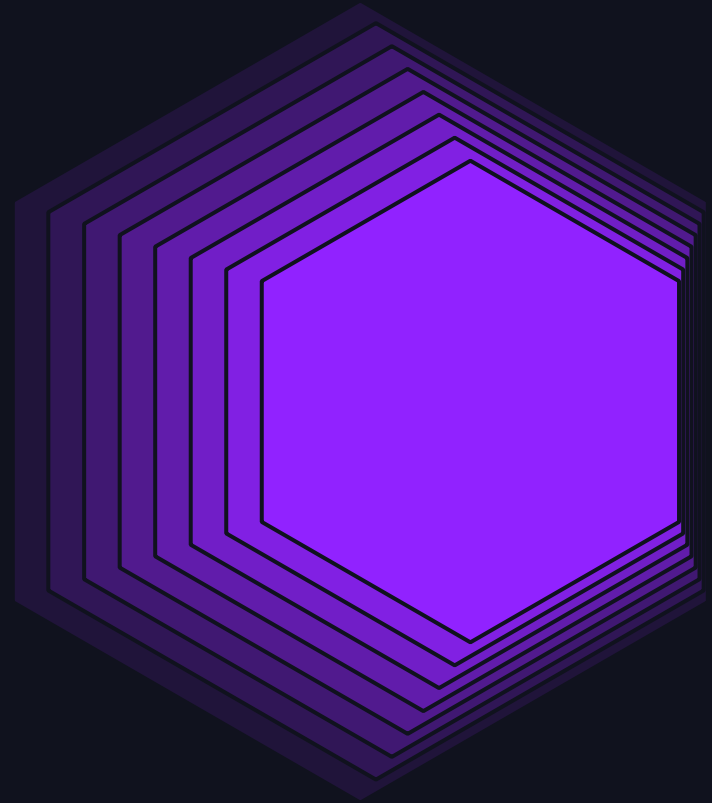


# RSTUDIO, POSIT WORKBENCH, AND DATABRICKS

---

James Blair  
2024-06-11





Posit was founded with the mission to create **open source software** for data science, scientific research, and technical communication.

databricks-ufo - main - RStudio

Go to file/function Addins

ufo-eda.qmd x

Render on Save Render Run

Source Visual B I <> Normal Format Insert Table Outline

```
title: "UFO EDA"
format: html
editor: visual
```

EDA  
Close Connections

### Simple exploration of the UFO data stored on Databricks

```
{r setup}
# Packages
library(sparklyr)
library(pysparklyr)
library(tidyverse)
library(dbplyr)
libraryr(DRT)
```

(Top Level) Copilot: No completions available. Quarto

Console Terminal Background Jobs

R 4.4.0 · ~/Documents/Posit/repos/databricks-ufo/

Natural language support but running in an English locale

R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.

> |

databricks-ufo

Environment History Connections Git

193 MIB

R Global Environment

Environment is empty

Files Plots Packages Help Viewer Pr

me > Documents > Posit > repos > databricks-ufo

Name	Size
..	
.gitignore	676 B
.Renviron	425 B
.Rhistory	914 B
databricks-ufo.Rproj	257 B
deploy.R	579 B
eda	
README.md	5.1 KB
slides	
ufo-dashboard	
ufo-shiny	



sparklyr

tidyverse

dbplyr

connections

GT

# CONNECTING TO DATABRICKS

## Open Source R Packages

### R (Spark)

```
library(sparklyr)
sc <- spark_connect(
  cluster_id = Sys.getenv("DATABRICKS_CLUSTER_ID"),
  version = "14.0",
  method = "databricks_connect"
)
#> ! Changing host URL to: https://rstudio-partner-posit-
default.cloud.databricks.com
#> i Retrieving info for cluster:'0907-051241-shqxcbmx'
#> ✓ Cluster: '0907-051241-shqxcbmx' | DBR: '14.0' [437ms]
#>
#> i Attempting to load 'r-sparklyr-databricks-14.0'
#> ✓ Python environment: 'r-sparklyr-databricks-14.0' [1.9s]
#>
#> i Connecting to 'Test Cluster'
#> ✓ Connected to: 'Test Cluster' [7ms]
```

