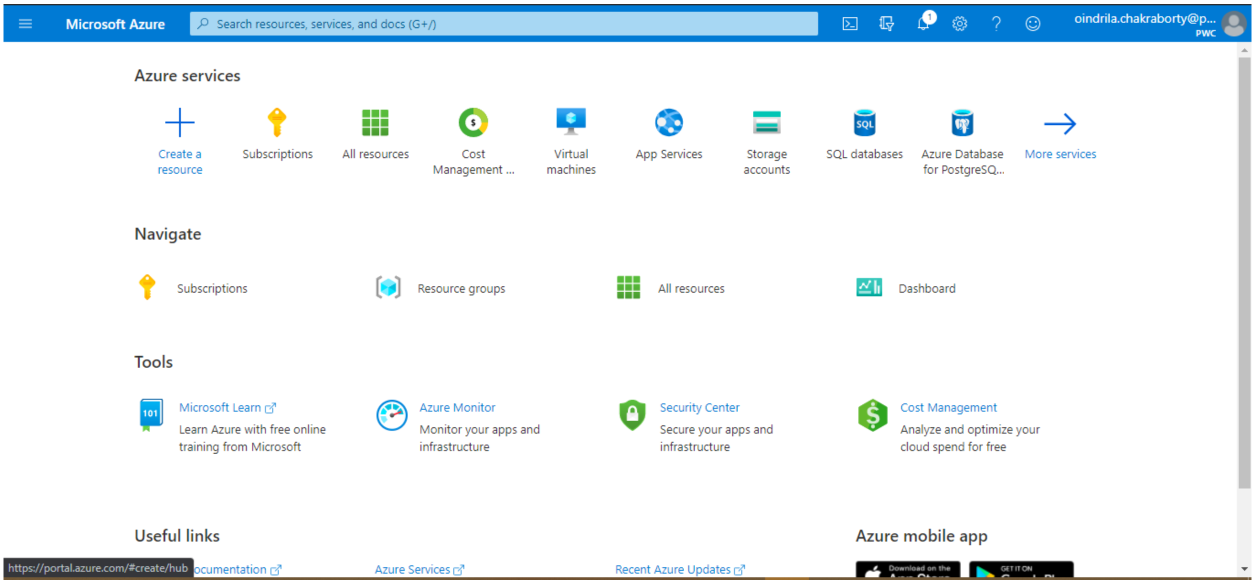
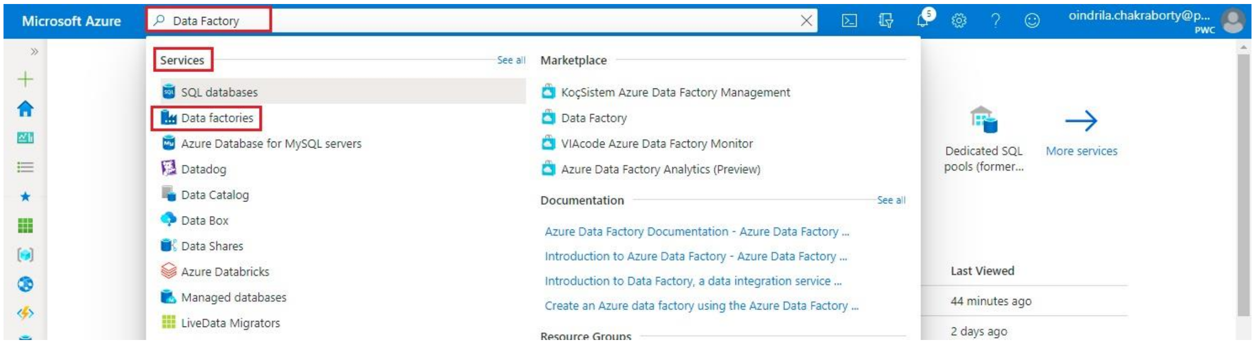
DATABRICKS NOTEBOOK WITH INPUT PARAMETER

Run Databricks Notebook with Parameter from Azure Data Factory Notebook Activity

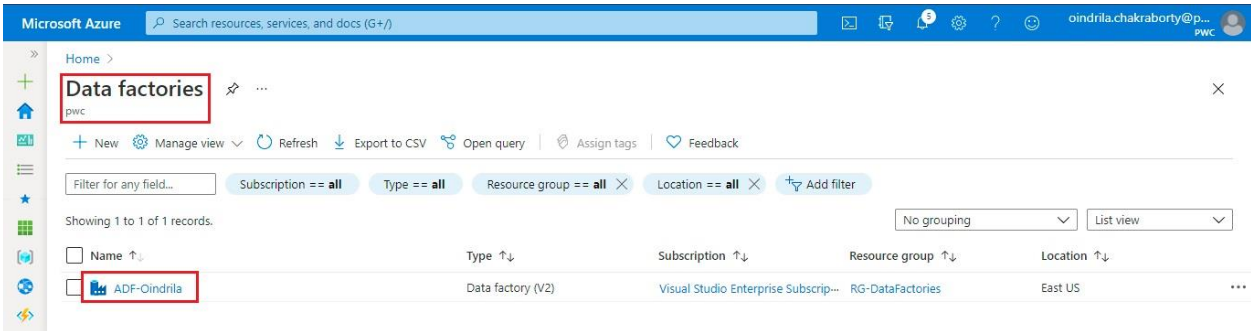
- An *Azure Data Factory Pipeline* can be *created* that *executes* a *Databricks Notebook against* the *Databricks Job Cluster*. The *Pipeline* can *also pass Azure Data Factory Parameters* to the *Databricks Notebook during execution*.
- *Create an Azure Databricks Linked Service* -
  - *Step 1* - *Open* the *Azure portal* (*portal.azure.com*).



  - *Step 2* - *Type* "*Data Factory*" in the *global search bar* in the *home page* of the *Azure portal*. *Click on* the *second search result* "*Data factories*" under "*Services*" in the *left side*.


Microsoft Azure
Data Factory
oindrila.chakraborty@p...
PWC
Services
See all
SQL databases
Data factories
Azure Database for MySQL servers
Datadog
Data Catalog
Data Box
Data Shares
Azure Databricks
Managed databases
LiveData Migrators
Marketplace
KoçSistem Azure Data Factory Management
Data Factory
VIAcode Azure Data Factory Monitor
Azure Data Factory Analytics (Preview)
Documentation
See all
Azure Data Factory Documentation - Azure Data Factory ...
Introduction to Azure Data Factory - Azure Data Factory ...
Introduction to Data Factory, a data integration service ...
Create an Azure data factory using the Azure Data Factory ...
Resource Groups
Dedicated SQL pools (former...
More services
Last Viewed
44 minutes ago
2 days ago

  - *Step 3* - *In* the "*Data factories*" *page*, *click on* the *Azure Data Factory resource* "*ADF-Oindrila*".



Microsoft Azure | Search resources, services, and docs (G+/) | oindrila.chakraborty@p... PWC

Home >

Data factories

pwc

+ New | Manage view | Refresh | Export to CSV | Open query | Assign tags | Feedback

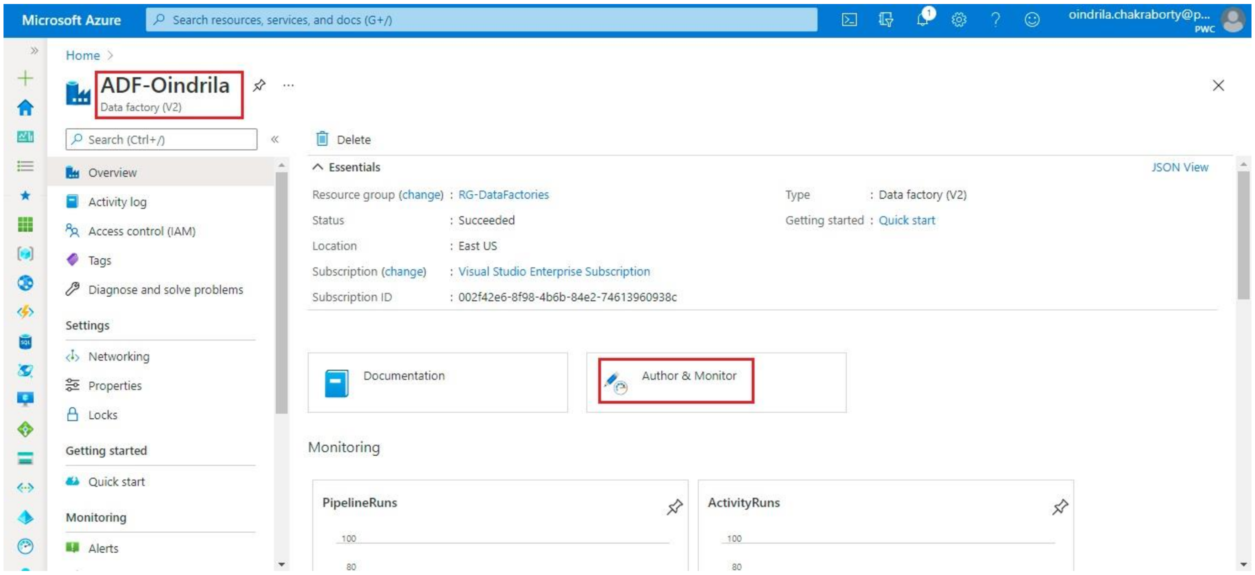
Filter for any field... | Subscription == all | Type == all | Resource group == all | Location == all | Add filter

Showing 1 to 1 of 1 records.

No grouping | List view

Name	Type	Subscription	Resource group	Location
ADF-Oindrila	Data factory (V2)	Visual Studio Enterprise Subscrip...	RG-DataFactories	East US

➢ ***Step 4*** - *Click on* the “*Author & Monitor*” *link*.



