Lab 04: Create and schedule a Job

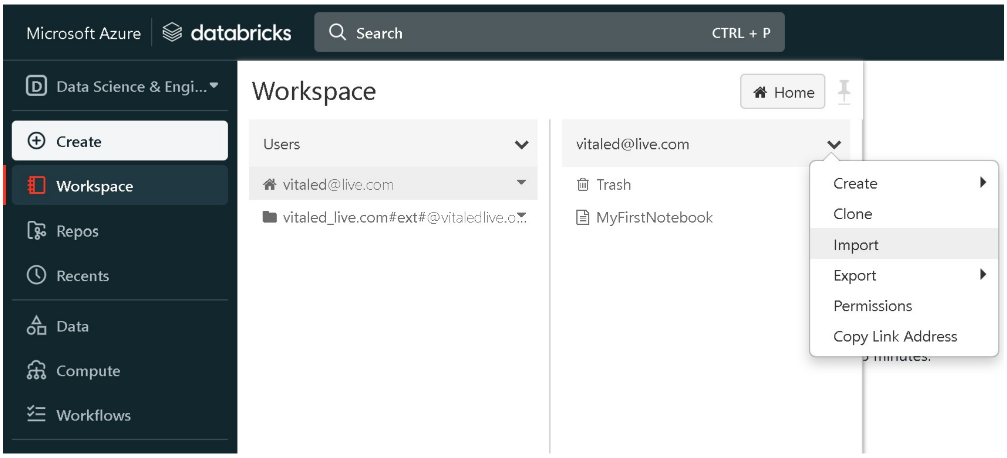
Goal

During this lab you will learn how to create and schedule a Job in a Azure Databricks

