Loading data into HDFS

UK Data Service

Author: UK Data Service
Created: April 2016
Version: 1

Contents

1. Introduction

The aim of this short guide is to provide detailed instructions of how to load a dataset from a PC into a Hadoop system.[1] In these instructions we will assume that the Hadoop system is running in a Hortonworks provided HDP (Hortonworks Data Platform) VM (Virtual Machine) Sandbox on the same PC. Details of how to get and install the Hortonworks HDP VM Sandbox are given in the Obtaining and downloading the HDP Sandbox guide available from the UK Data Service website. It doesn't really matter where the Hadoop system is running, it could be a cloud based system or an on a dedicated server, you only need to know the IP address of the Hadoop system and have permission to login as the Hive user.
In order to carry out these instructions, some software tools will be required, you may already have and be familiar with, or they may be completely new to you. We have assumed the latter, so we have included instructions on how they can be obtained and installed and when they are used, detailed instructions and screenshots are provided.

2. The tools you will need

In order to perform the necessary steps to load the file(s) following these instructions, you will need the following software tools (utility applications). They are all free and are easily downloadable from the Internet.

2.1. PuTTY

This tool allows you to access a remote system (in our case the Hadoop VM), login and be able to issue commands from a command line prompt. This is very similar to using the cmd application on a Windows system which is used to issue command line instructions to the PC. The ordinary is unlikely to need to use the command line in Windows, so it is possible that you have never seen it. You need to be able to issue commands directly to create directories and move files into the Hadoop system. The actual commands needed are detailed in the instructions below.
The software can be downloaded from PuTTY's website. PuTTY is a simple executable file (i.e. a self-contained program). Once you have downloaded it you can run it by simply double clicking the file. By default, it will download into your downloads folder, so you may wish to move it somewhere else before using it. It is only a small file so it could be put directly onto the desktop if required.

2.2. FileZilla or WinSCP

FileZilla or **WinSCP** are two tools known as FTP clients (File Transfer Programs). They both do the same job, so you only need to install one of them.
FileZilla can be downloaded from FileZilla's website and WinSCP from WinSCP's website; both

[1] There are other ways of doing these tasks via the provided web based tools, but in practice they have proved unreliable, particularly for large files.

of these programs require Administrator rights to install. In both cases you just need to double click the downloaded file and follow the installation instructions.
You need an FTP program to copy the datasets from the PC to the Sandbox VM. The actual procedures for accessing the Sandbox and transferring the files using FileZilla are in the instructions below.

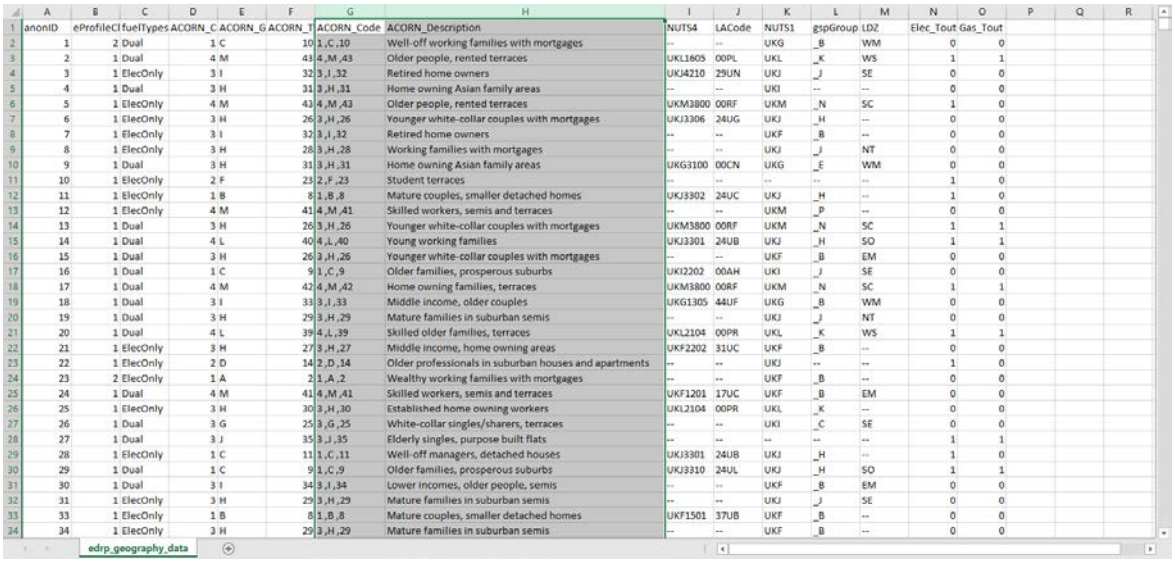
3. The data files we will be loading

For the purposes of this guide we will demonstrate the loading of two files available from the Energy Demand Research Project: Early Smart Meter Trials, 2007-2010, a set of trials on smart meter data available for download from the UK Data Service. To access the data, you must login/register with the UK Data Service. All users, including those outside the UK, can obtain a login – see our login and registration FAQs for more details.
After you have logged in, the files can be found by downloading the zip file. Once you have unzipped the folder, you will find several files two of which are edrp_gas.csv and another called edrp_geography_data.csv. The .csv suffix indicates that these files are in Comma Separated Values (CSV) format. Therefore the values for each column are separated from each other by a ‘,’. These are the two files which we will load into the Hadoop file system (HDFS).
The instructions are of course equally applicable to any other file(s) that you may wish to load. You would only need to change the filenames and the folder names where you choose to place them.

