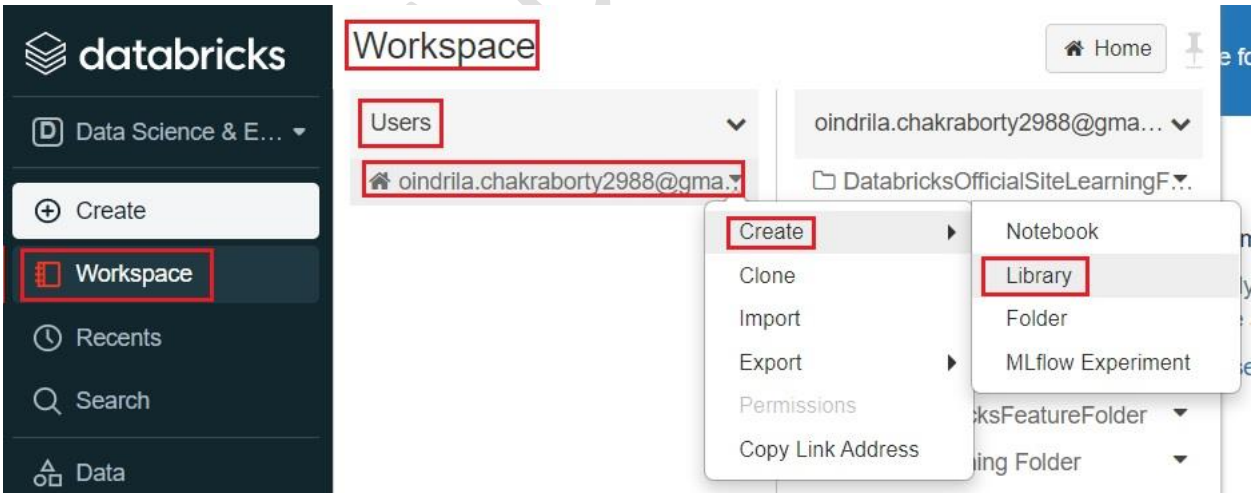
INTRODUCTION TO WORKSPACE LIBRARY

What is Workspace Library

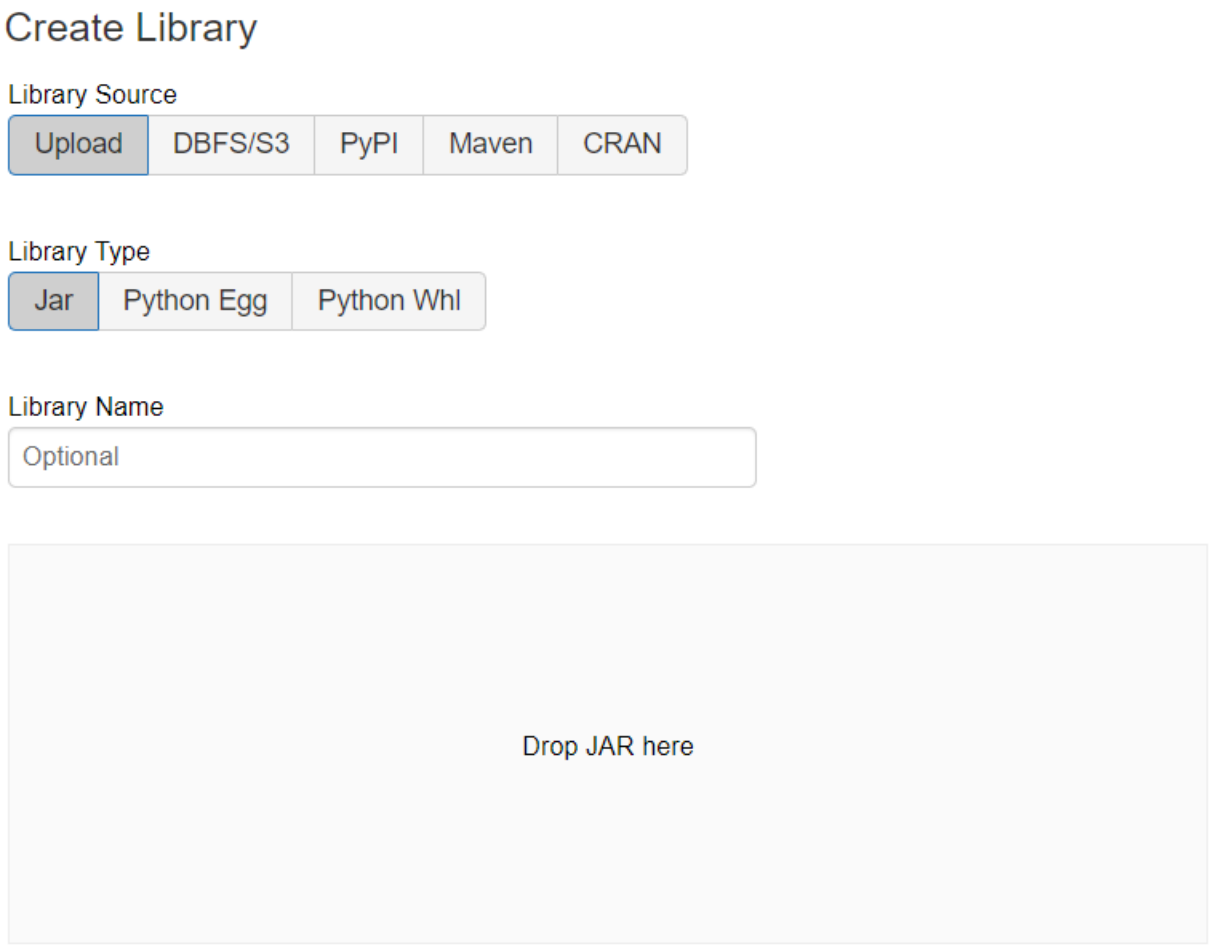
- "Workspace Libraries" serve as a "Local Repository" from which the "Cluster-Installed Libraries" can be created.
- A "Workspace Library" can be created in the following two ways -
  - "Custom-Code Created" by an "Organization".
  - "Version" of an "Open-Source Library" that an "Organization" has "Standardized" on.
- A "Workspace Library" must be "Installed" on a "Cluster" before it can be used in a "Notebook" or a "Job".
- "Workspace Libraries" in the "Shared" Folder are available to all the "Users" in a "Workspace".
- "Workspace Libraries" in a "User" Folder are available only to that "User".

Create a Workspace Library

- Perform a "Right Click" on the "Workspace Folder", e.g., "Shared", where to "Store" the "Library".
- First, select "Create". Then select "Library".



- The "Create Library" Dialog Box appears.



Select the desired "Library Source" option and follow the appropriate procedure.

- ***Upload a Library*** -
  - In this option, select "Upload" from the "Library Source" button list.
  - From the "Library Type" button list, select one of the following -
    - ✓ JAR
    - ✓ ***Python Egg*** - A "Python Egg" is a "Logical Structure" embodying the "Release" of a specific "Version" of a "Python Project", comprising its "Codes", "Resources" and "Metadata".
      There are "Multiple Formats" that can be used to "Physically Encode" a "Python Egg". However, a key principle of "Python Eggs" is that these should be "Discoverable" and "Importable".

The ".egg" File is a "Distribution Format" for "Python Packages". It is a ".zip" File with some "Metadata" Files, renamed as ".egg" for "Distributing Codes" as "Bundles". ".egg" Format in Python is similar to ".jar" Format in Java.

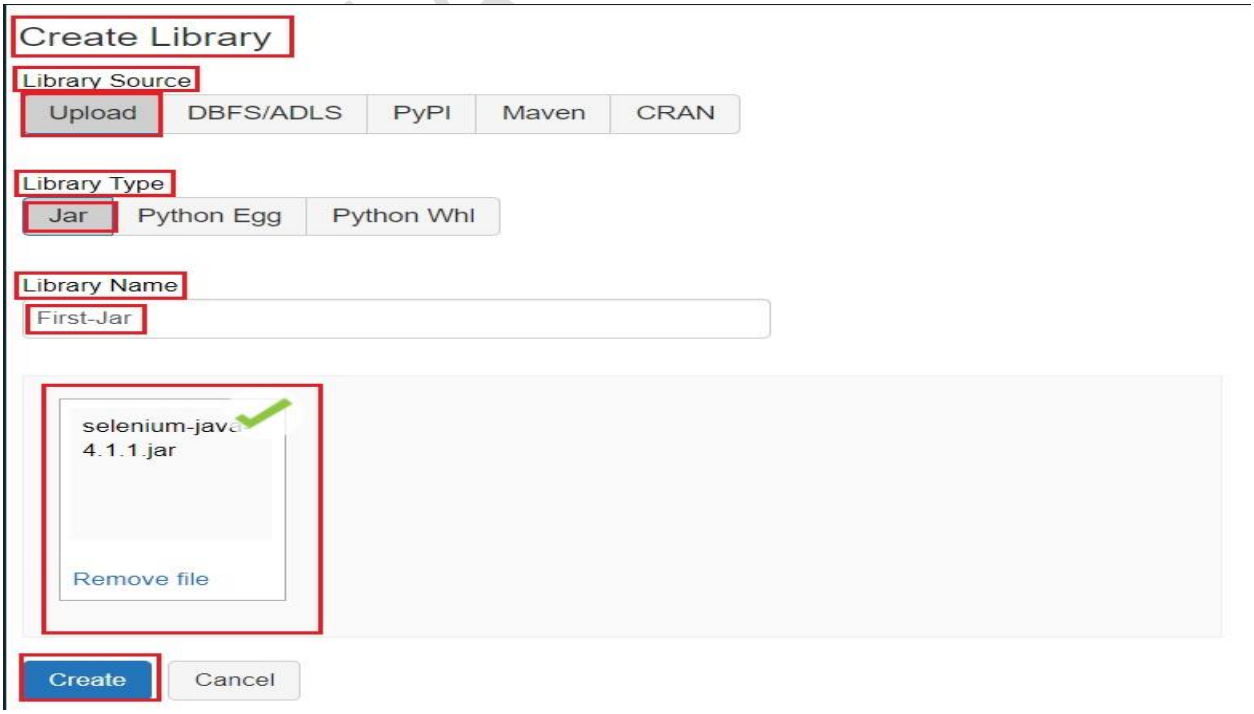
The ".egg" Format is well-suited to "Distribution", since the "Python Project" is essentially "Self-Contained" within a "Single Directory" or "File", unmingled with any other "Python Project's" Code or Resources. This advantage also makes it possible to have "Multiple Versions" of the same "Python Project" "Simultaneously Installed", such that "Individual "Python Programs" can select the desired "Version".

Since, an ".egg" File is essentially a ".zip" File, if the "Extension" is "Changed" from ".egg" to ".zip", it is possible to see the Contents as "Folders" inside the "archive" Folder.

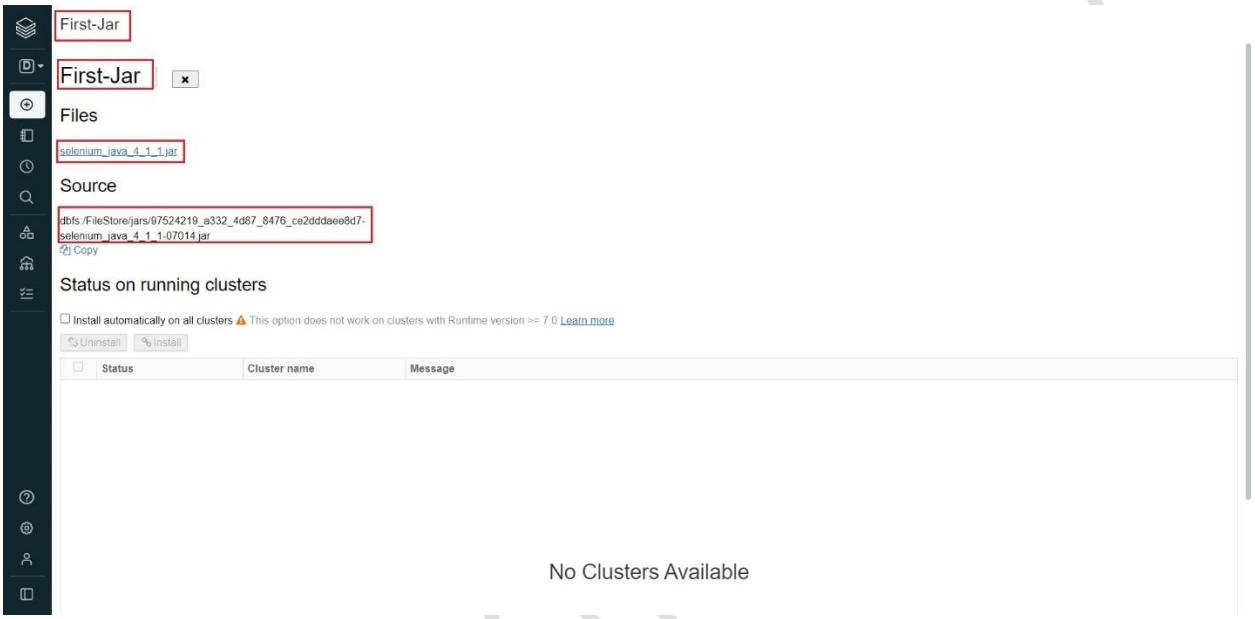
"Egg" Package has been "Suspended" by "Wheel" Package.

- ✓ ***Python Whl*** - "Wheels" are the new standard of "Python Distribution" and are intended to replace "Eggs".

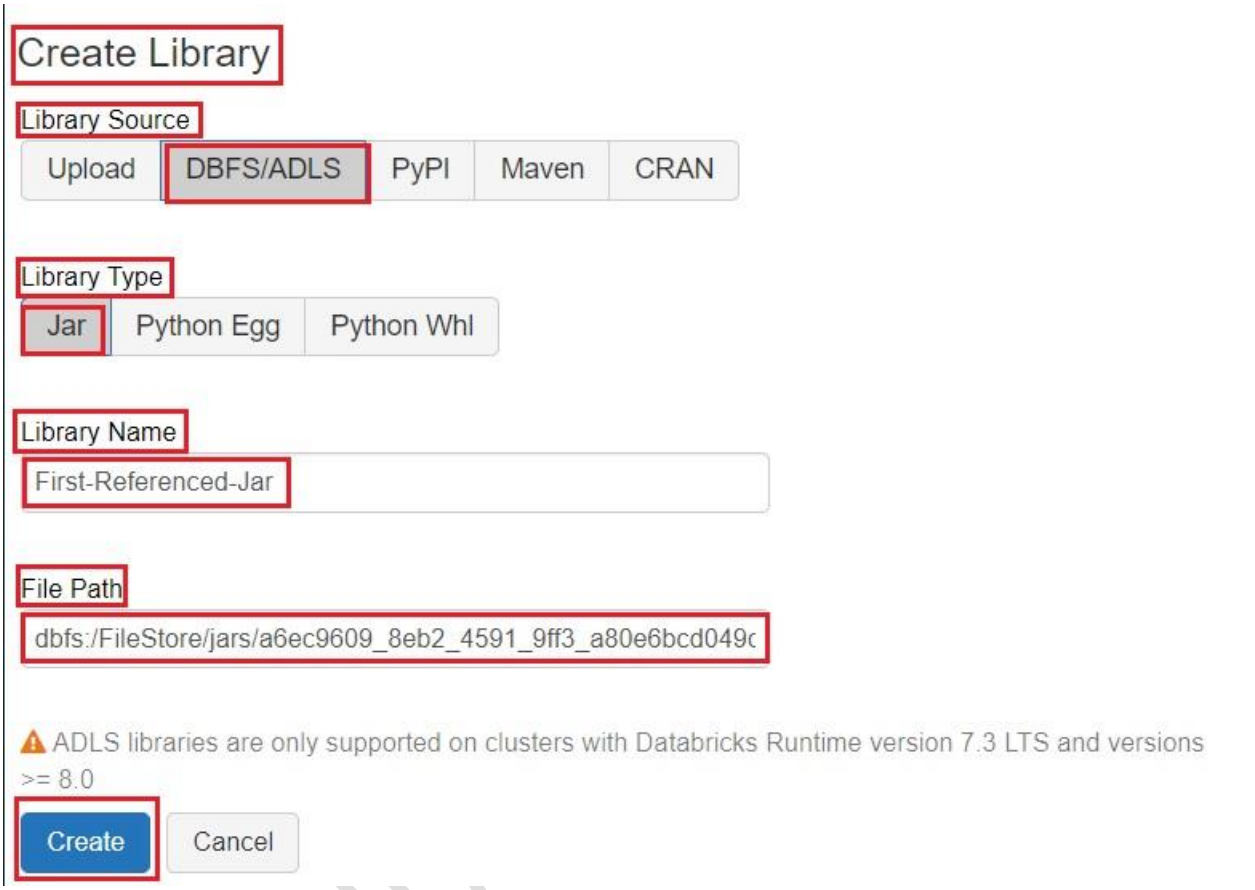
- ➢ Optionally, enter a name for the "Library" in the "Library Name" textbox.
- ➢ "Drag" the "JAR", "EGG" or "WHL" File to the "Drop Box" or "Click" on the "Drop Box" and "Navigate" to the "JAR", "EGG" or "WHL" File. Click on the "Create" button.



