





## Lab 3: Data Flow Aggregation

In this lab, we take a file from our ADLSgen2, perform data transformation, then put the result to an Azure SQL DB Table.

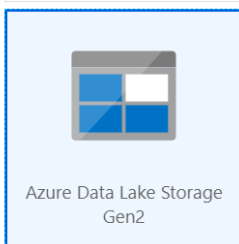
### Step 1: Build a dataset for source CSV

Confirm you have the file in ADLSgen2 after Lab2,

<div> <div>← → ∨ ↑</div> <div>adlsgen2 &gt; demodest</div> </div>	
Name	Last Modified
 SacramentoCrimeJanuary2006.csv	8/22/2019, 9:40:04 AM
 SacramentoRealEstateTransactions.csv	8/22/2019, 9:40:05 AM
 SalesJan2009.csv	8/22/2019, 9:40:06 AM
 TechCrunchContinentalUSA.csv	8/22/2019, 9:40:05 AM

In ADF portal, create a dataset, reuse the ADLSGen2 Linked Services, define a data set:

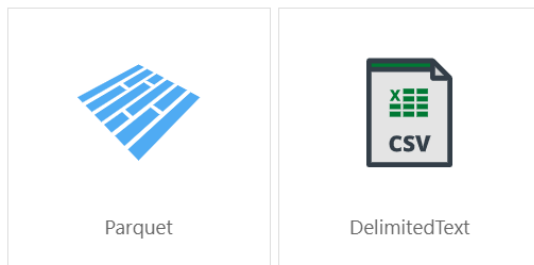
### New Dataset



Select DelimitedText in the box,

### ← Select Format

Choose the format type of your data



Define below properties with your own ADLS, clicking the First row as header will make your life easier.

← Set Properties
×

---

Name

Linked service \*

[Edit Connection](#)

File path  
 /  /  [Browse](#)

First row as header ☐

Import schema  
☒ From connection/store ☐ From sample file ☐ None

## Step 2: Define the required DataFlow transformation

Drag an activity [Data Flow (Preview)] to use

**Activities**
⌵
⏪

Move & Transform

Copy Data

Data Flow (Preview)

Saved
Save as template
Validate

In Settings, click [+New]

General
Settings
Parameters
User Properties

Data Flow \*
AggregateCSV\_To\_Table
Edit
New

Run on \*
AutoResolveIntegrationRuntime

▶ PolyBase ⓘ

Once in the DataFlow panel click on the first slot as source. Remember you defined a CSV on ADLSgen2, now we use it here in Source dataset.

Saved

Validate

source1

Columns: 10 total

+

Aggregate1

Aggregating data by 'Column\_2' producing columns 'Column\_8'

+

sink1

Export data to AzureSqlTable1

Add Source

Source Settings

Source Options

Projection

Optimize

Inspect

Data Preview

Output stream name \*

source1

Documentation

Source dataset \*

TechCrunch

Edit

New

Options

Allow schema drift

Infer drifted column types

Validate schema

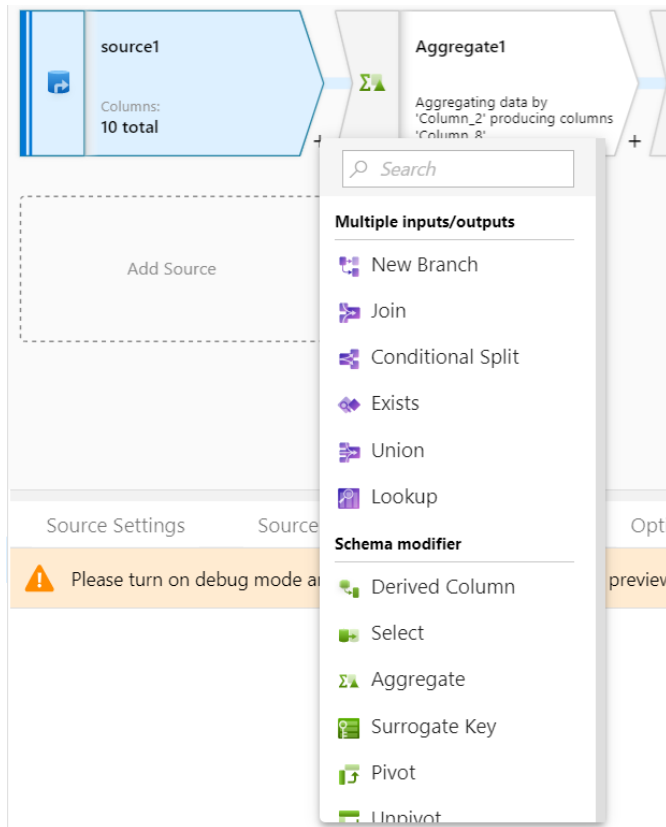
Skip line count

Sampling \*

Enable

Disable

Then go to the small [+] sign underneath,



Choose Aggregate, In Aggregate Settings [Group by ] tab, define source1's column as "Column\_2"

