Spark Job in Synapse Spark Pool

I have taken a twitter dataset from Kaggle, and downloaded it.
https://www.kaggle.com/datasets/kazanova/sentiment140?resource=download

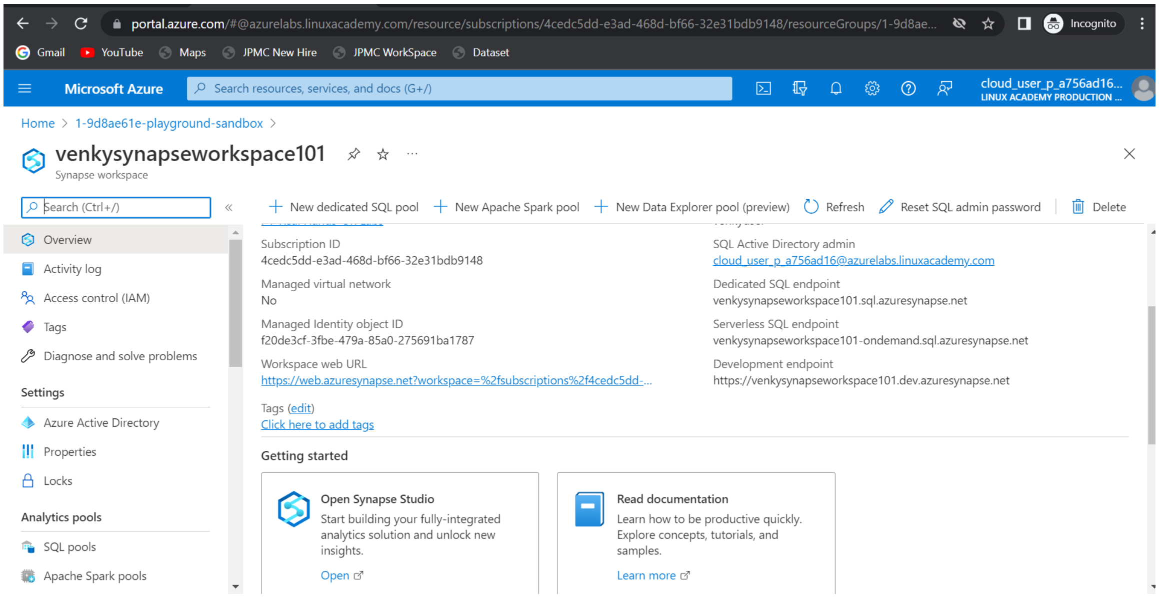
I have a spark program that takes in this CSV and loads it into avro and then reads the avro dataset, and writes out a parquet file.

The code is here : https://github.com/SowmyaVenky/Azure-DP-203/blob/main/SparkExamples/src/main/java/com/gssystems/spark/AvroInsideSynapse.java

Create a Synapse analytics workspace. The following powershell script can use an ARM template and creates the required synapse spark pools.

https://github.com/SowmyaVenky/Azure-DP-203/blob/main/1005-Create-Synapse-workspace.ps1

Once the powershell us submitted, it uses the ARM template and creates a synapse workspace.



portal.azure.com/#@azurelabs.linuxacademy.com/resource/subscriptions/4cedc5dd-e3ad-468d-bf66-32e31bdb9148/resourceGroups/1-9d8ae...

Home > 1-9d8ae61e-playground-sandbox >

venkysynapseworkspace101
Synapse workspace

Subscription ID
4cedc5dd-e3ad-468d-bf66-32e31bdb9148

Managed virtual network
No

Managed Identity object ID
f20de3cf-3fbe-479a-85a0-275691ba1787

Workspace web URL
https://web.azuresynapse.net?workspace=%2fsubscriptions%2f4cedc5dd-...

