

ELG 5166 – Cloud Analytics

Assignment #3

Submitted to: Dr. Benjamin Eze

Submitted by: Group - 11

Name	ID	Email
Youssef Metwally	300267001	ymetw027@uottawa.ca
Hussien Khaled	300266718	hkhal023@uottawa.ca
Abdelrahman Basha	300266859	abash023@uottawa.ca
Khadija Hesham	300266888	khesh072@uottawa.ca

November 28, 2021

Personal Ethics & Academic Integrity Statement

Student name:	Youssef Metwally	Student ID:	300267001
Student Name:	Hussien Khaled	Student ID:	300266718
Student Name:	Abdelrahman Basha	Student ID:	300266859
Student Name:	Khadija Hesham	Student ID:	300266888

By typing in my name and student ID on this form and submitting it electronically, I am attesting to the fact that I have reviewed not only my work but the work of my team member, in its entirety.

I attest to the fact that my work in this project adheres to the fraud policies as outlined in the Academic Regulations in the University's Graduate Studies Calendar. I further attest that I have knowledge of and have respected the "Beware of Plagiarism" brochure for the university. To the best of my knowledge, I also believe that each of my group colleagues has also met the aforementioned requirements and regulations. I understand that if my group assignment is submitted without a completed copy of this Personal Work Statement from each group member, it will be interpreted by the school that the missing student(s) name is confirmation of non-participation of the aforementioned student(s) in the required work.

We, by typing in our names and student IDs on this form and submitting it electronically,

- warrant that the work submitted herein is our own group members' work and not the work of others
- acknowledge that we have read and understood the University Regulations on Academic Misconduct
- acknowledge that it is a breach of University Regulations to give or receive unauthorized and/or unacknowledged assistance on a graded piece of work

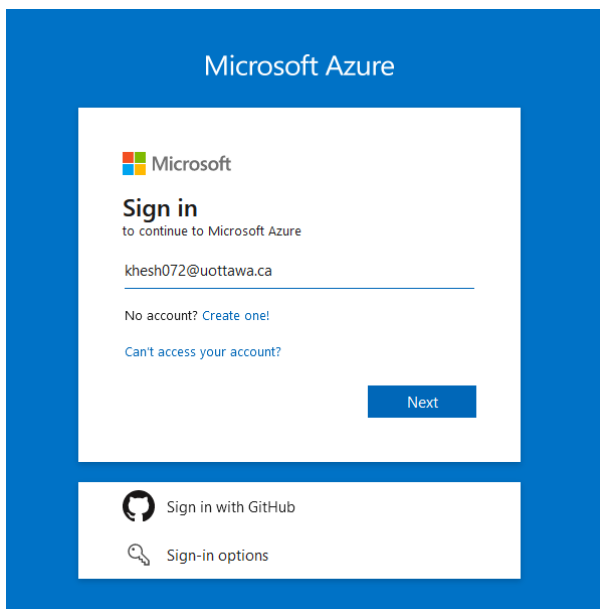
Part 1 - Event Hubs Analytics:

a) Top 20 zip codes where most bikes were rented from.

Part 1 – Event Hubs Analytics:

First of all, we set up our azure account.

I have signed in with my uOttawa account.



**Khadija Hesham**

kesh072@uottawa.ca

[View account](#)[Switch directory](#)

...




Sign in with a different account

A screenshot from my azure.

Microsoft Azure Search resources, services, and docs (G+/)


Welcome to Azure!

Don't have a subscription? Check out the following options.




Start with an Azure free trial
Get \$200 free credit toward Azure products and services, plus 12 months of popular [free services](#).

[Start](#) [Learn more](#)



Manage Azure Active Directory
Manage access, set smart policies, and enhance security with Azure Active Directory.

[View](#) [Learn more](#)



Access student benefits
Get free software, Azure credit, or access Azure Dev Tools for Teaching after you verify your academic status.

[Explore](#) [Learn more](#)

We click on the student tap to claim the students voucher.

Microsoft Azure Search resources, services, and docs (G+/)

Home >

Education | Overview

Are you a Student? Redeem the Azure for Students offer and get additional free benefits. →

Overview

Get started

Learning resources

Roles

Start building the future with Azure for Students!

Students are eligible for \$100 credit from Azure for Students.

[Claim your Azure credit now](#)

Identity Verification by phone



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

Country code

Egypt (+20)



Phone number

12 822 49616

Text me

Call me

Verification code

Verify code

Your profile



Country/Region ⓘ

Egypt



Choose the location that matches your billing address. You cannot change this selection later. If your country is not listed, the offer is not available in your region. [Learn More](#)

First name

Khadja

Last name

Hesham

Email address for important notifications ⓘ

khesham072@uottawa.ca

Phone

12 822 49616

Please do not enter country code in your phone number.

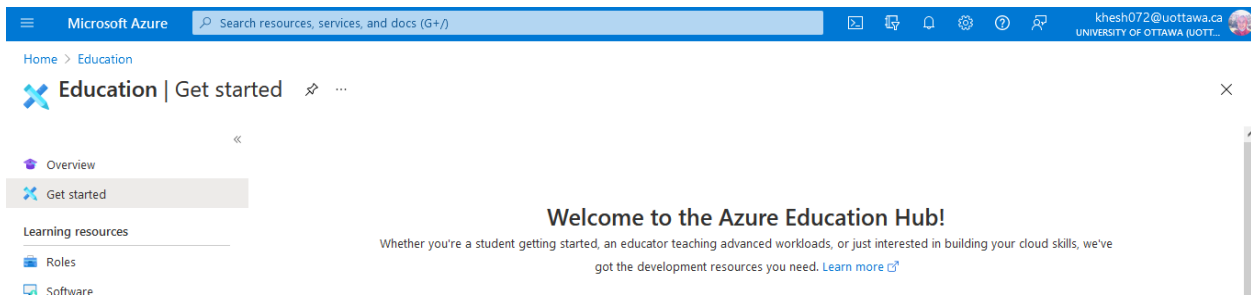
By proceeding you acknowledge that if you use your organization's email, your organization may have rights to access and manage your data and account. [Learn more](#)

- ☒ I agree to the [customer agreement](#) and [privacy agreement](#).
- ☐ I would like information, tips, and offers about Azure, including Azure Newsletter, Pricing updates, and other Microsoft products and services.
- ☐ I would like Microsoft to share my information with select partners so I can receive relevant information about their products and services.

Sign up

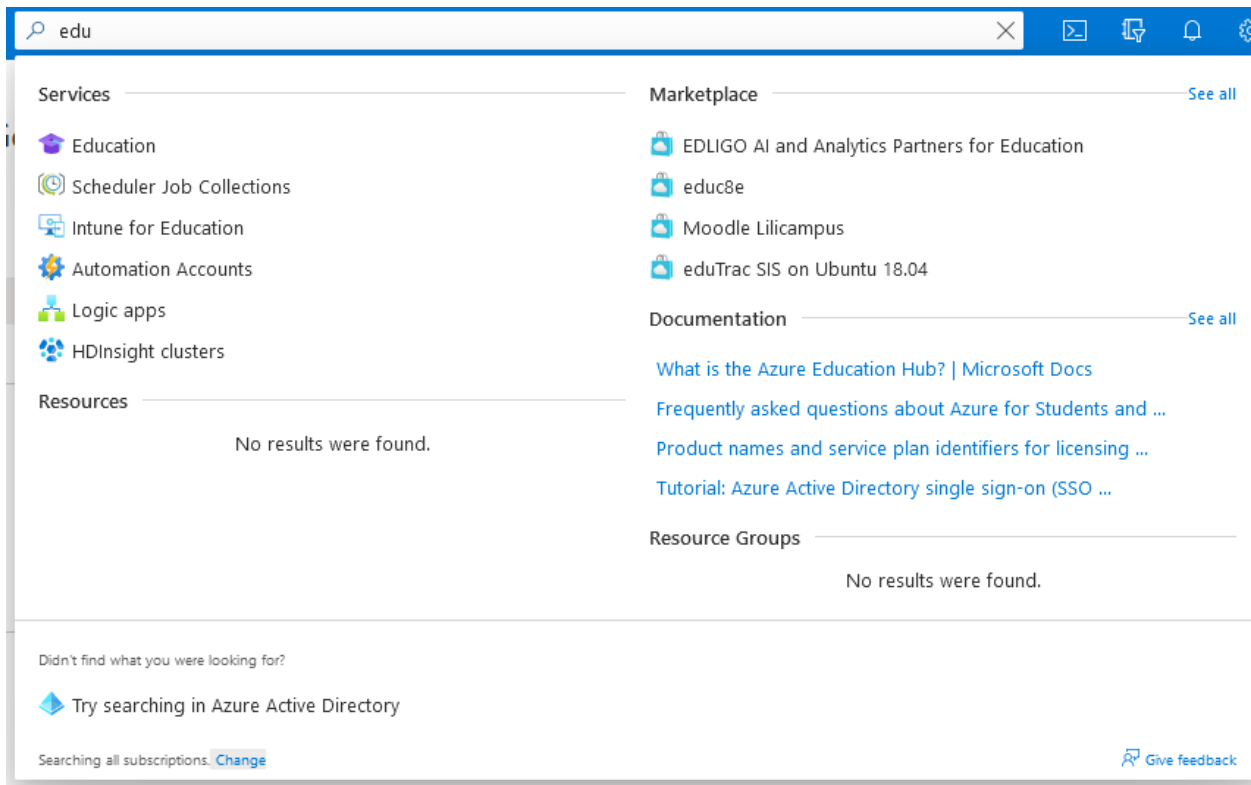
Azure for Students

Get \$100 in Azure credits and free access to popular cloud services plus developer tools like Visual Studio Code



Install Visual Studio:

We choose the education services.



We got the 100\$ voucher.

Home >

Education | Overview

Overview

Get started

Learning resources

- Roles
- Software
- Learning
- Templates

Need help?

- Support

Explore the capabilities of the cloud with Azure.


Easily set up your personal website, unlock the possibilities of AI or securely store your data in the cloud!

Click here to complete your student profile

Azure credits

Student offer: \$100
\$100 unused credit expires on November 27, 2022
0
\$100

Explore Azure roles [See all roles](#)



We search for Visual Studio Community 2019 in the software section in the menu at left.

Home > Education

Education | Software

Overview

Get started

Learning resources

- Roles
- Software
- Learning
- Templates

Need help?

- Support

Visual Studio Community 2019

Product category : All Operating System : All System type : 64 bit

Product language : English, Multilanguage

1 Item

Name ↑↓	Product category ↑↓	Operating System ↑↓	System type ↑↓	Language ↑↓
Visual Studio Community 2019 (v...	Developer Tools	Windows	64 bit	Multilanguage

And choose to download it.

Software

Education



Visual Studio Community 2019 (version 16.0)

A free, fully featured, and extensible solution for individual developers to create applications for Android, iOS, Windows, and the web.

Operating System

Windows

Product language

Multilanguage

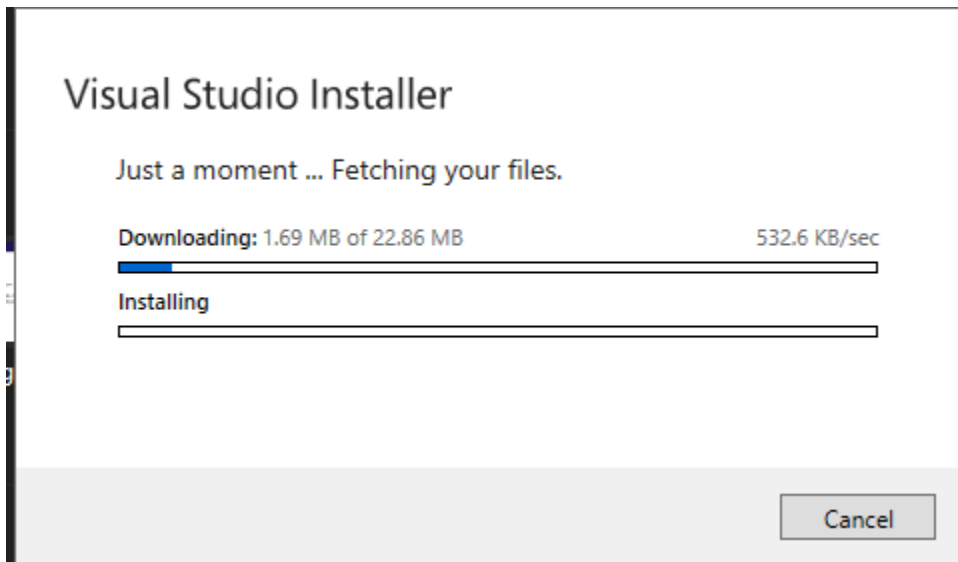
System

64 bit

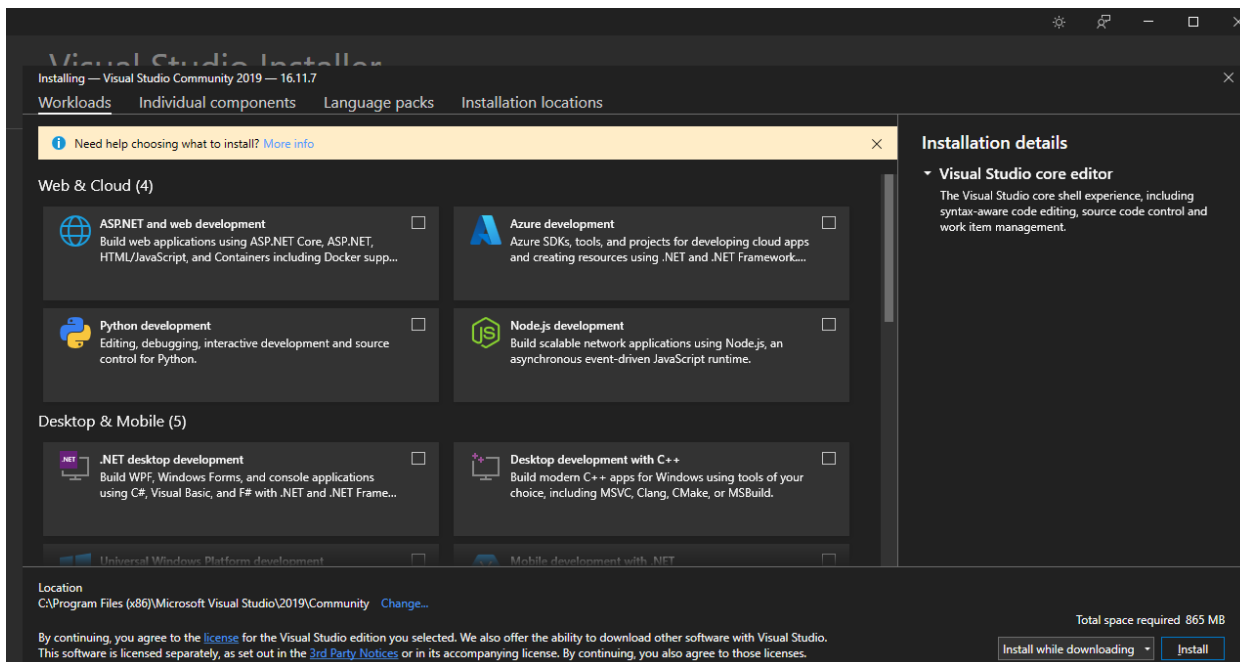
Download

Cancel

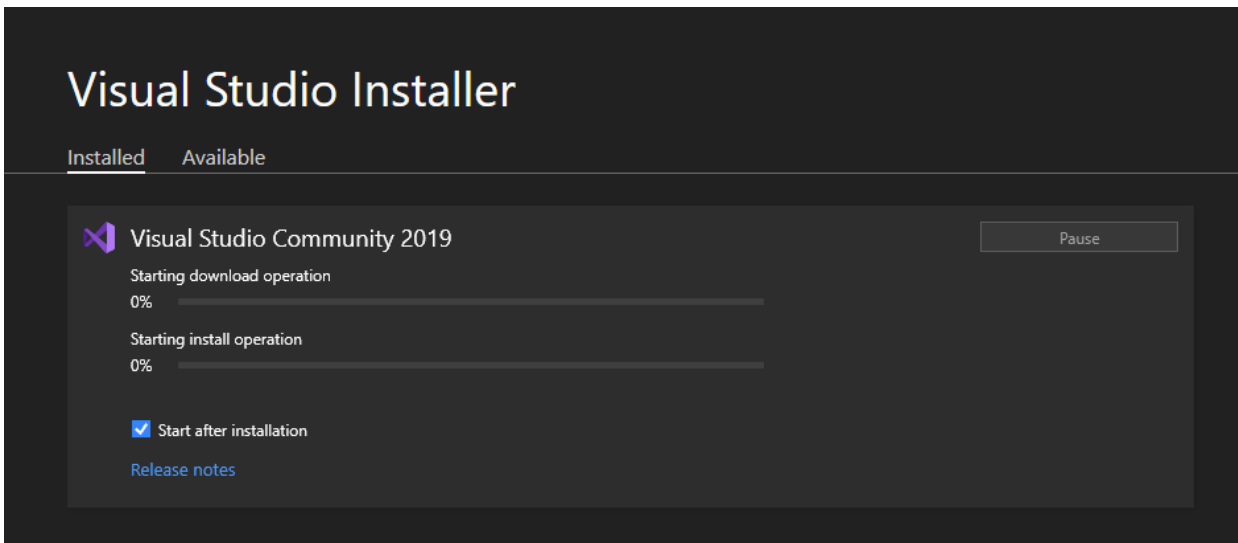
Download is started.