### R (ODBC)

```
library(DBI)
con <- dbConnect(
  odbc::databricks(),
  HTTPPath = Sys.getenv("DATABRICKS_HTTP_PATH"),
  timeout = 10
)
con
#> <OdbcConnection> token@Spark SQL
#> Database: hive_metastore
#> Spark SQL Version: 3.1.1
```

ufo-eda.qmd x

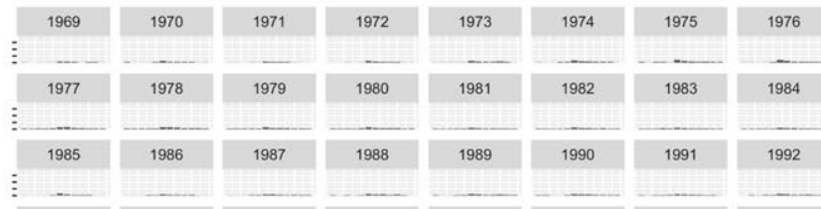
Render on Save Render

Source Visual Normal Format Table

```
{r}
ufo_by_day <- nuforc >
  filter(!is.na(date_time)) >
  count(date = as.character(as.Date(date_time))) >
  collect() >
  mutate(date = as.Date(date),
         month = month(date),
         year = year(date))

ufo_by_day >
  ggplot(aes(x = month, y = n)) +
  facet_wrap(~year) +
  geom_col() +
  theme(axis.text.x = element_blank(),
        axis.text.y = element_blank()) +
  labs(title = "UFO Sightings by Month",
       x = "",
       y = "")
```

## UFO Sightings by Month



Chunk 5

Copilot: No completions available. Quarto

Console

Environment History Connections Git Tutorial

SQL Help Refresh Connection Data

Test Cluster (0907-051241-shqxcbm)

Spark

```
databricks_llama_2_models
demos
  default
  information_schema
  nuforc
    nuforc_reports
      summary : chr MADAR Node 100 Steady flash ...
      country : chr USA USA USA USA France USA
      city : chr Mountlake Terrace Hamden Ch...
      state : chr WA CT VA MI NA CA
      date_time : dttm 2019-06-23 12:53:00 2019-06 ...
      shape : chr NA light circle light cigar ...
      duration : chr NA 5 hours 15 seconds 2 min ...
      stats : chr Occurred : 6/23/2019 18:53 ...
      report_link : chr http://www.nuforc.org/webre ...
      text : chr MADAR Node 100 Steady flash ...
      posted : dttm 2019-06-26 18:00:00 2019-06 ...
      city_latitude : num 47.7941 41.3739408 38.05596 ...
      city_longitude : num -122.3066 -72.9213248
                    -78.4 ...
      city_location : chr POINT(-122.3066 47.7941) PO ...
    testing
```

Files Plots Packages Help Viewer Presentation

Folder Blank File Delete Rename

Home &gt; Documents &gt; Posit &gt; repos &gt; databricks-ufo

	Name	Size	Modified
	..		
<input type="checkbox"/>	.gitignore	676 B	May 31, 2024, 2:11
<input type="checkbox"/>	.Renv	468 B	May 31, 2024, 2:21
<input type="checkbox"/>	.Rhistory	914 B	May 31, 2024, 2:21
<input type="checkbox"/>	databricks-ufo Rnri	257 B	May 31, 2024, 2:21

ufo-eda.qmd x

Render on Save

Source Visual

Outline

```
{r}
ufo_by_day <- nuforc >
  filter(!is.na(date_time)) >
  count(date = as.character(as.Date(date_time))) >
  collect() >
  mutate(date = as.Date(date),
         month = month(date),
         year = year(date))

ufo_by_day >
  ggplot(aes(x = month, y = n)) +
  facet_wrap(~year) +
  geom_col() +
  theme(axis.text.x = element_blank(),
        axis.text.y = element_blank()) +
  labs(title = "UFO Sightings by Month",
       x = "",
       y = "")
```