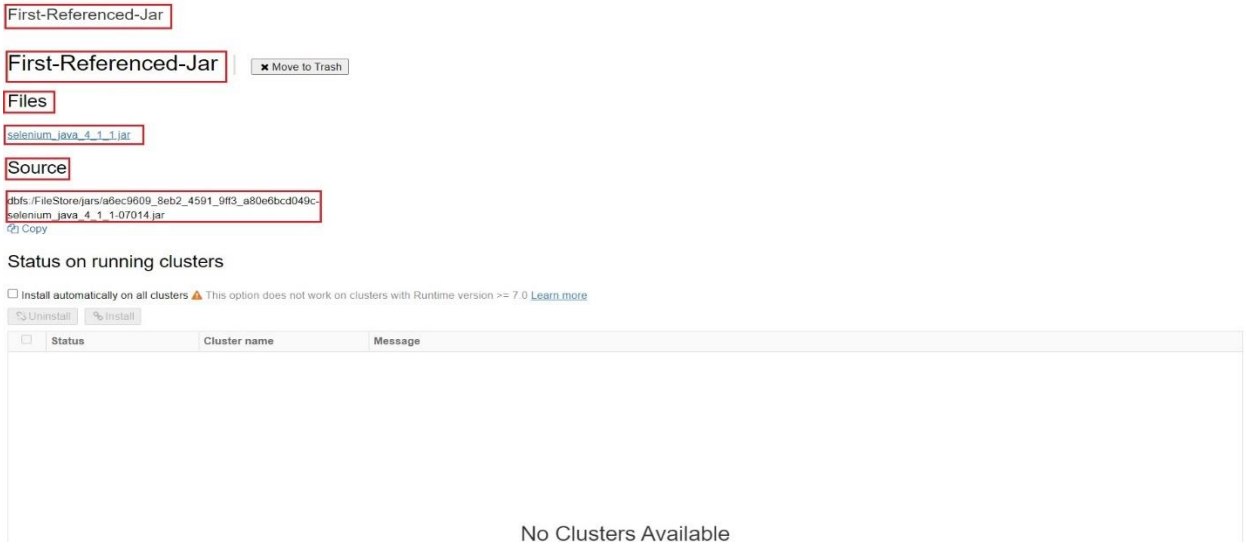
- The “Library Status” page is displayed. It can be seen that the “Package File” is “Uploaded” to the Path “*dbfs:/FileStore/jars*”.



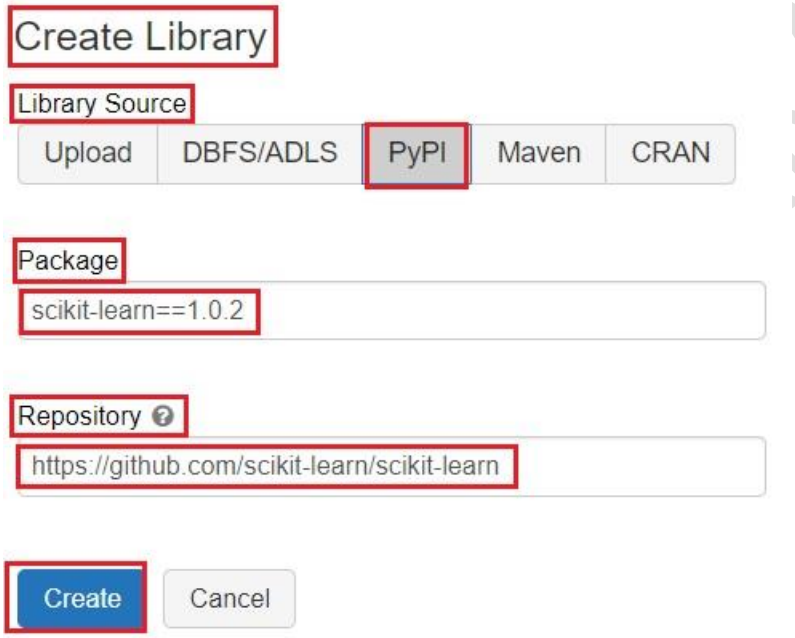
- ***Reference an Uploaded Library*** - If, already any “JAR”, “Python Egg” or “Python Wheel” File is “Uploaded” to the “Object Storage”, that File can be “Referenced” in a “Workspace Library”. It is possible to choose a “Library” in “DBFS” or one “Stored” in “Amazon S3”.
  - In this option, select “DBFS/S3” from the “Library Source” button list.
  - From the “Library Type” button list, select one of the following -
    - ✓ JAR
    - ✓ Python Egg
    - ✓ Python Whl
  - Optionally, enter a name for the “Library” in the “Library Name” textbox.
  - Specify the “DBFS” or “S3” Path where the “Library” is “stored” in the “File Path” textbox. Click on the “Create” button.



- The "Library Status" page is displayed. It can be seen that the respective File is "Uploaded" to the Path "dbfs:/FileStore/jars".

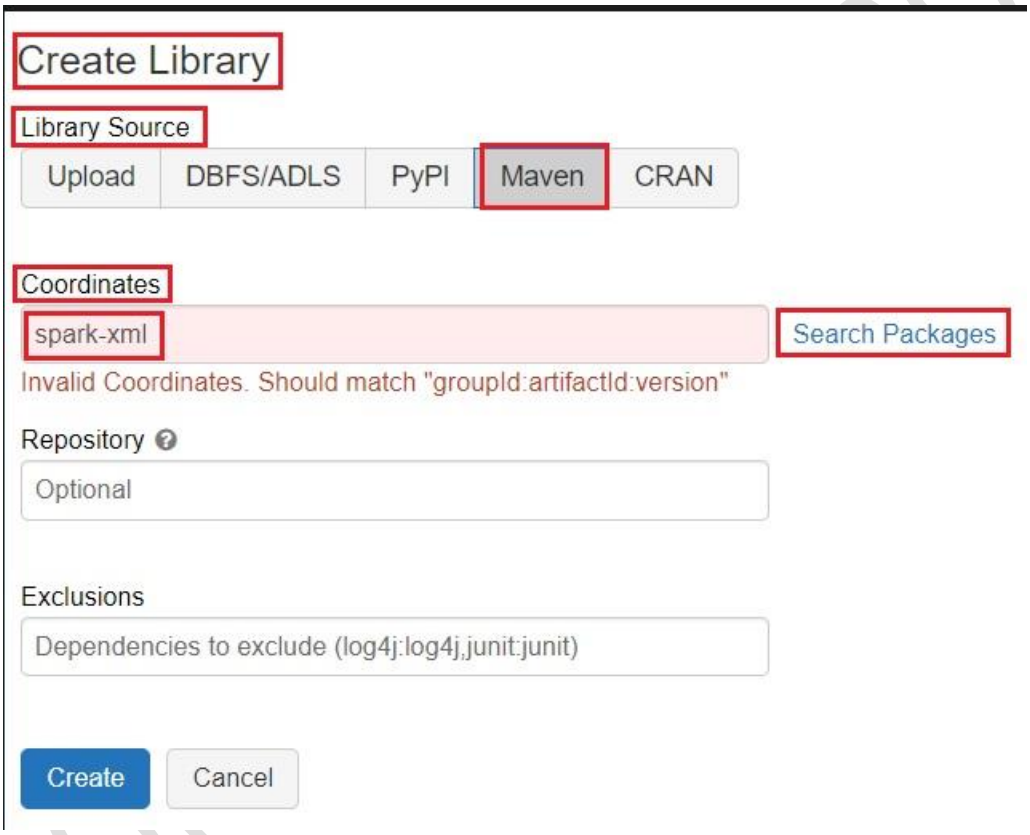


- ***PyPI Package*** -
  - In this option, select "PyPI" from the "Library Source" button list.
  - Enter the "Name" of the "PyPI Package" to "Install" in the "Package" textbox. To "Install" a specific "Version" of a "Library", use the following "Format" for the "Library" –
    "*<library>==<version>*".
    Example - "*scikit-learn==1.0.2*".
  - In the "Repository Field" textbox, optionally enter a "PyPI Repository URL". Click on the "Create" button.

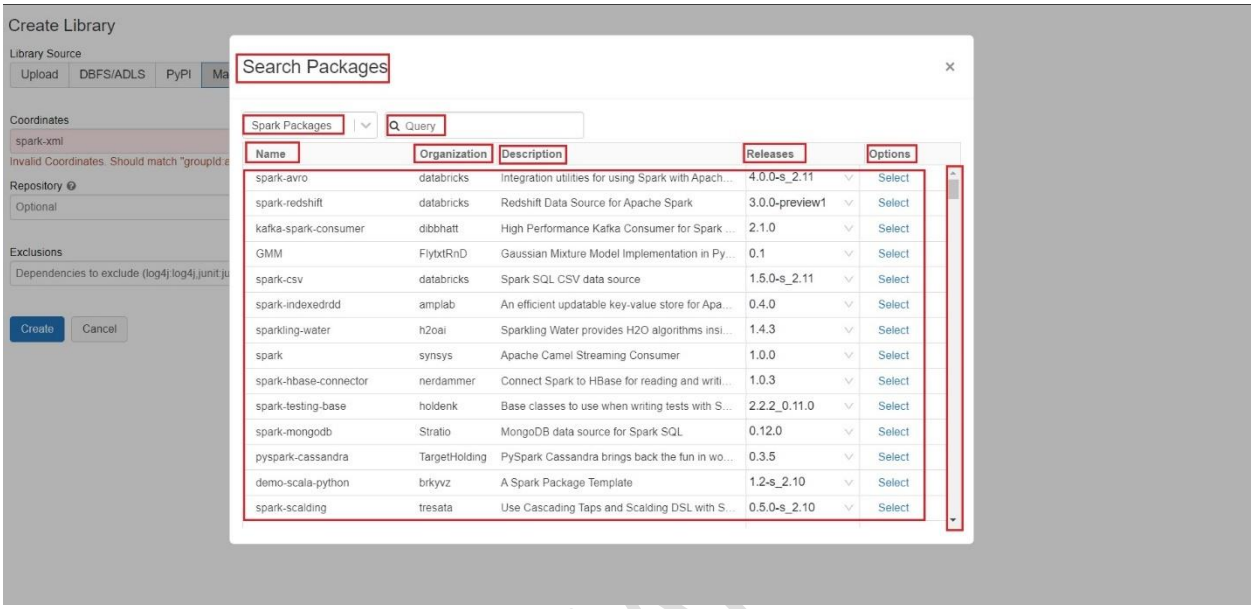


  - The "Library Status" page is displayed.

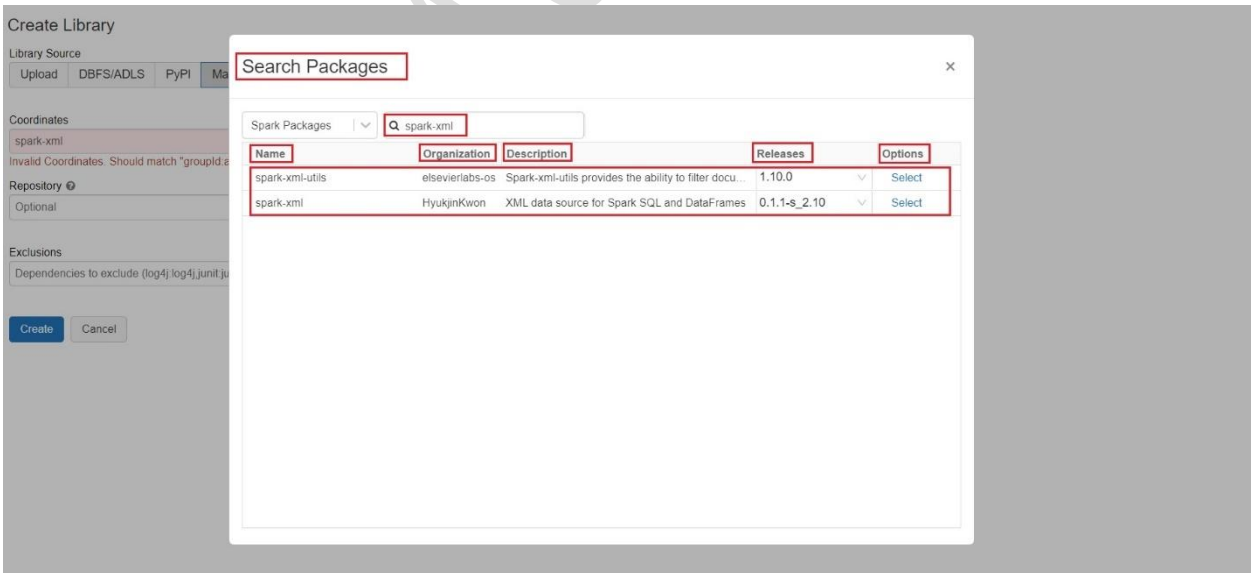
- ***Maven Package*** -
  - In this option, select "Maven" from the "Library Source" button list.
  - In the "Coordinates" textbox, enter the "Maven Coordinate" of the "Maven Library" to "Install". "Maven Coordinates" are in the form "*groupId:artifactId:version*". Example - "*com.databricks:spark-avro_2.10:1.0.0*".
    If the exact "Coordinate" is not known, enter the name of the "Maven Library" to "Install" in the "Coordinates" textbox and click on the "Search packages" link.



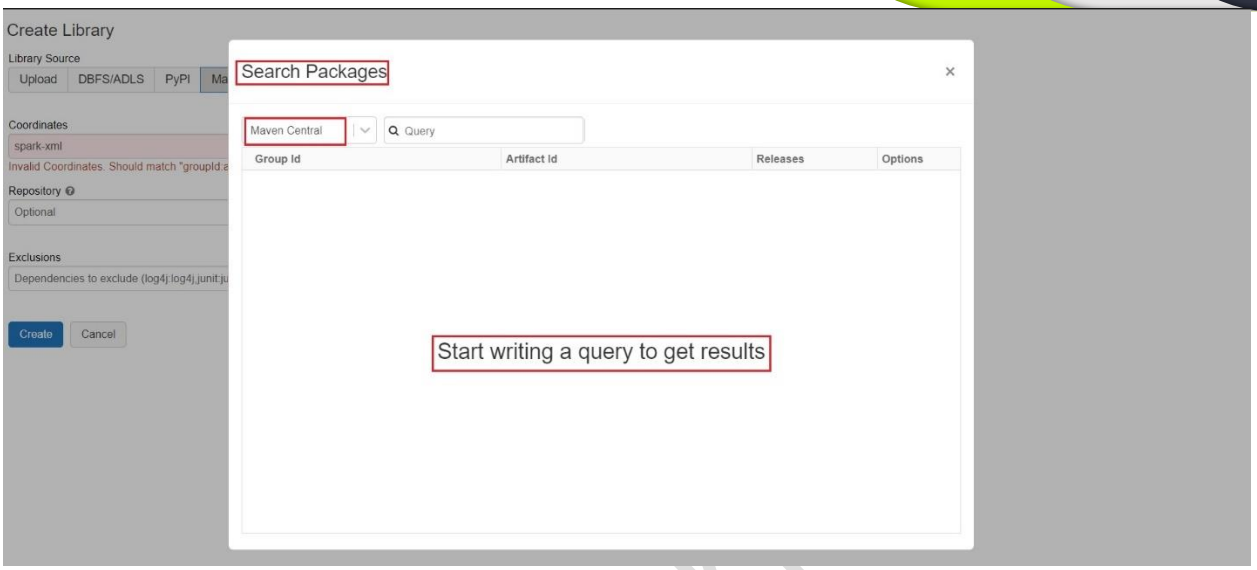
In the "Search Packages" Dialog Box, a list of Matching "Spark Packages" will be displayed.



