# Setting up a standalone Spark KNIME cluster

#### Kyle Monahan, Statistics and Research Technology Specialist

*KNIME Analytics Platform* is a software for graphically representing analytics and computation, and *Spark* is an efficient engine for large-scale data processing. Connecting KNIME to Spark provides a more efficient computational platform for highly distributed data. This was developed in support of the Computational Methods for the Digital Humanities course.

**Goal**: This document provides a walkthrough of setting up the new KNIME Extension for local big data environments. This will provide a local option for using KNIME with a Spark cluster, without requiring a connection to the larger Spark cluster available at Tufts.

## Installing the Big Data environment

In order to produce a local Spark cluster, we will install the Big Data environment provided by KNIME. The steps to do this are below.

1. First, download the KNIME Analytics Platform nightly build. You can access the nightly builds by going to the link below:
   1. <https://www.knime.com/nightly-build-downloads>
   2. You can also access a backup of the tested nightly builds that worked for this install here: <https://tufts.box.com/v/KNIMENightlyInstallers>
   3. From there, download the KNIME Analytics Platform (nightly build) for your version of MacOS, Windows or Linux.
   4. Note that these are not production-quality versions of KNIME, and any critical workflows should be developed with the most stable release
2. Double-click on the installer to go through the process of installing the program. You can accept all defaults for install location, analytics collection, etc.
3. Once the installer is complete, you can go to the Windows search bar to find and open KNIME Analytics.
4. At startup, select your working directory to the default setting and continue to the main screen (Fig 1).
   1. You’ll note that it says “*You are using a nightly build*!” in large letters. This is entirely normal, and just reminding you this is a more experimental version.