## UFO Sightings by Month



Chunk 5

Copilot: No completions available. Quarto

Console

EDA

Close Connections

Environment History Connections Git Tutorial

SQL Help Refresh Connection Data

Test Cluster (0907-051241-shqxcbmx)

Spark

```
databricks_llama_2_models
demos
  default
  information_schema
  nuforc
    nuforc_reports
      summary : chr MADAR Node 100 Steady flash ...
      country : chr USA USA USA USA France USA
      city : chr Mountlake Terrace Hamden Ch...
      state : chr WA CT VA MI NA CA
      date_time : dttm 2019-06-23 12:53:00 2019-06 ...
      shape : chr NA light circle light cigar ...
      duration : chr NA 5 hours 15 seconds 2 min ...
      stats : chr Occurred : 6/23/2019 18:53 ...
      report_link : chr http://www.nuforc.org/webre ...
      text : chr MADAR Node 100 Steady flash ...
      posted : dttm 2019-06-26 18:00:00 2019-06 ...
      city_latitude : num 47.7941 41.3739408 38.05596 ...
      city_longitude : num -122.3066 -72.9213248
                    -78.4 ...
      city_location : chr POINT(-122.3066 47.7941) PO ...
    testing
```

Files Plots Packages Help Viewer Presentation

Folder Blank File Delete Rename

Home Documents Posit repos databricks-ufo

Name	Size	Modified
..		
.gitignore	676 B	May 31, 2024, 2:24
.Renviron	468 B	May 31, 2024, 2:24
.Rhistory	914 B	May 31, 2024, 2:24
databricks-ufo Rrnoi	257 B	May 31, 2024, 2:24

ufo-eda.qmd x

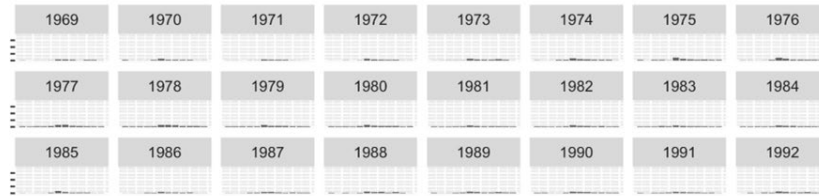
Render on Save ABC Render Run Outline

Source Visual B I <> Normal Format Insert Table

```
{r}
ufo_by_day <- nuforc >
  filter(!is.na(date_time)) >
  count(date = as.character(as.Date(date_time))) >
  collect() >
  mutate(date = as.Date(date),
         month = month(date),
         year = year(date))

ufo_by_day >
  ggplot(aes(x = month, y = n)) +
  facet_wrap(~year) +
  geom_col() +
  theme(axis.text.x = element_blank(),
        axis.text.y = element_blank()) +
  labs(title = "UFO Sightings by Month",
       x = "",
       y = "")
```

## UFO Sightings by Month



Chunk 5

Copilot: No completions available. Quarto

Console

Environment History Connections Git Tutorial

SQL ? Help Refresh Connection Data

Test Cluster (0907-051241-shqxcmbx)

Spark

databricks\_llama\_2\_models

demos

- default
- information\_schema
- nuforc
  - nuforc\_reports
    - summary : chr MADAR Node 100 Steady flash ...
    - country : chr USA USA USA USA France USA
    - city : chr Mountlake Terrace Hamden Ch...
    - state : chr WA CT VA MI NA CA
    - date\_time : dtm 2019-06-23 12:53:00 2019-06 ...
    - shape : chr NA light circle light cigar ...
    - duration : chr NA 5 hours 15 seconds 2 min ...
    - stats : chr Occurred : 6/23/2019 18:53 ...
    - report\_link : chr http://www.nuforc.org/webre ...
    - text : chr MADAR Node 100 Steady flash ...
    - posted : dtm 2019-06-26 18:00:00 2019-06 ...
    - city\_latitude : num 47.7941 41.3739408 38.05596 ...
    - city\_longitude : num -122.3066 -72.9213248 -78.4 ...
    - city\_location : chr POINT(-122.3066 47.7941) PO ...
- testing