3.1. Initial editing of the Geography file

The Geography file is a small file which can easily be loaded into Excel. Before loading it into HDFS we are going to edit the file in Excel to remove some of the columns that we will not be using.

You can load the edrp_geography_data.csv file into Excel by double-clicking it in File Explorer.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	anonID	eProfileCl	fuelTypes	ACORN_C	ACORN_G	ACORN_T	ACORN_Code	ACORN_Description	NUTS4	LACode	NUTS1	gspGroup	LDZ	Elec_Tout	Gas_Tout			
2	1	2	Dual	1	C	10	1,C,10	Well-off working families with mortgages	--	--	UKG	_B	WM	0	0			
3	2	1	Dual	4	M	43	4,M,43	Older people, rented terraces	UKL1605	00PL	UKL	_K	WS	1	1			
4	3	1	ElecOnly	3	I	32	3,I,32	Retired home owners	UKJ4210	29UN	UKJ	_J	SE	0	0			
5	4	1	Dual	3	H	31	3,H,31	Home owning Asian family areas	--	--	UKI	--	--	0	0			
6	5	1	ElecOnly	4	M	43	4,M,43	Older people, rented terraces	UKM3800	00RF	UKM	_N	SC	1	0			
7	6	1	ElecOnly	3	H	26	3,H,26	Younger white-collar couples with mortgages	UKJ3306	24UG	UKJ	_H	--	0	0			
8	7	1	ElecOnly	3	I	32	3,I,32	Retired home owners	--	--	UKF	_B	--	0	0			
9	8	1	ElecOnly	3	H	28	3,H,28	Working families with mortgages	--	--	UKJ	_J	NT	0	0			
10	9	1	Dual	3	H	31	3,H,31	Home owning Asian family areas	UKG3100	00CN	UKG	_E	WM	0	0			
11	10	1	ElecOnly	2	F	23	2,F,23	Student terraces	--	--	--	--	--	1	0			
12	11	1	ElecOnly	1	B	8	1,B,8	Mature couples, smaller detached homes	UKJ3302	24UC	UKJ	_H	--	1	0			
13	12	1	ElecOnly	4	M	41	4,M,41	Skilled workers, semis and terraces	--	--	UKM	_P	--	0	0			
14	13	1	Dual	3	H	26	3,H,26	Younger white-collar couples with mortgages	UKM3800	00RF	UKM	_N	SC	1	1			
15	14	1	Dual	4	L	40	4,L,40	Young working families	UKJ3301	24UB	UKJ	_H	SO	1	1			
16	15	1	Dual	3	H	26	3,H,26	Younger white-collar couples with mortgages	--	--	UKF	_B	EM	0	0			
17	16	1	Dual	1	C	9	1,C,9	Older families, prosperous suburbs	UKI2202	00AH	UKI	_J	SE	0	0			
18	17	1	Dual	4	M	42	4,M,42	Home owning families, terraces	UKM3800	00RF	UKM	_N	SC	1	1			
19	18	1	Dual	3	I	33	3,I,33	Middle income, older couples	UKG1305	44UF	UKG	_B	WM	0	0			
20	19	1	Dual	3	H	29	3,H,29	Mature families in suburban semis	--	--	UKJ	_J	NT	0	0			
21	20	1	Dual	4	L	39	4,L,39	Skilled older families, terraces	UKL2104	00PR	UKL	_K	WS	1	1			
22	21	1	ElecOnly	3	H	27	3,H,27	Middle income, home owning areas	UKF2202	31UC	UKF	_B	--	0	0			
23	22	1	ElecOnly	2	D	14	2,D,14	Older professionals in suburban houses and apartments	--	--	UKJ	--	--	1	0			
24	23	2	ElecOnly	1	A	2	1,A,2	Wealthy working families with mortgages	--	--	UKF	_B	--	0	0			
25	24	1	Dual	4	M	41	4,M,41	Skilled workers, semis and terraces	UKF1201	17UC	UKF	_B	EM	0	0			
26	25	1	ElecOnly	3	H	30	3,H,30	Established home owning workers	UKL2104	00PR	UKL	_K	--	0	0			
27	26	1	Dual	3	G	25	3,G,25	White-collar singles/sharers, terraces	--	--	UKI	_C	SE	0	0			
28	27	1	Dual	3	J	35	3,J,35	Elderly singles, purpose built flats	--	--	--	--	--	1	1			
29	28	1	ElecOnly	1	C	11	1,C,11	Well-off managers, detached houses	UKJ3301	24UB	UKJ	_H	--	1	0			
30	29	1	Dual	1	C	9	1,C,9	Older families, prosperous suburbs	UKJ3310	24UL	UKJ	_H	SO	1	1			
31	30	1	Dual	3	I	34	3,I,34	Lower incomes, older people, semis	--	--	UKF	_B	EM	0	0			
32	31	1	ElecOnly	3	H	29	3,H,29	Mature families in suburban semis	--	--	UKJ	_J	SE	0	0			
33	33	1	ElecOnly	1	B	8	1,B,8	Mature couples, smaller detached homes	UKF1501	37UB	UKF	_B	--	0	0			
34	34	1	ElecOnly	3	H	29	3,H,29	Mature families in suburban semis	--	--	UKF	_B	--	0	0			