Microsoft Azure | Search resources, services, and docs (G+/) | oindrila.chakraborty@p... PWC

Home >

ADF-Oindrila

Data factory (V2)

Search (Ctrl+/)

- Overview
- Activity log
- Access control (IAM)
- Tags
- Diagnose and solve problems

Settings

- Networking
- Properties
- Locks

Getting started

- Quick start

Monitoring

- Alerts

Delete

Essentials | JSON View

Resource group (change) : RG-DataFactories

Status : Succeeded

Location : East US

Subscription (change) : Visual Studio Enterprise Subscription

Subscription ID : 002f42e6-8f98-4b6b-84e2-74613960938c

Type : Data factory (V2)

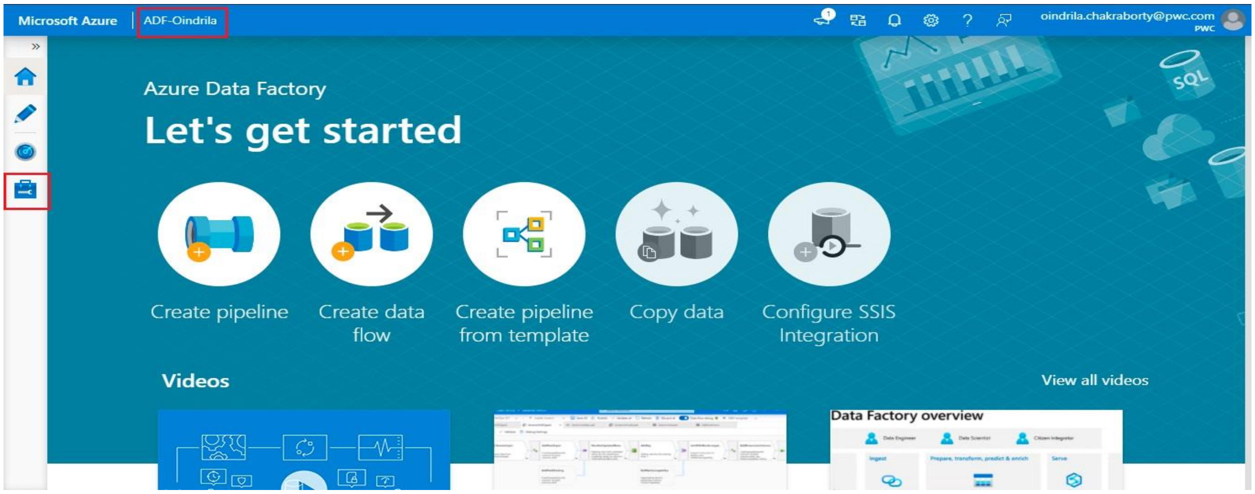
Getting started : Quick start

Documentation | Author & Monitor

Monitoring

PipelineRuns | ActivityRuns

➢ ***Step 5*** - *Click on* the “*Manage*” *button*.

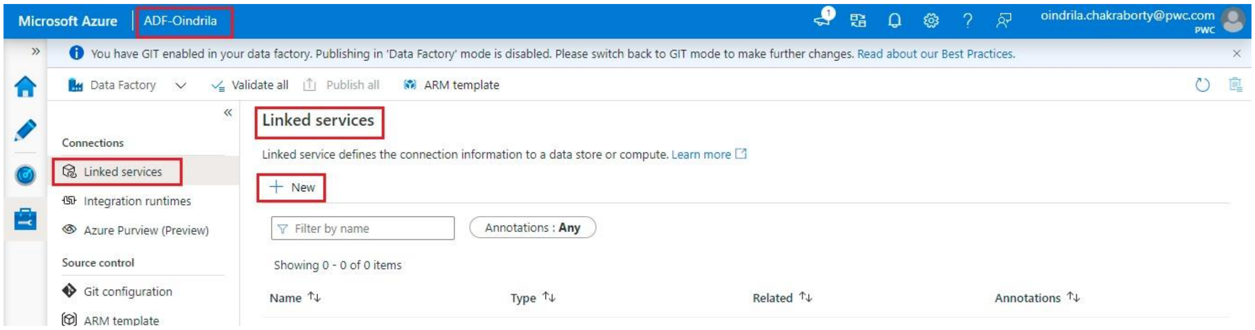


Microsoft Azure | ADF-Oindrila | oindrila.chakraborty@pwc.com PWC

Azure Data Factory

Let's get started

- ***Step 6*** - ***Under*** the "***Connections***" ***left menu***, ***click on*** the ***menu option*** "***Linked services***". Then, ***click on*** the "***+ New***" ***button*** to ***add*** a ***new Linked Service***.



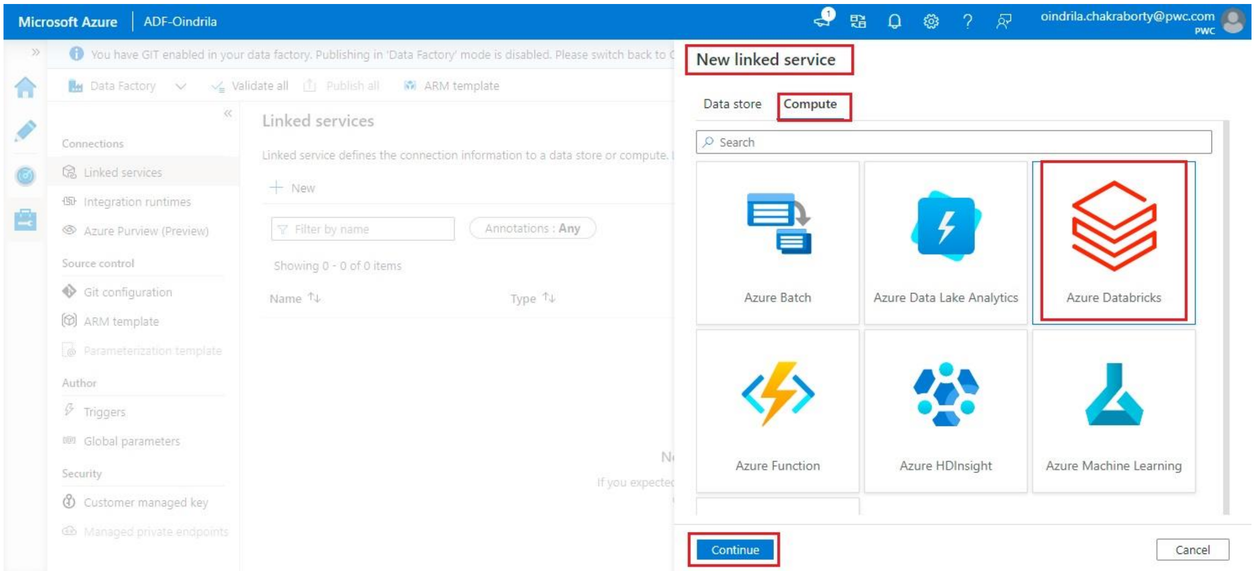
You have GIT enabled in your data factory. Publishing in 'Data Factory' mode is disabled. Please switch back to GIT mode to make further changes. Read about our Best Practices.

Linked services

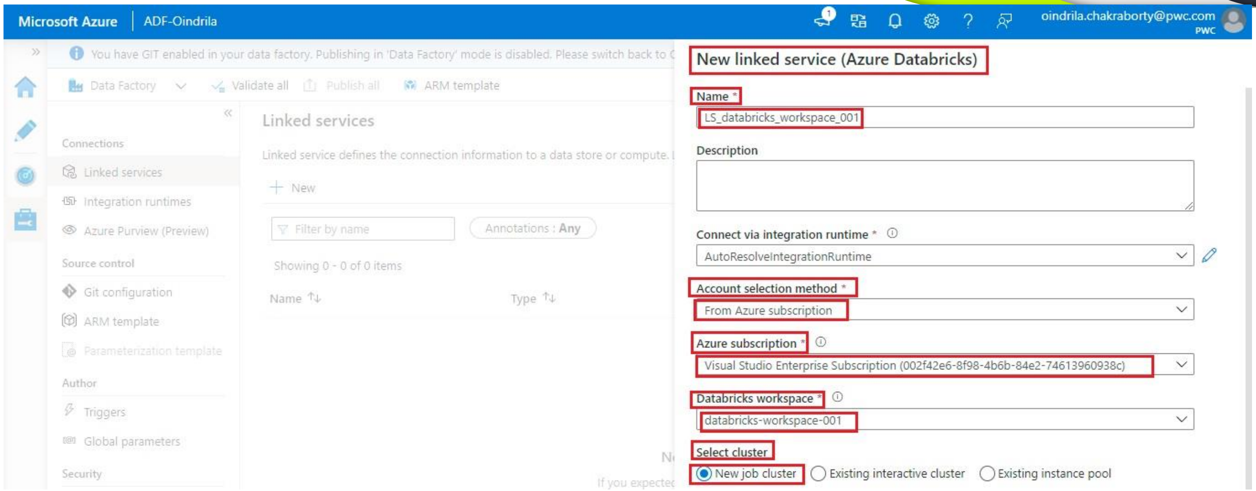
Linked service defines the connection information to a data store or compute. Learn more

Showing 0 - 0 of 0 items

- ***Step 7*** - ***In*** the "***New Linked Service***" ***window***, ***click on*** the "***Compute***" ***tab***. Then, ***select*** the "***Azure Databricks***" and ***click on*** the "***Continue***" ***button***.



- ***Step 8*** - ***In*** the "***New Linked Service (Azure Databricks)***" ***window***, ***provide*** "***LS_databricks_workspace_001***" in the "***Name***" ***textbox***.
  Select the ***option*** "***From Azure Subscription***" ***from*** the ***dropdown*** "***Account selection method***".
  Select the ***proper Azure Subscription from*** the ***dropdown*** "***Azure subscription***".
  Select the ***appropriate Databricks Workspace*** that will ***run*** the ***Notebooks***, ***from*** the ***dropdown*** "***Databricks workspace***", e.g., "***databricks-workspace-001***".
  Select the ***radio button option*** "***New job cluster***" ***for*** the ***property*** "***Select cluster***".