Files Plots Packages Help Viewer Presentation

Folder Blank File Delete Rename

Home Documents Posit repos databricks-ufo

Name	Size	Modified
..		
.gitignore	676 B	May 31, 2024, 2:24
.Renvirom	468 B	May 31, 2024, 2:24
.Rhistory	914 B	May 31, 2024, 2:24
databricks-ufo.Rproj	257 B	May 31, 2024, 2:24



- New
- Workspace
- Recents
- Catalog
- Workflows
- Compute
- SQL
  - SQL Editor
  - Queries
  - Dashboards
  - Alerts
  - Query History
  - SQL Warehouses
- Data Engineering
  - Job Runs
  - Data Ingestion
  - Delta Live Tables
- Machine Learning
  - Playground
  - Experiments
  - Features
  - Models
  - Serving
- Marketplace
- Partner Connect

Compute >

## Test Cluster ✓

More ⋮ Terminate Edit

Configuration Notebooks (0) Libraries Event log **Spark UI** Driver logs Metrics Apps Spark compute UI - Master ⌵

Open in new tab ↗

Jobs Stages Storage Environment Executors **SQL / DataFrame** JDBC/ODBC Server Structured Streaming Connect

### SQL / DataFrame

Completed Queries: 992

Completed Queries (992)

Page: 1 2 3 4 5 6 7 8 9 10 >

10 Pages. Jump to  . Show  items in a page.

ID	Description	Submitted	Duration	Job IDs	Sub Execution IDs
1048	Spark Connect - session_id: "ede03f6b-8e4a-4f3c-9fc2-375e483bf696" user_context { user_id: "548199642521214... <pre>sql {   query: "SELECT `date`, COUNT(*) AS `n` FROM ( SELECT `nuforc_reports`.*,     CAST(CAST(`date_time` AS DATE) AS STRING) AS `date` FROM `demos`.`nuforc`.`nuforc_reports` WHERE (NOT(`date_time` IS NULL)))` `q01` GROUP BY `date`" }</pre>	2024/05/31 20:48:56	1 s	[323][324]	
1047	Spark Connect - session_id: "ede03f6b-8e4a-4f3c-9fc2-375e483bf696" user_context { user_id: "548199642521214... +details	2024/05/31 20:48:56	20 ms		
1046	Spark Connect - session_id: "ede03f6b-8e4a-4f3c-9fc2-375e483bf696" user_context { user_id: "548199642521214... +details	2024/05/31 20:48:56	20 ms		
1045	Spark Connect - session_id: "ede03f6b-8e4a-4f3c-9fc2-375e483bf696" user_context { user_id: "548199642521214... +details	2024/05/31 20:48:55	50 ms		
1044	Spark Connect - session_id: "ede03f6b-8e4a-4f3c-9fc2-375e483bf696" user_context { user_id: "548199642521214... +details	2024/05/31 20:48:54	47 ms		
1043	Spark Connect - session_id: "ede03f6b-8e4a-4f3c-9fc2-375e483bf696" user_context { user_id: "548199642521214... +details	2024/05/31 20:48:20	22 ms		
1042	Spark Connect - session_id: "ede03f6b-8e4a-4f3c-9fc2-375e483bf696" user_context { user_id: "548199642521214... +details	2024/05/31 20:48:09	22 ms		

- New
- Workspace
- Recents
- Catalog
- Workflows
- Compute
- SQL
  - SQL Editor
  - Queries
  - Dashboards
  - Alerts
  - Query History
  - SQL Warehouses
- Data Engineering
  - Job Runs
  - Data Ingestion
  - Delta Live Tables
- Machine Learning
  - Playground
  - Experiments
  - Features
  - Models
  - Serving
  - Marketplace
  - Partner Connect

Compute > **Test Cluster** ✓

Configuration **Notebooks (0)** Libraries Event log **Spark UI** Driver logs Metrics Apps Spark compute UI - Master ▼