SQL Active Directory admin
cloud_user_p_a756ad16@azurelabs.linuxacademy.com

Dedicated SQL endpoint
venkysynapseworkspace101.sql.azuresynapse.net

Serverless SQL endpoint
venkysynapseworkspace101-ondemand.sql.azuresynapse.net

Development endpoint
https://venkysynapseworkspace101.dev.azuresynapse.net

Tags (edit)
Click here to add tags

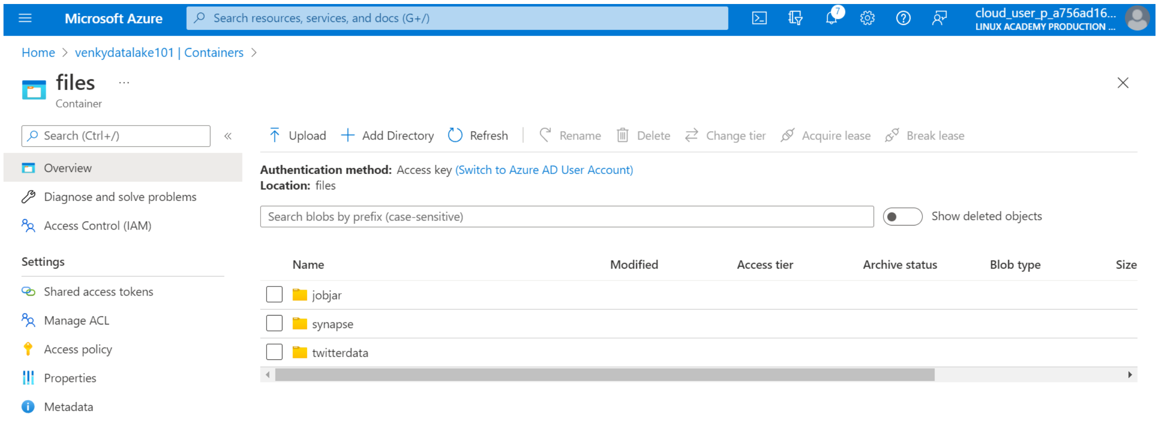
Getting started

Open Synapse Studio
Start building your fully-integrated analytics solution and unlock new insights.
Open

Read documentation
Learn how to be productive quickly. Explore concepts, tutorials, and samples.
Learn more

I have compiled the code with JDK 1.8. If we compile the code with a higher level JDK, we will get an error when we submit the job inside Synapse Spark.

I have uploaded the jar file after the maven package into the ADLS storage



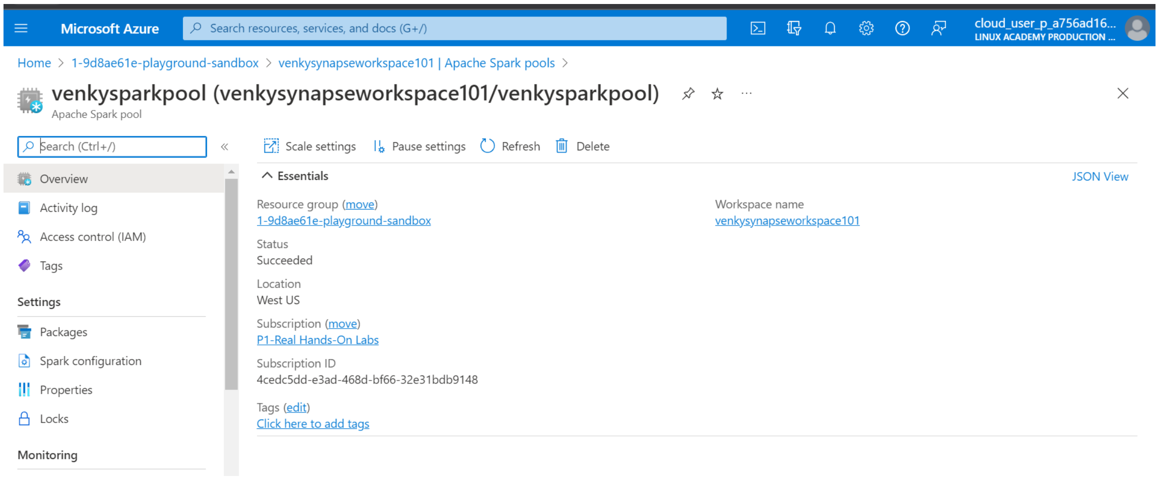
Authentication method: Access key (Switch to Azure AD User Account)
Location: files

	Name	Modified	Access tier	Archive status	Blob type	Size
□	jobjar					
□	synapse					
□	twitterdata					

I have also uploaded the csv file that we downloaded from kaggle, and put it in the twitterdata folder.

Once we have these 2 things in place, we are ready to fire the spark job from our computer.

Go to Synapse Studio, and note down the parameters required, workspace name, spark pool name, and use these to fire the spark job.



