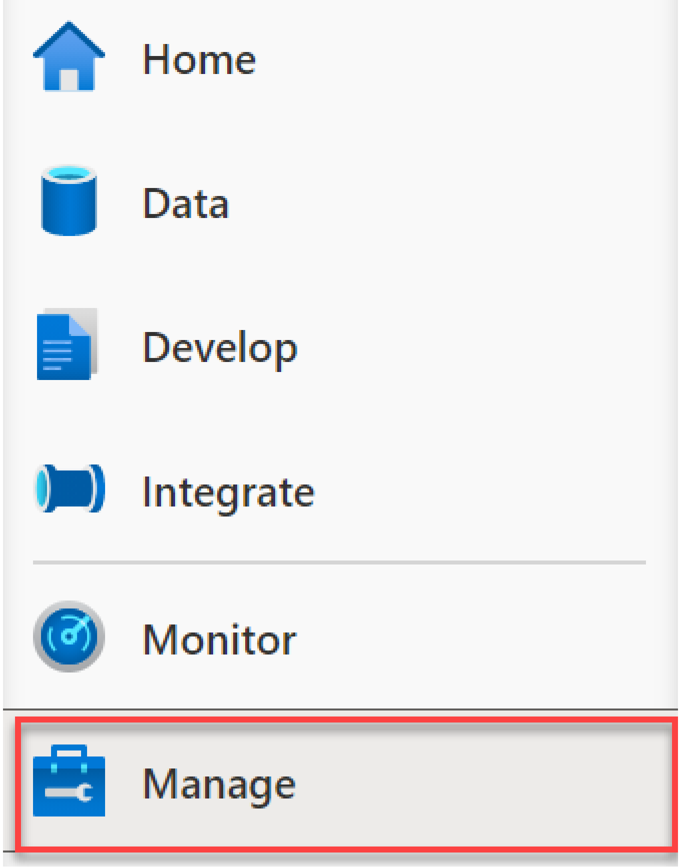
Working with Synapse Spark

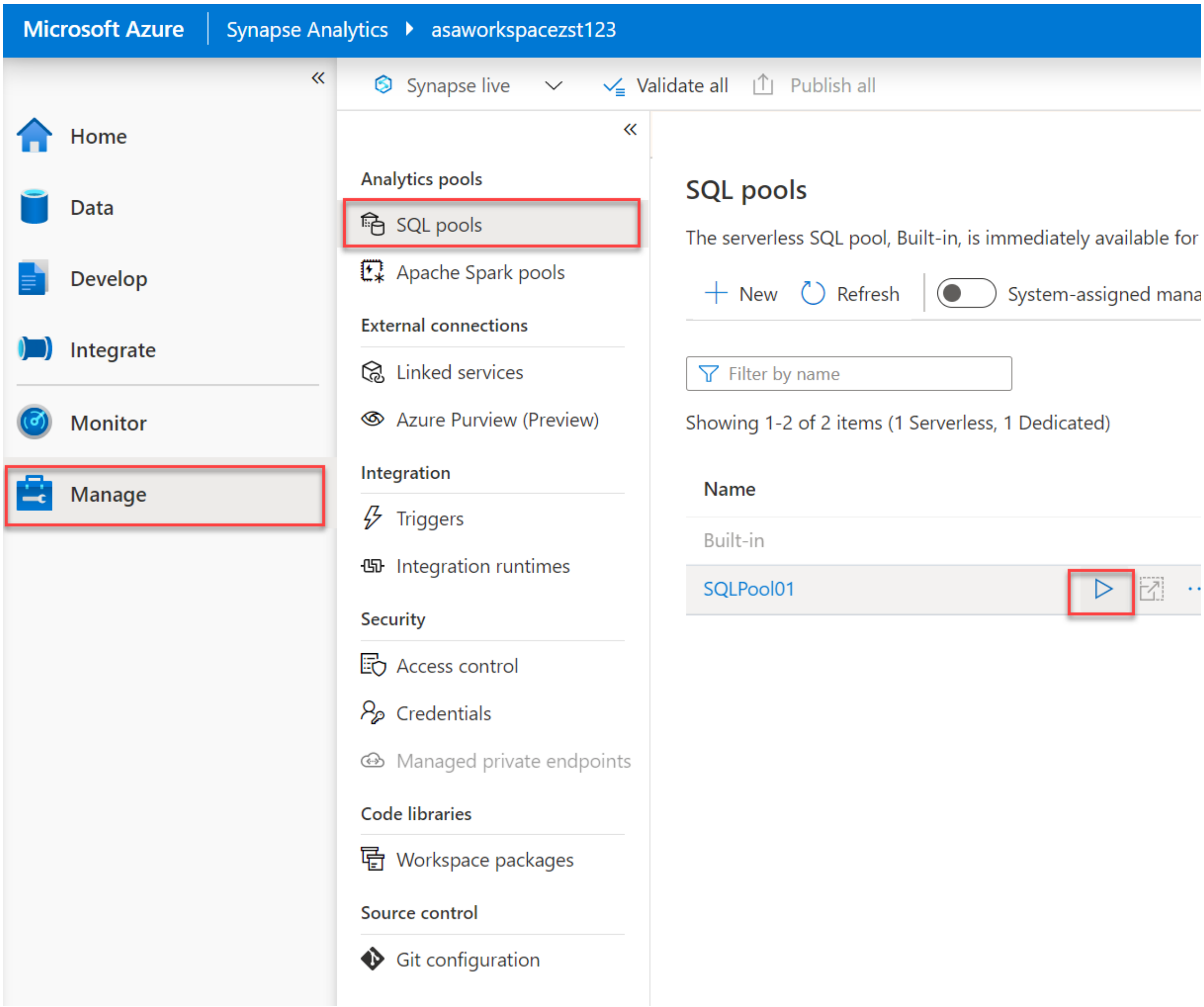
Lab pre-requisite

Start the SQL Pool in your lab environment.

1. Open the Synapse Studio workspace and navigate to the **Manage** hub.



2. From the center menu, select **SQL pools** from beneath the **Analytics pools** heading. Locate `SQLPool01`, and select the **Resume** button.



Exercise 1 - Working with Spark DataFrames in Synapse Spark

In this exercise you will learn how to work with Spark DataFrames in Synapse Spark, including:

- Working with schemas and lake databases
- Performing dataframe operations
- Working with dataframe partitions

1. Open Synapse Analytics Studio, and then navigate to the `Develop` hub.
2. Under **Notebooks**, select the notebook called `Lab 07 - Part 1 - Spark DataFrames`.
3. Read through the notebook and execute the cells as instructed in the notebook. When you have finished in the notebook, you have completed this lab.

> **IMPORTANT!**
>
> Once you complete the steps in the notebook, make sure you stop the Spark session when closing the notebook. This will free up the necessary compute resources to start the Spark sessions for the other exercises in this lab.

Exercise 2 - Delta Lake features in Synapse Spark

In this exercise you will learn how to work with Delta Lake and `mssparkutils` in Synapse Spark.

1. Open Synapse Analytics Studio, and then navigate to the `Develop` hub.

2. Under **Notebooks**, select the notebook called `Lab 07 - Part 2 - Spark Delta Lake`.

3. Read through the notebook and execute the cells as instructed in the notebook. When you have finished in the notebook, you have completed this lab.

> **IMPORTANT!**
>
> Once you complete the steps in the notebook, make sure you stop the Spark session when closing the notebook. This will free up the necessary compute resources to start the Spark sessions for the other exercises in this lab.

Exercise 3 - Indexing in Synapse Spark with Hyperspace

In this exercise you will learn how to work with Hyperspace in Synapse Spark.

1. Open Synapse Analytics Studio, and then navigate to the `Develop` hub.

2. Under **Notebooks**, select the notebook called `Lab 07 - Part 3 - Spark Hyperspace`.

3. Read through the notebook and execute the cells as instructed in the notebook. When you have finished in the notebook, you have completed this lab.

