ukdataservice.ac.uk

Obtaining and downloading the HDP Sandbox

UK Data Service

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

Contents

1. What is the Hortonworks HDP Sandbox VM?

Hortonworks is a commercial company which specialises in data platforms based on open source software for big data, in particular Hadoop. HDP is an acronym for the Hortonworks Data Platform which is an implementation of a Hadoop cluster (many computers working together) and a range of associated big data products which run in the Hadoop environment.

A Sandbox is a general term, not used exclusively in IT environments, to represent an environment which is safe - safe in the sense that no matter what you do in the Sandbox, it will not affect anything outside of the Sandbox. If something goes wrong in the Sandbox, you can simply delete it and re-create a new pristine version to start again in. For the rest of this guide the Hortonworks HDP Sandbox VM will simply be referred to as the 'Sandbox'.

VM stands for Virtual Machine. The term machine refers to any computer, whether it is a PC like your desktop or large server at a data centre. Virtual in the way it is used here refers to simulation. A Virtual Machine is a complete PC which is simulated entirely by software and data files within your real PC. You start (power on) your VM by running a Virtualisation application (covered in later section) on your real PC and telling it what data files (which represent your VM) to use. When you have finished using the VM, you simply close the Virtualisation application and the VM stops running and ceases to exist, but of course the data files which represent the VM do still exist, so you can start the VM again any time you want.

2. Why do I want the Sandbox?

Although the Sandbox isn't really a Hadoop cluster with thousands of computers, it still behaves as if it were in two very important respects:

1. It will allow you to process datasets (files) far larger than you could in a normal desktop application. The actual size of dataset you could process is of course still restricted to what will fit into the Sandbox. But you may just want to process these big datasets so as to reduce them in size and then move them back to your desktop application. The actual data capacity of the Sandbox will depend on the total amount of data stored as well as how you process it. If you aim to load no more than 20Gb of data and remember to delete unwanted files and tables regularly, you should be OK

2. In a Hadoop cluster all of the complexities of storing and processing big datasets can be hidden from the end user. When a user stores a file in the Hadoop file system, it is just a file stored in a directory. When a user writes a query to explore or manipulate the data[1] and runs it, they don't need to know the internal processes which actually take place in order to return the results. The Sandbox behaves exactly like a Hadoop

[1] For example, an SQL-like query in Hive

cluster, but with only one computer in the cluster. So, although this means jobs (your queries) will run a lot slower than if they were running on a Hadoop cluster, it also means that commands you use to move your files around and the queries you write to process the data are exactly the same as you would have written had you been using a real Hadoop cluster with thousands of computers. This makes the Sandbox an excellent training ground for learning about big data techniques and software products.

3. What hardware do I need to run the Sandbox?

The hardware specification needed to run the Sandbox is provided in the Hortonworks Installation guides (see later section). However, in brief, you will need a PC/Laptop with:

- A minimum of 8GB of Ram (the more the merrier)
- Up to 50GB of free Hard disc space (The initial VM files are smaller than this, but they grow as you move data into the VM)
- A CPU which supports Virtualisation; in practice almost all processors in PC or Laptops (Not tablets) less than 5 years old will support virtualisation

4. What software do I need to run the Sandbox?

The Sandbox VM is essentially a set of data files (discussed in next section) which need a Virtualisation application to run them. A Virtualisation application is an application (program) which runs on your desktop and processes the VM data files to create a VM which behaves as a complete PC. You essentially have two choices of Virtualisation software both of which are available as free downloads for the PC.

4.1. VM Player

The VM Player is provided by VMWare. You can download the software from http://www.vmware.com/products/player/. Installation just involves double-clicking the downloaded file and following the instructions. You will however require Administrator rights on your machine in order to complete the install.

4.2. Virtualbox

The Virtualbox is provided by Oracle. You can download the software from https://www.virtualbox.org/. Documentation is also available from the site, although installation just involves double-clicking the downloaded file and following the instructions. You will however require Administrator rights on your machine in order to complete the install.