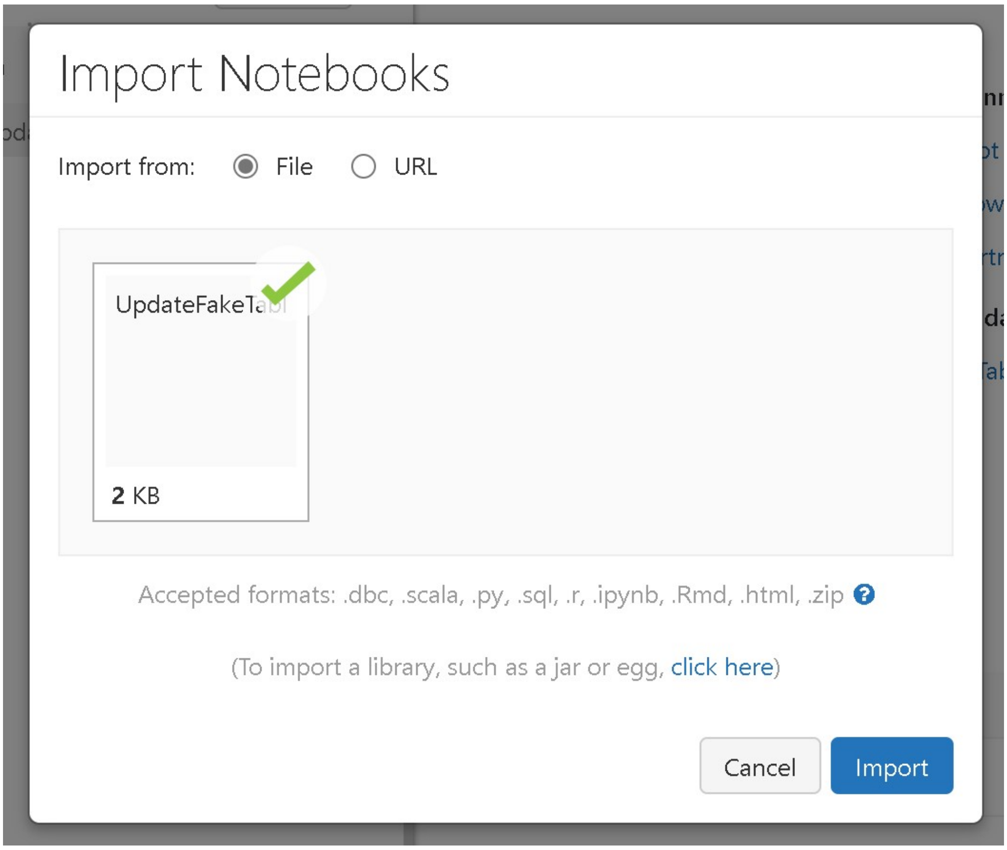
Tasks

Task 1: Import notebook to schedule

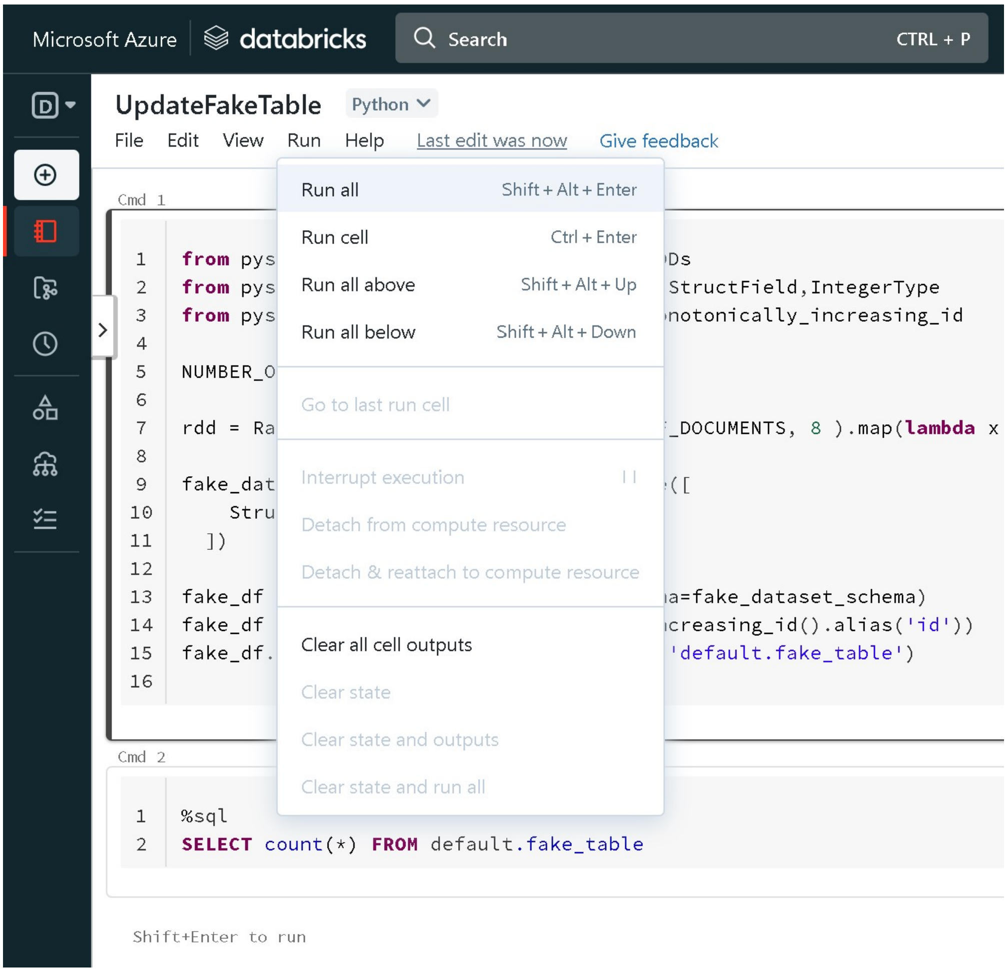
Click on **Workspace** navigate to your home directory and click on **Import**



Click on browse, select the file *UpdateFakeTable.ipynb* and click on import


Import Notebooks
Import from: ● File ○ URL
UpdateFakeTabl
2 KB
Accepted formats: .dbc, .scala, .py, .sql, .r, .ipynb, .Rmd, .html, .zip
(To import a library, such as a jar or egg, click here)
Cancel
Import