edrp_geography_data

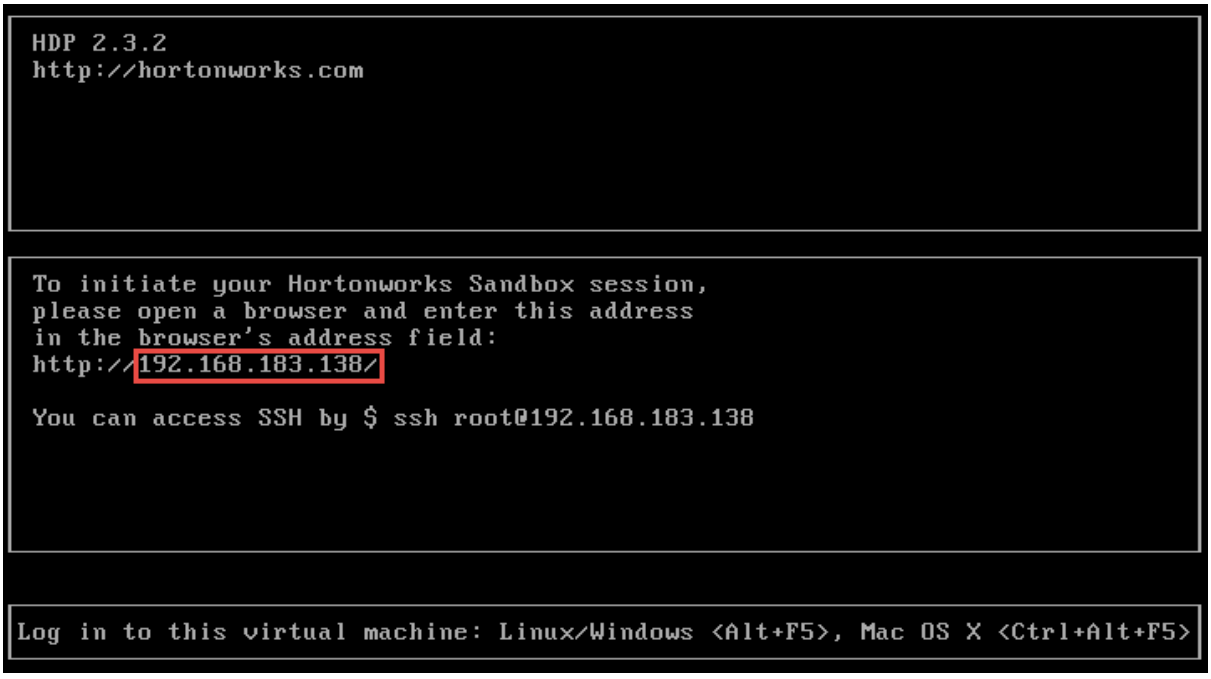
The columns we are going to delete are ACORN_Code and ACORN_Description. Simply select the two columns, right mouse click and select *Delete* from the context menu which appears.

Neither of these columns are needed for the analysis we intend to do. The ACORN description is just a description of the ACORN_Code and the ACORN_Code is just the concatenation of the ACORN Category, Group and Type columns to the left.