Microsoft Azure | ADF-Oindrila

You have GIT enabled in your data factory. Publishing in 'Data Factory' mode is disabled. Please switch back to G

New linked service (Azure Databricks)

Name *: LS_databricks_workspace_001

Description:

Connect via integration runtime *: AutoResolveIntegrationRuntime

Account selection method *: From Azure subscription

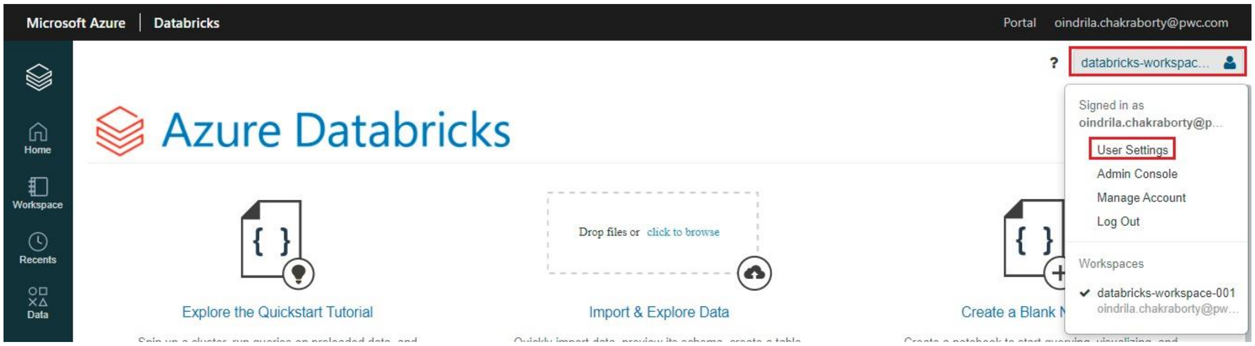
Azure subscription *: Visual Studio Enterprise Subscription (002f42e6-8f98-4b6b-84e2-74613960938c)

Databricks workspace *: databricks-workspace-001

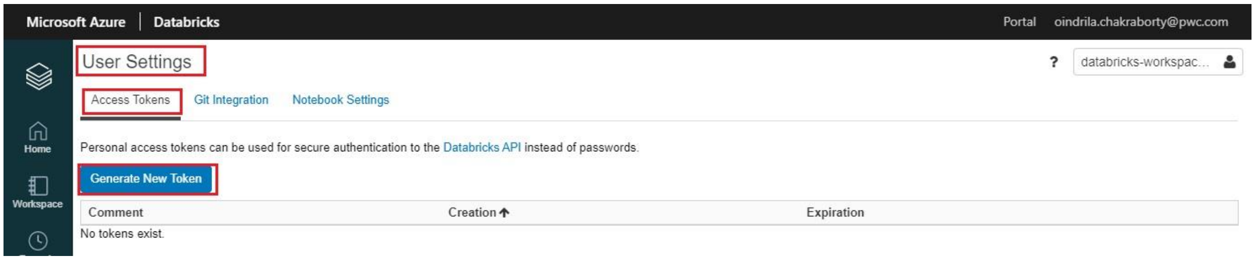
Select cluster: ● New job cluster ○ Existing interactive cluster ○ Existing instance pool

The *Databricks Workspace URL* is *auto populated*.
Select "*Access Token*" *from* the *dropdown* "*Authentication type*".
To *authenticate* to and *access Databricks REST APIs*, the *Databricks Personal Access Token*, or, *Databricks Password* can be *used*. It is *recommended* to *use Databricks Personal Access Token*. *Token-Based Authentication* is *enabled by default for all Databricks Accounts*, *launched after January 2018*. The *number* of *Databricks Personal Access Tokens per User* is *limited* to *600 per Workspace*. *Following* are the *steps* to *generate* the *Databricks Personal Access Token* for the *selected Databricks Workspace* -

- ✓ *__Step 8.1__* - *Click on* the *User Profile Icon* in the *upper right corner* of the *Databricks Workspace*. *Click on* the *menu option* "*User Settings*".



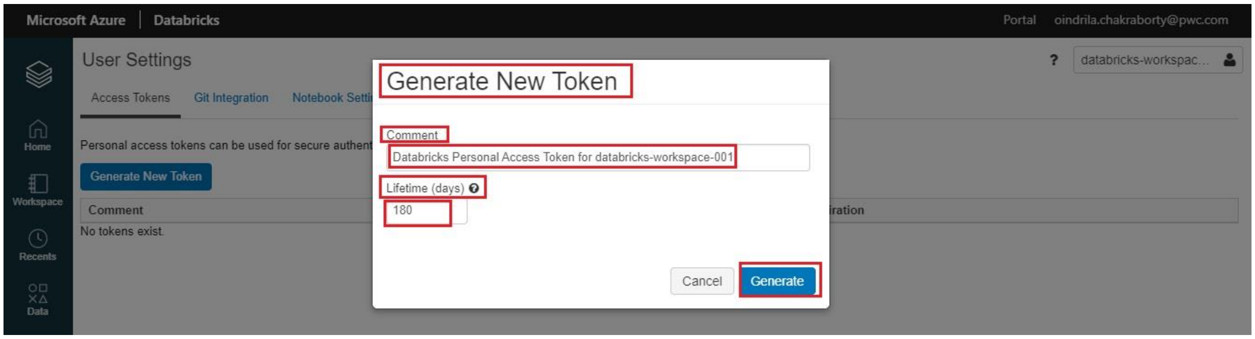
- ✓ *__Step 8.2__* - *Go to* the "*Access Tokens*" *tab*, and, *click on* the "*Generate New Token*" *button*.



Personal access tokens can be used for secure authentication to the Databricks API instead of passwords.

No tokens exist.

✓ *Step 8.3* - *In* the "*Generate New Token*" *pop up*, *provide* a *comment*, e.g., "*Databricks Personal Access Token for databricks-workspace-001*" in the *textbox* "*Comment*", and, *provide* "*180*" in the *textbox* "*Lifetime (days)*". *Finally*, *click on* the "*Generate*" *button*.

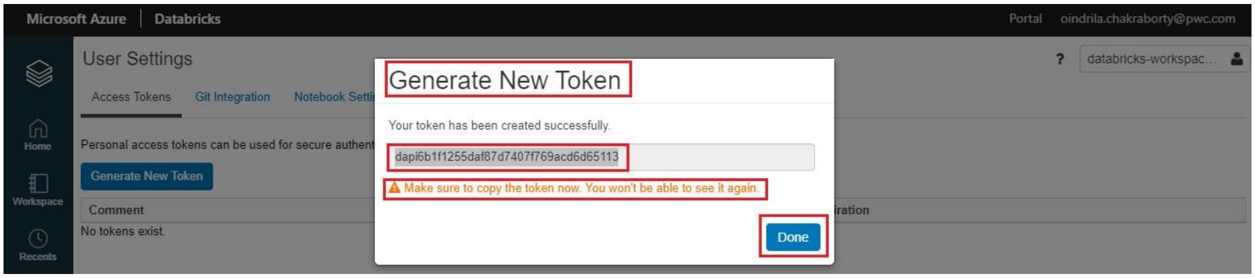


Generate New Token

Comment: Databricks Personal Access Token for databricks-workspace-001

Lifetime (days): 180

✓ *Step 8.4* - *Copy* the *generated Token*, i.e., "*dapi6b1f1255daf87d7407f769acd6d65113*", and *save* it in a *secured location*, *as*, the *Token* will *not* be *displayed again*. *Click on* the "*Done*" *button*.



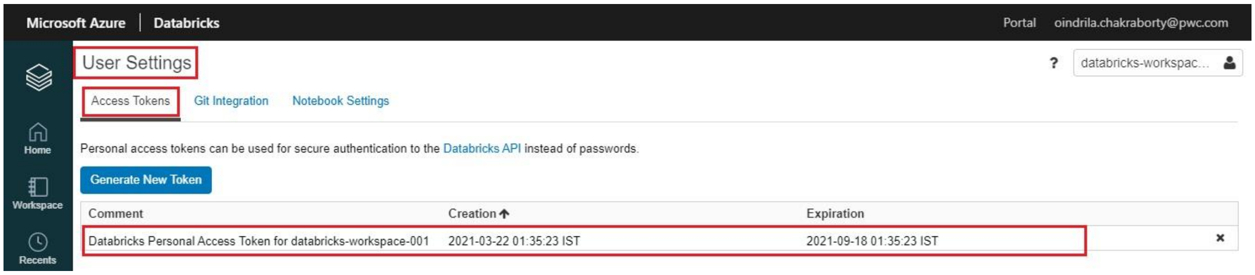
Generate New Token

Your token has been created successfully.

dapi6b1f1255daf87d7407f769acd6d65113

⚠ Make sure to copy the token now. You won't be able to see it again.

✓ *Step 8.5* - The *Databricks Personal Access Token for* the *Databricks Workspace* "*databricks-workspace-001*" is *created*.



User Settings

Personal access tokens can be used for secure authentication to the Databricks API instead of passwords.

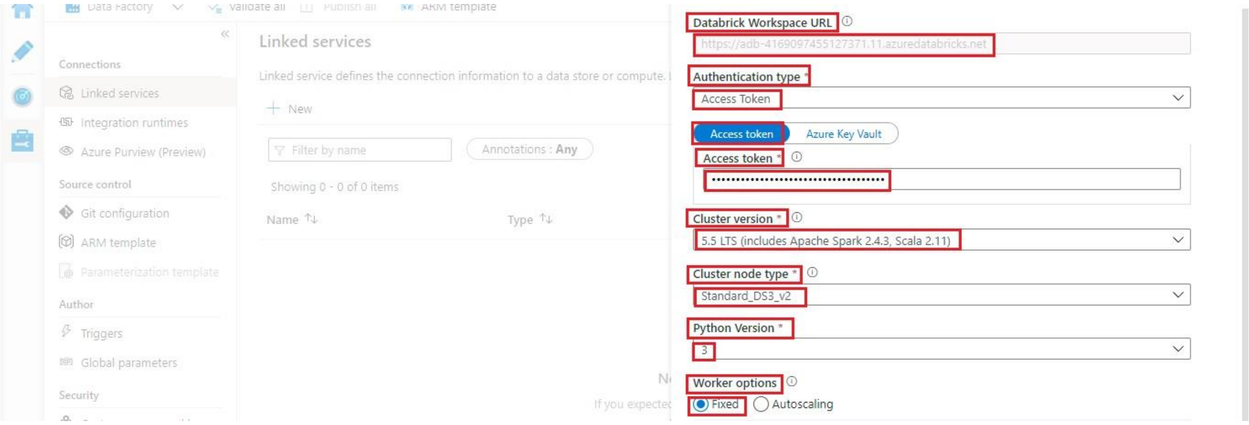
Comment	Creation	Expiration
Databricks Personal Access Token for databricks-workspace-001	2021-03-22 01:35:23 IST	2021-09-18 01:35:23 IST

Provide the *created Databricks Personal Token* “*dapi6b1f1255daf87d7407f769acd6d65113*” in the *textbox* “*Access token*”. *Select* the *option* “*5.5 LTS (includes Apache Spark 2.4.3, Scala 2.11)*” *from* the *dropdown* “*Cluster version*”.

Select the *option* “*Standard_DS3_V2*” *under* the “*General purpose*” *category from* the *dropdown* “*Cluster node type*”.

Select the *option* “*3*” *from* the *dropdown* “*Python Version*”.

