
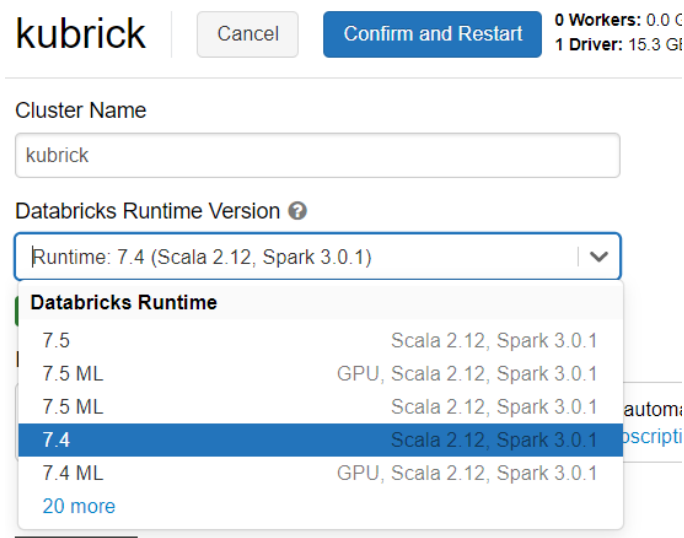


Reading XML data in Spark

XML files can be processed in Spark by using this open-source package published by Databricks:
<https://github.com/databricks/spark-xml>

You need the right version of this library for your cluster. The version 2.12:0.x of the above spark-xml package works with version 7.4 of Databricks (Spark 3.0.1, Scala 2.12) but apparently not with version 7.5 (released December 16, 2020). To install it in Databricks:

1. Click on Clusters,  and choose your cluster. The Community version of Databricks only allows one active cluster at a time.
2. Check the version of your cluster (Databricks Runtime Version). If later than 7.4 then you may have to change it to 7.4 to get the following package to work. To change the version, click Edit, choose the required version from the dropdown menu and then click Confirm and Restart.



kubrick | Cancel | Confirm and Restart | 0 Workers: 0.0 C | 1 Driver: 15.3 Gi

Cluster Name

kubrick

Databricks Runtime Version ?

Runtime: 7.4 (Scala 2.12, Spark 3.0.1) | v

Databricks Runtime

7.5	Scala 2.12, Spark 3.0.1
7.5 ML	GPU, Scala 2.12, Spark 3.0.1
7.5 ML	Scala 2.12, Spark 3.0.1
7.4	Scala 2.12, Spark 3.0.1
7.4 ML	GPU, Scala 2.12, Spark 3.0.1
20 more	

3. Select Libraries => Install New => Select Library Source = "Maven" => Coordinates => Search Packages => Select **Maven Central** => Search
Enter "spark-xml" and find the following package:
com.databricks:spark-xml_2.12:0.11.0
(see [here](#) for a step-by-step example of how to do this)

Now try it out:

4. Download this example file and load it to the Databricks file system:
<https://raw.githubusercontent.com/databricks/spark-xml/master/src/test/resources/books.xml>

(continued . . .)

5. Run the following script:

```
from pyspark.sql.functions import explode

dcat = spark.read.format('xml')\
    .options(rowTag='catalog')\
    .load('/FileStore/tables/books.xml')

book = dcat.withColumn('bookdetail', explode('book')).select('bookdetail.*')
book.select('_id','title','author','genre','publish_date','price').show()

# The descriptions can have multiple lines
# Not a problem, but it's easier to read on the screen if the description isn't included
# book.select('_id','title','author','genre','publish_date','price','description').show()
```

Expected Result:

_id	title	author	genre	publish_date	price
bk101	XML Developer's G...	Gambardella, Matthew	Computer	2000-10-01	44.95
bk102	Midnight Rain	Ralls, Kim	Fantasy	2000-12-16	5.95
bk103	Maeve Ascendant	Corets, Eva	Fantasy	2000-11-17	5.95
bk104	Oberon's Legacy	Corets, Eva	Fantasy	2001-03-10	5.95
bk105	The Sundered Grail	Corets, Eva	Fantasy	2001-09-10	5.95
bk106	Lover Birds	Randall, Cynthia	Romance	2000-09-02	4.95
bk107	Splish Splash	Thurman, Paula	Romance	2000-11-02	4.95
bk108	Creepy Crawlies	Knorr, Stefan	Horror	2000-12-06	4.95
bk109	Paradox Lost	Kress, Peter	Science Fiction	2000-11-02	6.95
bk110	Microsoft .NET: T...	O'Brien, Tim	Computer	2000-12-09	36.95
bk111	MSXML3: A Compreh...	O'Brien, Tim	Computer	2000-12-01	36.95
bk112	Visual Studio 7: ...	Galos, Mike	Computer	2001-04-16	49.95