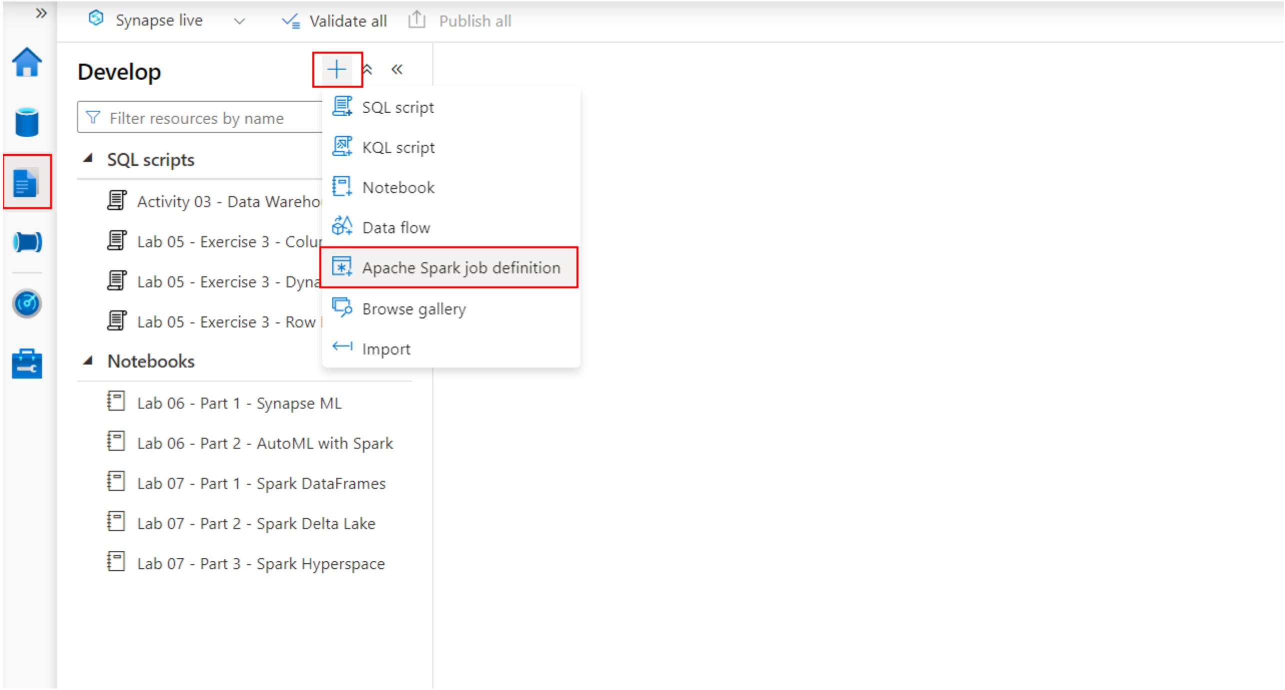
> **IMPORTANT!**
>
> Once you complete the steps in the notebook, make sure you stop the Spark session when closing the notebook. This will free up the necessary compute resources to start the Spark sessions for the other exercises in this lab.

Exercise 4 - Working with Synapse Spark job definitions

In this exercise you will learn how to create and run a Spark job in Synapse Spark. The job will perform the task of counting the words in a text file stored in the Synapse workspace data lake storage.

1. Open Synapse Analytics Studio, and then navigate to the `Develop` hub.

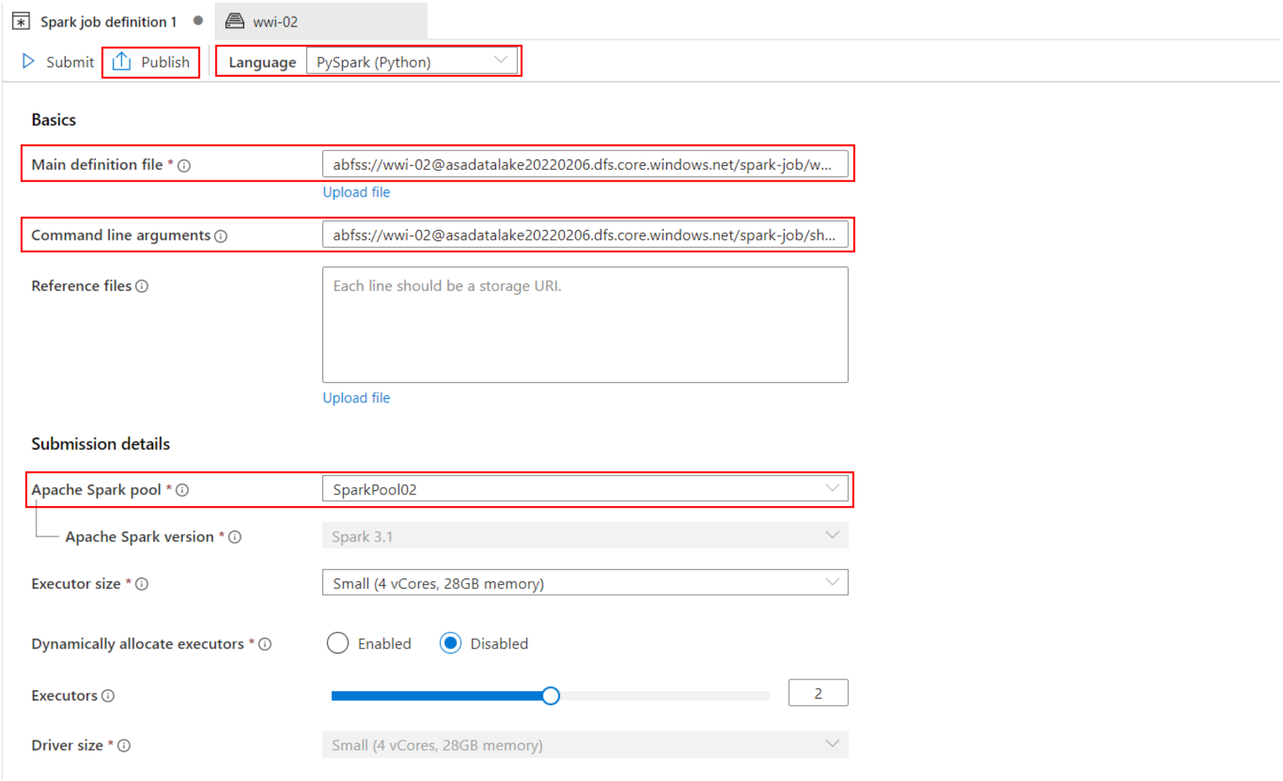
2. Select `+` and then select `Apache Spark job definition` to initiate the creation of a new Spark job.


