



BIG DATA
ACADEMY

LABORATORIO 22

INSTALACIÓN DE LIBRERÍAS

FORMADOR: ALONSO MELGAREJO
alonsoraulmgs@gmail.com

LABORATORIO 22 – INSTALACIÓN DE LIBRERÍAS

1. En Internet, buscamos un foro en donde nos den la librería para leer archivos XML, entramos al siguiente enlace:

<https://community.cloudera.com/t5/Support-Questions/Error-processing-XML-data-in-Spark-Unable-to-process-XML/td-p/187436>

Y encontramos los datos de la librería que soluciona el problema:

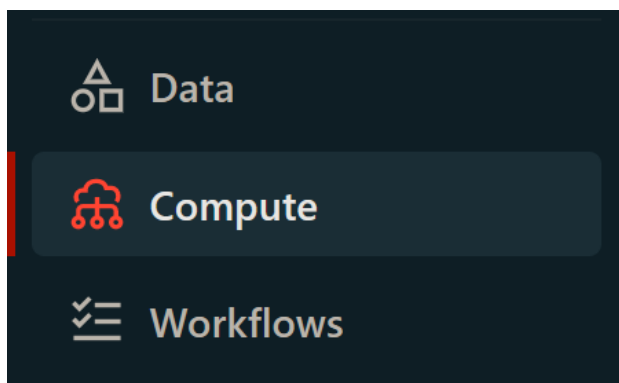
```
package for load the XML files. You can load the  
mand with spark-submit or spark-shell.  
  
ackages com.databricks:spark-xml_2.10:0.4.1
```

2. Según el foro, la librería que debemos instalar es la siguiente:

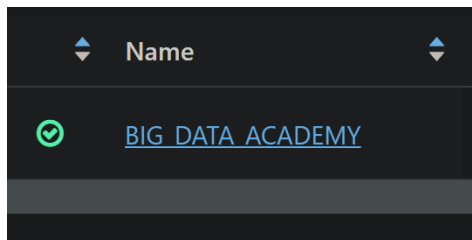
providerId	com.databricks
artifactId	spark-xml_2.10
version	0.4.1

Vemos que en el artifactId se incluye un número (spark-xml_2.10), este indica la versión de SCALA con el que la librería es compatible (Scala 2.10)

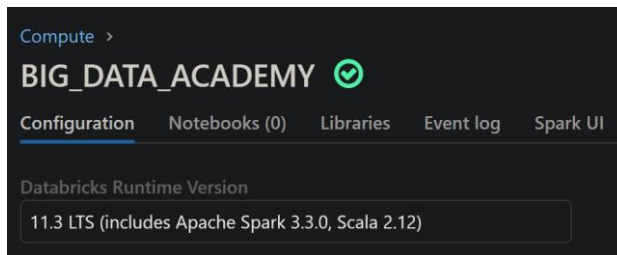
3. SPARK funciona sobre el lenguaje SCALA, cuando instalamos una librería debemos asegurarnos que la librería sea compatible con la versión de SCALA que tenemos en nuestro clúster. Para averiguar qué versión de SCALA tenemos, sobre DATABRICKS seleccionamos la opción "Compute"



4. Damos clic sobre el nombre de nuestro clúster



5. En la pestaña “Configuration” podemos encontrar que en nuestro clúster tenemos la versión “2.12” de SCALA



No podemos instalar la librería “spark-xml_2.10” en nuestro clúster ya que no es compatible con la versión de SCALA de nuestro clúster (2.12)

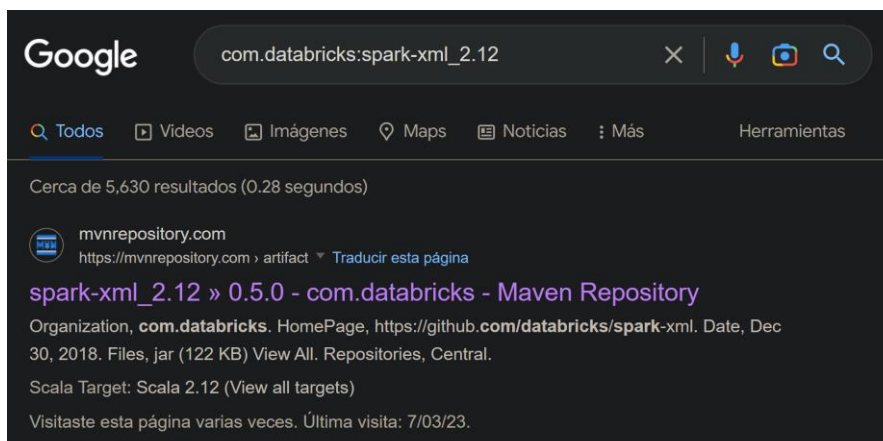
6. Buscaremos si el providerId (com.databricks) ha publicado una librería “spark-xml_2.12”, en el navegador buscamos:

com.databricks:spark-xml_2.12

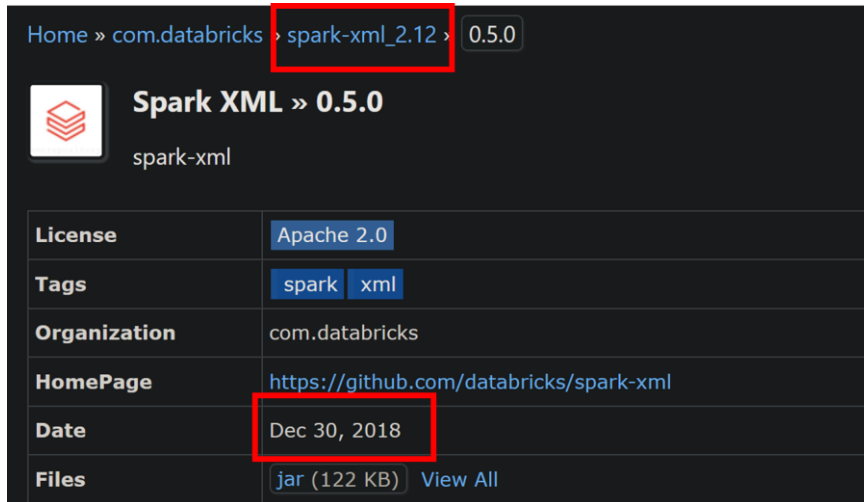
7. En el primer resultado encontremos la web de MAVEN (<https://mvnrepository.com>), el cual es el repositorio más grande de librerías, en este repositorio las empresas publican librerías gratuitas para que los desarrolladores puedan descargarlas. Vemos que se nos muestra el enlace MAVEN de la librería “spark-xml_2.12”, damos clic sobre él.

ENLACE

https://mvnrepository.com/artifact/com.databricks/spark-xml_2.12/0.5.0



8. Vemos que la librería “spark-xml_2.12” para su versión “0.5.0” fue publicada el 30 de diciembre del 2018, tal vez haya una versión más reciente, damos clic sobre el nombre de la librería.

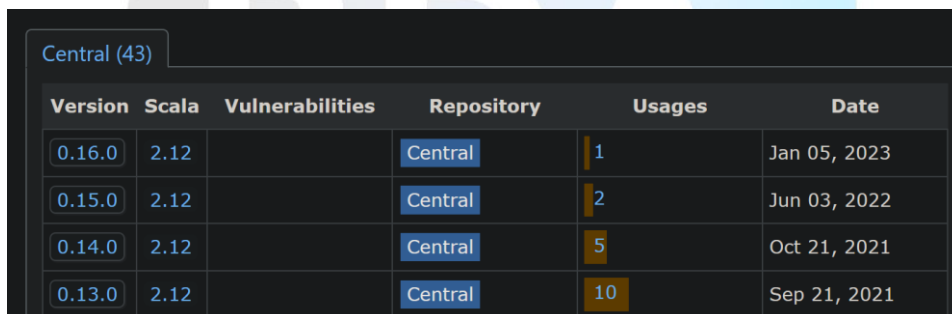


Home » com.databricks » spark-xml_2.12 » 0.5.0

Spark XML » 0.5.0
spark-xml

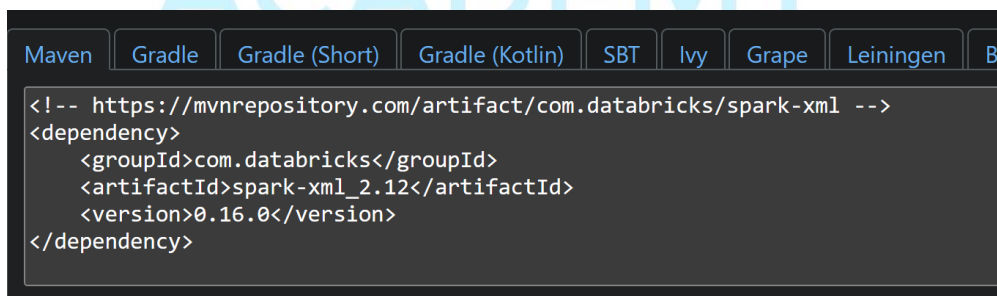
License	Apache 2.0
Tags	spark xml
Organization	com.databricks
HomePage	https://github.com/databricks/spark-xml
Date	Dec 30, 2018
Files	jar (122 KB) View All