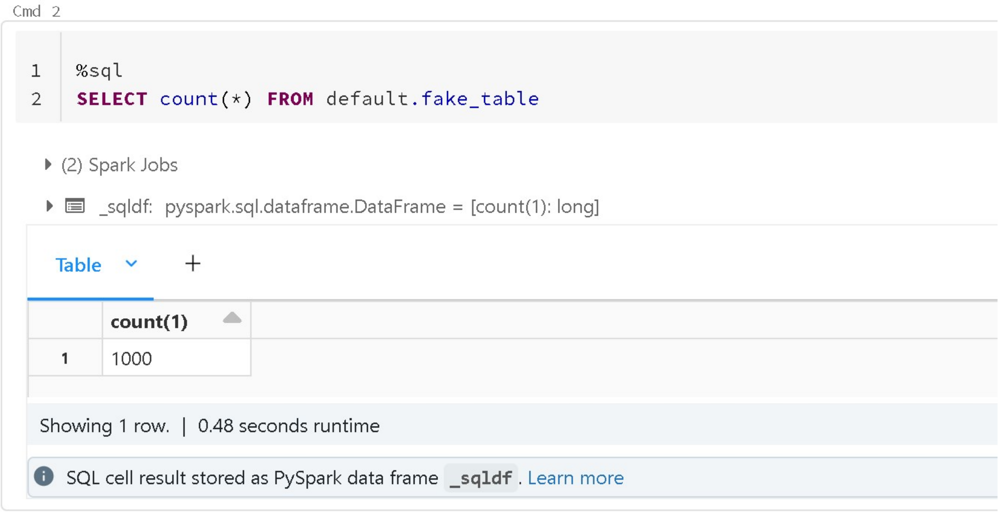
In the notebook page click on **Run** and then **Run all**



```
from pys                                   Ds
from pys                                   StructField,IntegerType
from pys                                   notonically_increasing_id

NUMBER_O

rdd = Ra                                   _DOCUMENTS, 8 ).map(lambda x

fake_dat                                   ([
    Stru
  ])

fake_df                                    a=fake_dataset_schema)
fake_df                                    creasing_id().alias('id'))
fake_df.                                   'default.fake_table')
```

```
%sql
SELECT count(*) FROM default.fake_table
```

This first execution will create the table *default.fake_table* and insert 1000 entries you can check this by looking at the result of the count statement



```
%sql
SELECT count(*) FROM default.fake_table
```

(2) Spark Jobs

_sqldf: pyspark.sql.dataframe.DataFrame = [count(1): long]

