
Virtual networks
Azure Active Directory
Monitor
Advisor
Security Center
Cost Management + Bill...
Help + support
Data explorer
edxdat278xstaff
Storage accounts
edxdat278xstaffadls (default)
catalog
course-kg
input
system
Catalog
edxdat278xstaff
edxdat278xstaffadls
Data Lake Storage Gen1
Filter
New folder
Upload
Access
Rename folder
Folder properties
edxdat278xstaffadls
course-kg
input
NAME
SIZE
No items.

4 Upload data files to Azure Data Lake Storage (ADLS) to prepare for smoke testing and Challenge Lab runs:

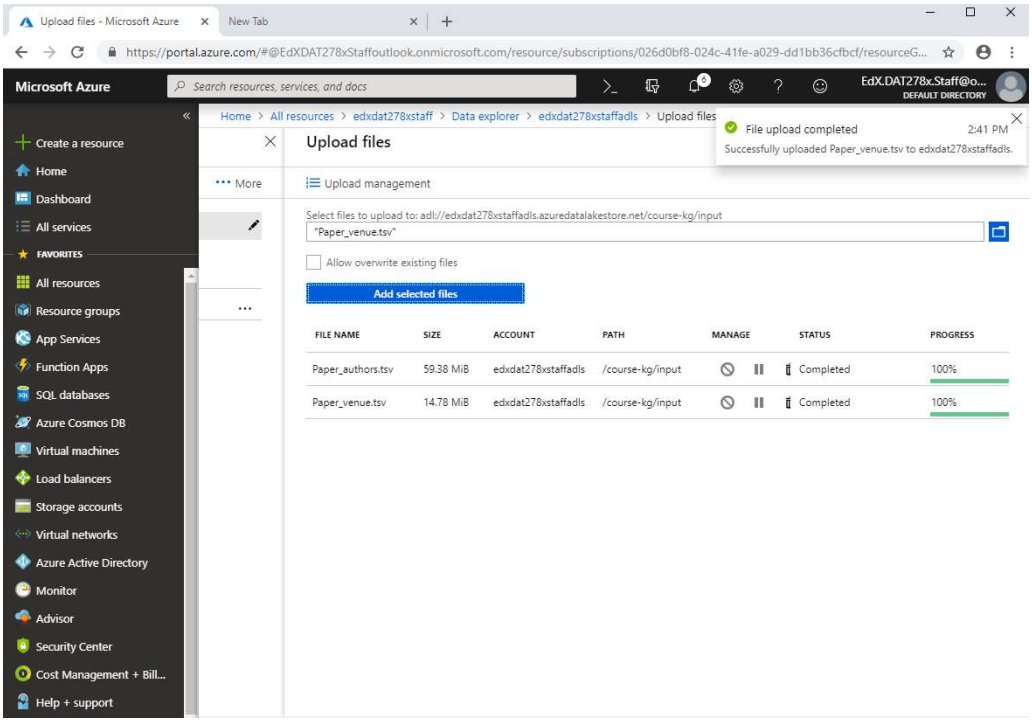
4.1 Upload Paper_authors.tsv and Paper_venue.tsv to the created "**/course-kg/input/**".





