Executing PIG Latin in HortonWorks Environment

Genoveva Vargas-Solar – CNRS/France
Plácido A. De Souza Neto – IFRN/Brazil

→ What is HortonWorks?

Hortonworks is a business computer software company based in Palo Alto, California. The company focuses on the development and support of Apache Hadoop, a framework that allows for the distributed processing of large data sets across clusters of computers.

HortonWorks also offers a pre-configured virtual machine to also run PIG and Hive.

→ Download and Install

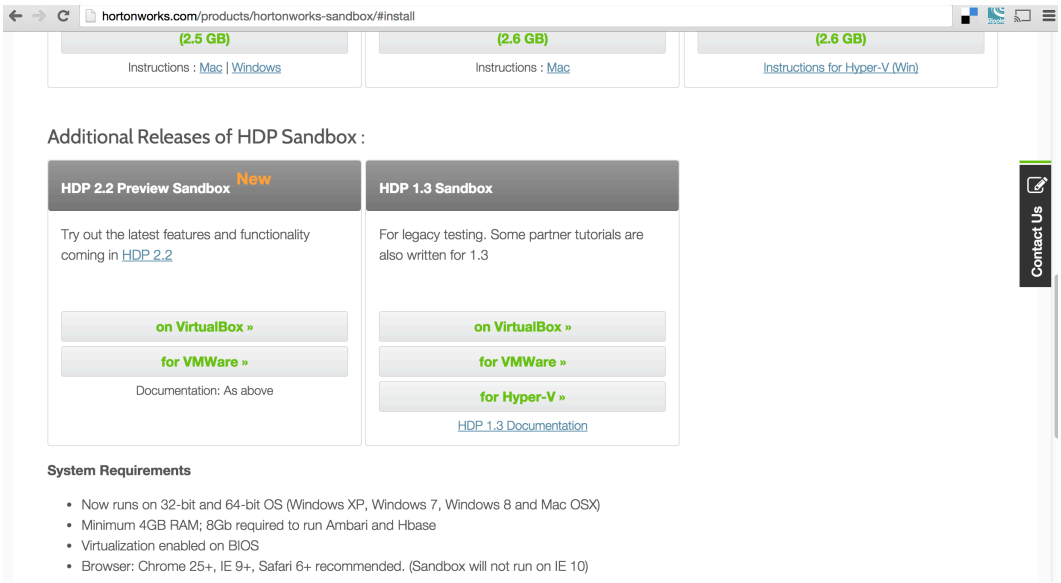
The HDP 2.2 Hortonworks Sandbox is virtual machine for initial use of PIG in a simple way. For this, it is necessary first install VirtualBox or VMWare.

VirtualBox and VMware provide cloud and virtualization software and services.

a. Link to Virtualbox: https://www.virtualbox.org/
b. Link to VMWare: http://www.vmware.com/

Once VirtualBox or VMWare installed, you must download the HortonWorks Sandbox.