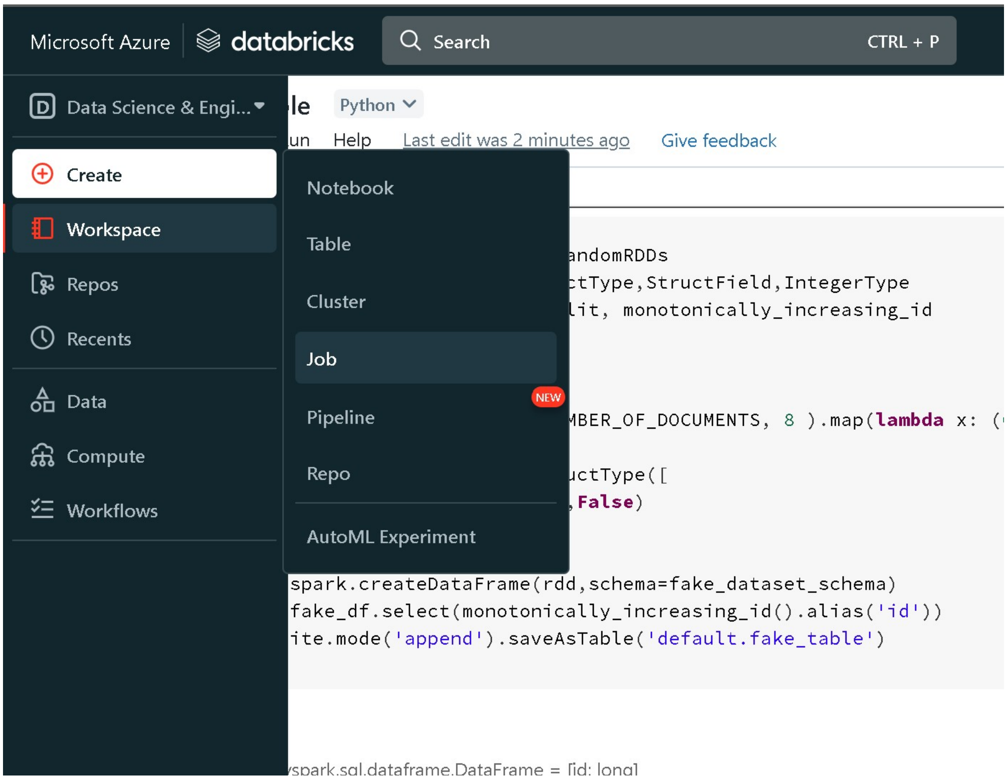
	count(1)
1	1000

Showing 1 row. | 0.48 seconds runtime

SQL cell result stored as PySpark data frame _sqldf. Learn more

Task 2: Create and Schedule a Job

Into the Azure Databrick portal from the lef-sided menu select **Create** + **Job**



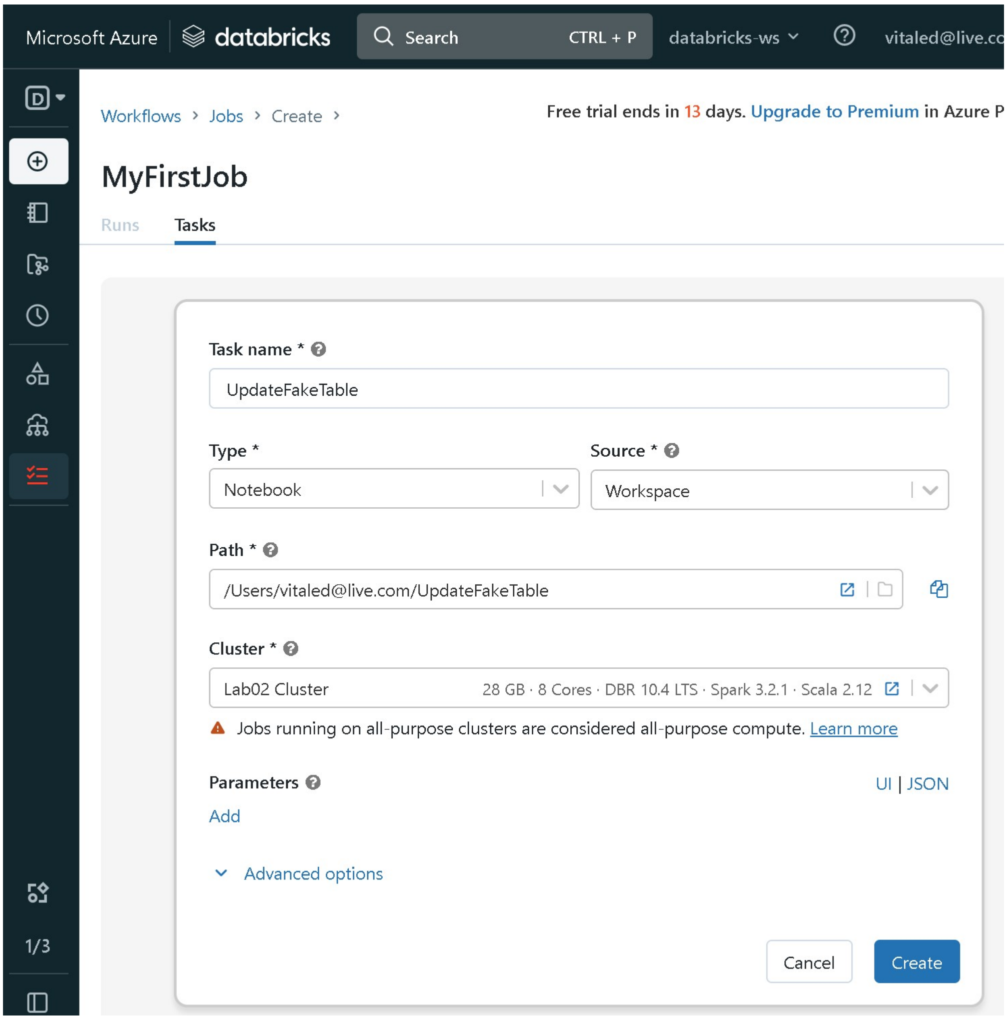
```
andomRDDs
ctType,StructField,IntegerType
lit, monotonically_increasing_id

MBER_OF_DOCUMENTS, 8 ).map(lambda x: (

uctType([
, False)

spark.createDataFrame(rdd,schema=fake_dataset_schema)
fake_df.select(monotonically_increasing_id().alias('id'))
ite.mode('append').saveAsTable('default.fake_table')
```

```
spark.sql.dataframe.DataFrame = [id: long]
```