More ⋮ Terminate Edit

Jobs Stages Storage Environment Executors **SQL / DataFrame** JDBC/ODBC Server Structured Streaming Connect Open in new tab

SQL / DataFrame

Completed Queries: 992

Completed Queries (992)

Page: 1 2 3 4 5 6 7 8 9 10 > 10 Pages. Jump to 1 . Show 100 items in a page. Go

ID	Description	Submitted	Duration	Job IDs	Sub Execution IDs
1048	Spark Connect - session_id: "ede03f6b-8e4a-4f3c-9fc2-375e483bf696" user_context { user_id: "548199642521214... <pre>sql {   query: "SELECT `date`, COUNT(*) AS `n` FROM ( SELECT `nuforc_reports`.*, CAST(CAST(`date_time` AS DATE) AS STRING) AS `date` FROM `demos`.`nuforc`.`nuforc_reports` WHERE (NOT(`date_time` IS NULL))) `q01` GROUP BY `date`" }</pre>	2024/05/31 20:48:56	1 s	[323][324]	
1047	Spark Connect - session_id: "ede03f6b-8e4a-4f3c-9fc2-375e483bf696" user_context { user_id: "548199642521214... +details	2024/05/31 20:48:56	20 ms		
1046	Spark Connect - session_id: "ede03f6b-8e4a-4f3c-9fc2-375e483bf696" user_context { user_id: "548199642521214... +details	2024/05/31 20:48:56	20 ms		
1045	Spark Connect - session_id: "ede03f6b-8e4a-4f3c-9fc2-375e483bf696" user_context { user_id: "548199642521214... +details	2024/05/31 20:48:55	50 ms		
1044	Spark Connect - session_id: "ede03f6b-8e4a-4f3c-9fc2-375e483bf696" user_context { user_id: "548199642521214... +details	2024/05/31 20:48:54	47 ms		
1043	Spark Connect - session_id: "ede03f6b-8e4a-4f3c-9fc2-375e483bf696" user_context { user_id: "548199642521214... +details	2024/05/31 20:48:20	22 ms		
1042	Spark Connect - session_id: "ede03f6b-8e4a-4f3c-9fc2-375e483bf696" user_context { user_id: "548199642521214... +details	2024/05/31 20:48:09	22 ms		

# REQUIREMENTS



## Authentication

- Personal access token
- .Renviron or other way to securely provide tokens and other connection details



## sparklyr

- Latest release of sparklyr package
- Local Python environment - Sparklyr will install for you
- DBR  $\geq$  13.0



## ODBC

- Latest release of odbc package
- Databricks ODBC Driver

# REQUIREMENTS



## Authentication

- Personal access token
- .Renviron or other way to securely provide tokens and other connection details



## sparklyr

- Latest release of sparklyr package
- Local Python environment - Sparklyr will install for you
- DBR  $\geq$  13.0



## ODBC

- Latest release of odbc package
- Databricks ODBC Driver





The logo icon consists of two overlapping, stylized 'X' shapes. The left 'X' is blue and the right 'X' is orange. They are positioned such that they appear to be intertwined.

**posit**<sup>TM</sup> Workbench

### Sessions


+ New Session  
Create a new session

#### New Session

 Jupyter Notebook     JupyterLab     RStudio Pro     VS Code

Session Name  
RStudio Pro Session

Session Credentials Edit Credentials

 Posit Azure

Cluster Options

Resource Profile  
Small

CPU's	Memory (GB)
1	1.95

Image  
rstudio-workbench-preview:dev-ubuntu2204-2024.07.0-daily-188.pro1 (default) Edit

Join session when ready     Notify when ready

Cancel Start Session

RStudio Pro interface showing a Databricks connection configuration and environment details.