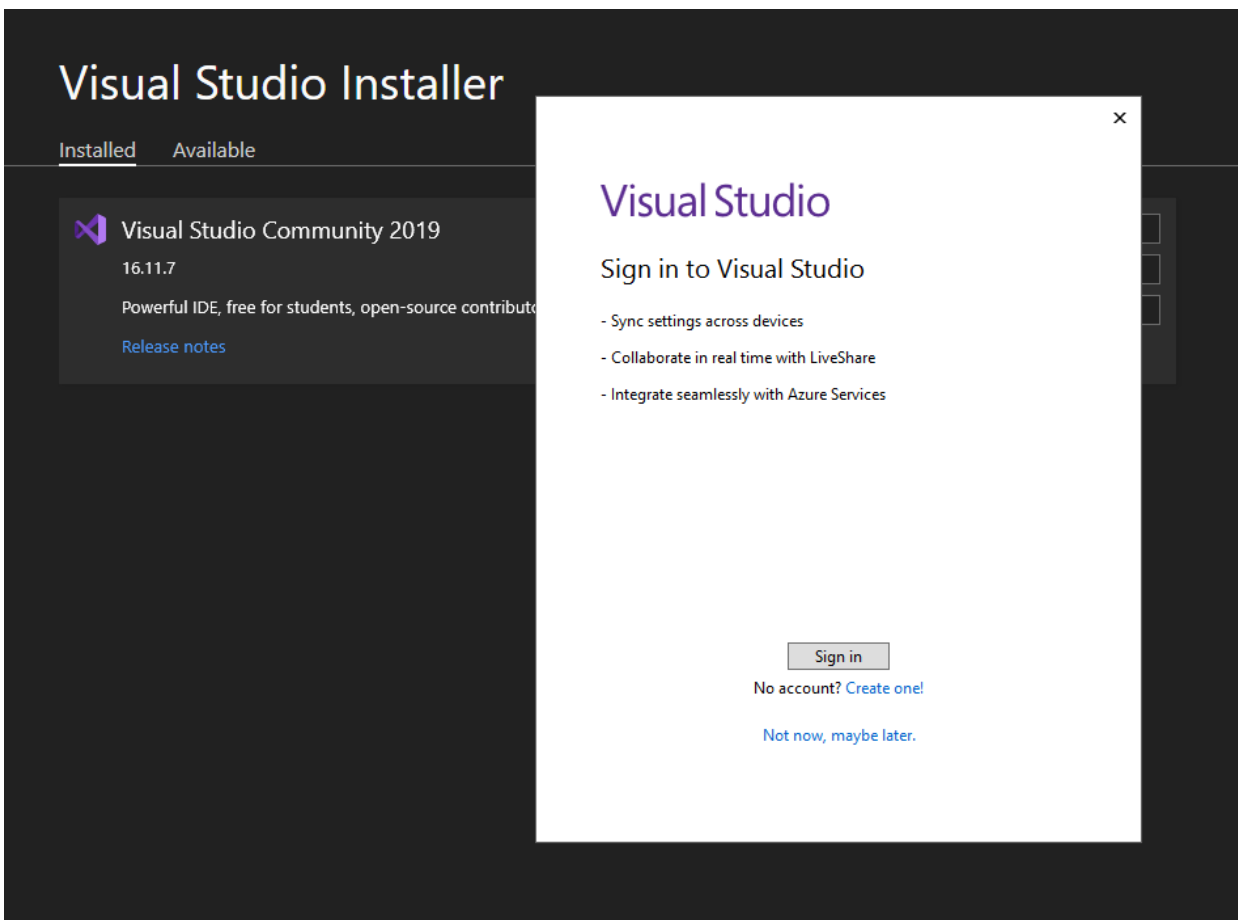
We start installing.



The installation is started.



Installation is complete, and now we are signing in.



We choose the uOttawa account.



Pick an account



Khadija Hesham
khesh072@uottawa.ca
Connected to Windows



Khadija Mahmoud Taha
sci.khadijamahmoud2018@alexu.edu.eg
Connected to Windows



Use another account

Back

We open up again the installer application to download tools for assignments.

Visual Studio Installer

Installed

Available



Visual Studio Community 2019

16.11.7

Powerful IDE, free for students, open-source contributors, and individuals

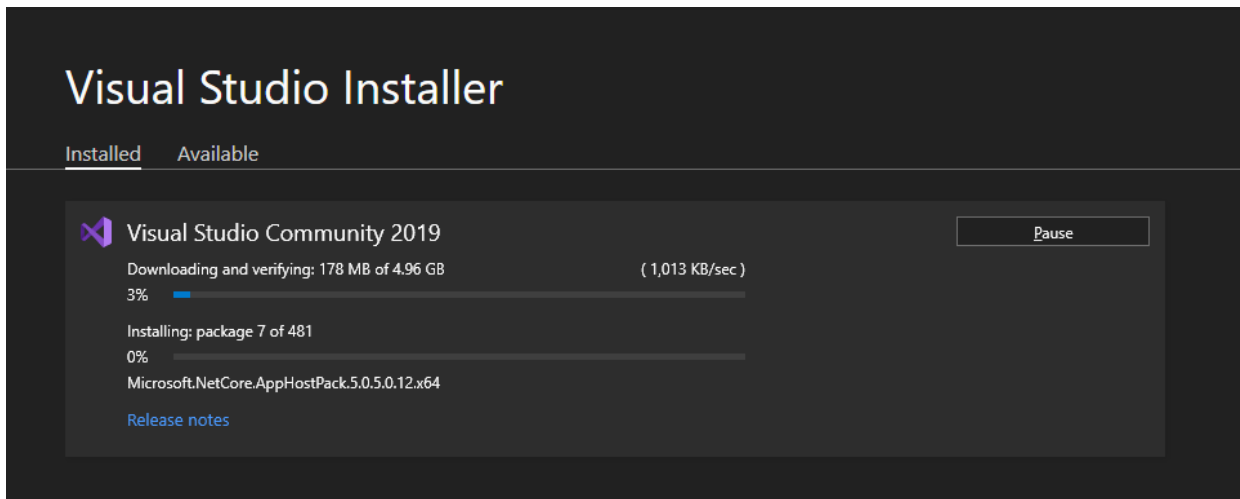
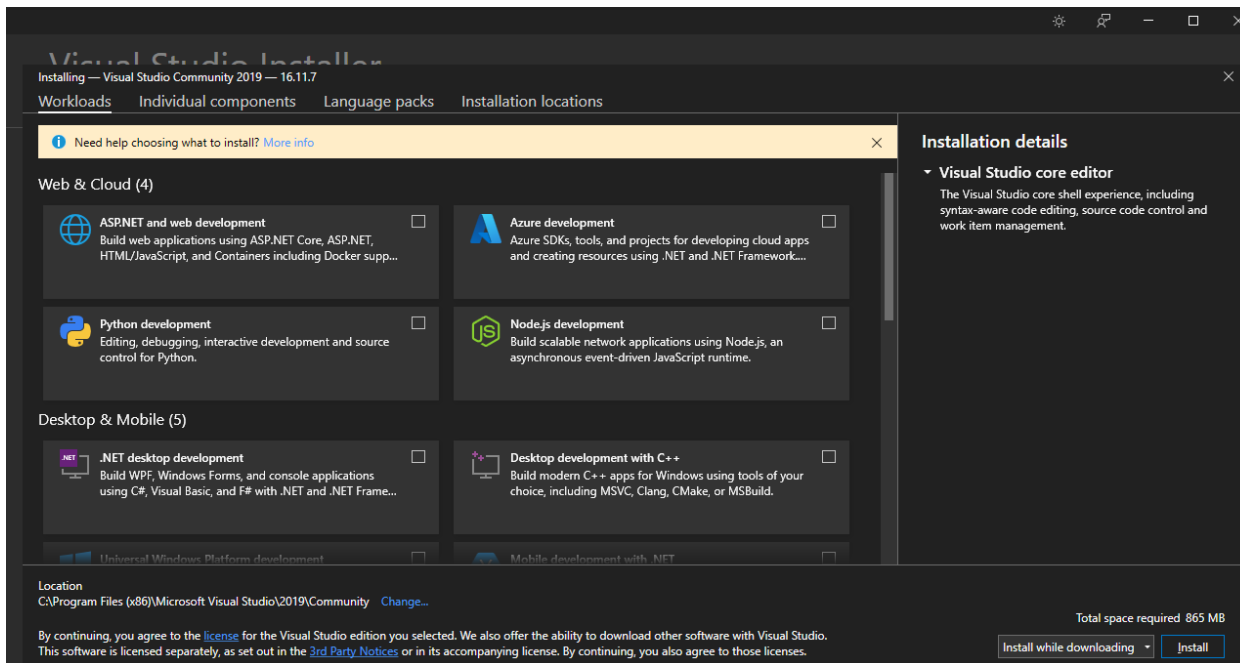
[Release notes](#)

Modify

Launch

More ▾

Then, we click on modify the program to install important workloads, we select some tools that you will use later such as Universal Windows Platform development, Azure Development, .NET desktop development, and Data storage and processing.



We click to start.

Visual Studio

Hello, Khadija Hesham



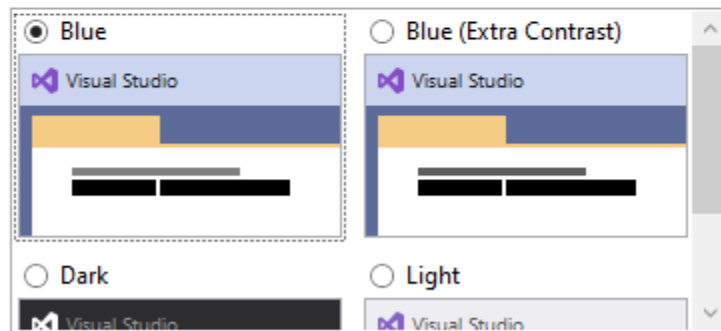
khesh072@uottawa.ca

[View your Visual Studio profile](#)

Start with a familiar environment

Development Settings: General

Choose your color theme



You can always change these settings later.

[Start Visual Studio](#)

Then, we create a new project to use U-SQL.

Visual Studio 2019

Open recent

As you use Visual Studio, any projects, folders, or files that you open will show up here for quick access.

You can pin anything that you open frequently so that it's always at the top of the list.

Get started



Clone a repository

Get code from an online repository like GitHub or Azure DevOps



Open a project or solution

Open a local Visual Studio project or .sln file



Open a local folder

Navigate and edit code within any folder



Create a new project

Choose a project template with code scaffolding to get started

[Continue without code →](#)

We search for U-SQL and then click next.

Create a new project

Recent project templates

A list of your recently accessed templates will be displayed here.



[Clear all](#)

All languages

All platforms

All project types



U-SQL Project

A project for creating U-SQL application

Query Language Azure Cloud



Class Library (For U-SQL Application)

A project for creating a C# class library(.dll) that can run on U-SQL.

Query Language Azure Cloud



U-SQL C# UDO Unit Test Project

A project that contains unit tests for U-SQL C# UDO.

Query Language Azure Cloud Test



U-SQL Sample Application

A project for creating a U-SQL Sample Project.

Query Language Azure Cloud



U-SQL C# UDO Unit Test Sample Project

A project that contains unit tests and UDO samples for U-SQL C# UDO

Query Language Azure Cloud Test



U-SQL Database Project

A project for creating a U-SQL database deployment system

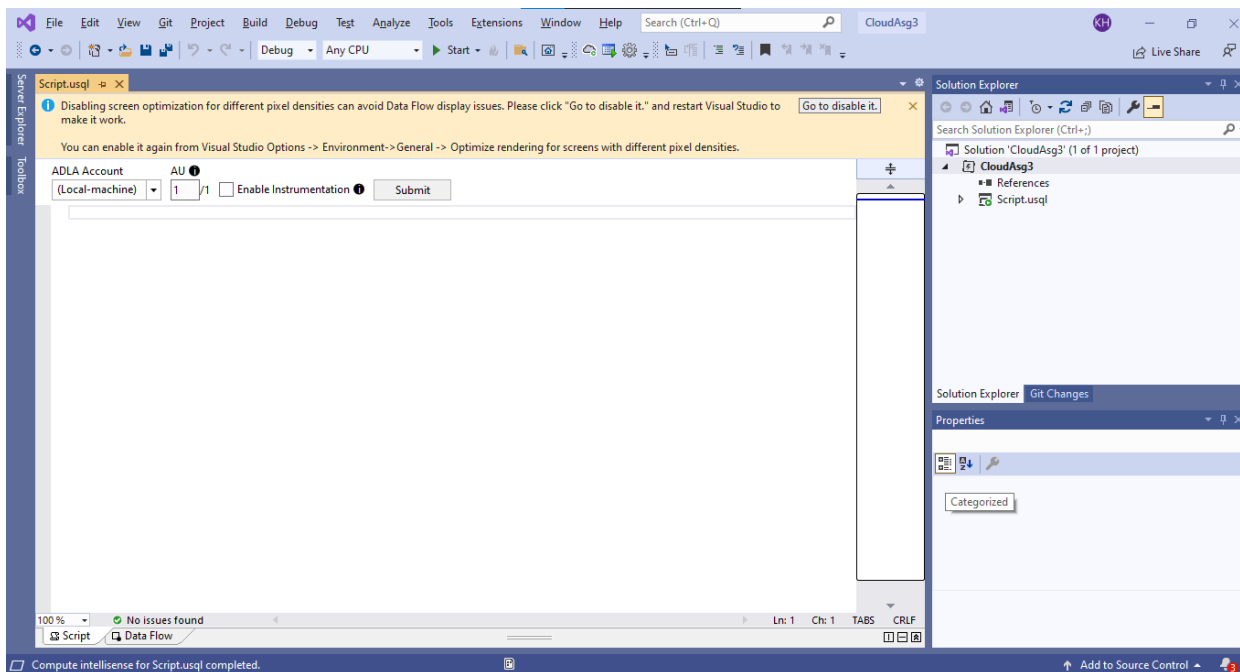
[Back](#)

[Next](#)

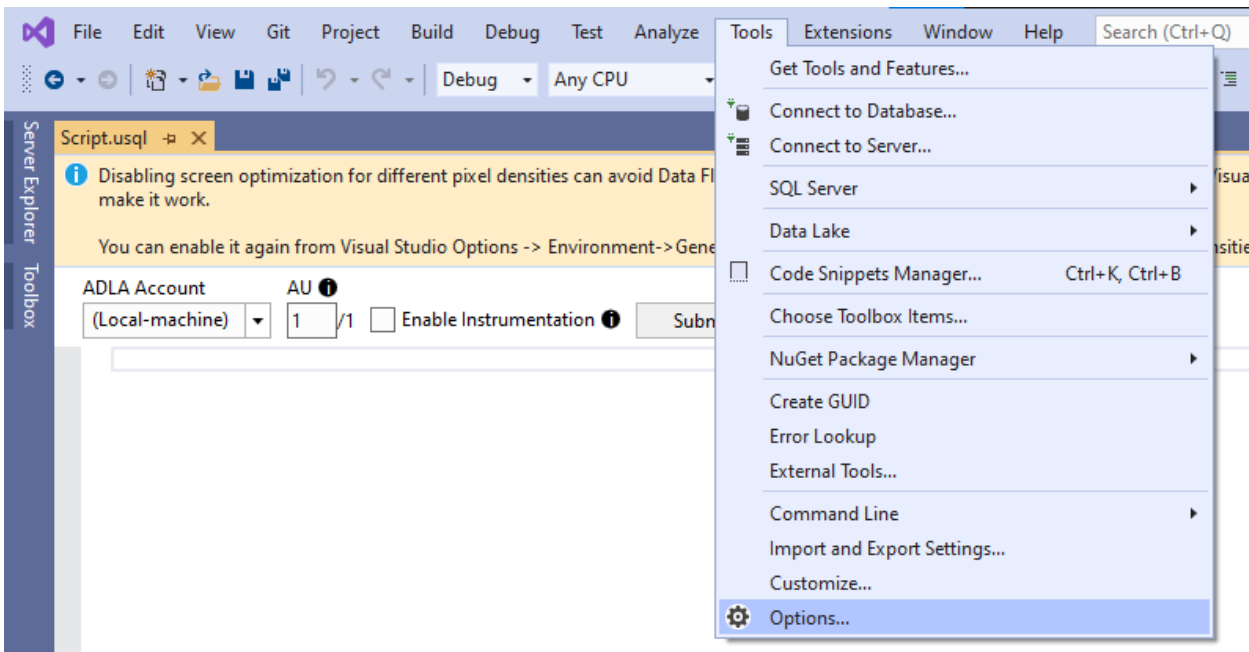
Specify the Visual Studio project name, location and solution, then we click create.