Select the *radio button option* “*Fixed*” *for* the *property* “*Worker options*”.



Linked service defines the connection information to a data store or compute.

Showing 0 - 0 of 0 items

Databrick Workspace URL: https://adb-4169097455127371.11.azuredatabricks.net

Authentication type *: Access Token

Access token *: ••••••••••••••••••••••••••••••••••••

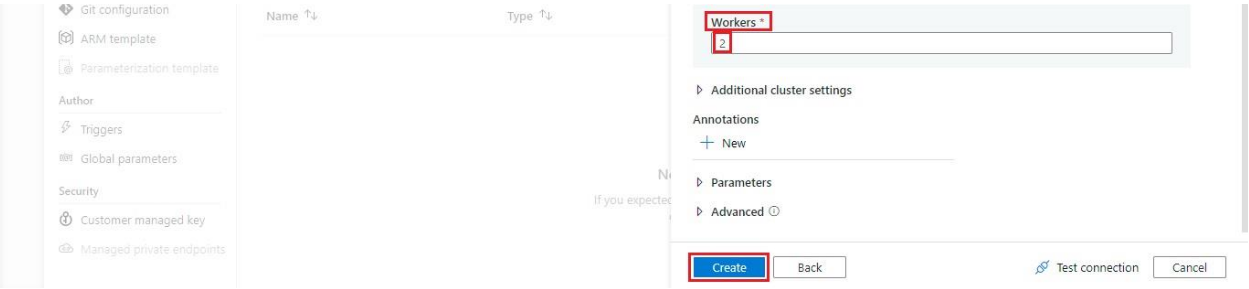
Cluster version *: 5.5 LTS (includes Apache Spark 2.4.3, Scala 2.11)

Cluster node type *: Standard_DS3_v2

Python Version *: 3

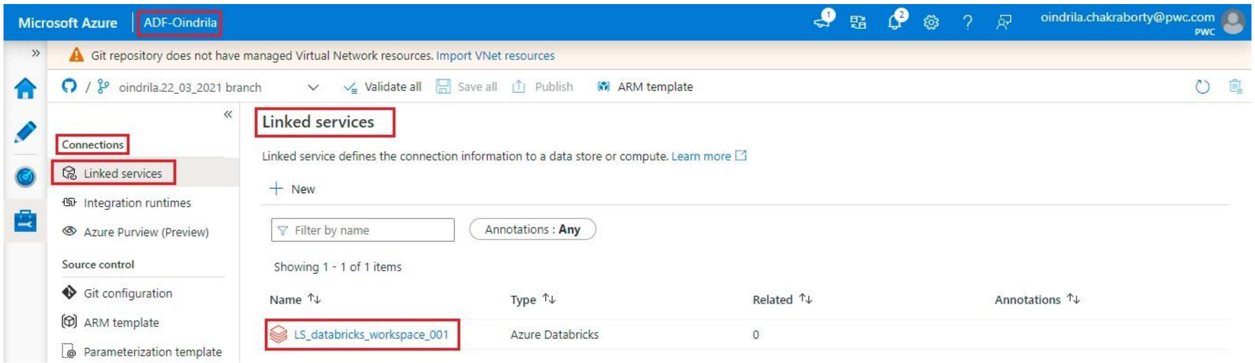
Worker options: ● Fixed ○ Autoscaling

Provide “*2*” in the *textbox* “*Workers*”, and, *finally*, *click on* the “*Create*” *button*.



Workers *: 2

- ➢ *Step 9* - The *Linked Service* “*LS_databricks_workspace_001*” is *created successfully*.



Git repository does not have managed Virtual Network resources. Import VNet resources

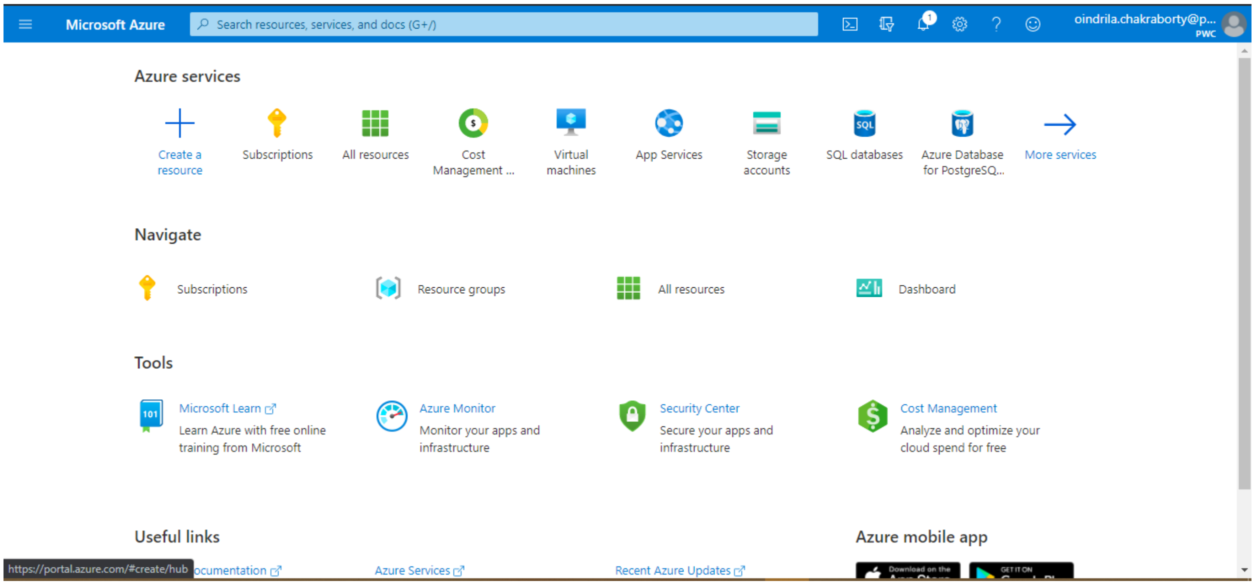
Linked services

Linked service defines the connection information to a data store or compute. Learn more

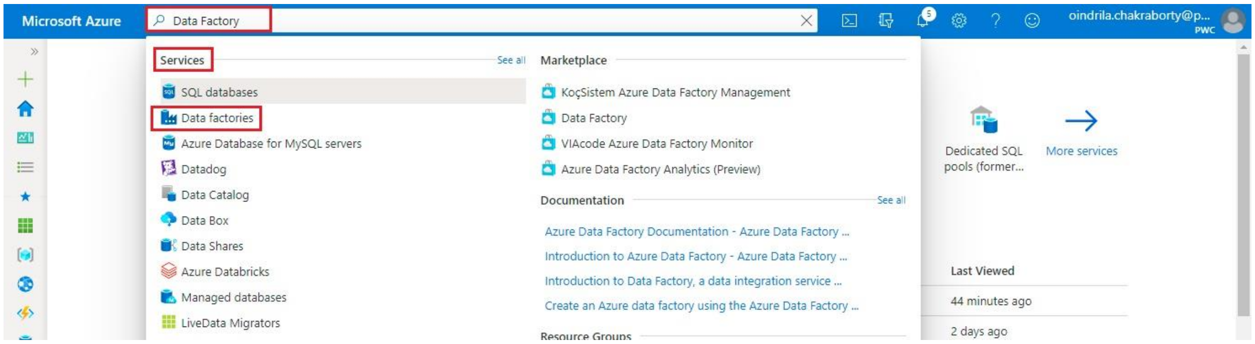
Showing 1 - 1 of 1 items

Name	Type	Related	Annotations
LS_databricks_workspace_001	Azure Databricks	0	

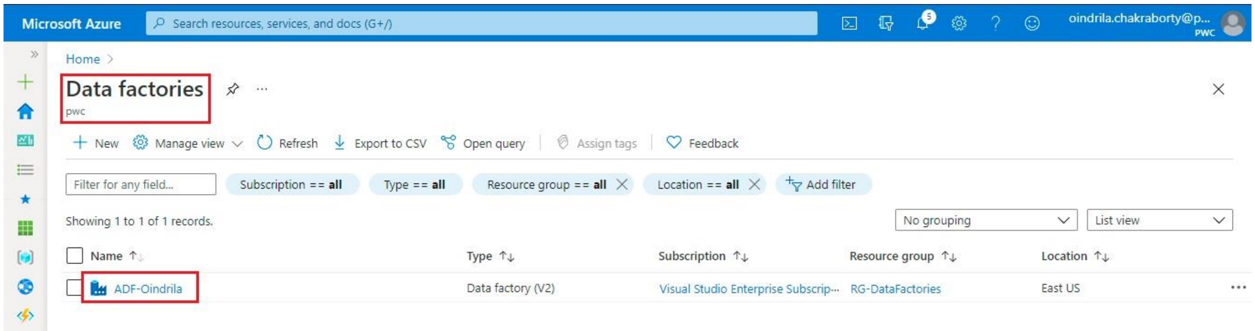
- ***Create a Pipeline*** -
  - ***Step 1*** - ***Open*** the ***Azure portal*** (***portal.azure.com***).



  - ***Step 2*** - ***Type*** "***Data Factory***" in the ***global search bar*** in the ***home page*** of the ***Azure portal***. ***Click on*** the ***second search result*** "***Data factories***" under "***Services***" in the ***left side***.