**Environment**

Clusters	Name	Runtime	Unity Catalog	Creator
⊗	Secondary Cluster	14.2 ML	Yes	james@posit.co
✔	Test Cluster	14.0.x-cpu-ml-scala2.12	Yes	james@posit.co

**Configuration**

Cluster ID	0907-051241-shqxcbmx
Policy	Personal Compute
Creator	james@posit.co
Source	UI
Access	Single User
Unity Catalog	Yes

**Performance**

Runtime	14.0.x-cpu-ml-scala2.12
Node Type	i3.xlarge
Active Cores	4
Active Memory	30.5 GB
Started	1/22/2024, 8:37:51 PM

**Files**

Name	Size	Modified
.Rhistory	0 B	Jun 5, 2024, 9:31 AM
databricks-connections.qmd	1.6 KB	Jun 5, 2024, 9:39 AM
R		

**Code Editor**

```

---
title: "Databricks Connections"
format: html
editor: visual
---

{r setup}
# Packages ----
library(tidyverse)
library(dbplyr)
library(sparklyr)
library(DBI)

```

**Console**

```

R 4.4.0 - ~/
Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Connected to your session in progress, last started 2024-Jun-05 16:35:07 UTC (11 minutes ago)
>

```

RStudio Pro interface showing a Databricks connection configuration and a terminal window.

**Environment Panel:**

Name	Runtime	Unity Catalog	Creator
Secondary Cluster	14.2 ML	Yes	james@posit.co
Test Cluster	14.0.x-cpu-ml-scala2.12	Yes	james@posit.co

**Configuration:**

Cluster ID	0907-051241-shqxcbmx
Policy	Personal Compute
Creator	james@posit.co
Source	UI
Access	Single User
Unity Catalog	Yes

**Performance:**

Runtime	14.0.x-cpu-ml-scala2.12
Node Type	i3.xlarge
Active Cores	4
Active Memory	30.5 GB
Started	1/22/2024, 8:37:51 PM

**Code Editor:**

```

---
title: "Databricks Connections"
format: html
editor: visual
---

{r setup}
# Packages ----
library(tidyverse)
library(dbplyr)
library(sparklyr)
library(DBI)

```

**Terminal:**

```

R 4.4.0 - ~/
Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Connected to your session in progress, last started 2024-Jun-05 16:35:07 UTC (11 minutes ago)
>

```

**File Explorer:**

Name	Size	Modified
.Rhistory	0 B	Jun 5, 2024, 9:31 AM
databricks-connections.qmd	1.6 KB	Jun 5, 2024, 9:39 AM
R		



dev.databricks.posit.team/s/e2976cfc78a317cfea691/?launcher=1

File Edit Code View Plots Session Build Debug Profile Tools Help

Render on Save Render Run

Source Visual B I / < / > Normal Format Insert Table Outline

Sparklyr Connect

```
---
title: "Databricks Connections"
format: html
editor: visual
---
```

```
{r setup}
# Packages ----
library(tidyverse)
library(dbplyr)
library(sparklyr)
library(DBI)
```

### Sparklyr

Reference: <https://spark.rstudio.com/deployment/databricks>

### Connect

```
{r}
sc <- spark_connect(
  method = "databricks_connect",
  version = "14.0",
  cluster_id = "0907-051241-shqxcbm"
)
```

Chunk 2

Environment History Connections Tutorial Databricks

Clusters

Name	Runtime	Unity Catalog	Creator
Secondary Cluster	14.2 ML	Yes	james@posit.co
	14.0.x-cpu-ml-scala2.12	Yes	james@posit.co

Search by name or ID

Cluster ID: 0907-051241-shqxcbm  
✓ Found - Cluster's DBR is 14.0

Python Env: ✓ Found - Using 'r-sparklyr-databricks-14.0'

Host URL: rstudio-partner-posit-default.cloud.databricks.com

Connection: Connect from: R Console

```
library(sparklyr)
sc <- spark_connect(
  cluster_id = "0907-051241-shqxcbm",
  version = "14.0",
  method = "databricks_connect"
```

Using "Databricks Connect (DBR 13+)"

OK Cancel

Files Plots Packages Help Viewer Presentation

New Folder New Blank File Upload Delete Rename More

Home

Name	Size	Modified
.Rhistory	0 B	Jun 5, 2024, 9:31 AM
databricks-connections.qmd	1.6 KB	Jun 5, 2024, 9:39 AM
R		