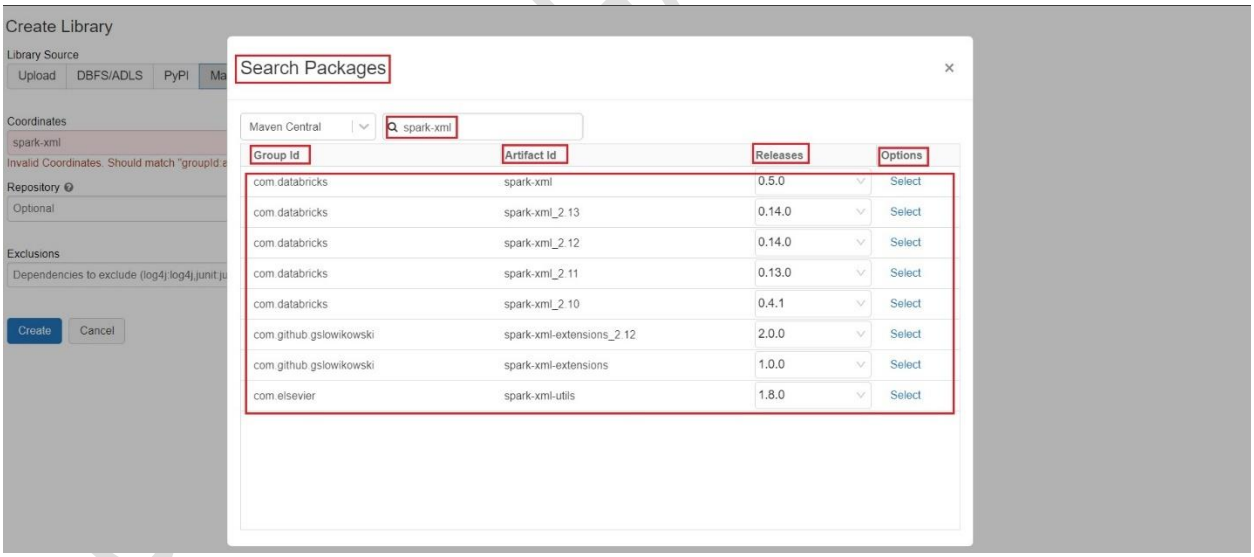
It is possible to “Filter” the results of displayed “Spark Packages” by writing a query in the “Search” textbox. The results “Refresh Automatically”.
Example - enter “*spark-xml*” in the “Search” textbox and immediately only the “Libraries” with the mentioned words in the respective names appear.



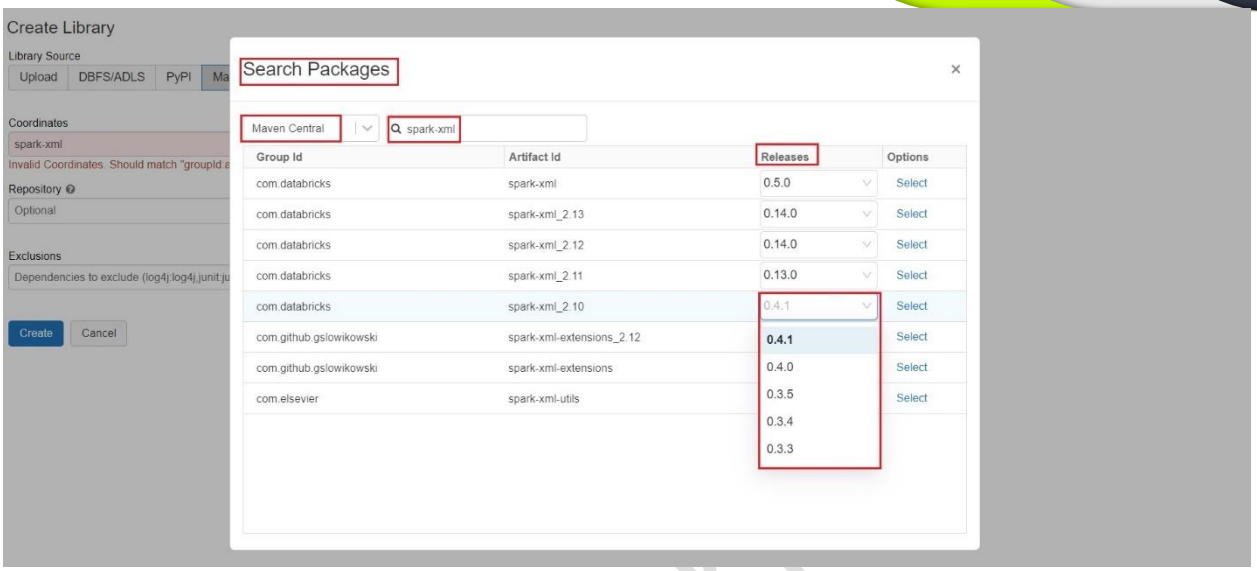
Select the option “Maven Central” in the “Drop Down List” at the top left of the “Search Packages” Dialog Box. It can be seen from the below image that to display any “Maven Library” as result, a query needs to be written in the “Search” textbox.



Enter "*spark-xml*" in the "Search" textbox and immediately only the "Libraries" with the mentioned words in the respective names appear.

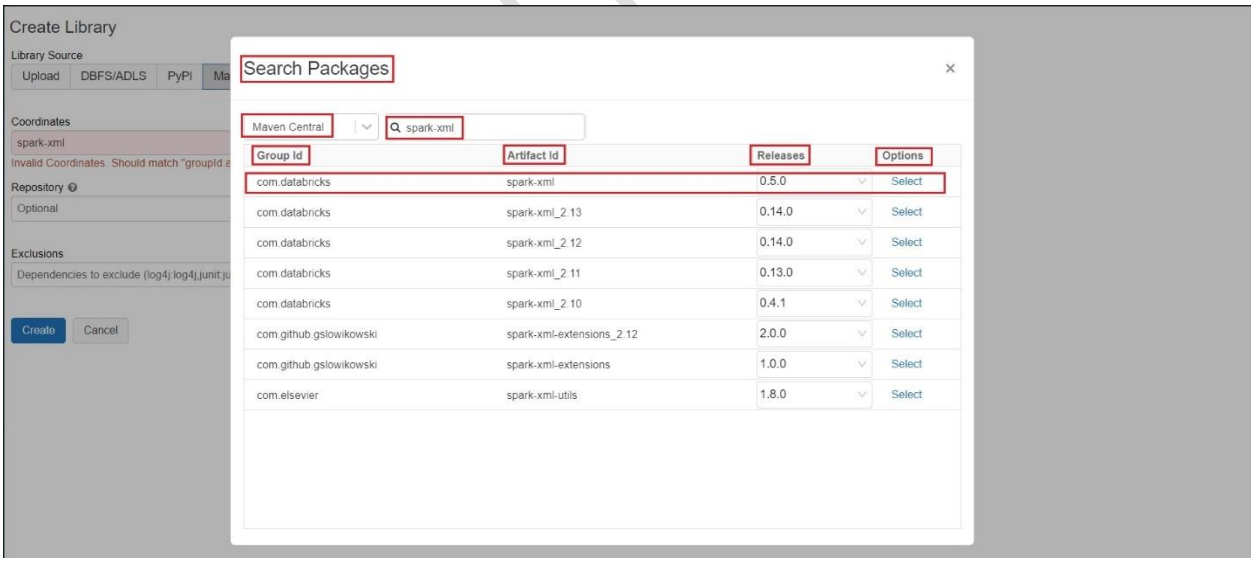


It is possible to optionally select the different "Version" of the "Maven Library" in the "Release" column.

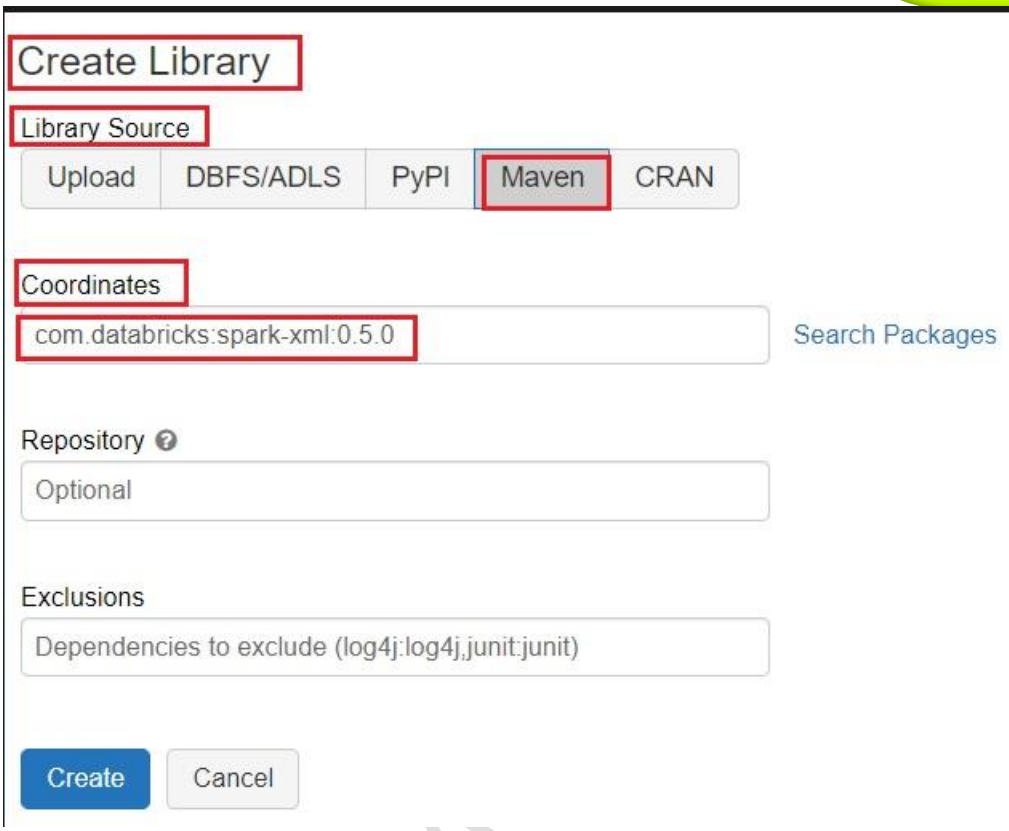


Click on the “Select” link of the desired “Maven Library”, with desired “Release Version”, in the column “Options”.

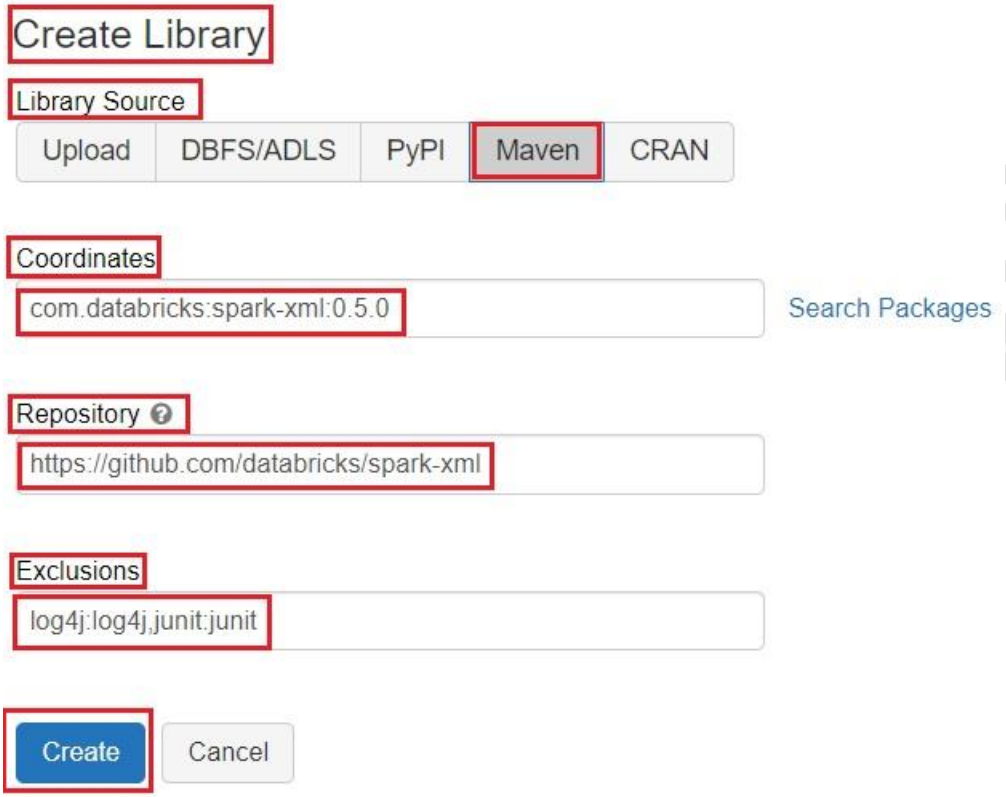
Example - select the “Maven Library” with “*com.databricks*” as “GroupId”, “*spark-xml*” as “ArtifactId”, “*0.5.0*” as “Release Version”.



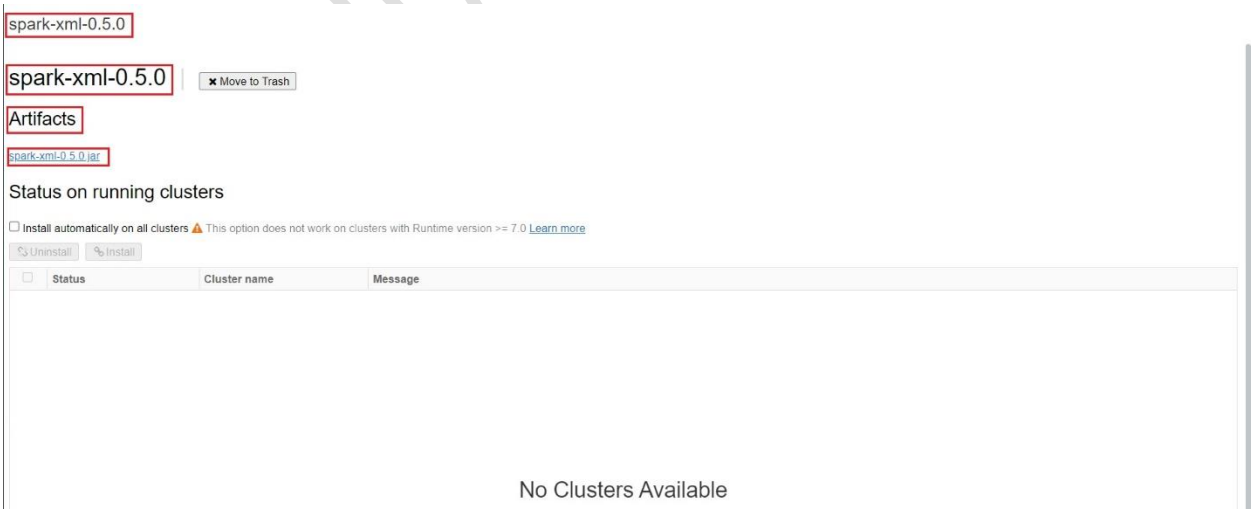
The “Coordinate” textbox is filled with the selected “Maven Library” in the format “*groupId:artifactId:version*”, i.e., “*com.databricks:spark-xml:0.5.0*”.



- In the "Repository" textbox, optionally enter a "Maven Repository URL". "Internal Maven Repositories" are "Not Supported".
- In the "Exclusions" textbox, optionally enter the "groupId" and "artifactId", separated by a colon ":" of the dependent "Maven Library" to "Exclude". It is possible to provide Multiple "Maven Libraries" to "Exclude". Example - "*log4j:log4j,junit:junit*". Finally, click on the "Create" button.