Synapse live
Validate all
Publish all
Develop
Filter resources by name
SQL scripts
Activity 03 - Data Wareho
Lab 05 - Exercise 3 - Colu
Lab 05 - Exercise 3 - Dyna
Lab 05 - Exercise 3 - Row
Notebooks
Lab 06 - Part 1 - Synapse ML
Lab 06 - Part 2 - AutoML with Spark
Lab 07 - Part 1 - Spark DataFrames
Lab 07 - Part 2 - Spark Delta Lake
Lab 07 - Part 3 - Spark Hyperspace
SQL script
KQL script
Notebook
Data flow
Apache Spark job definition
Browse gallery
Import

3. In the Spark job definition form, fill in the following properties:

   - Language: `PySpark (Python)`
   - Main definition file: `abfss://wwi-02@<your_data_lake_account_name>.dfs.core.windows.net/spark-job/wordcount.py` (where `<your_data_lake_account_name>` is the name of the Synapse workspace data lake account configured in your environment)
   - Command line arguments: `abfss://wwi-02@<your_data_lake_account_name>.dfs.core.windows.net/spark-job/shakespeare.txt abfss://wwi-02@<your_data_lake_account_name>.dfs.core.windows.net/spark-job/result` (where `<your_data_lake_account_name>` is the name of the Synapse workspace data lake account configured in your environment)
   - Apake Spark pool: `SparkPool02`

   Once all the properties mentioned above are filled in, select `Publish` to publish the new Spark job.



Spark job definition 1 ● | wwi-02

▷ Submit | Publish | Language: PySpark (Python)

Basics

Main definition file *: abfss://wwi-02@asadatalake20220206.dfs.core.windows.net/spark-job/w...
Upload file

Command line arguments: abfss://wwi-02@asadatalake20220206.dfs.core.windows.net/spark-job/sh...

