

Azure spark cluster (synapse) creation steps

Step1: Create an Azure account.

Step2: Go to azure and search for Synapse.

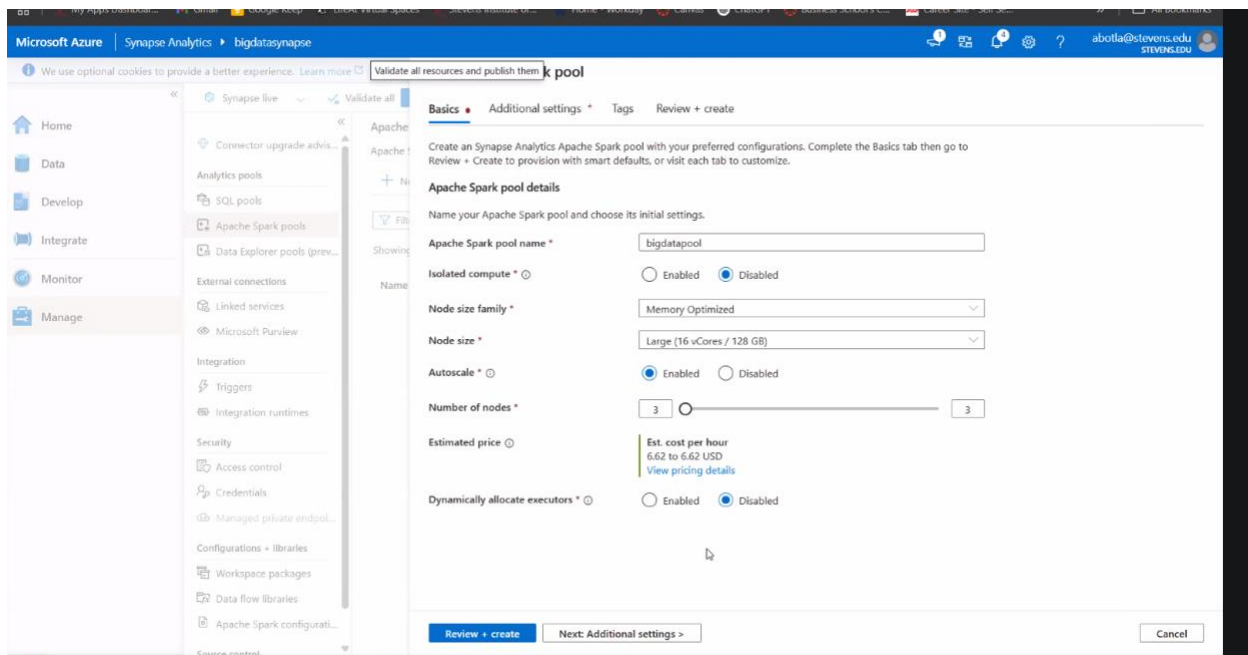
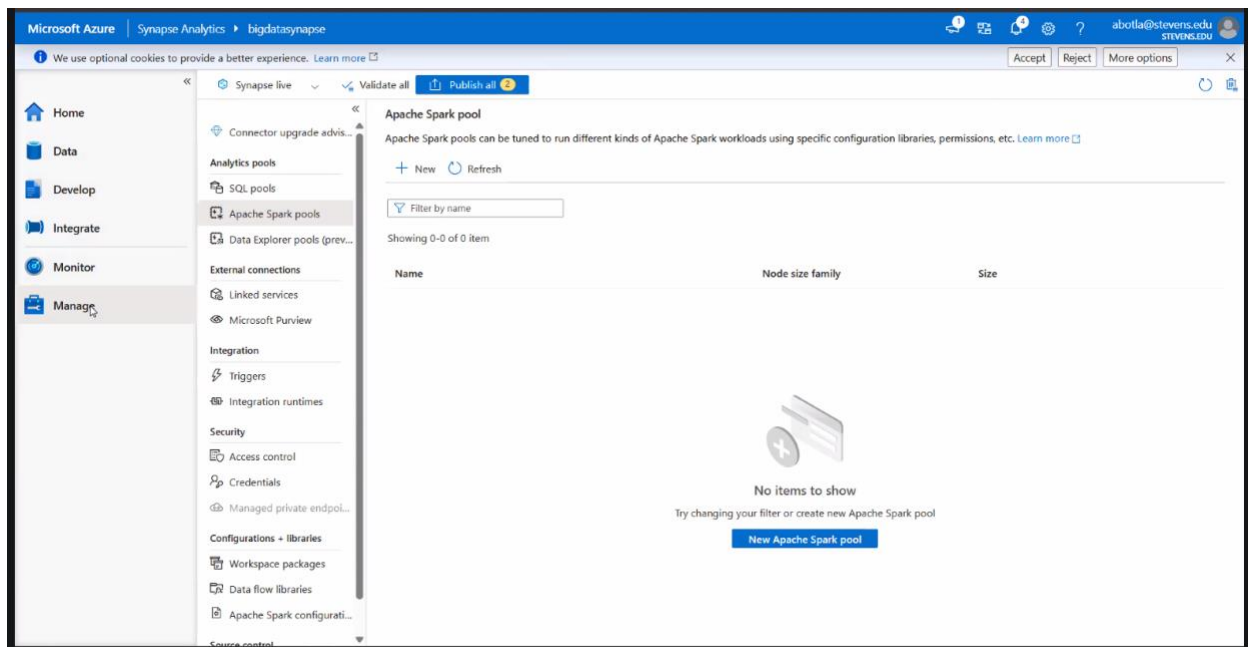
Step3: Create a Synapse resource.