The screenshot shows the 'Configure your new project' dialog in Visual Studio. The title bar indicates the window is titled 'CloudAsg3'. The dialog has a header with tabs: 'U-SQL Project', 'Query Language', 'Azure', and 'Cloud'. The 'U-SQL Project' tab is selected. Below the tabs, there are four main sections: 'Project name' with a text box containing 'CloudAsg3'; 'Location' with a dropdown menu showing 'D:\Cloud-A3\' and a 'Browse...' button; 'Solution name' with a text box containing 'CloudAsg3' and an information icon; and 'Framework' with a dropdown menu showing '.NET Framework 4.7.2'. Below the 'Solution name' section, there is a checkbox labeled 'Place solution and project in the same directory' which is currently unchecked. At the bottom right of the dialog, there are two buttons: 'Back' and 'Create'.

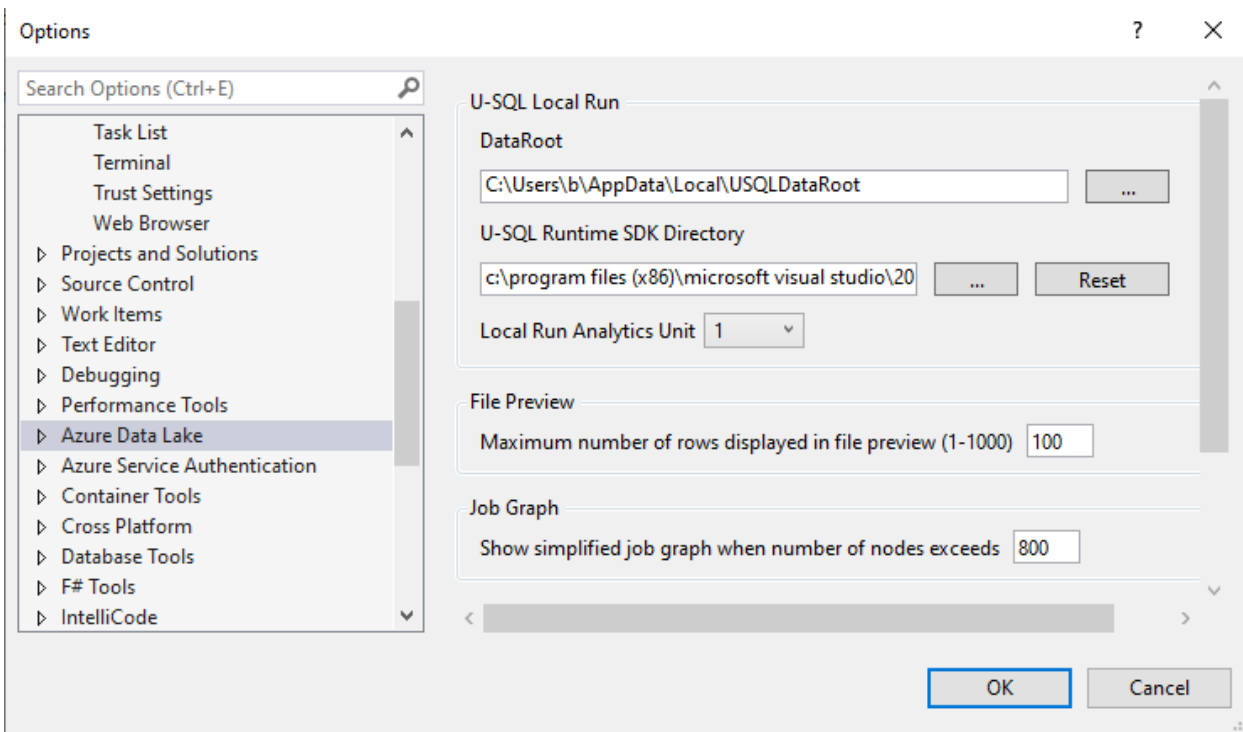
We wait until the project is created, after that we can see this screen.



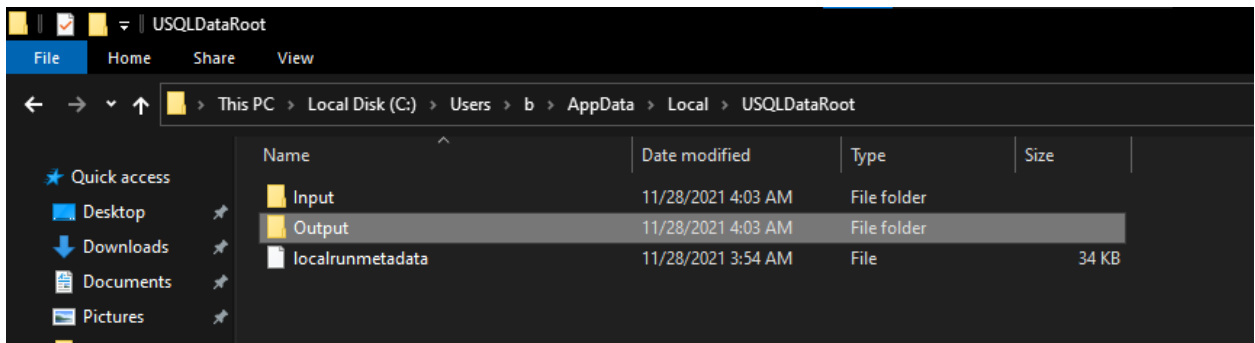
Before we write the script, we need to set the input and output folders path in Visual Studio.



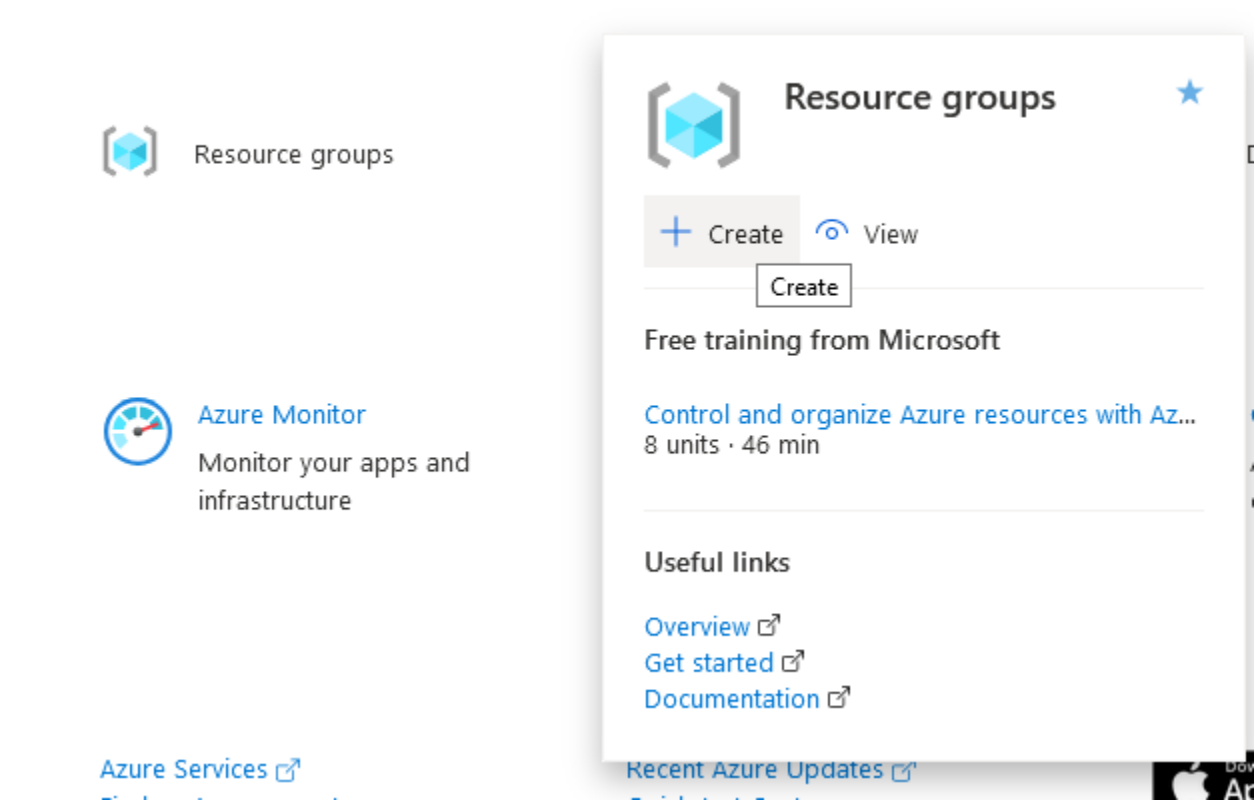
Now, browse to the data root directory and create two folders for input files and output files.



Now, browse to the data root directory and create two folders for input files and output files.




Create resource group







Create a resource group ...



[Basics](#) [Tags](#) [Review + create](#)

Resource group - A container that holds related resources for an Azure solution. The resource group can include all the resources for the solution, or only those resources that you want to manage as a group. You decide how you want to allocate resources to resource groups based on what makes the most sense for your organization. [Learn more](#) 

Project details


Subscription * 	<div>Azure for Students </div>
Resource group * 	<div>ELG5166_Resources </div>

Resource details

Region * 	<div>(Canada) Canada Central </div>
--	--

[Review + create](#)[< Previous](#)[Next : Tags >](#)

Notifications

[More events in the activity log →](#)[Dismiss all](#) 

Resource group created



Creating resource group 'ELG5166_Resources' in subscription 'Azure for Students' succeeded.

[Go to resource group](#)[Pin to dashboard](#)[Go to resource group](#)

a few seconds ago

Microsoft Azure Search resources, services, and docs (G+/J)

Home > ELG5166_Resources Resource group

Search (Ctrl+/) < + Create Edit columns Delete resource group Refresh Export to CSV Open query Assign tags Move Delete

Overview

- Activity log
- Access control (IAM)
- Tags
- Resource visualizer
- Events

Settings

- Deployments
- Security
- Policies
- Properties
- Locks
- Cost Management
- Cost analysis

Essentials

Subscription (Move): Azure for Students Deployments: No deployments

Subscription ID: 330d280a-9166-4af8-8912-c6ede2ed739e Location: Canada Central

Tags (Edit): Click here to add tags

Resources Recommendations

Filter for any field... Type == all Location == all Add filter

Showing 0 to 0 of 0 records. Show hidden types No grouping List view

Name ↑↓ Type ↑↓ Location ↑↓

No resources match your filters

Try changing or clearing your filters.

2 Creating the Event Hubs

eve

Services See all Marketplace See all

Event Hubs

- Event Grid Domains
- Event Grid Subscriptions
- Event Grid Topics
- Event Hubs Clusters
- Event Grid Partner Namespaces
- Event Grid Partner Registrations
- Event Grid Partner Topics
- Event Grid System Topics
- Time Series Insights event sources

Resources

No results were found.

EventTracker SIEM

Everlasting Fairytale QuickStart

Lansera Virtual Events Platform

E.F. Dream

Documentation See all

Perform network intrusion detection with open source tools ...

Ev3-series and Esv3-series - Azure Virtual Machines ...

Elevate access to manage all Azure subscriptions and ...

Site-to-Site VPN connections over ExpressRoute private ...

Resource Groups

No results were found.

Didn't find what you were looking for?

Try searching in Azure Active Directory

Searching all subscriptions. Change Give feedback

Click the “Create event hubs namespace” to create a namespace for your event hubs



No event hubs namespaces to display

Try changing or clearing your filters.

[Create event hubs namespace](#)


[Learn](#) [Create event hubs namespace](#)

- Provide a name for the namespace, for example, the namespace is ELG5166.

Microsoft Azure

Search resources, services, and docs (G+)

[Home](#) > [Event Hubs](#) >

 **Create Namespace** ...

Event Hubs

Subscription *

Azure for Students

Resource group *

ELG5166_Resources

[Create new](#)

Instance Details

Enter required settings for this namespace, including a price tier and configuring the number of units (capacity).


Namespace name *

ELG5166

.servicebus.windows.net

Location *

Canada Central

 The region selected supports Availability zones. Your namespace will have Availability Zones enabled. [Learn more.](#)

Pricing tier ([View full pricing details](#)) *

Basic (1 Consumer group, 100 Brokered connections)

Throughput Units *

1

[Review + create](#)

[< Previous](#)


[Next: Tags >](#)

After validation, we click on create.

Go back one page (Alt+Left Arrow)
Right-click or pull down to show history

Search resources, services, and docs (G+)

Home > Event Hubs >



Create Namespace

...
Event Hubs

Validation succeeded.

Basics Tags Review + create

Event Hubs Namespace
by Microsoft

Basics

Namespace name	ELG5166
Subscription	Azure for Students
Resource group	ELG5166_Resources
Location	Canada Central
Pricing tier	Basic
Throughput Units	1
Availability Zones (Zone Redundancy)	Enabled

Create


< Previous

Next >

Deployment is in progress.

Microsoft Azure Search resources, services, and docs (G+)

Home >

 **ELG5166 | Overview** ...

Deployment

Search (Ctrl+J) « Delete Cancel Redeploy Refresh

Overview

Inputs

Outputs

Template

Deployment is in progress

Deployment name: ELG5166
Subscription: [Azure for Students](#)
Resource group: [ELG5166_Resources](#)

Start time: 11/28/2021, 7:47:47 AM
Correlation ID: ded9d847-5e5a-423c-969f-cd8a65a6f566

Deployment details [\(Download\)](#)

Resource	Type	Status	Operation details
No results.			

Microsoft Azure Search resources, services, and docs (G+/)

Home > **ELG5166** | Overview Deployment

Search (Ctrl+/) Delete Cancel Redeploy Refresh

Overview

- Inputs
- Outputs
- Template

Your deployment is complete

Deployment name: ELG5166
Subscription: [Azure for Students](#)
Resource group: [ELG5166_Resources](#)

Start time: 11/28/2021, 7:47:47 AM
Correlation ID: ded9d847-5e5a-423c-969f-cd8a65a6f566

Deployment details (Download)

Next steps

[Go to resource](#)

Microsoft Defender for Cloud
Secure your apps and infrastructure
[Go to Azure security center >](#)

Free Microsoft tutorials
Start learning today >

Work with an expert
Azure experts are service provider partners

we have set up the Namespace for the Event Hubs, we can proceed to create the Event Hubs. We click on go to resource.

Microsoft Azure Search resources, services, and docs (G+/)

Home > ELG5166 > **ELG5166** Event Hubs Namespace

Search (Ctrl+/) + Event Hub Delete Refresh

Overview

- Activity log
- Access control (IAM)
- Tags
- Diagnose and solve problems
- Events
- Settings
- Shared access policies
- Scale
- Geo-Recovery
- Encryption
- Properties
- Locks

Essentials

Resource group (Move): [ELG5166_Resources](#)
Status: Active
Location: Canada Central
Subscription (Move): [Azure for Students](#)
Subscription ID: 330d280a-9166-4af8-8912-c6ede2ed739e
Host name: ELG5166.servicebus.windows.net

Created: Sunday, November 28, 2021, 07:47:53 GMT+2
Updated: Sunday, November 28, 2021, 07:48:46 GMT+2
Zone Redundancy: Enabled
Pricing tier: Basic
Throughput Units: 1 unit
Auto-inflate throughput ...: Not Supported
Local Authentication: Enabled

Tags (Edit): [Click here to add tags](#)

NAMESPACE CONTENTS: 0 EVENT HUBS
KAFKA SURFACE: NOT SUPPORTED
ZONE REDUNDANCY: ENABLED

Show data for the last: 1 hour 6 hours 12 hours 1 day 7 days 30 days

Requests: 100 90
Messages: 100 90
Throughput: 100B 90B

Choose event hubs then click on create event hub.

Home > ELG5166 > ELG5166

ELG5166 | Event Hubs Event Hubs Namespace

Search (Ctrl+/) + Event Hub Refresh

Search to filter items...

Name	Status	Message Retention	Partition Count
No Event Hubs yet.			

Entities

Event Hubs

Monitoring

We type the name of the Event Hub. In this case, the name is Bike_Data.

Home > ELG5166 > ELG5166 >

Create Event Hub ...

Event Hubs

Name * ⓘ

Bike_Data ✓

Partition Count ⓘ



2

Message Retention ⓘ



1

Capture ⓘ

On

Off

Create

Create

Event hub object is active.

Home > ELG5166 > ELG5166

ELG5166 | Event Hubs ...