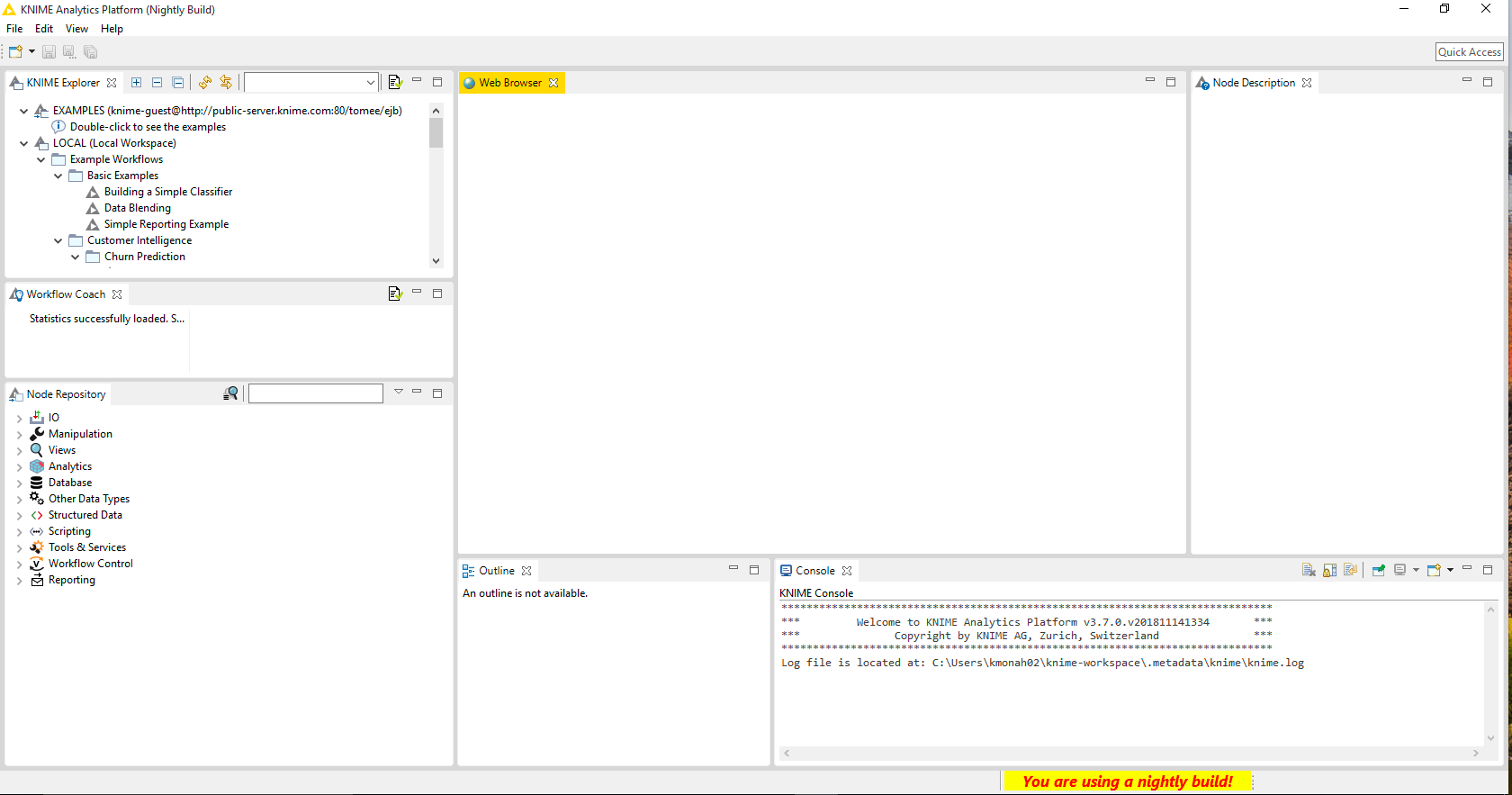


Figure 1 The view of KNIME after installing the nightly build.

1. You now have installed the KNIME Nightly build! In this build, we have new tools available for us to build and access a KNIME cluster locally. Next, we will install them.
2. To install these tools, go to *File > Install KNIME Extensions* (Fig. 2A).
3. Look up “local big data environment” in the search bar that appears (Fig 2B). Click on *KNIME Extension for Local Big Data Environments* and then click Next. Review the install details and click *Next* again. Review the license agreements, and if you agree, click “I accept” and then *Finish*. The packages will install, and this will take 10 - 15 minutes. When this is complete, you will need to restart KNIME.

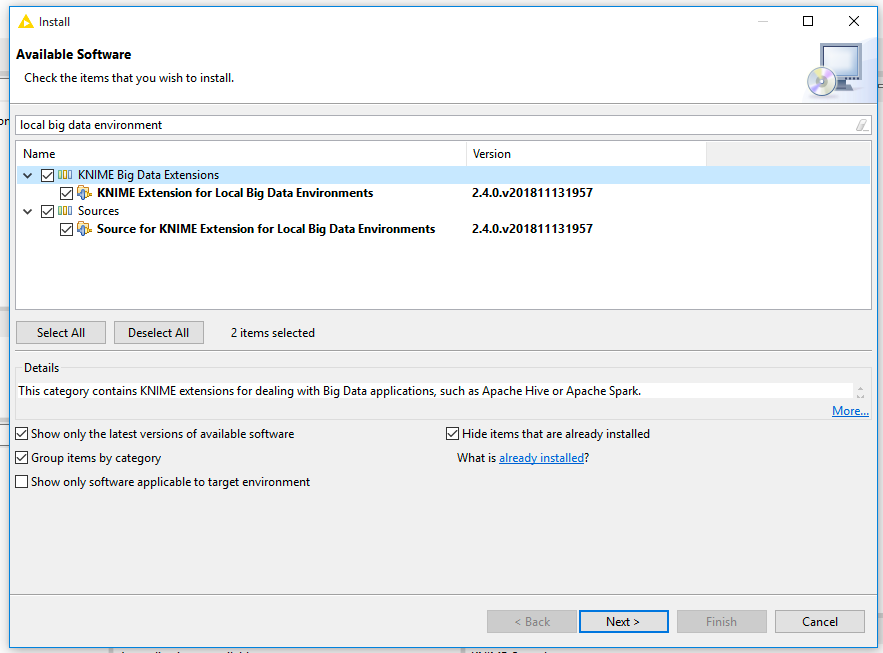
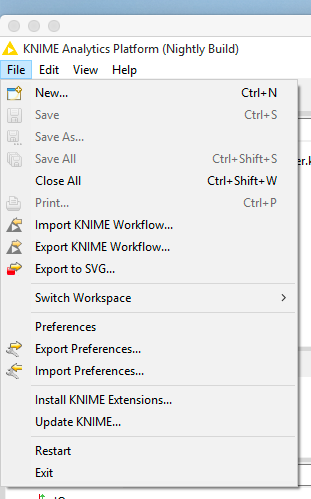


Figure 2 On left, Figure 2A shows the dropdown menu and on right Figure 2B shows the search menu.