The screenshot displays the Microsoft Azure portal interface. At the top, the navigation bar shows 'Microsoft Azure' and a search bar. Below the navigation bar, the breadcrumb trail indicates 'Home > Microsoft.Azure.SynapseAnalytics-20250511141348 | Overview >'. The main content area is titled 'bigdatasynapse Synapse workspace'. On the left, a sidebar lists various management options: Overview (selected), Activity log, Access control (IAM), Tags, Diagnose and solve problems, Resource visualizer, Settings, Microsoft Entra ID, Properties, Locks, Analytics pools, SQL pools, Apache Spark pools, Data Explorer pools (preview), Security, Monitoring, Automation, and Help. The 'Overview' section is expanded, showing 'Essentials' and 'Getting started' cards. The 'Essentials' section lists key properties: Resource group (Bigdata), Status (Succeeded), Location (East US), Subscription (Azure subscription 1), Subscription ID (b4318afb-f248-4c53-a67d-19c1bbe78c68), Managed virtual network (No), Managed Identity object (921384dd-0679-4594-b41b-28cd7371b6db), Workspace web URL (https://web.azure.synapse.net/workspace=%2fsubscriptions%2fb4318afb-f248-4c53-a67d-19c1bbe78c68), and Tags (Add tags). The 'Getting started' section contains two cards: 'Open Synapse Studio' and 'Read documentation'. Below these, the 'Analytics pools' section is visible, featuring a search bar and a table with columns for Name, Type, and Size.

Name	Type	Size
------	------	------

Step4: Launch the synapse workspace through Workspace URL.

Step5: Now go to manage and create an Apache Spark Pools.



Step6: Once a pool is created then go to Develop page and create/import an notebook.

Step7: Select the pool which you created.

Microsoft Azure | Synapse Analytics | bigdatasync

We use optional cookies to provide a better experience. [Learn more](#)

Accept Reject More options

Home Data Develop Integrate Monitor Manage

Develop

Filter resources by name

Notebooks

Random_Forest_2E_100D

Random_Forest_2E_100D

Run all Undo Publish Outline Attach to bigdatapool

Not started

bigdatapool
Large, 3 to 3 nodes
Manage pools

Big Data Project - Model2: Random Forest

This file contain code and result of Random Forest trained on review of books.jsonl file with 2 executor on 100 percent of actual dataset

```
1 import pandas as pd
2 import pyspark.sql.functions as F
3 from pyspark.sql.functions import col, count, when
```

--Creating and building spark session with 2 maxExecutors with 50 partition

Properties

General Related (0)

Name *
Random_Forest_2E_100D

Description

Type
.ipynb notebook

Size
75,660 bytes

Notebook settings

☒ Include cell output when saving
☐ Enable unpublished notebook reference

Session
[Configure session](#)

Step7: Use your Azure blob storage and get the SAS token like below to load the dataset.