Event Hubs Namespace

Search (Ctrl+/)

+ Event Hub Refresh

Locks

Entities

Event Hubs

Monitoring

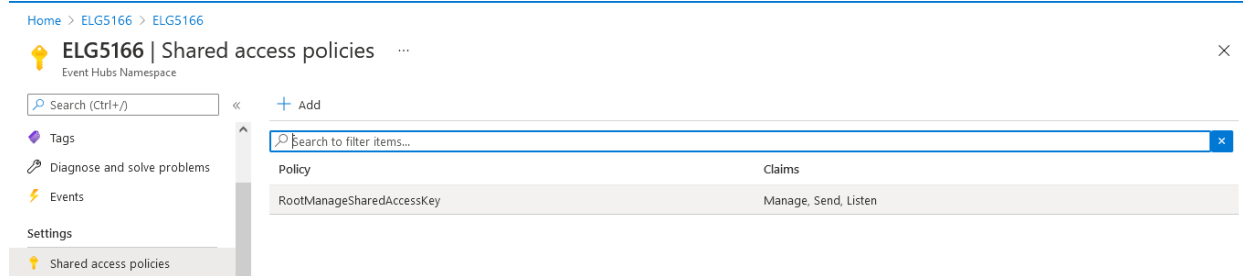
Alerts

Search to filter items...

Name	Status	Message Retention	Partition Count
bike_data	Active	1 day	2

To create a connection string. This is done using the event hub namespace shared access policies. The shared access policies can be found under the Settings category of the left menu.

Click on the RootManageSharedAccessKey



We will use the Connection String-Primary Key for creating the Event Hubs request generator.

SAS Policy: RootManageShare... ✕

Save ✕ Discard Delete ...

☒ Manage

☒ Send

☒ Listen

Primary key

1JSgY8Tdy0/dpkDQ5Jb1P8kmc1Y+0PcXObqdS5eYYzY=

Secondary key

13XnWGoYcBYKR4n5J2XhOq9W17tq3dDizMZCua1K3uU=

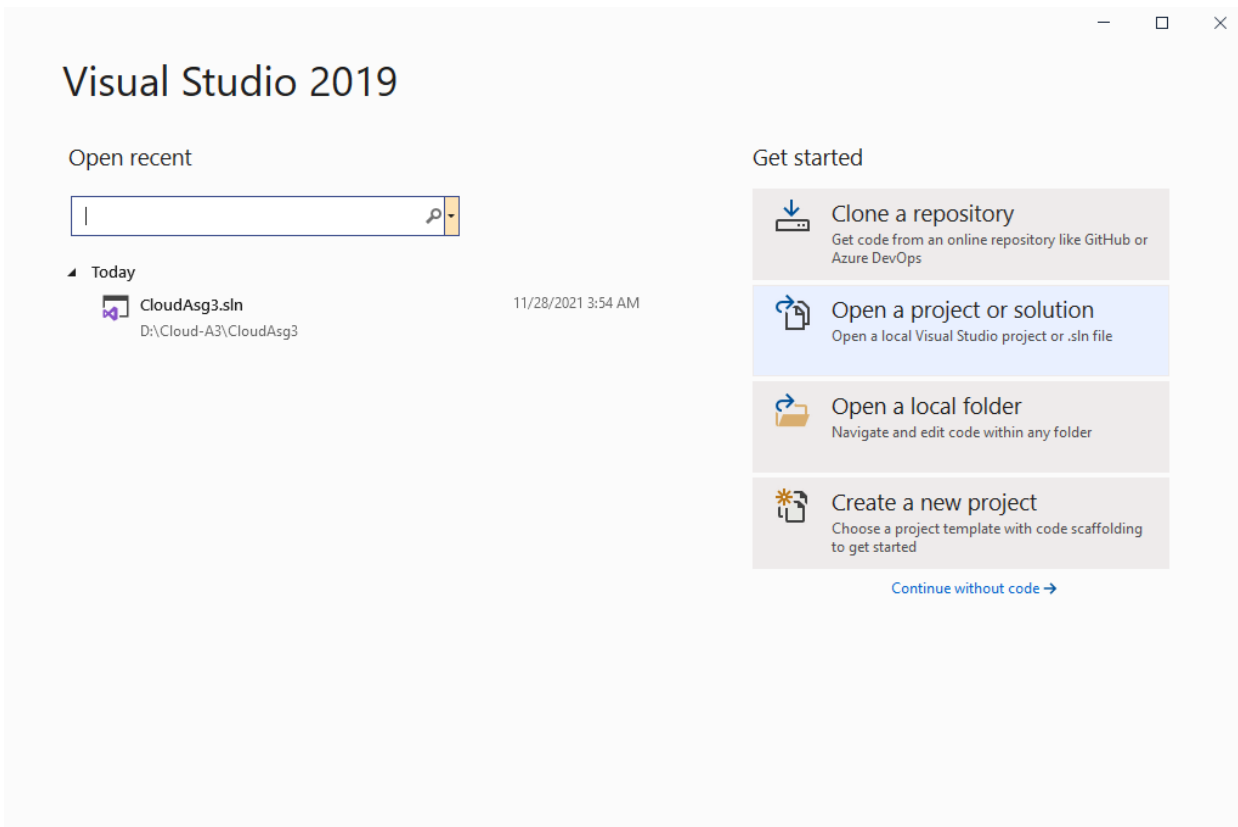
Connection string–primary key

Endpoint=sb://elg5166.servicebus.windows.net;/SharedAcce ...

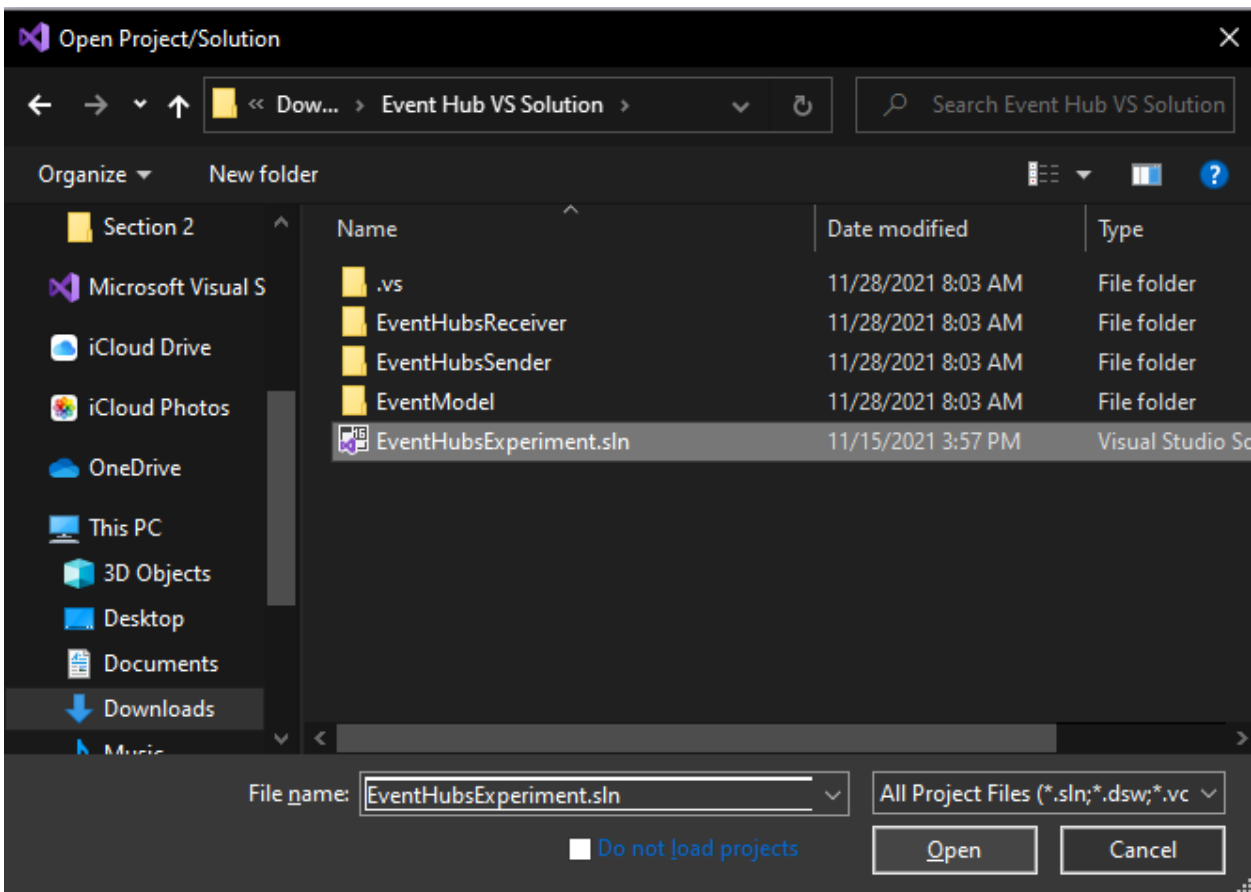
Connection string–secondary key

Endpoint=sb://elg5166.servicebus.windows.net;/SharedAcce ...

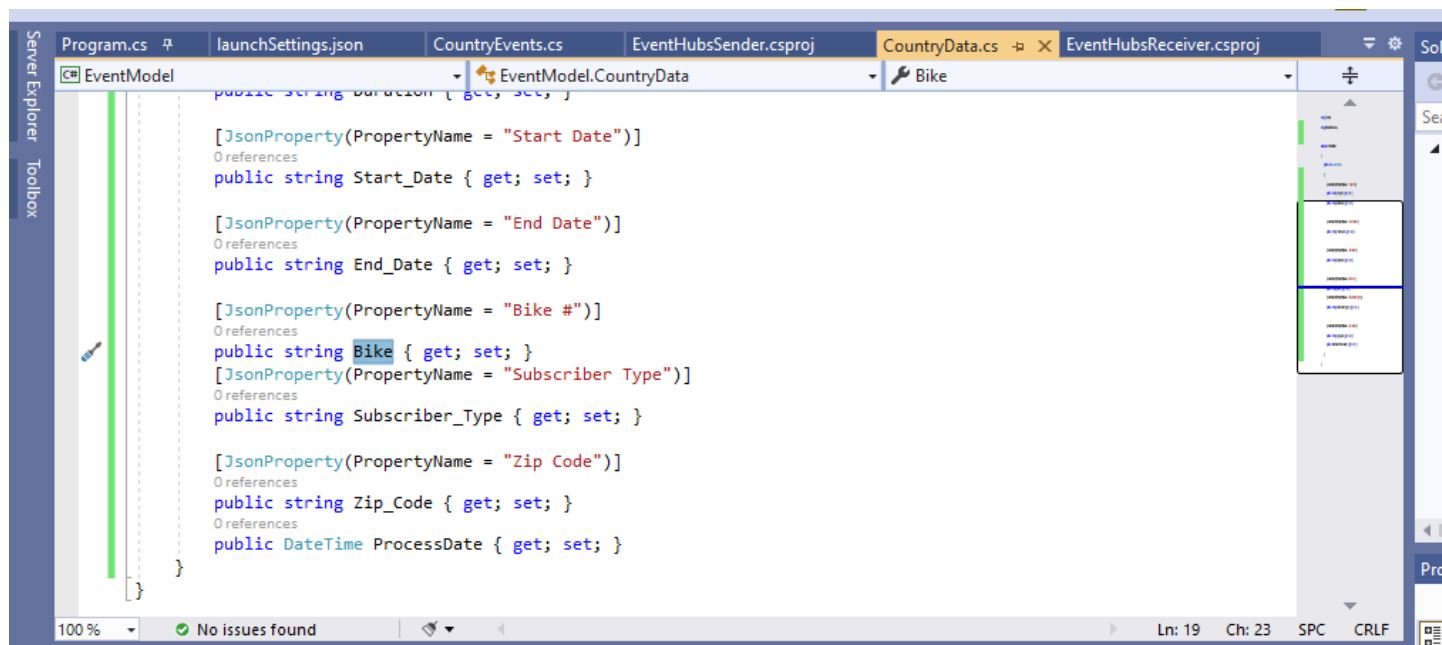
Open Visual Studio and choose to open a solution.



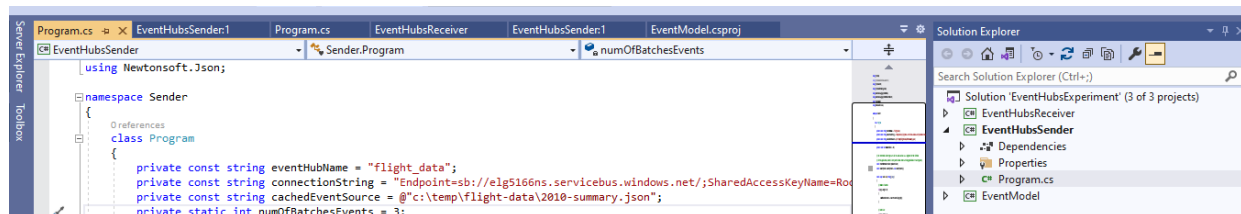
We choose the lab solution.



First of all, we edit the data schema with our bike data schema.



Go to event hub sender project and choose the program.cs file.



SAS Policy: RootManageShare... ×


 Save  Discard  Delete ...

☒ Manage


☐ Send

☐ Listen

Primary key

1JSgY8Tdy0/dpkDQ5Jb1P8kmc1Y+0PcXObqdS5eYYzY= 


Secondary key

13XnWGoYcBYKR4n5J2XhOq9W17tq3dDizMZCua1K3uU= 

Connection string–primary key

Endpoint=sb://elg5166.servicebus.windows.net/;SharedAcce ... 

Connection string–secondary key

Endpoint=sb://elg5166.servicebus.windows.net/;SharedAcce ... 

Copy the connection string primary key and paste it into the code.

```
connectionString = "Endpoint=sb://elg5166.servicebus.windows.net/;SharedAccessKeyName=Ro...
```

Edit the event hub name.

 Search to filter items...

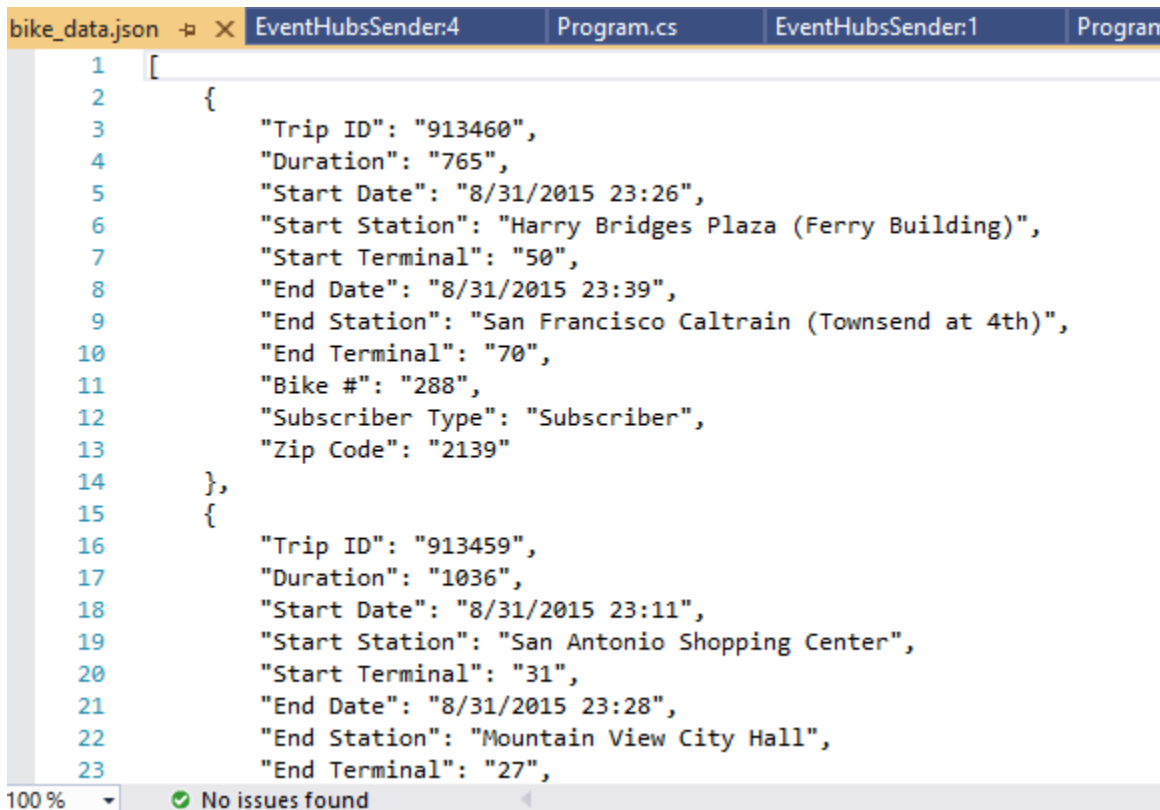
Name	Status	Message Retention	Partition Count
bike_data	Active	1 day	2

```
private const string eventHubName = "bike_data";
```

Edit the path of the data by writing our data path.

```
private const string cachedEventSource = @"D:\Cloud-A3\bike_data";
```

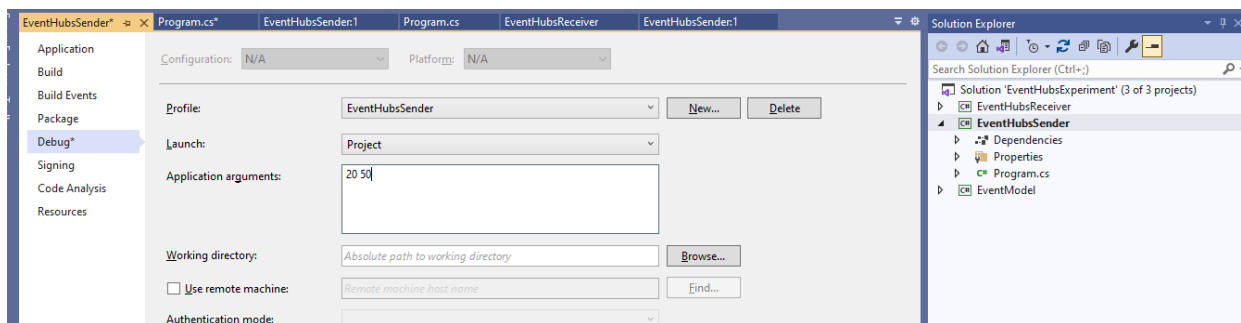
We can see our data.