1. We are now ready to connect to the local Spark cluster!

## Connecting to the local Spark cluster

1. So far, we have installed all the tools we need to work with Spark and KNIME. First, we will create new KNIME workflow called Big Data Spark Workflow. Click on the new workflow button () and title the workflow “Big Data Spark Workflow.”
2. Once the workflow is open, go to the Node Repository and search for Create Local Big Data Environment. Drag that node to the Workflow you created. Right click on the node, and go to Configure to set the number of threads to match for your machine, for this case, two is fine.

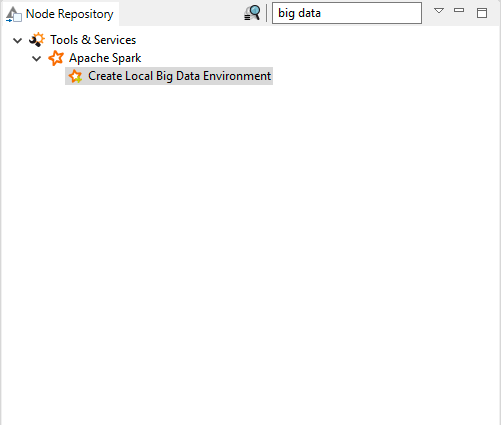


Figure 3 The location of the big data environment node.

1. Next, we need to copy the files in C:/Program Files/KNIME to a new folder that we create in the C:/ directory called C:/Program%20Files/KNIME. This is due to an error in the encoding of paths for modules in Windows 10 with KNIME, which I have reported.
2. After copying those files, restart KNIME and be sure to **Run as Administrator.**
3. Now, when you open KNIME, you’ll be able to run the Big Data Spark Workflow node. Congratulations! Our Spark cluster and Hadoop Distributed File System (HDFS) are available.

## Using a sample dataset

To help start users out with a good example, I’ve saved a workflow that creates a random data set containing a few million records (Data Generator, Fig 4), filters out the columns that we need (Spark Column Filter), and then performs a correlation between those columns (Spark Linear Correlation). As this is random data, we would expect this correlation to be low (which it is; r = 0.002).

Download this sample workflow here: <https://tufts.box.com/v/KNIMEBigDataWorkflow>

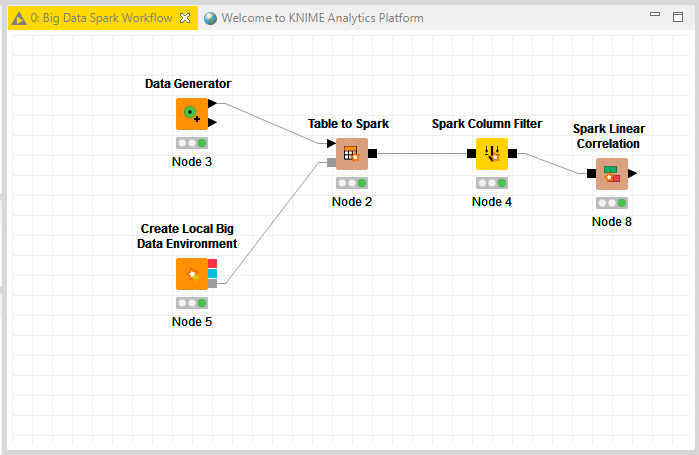


Figure 4 Sample workflow created for this tutorial

## Moving forward with a remote Spark cluster

There are a few security issues which must be addressed before we can mount to the actual Tufts Spark cluster. Instructions to connect to the Spark cluster are available here: <https://tufts.box.com/v/KNIME-SparkCluster-Tutorial>