Although these products behave in a very similar manner, the **VM Player** product seems to provide more reliable networking facilities out of the box, so I would prefer to use that. The networking is all set up for you automatically and is needed to allow you to connect (talk) to your VM from your PC.

5. How do I get a copy of the Sandbox?

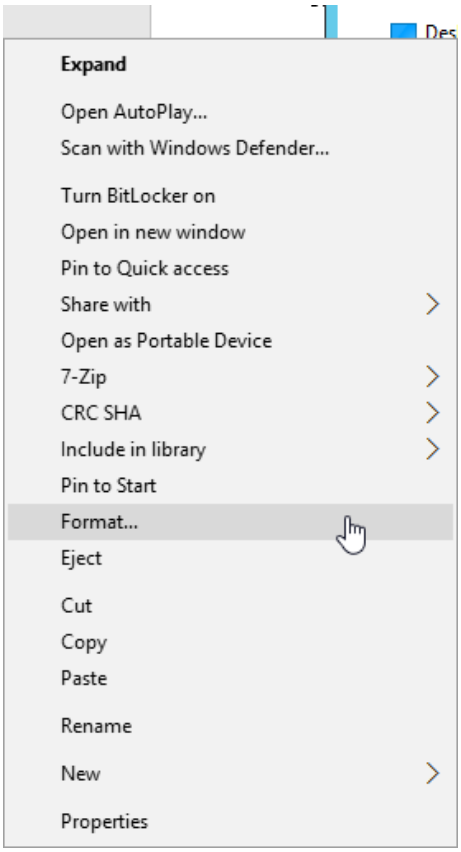
The Sandbox is packaged by Hortonworks as a single file. Hortonworks provide a version for either VM Player or Virtual box. You can download the file you need for your virtualisation software from the Hortonworks website. At the time of writing, the latest version of the Sandbox is 2.4.0. This has proved to be a bit unpredictable with our testing and so we would recommend downloading the earlier version 2.3.2., which is available from the Archive page.

5.1. Download sizes and times

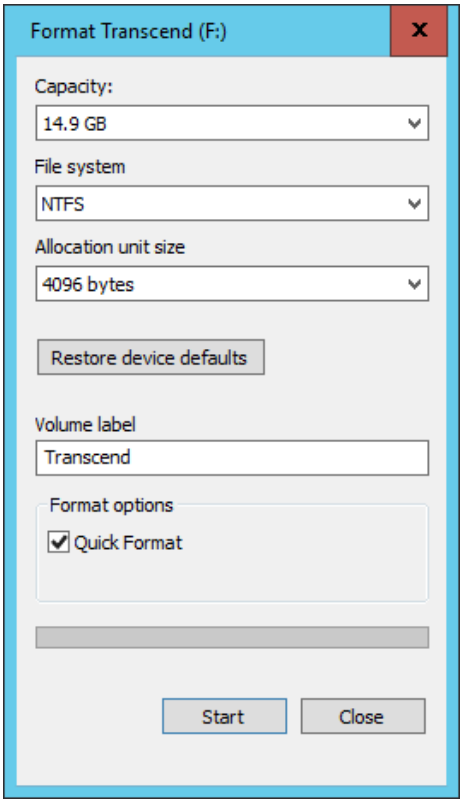
In both cases the files are about 9Gb in size - you may need to take this into account as downloading will take some time. The actual time is not just a function of your ISP (Internet Service Provider) download speeds but also that being offered by Hortonworks' provider as well as on general loading of the internet at the time. Expect it to take several hours.

5.2. Copying the data file

Once downloaded, you can treat the file much like any other. Copying it to an external hard drive should not present any problem, however if you want to copy the file onto a USB memory stick (16Gb at least), then you will probably have to re-format the memory stick first. This will wipe out any existing data on it. The reason you will need to reformat it is because by default memory sticks have been pre-formatted using a file system type known as FAT32. FAT32 can only deal with individual file sizes up to 4GB and this clearly isn't going to be enough.
In Windows you can re-format a USB stick by selecting it in the left hand pane of the File Explorer, right mouse click and select format as shown below:



In the format window change the File System from FAT32 to NTFS:



And then start the format process. Quick format is quite OK.