Microsoft Azure
Data Factory
Services
See all
SQL databases
Data factories
Azure Database for MySQL servers
Datadog
Data Catalog
Data Box
Data Shares
Azure Databricks
Managed databases
LiveData Migrators
Marketplace
KoçSistem Azure Data Factory Management
Data Factory
VIAcode Azure Data Factory Monitor
Azure Data Factory Analytics (Preview)
Documentation
See all
Azure Data Factory Documentation - Azure Data Factory ...
Introduction to Azure Data Factory - Azure Data Factory ...
Introduction to Data Factory, a data integration service ...
Create an Azure data factory using the Azure Data Factory ...
Resource Groups
Dedicated SQL pools (former...
More services
Last Viewed
44 minutes ago
2 days ago
oindrila.chakraborty@p...
PWC

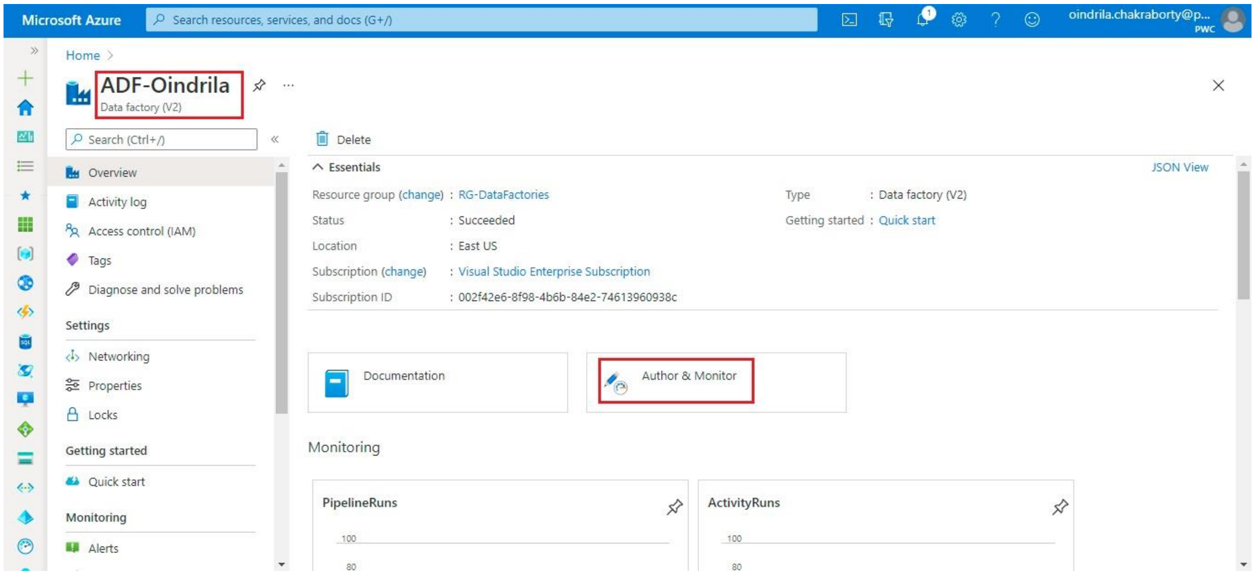
➢ ***Step 3*** - *In* the "*Data factories*" *page*, *click on* the *Azure Data Factory resource* "*ADF-Oindrila*".



Showing 1 to 1 of 1 records.

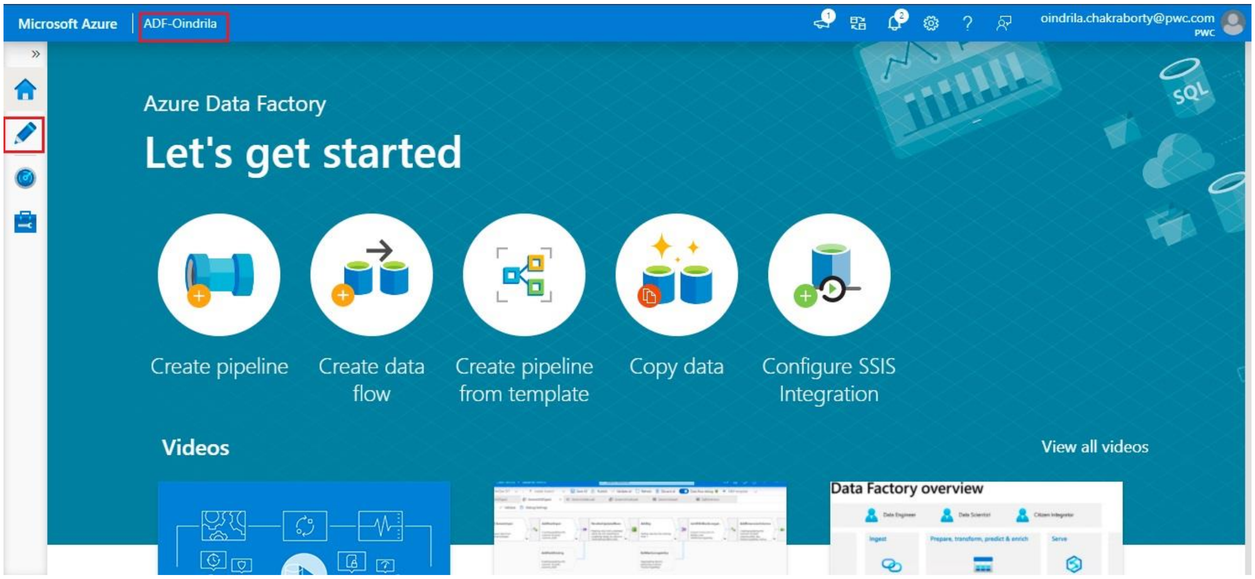
Name	Type	Subscription	Resource group	Location
ADF-Oindrila	Data factory (V2)	Visual Studio Enterprise Subscrip...	RG-DataFactories	East US

➢ ***Step 4*** - *Click on* the "*Author & Monitor*" *link*.

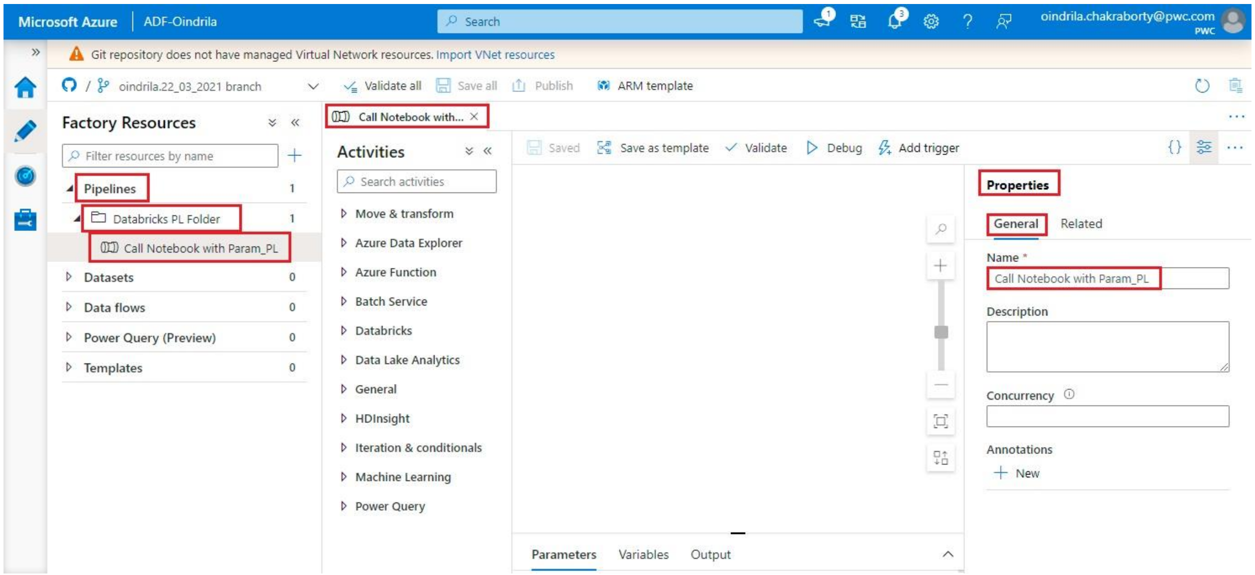


Resource group (change) : RG-DataFactories
Status : Succeeded
Location : East US
Subscription (change) : Visual Studio Enterprise Subscription
Subscription ID : 002f42e6-8f98-4b6b-84e2-74613960938c
Type : Data factory (V2)
Getting started : Quick start

➢ ***Step 5*** - *Click on* the "*Author*" *button*.

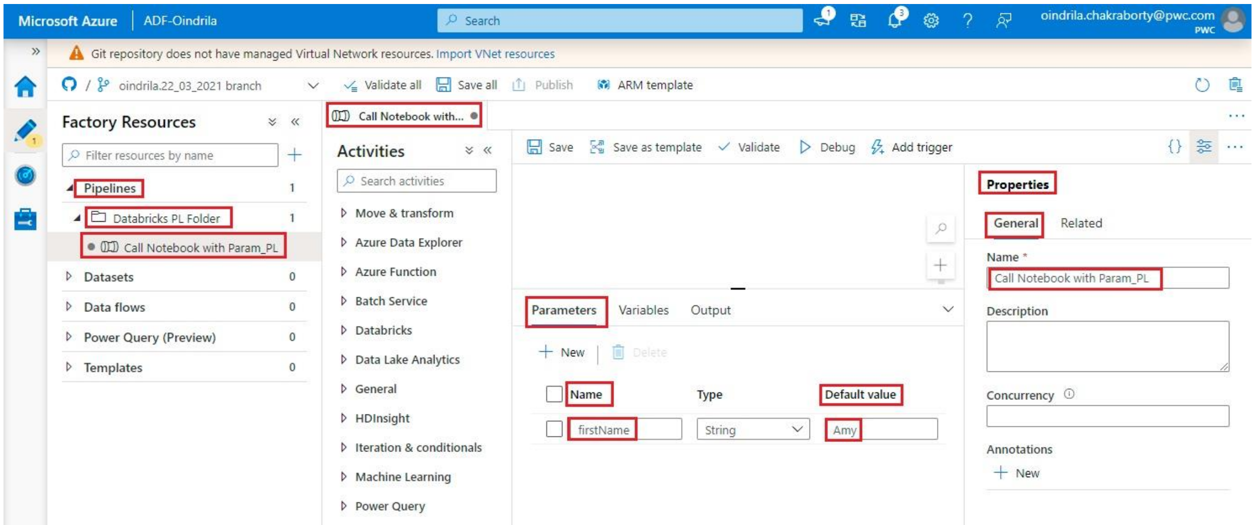


➢ *__Step 6__* - ***Create*** a ***Pipeline*** *"Call Notebook with Param_PL"*.



Name * Call Notebook with Param_PL

➢ *__Step 7__* - ***Create*** a ***Pipeline Parameter*** by the ***name*** *"firstName"*, ***with*** the ***Default Value*** *"Amy"*. This ***Pipeline Parameter*** will be ***passed*** to the ***Databricks Notebook***.