Resource group (move)
1-9d8ae61e-playground-sandbox

Status
Succeeded

Location
West US

Subscription (move)
P1-Real Hands-On Labs

Subscription ID
4cedc5dd-e3ad-468d-bf66-32e31bdb9148

Tags (edit)
Click here to add tags

Workspace name
venkysynapseworkspace101

As we can see venkysynapseworkspace101 and venkysparkpool are the relevant parameters here.

We can submit the job directly from powershell using this command.

```
Synapse wants a jar with JDK 1.8


set JAVA_HOME=C:\Venky\jdk-8.0.342.07-hotspot

set PATH=%PATH%;c:\Venky\spark\bin;c:\Venky\apache-maven-3.8.4\bin
set SPARK_HOME=c:\Venky\spark
```

```
SET HADOOP_HOME=C:\Venky\DP-203\Azure-DP-203\SparkExamples
```

```
mvn clean package (builds the jar required to upload).
```

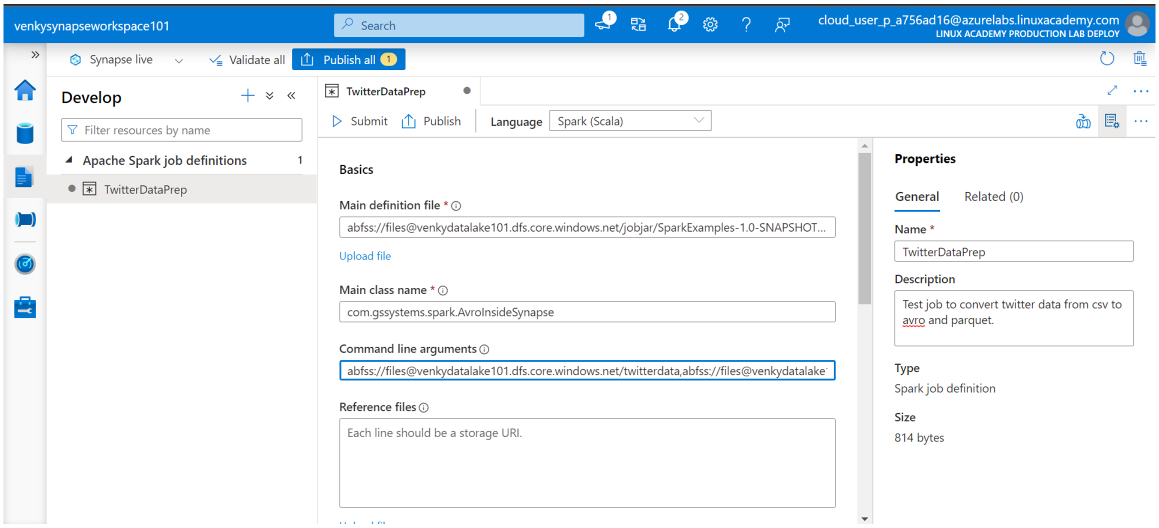
```
Submit-AzSynapseSparkJob -WorkspaceName venkysynapseworkspace101
-SparkPoolName venkysparkpool -Language Spark -Name TwitterDataPrep
-MainDefinitionFile
abfss://files@venkydatalake101.dfs.core.windows.net/jobjar/SparkExamples-1.0-
SNAPSHOT.jar -MainClassName com.gssystems.spark.AvroInsideSynapse
-CommandLineArgument
abfss://files@venkydatalake101.dfs.core.windows.net/twitterdata,abfss://files@
venkydatalake101.dfs.core.windows.net/twitterdataavro,abfss://files@venkydatal
ake101.dfs.core.windows.net/twitterdataparquet -ExecutorCount 2 -ExecutorSize
Small
```

Once we do this, we can check the status of the job.

```
Synapse upload file and submit spark job
```

```
Get-AzSynapseSparkJob -WorkspaceName venkysynapseworkspace101 -SparkPoolName
venkysparkpool
```

If we want to do it manually, we can define the job on the synapse studio UI



Language: Spark (Scala)

Basics

Main definition file *: abfss://files@venkydatalake101.dfs.core.windows.net/jobjar/SparkExamples-1.0-SNAPSHOT...