Reference files: Each line should be a storage URI.
Upload file

Submission details

Apache Spark pool *: SparkPool02

Apache Spark version *: Spark 3.1

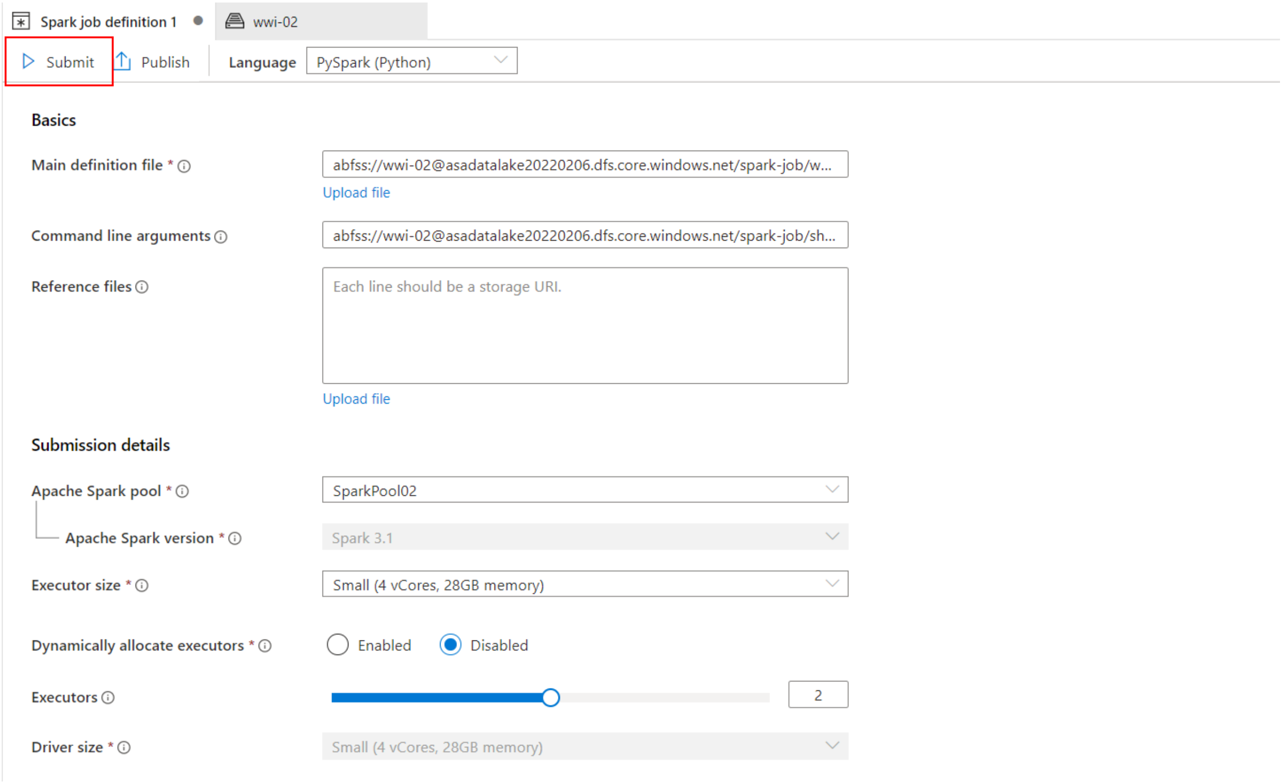
Executor size *: Small (4 vCores, 28GB memory)

Dynamically allocate executors *: ○ Enabled ● Disabled

Executors: 2

Driver size *: Small (4 vCores, 28GB memory)

4. When the publishing is finished, select `Submit` to start the new Spark job.



Spark job definition 1 ● | wwi-02

▷ Submit | Publish | Language: PySpark (Python)

Basics

Main definition file *: abfss://wwi-02@asadatalake20220206.dfs.core.windows.net/spark-job/w...
Upload file

Command line arguments: abfss://wwi-02@asadatalake20220206.dfs.core.windows.net/spark-job/sh...

Reference files: Each line should be a storage URI.
Upload file

Submission details

Apache Spark pool *: SparkPool02

Apache Spark version *: Spark 3.1

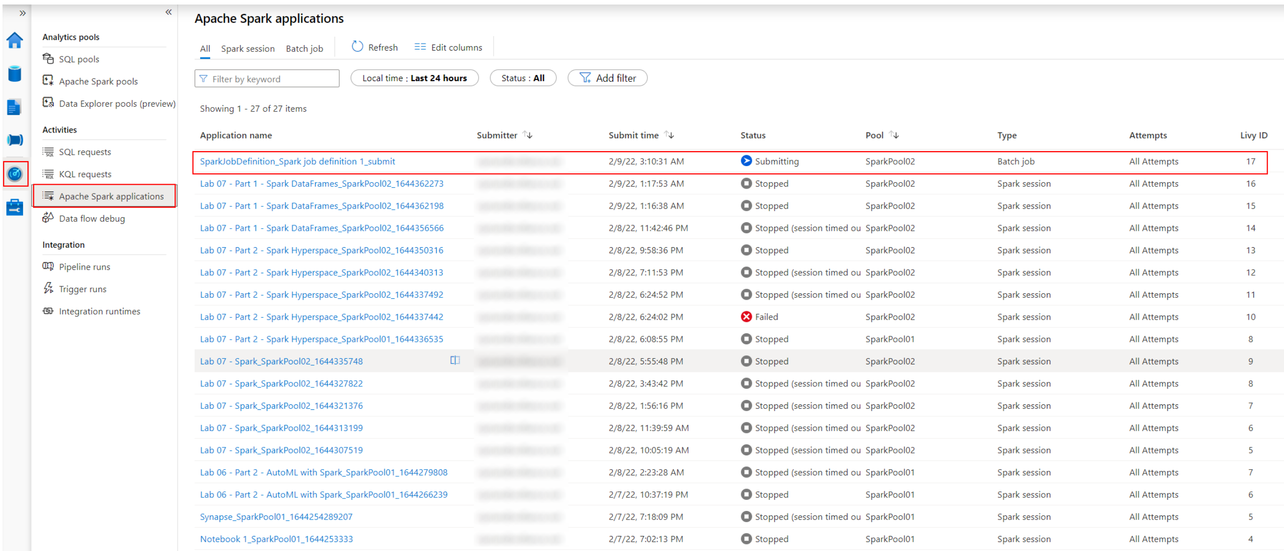
Executor size *: Small (4 vCores, 28GB memory)