4. Detailed Instructions

4.1. Starting the Sandbox

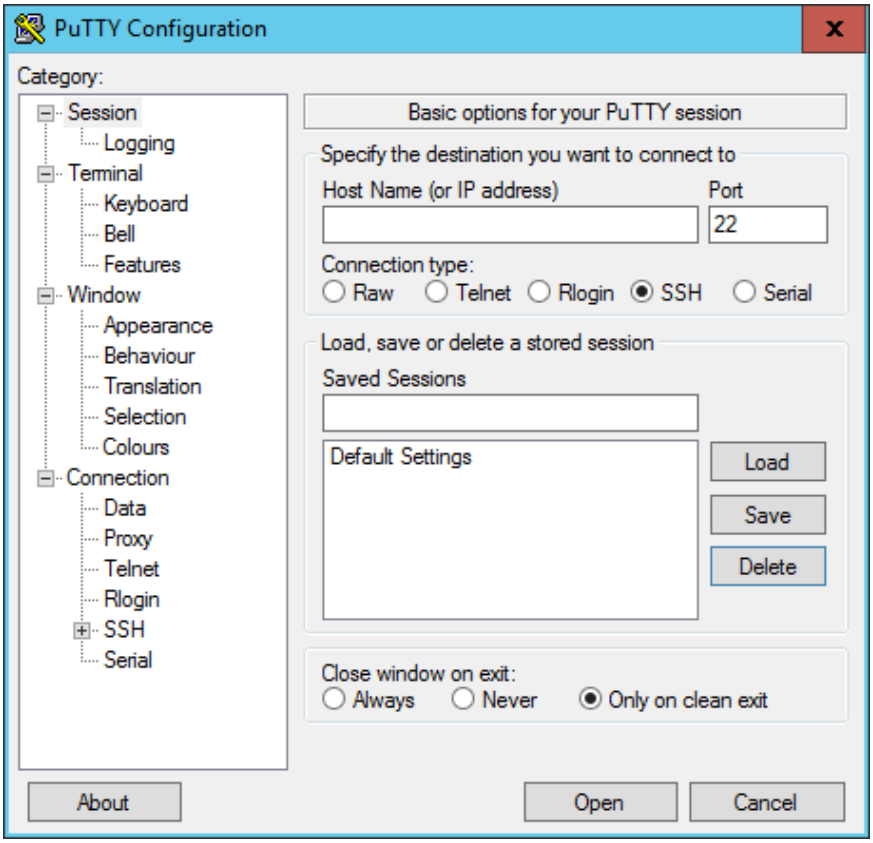
Before you can transfer files to the Sandbox VM you need to ensure that it is running. Details of how to do this are included in the Installing the Sandbox guide. The final screen of the load process will look something like this.



The IP address of the Sandbox is highlighted in the red box.

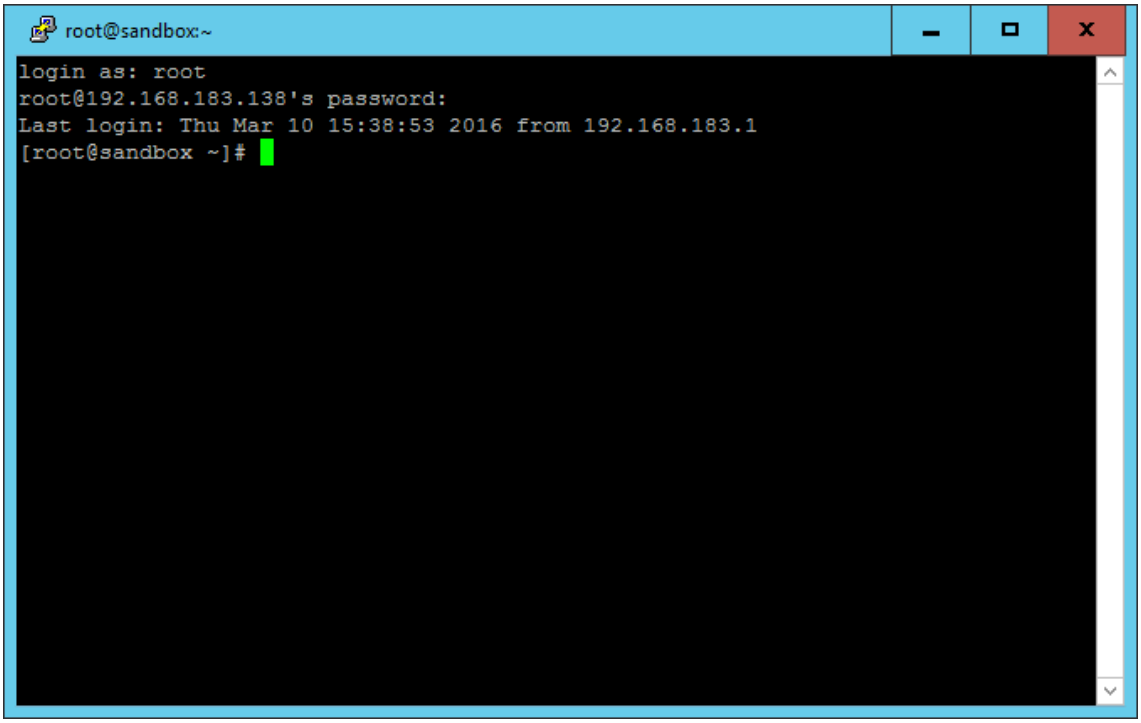
4.2. Run PuTTY (to login as root)

This step is only needed to change the password of the Hive account. We need to login as the Hive user in later steps and the password for the Hive account in the installed Sandbox is not known. The change is permanent so you only have to run this step once.
When you run PuTTy, the initial dialog will look like this:



In the Host Name (or IP address) box, type in the IP address of your Sandbox VM and click the open button.
If this is the first time you have used PuTTY to access your Sandbox you will get a warning message querying whether or not you are connecting to the machine you are, in this case we are, so you can click the Yes button. A new window will then open and there will be a login prompt as shown below.