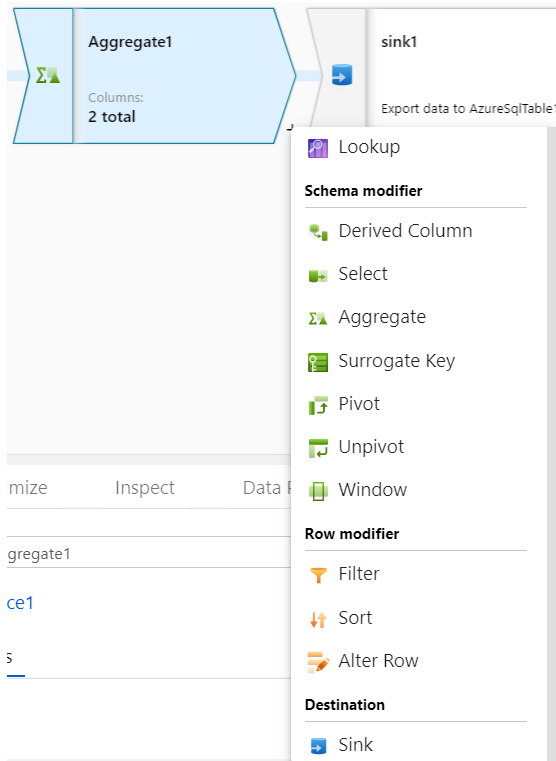
Group by		Aggregates	
<b>source1's column</b>		<b>Name as</b>	
abc Column_2		Column_2	

In Aggregates Table, select column 8 and sum(column\_8) per below:

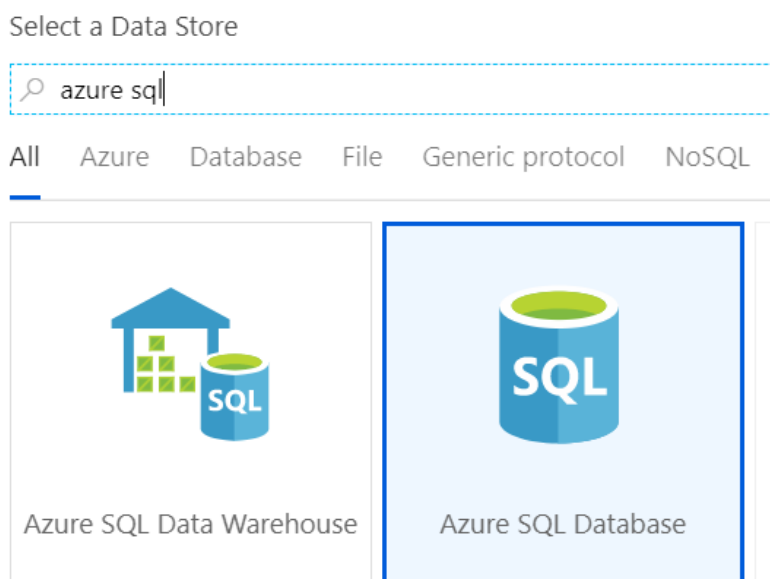
Incoming stream * source1	
Group by	Aggregates
Grouped by: Column_2	
Column_8	sum(Column_8) 121

### Step 3: Define output table

Once the Aggregation is read, click on small [+] again and scroll down to sink



In the Sink tab, new a dataset and select [Azure SQL Database]



In the Dataset properties, setup per below

← Set Properties
×

Name

Linked service \*

[Edit Connection](#)  
☐ Select from existing table    ☒ Create new table

Schema and table name  
 .

Your pipeline is ready to be executed and below is the execution result:

```

/***** Script for SelectTopNRows command from SSMS *****/
SELECT TOP (1000) [Column_2]
, [Column_8]
FROM [dbo].[TechCrunchAggregated]
order by 1;

```

100 %

Results Messages

	Column_2	Column_8
235	fabrik	51300000
236	Facebook	495700000
237	FameCast	4500000
238	Farecast	20600000
239	fatdoor	7000000
240	Fathom Online	6000000
241	Fave Media	1600000
242	Federated Media	54500000
243	FeedBurner	8000000
244	ffwd	1700000
245	Fifth Generation Syste...	10550000
246	FilmLoop	5600000
247	Filtrbox	515000
248	Firefly Energy	29000000