Dynamically allocate executors *: ○ Enabled ● Disabled

Executors: 2

Driver size *: Small (4 vCores, 28GB memory)

5. Navigate to the `Monitor` hub and select the `Apache Spark applications` section. Identify the Spark application corresponding to your Spark job.

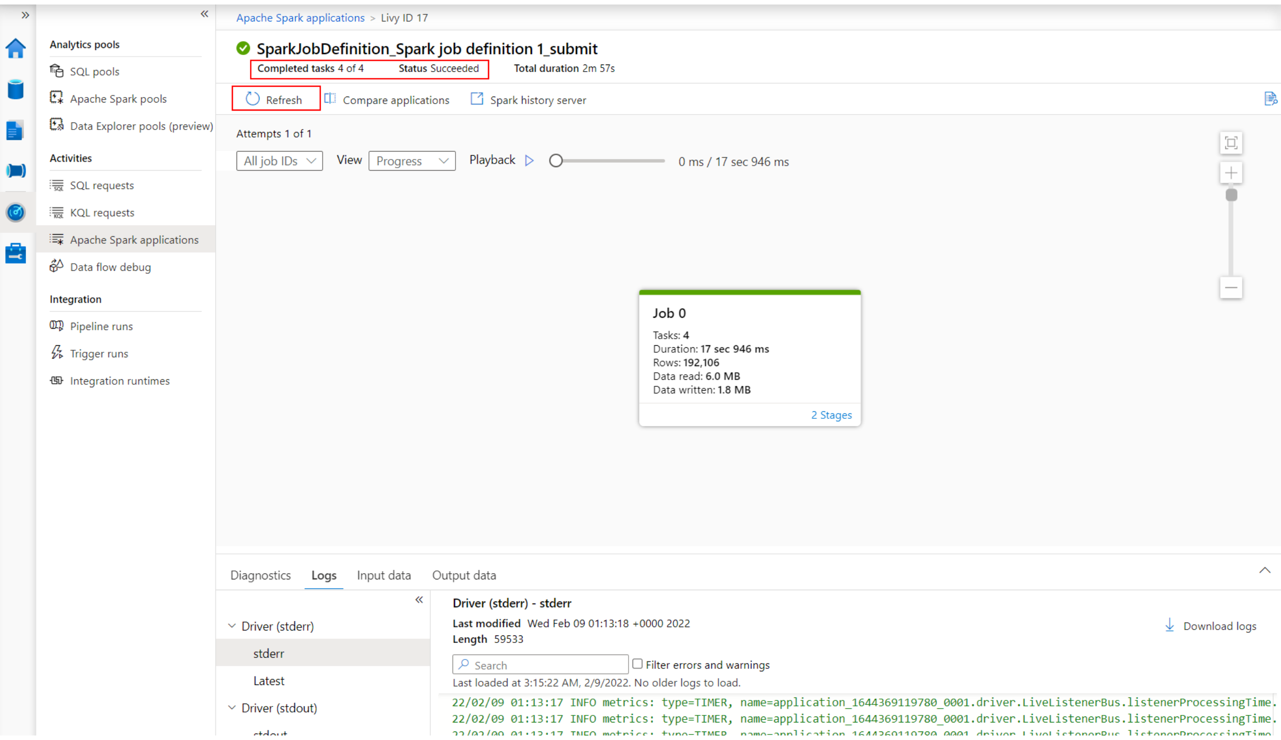


Apache Spark applications

Showing 1 - 27 of 27 items

Application name	Submitter	Submit time	Status	Pool	Type	Attempts	Livy ID
SparkJobDefinition_Spark job definition 1_submit	[redacted]	2/9/22, 3:10:31 AM	Submitting	SparkPool02	Batch job	All Attempts	17
Lab 07 - Part 1 - Spark DataFrames_SparkPool02_1644362273	[redacted]	2/9/22, 1:17:53 AM	Stopped	SparkPool02	Spark session	All Attempts	16
Lab 07 - Part 1 - Spark DataFrames_SparkPool02_1644362198	[redacted]	2/9/22, 1:16:38 AM	Stopped	SparkPool02	Spark session	All Attempts	15
Lab 07 - Part 1 - Spark DataFrames_SparkPool02_1644356566	[redacted]	2/8/22, 11:42:46 PM	Stopped (session timed ou	SparkPool02	Spark session	All Attempts	14
Lab 07 - Part 2 - Spark Hyperspace_SparkPool02_1644350316	[redacted]	2/8/22, 9:58:36 PM	Stopped	SparkPool02	Spark session	All Attempts	13
Lab 07 - Part 2 - Spark Hyperspace_SparkPool02_1644340313	[redacted]	2/8/22, 7:11:53 PM	Stopped (session timed ou	SparkPool02	Spark session	All Attempts	12
Lab 07 - Part 2 - Spark Hyperspace_SparkPool02_1644337492	[redacted]	2/8/22, 6:24:52 PM	Stopped (session timed ou	SparkPool02	Spark session	All Attempts	11
Lab 07 - Part 2 - Spark Hyperspace_SparkPool02_1644337442	[redacted]	2/8/22, 6:24:02 PM	Failed	SparkPool02	Spark session	All Attempts	10
Lab 07 - Part 2 - Spark Hyperspace_SparkPool01_1644336535	[redacted]	2/8/22, 6:08:55 PM	Stopped	SparkPool01	Spark session	All Attempts	8