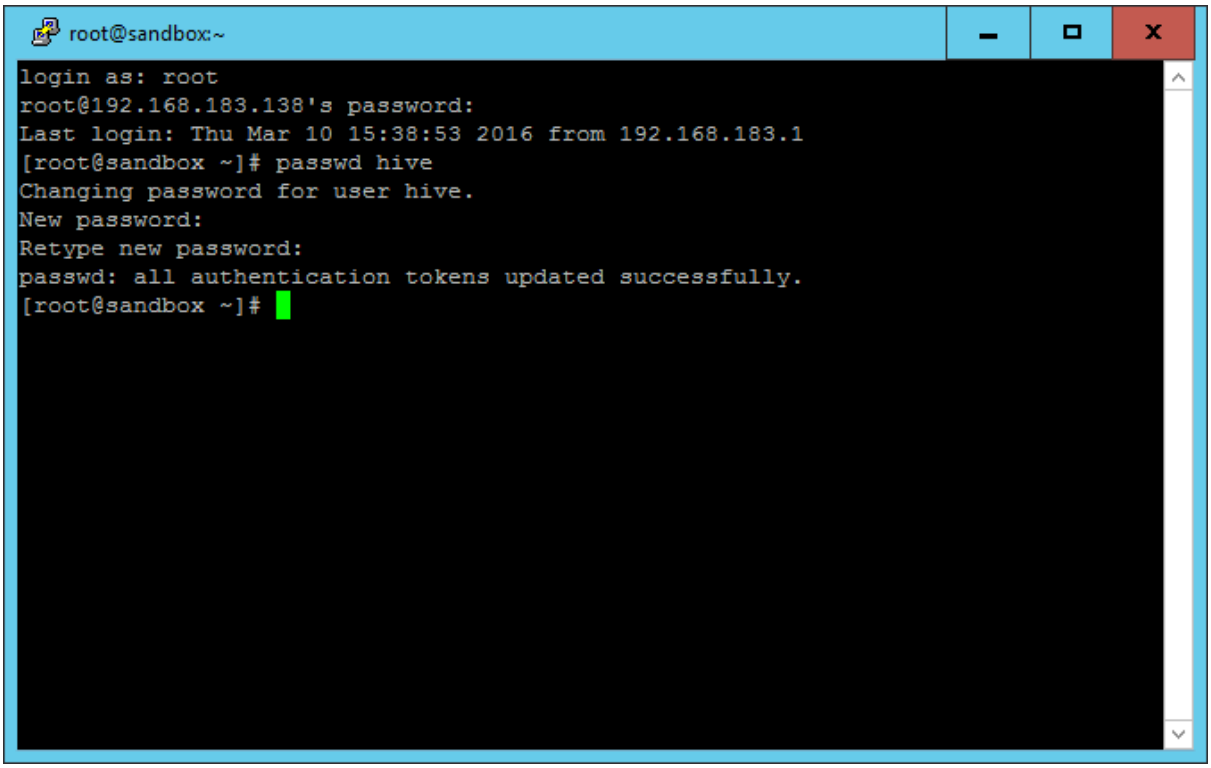
In this case we need to login as the user root. In a Linux system (which the Sandbox is based on) the root user account is the Superuser, which is allowed to issue all commands, like changing the passwords of other users. After you type in root and hit enter, you will be prompted for a password. The default (initial) password for the root user is 'hadoop'. Again, if this is the first time you have tried to login as root, you will be prompted to change the password for the root user. You can pick your own password at this point. The sequence is; you need to provide the current password ('hadoop') and then provide the new password and then confirm the new password. After this you will be left with a normal Linux command line prompt like this:



The only reason we need to login as root is to change the password for the hive user account. To do this we use the following command:

```
passwd hive
```