### Sessions

+ New Session  
Create a new session

#### New Session

Jupyter Notebook  JupyterLab  RStudio Pro  VS Code

Session Name  
VS Code Session

Session Credentials Edit Credentials

Posit AWS

Cluster Options

Resource Profile  
Small

CPU's	Memory (GB)
1	1.95

Image  
rstudio-workbench-preview:dev-ubuntu2204-2024.07.0-daily-188.pro1 (default) Edit

Join session when ready  Notify when ready

Cancel Start Session

localhost:8787/s/0da293c9ab5a7a60b9627/?folder=/home/james/databricks-tour

DATABRICKS: CONF... databricks-python.qmd U X

Workspace

- Auth Profile 'workbench'
- User james@posit.co
- Host https://rstudio-par...

Cluster

- Name Test Cluster
- Cluster ID 0907-05124...
- Driver I3.xlarge
- Worker None (single no...)
- Databricks Runtime 14...
- State RUNNING
- Creator james@posit.co

Sync Destination

- Name databricks-to...
- State STOPPED
- Path /Users/james@posi...

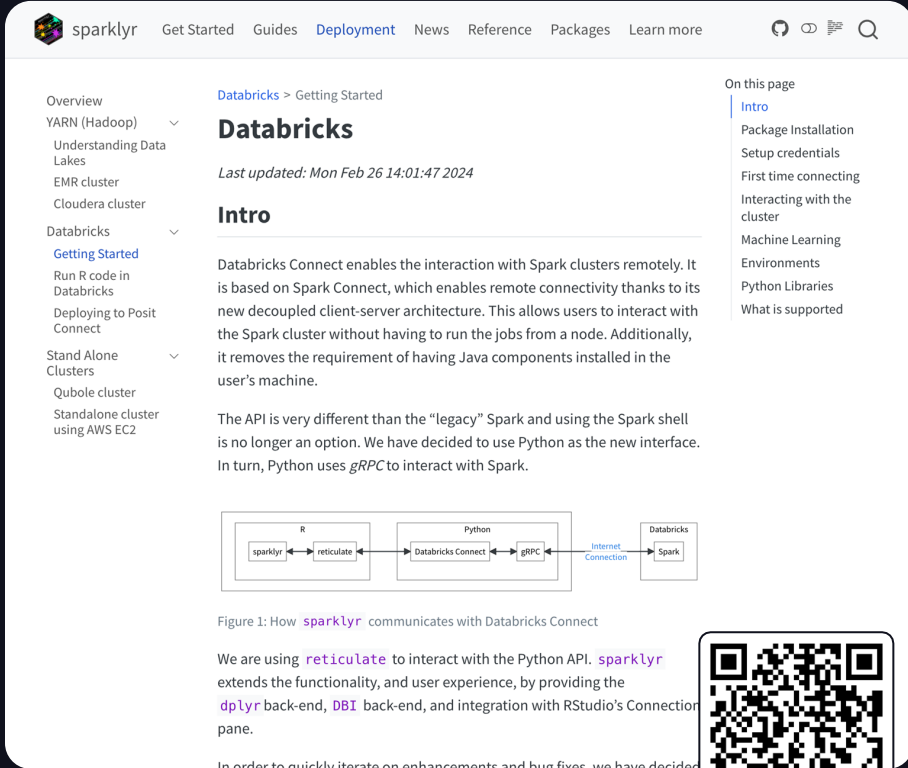
```
python > databricks-python.qmd > Python, Databricks and VS Code
1 # Python, Databricks and VS Code
2
3 ## Databricks Connect
4 {python}
5 from databricks.connect import DatabricksSession
6
7 spark = DatabricksSession.builder.getOrCreate()
8
9 nuforc = spark.read.table("demos.nuforc.nuforc_reports")
10 nuforc.show(5)
11
12
```

main\* Databricks Connect enabled Posit Workbench R: (not attached) Ln 1, Col 1 Spaces: 4 UTF-8 LF Quarto Layout: U.S.