Main class name *: com.gssystems.spark.AvroInsideSynapse

Command line arguments: abfss://files@venkydatalake101.dfs.core.windows.net/twitterdata,abfss://files@venkydatalake

Reference files: Each line should be a storage URI.

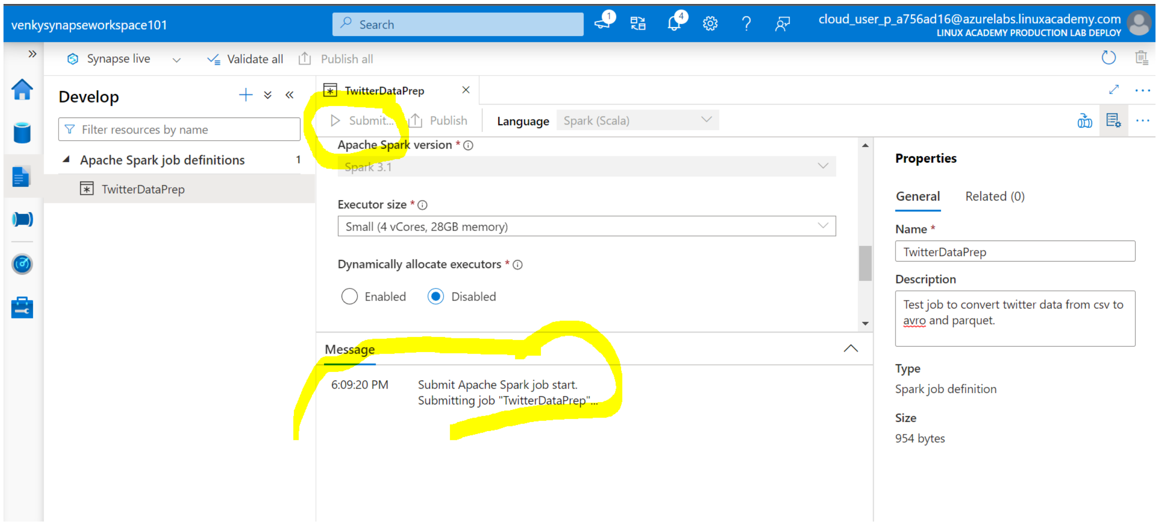
Name *: TwitterDataPrep

Description: Test job to convert twitter data from csv to avro and parquet.

Type: Spark job definition

Size: 814 bytes

We can select the spark pool to use, and publish



Apache Spark version * Spark 3.1

Executor size * Small (4 vCores, 28GB memory)

Dynamically allocate executors * ○ Enabled ● Disabled

Message

6:09:20 PM Submit Apache Spark job start.
Submitting job "TwitterDataPrep"...

Properties

General | Related (0)

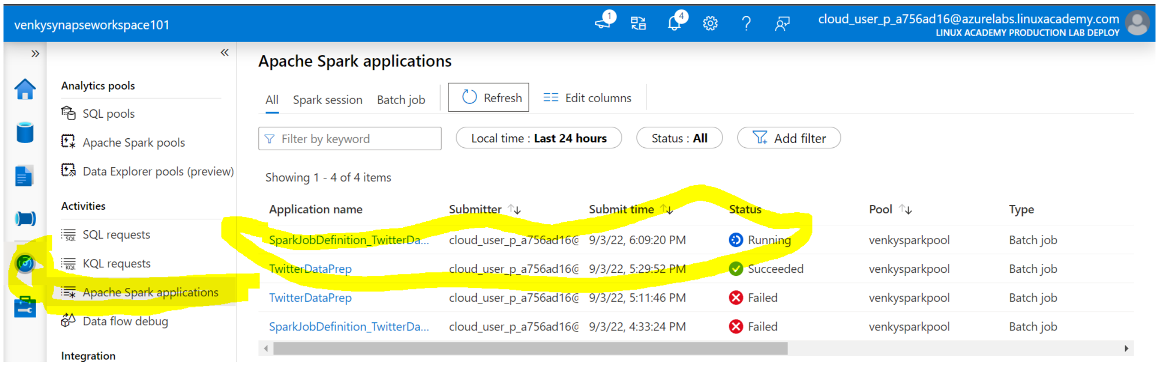
Name * TwitterDataPrep

Description
Test job to convert twitter data from csv to avro and parquet.

Type
Spark job definition

Size
954 bytes

Once published, we can submit it.

