

# BDA Lab 13

## SPARK Data ingestion from XML

Download the XML plugin provided by Databricks from:

<https://github.com/databricks/spark-xml>

The plugin uses sbt for building from source you can setup sbt for mac using

```
brew install sbt
```

Now build the plugin from source files using:

```
sbt package
```

```
i (ignored the warnings)
```

Once all done follow the process showed in the lab to read the XML in Spark DataFrame and then save it to a CSV file

## Requirements:


```
[error] at sbt.MainLoop$.runLoggedLoop(MainLoop.scala:44)
[error] at sbt.MainLoop$.runLogged(MainLoop.scala:35)
[error] at sbt.StandardMain$.runManaged(Main.scala:138)
[error] at sbt.xMain.run(Main.scala:89)
[error] at xsbt.boot.Launch$$anonfun$run$1.apply(Launch.scala:111)
[error] at xsbt.boot.Launch$.withContextLoader(Launch.scala:131)
[error] at xsbt.boot.Launch$.run(Launch.scala:111)
[error] at xsbt.boot.Launch$$anonfun$apply$1.apply(Launch.scala:37)
[error] at xsbt.boot.Launch$.launch(Launch.scala:120)
[error] at xsbt.boot.Launch$.apply(Launch.scala:20)
[error] at xsbt.boot.Boot$.runImpl(Boot.scala:56)
[error] at xsbt.boot.Boot$.main(Boot.scala:18)
[error] at xsbt.boot.Boot.main(Boot.scala)
[error] java.io.IOException: java.lang.RuntimeException: /packages cannot be represented as URI
[error] Use 'last' for the full log.
Project loading failed: (retry, (e)xit, (l)ast, or (i)gnore? i
[warn] Ignoring load failure: no project loaded.
[error] java.lang.RuntimeException: Session not initialized.
[error] at scala.sys.package$.error(package.scala:26)
[error] at sbt.Project$.sanonfun$getError$1(Project.scala:442)
[error] at scala.Option.getOrElse(Option.scala:121)
[error] at sbt.Project$.getError(Project.scala:442)
[error] at sbt.Project$.session(Project.scala:448)
[error] at sbt.Project$.extract(Project.scala:453)
[error] at sbt.BuiltinCommands$.notifyUsersAboutShell(Main.scala:928)
[error] at sbt.BuiltinCommands$.sanonfun$notifyUsersAboutShell$1$3(Main.scala:937)
[error] at sbt.Command$.sanonfun$Command$2(Command.scala:91)
[error] at sbt.Command$.process(Command.scala:181)
[error] at sbt.MainLoop$.processCommand(MainLoop.scala:151)
[error] at sbt.MainLoop$.sanonfun$next$2(MainLoop.scala:139)
[error] at sbt.State$$anon$1.runCmd$1(State.scala:246)
[error] at sbt.State$$anon$1.process(State.scala:250)
[error] at sbt.MainLoop$.sanonfun$next$1(MainLoop.scala:139)
[error] at sbt.internal.util.ErrorHandling$.wideConvert(ErrorHandling.scala:16)
[error] at sbt.MainLoop$.next(MainLoop.scala:139)
[error] at sbt.MainLoop$.run(MainLoop.scala:132)
[error] at sbt.MainLoop$.sanonfun$runWithNewLog$1(MainLoop.scala:110)
[error] at sbt.io.Using$.apply(Using.scala:22)
[error] at sbt.MainLoop$.runWithNewLog(MainLoop.scala:104)
[error] at sbt.MainLoop$.runAndPostLast(MainLoop.scala:59)
[error] at sbt.MainLoop$.runLoggedLoop(MainLoop.scala:44)
[error] at sbt.MainLoop$.runLogged(MainLoop.scala:35)
[error] at sbt.StandardMain$.runManaged(Main.scala:138)
[error] at sbt.xMain.run(Main.scala:89)
[error] at xsbt.boot.Launch$$anonfun$run$1.apply(Launch.scala:111)
[error] at xsbt.boot.Launch$.withContextLoader(Launch.scala:131)
[error] at xsbt.boot.Launch$.run(Launch.scala:111)
[error] at xsbt.boot.Launch$$anonfun$apply$1.apply(Launch.scala:37)
[error] at xsbt.boot.Launch$.launch(Launch.scala:120)
[error] at xsbt.boot.Launch$.apply(Launch.scala:20)
[error] at xsbt.boot.Boot$.runImpl(Boot.scala:56)
[error] at xsbt.boot.Boot$.main(Boot.scala:18)
[error] at xsbt.boot.Boot.main(Boot.scala)
[error] Session not initialized.
[error] Use 'last' for the full log.
saddheseem@Macbook-Pro ~/Documents/FAST/FAST-NUCES-LABS/BDA/Labs/BDA-Lab-13/spark-xml-master *master*
```

Plugin build process. Mine shows **error** because I installed sbt using brew so I had to ignore the warnings and minor errors but the build was successful your output might be different. If your build has fatal errors, it might fail entirely and you would have to fix the problem first ("RollNo-1.png")

```
Out[5]: SparkSession - in-memory
SparkContext

Spark UI
Version
v3.1.1
Master
local[*]
AppName
Saad Naeem i19-1207 XML to Dataframe
```

PySpark's SparkContext ("RollNo-2.png")

3.1.1

JobsStagesStorageEnvironmentExecutorsSQL

Saad Naeem i19-1207 XML to Dataf... application UI

### Executors

[Show Additional Metrics](#)