9. Encontramos que hay una versión más reciente, damos clic sobre la versión más reciente para ver el detalle.



Version	Scala	Vulnerabilities	Repository	Usages	Date
0.16.0	2.12		Central	1	Jan 05, 2023
0.15.0	2.12		Central	2	Jun 03, 2022
0.14.0	2.12		Central	5	Oct 21, 2021
0.13.0	2.12		Central	10	Sep 21, 2021

10. Podemos encontrar los detalles de instalación de la librería:



Maven Gradle Gradle (Short) Gradle (Kotlin) SBT Ivy Grape Leiningen B

```
<!-- https://mvnrepository.com/artifact/com.databricks/spark-xml -->
<dependency>
  <groupId>com.databricks</groupId>
  <artifactId>spark-xml_2.12</artifactId>
  <version>0.16.0</version>
</dependency>
```

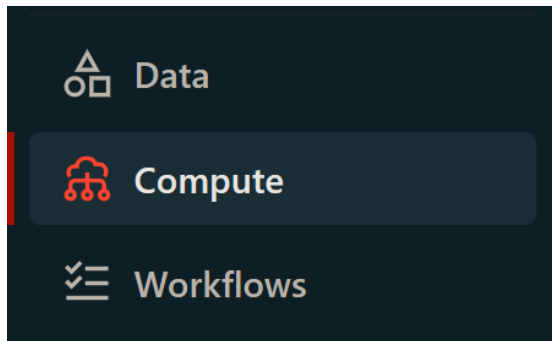
providerId	com.databricks
artifactId	spark-xml_2.12
version	0.16.0

La cadena de instalación de la librería compatible con nuestro clúster es:

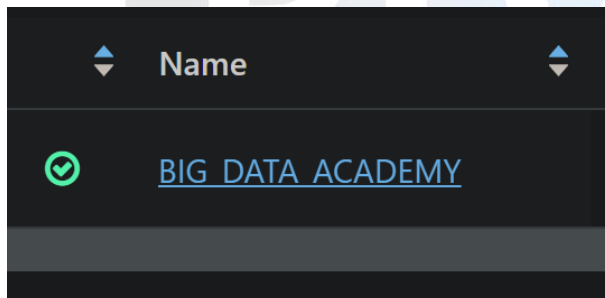
```
com.databricks:spark-xml_2.12:0.16.0
```

IMPORTANTE: Siempre debemos de adaptar la librería a la versión exacta de SCALA que tengamos en el clúster, sino la librería no funcionará.

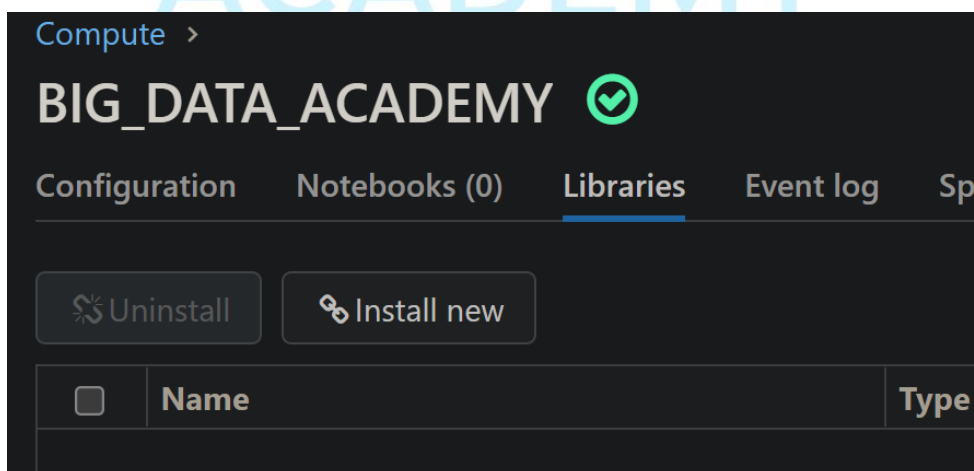
11. Para instalar la librería, sobre DATABRICKS seleccionamos la opción “Compute”