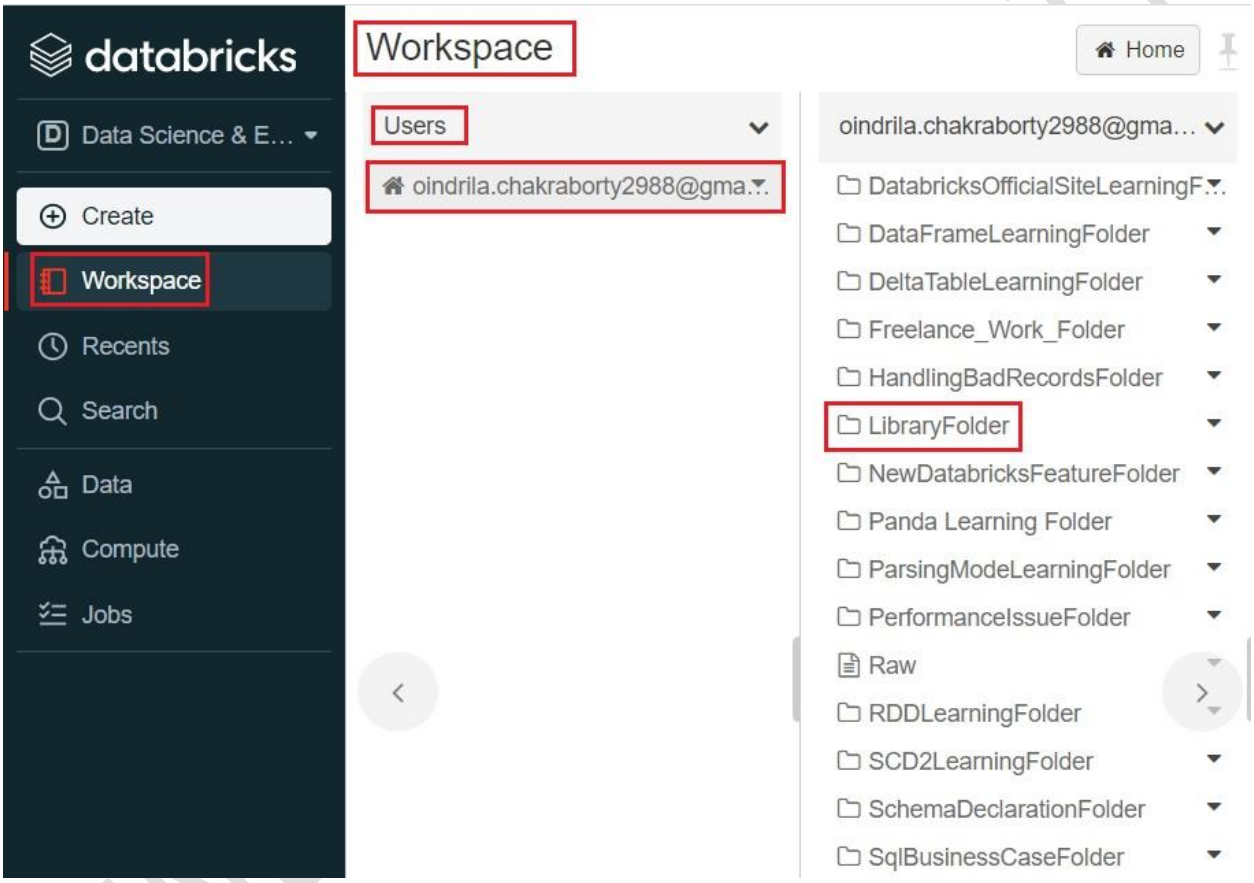
- The "Library Status" page is displayed.



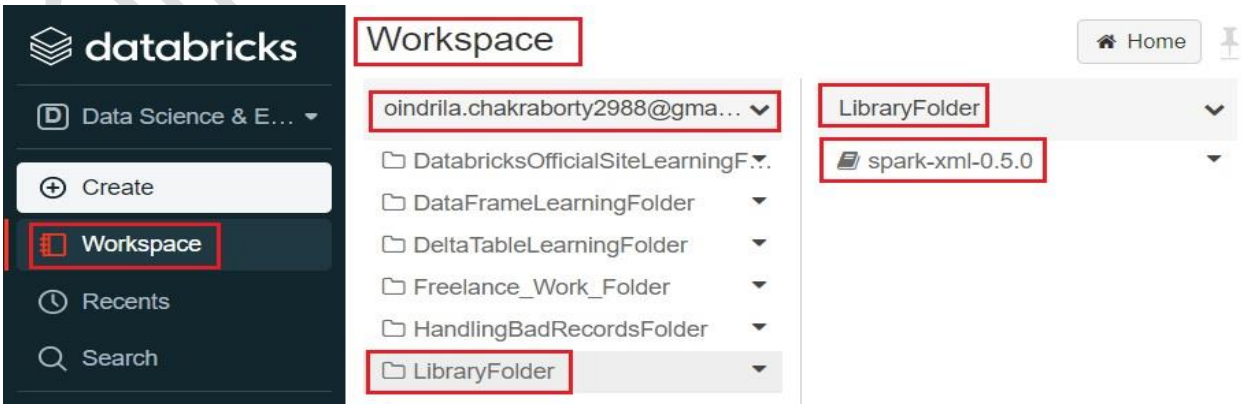
- ***CRAN Package***

View Installed Workspace Library Details

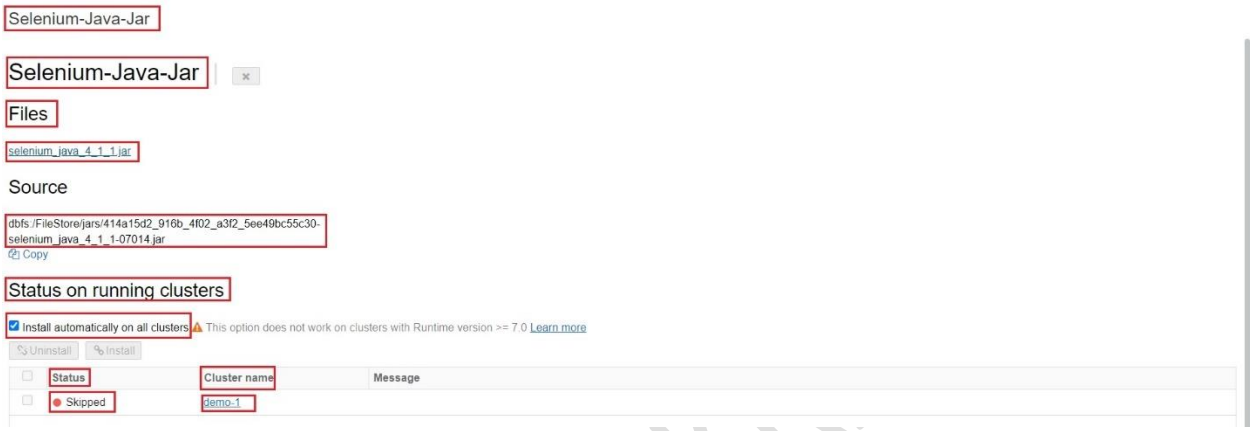
- To view the details of the Installed “Workspace Library”, perform the following steps –
  - Go to the “Workspace Folder” containing the “Library”.



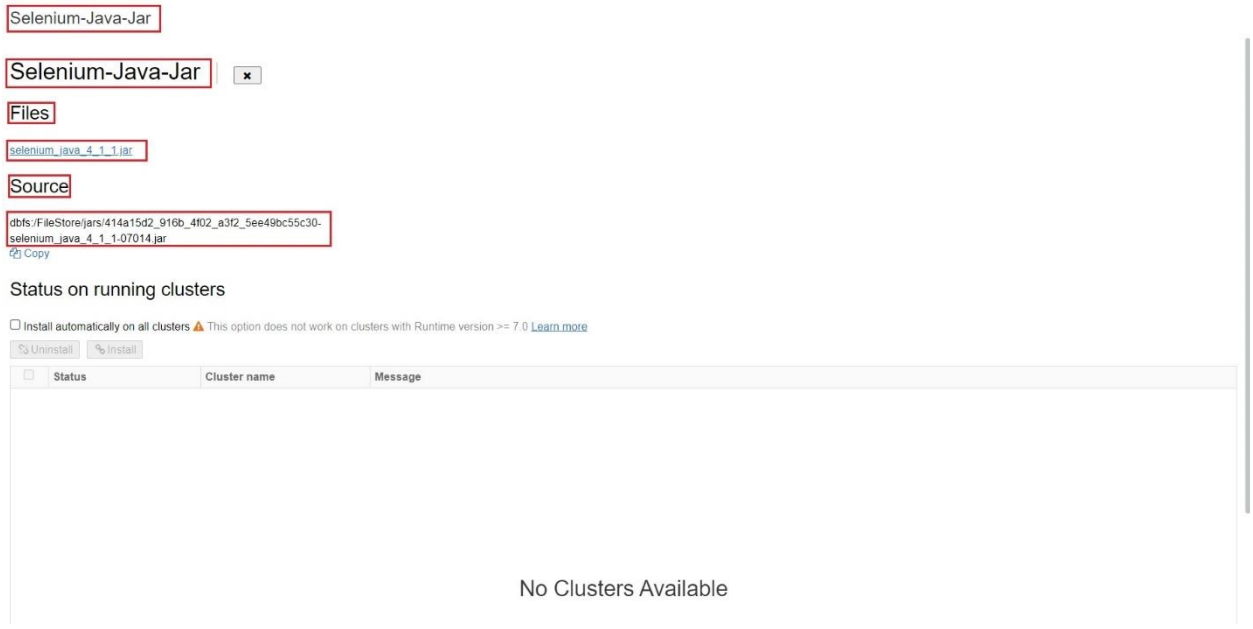
  - Click the name of the “Library”.



- The “Library Status” page is displayed, which shows the “Running Clusters” and the “Install Status” of the Installed “Library”.



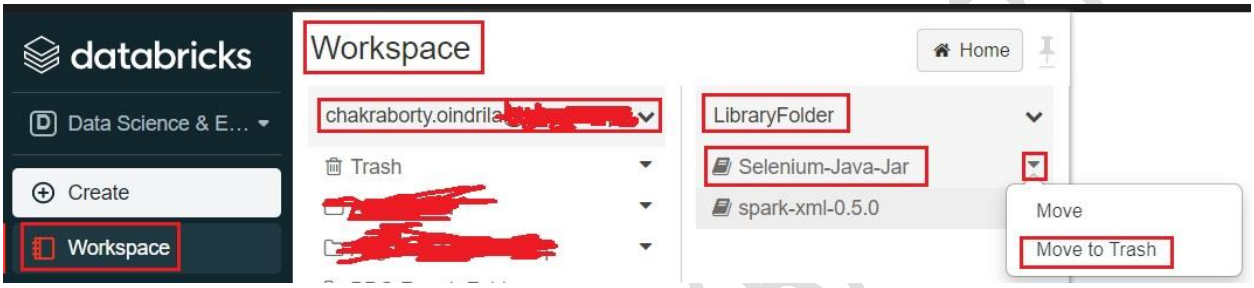
- If the “Library” was “Uploaded”, the “Library Status” page contains a “Path” under the “Source” section that shows the “Path” of the “Uploaded Package File in DBFS”. This path is “*dbfs:/FileStore/jars/<jar-file>*”.



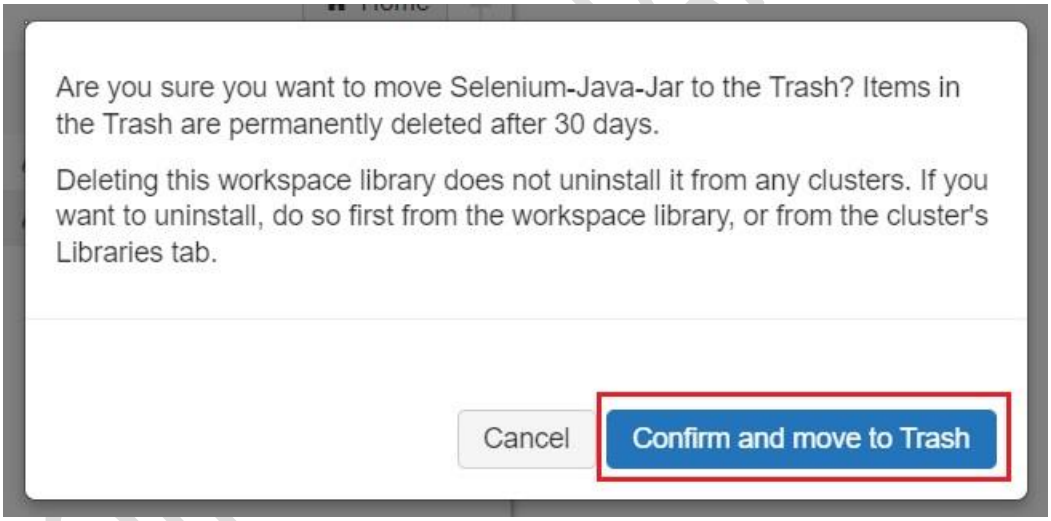
Delete a Workspace Library

- Before deleting a “Workspace Library”, it needs to be made sure that the “Workspace Library” is “Uninstalled” from all the “Clusters”.
- To delete a “Workspace Library”, perform the following steps -

- ***Move the "Workspace Library" to the "Trash" Folder*** –
  - ✓ Go to the "Workspace Folder", containing the "Workspace Library".
  - ✓ Click on the "Drop Down Arrow" to the "Right" of the name of the "Workspace Library".
  - ✓ Select "Move to Trash".

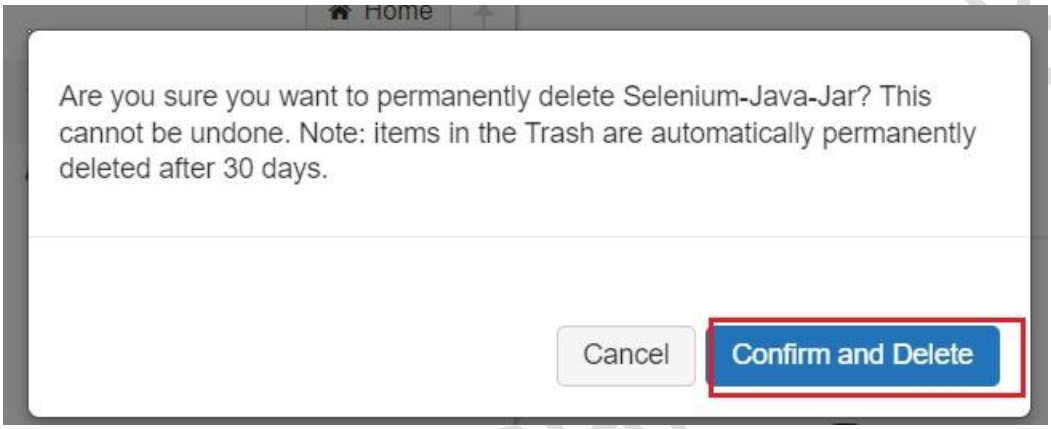


  - ✓ A Dialog Box is displayed. Click on the "Confirm and move to Trash" button.



- ***Delete the "Workspace Library" in the "Trash" Folder*** -
  - ✓ Go to the "Trash" Folder.
  - ✓ Click on the "Drop Down Arrow" to the "Right" of the name of the deleted "Workspace Library".
  - ✓ Select "Delete Immediately".

✓ A Dialog Box is displayed. Click on the “Confirm and Delete” button.



