

😊 Loading data from a file

In this tutorial, we're setting up a notebook in Azure Databricks to load data from a Parquet file. The end goal is to enable users to efficiently access and analyze data stored in the file within the Databricks environment. By following the steps outlined, users can seamlessly upload the file, write the necessary code in Scala to load the data and visualize it in a tabular format. Ultimately, this facilitates data exploration, analysis, and further processing within Azure Databricks.

1. Once you have your cluster in place. Then you are going to create a notebook.

[Compute](#) >

Data Cluster

[Configuration](#)

[Notebooks \(0\)](#)

[Libraries](#)

[Event log](#)

Policy ⓘ

Unrestricted

☐ Multi node ☒ Single node

2. Before that you have to turn on a feature called DBFS file browser.
3. For that you need to open the settings in your workspace and then go to advanced and in there, you have to scroll down to the DBFS file browser.
4. Now you are going to turn on this feature. After enabling this feature you have to refresh the entire page.

Settings

 Workspace admin

Appearance

Identity and access

Security

Compute

Development

Notifications

Advanced

 User

Profile

Preferences

Developer

Default catalog for the workspace:
databricksworkspace120

[► More info](#)

Azure AI services-powered assistive features

Enables access to product features, such as the Databricks Assistant, that use Azure AI services for this workspace. Azure AI services do not access or retain your data.

Default (enabled) ▾

DBFS File Browser

Enable or disable DBFS File Browser

Off ☐

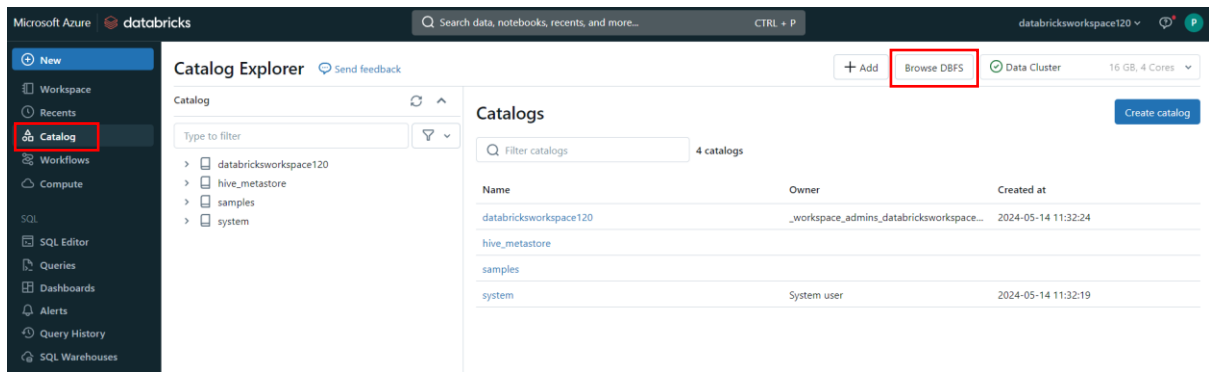
Databricks Autologging

Enable or disable Databricks Autologging for this workspace. When enabled, ML model training runs executed interactively on clusters with supported versions of the Databricks Runtime for Machine Learning will automatically be logged to MLflow.

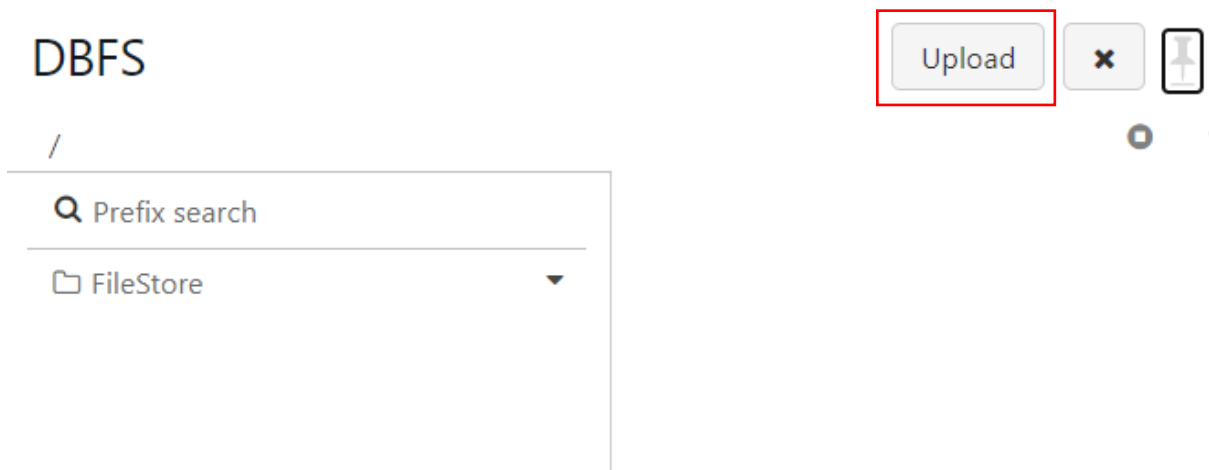
On ☒

[Learn more](#)

5. Then you need to go to the catalog and click on Browse DBFS.



6. Now you need to click on Upload, so that we can upload our Parquet file.



7. Then you can give the folder name as parquet and choose the file from your system.


Upload Data to DBFS

DBFS Target Directory 

/FileStore

Files uploaded to DBFS are accessible by everyone who has access to this workspace. [Learn more](#)

Files 

log.parquet 


0.7 MB
[Remove file](#)


✓ File uploaded to /FileStore/Parquet/log.parquet

Done



- Now click on new and then choose notebook from the menu. It will create your first notebook where you will be writing the code.

