



Help

**Databricks** > **Data Engineering** > Selenium chrome driver on databricks driver On the...

Options ⋮

## Selenium chrome driver on databricks driver On the databricks community, I see repeated problems regarding the selenium installation on the databricks...

**Hubert\_Dudek1**

Esteemed Contributor III



11-09-2022 06:12 AM

### Selenium chrome driver on databricks driver

On the databricks community, I see repeated problems regarding the selenium installation on the databricks driver. Installing selenium on databricks can be surprising, but for example, sometimes we need to grab some datasets behind fancy authentication, and selenium is the most accessible tool to do that. Of course, always remember to check the most uncomplicated alternatives first. For example, if we need to download an HTML file, we can use `SparkContext.addFile()` or just use the requests library. If we need to parse HTML without simulating user actions or downloading

complicated pages, we can use BeautifulSoup. Please remember that selenium is running on the driver only (workers are not utilized), so just for the selenium part single node cluster is the preferred setting.

## Installation

The easiest solution is to use apt-get to install ubuntu packages, but often version in the ubuntu repo is outdated. Recently that solution stopped working for me, and I decided to take a different approach and to get the driver and binaries from chromium-browser-snapshots <https://commondatastorage.googleapis.com/chromium-browser-snapshots/index.html> Below script download the newest version of browser binaries and driver. Everything is saved to /tmp/chrome directory. We must also set the chrome home directory to /tmp/chrome/chrome-user-data-dir. Sometimes, chromium complains about missing libraries. That's why we also install libgbm-dev. The below script will create a bash file implementing mentioned steps.

```
dbutils.fs.mkdirs("dbfs:/databricks/scripts/")
dbutils.fs.put("/databricks/scripts/selenium-install.sh", "")
#!/bin/bash
%sh
LAST_VERSION="https://www.googleapis.com/download/storage/v1/b/chromium-browser-
snapshots/o/Linux_x64%2FLAST_CHANGE?alt=media"
VERSION=$(curl -s -S $LAST_VERSION)
if [ -d $VERSION ] ; then
    echo "version already installed"
    exit
fi

rm -rf /tmp/chrome/$VERSION
mkdir -p /tmp/chrome/$VERSION

URL="https://www.googleapis.com/download/storage/v1/b/chromium-browser-snapsho
t/o/Linux_x64%2F$VERSION%2Fchrome-linux.zip?alt=media"
ZIP="{VERSION}-chrome-linux.zip"

curl -# $URL > /tmp/chrome/$ZIP
unzip /tmp/chrome/$ZIP -d /tmp/chrome/$VERSION

URL="https://www.googleapis.com/download/storage/v1/b/chromium-browser-snapsho
t/o/Linux_x64%2F$VERSION%2Fchromedriver_linux64.zip?alt=media"
ZIP="{VERSION}-chromedriver_linux64.zip"

curl -# $URL > /tmp/chrome/$ZIP
unzip /tmp/chrome/$ZIP -d /tmp/chrome/$VERSION

mkdir -p /tmp/chrome/chrome-user-data-dir

rm -f /tmp/chrome/latest
ln -s /tmp/chrome/$VERSION /tmp/chrome/latest

# to avoid errors about missing libraries
sudo apt-get update
sudo apt-get install -y libgbm-dev
"", True)
display(dbutils.fs.ls("dbfs:/databricks/scripts/"))
```

The script was saved to DBFS storage as `/dbfs/databricks/scripts/selenium-install.sh`. We can set it as an init script for the server. Click your cluster in "compute" -> click "Edit" -> "configuration" tab -> scroll down to "Advanced options" -> click "Init Scripts" -> select "DBFS" and set "Init script path" as `/dbfs/databricks/scripts/selenium-install.sh` -> click "add".