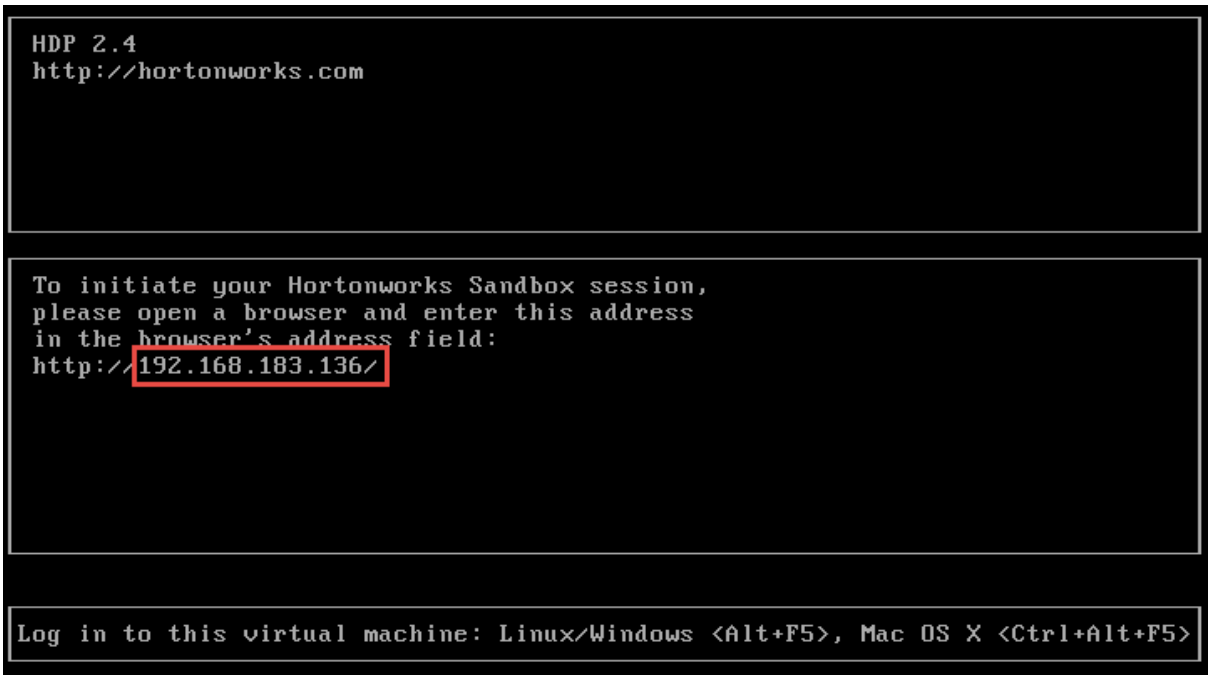
6. How do I install the Sandbox?

On the same Hortonworks webpage from which you downloaded the Sandbox file, you can also download installation guide documents provided by Hortonworks. Again there is a separate guide for each virtualisation software product. These documents are quite comprehensive and easy to follow. We will not make any attempt to reproduce them here.