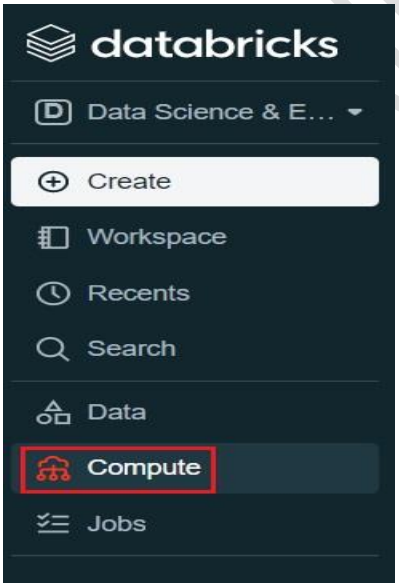
INTRODUCTION TO CLUSTER LIBRARY

What is Cluster Library

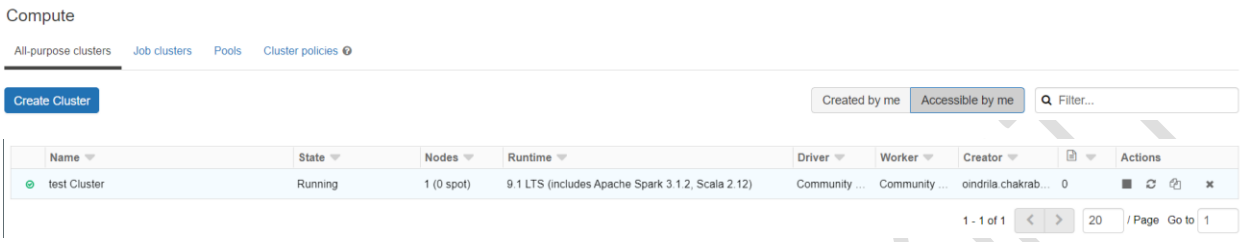
- "Cluster Libraries" can be used by all "Notebooks" running on the "Cluster" on which the "Libraries" are "Installed". "Cluster Libraries" can be "Installed" directly from a "Public Repository", such as - "PyPI", or, "Maven", using a previously installed "Workspace Library" or using an "INIT Script".
- When a "Library" is "Installed" on a "Cluster", a "Notebook" that is already attached to that "Cluster" will Not Immediately see the "New Library". The "Notebook" must be first "De-Attached" and then "Re-Attached" to the "Cluster".
- There are three ways to "Install" a "Library" on a "Cluster" -
  - Install a "Workspace Library" that has been already "Uploaded" to the "Workspace".
  - Install a "Library" for Use with a Specific "Cluster" only.
  - Using "INIT" Script at the time of creating the "Cluster".

Create a Cluster Library Using Already Installed Workspace Library

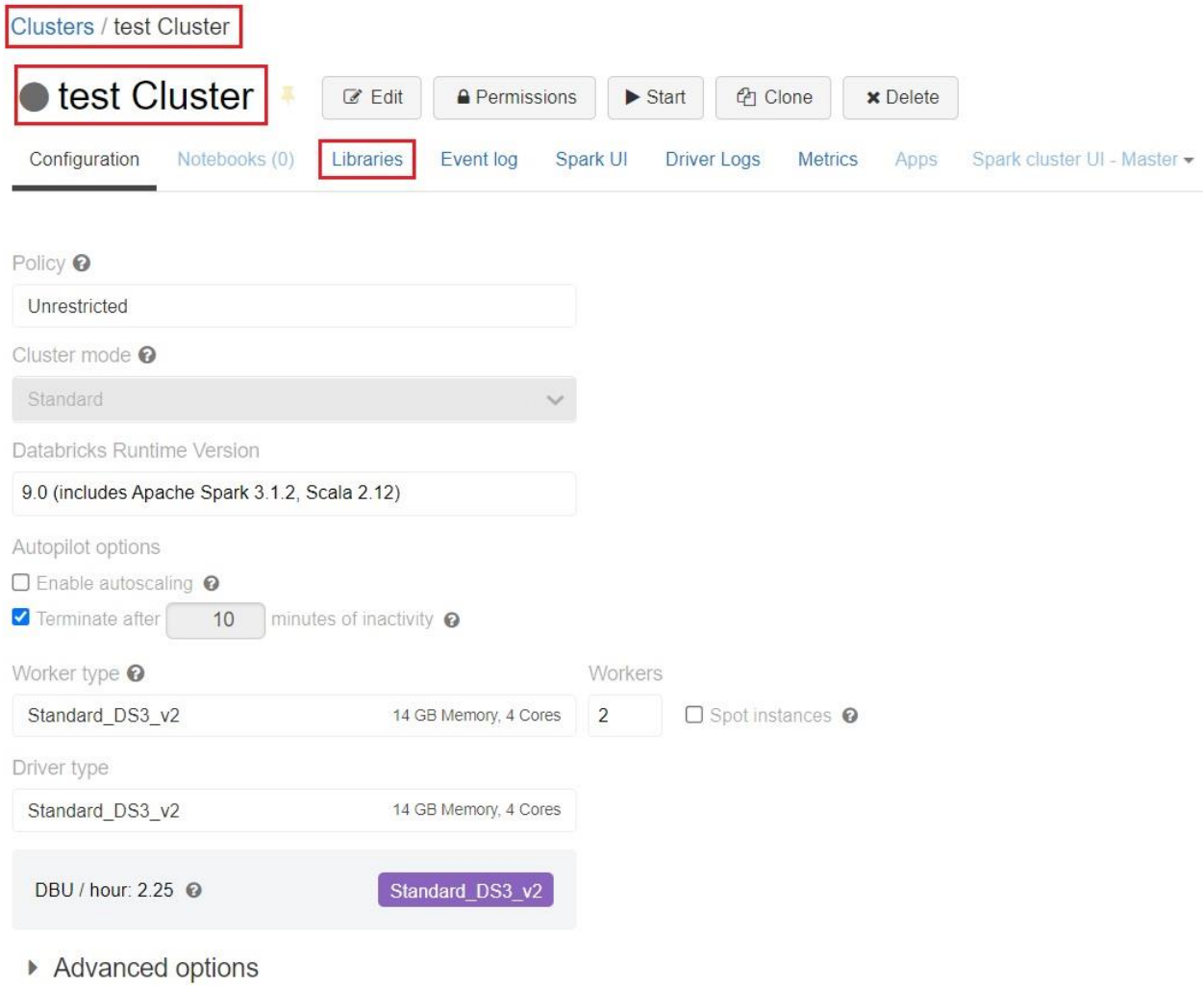
- To "Install" a "Library", that already exists in the "Workspace", in a "Cluster", either start from the "Cluster UI" or from the "Library UI".
- ***Starting from "Cluster UI"*** -
  - Click on the "Compute" icon on the Sidebar.



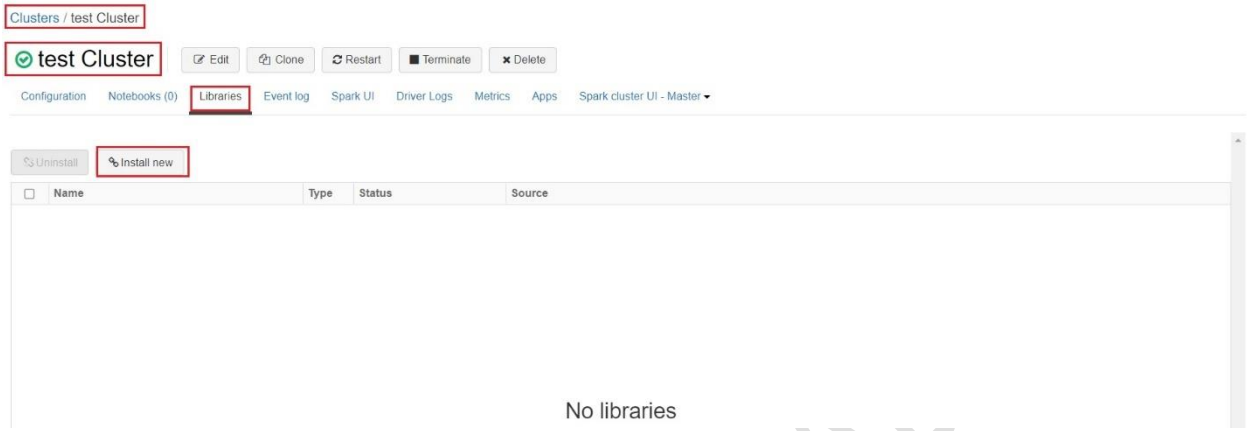
- Click on the name of any “Cluster” displayed.



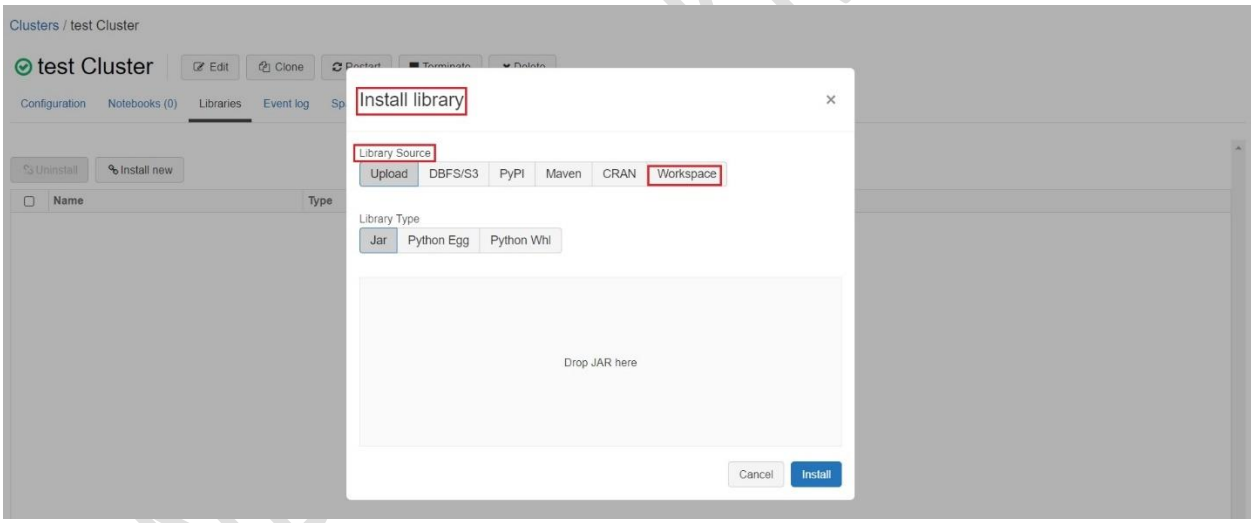
- The details page of the respective “Cluster” opens. Click on “Libraries” tab.



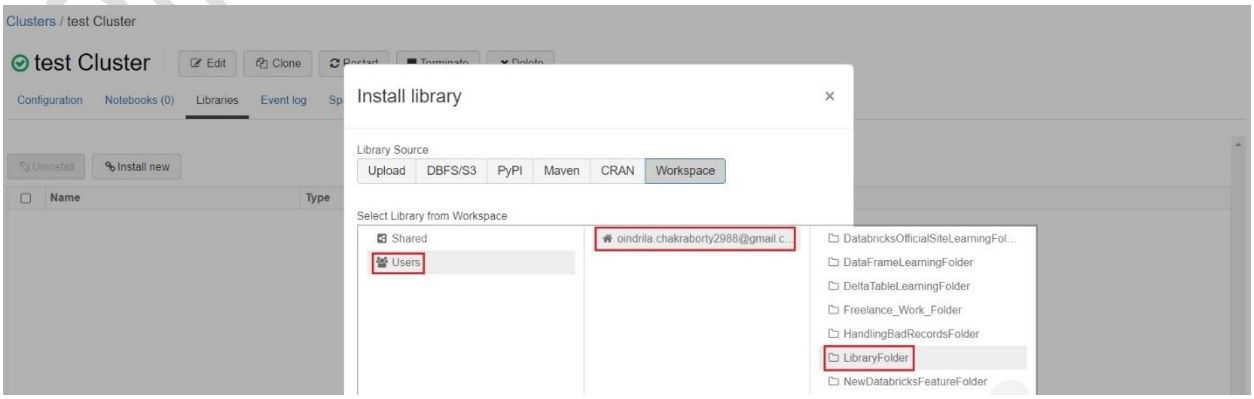
- Click on the “Install new” button. This button will be enabled only if the “Cluster” is in “Running” state.



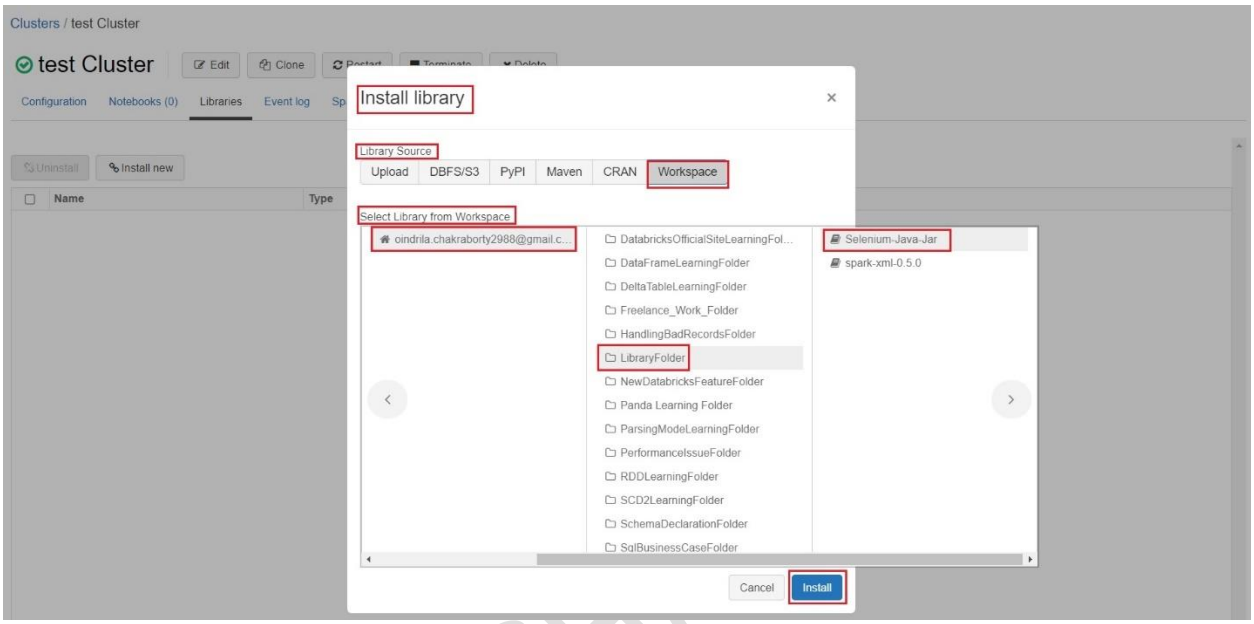
- The “Install Library” Dialog Box opens. From the options of “Library Source” button list, select “Workspace”.



- Select the desired “User”/”Shared” “Workspace Folder”.



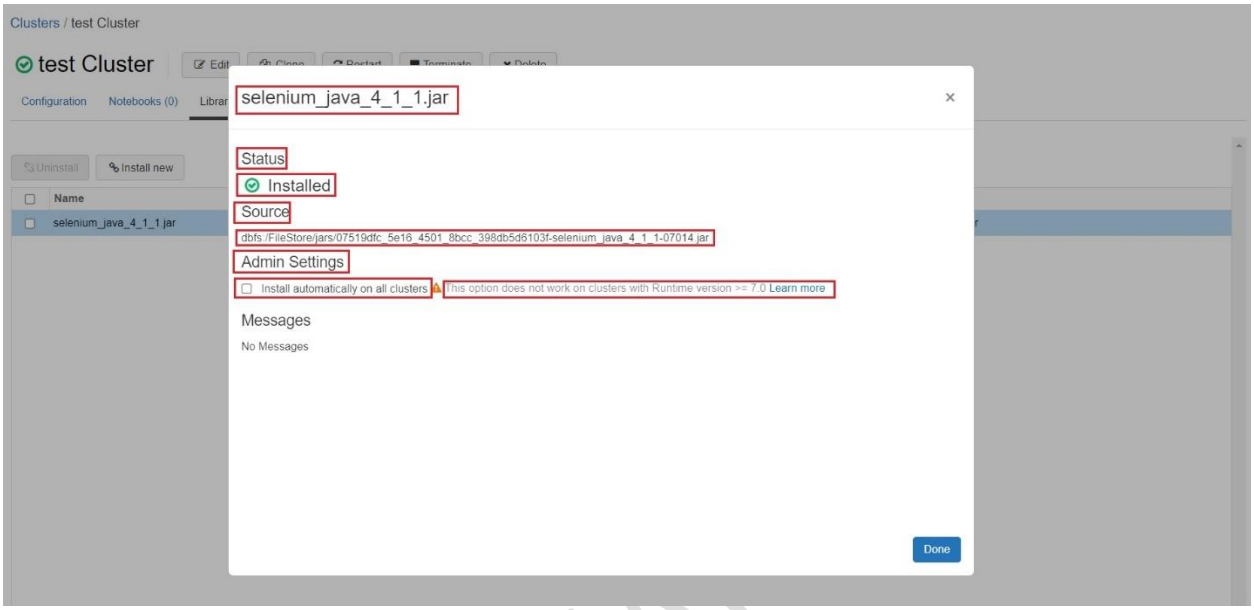
- "Navigate" to the desired "Workspace Library". Click on the "Install" button.



