**Apache Drill Demo Project**



**Submitted and Created By: Paul W. Nichols**

Contents

[Statement of Purpose 3](#_Toc532737175)

[Scope 3](#_Toc532737176)

[Supported Platforms 3](#_Toc532737177)

[Required Software 3](#_Toc532737178)

[What is Apache Drill? 3](#_Toc532737179)

[How to Install and Use this software 4](#_Toc532737180)

[A. Running the native executable jar file (Out of the Box) 4](#_Toc532737181)

[1. You must have Java SDK 1.8.x installed on your machine 4](#_Toc532737182)

[2. You must install Apache Drill under a root folder called /export/programs 4](#_Toc532737183)

[3. You must copy the free-zipcode-database.csv file into the /export/programs/apache-drill-1.14.0/sample-data folder. 4](#_Toc532737184)

[4. You must start the apache-drill program using the drill-embedded start file script 4](#_Toc532737185)

[5. Start the apache-drill=embedded.jar 5](#_Toc532737186)

[B. Changing the jar file to meet your installation needs. 6](#_Toc532737187)

[1. Open the jar file in a zip tool (like archive manager on Limux) 6](#_Toc532737188)

[2. Go inside the META-INF folder and select the Classes folder. Right click and open the application.yml file 7](#_Toc532737189)

[3. Change the following sections 9](#_Toc532737190)

[4. Run the File as usual once changes have been made. 10](#_Toc532737191)

[C. Running from Source Code 10](#_Toc532737192)

[1. Running within your IDE 10](#_Toc532737193)

[Final Notes 11](#_Toc532737194)

# Statement of Purpose

**This project demo is submitted to show the versatility and usefulness of the Apache Drill application and utility.**

# Scope

**The scope of this project demo is to demonstrate the power of Apache Drill using simple examples coded in Java 8 and SpringBoot.**

# Supported Platforms

**This Demo was created on Linux Mint 19 and has only been tested on Linux. However, once you install the prerequisites and change the necessary settings in the /src/main/resources/application.properties file, it should run fine on Microsoft Windows, or Apple MAC OS.**

# Required Software

**In order to run this demo, you will need the following ser vices installed on your machine**

1. **Java JRE 1.8.x**
2. **Apache Drill version 1.14.0 (You can download it** [**here**](https://drill.apache.org/download/)**.)**

**You can download Apache Drill from the following URL:**

1. **The free-zipcode-database.csv file, which must be placed in the sample-data folder of the Apache Drill 1.14.0 installed folder[[1]](#footnote-2).**
2. **Apache Maven if you want to build the demo from source code.**
3. **We have also created a self-contained jar file that you can run once Apache Drill has been installed and the .csv file has been deployed to the sample-data folder.**

# What is Apache Drill?

**Apache Drill** is an [open-source](https://en.wikipedia.org/wiki/Open-source_software) [software framework](https://en.wikipedia.org/wiki/Software_framework) that supports data-intensive distributed applications for interactive analysis of large-scale datasets. Drill is the open source version of Google's [Dremel](https://en.wikipedia.org/wiki/Dremel_(software)) system which is available as an [infrastructure service](https://en.wikipedia.org/wiki/Infrastructure_as_a_service) called Google [***BigQuery***](https://en.wikipedia.org/wiki/BigQuery). One explicitly stated [design goal](http://wiki.apache.org/incubator/DrillProposal) is that Drill is able to scale to 10,000 servers or more and to be able to process petabytes of data and trillions of records in seconds. Drill is an Apache top-level project.[[1]](https://en.wikipedia.org/wiki/Apache_Drill#cite_note-1)