7. How do I run the Sandbox and check that it is working?

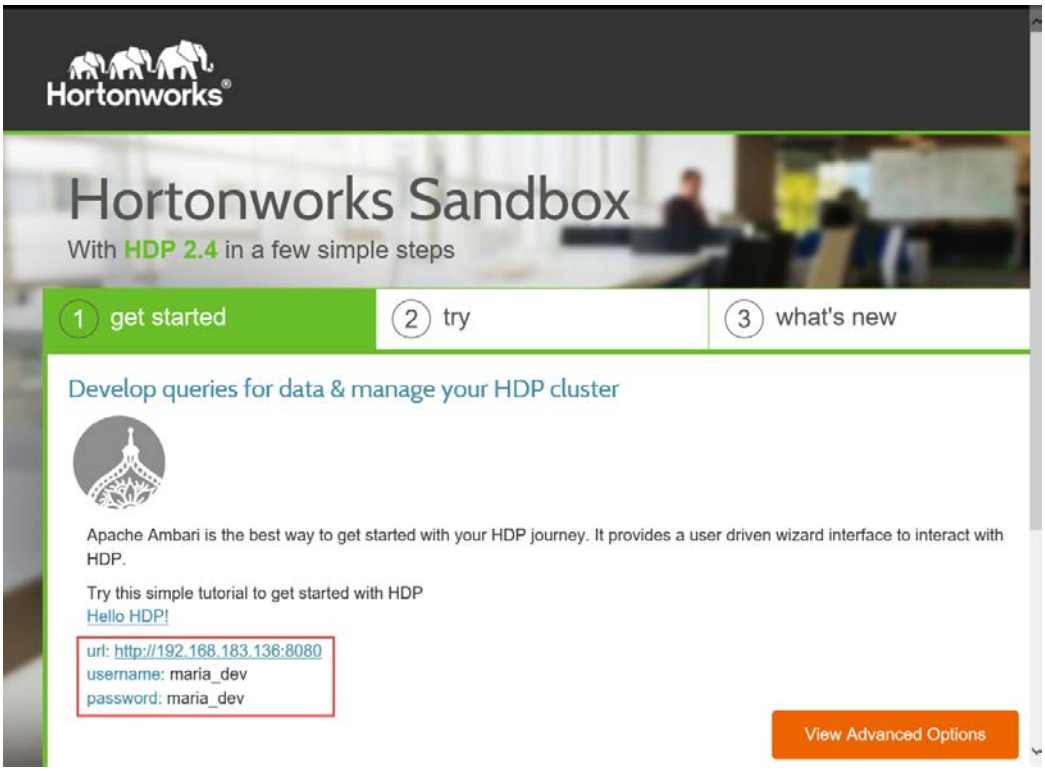
7.1. Run the Sandbox

When have completed the install of the Sandbox by following the Hortonworks instructions, it will start to run straight away. It will take about 3-4 minutes to load completely (depending on the size of your PC/Laptop). When it has finished loading you should be left with a screen (within the window of your virtualisation application) which looks something like this.