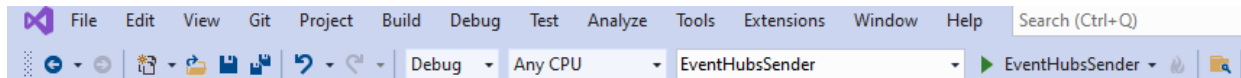
```
bike_data.json  EventHubsSender:4  Program.cs  EventHubsSender:1  Program
1  [
2      {
3          "Trip ID": "913460",
4          "Duration": "765",
5          "Start Date": "8/31/2015 23:26",
6          "Start Station": "Harry Bridges Plaza (Ferry Building)",
7          "Start Terminal": "50",
8          "End Date": "8/31/2015 23:39",
9          "End Station": "San Francisco Caltrain (Townsend at 4th)",
10         "End Terminal": "70",
11         "Bike #": "288",
12         "Subscriber Type": "Subscriber",
13         "Zip Code": "2139"
14     },
15     {
16         "Trip ID": "913459",
17         "Duration": "1036",
18         "Start Date": "8/31/2015 23:11",
19         "Start Station": "San Antonio Shopping Center",
20         "Start Terminal": "31",
21         "End Date": "8/31/2015 23:28",
22         "End Station": "Mountain View City Hall",
23         "End Terminal": "27",
```

Broadcast bike rental events in batches of 20 trip entries and publish at least 50 events at 2-second intervals per batch.

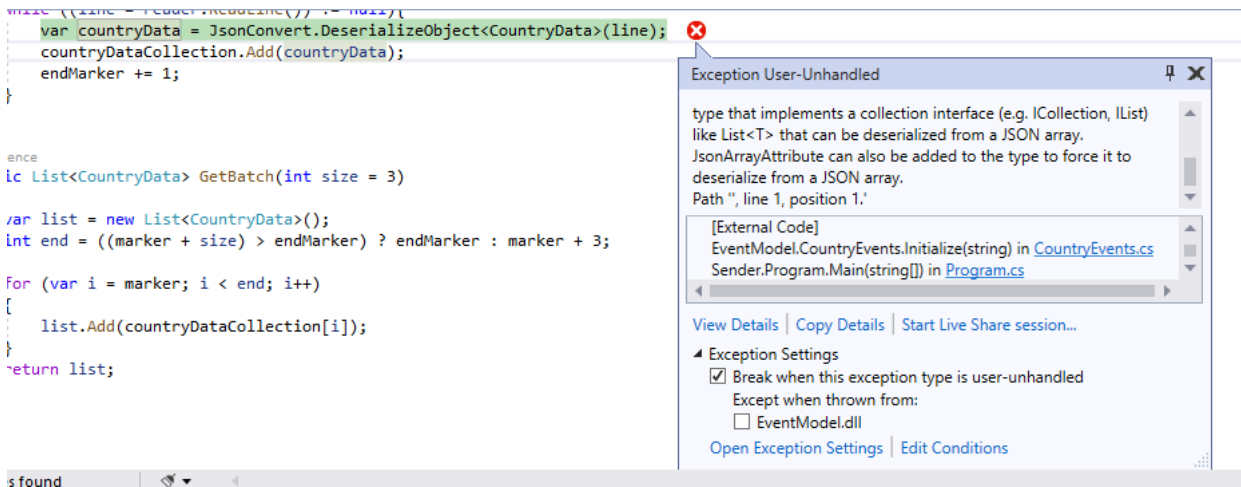
We highlight the event hub sender and right click on the properties, then we write a number of batches of 20 and let the number of events be 50.



Then we run the application

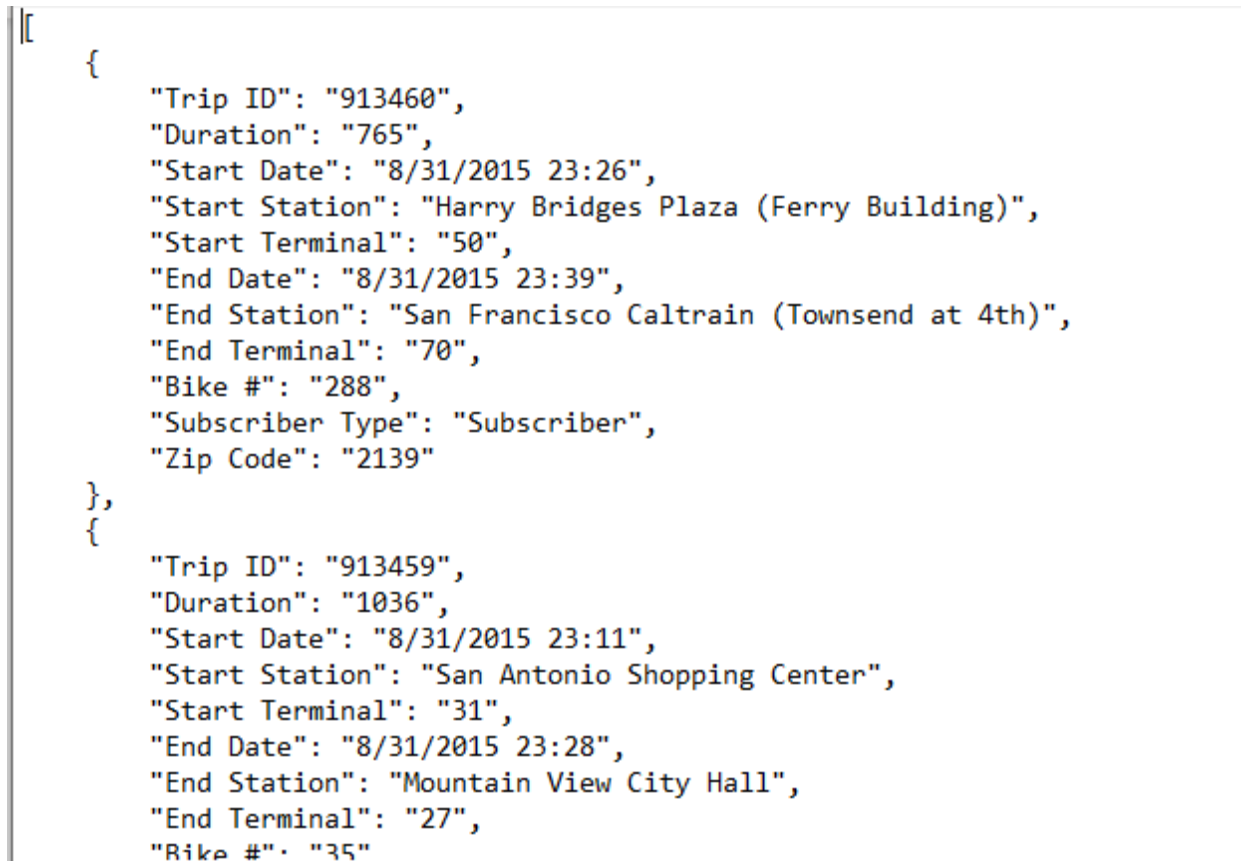


We got an error



The data needs to be formatted in this way, to remove the “,” from each observation , remove the brackets “[“ in the beginning and “]” at last, remove empty json objects, and to reformat the data with removing the white space by the help of this website: <http://jsonviewer.stack.hu/>.

Data before formatting:

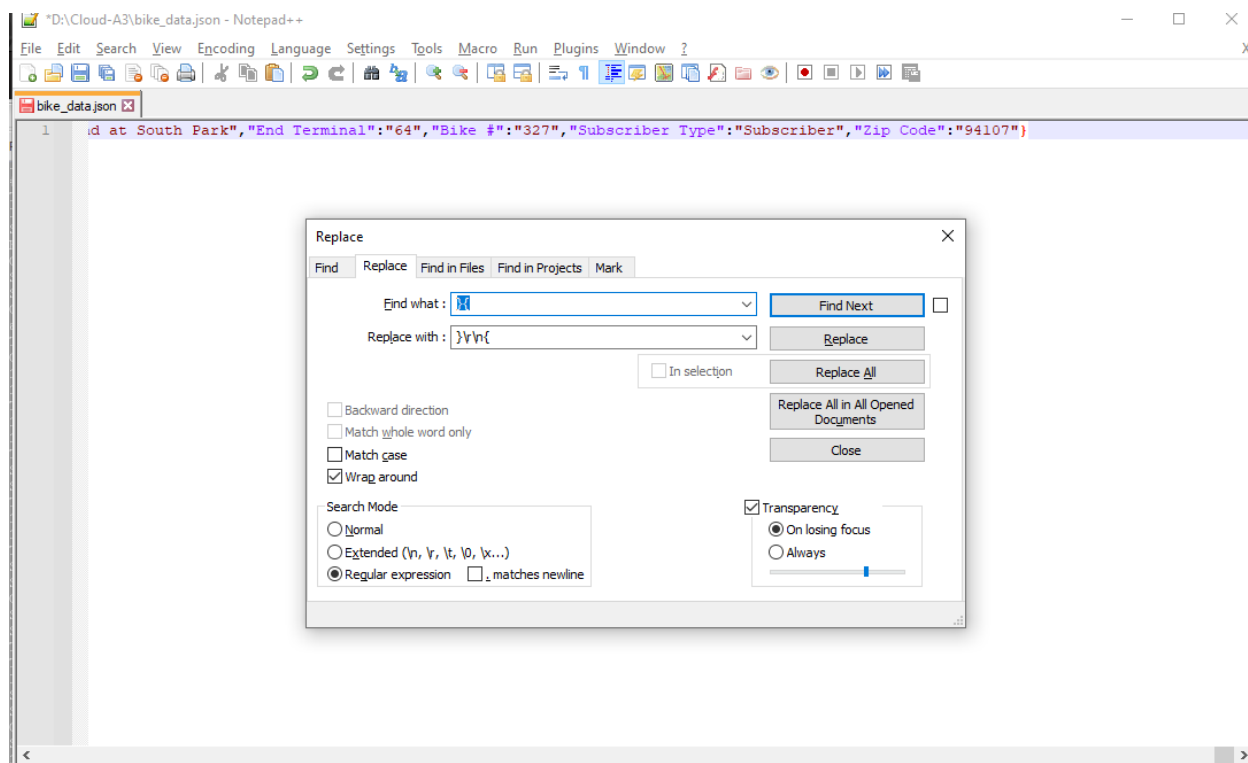


```

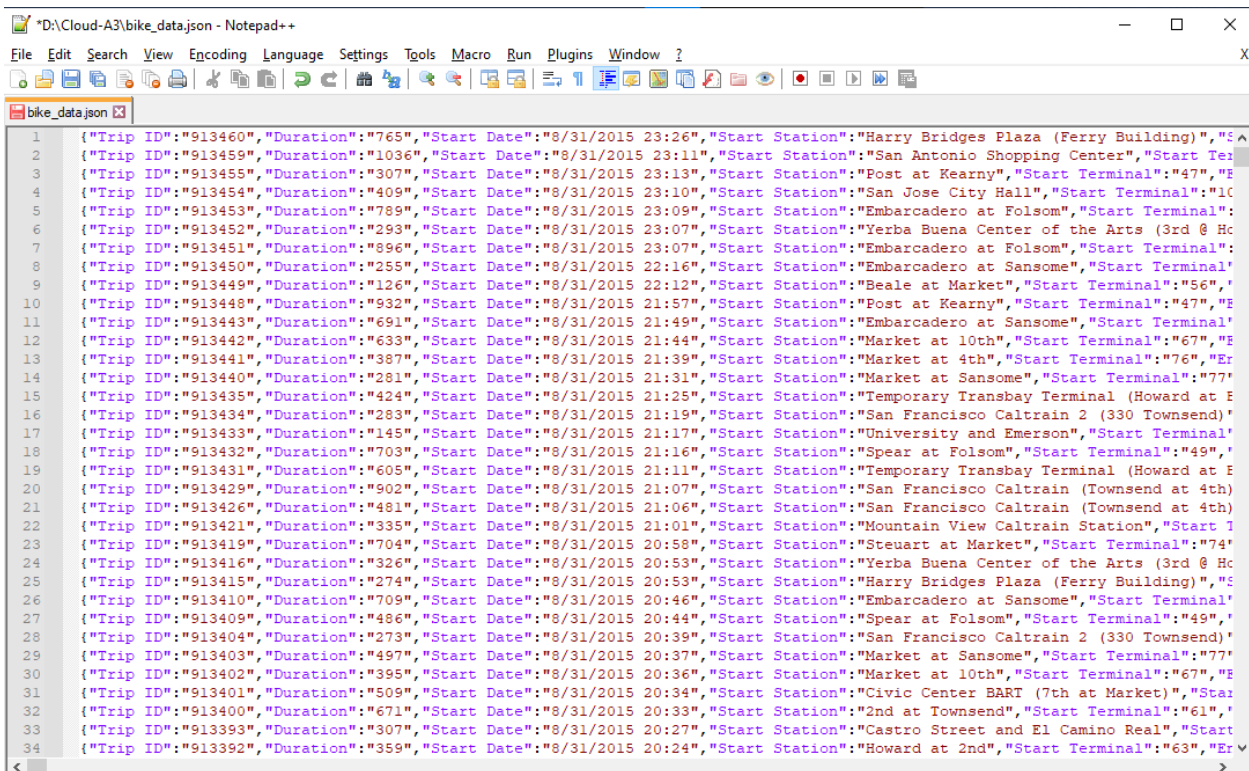
{
  "Trip ID": "",
  "Duration": "",
  "Start Date": "",
  "Start Station": "",
  "Start Terminal": "",
  "End Date": "",
  "End Station": "",
  "End Terminal": "",
  "Bike #": "",
  "Subscriber Type": "",
  "Zip Code": ""
}
{
  "Trip ID": "",
  "Duration": "",
  "Start Date": ""

```

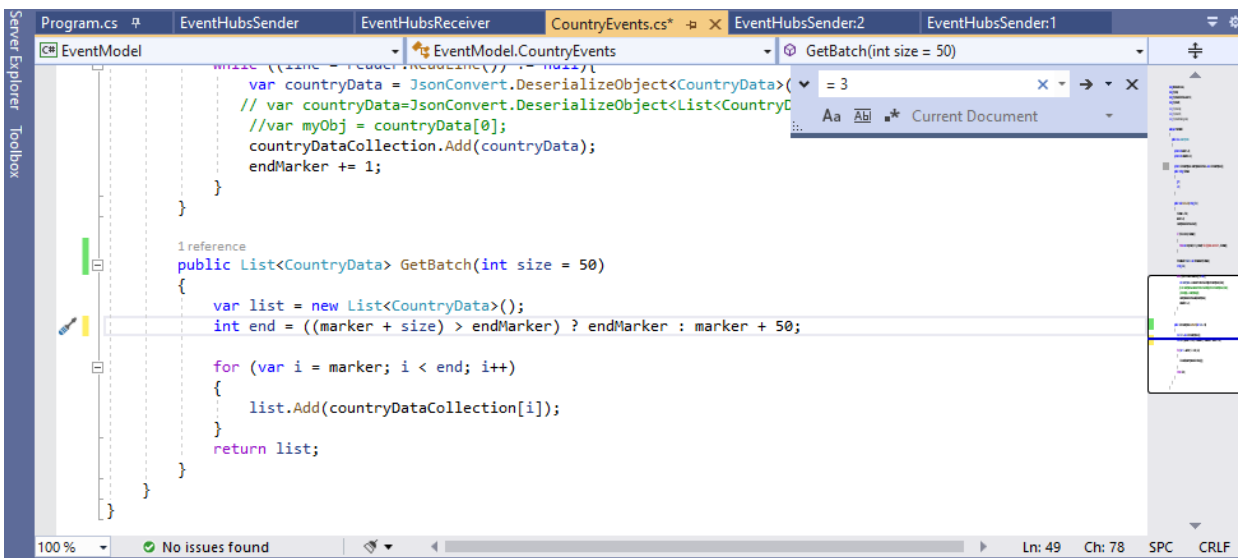
Formating:



Data after reformatting:



We edit the batch size and set the number of events =50;



Then we run the code:

```
C:\Users\b\Downloads\Event Hub VS Solution\EventHubsSender\bin\Debug\netcoreapp3.1\EventHubsSender.exe
A batch of 50 events has been published.
A batch of 50 events has been published.
A batch of 50 events has been published.
```

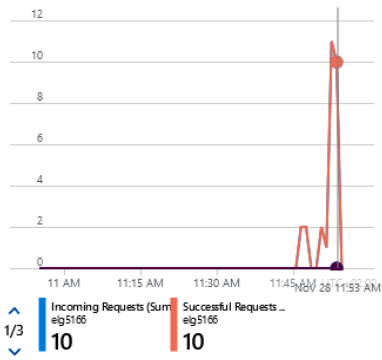
We look into azure, it seems that the connection has been successfully established.



