

You have 1 free story left this month. Sign up and get an extra one for free.

# How to make XGBoost available in the Spark Notebook



This is a step by step tutorial on how to install <u>XGBoost</u> (an efficient implementation of gradient boosting) on the <u>Spark Notebook</u> (tool for doing Apache Spark and Scala analysis and plotting graphs, similar to the Jupyter notebook).

If you don't have the Spark Notebook installed you can follow this <u>quick</u> <u>guide</u>.

# Step 1: Build XGBoost

For this step we need to clone the repository from github and build the project:

Next we need to go into the newly cloned repository and build the project:

```
cd xgboost
make -j4
```

For Mac users you have to do an additional step, before building with make:

```
cp make/config.mk ./config.mk
```

## Step 2: Create the XGBoost jar

Before we are able to build the jar we need to make sure we have JAVA\_HOME set up and pointing to the JDK directory. We also need maven installed.

Next we need to run maven in the xgboost directory and publish the artifact on your local maven repository:

```
mvn install
```

This command will also run some tests, but we can skip that to make the installation faster. If you wish to do that simply run this command instead:

## Step 3: Include the XGBoost jar into the Spark Notebook

As we now have the jar available in our local repository we can include it as a dependency in the notebook. In order to do that we need to go to the notebook and open the metadata window in the menu:

```
Edit -> Edit Notebook Metadata
```

A window will open containing a configuration JSON file. We will add the xgboost dependency in the customDeps property as shown in the screenshot below:

```
"customDeps": [
   "ml.dmlc:xgboost4j-spark:0.8-SNAPSH0T"
]
```

#### Edit Notebook Metadata

 $\times$ 

Manually edit the JSON below to manipulate the metadata for this Notebook. We recommend putting custom metadata attributes in an appropriately named sub-structure, so they don't conflict with those of others.

```
1 {
2  "id": "1281640a-8a9e-4604-998f-96bc8c0cbc44",
3  "name": "xgboost_tutorial",
```

```
"user_save_timestamp": "1970-01-01T01:00:00.000Z",
     "auto save timestamp": "1970-01-01T01:00:00.000Z",
     "language info": {
      "name": "scala",
      "file_extension": "scala",
      "codemirror mode": "text/x-scala"
9
10
11
     "trusted": true,
12
     "sparkNotebook": null,
    "customLocalRepo": null,
14
    "customRepos": null,
    "customDeps": [
15
    "ml.dmlc:xgboost4j-spark:0.8-SNAPSHOT"
16
17 ],
18
    "customImports": null,
    "customArgs": null,
19
    "customSparkConf": null,
    "customVars": null,
21
22
     "kernelspec": {
     "name": "spark",
23
                                                                       Cancel
```

Please note that at the moment this tutorial was created I was using xgboost version 0.8.

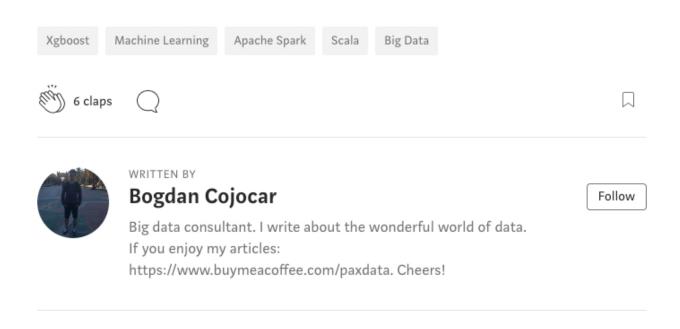
The last step is to restart the kernel of the notebook and we should have the dependency available. To restart the kernel go in the menu at the kernel tab and run restart:

```
Kernel -> Restart
```

Now we should be able to import XGBoost:



When we add import ml.dmlc.xgboost4j.scala.spark into a cell of the notebook and we run the code (in the menu Cell -> Run ) it should run succefully.



## More From Medium

How to run a realtime pipeline in AWS Kinesis using PySpark Structured Streaming

Bogdan Cojocar in Towards Data Science



Delta lake with Spark: What and Why?

Jyoti Dhiman in Towards Data Science



Predict Customer Churn using PySpark Machine Learning

Marvin Lüthe in Towards Data Science







Gülcan Öğündür in The Startup



PySpark Under the Hood: RandomSplit() and Sample() Inconsistencies Examined

Meltem Tutar in Udemy Tech Blog



Binary Classifier Evaluation made easy with HandySpark

Daniel Godoy in Towards Data Science



Deploy a Python model (more efficiently) over Spark

Schaun Wheeler in Towards Data Science



Want to Master Your Data? Here's Why You Should Care About Metadata



Nam Nguyen in Towards Data Science

### **Discover Medium**

Welcome to a place where words matter. On Medium, smart voices and original ideas take center stage with no ads in sight. <u>Watch</u>

## **Make Medium yours**

Follow all the topics you care about, and we'll deliver the best stories for you to your homepage and inbox. <u>Explore</u>

#### Become a member

Get unlimited access to the best stories on Medium
— and support writers while you're at it. Just
\$5/month. <u>Upgrade</u>

Medium

About

Help

Legal