

## STORING THE DATA

Before you can start querying the data, you need to load it into a triple store for persistent storage (Apache Jena TDB). With a smaller dataset, you could skip this step and query the data using another Apache Jena tool, ARQ. However, because the way ARQ works involves storing the dataset in RAM, this dataset is far too large and will exceed the capacity of any normal system (i.e., attempting to use ARQ will result in memory errors and may crash your machine).

### *Setting up Apache Jena*

If you do not already have Apache Jena installed on your machine, now is the time to do so. It is fairly simple, and there are a number of videos on the Web that can walk you through the steps. In a nutshell, the process looks like this:

1. Download the zip file appropriate for your machine from: <http://jena.apache.org/download/index.cgi>. For the purposes of this tutorial, you can choose the option *without* the Fuseki server.
2. Unzip the file – remember the directory where you are saving it.
3. Optional - add Environment Variable “JENA\_HOME” to your system path.

SEE: <https://jena.apache.org/documentation/tdb/commands.html#scripts>

### *Bulk loading the dataset*

The easiest way to get a large dataset into the TDB triple store without writing Java code is to use the bulk loader, which is invoked from the command line using “*tdbloader*”. NOTE: A newer version, “*tdbloader2*”, exists but is not available for Windows users.

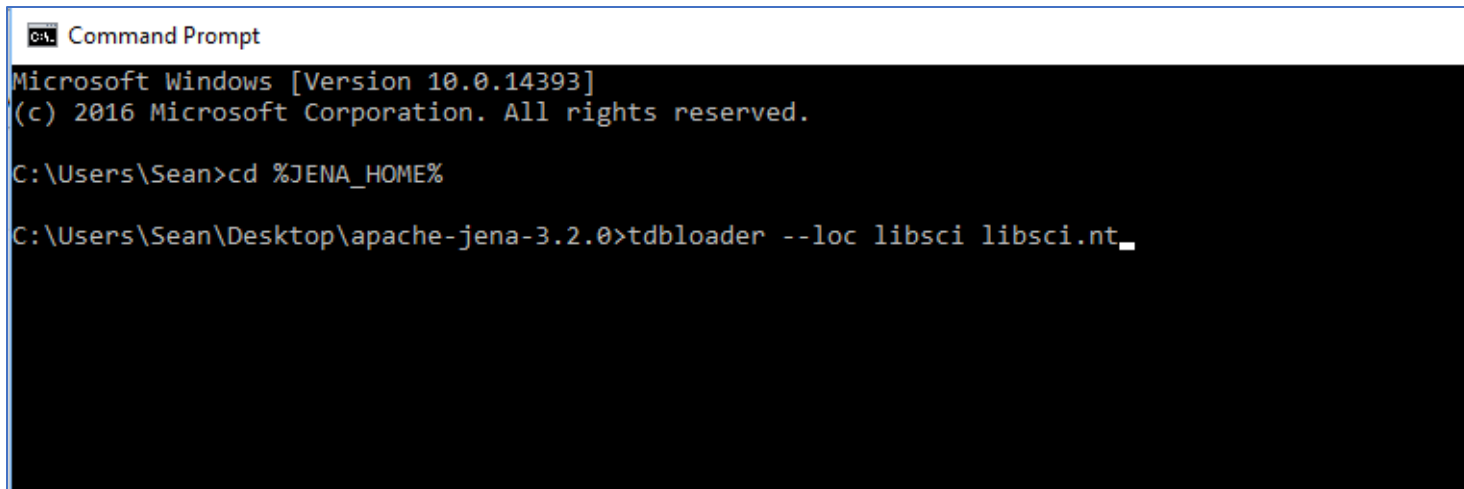
If you set the environment variable when installing Apache Jena, you can use the shortcut:

```
cd %JENA_HOME%
```

The bulk load can then be accomplished with one short command line:

```
tdbloader --loc libsci libsci.nt
```

This command line script is worth breaking down to highlight a few details (**Figure 2**).



```
Command Prompt
Microsoft Windows [Version 10.0.14393]
(c) 2016 Microsoft Corporation. All rights reserved.

C:\Users\Sean>cd %JENA_HOME%

C:\Users\Sean\Desktop\apache-jena-3.2.0>tdbloader --loc libsci libsci.nt_
```

**Figure 2- Bulk loading dataset from the command line**

The script starts with *“tdbloader”*, which tells Jena which utility to use for the task at hand- loading the data. The *“loc”* parameter is extremely important, because it creates the directory where the dataset will be stored. For its value, the user can supply any name they wish (e.g., *“libsci”*). This directory will need to be called upon when querying the data later.

Finally, the name of the file to be loaded (e.g., *“libsci.nt”*) finishes the command. No parameter name precedes it.

Due to the size of the dataset, the loading process is going to take between 60-80 minutes, depending on your machine. The command line will continuously issue updates on its progress as the contents of the file are loaded. Remember, this bulk load only needs to be performed one time, then the data will be *persistently stored* and can be queried at any time.