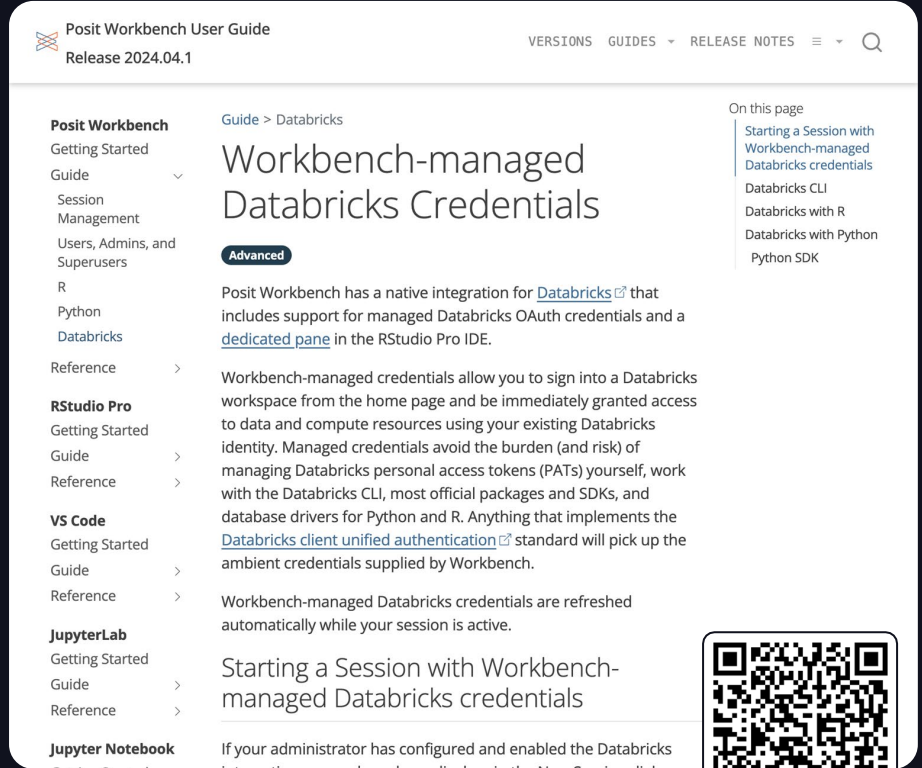


**COMING SOON**

# LEARN MORE



The screenshot shows the sparklyr documentation page. The navigation bar includes 'sparklyr', 'Get Started', 'Guides', 'Deployment', 'News', 'Reference', 'Packages', and 'Learn more'. The left sidebar lists various topics like 'Overview', 'YARN (Hadoop)', 'EMR cluster', 'Cloudera cluster', 'Databricks', 'Getting Started', 'Run R code in Databricks', 'Deploying to Posit Connect', 'Stand Alone Clusters', and 'Qubole cluster'. The main content area is titled 'Databricks' and 'Intro', with a sub-header 'Getting Started'. It explains that Databricks Connect enables interaction with Spark clusters remotely. A diagram illustrates the communication flow between sparklyr, reticulate, Databricks Connect, gRPC, and Spark. A QR code is located at the bottom right of the page.



The screenshot shows the Posit Workbench User Guide page for Release 2024.04.1. The navigation bar includes 'Posit Workbench User Guide', 'Release 2024.04.1', 'VERSIONS', 'GUIDES', and 'RELEASE NOTES'. The left sidebar lists sections like 'Posit Workbench', 'Getting Started', 'Guide', 'Session Management', 'Users, Admins, and Superusers', 'R', 'Python', 'Databricks', 'Reference', 'RStudio Pro', 'Getting Started', 'Guide', 'Reference', 'VS Code', 'Getting Started', 'Guide', 'Reference', 'JupyterLab', 'Getting Started', 'Guide', 'Reference', and 'Jupyter Notebook'. The main content area is titled 'Workbench-managed Databricks Credentials' and is marked as 'Advanced'. It explains that Posit Workbench has a native integration with Databricks that includes support for managed Databricks OAuth credentials. A QR code is located at the bottom right of the page.