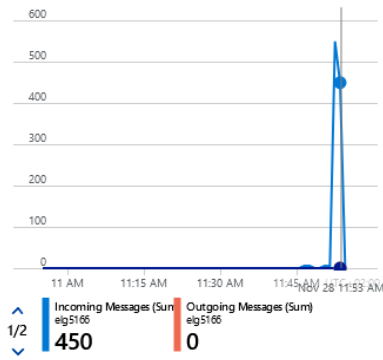
After 5 minutes:

Show data for the last: **1 hour** 6 hours 12 hours 1 day 7 days 30 days

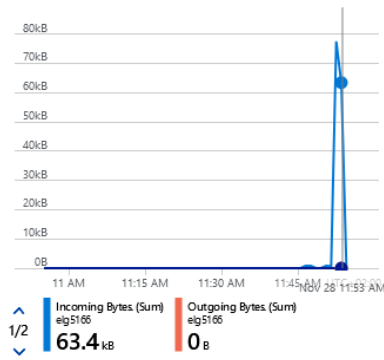
Requests



Messages



Throughput

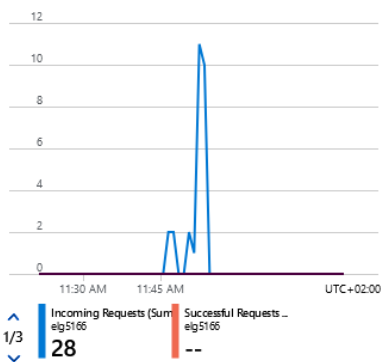


We wait until it's completed.

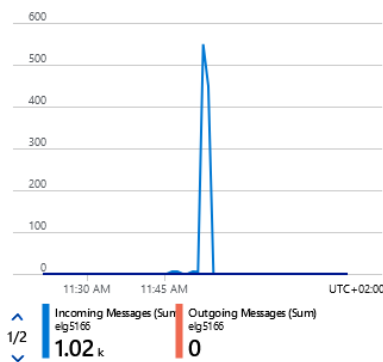
End of 20 batches:

```
C:\Users\b\Downloads\Event Hub VS Solution\EventHubsSender\bin\Debug\netcoreapp3.1\EventHubsSender.exe
A batch of 50 events has been published.
A batch of 50 events has been published.
A batch of 50 events has been published.
A batch of 50 events has been published.
A batch of 50 events has been published.
A batch of 50 events has been published.
A batch of 50 events has been published.
A batch of 50 events has been published.
A batch of 50 events has been published.
A batch of 50 events has been published.
A batch of 50 events has been published.
A batch of 50 events has been published.
A batch of 50 events has been published.
A batch of 50 events has been published.
A batch of 50 events has been published.
A batch of 50 events has been published.
A batch of 50 events has been published.
A batch of 50 events has been published.
```

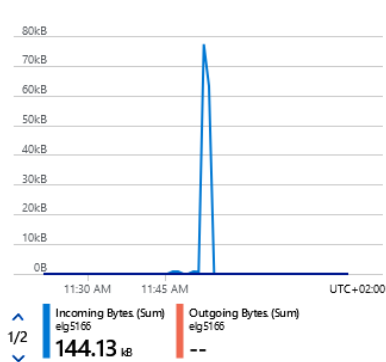
Requests



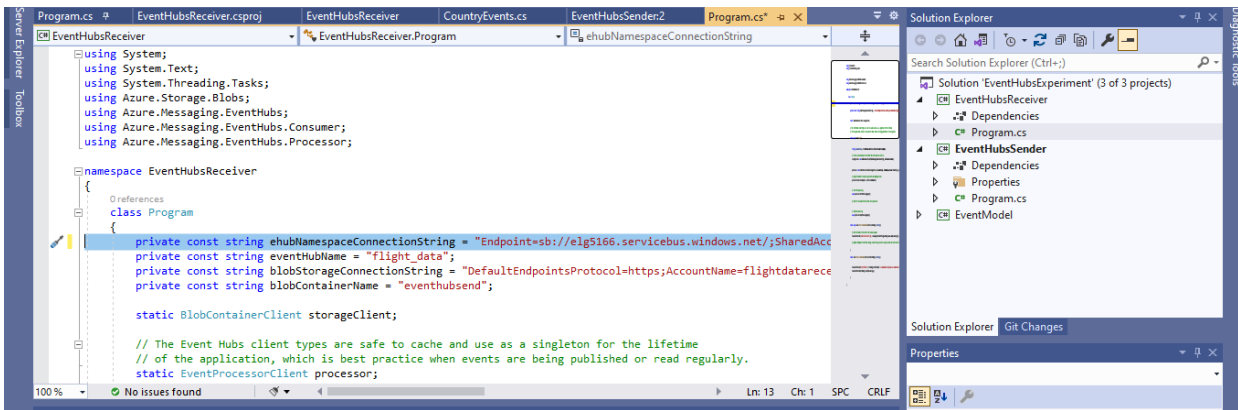
Messages



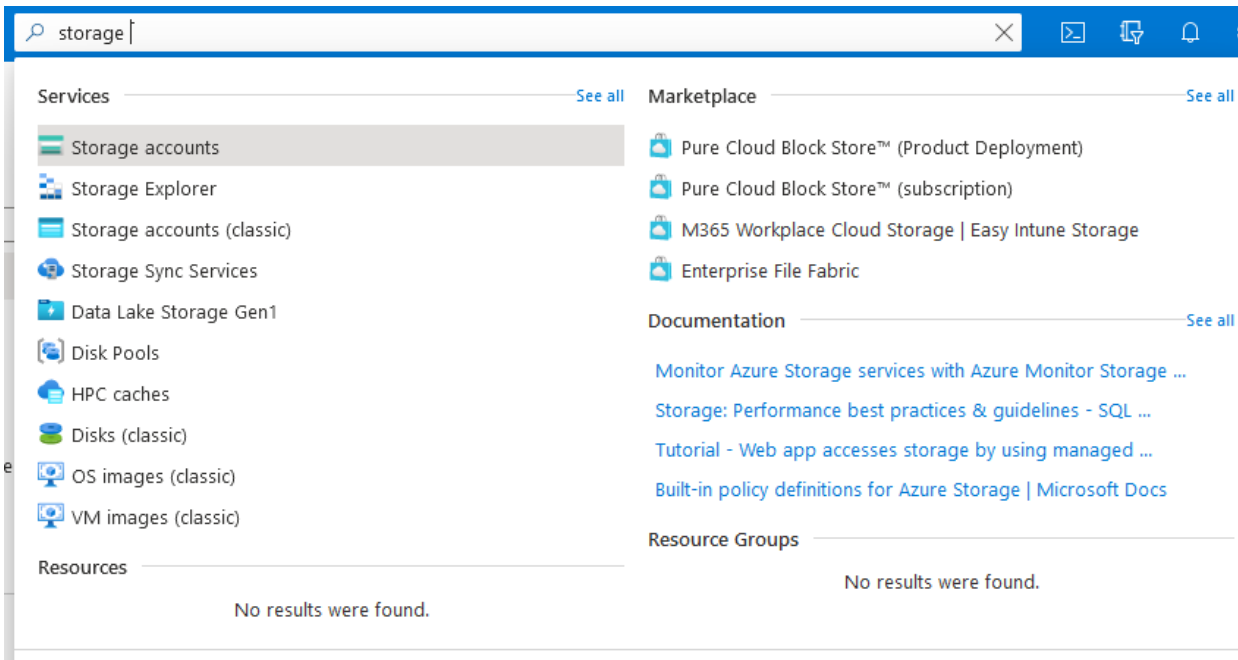
Throughput



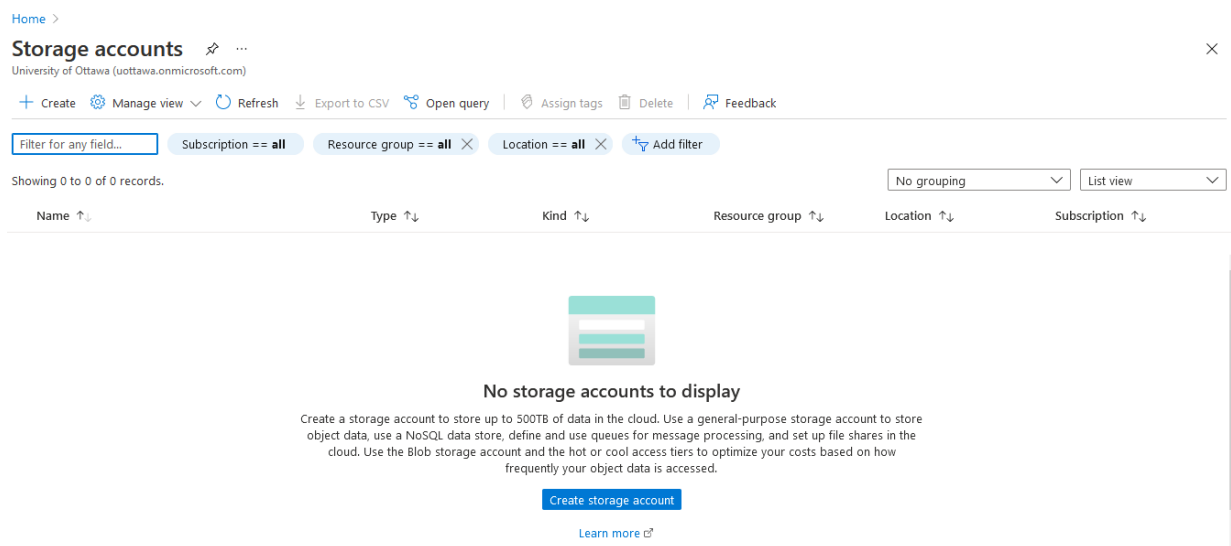
Then we copy the string primary key from the sender to the receiver.



We search for storage account.



Create storage account.



Microsoft Azure

Search resources, services, and docs (G+)

[Home](#) > [Storage accounts](#) >

Create a storage account ...

Basics

Advanced

Networking

Data protection

Tags

Review + create

If you need to create a legacy storage account type, please click [here](#).

Storage account name ⓘ *

bikedatarecievercloud

Region ⓘ *

(Canada) Canada Central

Performance ⓘ *

☒ Standard: Recommended for most scenarios (general-purpose v2 account)

☐ Premium: Recommended for scenarios that require low latency.

Redundancy ⓘ *

Geo-redundant storage (GRS)

☒ Make read access to data available in the event of regional unavailability.


Review + create

< Previous

Next : Advanced >

The we click create

Create a storage account ...

 Running final validation...

Basics Advanced Networking Data protection Tags Review + create

Deployment model	Resource manager
Performance	Standard
Replication	Read-access geo-redundant storage (RA-GRS)

Advanced

Secure transfer	Enabled
Allow storage account key access	Enabled
Allow cross-tenant replication	Enabled
Default to Azure Active Directory authorization in the Azure portal	Disabled
Infrastructure encryption	Disabled
Allow public access	Enabled

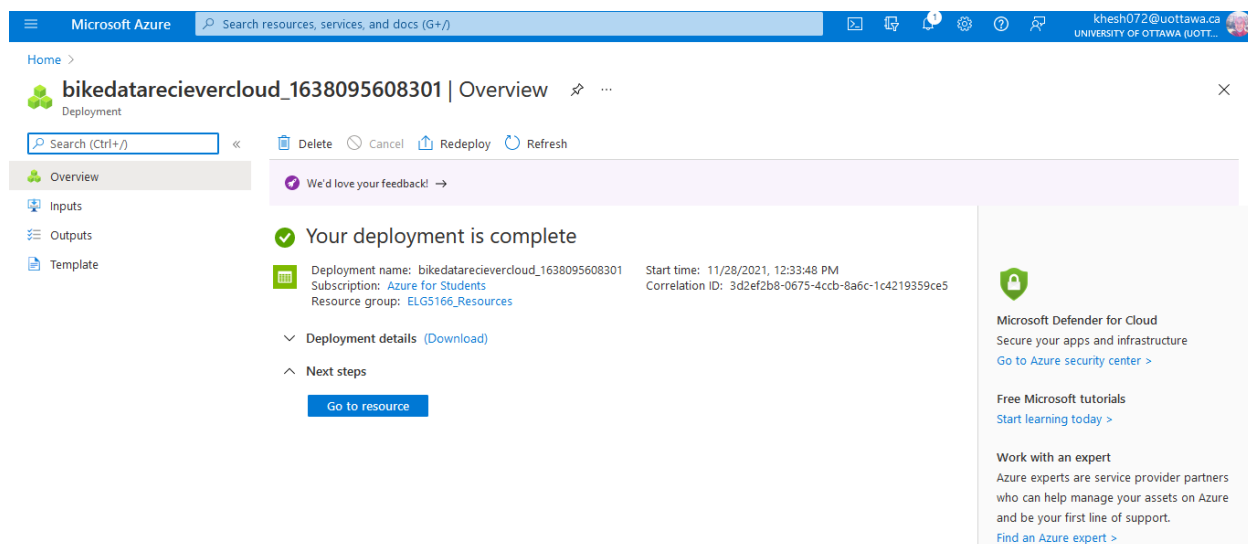
Create

< Previous

Next >

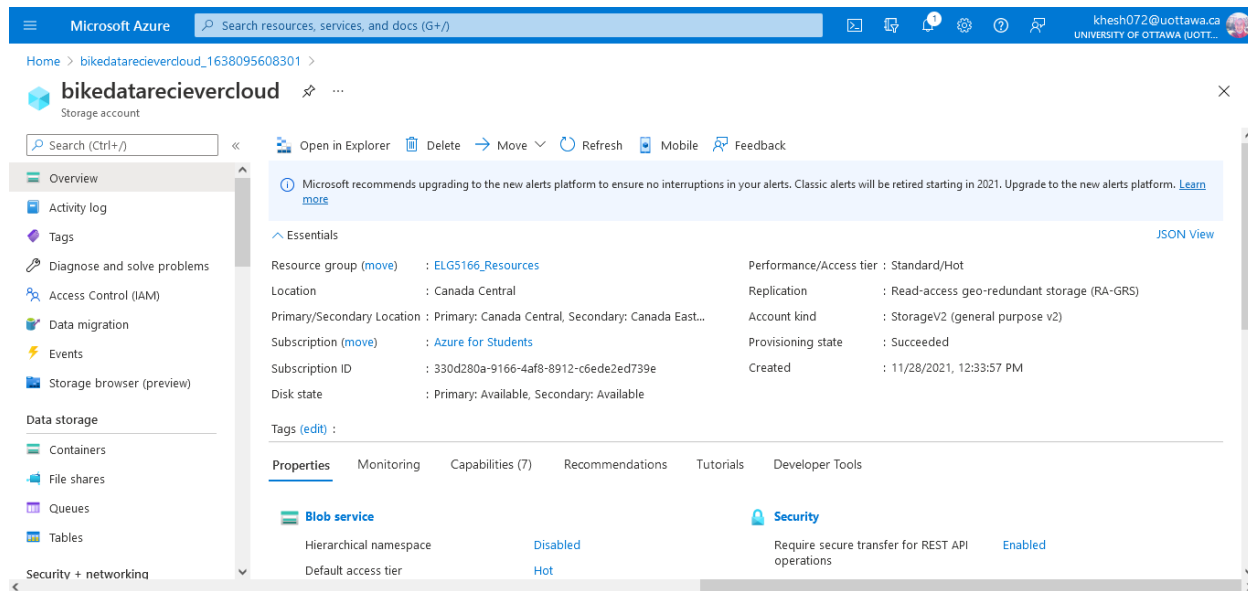
Download a template for a

The deployment is done.