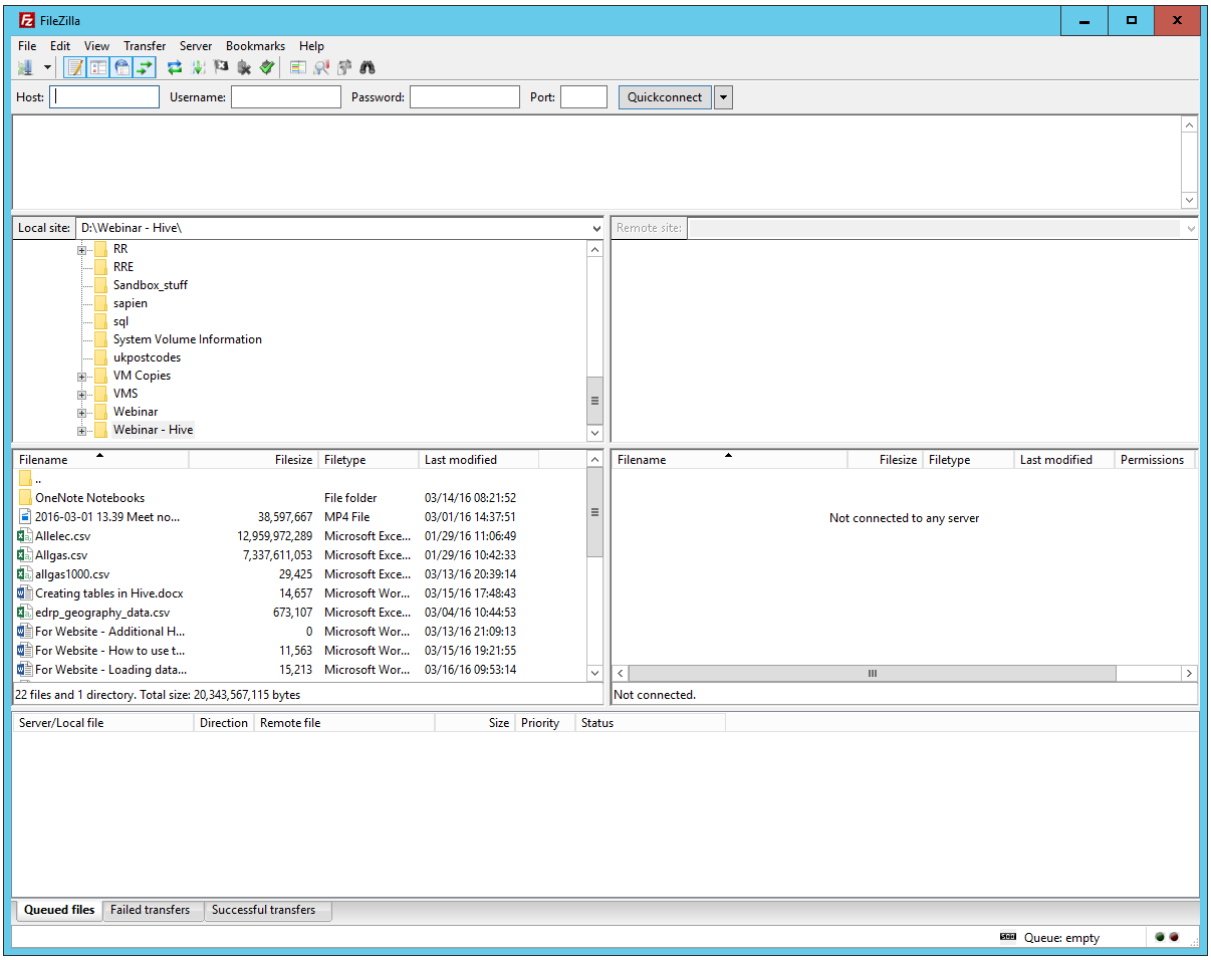
You will again be asked to provide a new password and retype it to confirm. Your screen will now look something like this:



(when you type the password it doesn't appear on the screen)
Having done this we are finished with the root user account and can close the PuTTY window by either using the red cross in the top right of the window or typing exit in the command line.

