Homework 2: HDFS & XML

DSCI 551 – Fall 2024

Due: 11:59pm, October 11, 2024, Friday

Points: 100

In this homework, you are asked to take the file system image file produced by hdfs Offline image viewer utility and write a Python program edfs.py to emulate how dfs ls command works. Your program should produce the output similar to that in HDFS (examples shown below).

Execution format:

python3 <yourname>_hw2.py <fsimage> -ls <object>

Where <fsimage> is a file system image file in XML format, <object> is a file system object, that is, file or directory (with complete path from root /).

See **in-class example** on the OIV utility. More details can be found here: https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsImageViewer.html

More information about file system shell command 'ls' can be found here:

https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/FileSystemShell.html#ls

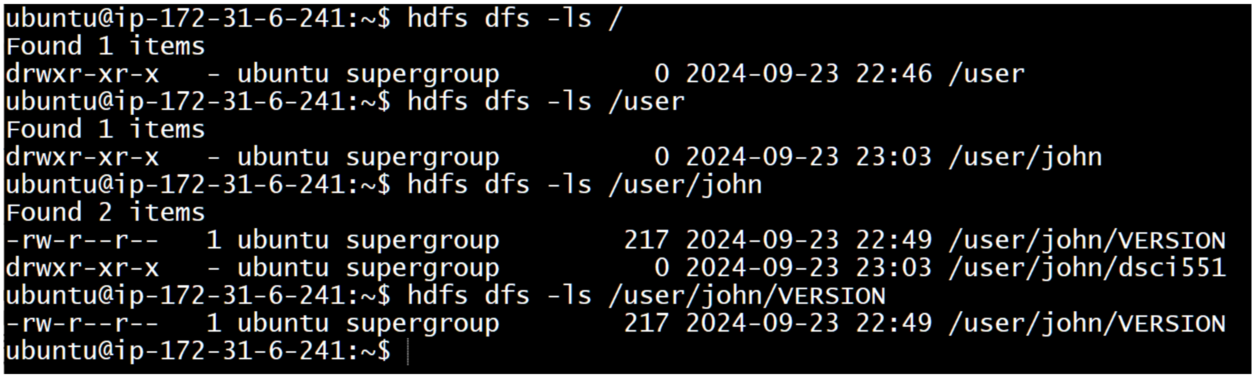
Requirements:

- Your code should use the xpath of lxml library to extract directory information and locate the information about a particular file system object.
- You need to convert the mtime (in epoch format) to date and time as shown in sample output.
- For file, it should return:
  ```
  permissions number_of_replicas userid groupid filesize
  modification_date modification_time filename
  ```

  note permissions need to be in 10-character format (see example below).

- For directory, the first character in the 10-character permission should be 'd'.
  The number_of_replicas for directory should show – and the size should show 0.

- Your code should run properly on any fsimage file other than the sample file provided to you.
- If the input object path you have provided is not available your code must return an empty string “”.



```
ubuntu@ip-172-31-6-241:~$ hdfs dfs -ls /
Found 1 items
drwxr-xr-x   - ubuntu supergroup          0 2024-09-23 22:46 /user
ubuntu@ip-172-31-6-241:~$ hdfs dfs -ls /user
Found 1 items
drwxr-xr-x   - ubuntu supergroup          0 2024-09-23 23:03 /user/john
ubuntu@ip-172-31-6-241:~$ hdfs dfs -ls /user/john
Found 2 items
-rw-r--r--   1 ubuntu supergroup        217 2024-09-23 22:49 /user/john/VERSION
drwxr-xr-x   - ubuntu supergroup          0 2024-09-23 23:03 /user/john/dsci551
ubuntu@ip-172-31-6-241:~$ hdfs dfs -ls /user/john/VERSION
-rw-r--r--   1 ubuntu supergroup        217 2024-09-23 22:49 /user/john/VERSION
ubuntu@ip-172-31-6-241:~$
```

Note:

- The outputs in the above screenshot are the sample outputs [similar in structure might differ in values] for the given fsimage.
- You are not required to process the “Found 1 items” line.

Permitted libraries:

- lxml, time, sys

lxml resource: lxml - Processing XML and HTML with Python

SUBMISSION INSTRUCTIONS:

1. A single python file with name: **[Student_Name]_hw2.py** [replace Student_Name with your name] Eg. John_Smith_hw2.py
2. Do not copy paste the commands from handout to terminal. Please rewrite the commands in the terminal. PDF format will encode special characters which are different from UTF-8 encoding in the terminal.
3. Do not modify any contents in the template. Just fill the template by reading the comments. Feel free to add helper functions as per your requirement.
4. The test script will accept the return data same as specified in the template.
5. Testing is done by test script with different test cases. So points will only be awarded if the method returns the expected result.
6. You will get 0 points if the code breaks for any syntax errors or any other problems. Please test the code thoroughly before submitting.