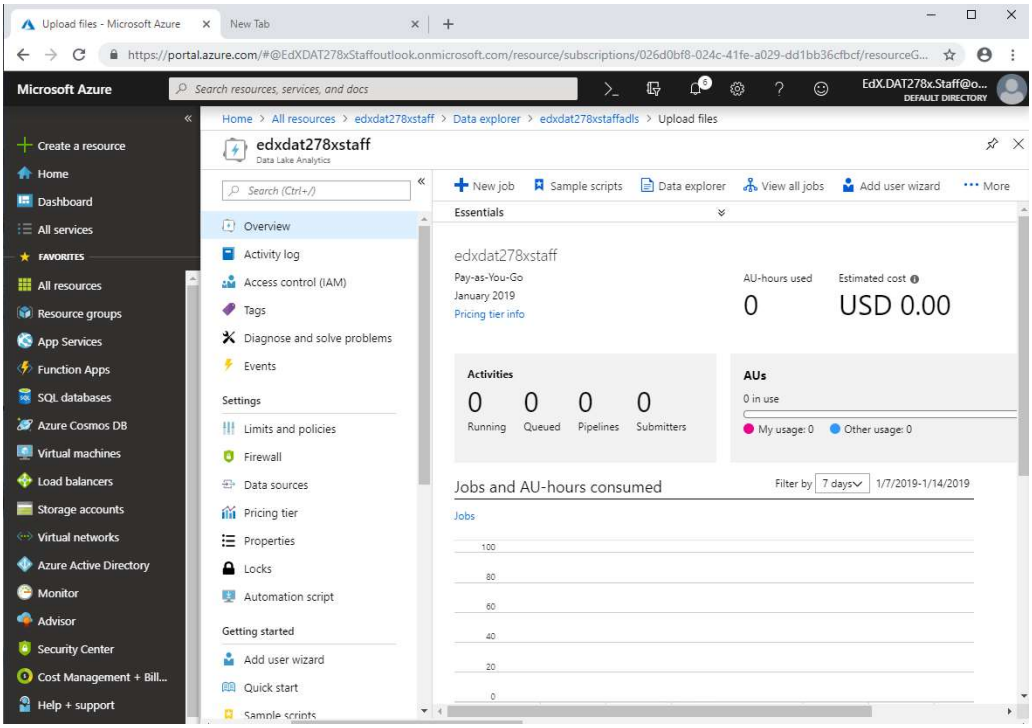


4.2 View uploaded files under "**course-kg/input**" folder.

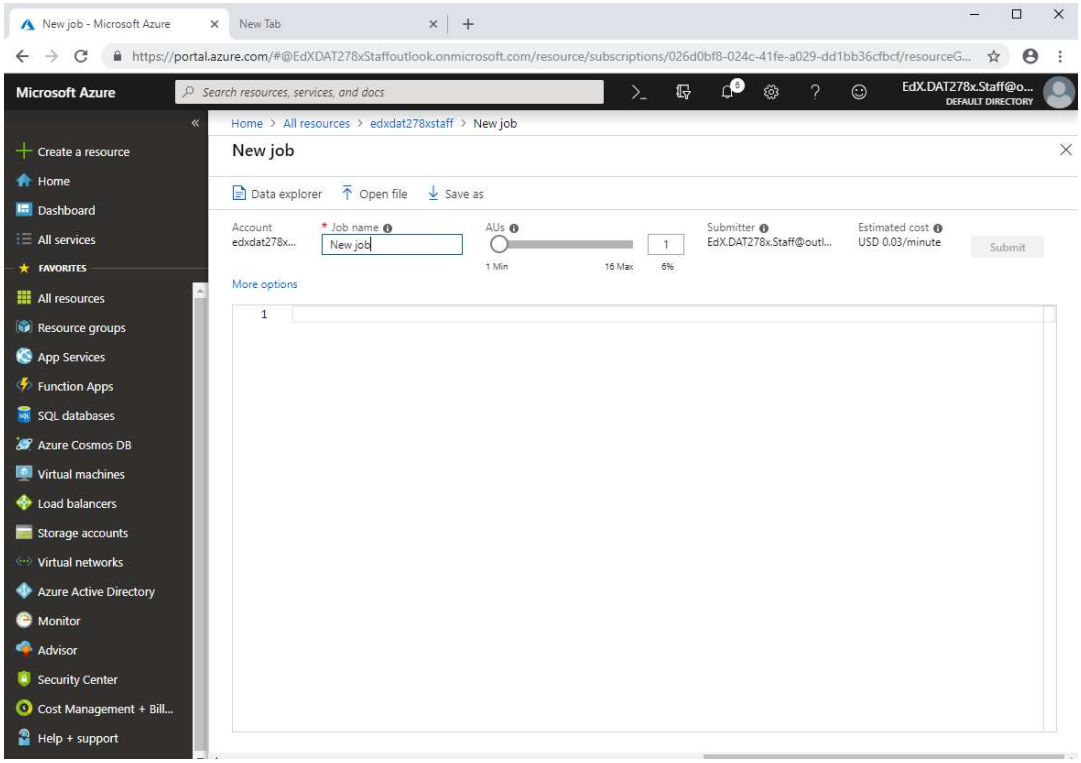


5 Run a smoke test U-SQL script to verify that the environment is setup properly.