Drill supports a variety of [NoSQL](https://en.wikipedia.org/wiki/NoSQL) databases and file systems, including [HBase](https://en.wikipedia.org/wiki/Apache_HBase), [MongoDB](https://en.wikipedia.org/wiki/MongoDB), [MapR](https://en.wikipedia.org/wiki/MapR)-DB, [HDFS](https://en.wikipedia.org/wiki/Apache_Hadoop#HDFS), [MapR-FS](https://en.wikipedia.org/wiki/MapR_FS), [Amazon S3](https://en.wikipedia.org/wiki/Amazon_S3), [Azure Blob Storage](https://en.wikipedia.org/wiki/Microsoft_Azure#Storage_services), [Google Cloud Storage](https://en.wikipedia.org/wiki/Google_Storage), [Swift](https://en.wikipedia.org/wiki/OpenStack#Swift), [NAS](https://en.wikipedia.org/wiki/Network-attached_storage) and local files. A single query can join data from multiple datastores. For example, you can join a user profile collection in [MongoDB](https://en.wikipedia.org/wiki/MongoDB) with a directory of event logs in [Hadoop](https://en.wikipedia.org/wiki/Apache_Hadoop).

Drill's data store-aware optimizer automatically restructures a query plan to leverage the datastore's internal processing capabilities. In addition, Drill supports [data locality](https://en.wikipedia.org/wiki/Data_locality), if Drill and the datastore are on the same nodes.

The beauty of Apache Drill is that it removes reliance on MapReduce, PIG, and Spark and allows people with standard SQL database skills to take advantage of the Big Data world, without having to invest heavily in learning these technologies in order to use and query HDFS and NO-SQL datastores.

In addition, those who must work with standard and non standard datasets such as CSV, Avro, Parquet, JSON, XML, etc. can treat these divergent sets of data and file formats just as if they were RDBMS datasets using the SQL they already know.

For developers, the ability to treat structured and unstructured data stores as if they were residing in a structured RDBMS system, can greatly increase their productivity in querying this data, migrating this data, and joining disparate data types in a quick and efficient manner.

# How to Install and Use this software

Installation of this software is straightforward. It is a standard Java 8 console application using the SpringBoot Console application.

If you wish to run the self-contained jar file there are some things that you must do to run it “out of the box”. You can also customize it according to your needs and operating system as required.

I will cover the native executable jar file first and then move to the source code section.

## Running the native executable jar file (Out of the Box)

To use the native executable jar file (apache-drill-embedded.jar), the following pre-requisites must be met.

### You must have Java SDK 1.8.x installed on your machine

### You must install Apache Drill under a root folder called /export/programs

* 1. Windows: C:\export\programs\ apache-drill-1.14.0
  2. Linux: /export/programs/apache-drill-1.14.0
  3. MAC OS: /export/programs/apache-drill-1.14.0

### You must copy the free-zipcode-database.csv file into the /export/programs/apache-drill-1.14.0/sample-data folder.

### You must start the apache-drill program using the drill-embedded start file script

Using the command, bin/drill-embedded (see screenshot below)

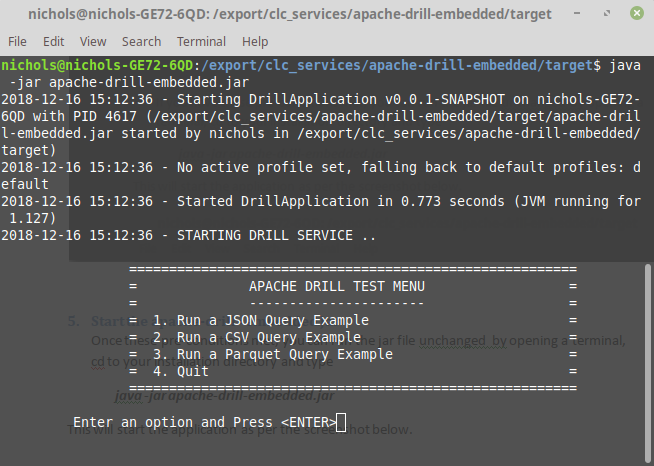
******

### Start the apache-drill=embedded.jar

Once these pre-conditions met, you can run the jar file unchanged by opening a terminal, cd to your installation directory and type

***java -jar apache-drill-embedded.jar***

This will start the application as per the screenshot below.