4.3. Run FileZilla to transfer the files from the Desktop to the Hive user in the Sandbox

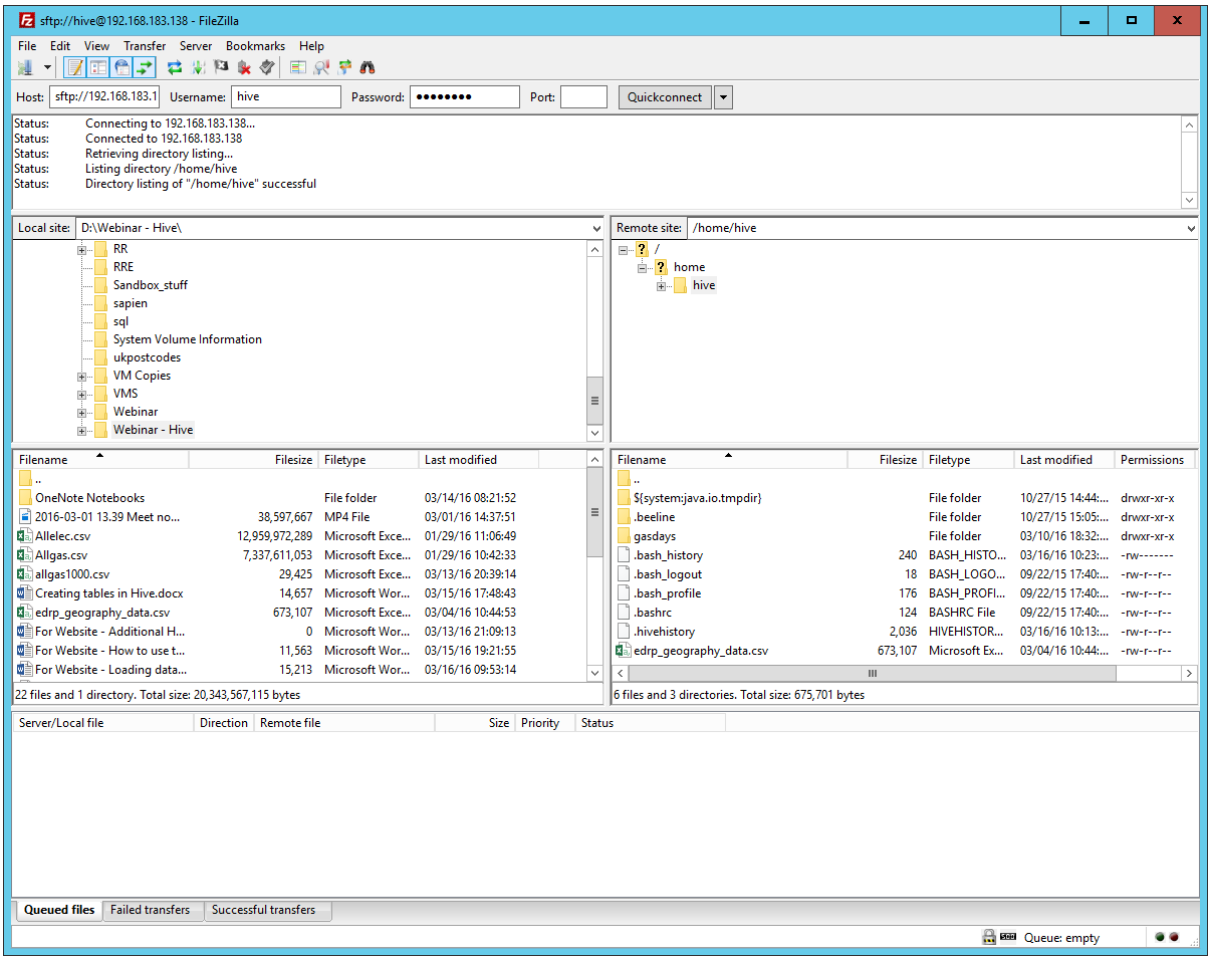
For these instructions I am going to use FileZilla, but the process of using WinSCP is similar. When you start Filezilla, the initial screen will look something like this:



The Host, Username, Password and Port boxes at the top are where you will fill in details of the system you wish to access.
In our case we need to type the following in these boxes:

Host : The IP address of the Sandbox
Username : hive
Password : the password for the hive user account
Port : 22 (this is always the same value)

You then click on *Quickconnect* (at the top by the port textbox). FileZilla will then connect to the Sandbox and login to it using the hive user account and password. The display will change to something like the screenshot below.



The left two panels in the middle work like File Explorer in Windows. You can navigate to drives and directories in the top pane and the lower panes shows the files in the selected directory. The two middle panes on the right behave in exactly the same way, except they show directories and files in the Sandbox. Because you logged into the Sandbox as the hive user, it is the home directory of the hive user which is displayed here. This where we want to copy the files to.