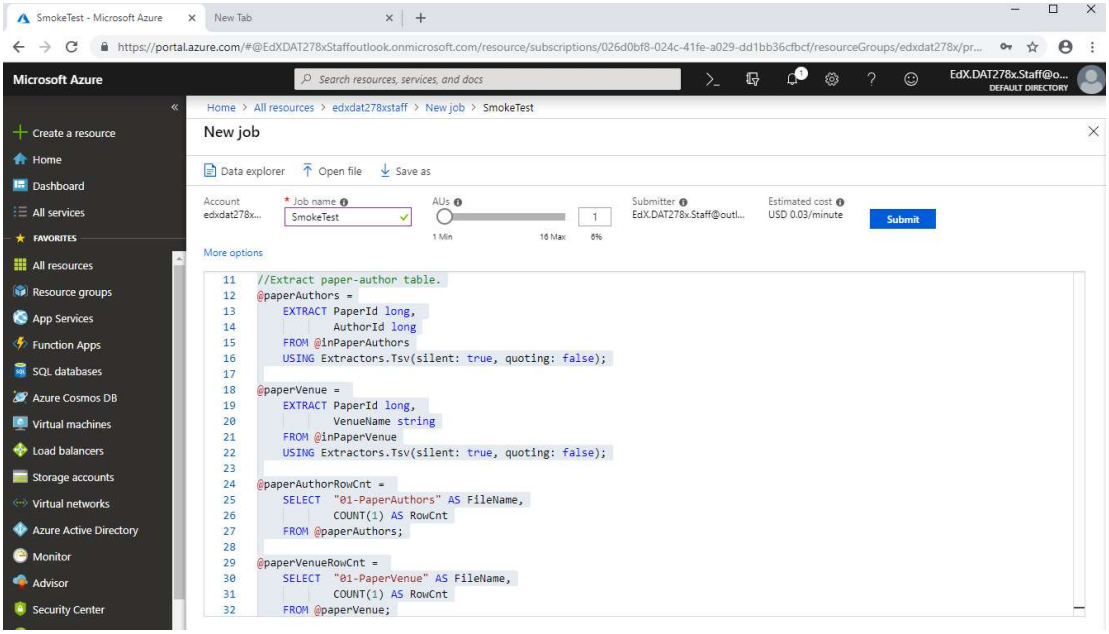
5.1 Go to Azure Data Lake Analytics account and click **New job** on the top menu.



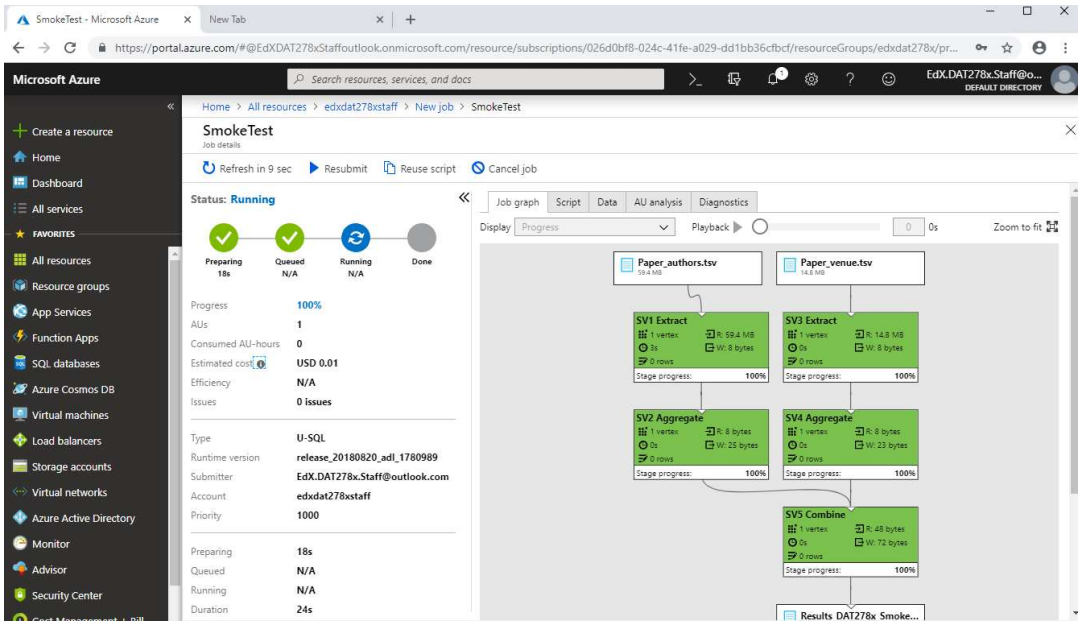
5.2 Click **Open file** to open the downloaded U-SQL script "DAT278x_SmokeTest.usql".