#### Summary

	<a href="#">RDD Blocks</a>	<a href="#">Storage Memory</a>	<a href="#">Disk Used</a>	<a href="#">Cores</a>	<a href="#">Active Tasks</a>	<a href="#">Failed Tasks</a>	<a href="#">Complete Tasks</a>	<a href="#">Total Tasks</a>	<a href="#">Task Time (GC Time)</a>	<a href="#">Input</a>	<a href="#">Shuffle Read</a>	<a href="#">Shuffle Write</a>	<a href="#">Excluded</a>
Active(1)	0	0.0 B / 366.3 MiB	0.0 B	8	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(1)	0	0.0 B / 366.3 MiB	0.0 B	8	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0

#### Executors

Show  entries

Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Thread Dump
driver	192.168.10.5:62079	Active	0	0.0 B / 366.3 MiB	0.0 B	8	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	<a href="#">Thread Dump</a>

Showing 1 to 1 of 1 entries

Previous1Next

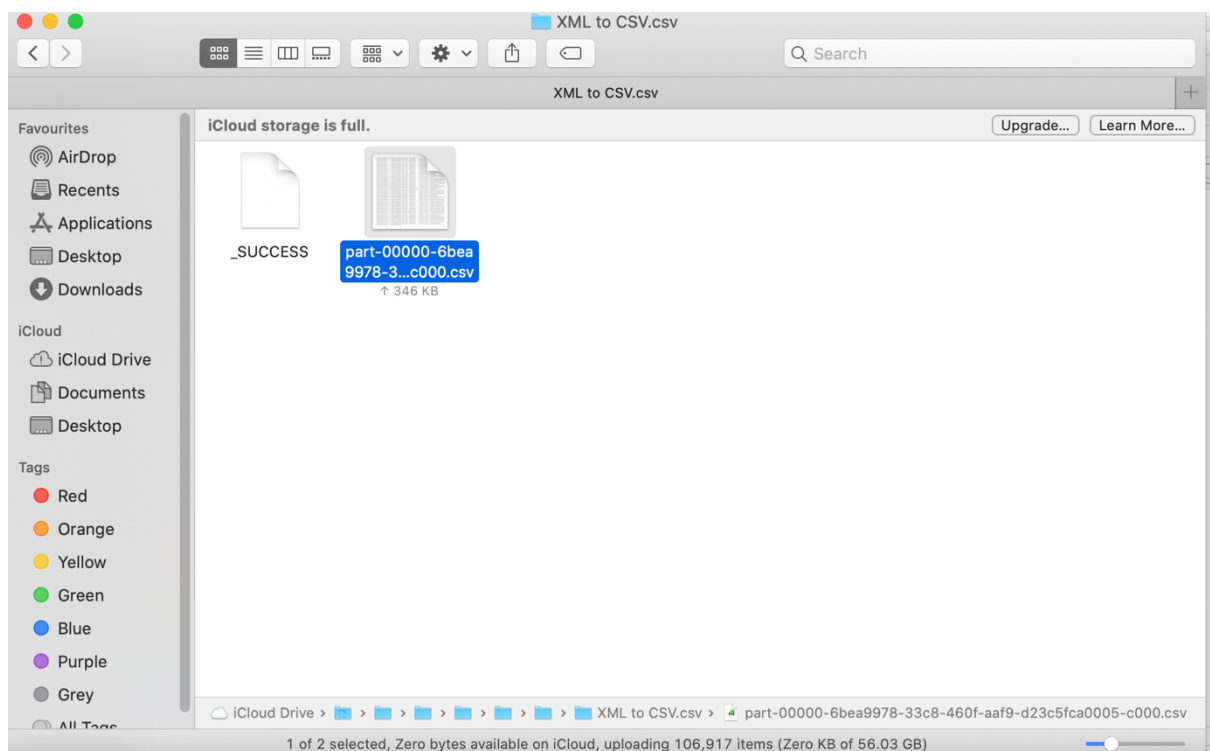
Cluster GUI showing your application name( "RollNo-3.png")

```

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|__address|__id|__position|__uid|application_sn|case_number|center|patent_|
expiration_date|patent_number|status|title|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|https://data.nasa...|407|407|2311F785-C00F-422...|13/033,085|KSC-12871|NASA Kennedy Spac...|
null|0|Application|Polyimide Wire In...|
|https://data.nasa...|1|1|BAC69188-84A6-4D2...|08/543,093|ARC-14048-1|NASA Ames Researc...|2015
-10-03T00:00:00|5694939|Issued|Autogenic-Feedbac...|
|https://data.nasa...|2|2|23D6A5BD-26E2-42D...|09/017,519|ARC-14231-1|NASA Ames Researc...|2017
-02-04T00:00:00|6109270|Issued|Multimodality Ins...|
|https://data.nasa...|3|3|F8052701-E520-43A...|10/874,003|ARC-14231-2DIV|NASA Ames Researc...|2024
-06-16T00:00:00|6976013|Issued|Metrics For Body ...|
|https://data.nasa...|4|4|20A4C4A9-EEB6-45D...|09/652,299|ARC-14231-3|NASA Ames Researc...|2017
-02-04T00:00:00|6718196|Issued|Multimodality Ins...|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 5 rows

```

Successful conversion from XML to dataframe (CSV) by df.show() ("RollNo-4.png")



Actual CSV with the success & file name showing the unique job id ("RollNo-5.png")