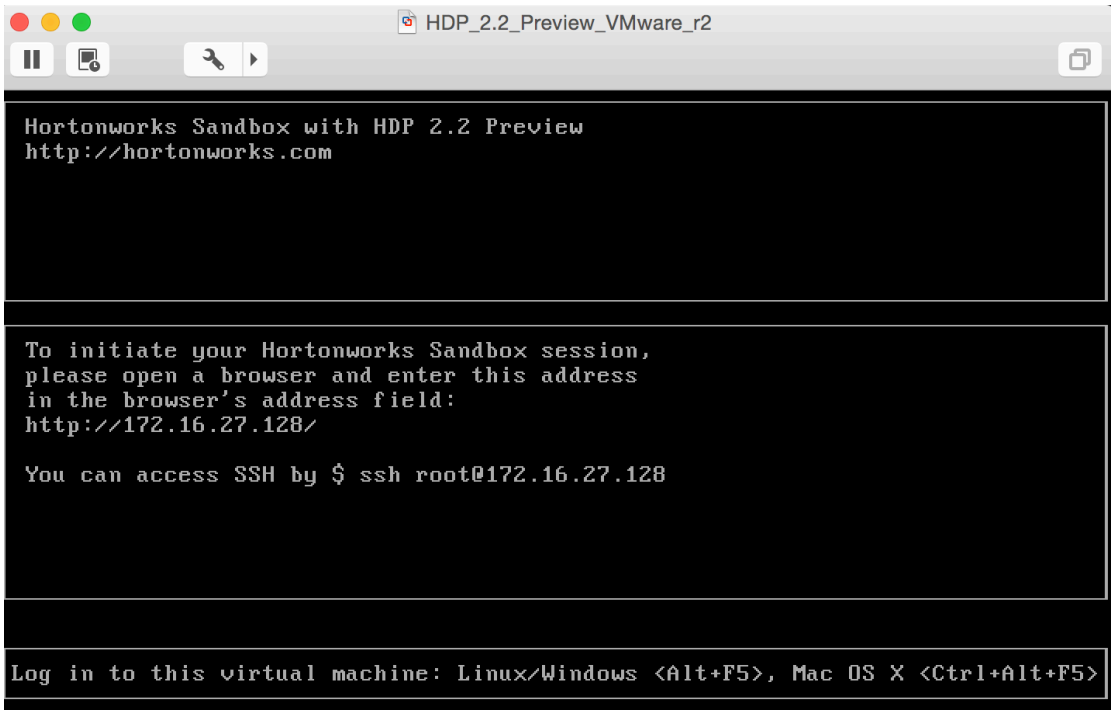
a. Link to Sandbox download:
http://hortonworks.com/products/hortonworks-sandbox
b. Choose HDP 2.2 Preview Sandbox as option for download.



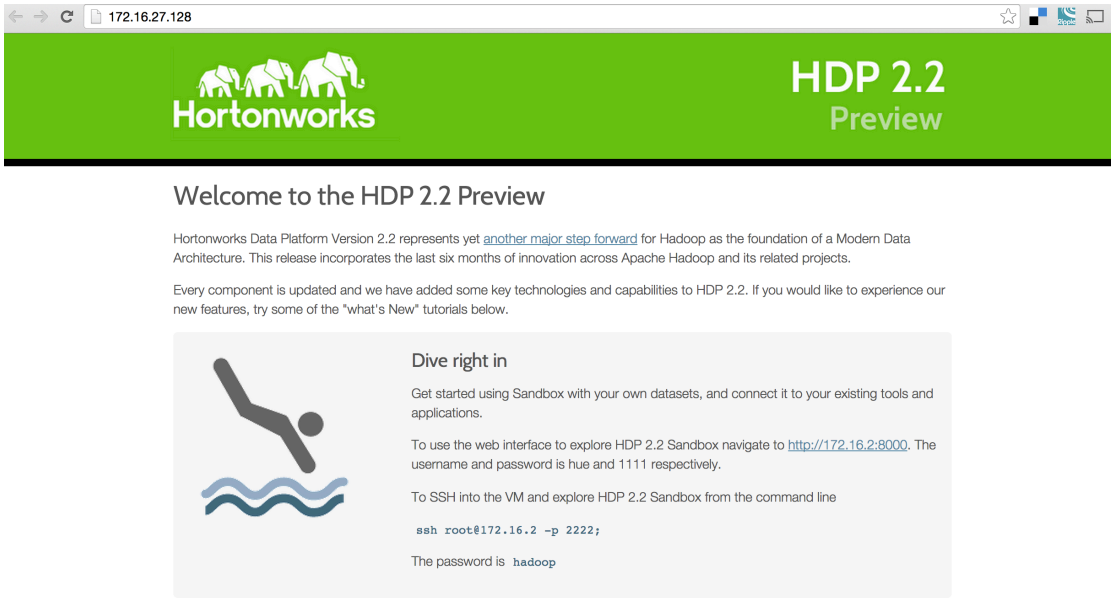
→ Executing HortonWork SandBox

When you run HortonWork SandBox virtual machine, you can analyse big data using, Hadoop, Pig or Hive.

The virtual machine will offer a IP to execute big data via HTTP browser connection or via SSH connection.