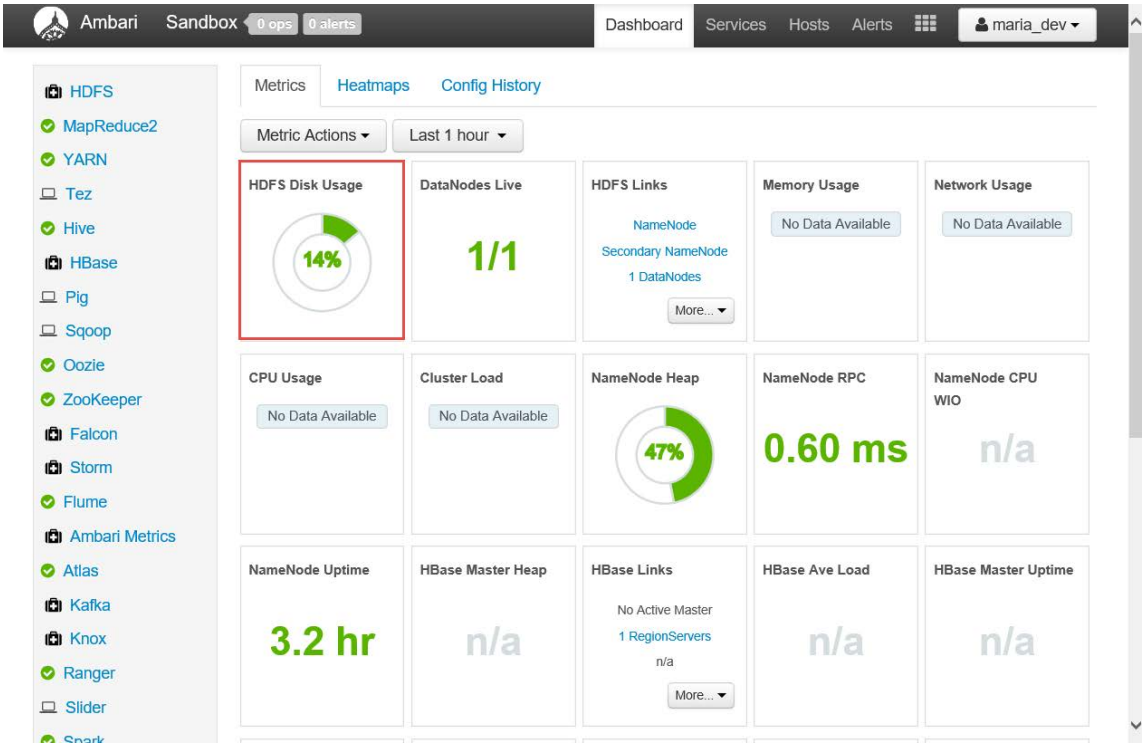


All you need to note from this screen is the IP address assigned to the Sandbox. This is highlighted as in the red box above. It will probably be different from the one above, but it will be the same every time you start the VM on the same PC/Laptop. This is quite convenient as it will allow you to store the related webpage addresses in the favourites of your web browser for future use.

When you put this IP address in your web browser of choice (any recent version of the popular browsers should be OK), you will get a Web page like that shown below.



The area highlighted in red is the IP address and port of the Ambari product within Hadoop. You can click on the IP address (which is a link) and then at the login prompt use the provided username (maria_dev) and password (maria_dev) to login. The screenshot is taken from the 2.4.0 version of the Sandbox. In you are using an earlier version you will note that the username and password are both set to admin and admin. The initial page of Ambari is essentially a dashboard as shown below:

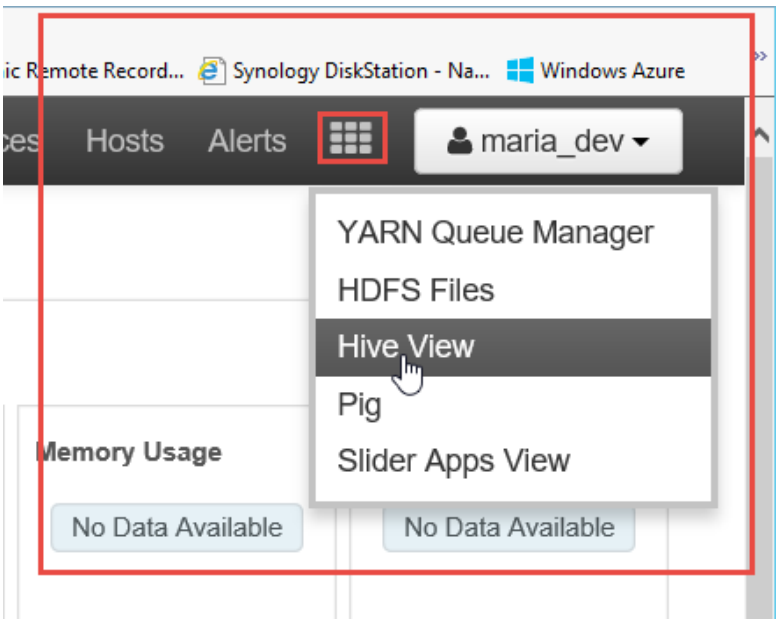


The only metric you are likely to be particularly interested in as you use the Sandbox is the HDFS[2] Disk usage, which effectively tells you how full the VM is.

7.2. Test the Sandbox is working

We have only logged into to Ambari so that we can access the Hive View which will allow us to run a simple query to check that the Sandbox is working. Hive is a tool provided within the Sandbox that is used to explore and manipulate data.

To access Hive, click on the highlighted 'grid' dropdown list button and select Hive View.



The Hortonworks Sandbox comes with a sample dataset: 'sample_07'. We will use this dataset to do a simple test to make sure that everything is working as expected.