Lab 07 - Spark_SparkPool02_1644335748	[redacted]	2/8/22, 5:55:48 PM	Stopped	SparkPool02	Spark session	All Attempts	9
Lab 07 - Spark_SparkPool02_1644327822	[redacted]	2/8/22, 3:43:42 PM	Stopped (session timed ou	SparkPool02	Spark session	All Attempts	8
Lab 07 - Spark_SparkPool02_1644321376	[redacted]	2/8/22, 1:56:16 PM	Stopped (session timed ou	SparkPool02	Spark session	All Attempts	7
Lab 07 - Spark_SparkPool02_1644313199	[redacted]	2/8/22, 11:39:59 AM	Stopped (session timed ou	SparkPool02	Spark session	All Attempts	6
Lab 07 - Spark_SparkPool02_1644307519	[redacted]	2/8/22, 10:05:19 AM	Stopped (session timed ou	SparkPool02	Spark session	All Attempts	5
Lab 06 - Part 2 - AutoML with Spark_SparkPool01_1644279808	[redacted]	2/8/22, 2:23:28 AM	Stopped (session timed ou	SparkPool01	Spark session	All Attempts	7
Lab 06 - Part 2 - AutoML with Spark_SparkPool01_1644266239	[redacted]	2/7/22, 10:37:19 PM	Stopped	SparkPool01	Spark session	All Attempts	6
Synapse_SparkPool01_1644254289207	[redacted]	2/7/22, 7:18:09 PM	Stopped (session timed ou	SparkPool01	Spark session	All Attempts	5
Notebook 1_SparkPool01_1644253333	[redacted]	2/7/22, 7:02:13 PM	Stopped	SparkPool01	Spark session	All Attempts	4

6. Select the Spark application corresponding to your job and wait until it finishes (you might need to select `Refresh` every minute or so to update the status).



SparkJobDefinition_Spark job definition 1_submit

Completed tasks 4 of 4 Status Succeeded Total duration 2m 57s

Attempts 1 of 1

Job 0
Tasks: 4
Duration: 17 sec 946 ms
Rows: 192,106
Data read: 6.0 MB
Data written: 1.8 MB

Driver (stderr) - stderr
Last modified Wed Feb 09 01:13:18 +0000 2022
Length 59533