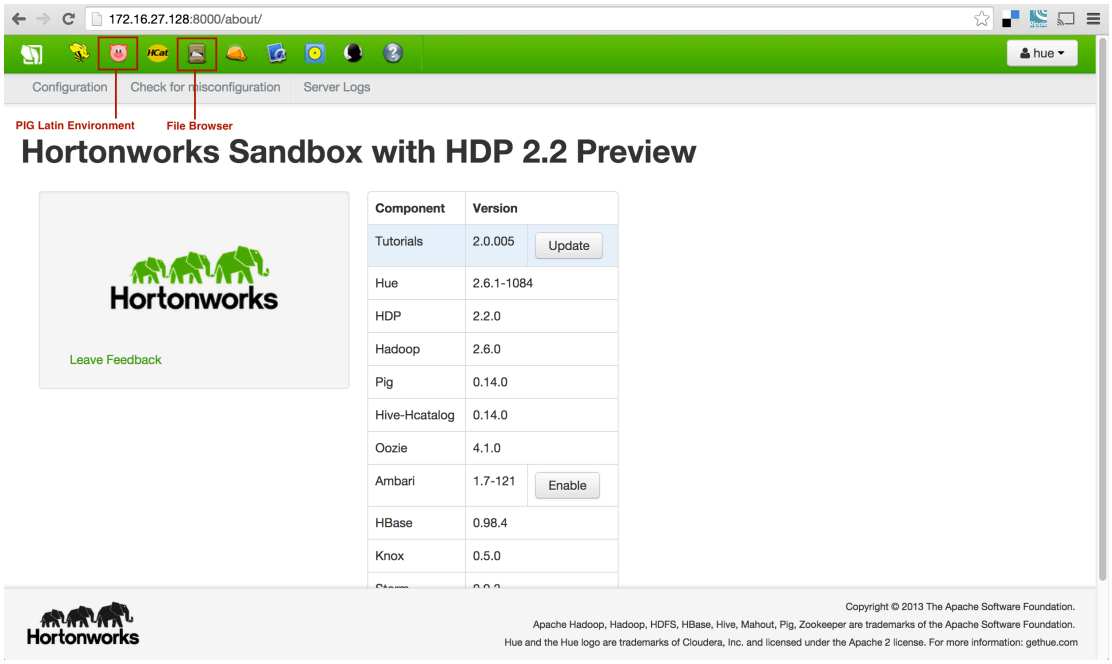
The figure presents the link ***http://172.16.27.128*** for http connection and ***root@172.16.27.128*** for ssh conection.



The http page also presents instructions for access the environment.

→ PIG Latin Environment on HortonWorks

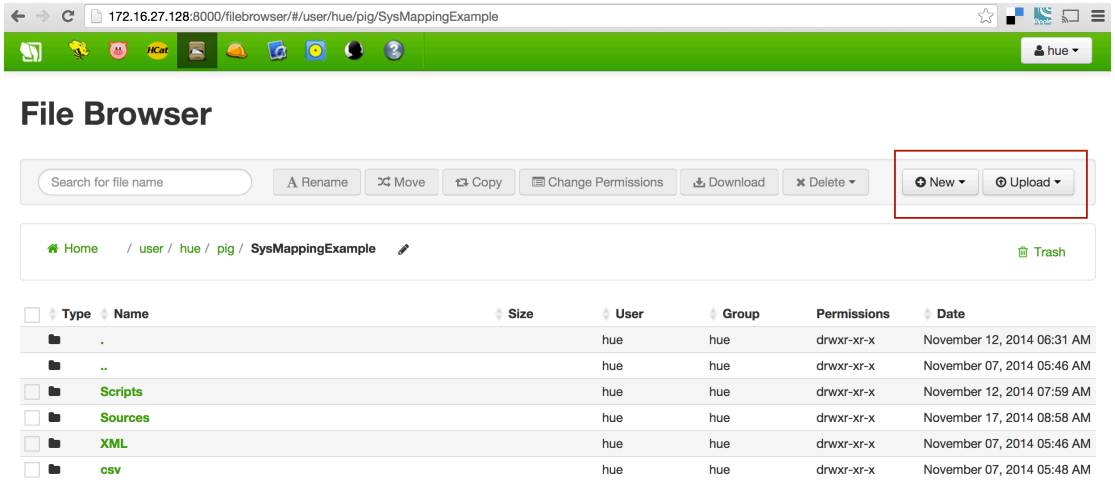
To access the PIG runtime environment it is necessary to insert in the browser ***172.16.27.128:8000 (or the IP presented in the sandbox virtual machine).***



Using this environment you can manage big files to be analyzed and run PIG scripts.

