Notice

FILES AND PROGRAMS ARE PROVIDED AS IS. INFOBRIGHT, INC. IS NOT RESPONSIBLE FOR ANY MALFUNCTION OR SUPPORT OF SUCH FILES AND PROGRAMS. IF YOU NEED HELP CONSIDER POSTING ON THE INFOBRIGHT COMMUNITY FORUMS AT http://www.infobright.org/Forums

Description of Files and Folders

Folders:

HDF5 – Folder that will contain the HDF5 formatted million song data file. Place the million song data files that you download into this folder.

Files:

complete_data_grabber.py – ETL script that extracts the complete data from the million song HDF5 files into a flat, csv file(complete.csv) . **NOTE** Running this script will produce a large file. Allow approximately 40 GB of storage space and ~24 hours of runtime per 10,000 songs.

metadata_data_grabber.py – ETL script that extracts the metadata from the million song HDF5 files into a flat, csv file(metadata.csv) . **NOTE Allow approximately 4 GB of storage space and ~20 minutes of runtime per 10,000 songs.

create_metadata_table.php – PHP script that will create the database `Music` if it does not exist and also create the table `Metadata` with the correct specifications according to the requirements documentation. **NOTE** Make sure to change the server variables at the top of this file to connect to your instance of ICE/IEE.

create_complete_table.php – PHP script that will create the database `Music` if it does not exist and also create the table `CompleteTable` with the correct specifications according to the requirements documentation. **NOTE** Make sure to change the server variables at the top of this file to connect to your instance of ICE/IEE.

genre_dict.pyc and hdf5_getters.pyc – Python libraries to aid the data grabber python scripts.

Installation Procedure

Downloading Required Files

Million Song Data Set

Download the million song data set at http://www.infochimps.com/collections/million-songs. We recommend downloading and running the scripts on one subset at a time due to the amount of memory and CPU time the large data set takes during conversion.

Python and Python Libraries

Python:

Download Python at http://www.python.org/download/. Follow the guide from http://docs.python.org/using/windows.html for windows users that have never used Python before.

Python Libraries:

Numpy:

For 32 bits OS users download from http://numpy.scipy.org/ and install the library.

For 64 bits OS users download from http://www.lfd.uci.edu/~gohlke/pythonlibs/ and install the library.

Numexpr:

Download Numexpr from http://code.google.com/p/numexpr/downloads/detail?name=numexpr-1.4.2.tar.gz and install the library.

Infobright Million Song Files:

Download the Infobright Million Song file archive from http://INSERT_URL_HERE and extract the package.

Infobright Community Edition:

Go to http://www.infobright.org/Download/ICE then download and install the version for your system.