The screenshot shows the Microsoft Azure portal interface. At the top, there's a navigation bar with the Microsoft Azure logo and a search bar. Below the navigation bar, the main content area displays the 'bikedatarecievercloud_1638095608301' deployment overview. The deployment is marked as 'Complete' with a green checkmark. The deployment details show the name 'bikedatarecievercloud_1638095608301', the subscription 'Azure for Students', and the resource group 'ELG5166_Resources'. The start time is '11/28/2021, 12:33:48 PM' and the correlation ID is '3d2ef2b8-0675-4ccb-8a6c-1c4219359ce5'. On the right side, there are three sections: 'Microsoft Defender for Cloud' with a link to 'Go to Azure security center', 'Free Microsoft tutorials' with a link to 'Start learning today', and 'Work with an expert' with a link to 'Find an Azure expert'.

Once we go to the resource, the storage group appears.



Microsoft Azure

Search resources, services, and docs (G+/I)

Home > bikedatarecievercloud_1638095608301 >

bikedatarecievercloud

Storage account

Search (Ctrl+/)

Open in Explorer Delete Move Refresh Mobile Feedback

Microsoft recommends upgrading to the new alerts platform to ensure no interruptions in your alerts. Classic alerts will be retired starting in 2021. Upgrade to the new alerts platform. [Learn more](#)

JSON View

Essentials

Resource group (move) : ELG5166_Resources

Location : Canada Central

Primary/Secondary Location : Primary: Canada Central, Secondary: Canada East...

Subscription (move) : Azure for Students

Subscription ID : 330d280a-9166-4af8-8912-c6ede2ed739e

Disk state : Primary: Available, Secondary: Available

Performance/Access tier : Standard/Hot

Replication : Read-access geo-redundant storage (RA-GRS)

Account kind : StorageV2 (general purpose v2)

Provisioning state : Succeeded

Created : 11/28/2021, 12:33:57 PM

Tags (edit) :

Properties Monitoring Capabilities (7) Recommendations Tutorials Developer Tools

Blob service

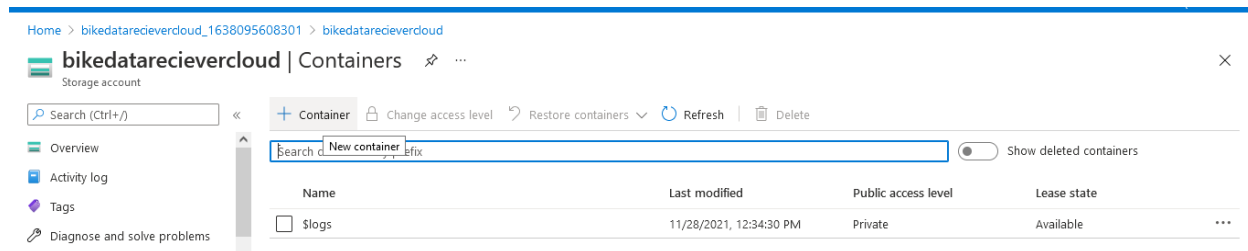
Hierarchical namespace : Disabled

Default access tier : Hot

Security

Require secure transfer for REST API operations : Enabled

Then, the storage needs a container. We go to the containers to create a new one.



Home > bikedatarecievercloud_1638095608301 > bikedatarecievercloud

bikedatarecievercloud | Containers

Storage account

Search (Ctrl+/)

+ Container Change access level Restore containers Refresh Delete

Search New container prefix

Show deleted containers

Name	Last modified	Public access level	Lease state	
\$logs	11/28/2021, 12:34:30 PM	Private	Available	...

We specify a name then click create.

?

khesh072@uottawa.ca
UNIVERSITY OF OTTAWA (UOTT...

New container

×

Name *

eventhubsend ✓

Public access level ⓘ

Private (no anonymous access) ▾

▽ Advanced

Create

Discard

Successfully created.

oud | Containers ⚙ ...

Successfully created storage container 'eventhubsend'.

+ Container

🔒 Change access level

↶ Restore containers ▾

🔄 Refresh

🗑 Delete

Search containers by prefix

● Show deleted containers

Name	Last modified	Public access level	Lease state	
<input type="checkbox"/> \$logs	11/28/2021, 12:34:30 PM	Private	Available	...
<input type="checkbox"/> eventhubsend	11/28/2021, 12:38:11 PM	Private	Available	...

We need to connect things, so that we go back to the receiver to go to the access keys.

Home > bikedatarecievercloud_1638095608301 > bikedatarecievercloud

bikedatarecievercloud | Access keys ...
Storage account

Search (Ctrl+/)

File shares
Queues
Tables
Security + networking
Networking
Azure CDN
Access keys
Shared access signature
Encryption
Security
Data management
Geo-replication
Data protection
Object replication

Show keys Set rotation reminder Refresh

Access keys authenticate your applications' requests to this storage account. Keep your keys in a secure location like Azure Key Vault, and replace them often with new keys. The two keys allow you to replace one while still using the other.

Remember to update the keys with any Azure resources and apps that use this storage account. [Learn more](#)

Storage account name
bikedatarecievercloud

key1
Last rotated: 11/28/2021 (0 days ago)
Rotate key
Key
.....
Connection string
.....

key2
Last rotated: 11/28/2021 (0 days ago)
Rotate key

We need to copy the connection key of the first key.

Hide keys Set rotation reminder Refresh

Access keys authenticate your applications' requests to this storage account. Keep your keys in a secure location like Azure Key Vault, and replace them often with new keys. The two keys allow you to replace one while still using the other.

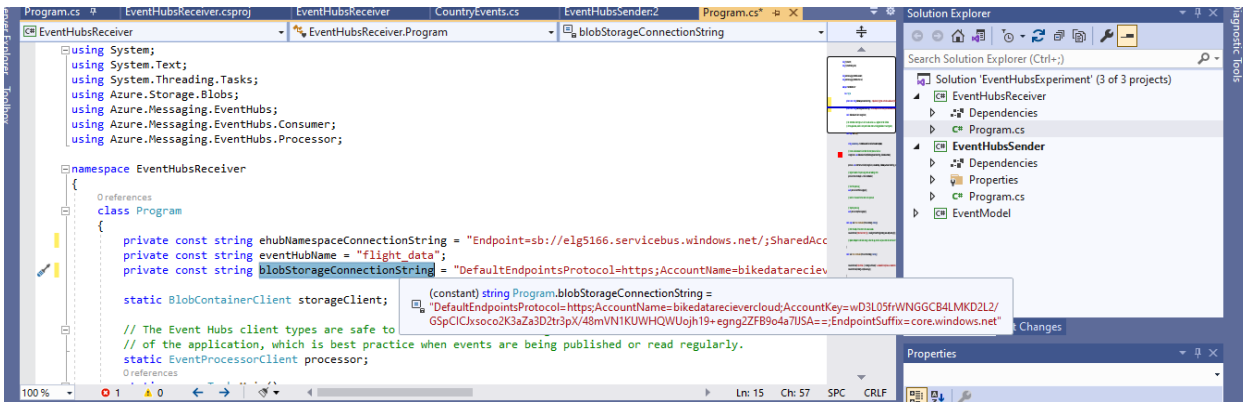
Remember to update the keys with any Azure resources and apps that use this storage account. [Learn more](#)

Storage account name
bikedatarecievercloud

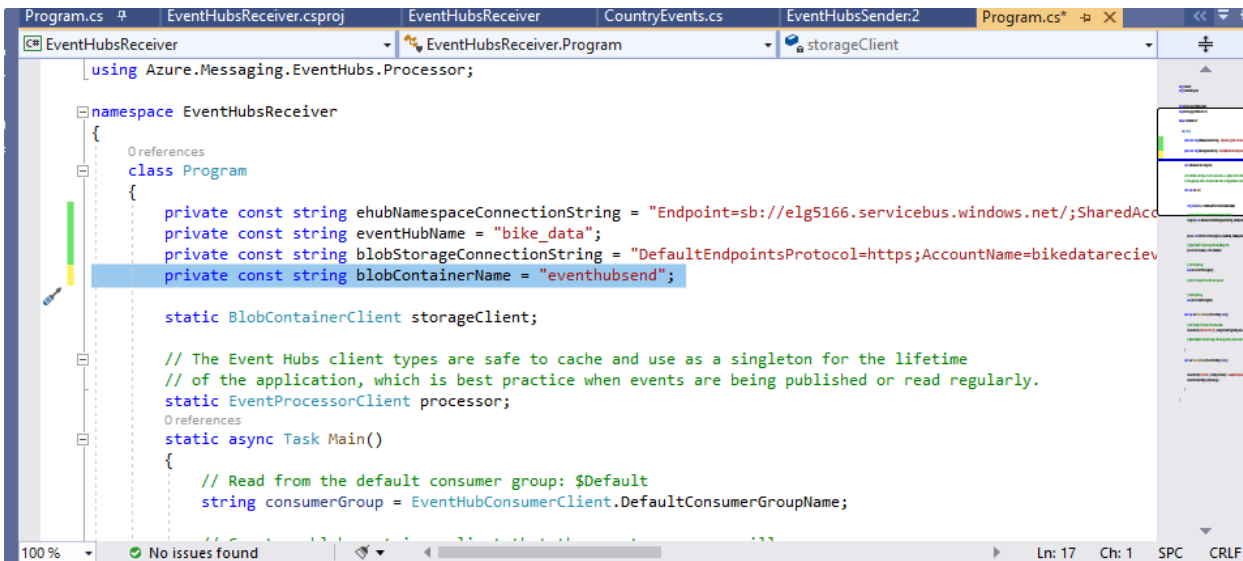
key1
Last rotated: 11/28/2021 (0 days ago)
Rotate key
Key
wD3L05frWNGGCB4LMKD2L2/GSpCICJxsoco2K3aZa3D2tr3pX/48mVN1KUWHQW ...
Connection string
DefaultEndpointsProtocol=https;AccountName=bikedatarecievercloud;AccountKey=wD3L05frWNGGCB4LMKD2L2/GSpCICJxsoco2K3aZa3D2tr3pX/48mVN1KUWHQW ...
Copied

key2

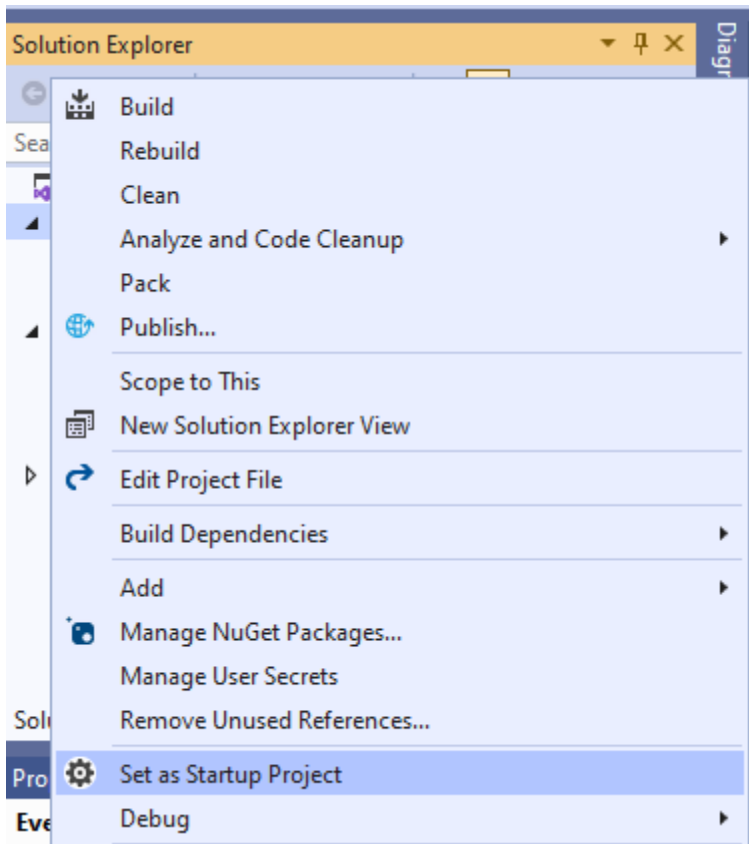
Then we go back to the receiver and paste the connection string to the blobStorageConnectionString



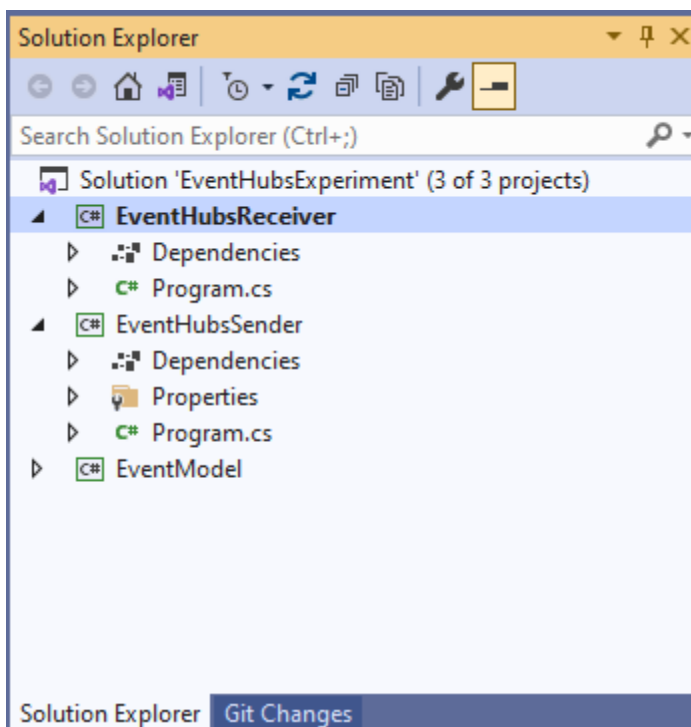
Set the name of the container as we have named it before and the name of the data.



Then we go to the receiver project and choose to set as a setup project.

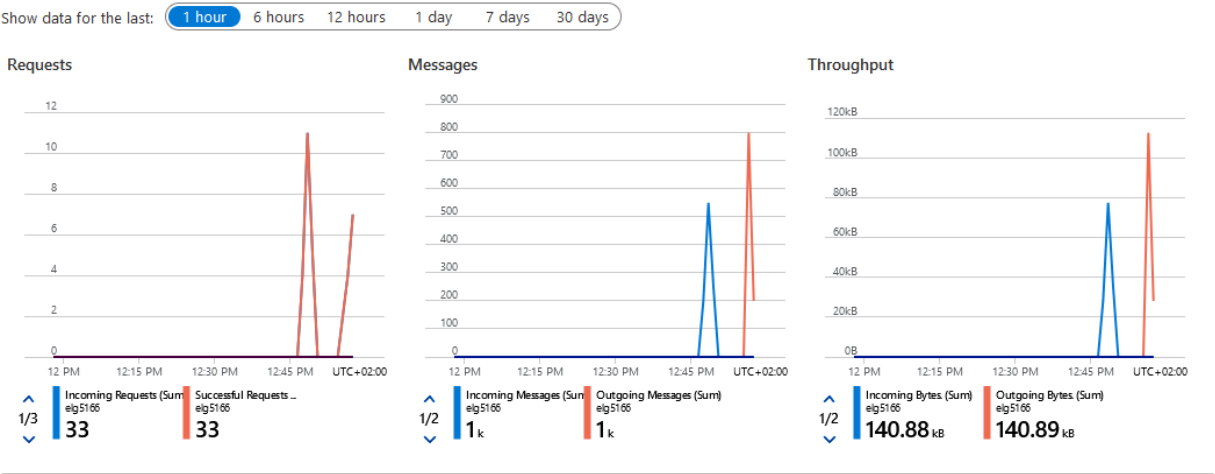


Then the project is being activated.



Receiving output:

Get back to the resource group, it seems that we consumed all the data we have.



Stream Analytics:

Go to the event-hub here then choose to process the data.



Home >

bike_data (ELG5166/bike_data) ...

Event Hubs Instance

Search (Ctrl+/) << + Consumer group Delete Refresh

Overview

- Access control (IAM)
- Diagnose and solve problems
- Settings
 - Shared access policies
 - Properties
 - Locks
- Entities
 - Consumer groups
- Features
 - Capture
 - Process data
- Automation
 - Task (preview)

Essentials [JSON View](#)

Resource group (Move) : [ELG5166_Resources](#) Status : [Active](#)

Location : Canada Central Namespace : [ELG5166](#)

Subscription (Move) : [Azure for Students](#) Created : Sunday, November 28, 2021, 05:53:42 GMT+2

Subscription ID : 330d280a-9166-4af8-8912-c6ede2ed739e Updated : Sunday, November 28, 2021, 05:53:42 GMT+2

Partition Count : 2 Message Retention : 1 day

Capture events

Use Capture to save your events to persistent storage.