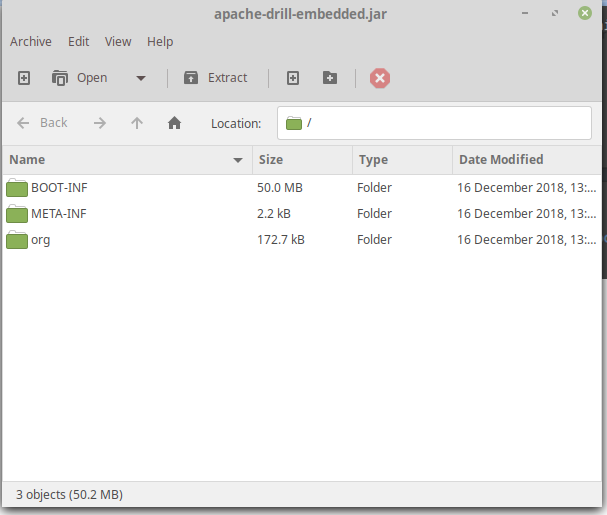


# Changing the jar file to meet your installation needs.

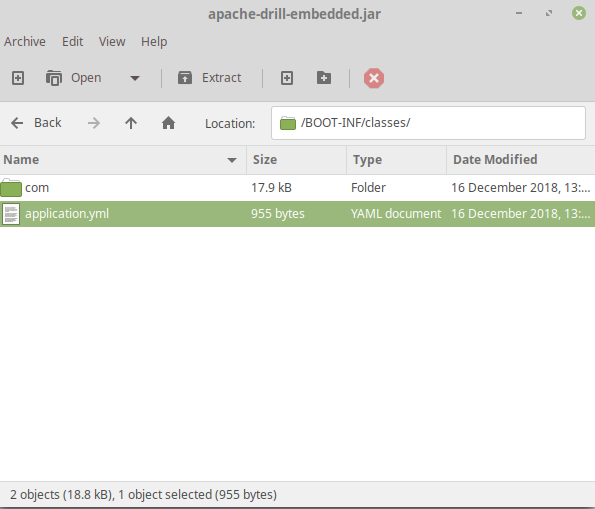
You can change the jar file to meet different installation directories by opening the jar file’s application.yml file and changing the actual directories to your installation directories.

This process is show below (This is on Linux, you will need to have appropriate tools per your operating system to change the application.yml file without deconstructing the jar file and rebuilding).

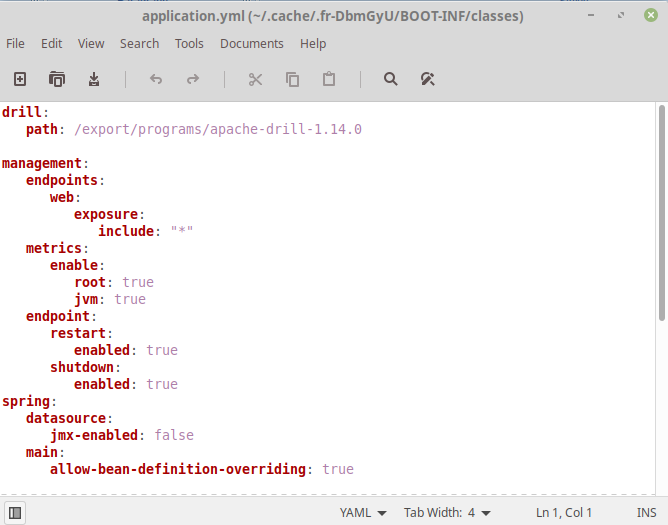
### Open the jar file in a zip tool (like archive manager on Limux)

****

### Go inside the META-INF folder and select the Classes folder. Right click and open the application.yml file



**Open the application.yml file for editing**

****

### Change the following sections

**datasource:**

**locations**:

**csv. `<Your apache drill installation directory/sample-data /free-zipcode-database.csv`**

**parquet: dfs. `<Your apache drill installation directory/sample-data /nationsMF/nationsMF.parquet`**

Once you have made the necessary changes, save the file and commit it to the same jar file per the tool you are using.