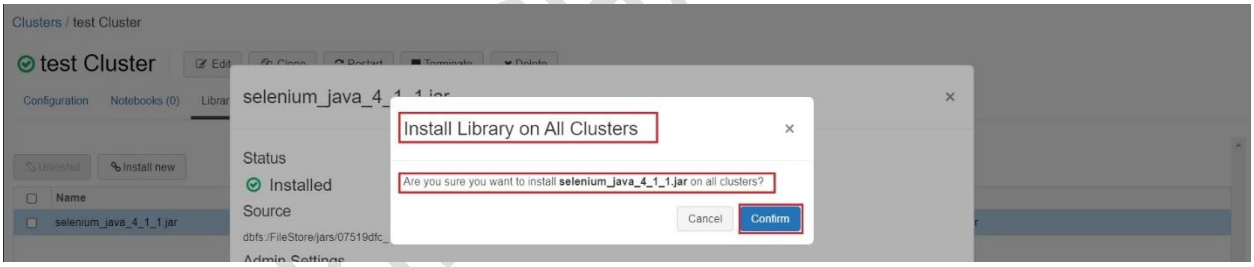
- In the details page of the "Cluster", it can be seen that the selected "Library" is "Installed" for that "Cluster".

- To "Configure" the "Installed" "Library" on all the "Clusters", perform the following steps -
  - ✓ Click on the name of the "Library".

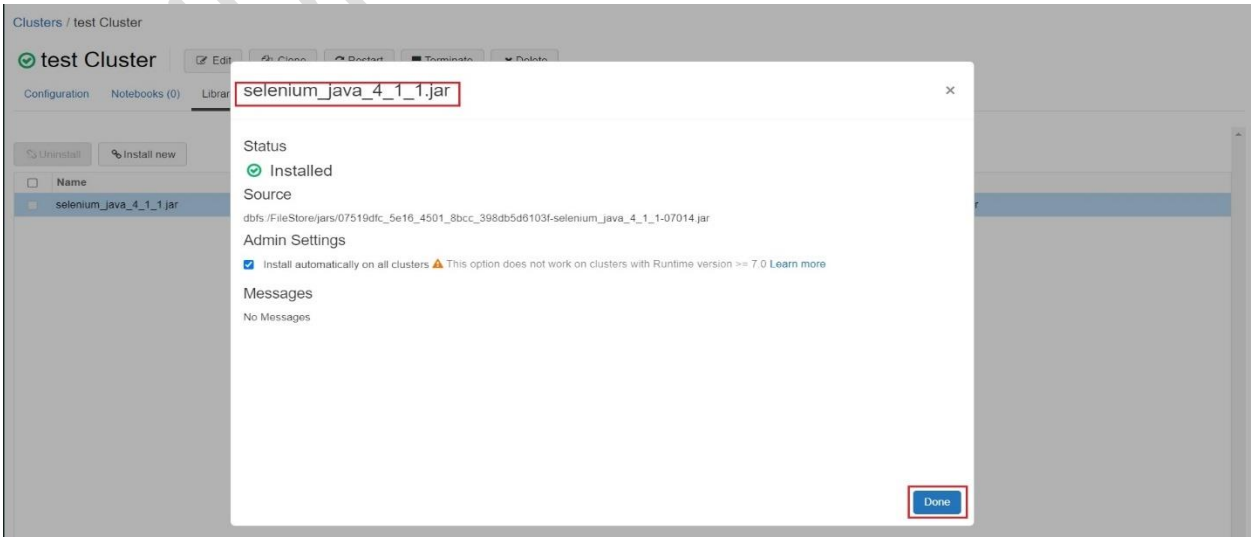
  - ✓ A Dialog Box with the name of the "Library" opens. Select the "Install automatically on all clusters" Checkbox.



✓ A Confirmation Dialog Box opens. Click on the “Confirm” button.

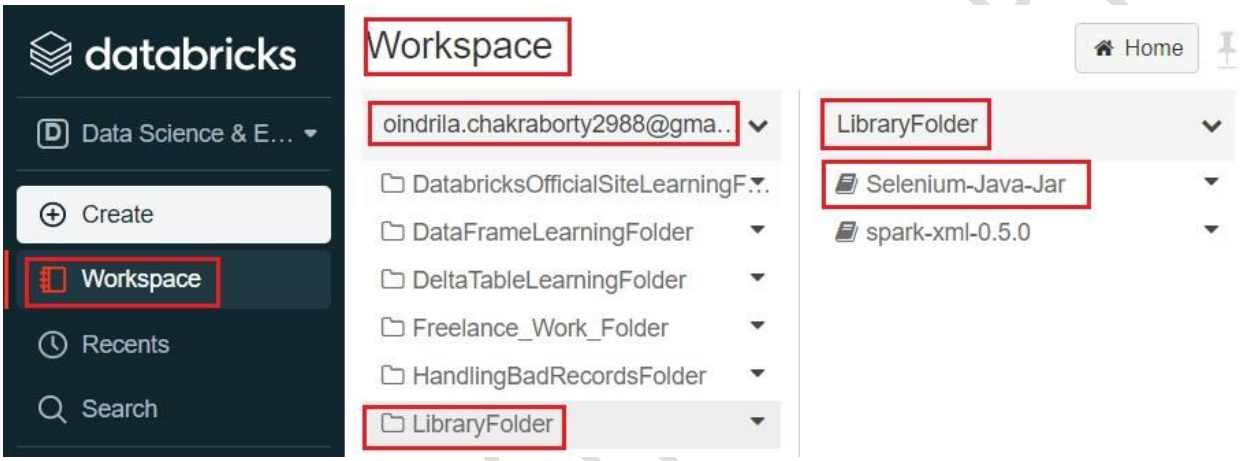


✓ Click on the “Done” button. The “Library” is “Installed” on the “Cluster”.

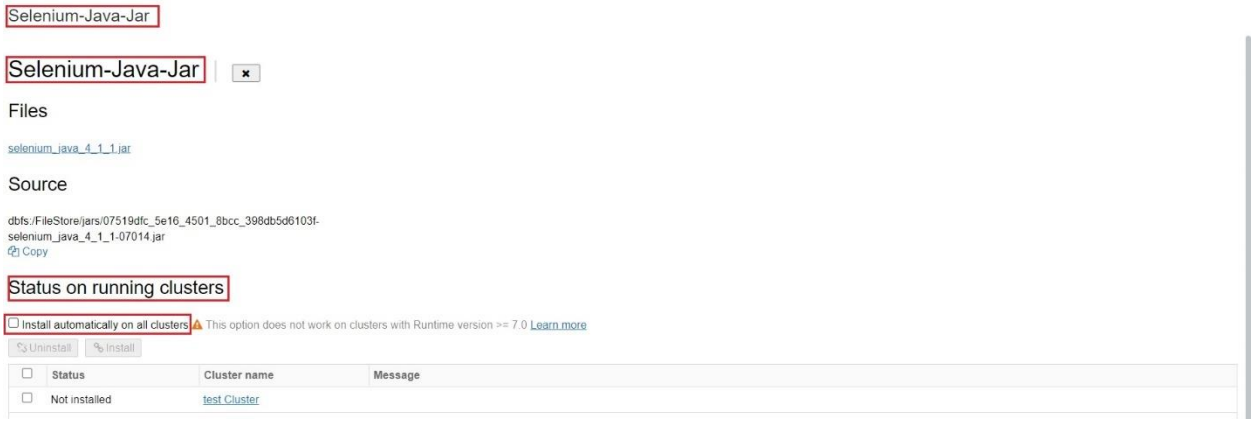


- ✓ This option does “Not Install” the “Library” on the “Clusters”, running with “Databricks Runtime Version 7.0” and “Above”.

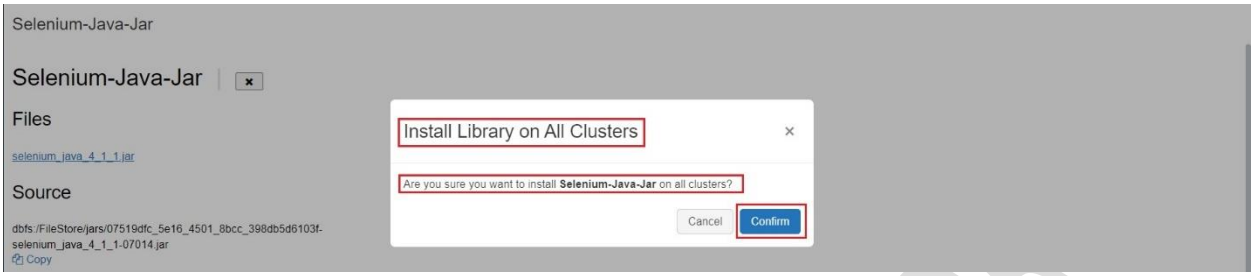
- ***Starting from “Library UI”*** -
  - ➢ Go to the desired “User”/”Shared” “Workspace Folder”, containing the “Workspace Library” to “Install”.
  - ➢ Click on the name of the “Library”.



  - ➢ Perform one of the following steps -
    - ✓ To “Configure” the “Installed” “Library” on all the “Clusters”, perform the following steps -
      - Click on the name of the “Library”.
      - The “Workspace Library” details page opens. Select the “Install automatically on all clusters” Checkbox.



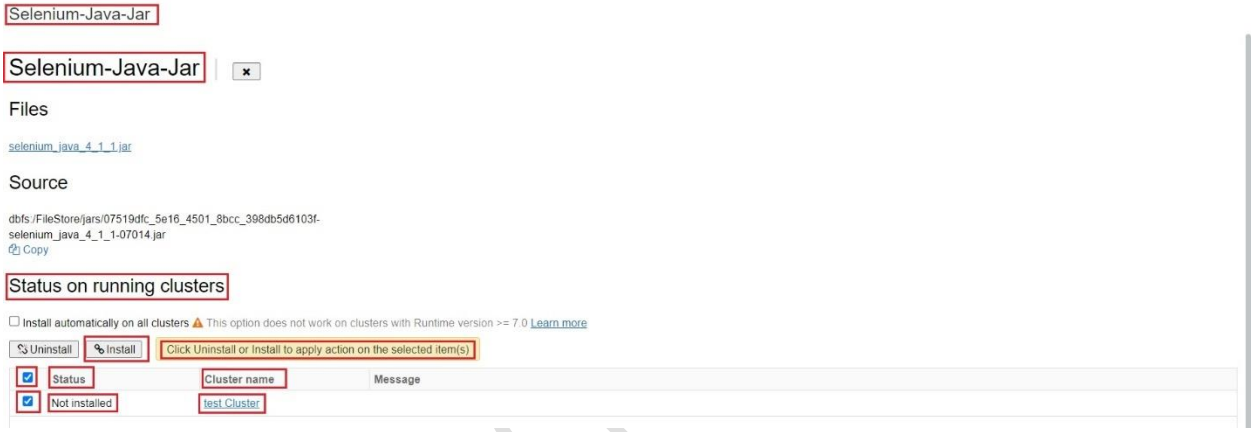
      - A Confirmation Dialog Box opens. Click on the “Confirm” button.



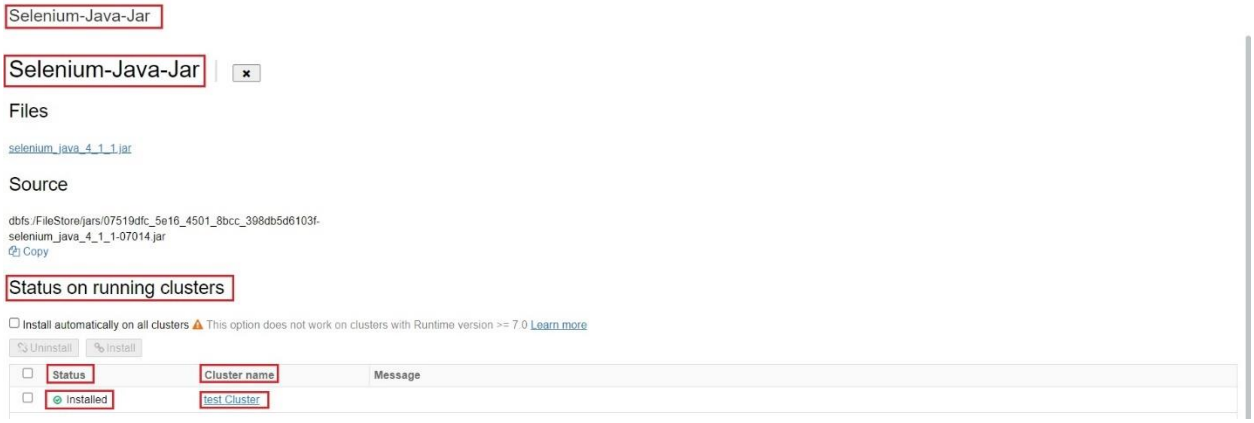
- "Restart" the "Cluster". Come back to the details page of the "Workspace Library" that was "Configured" to be "Installed" on all the "Clusters". The "Library" will be "Installed" on the all the "Clusters" and the respective "Status" will be displayed as "Installed", as follows -

- This option does "Not Install" the "Library" on the "Clusters", running with "Databricks Runtime Version 7.0" and "Above". For such "Clusters", the "Status" will be displayed as "Skipped", as follows -

- ✓ To “Configure” the “Installed” “Library” on the “Specific Clusters”, perform the following steps -
  - Click on the name of the “Library”.
  - The “Workspace Library” details page opens. Select the Checkbox, next to the desired “Cluster (s)” in which the “Workspace Library” needs to be “Installed”. Then, click on the “Install” button.

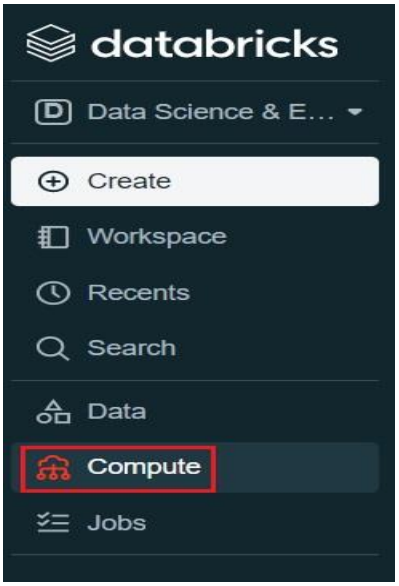


  - It can be seen from the below image that the “Workspace Library” is “Installed” in the selected “Cluster”.