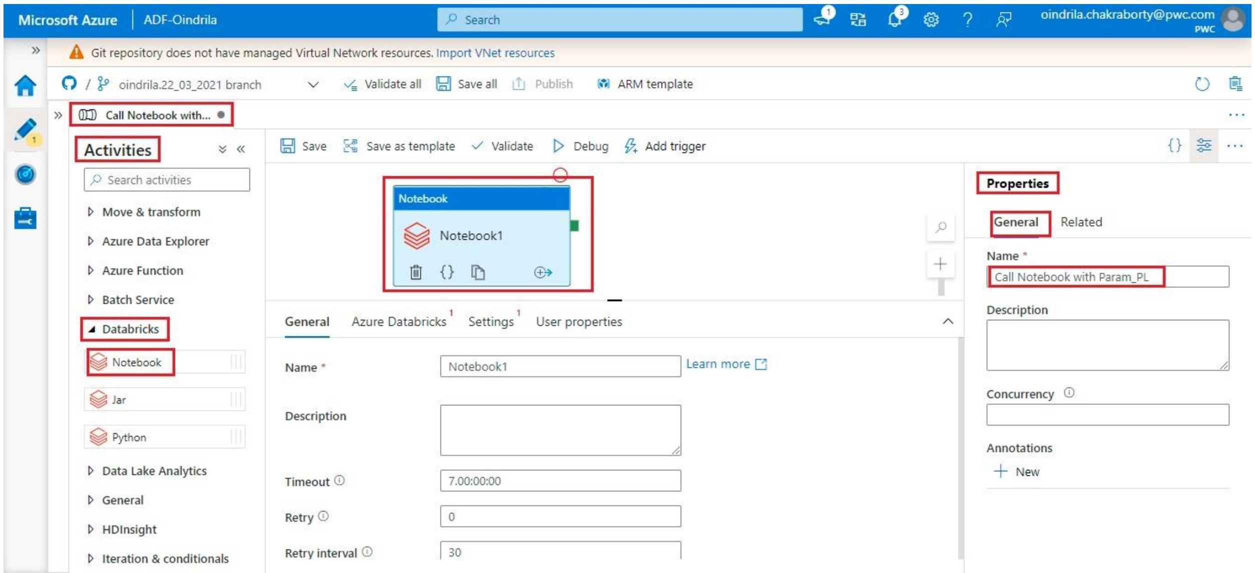


Name	Type	Default value
firstName	String	Amy

Name * Call Notebook with Param_PL

Description

- ➢ ***Step 8*** - *In* the "*Activities*" *toolbox*, *expand* "*Databricks*". *Drag* the "*Notebook*" *Activity from* the "*Activities*" *toolbox* to the *Pipeline Designer surface*.



Name * Notebook1

Description

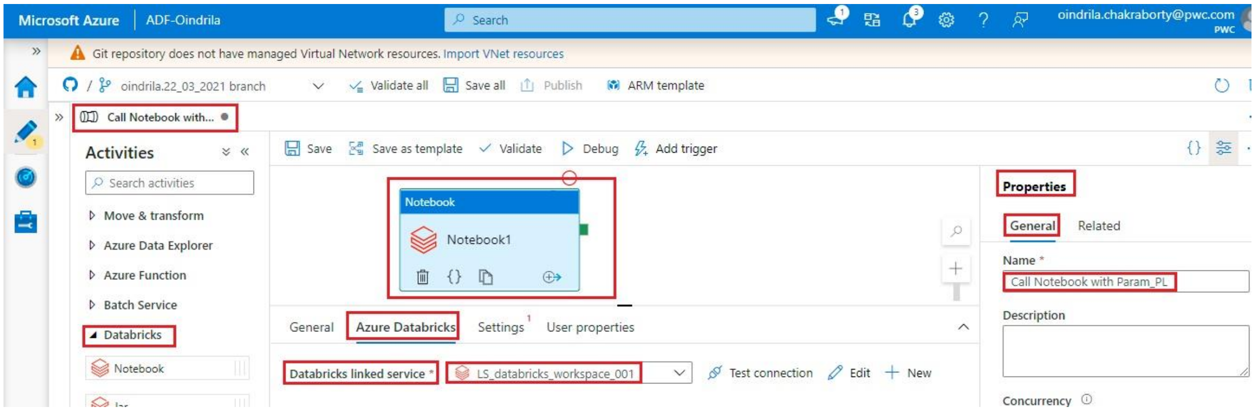
Timeout 7.00:00:00

Retry 0

Retry interval 30

Name * Call Notebook with Param_PL

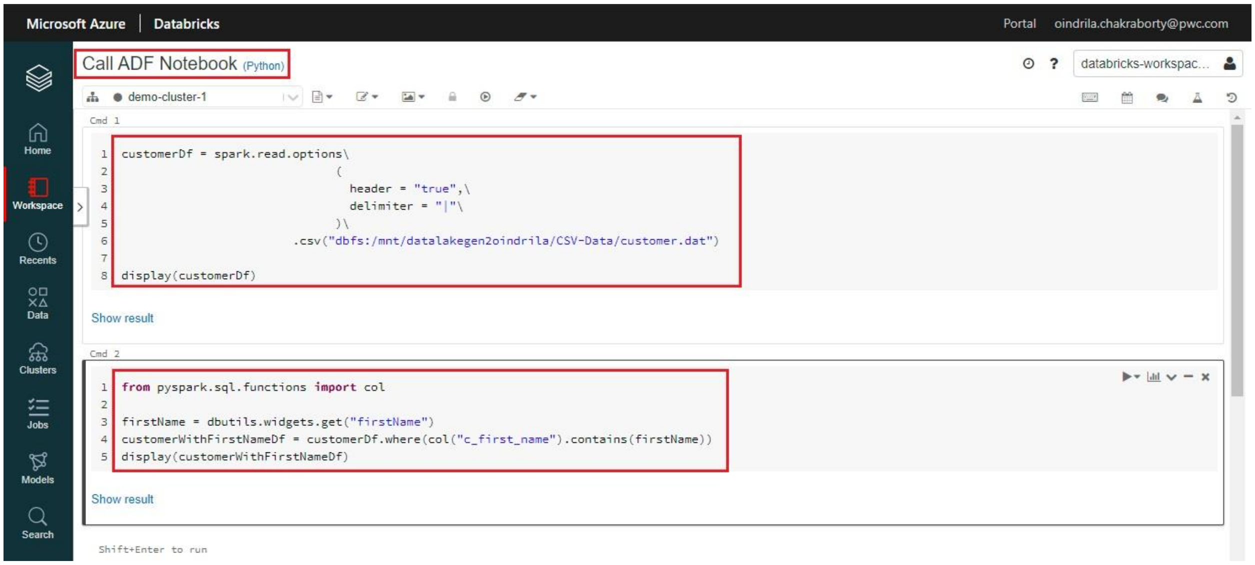
- ➢ ***Step 9*** - *In* the "*Notebook*" *Activity window*, *switch to* the "*Azure Databricks*" *tab*. *Select* the *created Linked Service* "*LS_databricks_workspace_001*" in the *dropdown* "*Databricks linked service*".



Name *: Call Notebook with Param_PL

Databricks linked service *: LS_databricks_workspace_001

- ➢ ***Step 10*** - ***Create*** a ***new Databricks Notebook*** by the ***name*** "***Call ADF Notebook***", ***inside*** the ***folder*** "***ADF_Notebook_Folder***", in the ***Databricks Workspace*** "***databricks-workspace-001***".



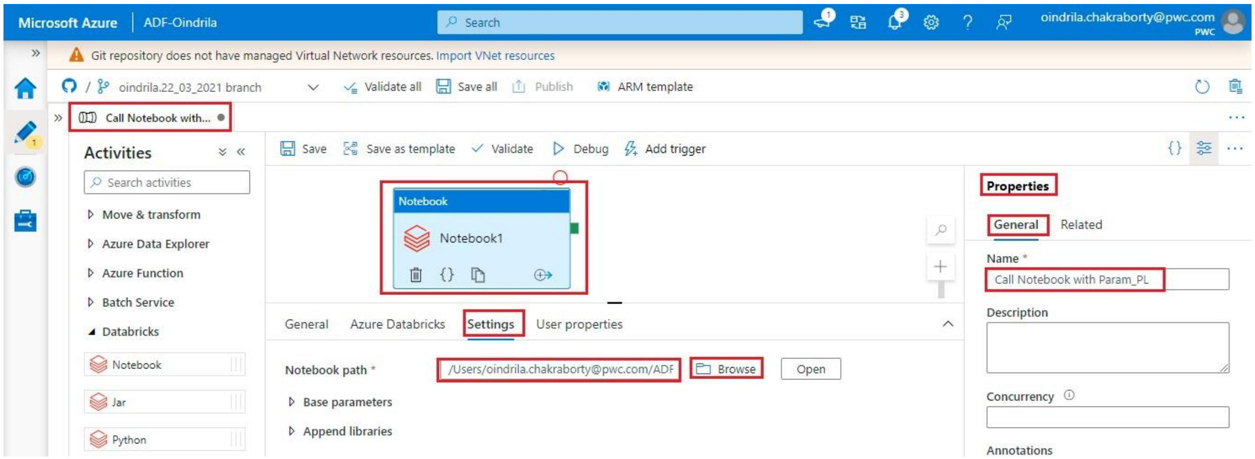
```
customerDf = spark.read.options\
                              (
                                header = "true",\
                                delimiter = "|"\
                              )\
                        .csv("dbfs:/mnt/datalakegen2oindrila/CSV-Data/customer.dat")

display(customerDf)
```

```
from pyspark.sql.functions import col

firstName = dbutils.widgets.get("firstName")
customerWithFirstNameDf = customerDf.where(col("c_first_name").contains(firstName))
display(customerWithFirstNameDf)
```

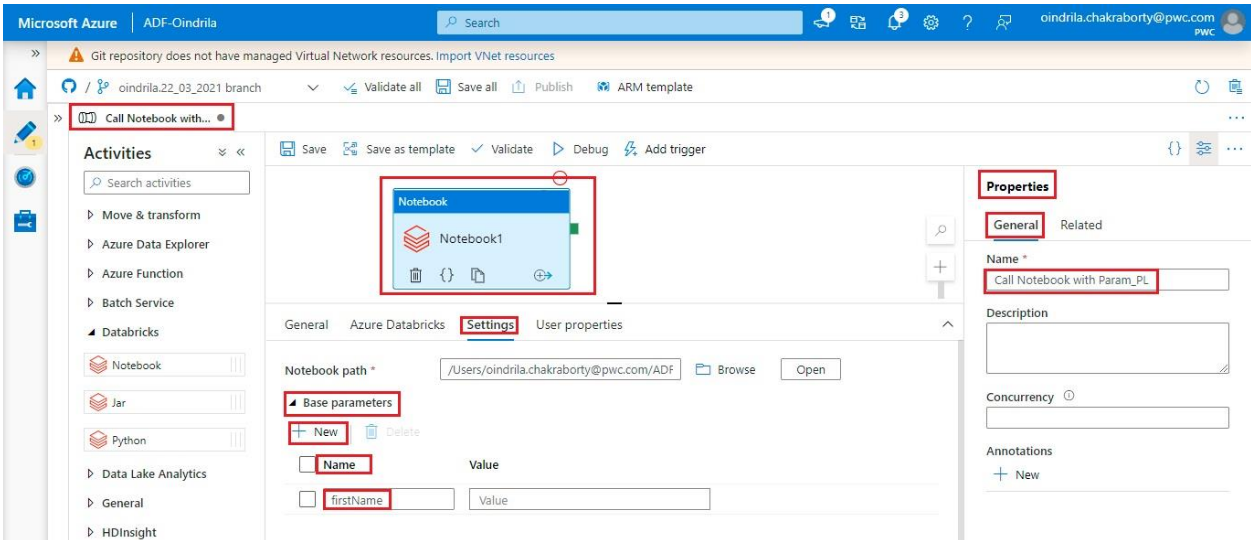
- ➢ ***Step 11*** - ***In*** the "***Settings***" ***tab***, ***browse*** to ***select*** the ***path*** of the ***Databricks Notebook***. The ***path*** of the ***Databricks Notebook*** is "***/Users/oindrila.chakraborty@pwc.com/ADF_Notebook_Folder/Call ADF Notebook***".



