

## 1. Created a new Stream Analytics job and chose to use 1 streaming Unit

### New Stream Analytics job ...

Azure Stream Analytics is a fully managed, SQL-based stream processing engine designed to help you tackle scenarios like streaming ETL to Azure Data Lake Storage, real-time dashboarding with Power BI, event driven applications with Azure SQL DB & Cosmos DB, remote monitoring, predictive maintenance, and more. [Learn more about Azure Stream Analytics](#)

#### Project details

Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription \* ⓘ Azure Pass - Sponsorship ✓

Resource group \* ⓘ b2\_group3 ✓  
[Create new](#)

#### Instance details

Name \* b2\_group3\_StreamJob ✓

Region \* ⓘ West Europe ✓

Hosting environment ⓘ  
☒ Cloud  
☐ Edge

#### Streaming unit details

Streaming units (SUs) represents the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job. The number of SUs can be modified once you create the job. You will be charged for the job's Streaming Units only when the job runs. [Learn more about streaming units](#)

Streaming units \* 1 ✓

[Review + create](#)

[< Previous](#)

[Next : Storage >](#)

## Then configured the Input for the stream:

15 | Overview > b2\_group3\_StreamJob  
Job | Inputs ☆ ...

+ Add input ⓘ Refresh

Alias ⓘ Source type

Event Hub

New input

Input alias \* input-from-eventhub ✓

☐ Provide Event Hub settings manually  
☒ Select Event Hub from your subscriptions

Subscription Azure Pass - Sponsorship ✓

Event Hub namespace \* ⓘ eventhub-b2-group3 ✓

Event Hub name \* ⓘ  
☐ Create new ☒ Use existing  
eventhub\_b2\_group3\_hub ✓

Event Hub consumer group \* ⓘ  
☐ Create new ☒ Use existing  
\$Default ✓

Authentication mode  
Create system assigned managed identity ✓  
The Azure Event Hubs Data Owner role will be granted to the Managed Identity for this Stream Analytics job when you click Save. If grant fails follow the manual grant steps [here](#) ⓘ.

Partition key ⓘ

Event serialization format \* ⓘ

[Save](#)

Remark: will we need a partition key? Not specified now.

## 3. Configured an output sink:

Created a storage account (storageb2group3) with a Blob Storage Container(blobstorageb2group3).

## 4. Configured the output in the Stream Analytics Job: StreamingOutputBlobStorage

5. Selected 'Run' on the Stream Analytics Job overview and took a first look at the results. Surprisingly, 2 datasets (JSON) were created in our BlobStorage for data with PartitionID 0 or 1...

6. Connection to PowerBI:

The screenshot shows the Azure Stream Analytics portal interface. On the left, a sidebar contains navigation options: Overview, Activity log, Access control (IAM), Tags, Diagnose and solve problems, Job topology, Inputs, Functions, Query (selected), Outputs, Settings, Environment, Storage account settings, Scale, Locale, Event ordering, Error policy, Compatibility level, Managed Identity, Properties, and Locks. The main area is titled 'b2\_group3\_StreamJob | Query' and shows a query editor with the following SQL:

```
1 SELECT
2   *
3 INTO
4   [StreamingOutput]
5 FROM
6   [input-from-eventhub]
7 WHERE Lat > 0
```

Below the query editor is an 'Input preview' section with a warning: 'No data was found for preview from 'input-from-eventhub'. Make sure the input has recently received data and the correct permissions are set.' Below this are buttons for 'Table', 'Raw', 'Refresh', 'Select time range', 'Upload sample input', and 'Download sample data'. At the bottom, a status bar reads: 'While sampling data, no data was received from '2' partitions.'

On the right, the 'Power BI' settings panel is open, showing a 'New output' configuration. The 'Output alias' is 'StreamingOutputForPowerBI'. The 'Provide Power BI settings manually' option is selected. The 'Group workspace' is '854bd1f8-052a-43e0-a1e9-149aa86fd186'. The 'Authentication mode' is 'Managed Identity: System assigned'. A warning message states: 'The test connection will fail because the Managed Identity for this Stream Analytics job doesn't have permission to this resource. You do not have the correct permissions to change the settings. Contact a authorized user to add the Contributor role for your Power BI. Learn more'. The 'Dataset name' is 'SailingDataStreamingOutput' and the 'Table name' is 'SailingDataPosition'. A 'Save' button is at the bottom.