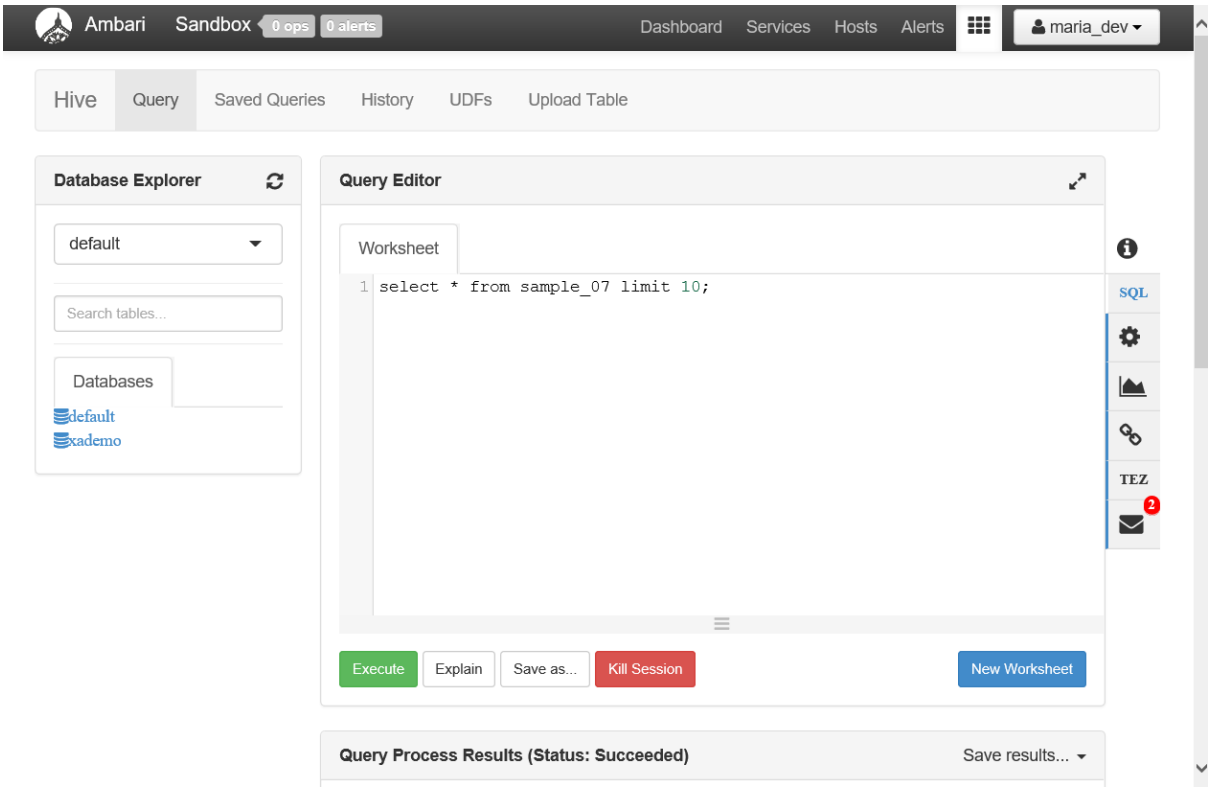
The following command[3] tells Hive to show the first 10 lines from the 'sample_07' table.

```
Select * from sample_07 limit 10;
```