Notebook path * /Users/oindrila.chakraborty@pwc.com/ADF

Name * Call Notebook with Param_PL

- ➢ ***Step 12*** - *Add* the *Pipeline Parameter* "*firstName*" to the "*Notebook*" *Activity*. *Expand* the *property* "*Base Parameters*". *Click on* the "*+ New*" *link*. *Provide* "*firstName*" in the *textbox* "*Name*".

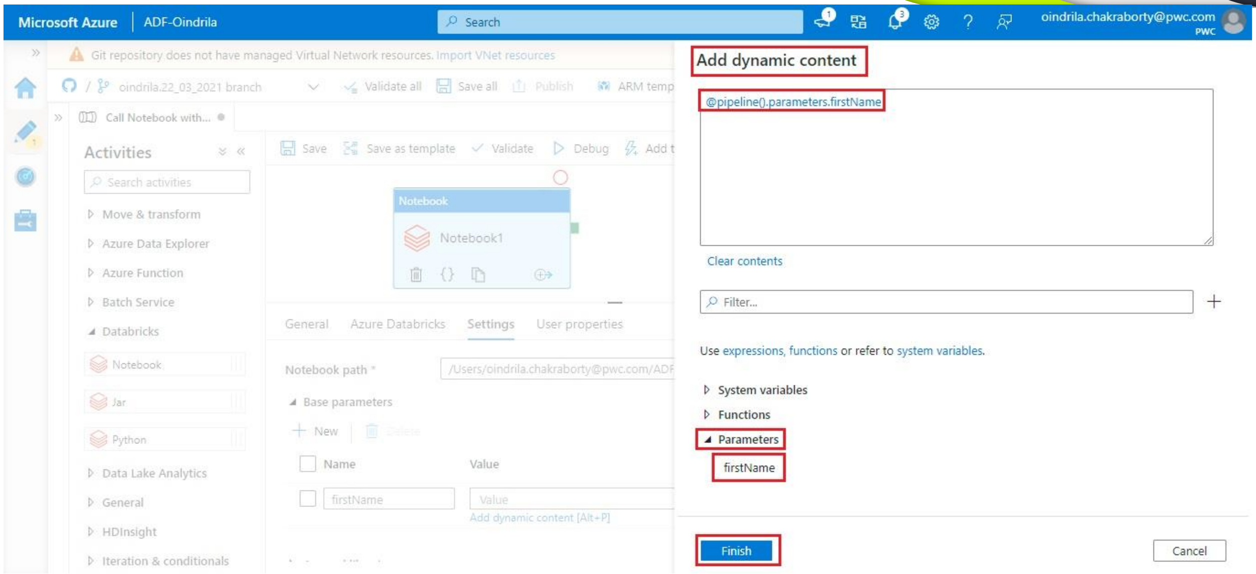


Notebook path * /Users/oindrila.chakraborty@pwc.com/ADF

Name	Value
firstName	Value

Name * Call Notebook with Param_PL

Click on the *textbox* "*Value*". This *enables* the *link* "*Add dynamic content [Alt+P]*" *to be displayed just below* the *textbox*. *Click* on that *link*. *Select* the *Pipeline Parameter* "*firstName*", and, *click on* the "*Finish*" *button*.



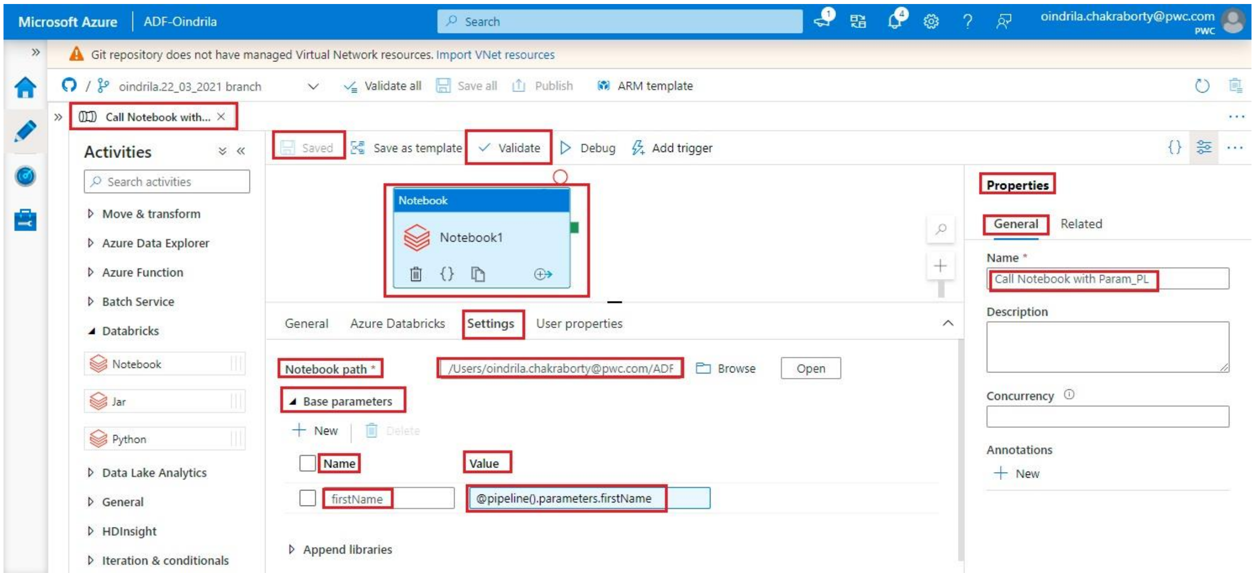
Add dynamic content

@pipeline().parameters.firstName

Use expressions, functions or refer to system variables.

- System variables
- Functions
- Parameters
  - firstName

The *Value* of the *Parameter* of the "*Notebook*" *Activity* "*firstName*" is the *expression* "*@pipeline().parameters.firstName*".



Settings

Notebook path *	/Users/oindrila.chakraborty@pwc.com/ADF

Base parameters

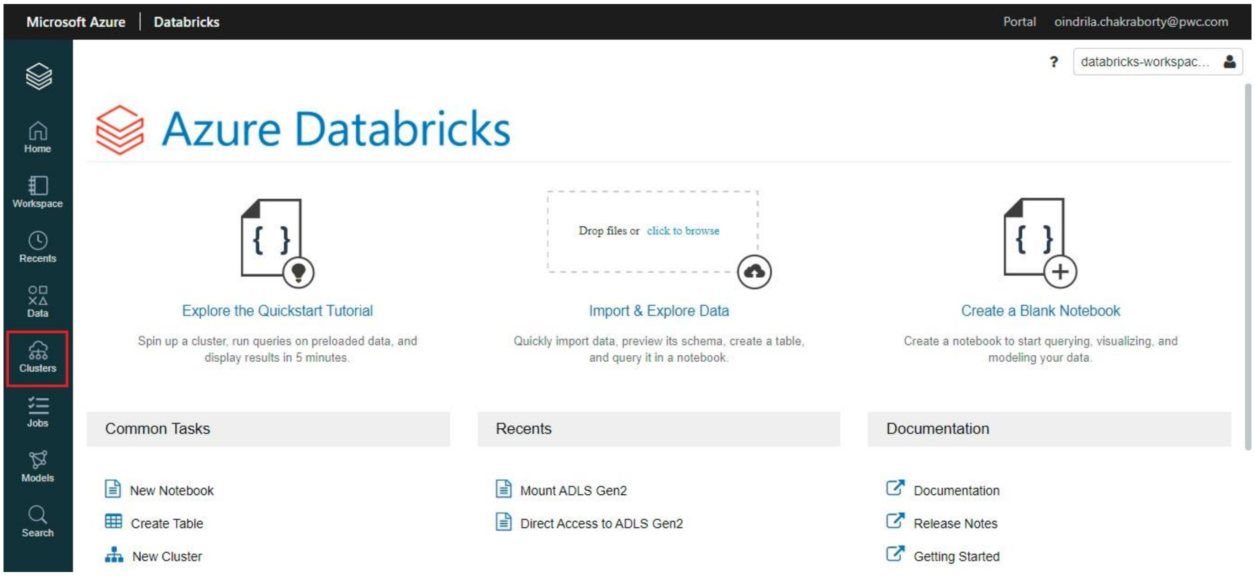
Name	Value
firstName	@pipeline().parameters.firstName

Properties — General

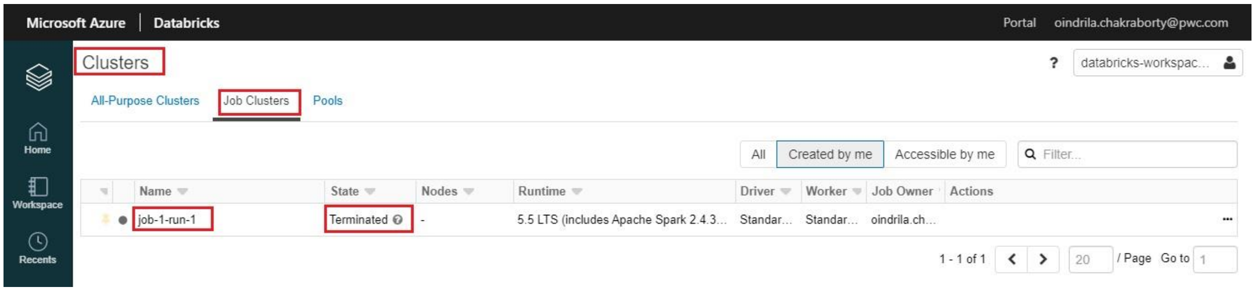
Name *: Call Notebook with Param_PL

- ➢ <u>*Step 13*</u> - *Save* the *Pipeline* and *Validate*. Then *run* the *Pipeline*, *by clicking on* the "*Debug*" *link*.