Edited the Query to only select Latitude and Longitude columns:

The screenshot shows the Azure Stream Analytics query editor with the following SQL:

```
1 SELECT
2   *
3 INTO
4   [StreamingOutput]
5 FROM
6   [input-from-eventhub]
7 WHERE
8   latitude > 0
9
10 SELECT
11   latitude
12   longitude
13 INTO
14   [StreamingOutputForPowerBI]
15 FROM
16   [input-from-eventhub]
17 WHERE
18   latitude > 0
```

PowerBI Streaming Settings:

Power BI

×

New output

Streaming output for Power BI

☒ Provide Power BI settings manually
 ☐ Select Power BI from your subscriptions

Group workspace \* ⓘ

854bd1f8-052a-43e0-a1e9-149aa86fd186

Authentication mode

User token

Dataset name \* ⓘ

SailingDataGroup3

Table name \*

LiveDataSailingGroup3

Currently authorized as [Gemma Schiphorst Preuper](#)  
(student\_gemma@techionista-academy.com)

**Authorize connection**  
 You'll need to authorize with Power BI to configure your output settings.

Authorize

**Note:** You are granting this output permanent access to your Power BI dashboard. Should you need to revoke this access in the future you can do one of the following:

1. Change the user account password.
2. Delete this output.
3. Delete this job.

Save

Idea: make a Gauge Chart to see how far they've come compared to the total distance  
 Idea: add speed of a few boats to see their speed over the last minutes

How to decide rank: by distance from starting point with a formula.

- add a column called 'distance'
- Add the formula

import math

latitude = 38.586562661789316

longitude = -9.429378108391333

latCascais = 38.69225437789037

lonCascais = -9.419236159278585



R = 6378.137 km

$$a = (\text{math.sin}() ** 2) * ((\text{latitude} - \text{latCascais})/2) + (\text{math.cos}(\text{latCascais})) * (\text{math.cos}(\text{latitude})) * (\text{math.sin}() ** 2) * ((\text{longitude} - \text{lonCascais})/2)$$

```
d = 2 * R * ((sin ** -1) * (sqrt(a)))
```




```
print(d)
```

## 7. Setup a Synapse Analytics Workspace

 Deployment name: Microsoft.Azure.SynapseAnalytics-20230331144... Start time: 31-3-2023 14:50:17  
Subscription: Azure Pass - Sponsorship Correlation ID: 7f32d5dd-8aa5-414b-8f22-9352022ecf29 

Resource group: b2\_group3

### Deployment details

Resource	Type	Status	Operation details
 sailingdatabatchprocessing	Microsoft.Synapse/workspaces	Created	<a href="#">Operation details</a>
 datalakegen2sailingdata/default/sailingdata	Microsoft.Storage/storageAccounts/blobSe...	Created	<a href="#">Operation details</a>
 datalakegen2sailingdata	Microsoft.Storage/storageAccounts	OK	<a href="#">Operation details</a>

## 8. Tweaked the streaming output to appear in Synapse



9. Created an Apache Spark Pool (SparkPoolB2Gr3), this is costly but only when running.


## New Apache Spark pool ...

Create a Synapse Analytics Apache Spark pool with your preferred configurations. Complete the Basics tab then go to Review + create to provision with smart defaults, or visit each tab to customize.

### Apache Spark pool details

Name your Apache Spark pool and choose its initial settings.

Apache Spark pool name *	<input type="text" value="SparkPoolB2Gr3"/> 
Node size family	MemoryOptimized
Node size *	<input type="text" value="Small (4 vCores / 32 GB)"/> 
Autoscale * ⓘ	<input type="radio"/> Enabled <input checked="" type="radio"/> Disabled
Number of nodes *	<input type="range" value="3"/> <input type="text" value="3"/>
Dynamically allocate executors ⓘ	<input type="radio"/> Enabled <input checked="" type="radio"/> Disabled
Estimated price ⓘ	<b>Est. cost per hour</b> 1.74 EUR <a href="#">View pricing details</a>

-  Contact an **Owner** of the storage account, and verify that the following role assignments have been made:
- Assign the workspace MSI to the **Storage Blob Data Contributor** role on the storage account
  - Assign you and other users to the **Storage Blob Data Contributor** role on the storage account

Once those assignments are made, the following Spark features can be used:  
(1) Spark Library Management, (2) Read and Write data to SQL pool databases via the Spark SQL connector, and (3) Create Spark databases and tables

We decided to do the calculation and adding of the 'Distance' column in Power BI instead of in a notebook because we couldn't get it to work.

Created a Notebook that fetches only the latest data for each boat

- The Notebook creates a table
- The Notebook runs only upon a trigger
- A trigger is set to run 'something' with a time interval of 5 minutes
- The trigger is connected to the Notebook using a Pipeline