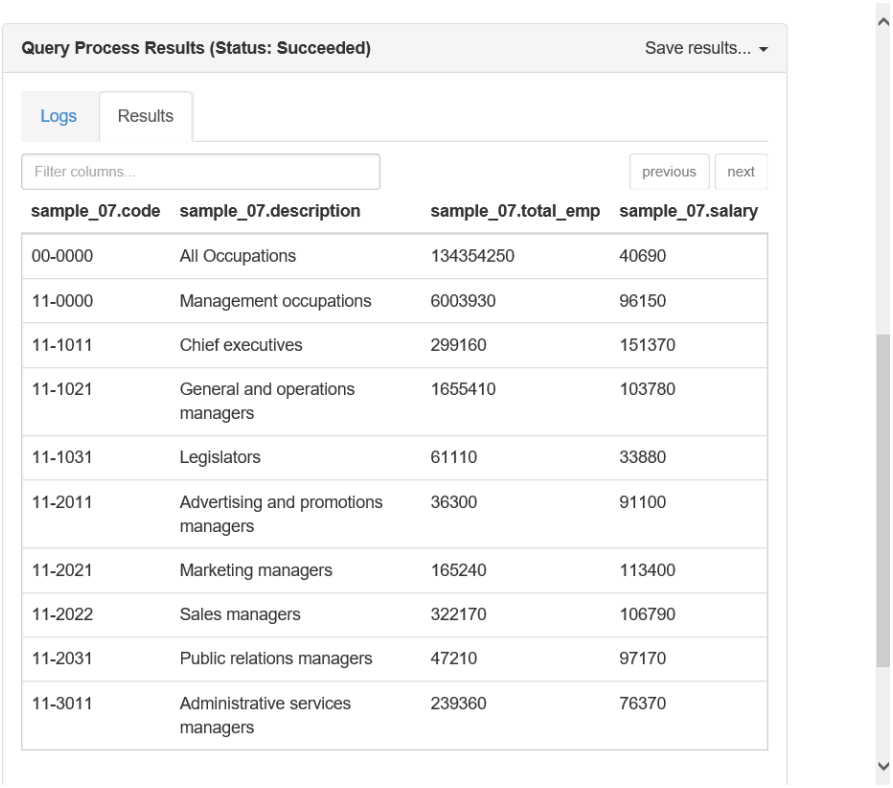
Type the command into the Worksheet tabbed area in the centre of the screen as shown below:

[2] HDFS = Hadoop Data File System

[3] The commands in Hive are in an SQL-like language, HiveQL



And click the Execute button. The query may take a few seconds to run but when completed, you should see results returned at the bottom of the screen.



sample_07.code	sample_07.description	sample_07.total_emp	sample_07.salary
00-0000	All Occupations	134354250	40690
11-0000	Management occupations	6003930	96150
11-1011	Chief executives	299160	151370
11-1021	General and operations managers	1655410	103780
11-1031	Legislators	61110	33880
11-2011	Advertising and promotions managers	36300	91100
11-2021	Marketing managers	165240	113400
11-2022	Sales managers	322170	106790
11-2031	Public relations managers	47210	97170
11-3011	Administrative services managers	239360	76370

8. What can I do with Hadoop in the Sandbox?

The simple test above is just to demonstrate that the system is installed and running as expected. The table sample_07 is provided with the Sandbox purely for testing. You will be far more interested in installing your own datasets into the Hadoop system and then using the supplied tools such as Hive, Spark and Zeppelin to help you manipulate and analyse your data.

The following guides and webinars are available to help you to get started with the Sandbox using some data from the UK Data Service:

A guide on Loading data into HDFS (Hadoop Distributed File System - the file system used by Hadoop) is available on the UK Data Service website. The datasets that you will be shown how to load are 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, and are those used in the 'What is Hive?' webinar.
Additionally, there is a HiveQL example queries document which includes all of the code used in the 'What is Hive?' webinar. Together they will allow you re-create the tables and much of the analysis demonstrated in the webinar.

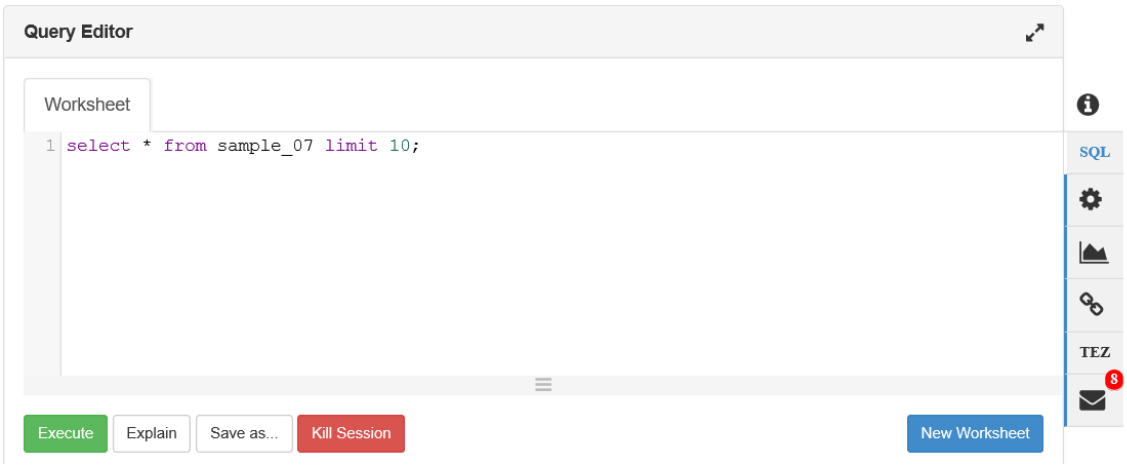
Once you have followed these guides, you should be in a position to adapt the data load instructions and the simple queries to load your own data and start analysing them using Hive.