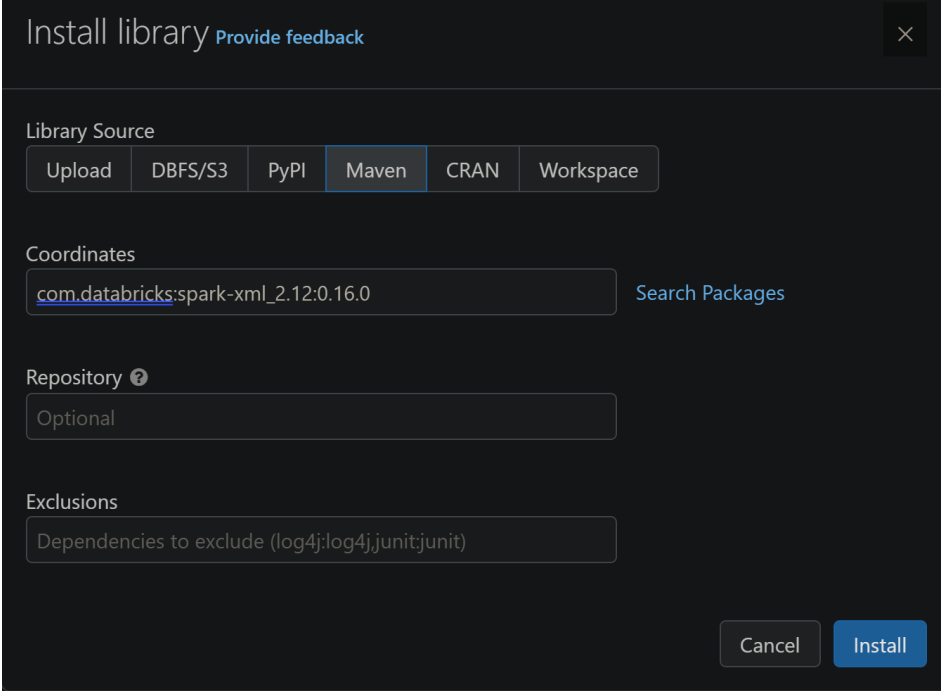
12. Damos clic sobre el nombre de nuestro clúster



13. En la pestaña “Libraries” damos clic sobre “Install new”

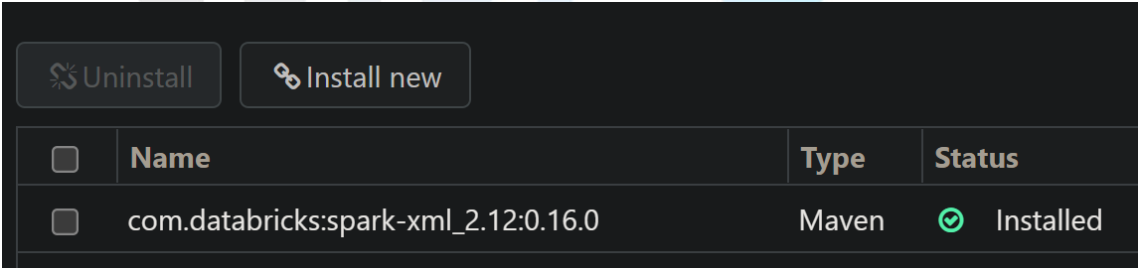





14. Seleccionamos la opción “Maven” y en el campo “Coordinates” ingresamos la cadena de instalación. Damos clic en “Install”.



The dialog box titled "Install library" has a "Provide feedback" link and a close button. It contains several sections: "Library Source" with tabs for Upload, DBFS/S3, PyPI, Maven (selected), CRAN, and Workspace; "Coordinates" with a text input field containing "com.databricks:spark-xml_2.12:0.16.0" and a "Search Packages" button; "Repository" with a dropdown menu set to "Optional"; and "Exclusions" with a text input field containing "Dependencies to exclude (log4j:log4junit:junit)". At the bottom right are "Cancel" and "Install" buttons.

15. La librería se descargará desde el repositorio MAVEN y se instalará en el clúster



 Uninstall		 Install new	
<input type="checkbox"/>	Name	Type	Status
<input type="checkbox"/>	com.databricks:spark-xml_2.12:0.16.0	Maven	 Installed