In the form provide the following details:

property	value
Job Name	*FirstJob*
Task name	*UpdateFakeTable*
Type	*Notebook*
Source	*Workspace*
Path	*path to the UpdateFakeTable notebook*
Cluster	*select cluster created in the Lab 02

then click on **Create**



MyFirstJob

Task name * UpdateFakeTable

Type * Notebook

Source * Workspace

Path * /Users/vitaled@live.com/UpdateFakeTable

Cluster * Lab02 Cluster 28 GB · 8 Cores · DBR 10.4 LTS · Spark 3.2.1 · Scala 2.12

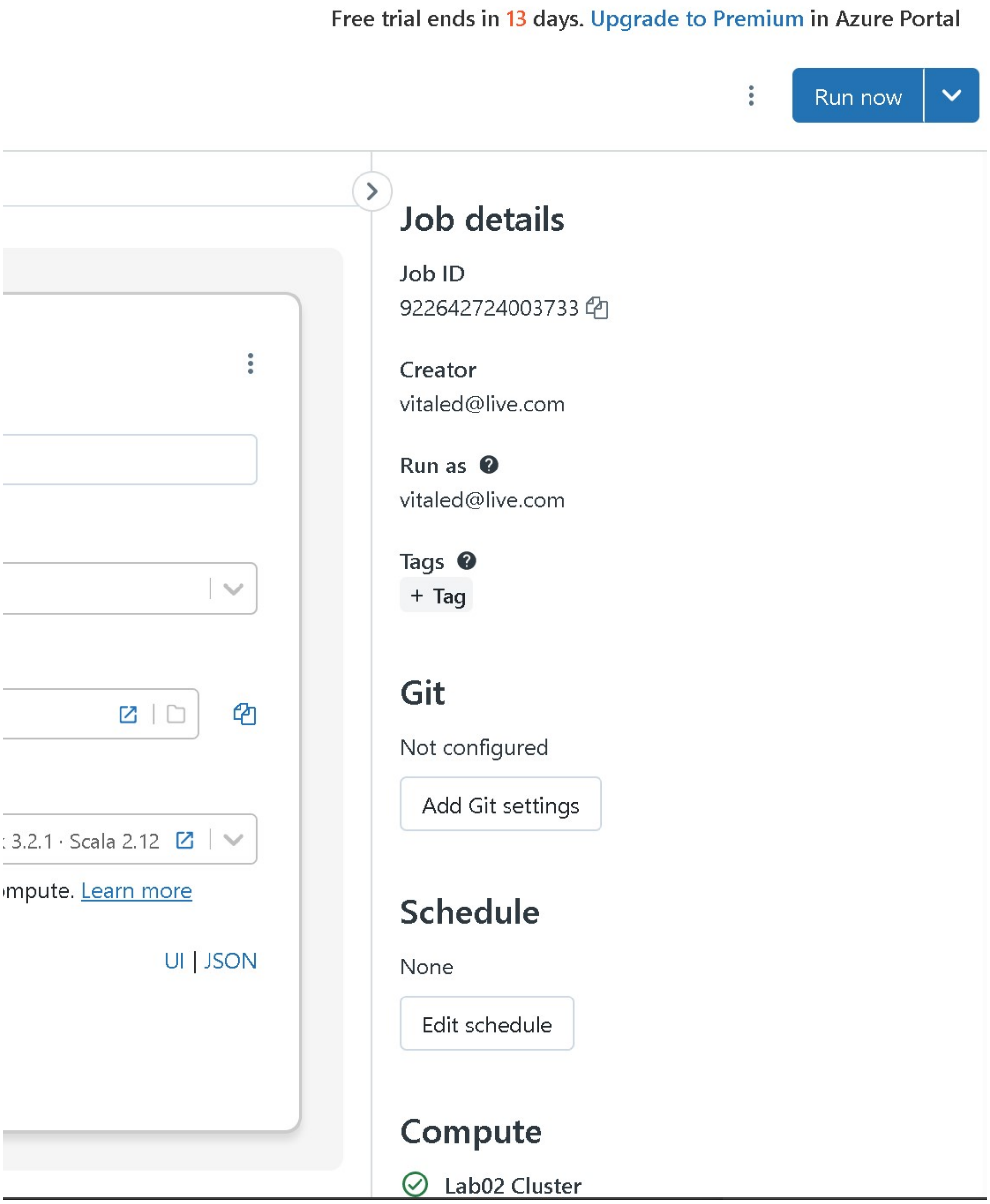
Jobs running on all-purpose clusters are considered all-purpose compute. Learn more