Microsoft Azure | Synapse Analytics | bigdatasync

We use optional cookies to provide a better experience. [Learn more](#)

Accept Reject More options

Home Data Develop Integrate Monitor Manage

Develop

Filter resources by name

Notebooks

Random_Forest_2E_100D

Random_Forest_2E_100D

Run all Undo Publish Outline Attach to bigdatapool Language PySpark (Python) Variables

Not started

Big Data Project - Model2: Random Forest

This file contain code and result of Random Forest trained on review of books.jsonl file with 2 executor on 100 percent of actual dataset

```
1 import pandas as pd
2 import pyspark.sql.functions as F
3 from pyspark.sql.functions import col, count, when
```

```
1 json_url = "https://bigdatastorage.blob.core.windows.net/bigdatacontainer/books.json.json?"
2
3 df = spark.read.option("multiline", True).json(json_url)
4
```

+ Code + Markdown

Properties

General Related (0)

Name *
Random_Forest_2E_100D

Description

Type
.ipynb notebook

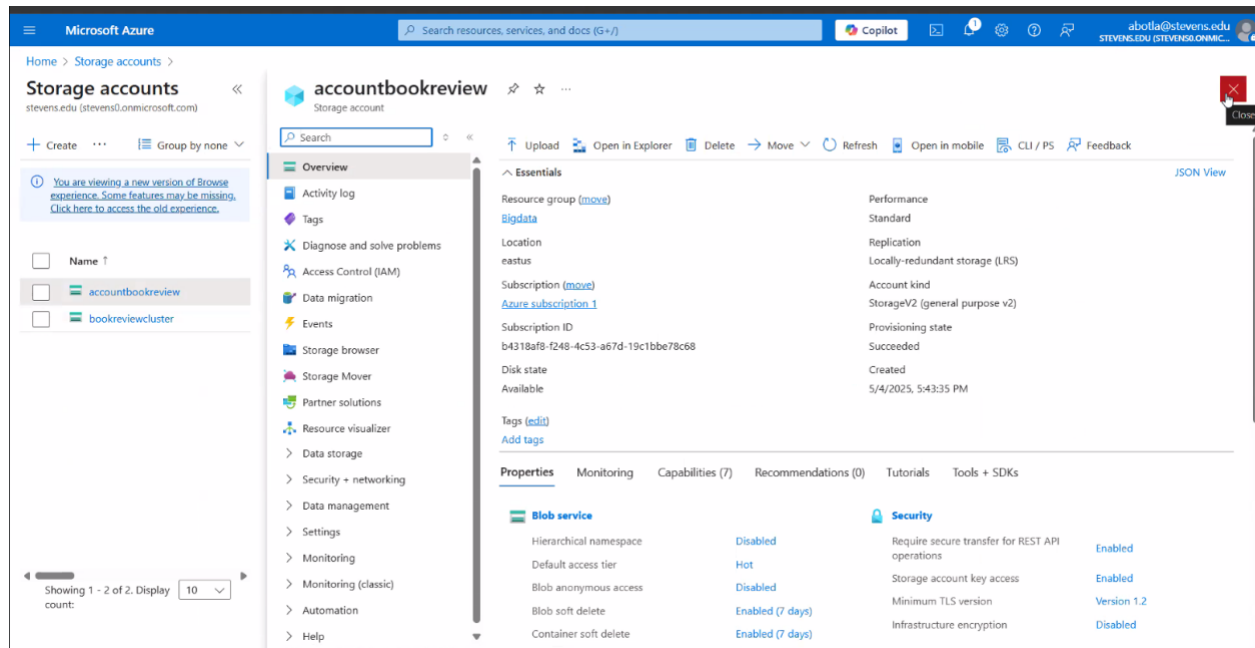
Size
75,919 bytes

Notebook settings

☒ Include cell output when saving
☐ Enable unpublished notebook reference

Session
[Configure session](#)

If you have not loaded the dataset then load the dataset in Azure storage account and it should be like below.



Step8: With the Spark pool configured and the data loaded, you can now run your PySpark code successfully within the Synapse notebook environment.

Thank You