Sandbox: Using The Tools

Pig

*What is it:* high-level language for analysis of data stored in HDFS or in Hive. Allows to write scripts with instructions for aggregating, filtering and cleaning data. With respect to Hive, it is more oriented on more complex tasks rather than interactive analysis.

*Access from HUE:* Click on "Pig" icon or go to "Hue shell" and choose "Pig Shell"

*Access from SSH:* Type "grunt" at the command line

Learning resources

R-Hadoop

*What is it:* R package that allows to write map-reduce programs in R, working directly on data stored in HDFS

*Access from HUE:* not available

*Access from SSH:* Type "R" at the command line

Learning resources

Hive

*What is it:* SQL interface to files stored in HDFS. Allows to issue SQL commands for analyzing data that has been previously stored in table structures. With respect to Pig, it is oriented to interactive analysis

*Access from HUE:* Click on "Hive" icon

*Access from SSH:* Type "hive" at the command line

Learning resources

Pentaho
What is it: business analytics platforms. Allows to build interactive visualizations on data that is either loaded from the web interface or stored in HDFS.

Access from Web: still to be configured

Learning resources: TBA