Microsoft Azure | databricks

 New

 Workspace

Compute >

Data Cluster  

- Now the first thing you have to do is change the name of your notebook and then change the language to scala.

Working with Databricks

Scala

☆

File Edit View Run Help Last edit was now New cell UI: ON

Workspace

← pulkitkumar2711@gmail.com

Sort: Name

Working with Databricks

Free trial ends in 14 days. Upgrade to Premium in Azure Portal

Start typing or generate with AI (Ctrl + I)...

10. Then we pasted our code, and on behalf of this code, we can pull the data from our parquet file and display the data in tabular format.

1 minute ago (2s)

```
import org.apache.spark.sql.types._
import org.apache.spark.sql.functions._

val file_location = "/FileStore/Parquet/log.parquet"
val file_type = "parquet"

val dataSchema = StructType(Array(
  StructField("Correlationid", StringType, true),
  StructField("Operationname", StringType, true),
  StructField("Status", StringType, true),
  StructField("Eventcategory", StringType, true),
  StructField("Level", StringType, true),
  StructField("Time", StringType, true),
  StructField("Subscription", StringType, true),
  StructField("Eventinitiatedby", StringType, true),
  StructField("Resourcetype", StringType, true),
  StructField("Resourcegroup", StringType, true),
  StructField("Resource", StringType, true)))

val df = spark.read.format(file_type).
  options(Map("header" -> "true")).
  schema(dataSchema).
  load(file_location)

display(df)
```

11. Once you click on run then you can see that data in place.

display(df)

(1) Spark Jobs

df: org.apache.spark.sql.DataFrame = [Correlationid: string, Operationname: string ... 9 more fields]

Table

New result table: ON

	Correlationid	Operationname	Status	Eventcategory	Level	Time	Subs
1	99fe9c3a-e36e-44e0-acd4-58272ab10c7e	Update SQL database	Succeeded	Administrative	Informational	2023-04-25T03:36:59.50...	6912d7a
2	99fe9c3a-e36e-44e0-acd4-58272ab10c7e	Create Deployment	Started	Administrative	Informational	2023-04-25T03:23:57.60...	6912d7a
3	99fe9c3a-e36e-44e0-acd4-58272ab10c7e	Create Deployment	Accepted	Administrative	Informational	2023-04-25T03:24:02.27...	6912d7a
4	99fe9c3a-e36e-44e0-acd4-58272ab10c7e	Registers the Microsoft SQL Database Resource Provider	Started	Administrative	Informational	2023-04-25T03:24:03.46...	6912d7a
5	99fe9c3a-e36e-44e0-acd4-58272ab10c7e	Registers the Microsoft SQL Database Resource Provider	Succeeded	Administrative	Informational	2023-04-25T03:24:07.17...	6912d7a
6	99fe9c3a-e36e-44e0-acd4-58272ab10c7e	Update SQL server	Started	Administrative	Informational	2023-04-25T03:24:13.50...	6912d7a
7	99fe9c3a-e36e-44e0-acd4-58272ab10c7e	'audit' Policy action.	Succeeded	Policy	Warning	2023-04-25T03:24:13.56...	6912d7a
8	99fe9c3a-e36e-44e0-acd4-58272ab10c7e	'auditIfNotExists' Policy action.	Started	Policy	Informational	2023-04-25T03:24:17.20...	6912d7a
9	99fe9c3a-e36e-44e0-acd4-58272ab10c7e	Update SQL server	Accepted	Administrative	Informational	2023-04-25T03:24:17.47...	6912d7a
10	99fe9c3a-e36e-44e0-acd4-58272ab10c7e	Update SQL server firewall rules	Started	Administrative	Informational	2023-04-25T03:25:37.98...	6912d7a
11	99fe9c3a-e36e-44e0-acd4-58272ab10c7e	Update SQL server firewall rules	Started	Administrative	Informational	2023-04-25T03:25:37.98...	6912d7a
12	99fe9c3a-e36e-44e0-acd4-58272ab10c7e	Update Server Connection Policy Create	Started	Administrative	Informational	2023-04-25T03:25:37.98...	6912d7a
13	99fe9c3a-e36e-44e0-acd4-58272ab10c7e	Update SQL database	Started	Administrative	Informational	2023-04-25T03:25:37.99...	6912d7a
14	99fe9c3a-e36e-44e0-acd4-58272ab10c7e	Update Server Connection Policy Create	Succeeded	Administrative	Informational	2023-04-25T03:25:38.10...	6912d7a

5,345+ rows | Truncated data | 1.66 seconds runtime

Refreshed 4 minutes ago