### ▼ Advanced options

#### Azure Data Lake Storage credential passthrough ?

☐ Enable credential passthrough for user-level data access

Spark   Logging   **Init Scripts**   Permissions

#### Init scripts ?

Type	File path

Destination   Init script path

DBFS | ▼

`/dbfs/databricks/scripts/selenium-install.sh`

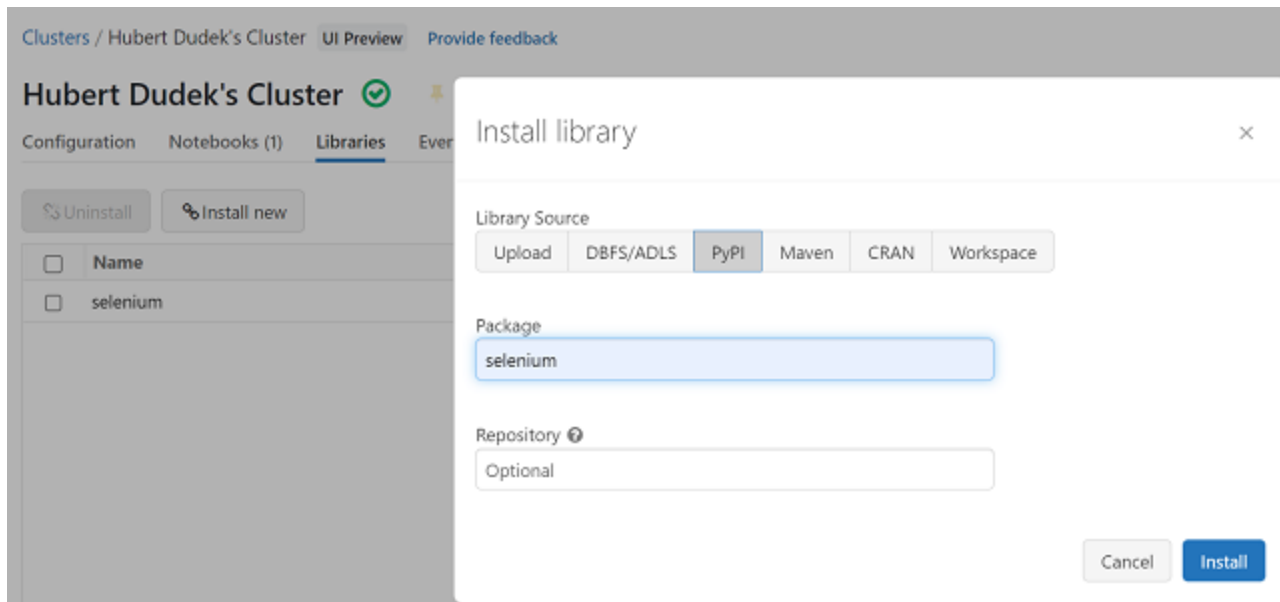
Add

If

you haven't set the init script, please run the below command.

```
%sh
/dbfs/databricks/scripts/selenium-install.sh
```

Now we can install selenium. Click your cluster in "compute" -> click "Libraries" -> click "Install new" -> click "PyPI" -> set "Package" as "selenium" -> click "install".



Alternatively (which is less convenient), you can install it every time in your notebook by running the below command.

```
%pip install selenium
```

So let's start webdriver. We can see that Service and binary\_location point to driver and binaries, which were downloaded and unpacked by our script.

```
from selenium import webdriver
from selenium.webdriver.chrome.service import Service
s = Service('/tmp/chrome/latest/chromedriver_linux64/chromedriver')
options = webdriver.ChromeOptions()
options.binary_location = "/tmp/chrome/latest/chrome-linux/chrome"
options.add_argument('headless')
options.add_argument('--disable-infobars')
options.add_argument('--disable-dev-shm-usage')
options.add_argument('--no-sandbox')
options.add_argument('--remote-debugging-port=9222')
options.add_argument('--homedir=/tmp/chrome/chrome-user-data-dir')
options.add_argument('--user-data-dir=/tmp/chrome/chrome-user-data-dir')
prefs = {"download.default_directory":"/tmp/chrome/chrome-user-data-di",
        "download.prompt_for_download":False
}
options.add_experimental_option("prefs",prefs)
driver = webdriver.Chrome(service=s, options=options)
```

Let's test webdriver. We will take the last posts from the databricks community and convert them to a dataframe.

```

from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
driver.execute("get", {'url': 'https://community.databricks.com/s/discussions?page=1&filter=All'})
date = [elem.text for elem in WebDriverWait(driver, 20).until(EC.visibility_of_all_elements_located((By.CSS_SELECTOR, "lightning-formatted-date-time")))]
title = [elem.text for elem in WebDriverWait(driver, 20).until(EC.visibility_of_all_elements_located((By.CSS_SELECTOR, "p[class='Sub-heading1']")))]

```

```

from pyspark.sql.types import StringType, StructType, StructField

schema = StructType([
    StructField("date", StringType()),
    StructField("title", StringType())
])
df = spark.createDataFrame(list(zip(date, title)), schema=schema)
display(df)

```

df: pyspark.sql.dataframe.DataFrame = [date: string, title: string]