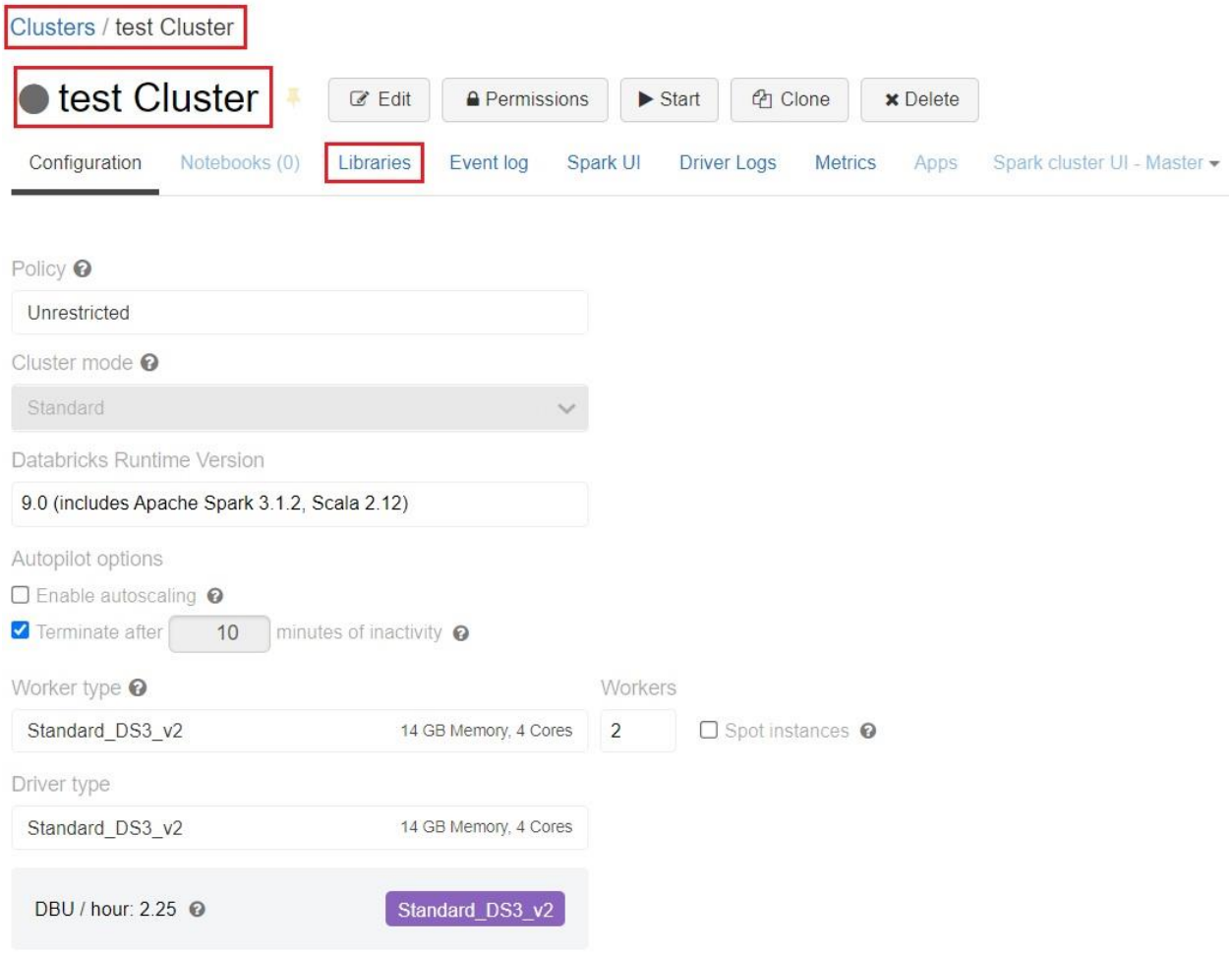
Create a Cluster Library

- It is possible to “Install” a “Library” on a specific “Cluster” without making it available as a “Workspace Library”.
- To “Install” a “Library” on a “Cluster”, first, click on the “Compute” icon on the Sidebar.

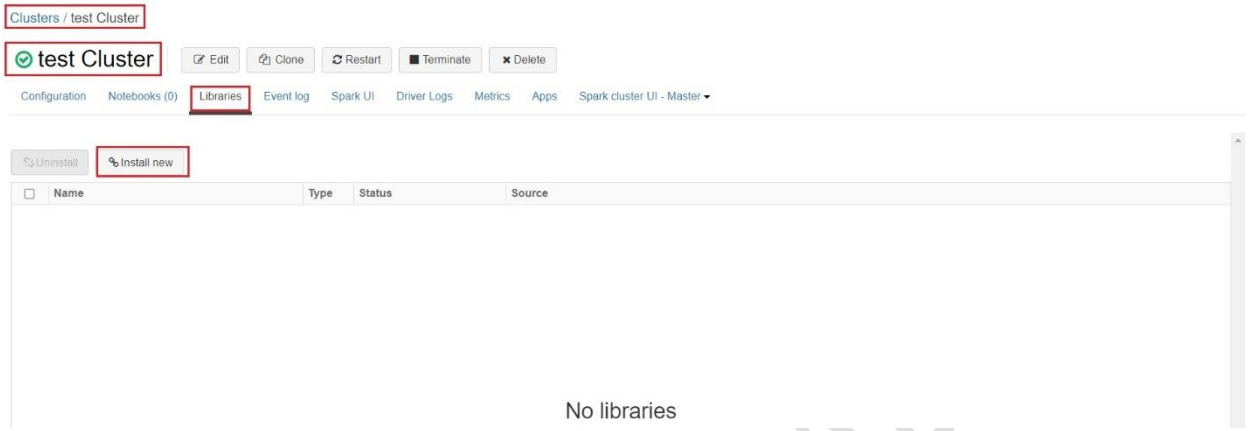


- Click on the name of any “Cluster” displayed.

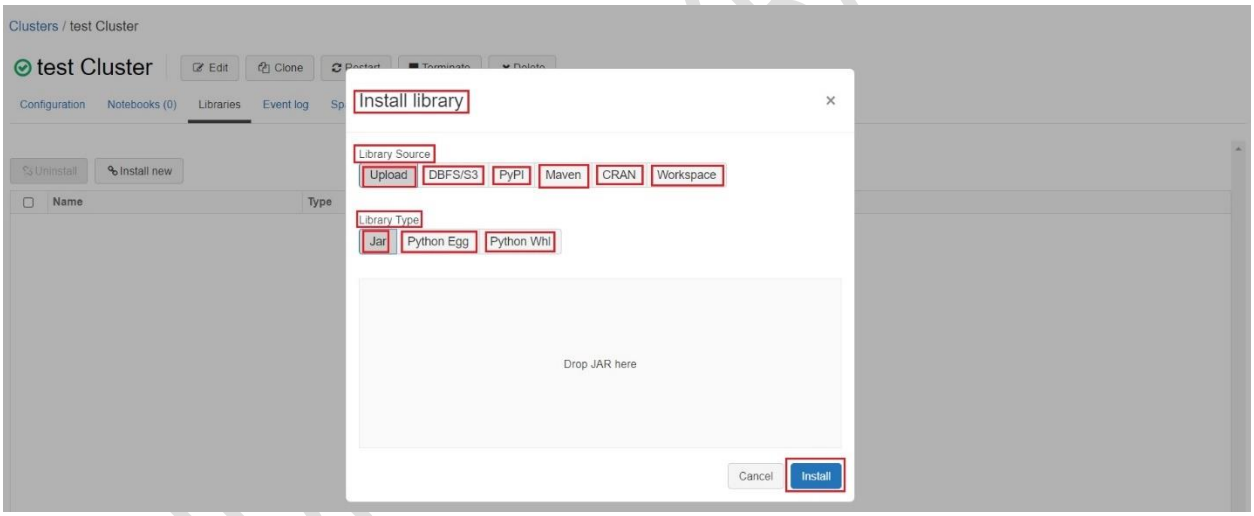
- The details page of the respective “Cluster” opens. Click on “Libraries” tab.



- Click on the “Install new” button. This button will be enabled only if the “Cluster” is in “Running” state.



- The “Install Library” Dialog Box opens.



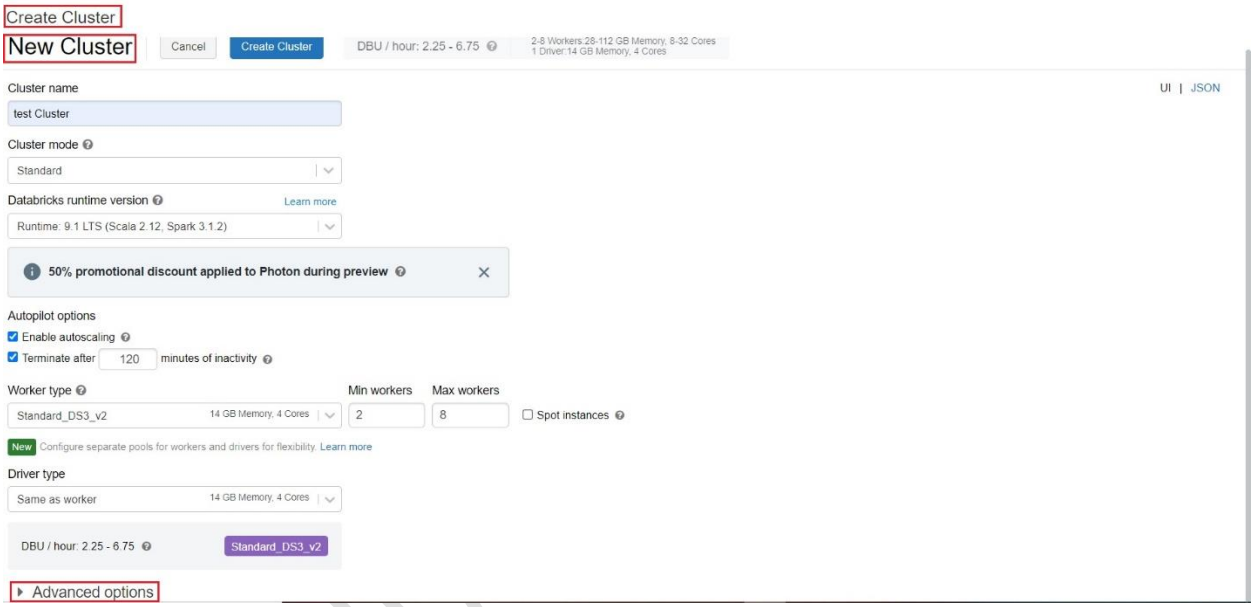
- Perform one of the methods for creating a “Workspace Library”. After the “Install” button is clicked, the respective “Library” will be “Installed” on the “Cluster”.

Create a Cluster Library Using INIT Script

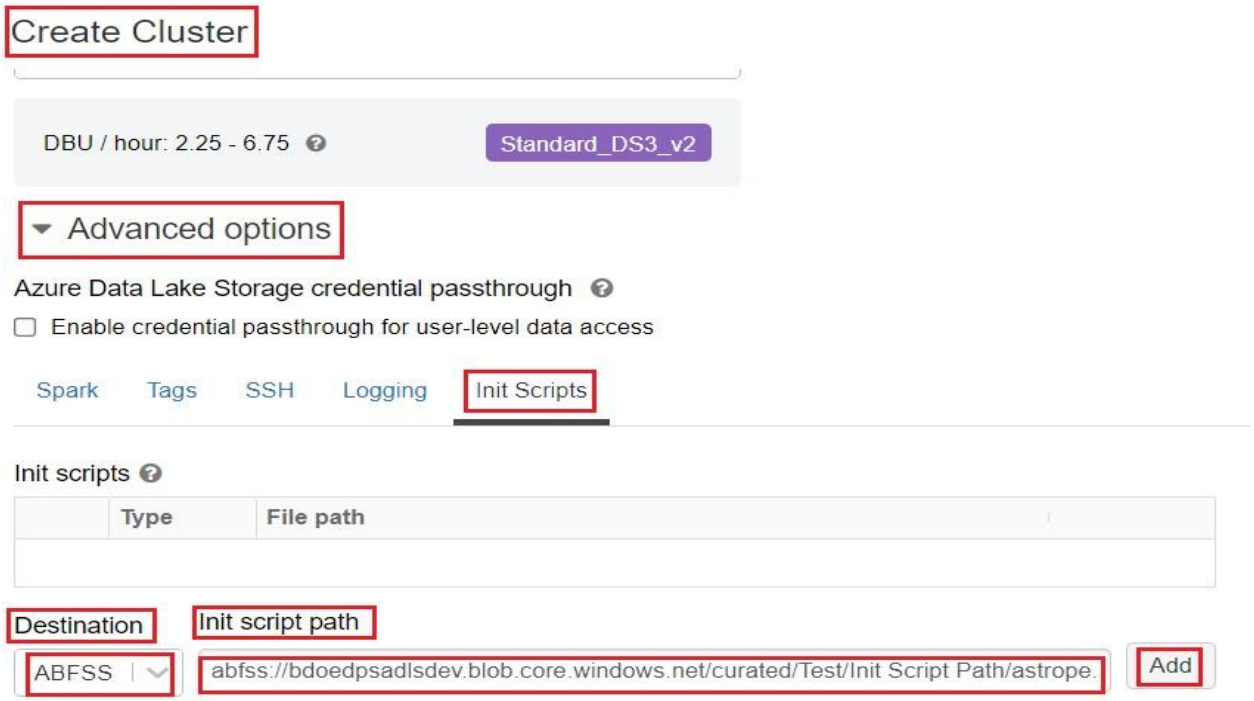
- If the “Library” to “Install” requires “Custom Configuration”, it would Not be Possible to “Install” using the “Workspace Library” or “Cluster Library” Interface. In such cases, the “Cluster Library” can be “Installed” using an “INIT Script” that “Runs” at the time of creating the “Cluster”.
- Following is an example of an “INIT” Script that uses “pip” to “Install” the “Python Libraries” -

#!/bin/bash
/databricks/python/bin/pip install astropy

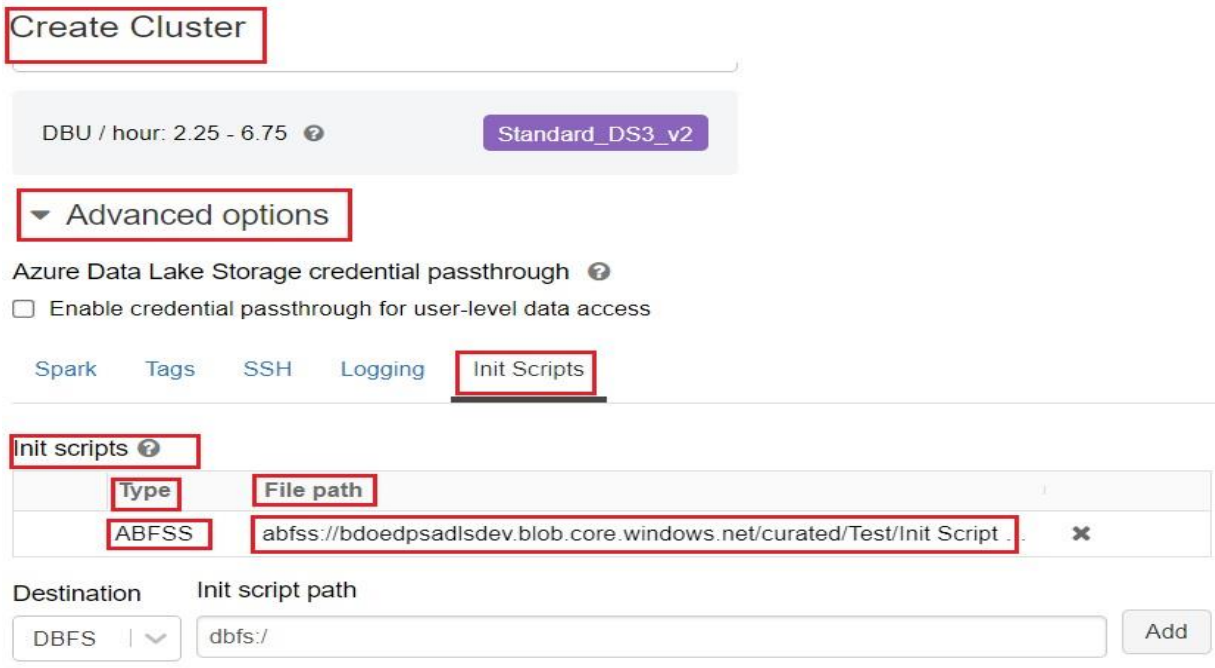
- This “INIT” Script can be used on a “Cluster” at the time of the creation of that “Cluster” using the following steps -
  - In the “New Cluster” page for creation of a “Cluster”, click on the “Advanced options” section header to expand it.



  - Go to the “Init Scripts” tab.
  - The “INIT” Script needs to “Stored” as a File in either “DBFS” or the “Object Storage”, attached to the “Databricks Workspace”, e.g., “Azure Data Lake Storage”. Select the desired Option from the “Drop Down” “Destination”.
  - Provide the respective “INIT” Script File Path in the “Init script path” textbox and click on the “Add” button.