Parameters UI | JSON

Add

Advanced options

In the Job details click on **Edit Schedule**



Free trial ends in 13 days. Upgrade to Premium in Azure Portal

Job details

Job ID
922642724003733

Creator
vitaled@live.com

Run as
vitaled@live.com

Tags

Git

Not configured

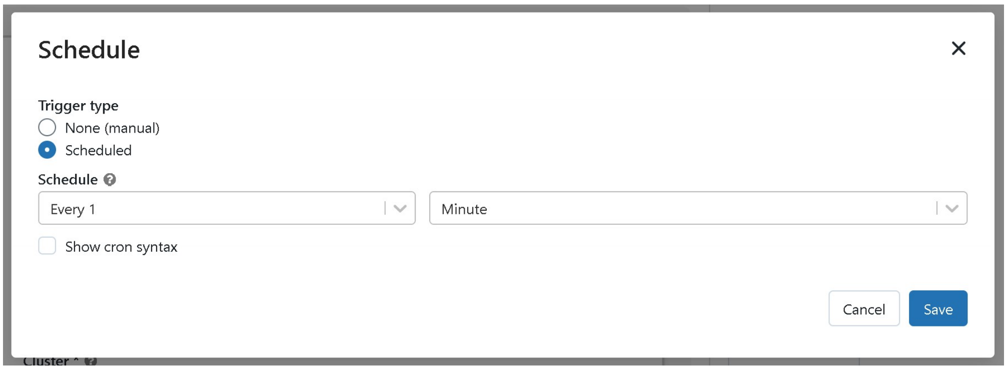
Schedule

None

Compute

Lab02 Cluster

Select **Scheduled** as Trigger type and **Every 1** and **minute** as other settings



Schedule

Trigger type

○ None (manual)