Table +

	date	title
1	Nov 09, 2022	new learner
2	Nov 09, 2022	authentication is not configured for provider
3	Nov 09, 2022	say, I want to download 2 files from this directory (dbfs:/databricks-datasets/abc-quality/) to my local filesystem, how do I do it? I understand that if those files are inside FileStore directory, it is much straightforward, which someone posts some solution here: <a href="https://medium.datadriveninvestor.com/how-to-download-a-file-from-databricks-filestore-to-a-local-machine-ae0c40f164f5">https://medium.datadriveninvestor.com/how-to-download-a-file-from-databricks-filestore-to-a-local-machine-ae0c40f164f5</a> Hence, now I am trying to see if it is possible to do a copy files from directory dbfs:/databricks-datasets/...
4	Nov 09, 2022	Spark SQL output multiple small files
5	Nov 09, 2022	How to do bucketing in Databricks?
6	Nov 09, 2022	Py4JError: com.xxx.TipsScore.score does not exist in the JVM
7	Nov 09, 2022	Issue converting the datasets into JSON

Showing all 10 rows. | 1.39 seconds runtime

We can see the latest posts in our dataframe. Now we can quit the driver.

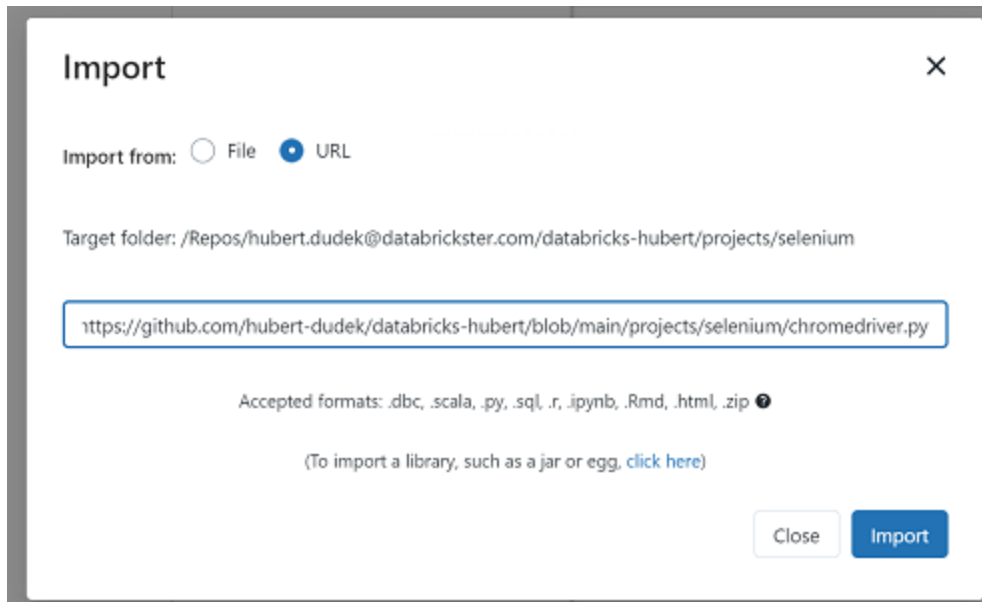
```
driver.quit()
```

The version of that article as ready to-run notebook is available at:

<https://github.com/hubert-dudek/databricks-hubert/blob/main/projects/selenium/chromedriver.py>

To import that notebook into databricks, go to the folder in your "workplace" -> from the arrow menu, select "URL" -> click "import" -> put

<https://raw.githubusercontent.com/hubert-dudek/databricks-hubert/main/projects/selenium/chromedriver...> as URL.



Labels: **Selenium**

**install\_library.png**   
37 KB

**results.png**   
114 KB

**init.png**   
30 KB

**import.png**   
30 KB



30 Kudos

Reply

## 10 REPLIES

[All forum topics](#) < [Previous Topic](#) [Next Topic](#) >



**swrd**

New Contributor III



I followed your article but got this error message:

config\_set\_up\_seleniumPython

FileEditViewRunHelpLast edit was 6 minutes agoGive feedback

Press F11 to exit full screen

Run allStephendit1/1



```
3 from webdriver_manager.chrome import ChromeDriverManager
4 s = Service('/tmp/chrome/latest/chromedriver_linux64/chromedriver')
5 options = webdriver.ChromeOptions()
6 options.binary_location = "/tmp/chrome/latest/chrome-linux/chrome"
7 options.add_argument("--start-maximized")
8 options.add_experimental_option('excludeSwitches', ['enable-logging'])
9 options.add_argument('--disable-infobars')
10 options.add_argument('--disable-dev-shm-usage')
11 options.add_argument('--no-sandbox')
12 options.add_argument('--remote-debugging-port=9222')
13 options.add_argument('--headless')
14 options.add_argument('--user-data-dir=/tmp/chrome/chrome-user-data-dir')
15 options.add_argument('--user-data-dir=/tmp/chrome/chrome-user-data-dir')
16 prefs = {"download.default_directory":"/tmp/chrome/chrome-user-data-dir",
17         "download.prompt_for_download":false}
18 }
19 options.add_experimental_option("prefs",prefs)
20 driver = webdriver.Chrome(service=s, options=options)

@WebDriverException: Message: unknown error: Chrome failed to start: exited abnormally.
(chrome not reachable)
(The process started from chrome location /tmp/chrome/latest/chrome-linux/chrome is no longer running, so ChromeDriver is assuming that Chrome has crashed.)
Stacktrace:
#0 0x55b679bd3ff2
#1 0x55b679b64803
#2 0x55b679b0910e1
#3 0x55b679927a5c
#4 0x55b6799236da
#5 0x55b679964617
#6 0x55b67995b903
#7 0x55b679926181
#8 0x55b67992f33e
#9 0x55b679b96ca1
#10 0x55b679baba17
#11 0x55b679bab2ff
#12 0x55b679bacc195
#13 0x55b679b084b3
#14 0x55b679bac4eb
#15 0x55b679b88def
#16 0x55b679bc7738
#17 0x55b679bc7857
#18 0x55b679be200f
#19 0x7f3d22e0e609 start_thread
#20 0x7f3d228b2133 clone

Command took 1.75 seconds --- by stephen.williams@databricks.com at 11/11/2022, 02:27:53 on Stephen Williams's Cluster

Cwd: /
1 from selenium.webdriver.common.by import By
2 from selenium.webdriver.support.ui import WebDriverWait
3 from selenium.webdriver.support import expected_conditions as EC
4 driver.execute("get", ('url': 'https://community.databricks.com/s/discussions?page=1&filter=All'))
5 date = [elem.text for elem in WebDriverWait(driver, 20).until(EC.visibility_of_all_elements_located((By.CSS_SELECTOR, "lightning-formatted-date-time")))]
6 title = [elem.text for elem in WebDriverWait(driver, 20).until(EC.visibility_of_all_elements_located((By.CSS_SELECTOR, "p[class='Sub-heading']")))]
```

How do I resolve?

 selenium\_not\_working.png   
186 KB



0 Kudos

Reply



fishjhu

New Contributor II



11-14-2022 12:20 PM



I am getting the error below. @S W have you solved yours?

```
import time as t
from datetime import datetime
from selenium import webdriver
from selenium.webdriver.chrome.service import Service

s = Service('/tmp/chrome/latest/chromedriver_linux64/chromedriver')

options = webdriver.ChromeOptions()
options.binary_location = "/tmp/chrome/latest/chrome-linux/chrome"
options.add_argument('headless')
options.add_argument('--disable-infobars')
options.add_argument('--disable-dev-shm-usage')
options.add_argument('--no-sandbox')
options.add_argument('--remote-debugging-port=9222')
options.add_argument('--homedir=/tmp/chrome/chrome-user-data-dir')
options.add_argument('--user-data-dir=/tmp/chrome/chrome-user-data-dir')
prefs = {"download.default_directory":"/tmp/chrome/chrome-user-data-dir",
        "download.prompt_for_download":False
}
options.add_experimental_option("prefs", prefs)

driver = webdriver.Chrome(service=s, options=options)

# chrome_options.add_argument('disable-notifications')
# chrome_options.add_argument("user-agent=UA")
# chrome_options.add_argument("--start-maximized")
# chrome_options.add_argument('window-size=2160x3840')
```

WebDriverException: Message: Service /tmp/chrome/latest/chromedriver\_linux64/chromedriver unexpectedly exited. Status code was: 127  
mand took 0.90 seconds -- by fisseha.berhane@flywheel.digital.com at 11/14/2022, 3:19:22 PM on webscreenshot

 Capture.png   
104 KB




0 Kudos

Reply



**fishjhu**

New Contributor II

 In response to **fishjhu**



11-14-2022 06:06 PM

Gray's script from the link below worked for me.