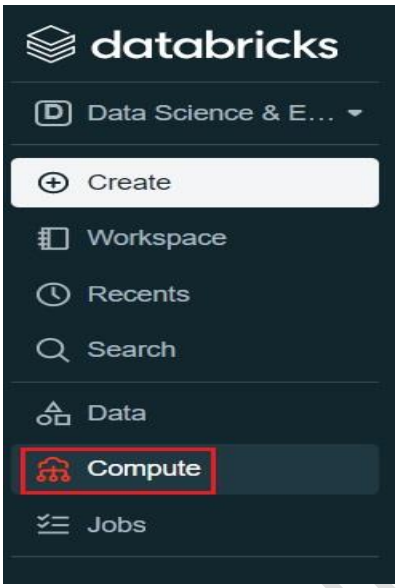
- The added "INIT" Script path along with from where the "INIT" Script File would be read, is displayed in the "Init scripts" section as follows -



- The respective “Custom Library” will be “Installed” on the “Cluster” after the “Cluster” itself gets created.

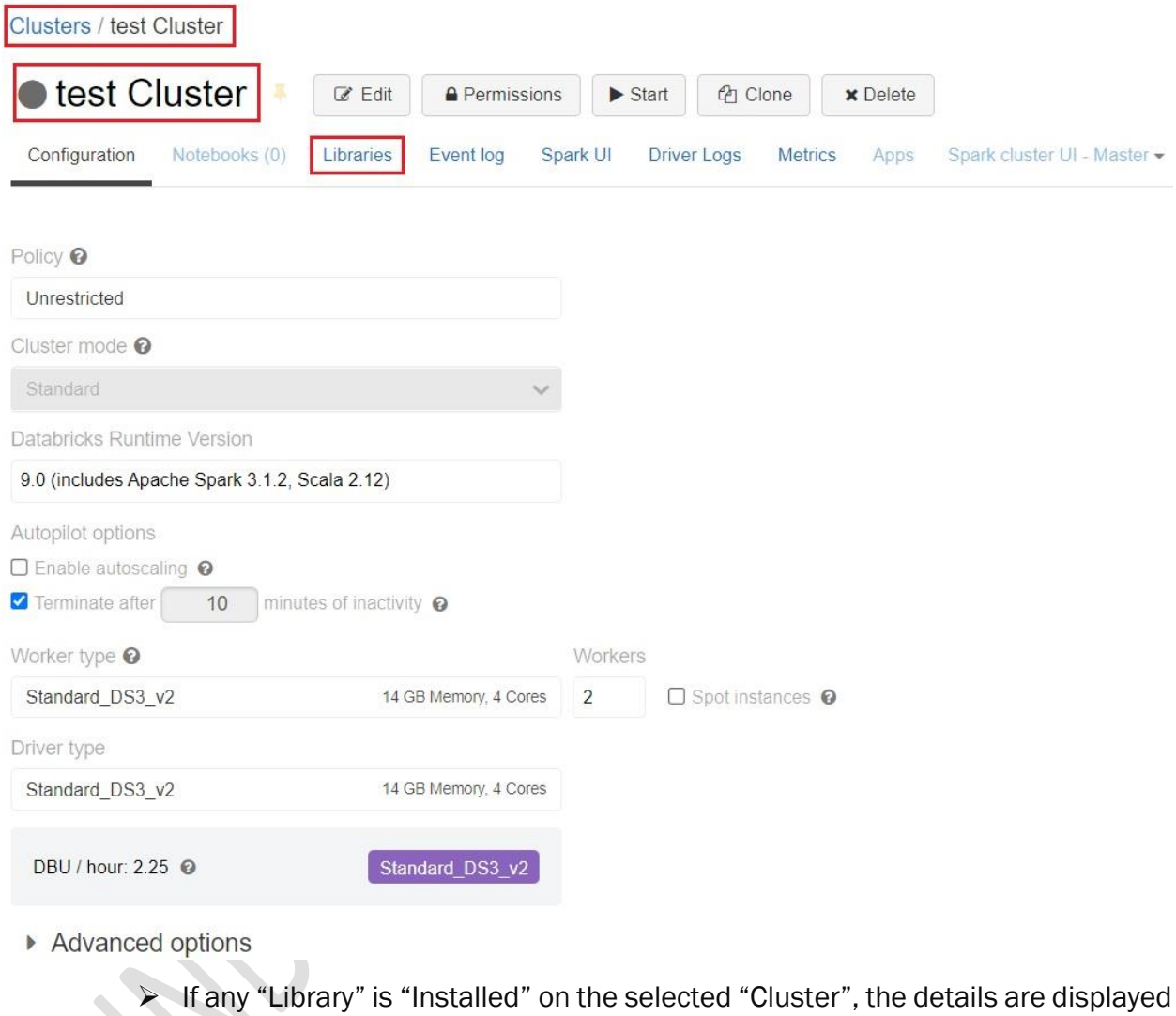
View Installed Cluster Library Details

- To view the details of the Installed “Cluster Library”, perform the following steps –
  - Click on the “Compute” icon on the Sidebar.



  - Click on the name of any “Cluster” displayed.

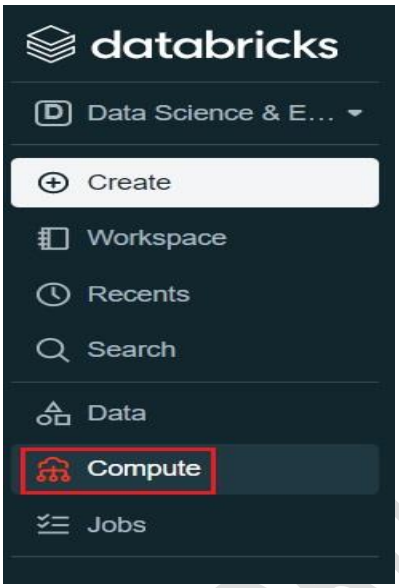
  - The details page of the respective “Cluster” opens. Click on “Libraries” tab.



- If any “Library” is “Installed” on the selected “Cluster”, the details are displayed in the “Tabular” form, as follows -

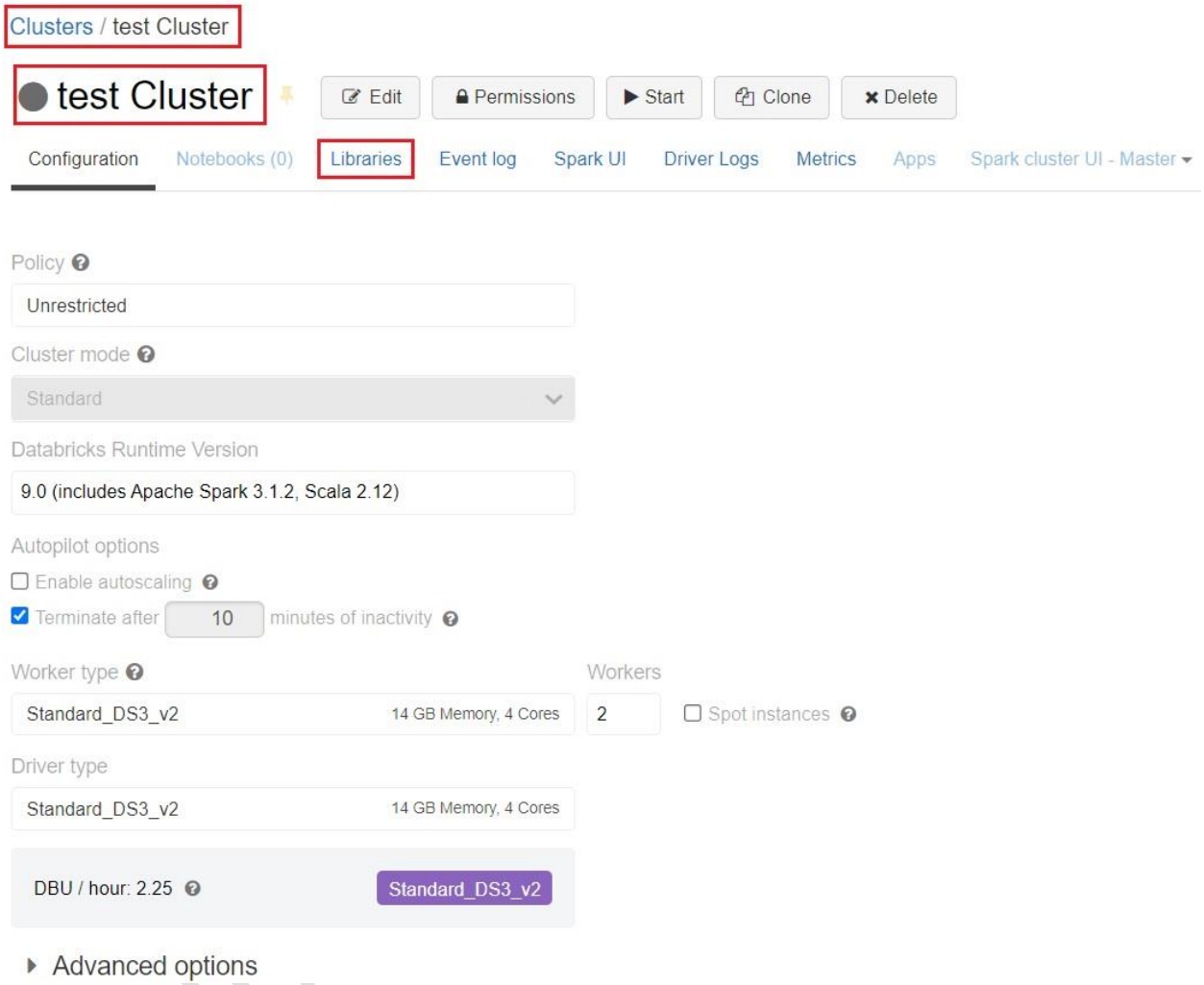
Uninstall a Library from a Cluster

- When a “Library” is “Uninstalled” from a “Cluster”, the “Library” is “Removed” only when the “Cluster” is “Re-Started”. Until the “Cluster” is “Re-Started”, the “Status” of the “Uninstalled” “Library” appears as “Uninstall pending restart”.
- To “Uninstall” a “Library” from a “Cluster”, either start from the “Cluster UI” or from the “Library UI”.
- ***Starting from “Cluster UI”*** -
  - Click on the “Compute” icon on the Sidebar.



  - Click on the name of any “Cluster” displayed.

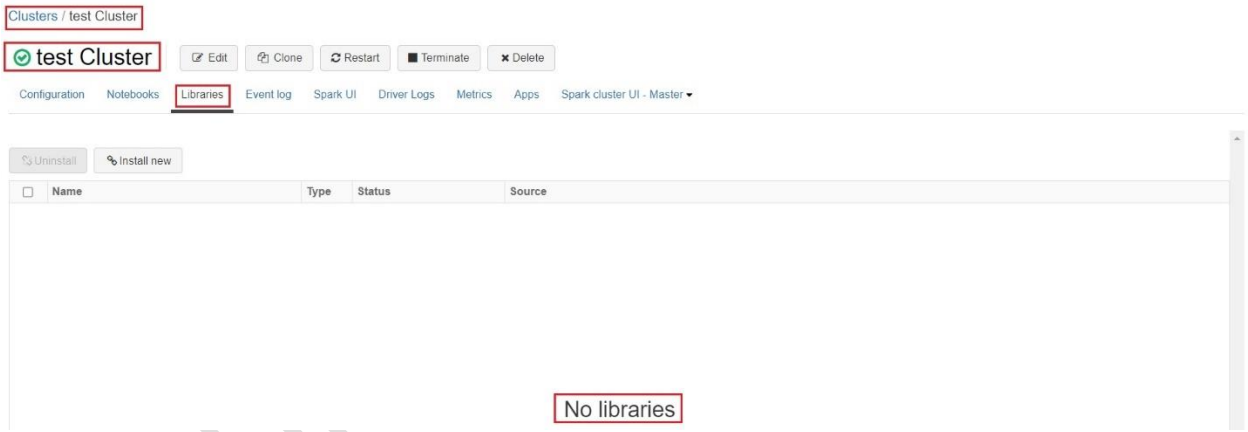
  - The details page of the respective “Cluster” opens. Click on “Libraries” tab.



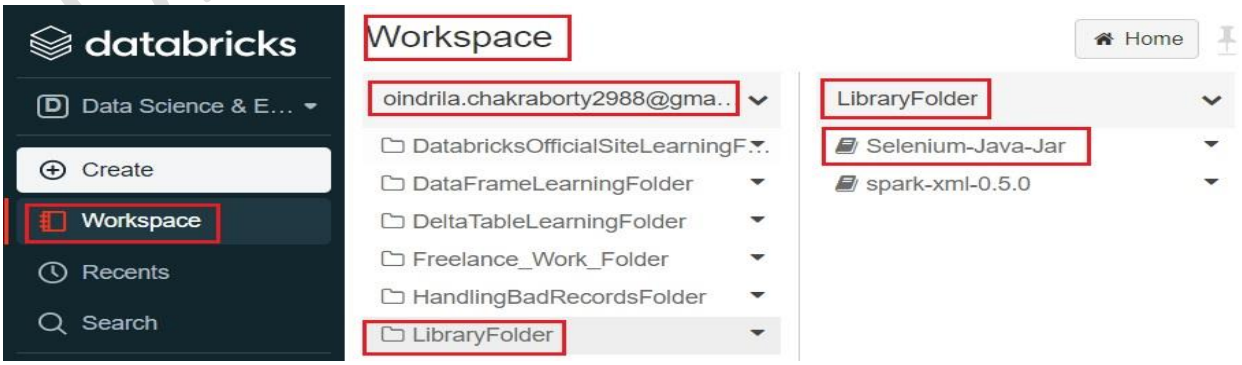
- Select the Checkbox/Checkboxes, next to the desired “Library” / ”Libraries” to “Uninstall” from the specific “Cluster”. Then, click on the “Uninstall” button.

- A Confirmation Dialog Box opens. Click on the “Confirm” button.

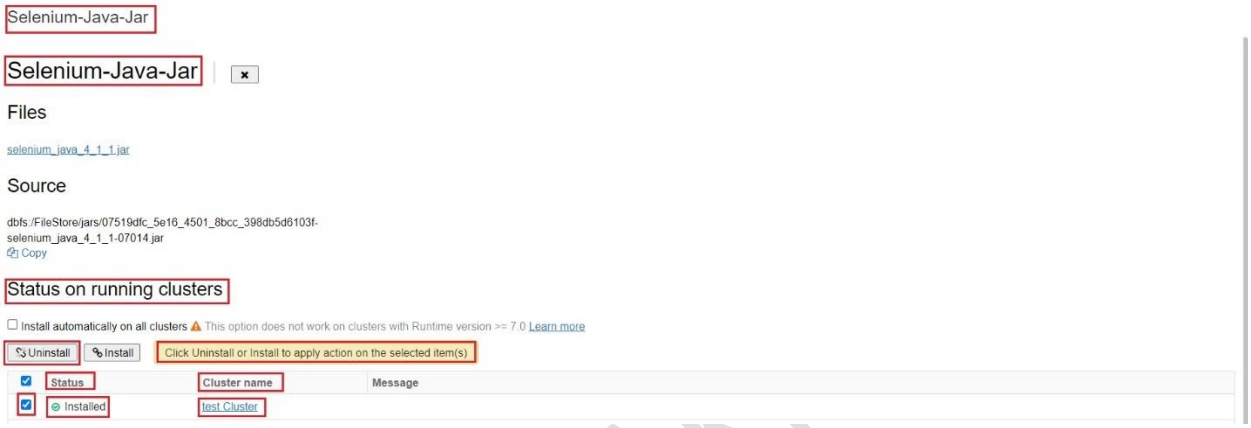
- "Restart" the "Cluster" and come back to the details page of the respective "Cluster". Go to the "Libraries" tab. It can be seen from the below image that the previously "Installed" "Library" on this "Cluster" is "No More Installed".



- ***Starting from "Library UI"*** -
  - Go to the desired "User"/"Shared" "Workspace Folder", containing the "Workspace Library" to "Install".
  - Click on the name of the "Library".



- The details page of the “Library” opens. Select the Checkbox/Checkboxes, next to the desired “Cluster (s)” from where the specific “Library” needs to be “Uninstalled”. Then, click on the “Uninstall” button.



- A Confirmation Dialog Box opens. Click on the “Confirm” button.

- “Restart” the “Cluster” and come back to the details page of the respective “Library”. It can be seen from the below image that the “Library” is “No More Installed” on the “Cluster (s)”.