### Run the File as usual once changes have been made.

***java –jar apache-drill-embedded.jar***

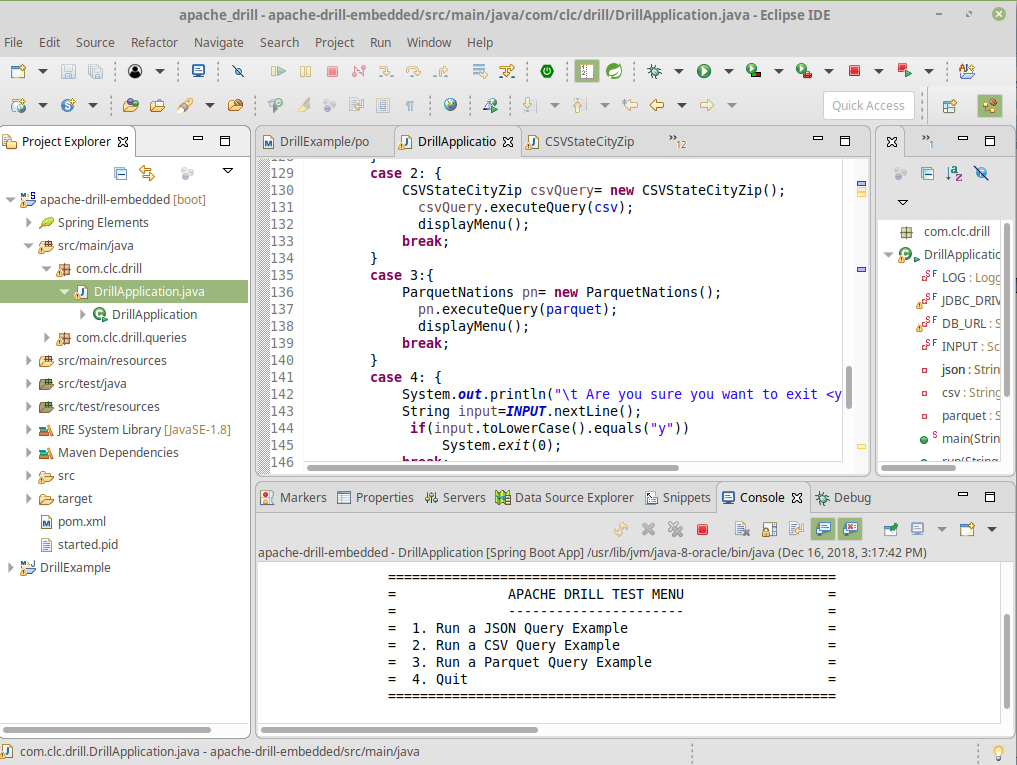
## Running from Source Code

We have tried to document the code with Java Docs that you can use and build. Look at the Java Docs / Comments sections to learn more about the source code.

### Running within your IDE

Using your favorite IDE, simply import the project into your IDE and use the Run As Java Application option.

The application is shown running in the Eclipse IDE below.



Feel free to modify the source code as you wish!

# Final Notes

We have only scratched the surface as to the power of Apache Drill. It is our hope that this Demo gets you started in enjoying this great tool to assist in your BIG DATA activities.

1. This would not be a requirement in a HDFS or Cloud repository. However, it is not the purpose of this demo to show how to configure Apache Drill. Installation and Configuration of Apache Drill can be found by consulting the Apache Drill Website ([https://drill.apache.org](https://drill.apache.org/)/) [↑](#footnote-ref-2)