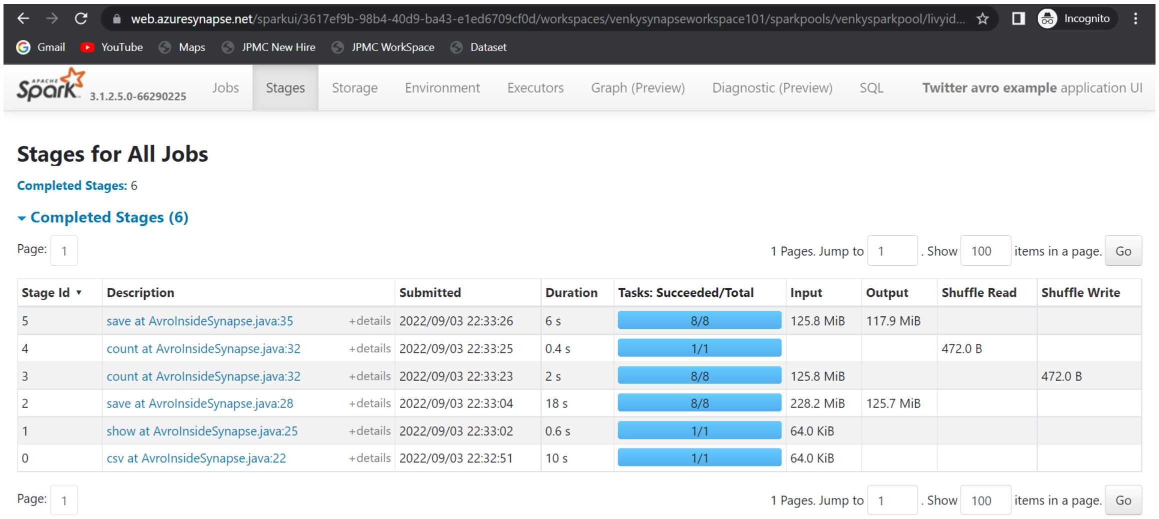


Apache Spark applications

Showing 1 - 4 of 4 items

Application name	Submitter	Submit time	Status	Pool	Type
SparkJobDefinition_TwitterDa...	cloud_user_p_a756ad16@	9/3/22, 6:09:20 PM	Running	venkysparkpool	Batch job
TwitterDataPrep	cloud_user_p_a756ad16@	9/3/22, 5:29:52 PM	✓ Succeeded	venkysparkpool	Batch job
TwitterDataPrep	cloud_user_p_a756ad16@	9/3/22, 5:11:46 PM	☒ Failed	venkysparkpool	Batch job
SparkJobDefinition_TwitterDa...	cloud_user_p_a756ad16@	9/3/22, 4:33:24 PM	☒ Failed	venkysparkpool	Batch job

We can see the spark UI by launching it from the job line.


web.azuresynapse.net/sparkui/3617ef9b-98b4-40d9-ba43-e1ed6709cf0d/workspaces/venkysynapseworkspace101/sparkpools/venkysparkpool/livyid...
Incognito
Gmail
YouTube
Maps
JPMC New Hire
JPMC WorkSpace
Dataset
3.1.2.5.0-66290225
Jobs
Stages
Storage
Environment
Executors
Graph (Preview)
Diagnostic (Preview)
SQL
Twitter avro example application UI

Stages for All Jobs

Completed Stages: 6

Completed Stages (6)

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

Stage Id ▾	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
5	save at AvroInsideSynapse.java:35 +details	2022/09/03 22:33:26	6 s	8/8	125.8 MiB	117.9 MiB		
4	count at AvroInsideSynapse.java:32 +details	2022/09/03 22:33:25	0.4 s	1/1			472.0 B	
3	count at AvroInsideSynapse.java:32 +details	2022/09/03 22:33:23	2 s	8/8	125.8 MiB			472.0 B
2	save at AvroInsideSynapse.java:28 +details	2022/09/03 22:33:04	18 s	8/8	228.2 MiB	125.7 MiB		
1	show at AvroInsideSynapse.java:25 +details	2022/09/03 22:33:02	0.6 s	1/1	64.0 KiB			
0	csv at AvroInsideSynapse.java:22 +details	2022/09/03 22:32:51	10 s	1/1	64.0 KiB			

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go