Homework #2: Exploring HDFS Metadata Using XML & XPath

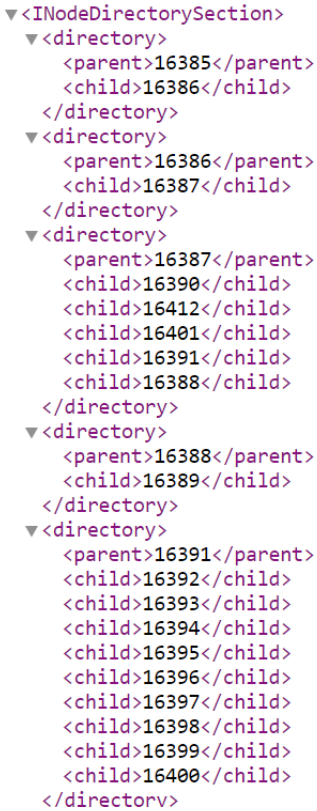
Due: February 18, Friday (11:59pm)
100 points

In this homework, we will explore the metadata stored in the namenode of HDFS. You can obtain such metadata by using the Offline Image Viewer (oiv) tool provided by Hadoop (https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsImageViewer.html).

For example,

<Your Hadoop-installation-dir>/bin/hdfs oiv -i /tmp/hadoop-ec2-user/dfs/name/current/fsimage_0000000000000000564 -o fsimage564.xml -p XML

will export the metadata stored in the specified fsimage (file system image) to an XML file called fsimage546.xml.



Fsimage has a INodeSection listing metadata about each inode and a INodeDirectorySection describing the directory structure, as show above. Note that id of inode is its inumber; and the directory nodes are represented by their inumbers, e.g., 16385.

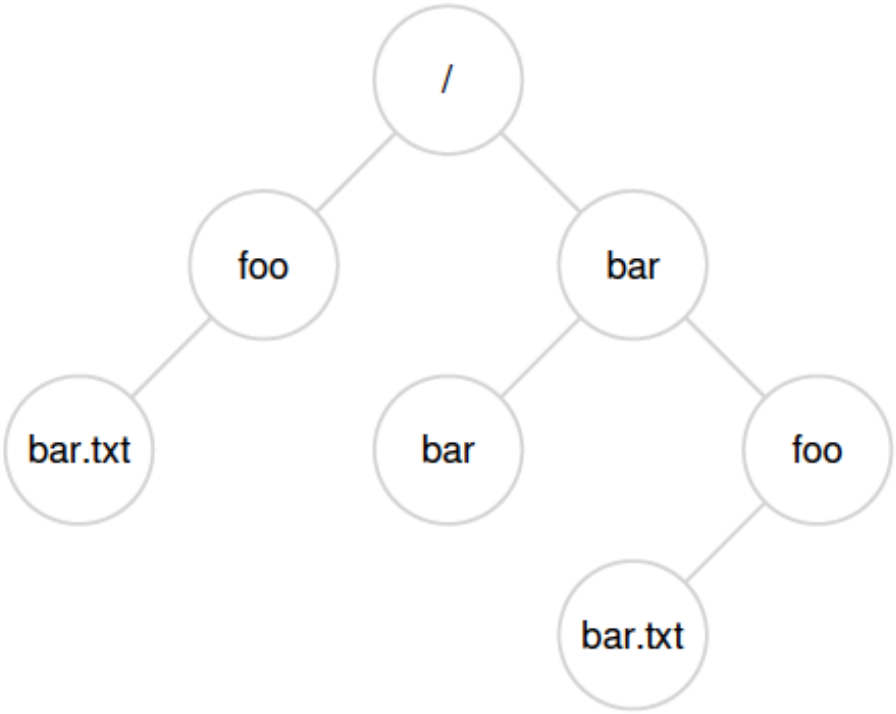
DSCI 551 – Spring 2022

Your task is to implement a Python program stats.py that takes a fsimage file in XML and outputs an JSON file that contains the following statistics about the file system:

{“number of files”: 5, “number of directories”: 10, “maximum depth of directory tree”: 4

“file size”: {“max”: 3518, “min”: 16}}

Note the maximum depth of directory tree is the number of levels of the tree, e.g., the maximum depth of the following directory tree is 4. Note if the file system does not contain any files, then you should not output the statistics about the “file size”.



Permitted libraries: lxml, sys, pandas.

Execution format: python3 stats.py <fsimage xml file> <stats json file>

e.g.,

python3 stats.py fsimage564.xml stats.json

Submission:

- Submit a ZIP file (file name => Lastname_Firstname_hw2.zip) which contains the following files –
  - stats.py
  - stats.json

NOTE: BE SURE to check all Piazza posts before submitting.