9. Troubleshooting

Experience of using the Sandbox has shown that things do not always work as you expect them to and quite often for no obvious reason. Below is one example that we have come across and suggestions for working around it.

9.1. Hive query starts to run but doesn't finish

Although this rarely happens, it seems to do so most often on the very first query run. If this is your first query after setting up the Sandbox, it can be particularly disconcerting.

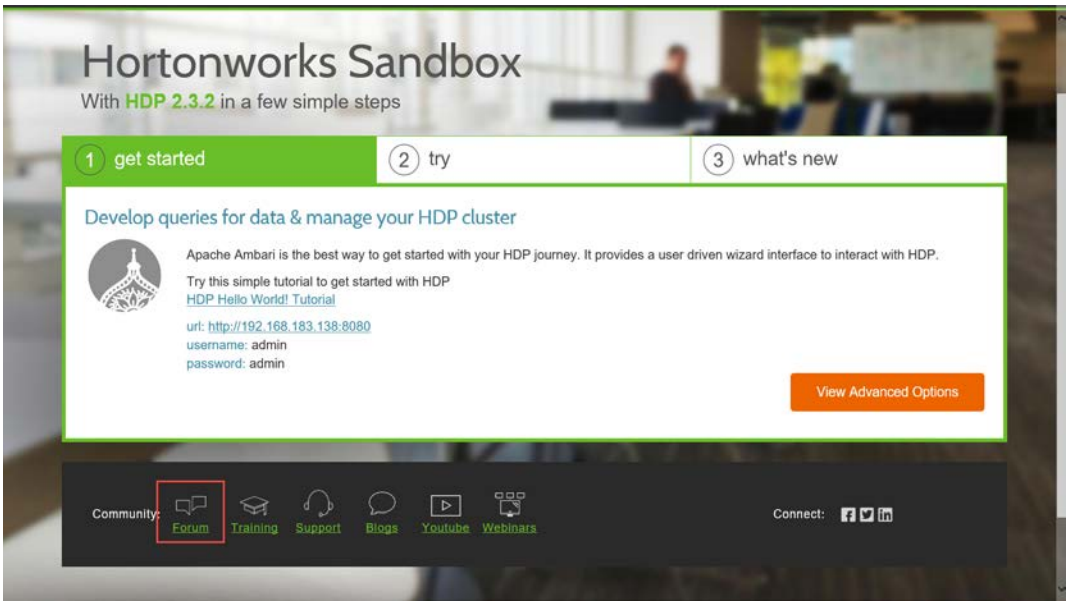


This screenshot shows the test query we used when checking the Sandbox setup. This query should take no more than a minute to run. If after a minute the message on this panel still says *Query Process Results (Status: Running),* then the chances are that it will not finish. In this case, you can copy the text of the query, click on New Worksheet; this will open a new tabbed pane and you can paste the query into it and Execute the query. The query should now run.

9.2. Other Problems

If you encounter other problems when using the Sandbox, Hortonworks provide a community forum which you can join (free) in which you may find answers or useful information about using the Sandbox or you can ask your own questions. If you are asking a question about a problem you are having you should always include as much information as possible; such as the version of the Sandbox, details of what you were trying to do and screenshots at least of any error messages you receive.

There is a link to the forum at the bottom of the initial Sandbox web page.



10. Next Steps

Having set up the Sandbox, the next step is to load data into it and to do some manipulation and analysis of the data. The following two guides available from the UK Data Service website will show you how:

- Loading Data into HDFS
- 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