● Scheduled

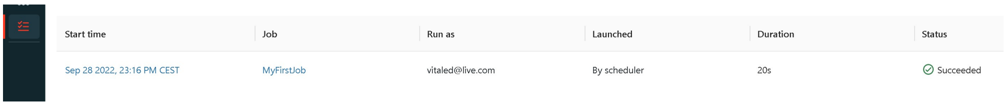
Schedule

Every 1 | Minute

□ Show cron syntax

Task 3: Check job results

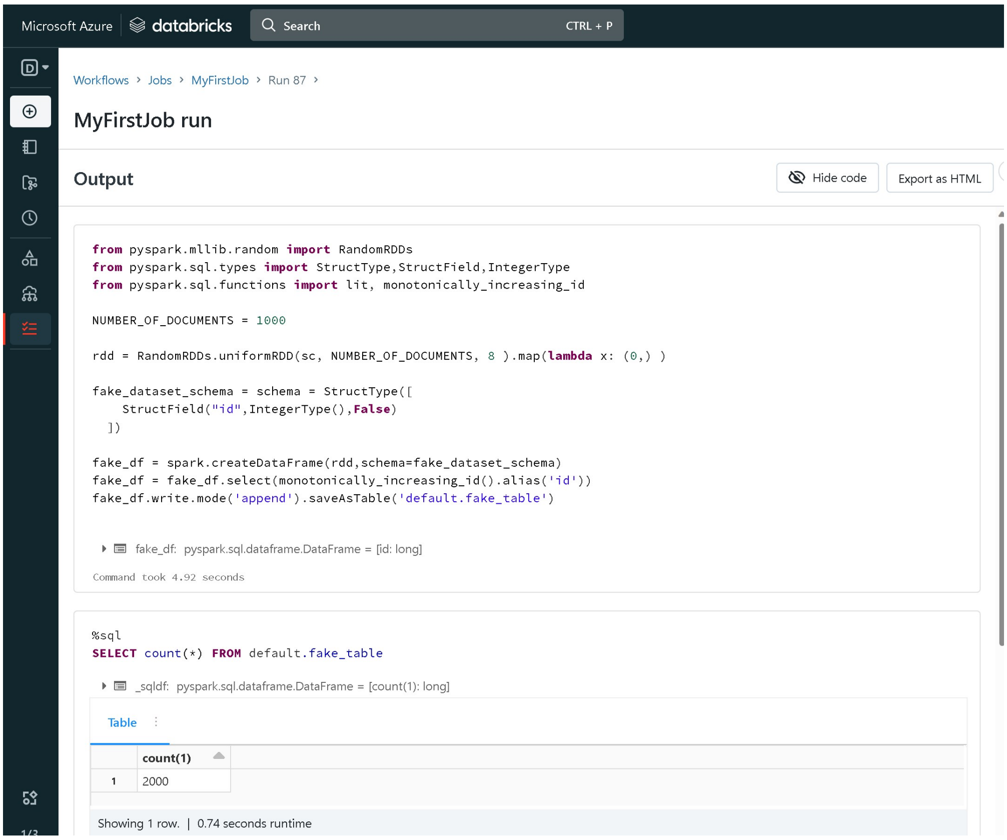
On the left-sided menu click on **Workflows** and then **Job Runs** after one minute you should see the succeded execution



Start time	Job	Run as	Launched	Duration	Status
Sep 28 2022, 23:16 PM CEST	MyFirstJob	vitaled@live.com	By scheduler	20s	✓ Succeeded

Click on the date under the *Start time* column

You will see the result of the notebook execution the count now should be **2000**



MyFirstJob run

Output

```
from pyspark.mllib.random import RandomRDDs
from pyspark.sql.types import StructType,StructField,IntegerType
from pyspark.sql.functions import lit, monotonically_increasing_id

NUMBER_OF_DOCUMENTS = 1000

rdd = RandomRDDs.uniformRDD(sc, NUMBER_OF_DOCUMENTS, 8 ).map(lambda x: (0,) )

fake_dataset_schema = schema = StructType([
    StructField("id",IntegerType(),False)
  ])

fake_df = spark.createDataFrame(rdd,schema=fake_dataset_schema)
fake_df = fake_df.select(monotonically_increasing_id().alias('id'))
fake_df.write.mode('append').saveAsTable('default.fake_table')
```

```
fake_df: pyspark.sql.dataframe.DataFrame = [id: long]
Command took 4.92 seconds
```

```
%sql
SELECT count(*) FROM default.fake_table
```

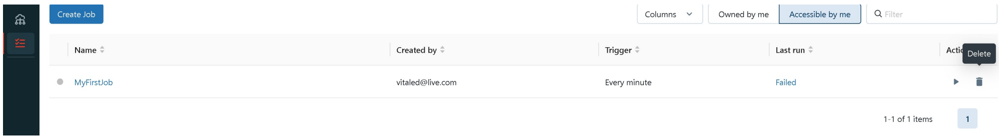
```
_sqldf: pyspark.sql.dataframe.DataFrame = [count(1): long]
```

	count(1)
1	2000

Showing 1 row. | 0.74 seconds runtime

Task 3: Delete the Job

On the left-sided menu click on **Workflows** and then **Jobs** click on the delete icon at then end of the job entry row



Name	Created by	Trigger	Last run	Acti
MyFirstJob	vitaled@live.com	Every minute	Failed	▶ 🗑

1-1 of 1 items 1

This Lab has been completed!