<https://community.databricks.com/s/question/OD58Y00009PIBaaSAF/errors-using-seleniumchromedriver-in-...>



0 Kudos

Reply

**swrd**

New Contributor III

↗ In response to fishjhu



11-18-2022 04:51 AM

@Fisseha Berhane Thanks, this worked for me!

However I can't get the browser to open – that would be vital so I can extract the relevant web elements for the automation script to work:

```
Microsoft Azure databricks Search CTRL + P adbresdiarydev stephen.williams@resdiary.com

Selenium Setup Python
File Edit View Run Help Last edit was 15 minutes ago Give feedback

libglb2.0-dev set to manually installed.
libglb2.0-dev-bin is already the newest version (2.64.6-1-ubuntu20.04.4).
libglb2.0-dev-bin set to manually installed.
libnss3 is already the newest version (2:3.49.1-1ubuntu1.8).
libnss3 set to manually installed.
The following additional packages will be installed:
binfmt-support bubblewrap ca-certificates mono-cil-common dbus-x11
gconf-service gconf-service-backend gconf2-common libdbus-glib-1-2
libmono-btls-interface4.0-cil libmono-corlib4.5-cil libmono-il2c-gc4.0-cil
libmono-rtls-interface4.0-cil libmono-security4.0-cil
Command took 2.29 seconds -- By stephen.williams@resdiary.com at 18/11/2022, 12:54:42 on Stephen Williams's Cluster

Cmd 11
1 from selenium import webdriver
2 from selenium.webdriver.chrome.service import Service as ChromiumService
3 from webdriver_manager.chrome import ChromeDriverManager
4 from webdriver_manager.core.utils import ChromeType
5
6 chrome_options = webdriver.ChromeOptions()
7 chrome_options.add_argument('--no-sandbox')
8 # chrome_options.add_argument('--headless')
9 chrome_options.add_argument('--disable-dev-shm-usage')
10 chrome_options.add_argument('--disable-dev-shm-usage')
11 chrome_options.add_argument('--disable-infobars')
12 chrome_options.add_argument('--remote-debugging-port=9222')
13 chrome_options.add_argument('--start-maximized')
14 chrome_options.add_experimental_option('excludeSwitches', ['enable-logging'])
15
16 driver = webdriver.Chrome(service=ChromiumService(ChromeDriverManager(chrome_type=ChromeType.CHROMIUM).install()), options=chrome_options)

Command took 0.67 seconds -- By stephen.williams@resdiary.com at 18/11/2022, 13:15:05 on Stephen Williams's Cluster

Cmd 12
1 driver.get("https://www.google.com")

Command took 0.28 seconds -- By stephen.williams@resdiary.com at 18/11/2022, 13:15:07 on Stephen Williams's Cluster

Cmd 13
1 driver.quit()

Command took 0.12 seconds -- By stephen.williams@resdiary.com at 18/11/2022, 13:01:16 on Stephen Williams's Cluster

Shift+Enter to run
```

Any ideas on how to get that done?

selenium\_not\_working\_v.3.0.png   
153 KB



0 Kudos

Reply

## Welcome to Databricks Community: Lets learn, network and celebrate together

Join our fast-growing data practitioner and expert community of 80K+ members, ready to discover, help and collaborate together while making meaningful connections.

Click **here to register** and join today!

Engage in exciting **technical discussions**, **join a group** with your peers and meet our Featured Members.

## Related Content

### Metastore creation – Azure Databricks – Internal Server Error

in **Data Governance** yesterday

---

### Hive metastore table access control End of Support

in **Data Engineering** yesterday

---

### Can't query delta tables, token missing required scope

in **Data Engineering** Saturday

---

### Ubuntu 22 ODBC Connectivity Issue with PHP – SQL error: [unixODBC][Driver Manager]Can't open lib

in **Administration & Architecture** Saturday

---

### Issue during testing SparkSession.sql() with pytest.

in **Data Engineering** Saturday

**Product**



**Learn & Support**



**Solutions**



**Company**



**Resources**

Databricks Inc.

160 Spear Street, 13th


Floor

San Francisco, CA 94105

1-866-330-0121

© Databricks 2023. All rights reserved. Apache, Apache Spark, Spark and the Spark logo are trademarks of the Apache Software Foundation.

---

[Privacy Notice](#) | [Terms of Use](#) | [Your Privacy Choices](#) | [Your California Privacy Rights](#) 

 [Top](#)