Process data

Produce insights with Azure's data processing services.

Connect

Authenticate with connection strings and SAS policies.

Checkpoint

Create consumer groups to checkpoint your events.

Then we click to explore the data.

Process Event Hub data with other Azure services

Transform your Event Hub data into actionable insights in near real-time by integrating with these Azure services

Enable real time insights from events

- Preview Event Hub data
- Analyze your data using SQL-like query
- Deploy query by creating a new Azure Stream Analytics job

[Explore](#)

More services coming soon!

We're working hard to bring more Azure integrations to your Event Hub.

Then we click on we need a permission.

[ut preview](#) [Test results](#)

We need your permission

Preview data we will create a new policy and, because this is a basic tier event hub instance, use the default consumer group. If this consumer group exceeds its maximum of five concurrent uses, it could impact other readers.

[Create](#)

[Create](#)

Now we can see the data after running the default provided query.

Test query

```
1 SELECT
2 | *
3 INTO
4 | [OutputAlias]
5 FROM
6 | [bikedata]
```

Input preview Test results

Showing events from 'bikedata'. This list of events might not be complete.

View in JSON  Table  Raw  Refresh  Download sample data

Trip ID	Duration	Start Date	End Date	Bike #	Subscriber Type	Zip Code
"913453"	"789"	"8/31/2015 23:09"	"8/31/2015 23:22"	"487"	"Customer"	"9069"
"913452"	"293"	"8/31/2015 23:07"	"8/31/2015 23:12"	"538"	"Subscriber"	"94118"
"913451"	"896"	"8/31/2015 23:07"	"8/31/2015 23:22"	"363"	"Customer"	"92562"
"913450"	"255"	"8/31/2015 22:16"	"8/31/2015 22:20"	"470"	"Subscriber"	"94111"

Now we can run the required query:

```
1 SELECT System.Timestamp AS WindowEnd,End_Date,sum(Bike) As TotalBikes,sum(Duration) AS TotalDuration
2 INTO [OutputAlias]
3 FROM [bikedata]
4 GROUP BY TumblingWindow(Duration(second,2)),End_Date
```

Result:

```
[{"WindowEnd":"2021-11-28T13:17:46.0000000Z","End_Date":null,"TotalBikes":12228,"TotalDuration":23356.0}]
```

Part 2 - Azure Data Lake Analytics:

a) Top 20 zip codes where most bikes were rented from.

Microsoft Azure

Search resources, services, and docs (G+/)

Home > elgstorageegen2 > for_rental_bike q1 >

File preview ...

rental-output1.csv

Format

Download

Rename file

Access

Properties

Set expiry

Delete file









Open in Explo

0	1
Zip Code	Total_Re...
94107	20228868
94105	8788689
94133	7077405
94103	6353674
94111	4782022
94102	4416376
nil	2805680
94109	2780737
94403	1814012
94611	1775741
94117	1770760
94501	1754683
94158	1722861
94602	1687228
94114	1527866
94010	1515129
95112	1438165
94025	1425836

b) Daily duration aggregate across the rental subscriber types.

file preview ...

rental-output2.csv

 Format  Download  Rename file  Access  Properties  Set expiry  Delete file  Open

0	1	2
Subscriber ...	Daily_Date	Total_D...
Customer	2014-09-01	869202
Subscriber	2014-09-01	116866
Customer	2014-09-02	411520
Subscriber	2014-09-02	624576
Customer	2014-09-03	215913
Subscriber	2014-09-03	722325
Customer	2014-09-04	285461
Subscriber	2014-09-04	680215
Customer	2014-09-05	427192
Subscriber	2014-09-05	722445
Customer	2014-09-06	667018
Subscriber	2014-09-06	115238
Customer	2014-09-07	745473
Subscriber	2014-09-07	149557
Customer	2014-09-08	890772
Subscriber	2014-09-08	704323
Customer	2014-09-09	513122
Subscriber	2014-09-09	831411

c) What are the top busiest terminals for bike pickup.

File preview ...

rental-output3.csv



Format



Download



Rename file



Access



Properties

0

1

Start... name

70 San Francisco Caltrain (Towns...

69 San Francisco Caltrain 2 (330 ...

50 Harry Bridges Plaza (Ferry Buil...

55 Temporary Transbay Terminal ...

60 Embarcadero at Sansome

61 2nd at Townsend

65 Townsend at 7th

74 Steuart at Market

67 Market at 10th

77 Market at Sansome

76 Market at 4th

39 Powell Street BART

64 2nd at South Park

56 Beale at Market

73 Grant Avenue at Columbus Av...

62 2nd at Folsom

72 Civic Center BART (7th at Mar...

57 5th at Howard

d) Which 5 terminal has the least drop-offs?

File preview






rental-output4.csv

Format Download Rename file Access Properties

0	1
End ...	name
24	Redwood City Public Libr...
21	Franklin at Maple
83	Mezes Park
23	San Mateo County Center
26	Redwood City Medical C...

e) What is the monthly summary of bike rentals (format - month/year ex. 06/2020).

We will show the monthly summary bike rentals assuming we want the trips count and duration of the bike rentals.

 Format  Download  Rename file  Access  Properties			
0	1	2	3
Month_y...	Total_bike...	Total_dur...	Trip_co...
2014-09	13310244	33160021	31682
2014-10	14386280	33401099	34220
2014-11	10752350	22454934	25516
2014-12	8382802	41131402	19677
2015-01	11756577	25611358	27840
2015-02	11139951	25633016	26401
2015-03	13444028	29892301	31626
2015-04	13313721	28031940	31363
2015-05	12574550	31584633	29540
2015-06	13447867	34481927	31907
2015-07	13718375	33983062	32476
2015-08	13564645	31088866	31904

Part 3 - Definitions: 1. Please compare briefly, based on at least 3 criteria, the differences in architecture between Apache Spark Structured Streaming and Azure Event Hubs & Stream Analytics? Ref [1, 2]

	Spark Structured Streaming	Azure Event Hubs	Azure Stream Analytics
Definition	A stream processing engines that allow users to process huge amounts of data using complex algorithms expressed with high-level functions like map, reduce, join, and window to be then pushed to filesystems, databases, or even back to Event Hubs.	A streaming platform and event ingestion service.	A real-time analytics service designed to analyze and process fast moving streams of data from multiple sources simultaneously that can be used to get insights, build reports or trigger alerts and actions.
Working mechanism	<ul style="list-style-type: none"> • Works on the architecture of polling the data after some duration, based on your trigger interval, but, there is no concept of a batch. • Each event is represented by a row in an unbounded table, and event-time is a column value in the row. • It process the data based on event-time when the timestamp of the event is included in the data received which handles late coming data and gets more accurate results. • Is based on Dataframe and Dataset Spark APIs which generate logically and 	<ul style="list-style-type: none"> • It ingests millions of events per second from any source to build dynamic data pipelines and immediately respond to business challenges. • It decouples event producers from event consumers using a partitioned consumer model. 	<ul style="list-style-type: none"> • It ingests data from Azure Event Hubs (including Azure Event Hubs from Apache Kafka), Azure IoT Hub, or Azure Blob Storage. • The query, which is based on SQL query language, can be used to easily filter, sort, aggregate, and join streaming data over a period of time.

	physically optimized query plan automatically.		
Processing of the output	<ul style="list-style-type: none"> It has a limited number of output sinks and with one sink only one operation can be performed and can't save the output at multiple external storages. 	Allows multiple applications to process the stream concurrently with a controlled processing speed.	Can write query result into multiple Azure services, including Azure Cosmos DB, Azure Synapse Analytics, Azure SQL DB, storage.
Keeping relative state based on time	<ul style="list-style-type: none"> Can maintain the intermediate state for partial aggregates for a long period of time such that late data can update aggregates of old windows correctly. For example, detect spike or get the maximum value all time. 	Has limited time retention buffer that prevents it from doing a direct query on past event data.	Can't store a relative state regardless of how much time has passed.
Use case	Build end-to-end continuous applications using Apache Spark that combine: streaming, batch, and interactive queries.	<ul style="list-style-type: none"> When there are multiple data sources and we need to feed them into one place for analysis. Connect with Azure Stream Analytics to build an end-to-end serverless streaming solution. Capture real-time data in an Azure Blob storage or Azure Data Lake Storage for long-term retention or micro-batch processing. 	<ul style="list-style-type: none"> Connect Azure event hubs with Azure Cosmos DB since Azure event hubs can't direct its output into Azure Cosmos DB, only a few Azure services can direct it. Find patterns and relationships that can be used to trigger actions and initiate workflows such as creating alerts, feeding information to a reporting tool, or storing transformed data for later use. Filter, sort,

			aggregate, and join streaming data over a period of time easily through simple language constructs and/or configurations.
Reusability	Can reuse the code and move its results to a different cloud provider with some work.	Can't take the code out of Azure and reuse it somewhere else, because it is a proprietary language and solution.	Can't take the code out of Azure and reuse it somewhere else, because it is a proprietary language and solution.
Programming Languages	Can use the Dataset/DataFrame API in Scala, Java, Python or R to express streaming aggregations, event-time windows, stream-to-batch joins, etc.	Not applicable.	Can extend its SQL language with JavaScript and C# user-defined functions (UDFs).
Recovery	Ensure end-to-end exactly-once semantics under any failure with the help of checkpointing and write-ahead logs mechanisms.	Keep processing data during emergencies using the geo-disaster recovery and geo-replication features.	<ul style="list-style-type: none"> Has built-in checkpoints to maintain the state of your job and provides repeatable results in case the delivery of an event fails. Guarantees exactly once event processing and at-least-once delivery of events, so events are never lost.
Scalability	Is scalable and received data in a trigger is appended to a continuously flowing unbounded data stream.	Easy scaling from streaming megabytes of data to terabytes while keeping control over when and how much to scale.	Adding more Streaming Units to a job easily. But it can't do auto scaling.
Security	By default, it's not enabled and need additional configurations.	It supports the option of encrypting data at rest with either Microsoft-managed keys or customer-managed keys.	It encrypts all incoming and outgoing communications and built-in checkpoints.

2. Describe briefly 3 benefits of Azure U-SQL over Apache Spark. Illustrate them briefly with some use cases. Ref [3]

Azure U-SQL is a Microsoft query language that combines a declarative SQL-like syntax with C# programming, enabling it to be used to process both structured and unstructured data in big data environments with batch mode only.

Benefits of Azure U-SQL over Apache Spark:

1. With Azure U-SQL, you can query data where it lives instead of having to move/copy all the data to one location. For external systems, such as Azure SQL DB/DW and SQL Server in a VM, this is achieved using federated queries against those data sources where the query is “pushed down” to the remote data source and executed on that data source, with only the results being returned.
2. In Microsoft’s Azure cloud, Data Lake Analytics offers U-SQL as a serverless job service which can be cheaper because you aren't charged while it's provisioning. Unlike Apache Saprk that costs money as long as the cluster is running (cluster service). Also, with Apache Spark, you have to take into account the implications of creating, sizing, scaling, and decommissioning the clusters.
3. U-SQL allows writing custom code easily in C# to express complex business requirements:
 - Inline C# functions,
 - User-Defined Functions (UDF),
 - User- Defined Operators (UDO) such as extractors, outputters, reducers, processors, appliers, and combiners,
 - and User-Defined Aggregators (UDAGG) such as aggregate to perform custom math calculations or manipulations with strings.

Unlike Spark that doesn't offer the same extensibility model for operators, but has equivalent capabilities for some. Also, if the developer is familiar with C# more than Scala or Python, this would be an additional benefit.

4. It unifies querying structured and unstructured data, data viewing (regardless of physical location), and data copying.

3. What are the 5 characteristics of Azure Data Lake that distinguish it from other Distributed Dataset Storage infrastructure such as Hadoop?

- Security and Encryption:
 - o Data is encrypted and secured both in transit and rest. Data Lake safeguards the data assets while allowing you to simply extend your on-premises security and governance policies to the cloud. Azure Active Directory has features like single sign-on (SSO) and multi-factor authentication. Unlike Hadoop is designed without considering security of data. Data stored at HDFS is in plaintext. This data is prone to be accessed by unauthorized user.
- Combining data from multiple sources into a single location:
 - o Azure Data Lake combines all of big data from many sources across cloud and on-premises infrastructures into a single location. Unlike Hadoop, the analysis of data across different sources, can be extremely difficult to achieve
- Enabling Hadoop for the Cloud:
 - o Because of the perceived constraints of cloud infrastructure, many organisations select for an on-premises Hadoop solution. Azure Data Lake brings Hadoop's cloud of constraints closer to typical deployments.
- Unlimited Data Size:
 - o File and account sizes are limited to a few terabytes in other cloud storage choices. This may appear to be a large number, but in big data situations, it is actually modest. This limitation is removed with Azure Data Lake: There is no fixed file or account size limit. Hadoop can store and process single or many files at terabyte and petabyte scales. This is a perfect scenario because Hadoop shines when dealing with massive files.
- Very High-Speed Throughput:
 - o Internet of Things (IOT) devices such as phones, sensors and equipment are streaming data in very small transactions at very high volume. At massive scale, Azure Data Lake supports incredibly high-speed data ingestion.
- Developed with Parallel Processing:
 - o Azure Data Lake is built to facilitate parallel processing throughput in order to deliver performance comparable to an on-premises Hadoop solution.

4. What 2 factors influence the use of Azure AUs for U-SQL query processing. Describe with a simple example.

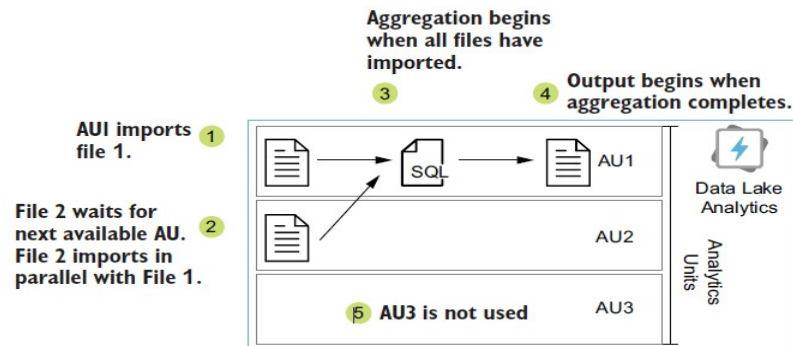


Figure 2.10 Parallel processing in Azure Data Lake Analytics jobs with multiple analytics units

Figure 1: Source: NUCKOLIS, R. L. (2020). "Azure Storage, Streaming, and Batch Analytics. A guide for data engineers". Manning Publications Co. ISBN: 9781617296307

The processing can be done by one, two, or more analytics units:

- Single Unit:
 - o Both files will be imported consecutively using one unit, the consolidation step will be run, and the summary file will be created.
- Two or more Units:
 - o Both files will be imported in parallel using two units, then the aggregation step will be run, and finally the summary file will be created.

2 Factors influence the use of Azure AUs for U-SQL:

- Time:
 - o The time which the Analytical Unit takes. So if we want to accelerate the process we can use two or many Analytical Units which the files will be imported in parallel.
- Cost:
 - o The cost is the same for running one or more Analytical Units.

Example:

- Time:
 - o If we run a job that takes 16 hours with one AUs
 - o It would take 8 hours with two AUs.

- o It would take one hour with 16 AUs.
- Cost:
 - o All three above runs will cost the same.
 - o $16\text{hr} \cdot 1\text{AU} = 8\text{hr} \cdot 2\text{AU} = 1\text{hr} \cdot 16\text{AU} = \text{the same cost.}$

References:

[1] Lab code and lecture notes.

[2] <https://stackoverflow.com/questions/52135478/deserialize-array-of-arrays-in-c-sharp/52135640>

[3] <https://superuser.com/questions/34451/notepad-find-and-replace-string-with-a-new-line>

[4] Nuckolls, R. L. (2020). "Azure Storage, Streaming, and Batch Analytics. A guide for data engineers". *Manning Publications Co.* ISBN: 9781617296307

[5] <https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-introduction>

[6] <https://azure.microsoft.com/en-us/services/event-hubs/#overview>

[7] <https://devblogs.microsoft.com/visualstudio/introducing-u-sql-a-language-that-makes-big-data-processing-easy/>

[8] <https://xo.xello.com.au/blog/azure-data-lake-benefits>

[9] <https://www.bluegranite.com/blog/6-key-features-from-microsoft-s-azure-data-lake>

[10] <https://www.techopedia.com/2/31730/trends/big-data/hadoop-analytics-not-so-easy-across-multiple-data-sources>

[11] <https://stackoverflow.com/questions/6467216/is-it-possible-to-use-aggregate-function-in-a-select-statement-without-using-group-by>