- <u>*Verify the Output*</u> - *Open* the *Azure Databricks Workspace* "*databricks-workspace-001*". *Click on* the "*Clusters*" *left menu link*.

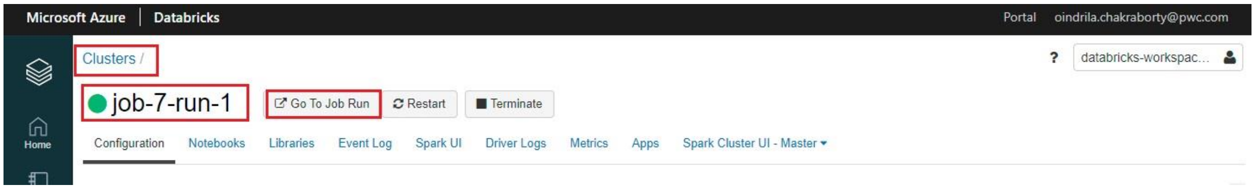

Microsoft Azure | Databricks
Portal oindrila.chakraborty@pwc.com
?
databricks-workspac...
Home
Workspace
Recents
Data
Clusters
Jobs
Models
Search
Azure Databricks
Explore the Quickstart Tutorial
Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.
Drop files or click to browse
Import & Explore Data
Quickly import data, preview its schema, create a table, and query it in a notebook.
Create a Blank Notebook
Create a notebook to start querying, visualizing, and modeling your data.
Common Tasks
New Notebook
Create Table
New Cluster
Recents
Mount ADLS Gen2
Direct Access to ADLS Gen2
Documentation
Documentation
Release Notes
Getting Started

Switch to the "***Job Clusters***" ***tab***. The ***Status*** of the ***Job*** can be "***Pending***", "***Execution***", "***Running***", or, "***Terminated***". ***Since***, the ***Job*** is ***finished executing***, its ***Status*** is ***displayed*** as "***Terminated***".

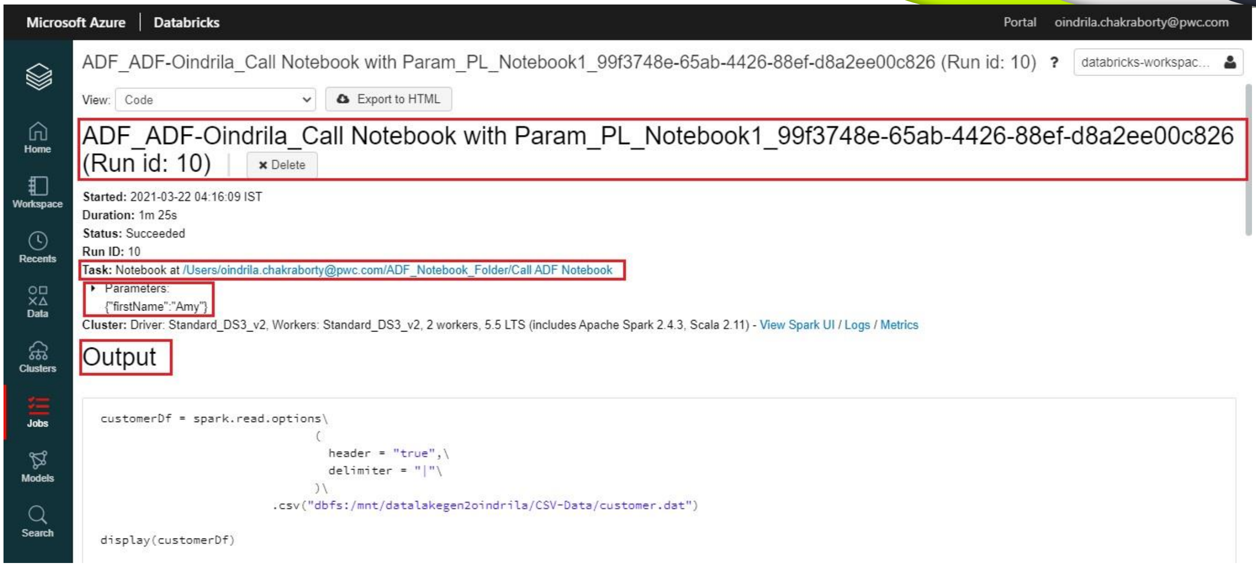


	Name	State	Nodes	Runtime	Driver	Worker	Job Owner	Actions
●	job-1-run-1	Terminated	-	5.5 LTS (includes Apache Spark 2.4.3...	Standar...	Standar...	oindrila.ch...	···

Click on the ***Job Name***, then, ***click on*** the ***button*** "***Go To Job Run***".



On successful run, the ***parameter passed***, and, the ***output*** of the ***Notebook*** can be ***validated***.



ADF_ADF-Oindrila_Call Notebook with Param_PL_Notebook1_99f3748e-65ab-4426-88ef-d8a2ee00c826 (Run id: 10)

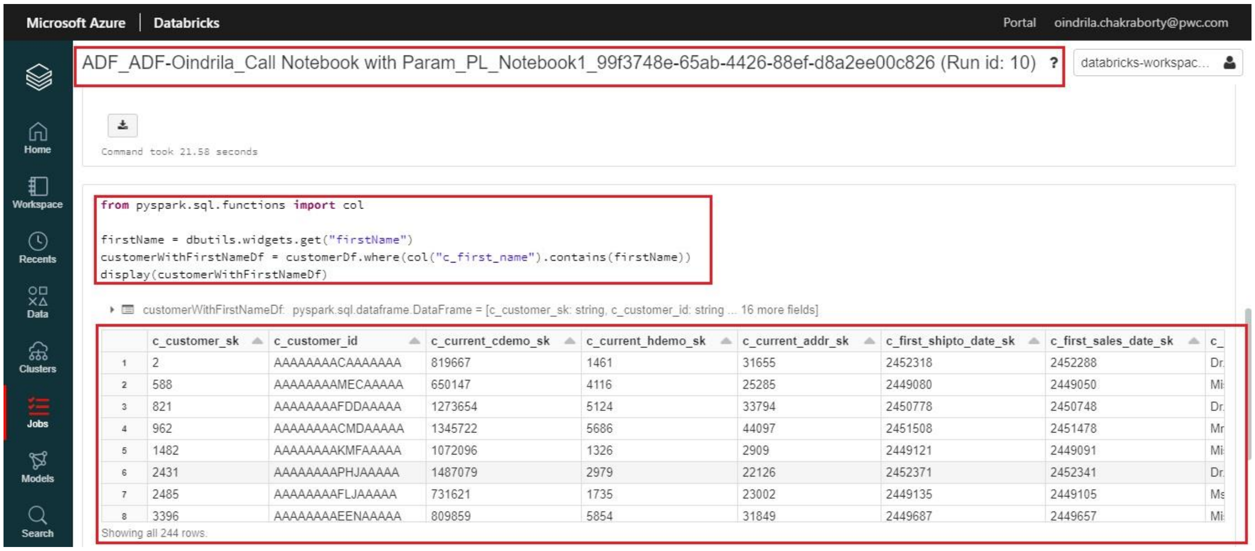
ADF_ADF-Oindrila_Call Notebook with Param_PL_Notebook1_99f3748e-65ab-4426-88ef-d8a2ee00c826 (Run id: 10)

Started: 2021-03-22 04:16:09 IST
Duration: 1m 25s
Status: Succeeded
Run ID: 10
Task: Notebook at /Users/oindrila.chakraborty@pwc.com/ADF_Notebook_Folder/Call ADF Notebook
Parameters:
{"firstName":"Amy"}
Cluster: Driver: Standard_DS3_v2, Workers: Standard_DS3_v2, 2 workers, 5.5 LTS (includes Apache Spark 2.4.3, Scala 2.11) - View Spark UI / Logs / Metrics

Output

```
customerDf = spark.read.options\
                              (
                                header = "true",\
                                delimiter = "|"\
                              )\
                        .csv("dbfs:/mnt/datalakegen2oindrila/CSV-Data/customer.dat")

display(customerDf)
```



ADF_ADF-Oindrila_Call Notebook with Param_PL_Notebook1_99f3748e-65ab-4426-88ef-d8a2ee00c826 (Run id: 10)

```
Command took 21.58 seconds
```

```
from pyspark.sql.functions import col

firstName = dbutils.widgets.get("firstName")
customerWithFirstNameDf = customerDf.where(col("c_first_name").contains(firstName))
display(customerWithFirstNameDf)
```

customerWithFirstNameDf: pyspark.sql.dataframe.DataFrame = [c_customer_sk: string, c_customer_id: string ... 16 more fields]

	c_customer_sk	c_customer_id	c_current_cdemo_sk	c_current_hdemo_sk	c_current_addr_sk	c_first_shipto_date_sk	c_first_sales_date_sk	c_
1	2	AAAAAAAACAAAAAAA	819667	1461	31655	2452318	2452288	Dr.
2	588	AAAAAAAAMECAAAAA	650147	4116	25285	2449080	2449050	Mi
3	821	AAAAAAAAFDDAAAAA	1273654	5124	33794	2450778	2450748	Dr.
4	962	AAAAAAAACMDAAAAA	1345722	5686	44097	2451508	2451478	Mr
5	1482	AAAAAAAAKMFAAAAA	1072096	1326	2909	2449121	2449091	Mi
6	2431	AAAAAAAAPHJAAAAA	1487079	2979	22126	2452371	2452341	Dr.
7	2485	AAAAAAAAFLJAAAAA	731621	1735	23002	2449135	2449105	Ms
8	3396	AAAAAAAAEENAAAAA	809859	5854	31849	2449687	2449657	Mi

Showing all 244 rows.