→ File Browser

In this environment you can create and organize your files, scripts and folders, and upload stuff to run on Hortonworks.

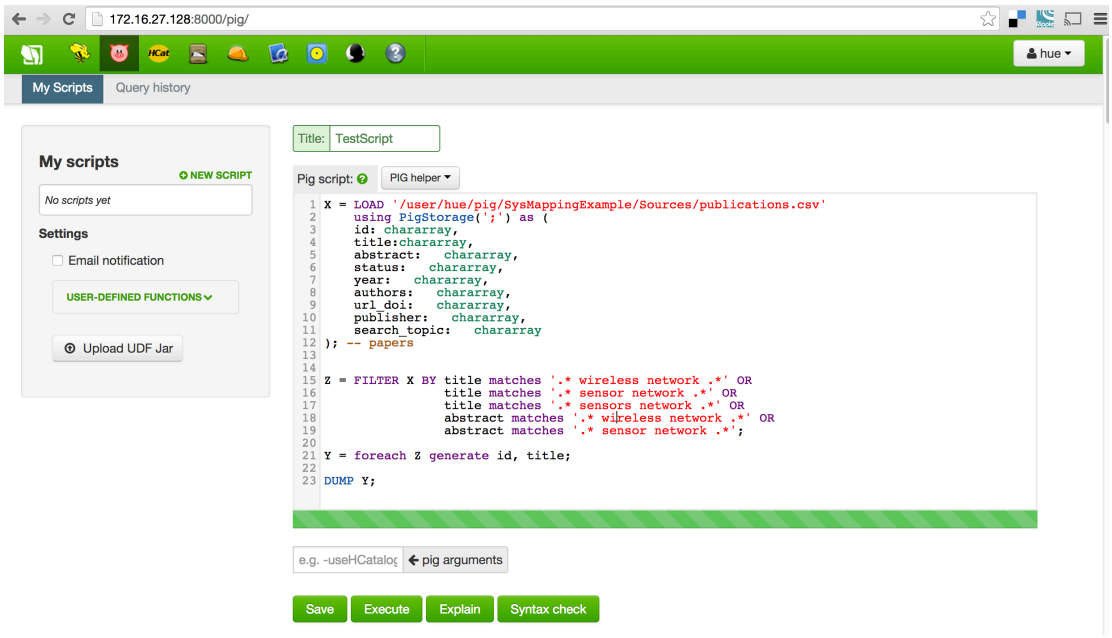


In this example a folder called **pig/SystematicMapping** was created in the standard user account **hue.** This file system is used for inputs and outputs of PIG scripts.

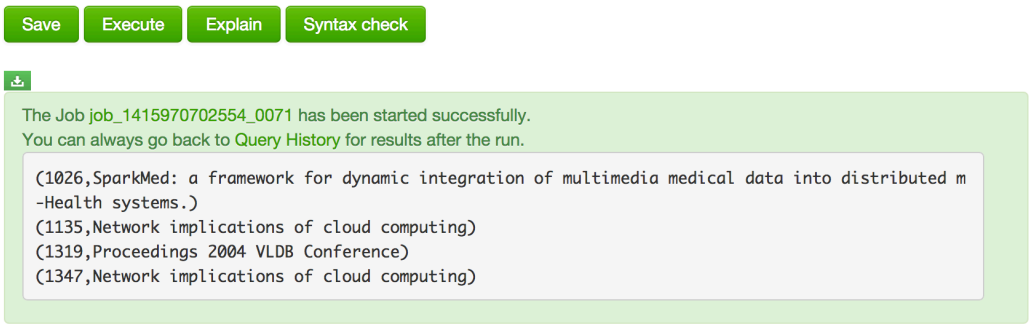
→ Executing PIG Latin

The environment for performing PIG is quite simple. Just insert the script in specific box and click **Execute**.

This script example look for papers that are described in a publications.cvs file that contains the words *wireless network* or *sensor network* in the *abstract* or *title* from a list of over 1300 publications.



Upon execution, the environment will print the IDs and titles of each paper found.



→ Files to Run the Example

You must use the *TestScript.pig* and *publications.cvs* files and to run the example.