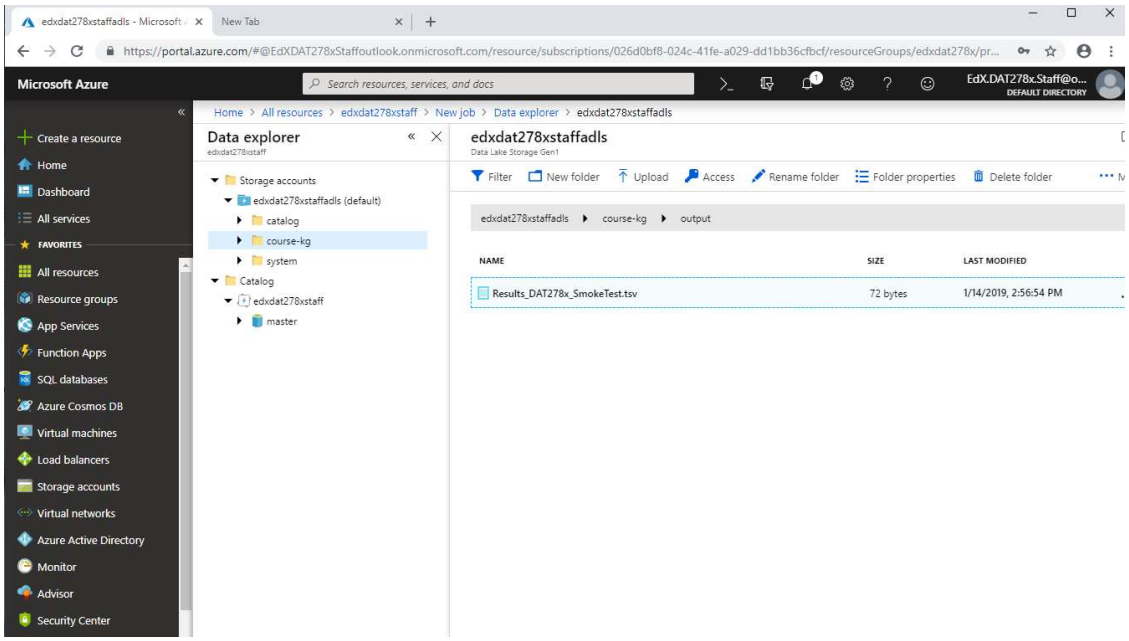


5.3 Click **Submit** to submit the job. The computing cost will be covered by Azure credit which you received when registering the free trial account.

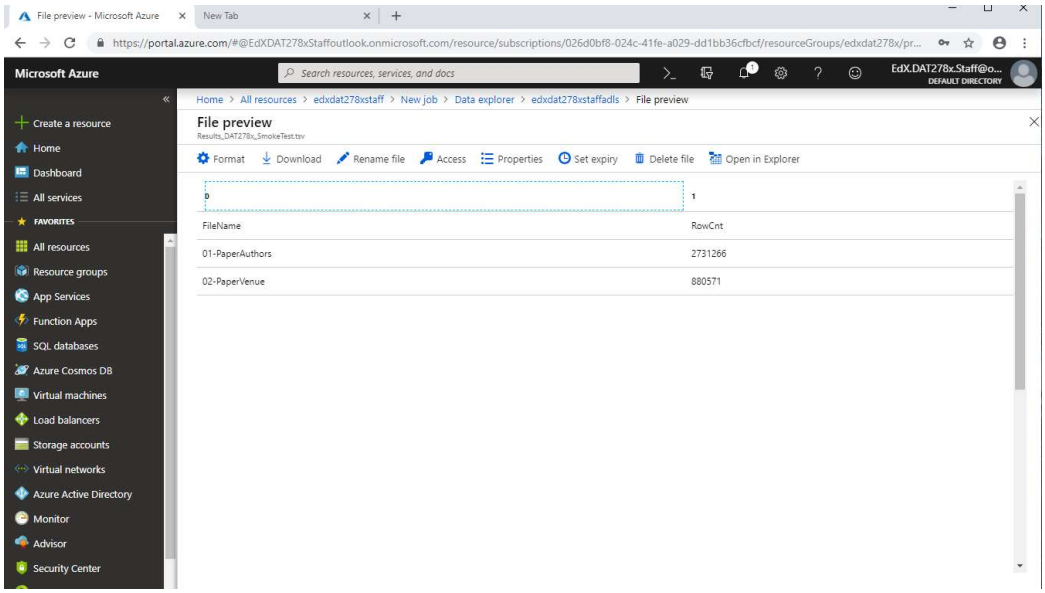


5.4 View the job progress as follow. After running the script successfully, the result file will be outputted to "**/course-kg/output/**" folder.





5.5 Click the output file “Results_DAT278x_SmokeTest.tsv”. Verify it contains the same contents as the following snapshot.



If you can follow above steps without issues and generate the same resulting contents for the smoke testing script, **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.