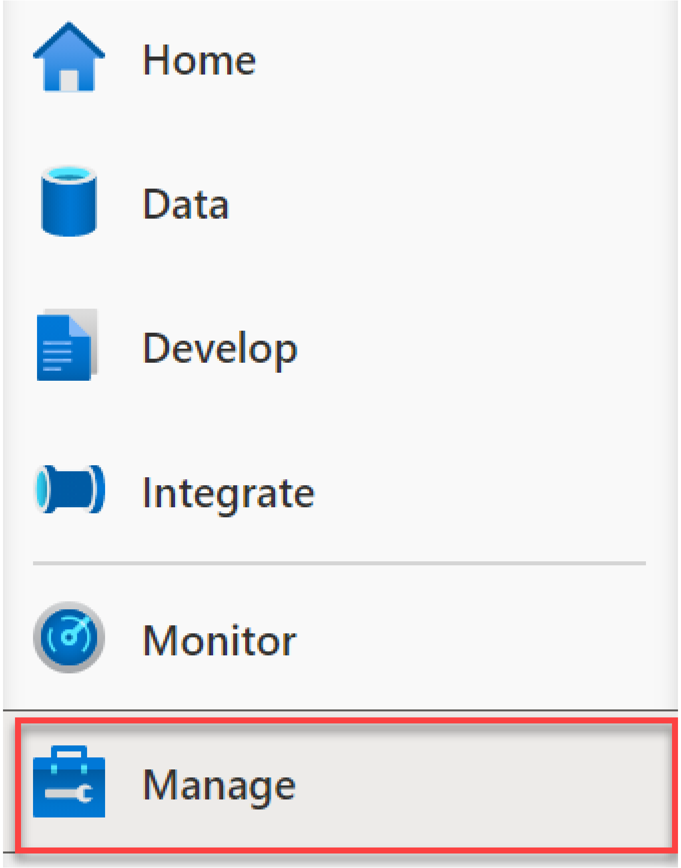
Last loaded at 3:15:22 AM, 2/9/2022. No older logs to load.

```
22/02/09 01:13:17 INFO metrics: type=TIMER, name=application_1644369119780_0001.driver.LiveListenerBus.listenerProcessingTime.
22/02/09 01:13:17 INFO metrics: type=TIMER, name=application_1644369119780_0001.driver.LiveListenerBus.listenerProcessingTime.
```

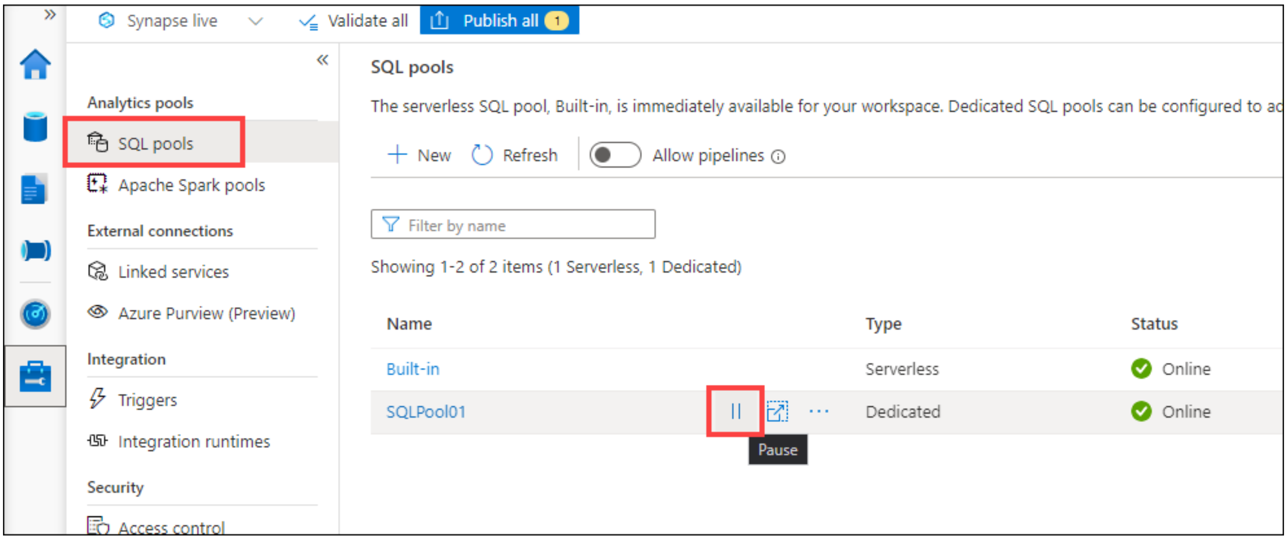
7. Once the Spark job finishes successfully, check the `/spark-job/result` folder located in the `wwi-02` container on the Synapse workspace data lake storage account. The files in the folder are text files containing the word counting results.

Cleanup: Pause the dedicated SQL pool

1. Navigate to the **Manage** hub.



2. From the center menu, select **SQL pools** from beneath the **Analytics pools** heading. Locate `SQLPool01`, and select the **Pause** button.



The serverless SQL pool, Built-in, is immediately available for your workspace. Dedicated SQL pools can be configured to ac

Showing 1-2 of 2 items (1 Serverless, 1 Dedicated)

Name	Type	Status
Built-in	Serverless	Online
SQLPool01	Dedicated	Online

3. When prompted, select **Pause**.