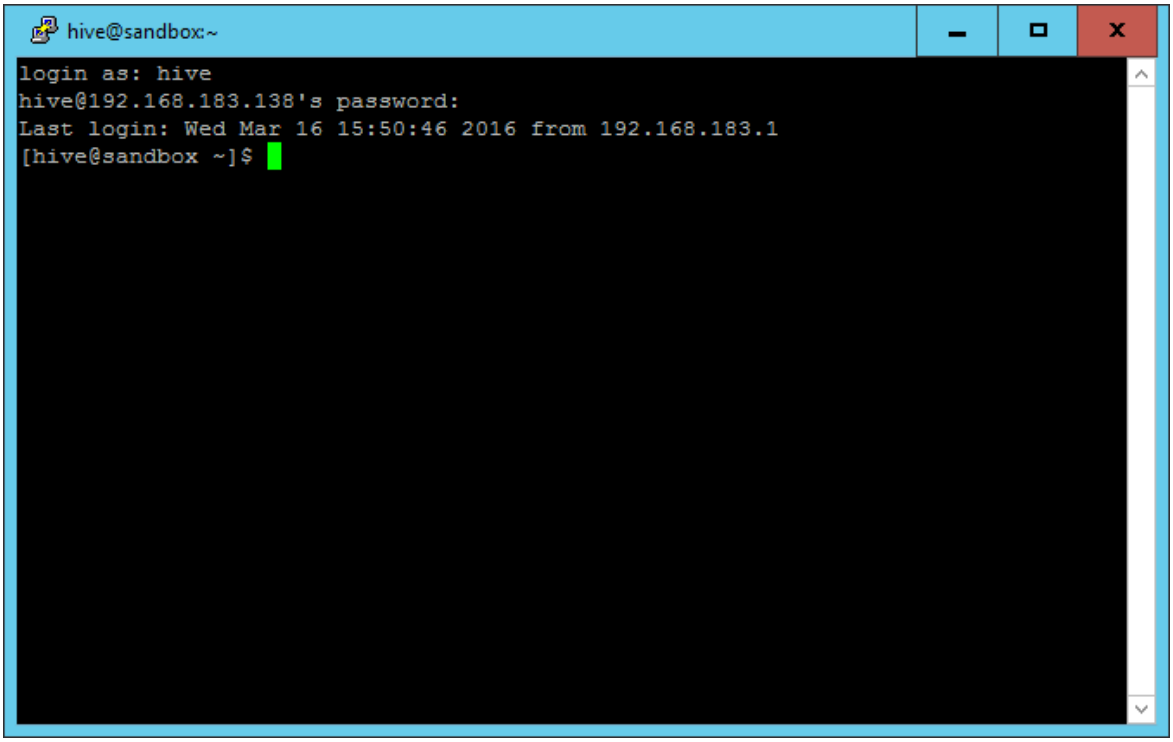
In FileZilla to copy a file you just need to double click the file. Double clicking on a file in the Windows pane will copy it to the Sandbox and vice versa. (In WinSCP you have similar panes but rather than double-clicking you use a drag-drop operation).

So all you have to do is navigate to the files you wish to copy in the left hand panes and double-click them. The screenshot above shows the edrp_geography_data.csv already copied to the Sandbox. The edrp_gas.csv file is 6.8Gb and takes several minutes to copy. During a copy action, the progress can be seen at the bottom of the Window.

Once the files have been copied you can close FileZilla.

4.4. Run PuTTY (to login as Hive)

Now that the hive password has been changed to something we know, we can use PuTTY to login to the Sandbox as the Hive user.
The process is the similar to before; run PuTTY, provide the Sandbox IP address, at the login prompt type hive as the user and provide the newly set password for Hive.



Once you have logged in as hive, you need to run the following set of commands. The # symbol denotes a comment rather than an actual command so you don't actually need to type them in.

```
# Create directories in hdfs for the data files
# command 1
hdfs dfs -mkdir /user/hive/geography
# command 2
hdfs dfs -mkdir /user/hive/energy
# to check that the directories have been created OK
# command 3
hdfs dfs -ls /user/hive
# check that your files to be loaded into hdfs are in the right
place
# command 4
ls -l
#Move the datasets into hdfs (you don't want copies left lying
#around using large amounts of space on the sandbox)
# command 5 -  Type in as a single line!!!!
hdfs dfs -moveFromLocal ~/edrp_geography_data.csv
/user/hive/geography
# command 6
dfs dfs -moveFromLocal ~/Allgas.csv /user/hive/energy
```

Commands 1 and 2 create directories in HDFS, which is the file system within Hadoop. These commands do not return any information; you will just see the normal prompt when they complete
Command 3 should show you that your directories have been indeed been created
Command 4 is just a check that the files that you want to move are in the local (i.e. not Hadoop) folder.
Commands 5 and 6 perform the actual moving of the files from the local file system to the Hadoop file system (HDFS). The files are moved rather than copied so as to save space in the VM. By default, the total size of the Sandbox VM is only 50GB, not all of which is available to you. As the files were only placed in the home directory of the hive user as a staging area before moving them to HDFS, there is no benefit in leaving copies of them there which just reduce the amount of space left for you to use in HDFS.

To check that the files have been moved into the HDFS system, you can run the following two commands:

```
hdfs dfs -ls /user/hive/geography
hdfs dfs -ls /user/hive/energy
```

In each case you should see a single file being listed.

The files have now been moved into HDFS. You can close the PuTTY session.

5. Next Steps

Now that you have data in your Sandbox, you are ready to perform some manipulation and analysis on it. The following guide, available from the UK Data Service website, contains some example HiveQL queries:

- HiveQL example queries

April 2016

T +44 (0) 1206 872143
E help@ukdataservice.ac.uk
W ukdataservice.ac.uk

The UK Data Service provides the UK's largest collection of social, economic and population data resources

UK Data Service

E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL