CS226 Assignment 1: HDFS

Overview

The goals of this assignment are:

1. Setup the development environment for Hadoop
2. Understand and use the APIs for HDFS
3. Compare the performance of HDFS to the local file system

Description

Write a Java program that makes a copy of a file. The source and target files could be either in the local file system or HDFS. It should run from the command line and it takes two command line arguments that represent the source and target files. It should read the source file and write all its contents to the target file. Notice that either file could be in HDFS or the local file system so your program should be designed to deal with both file system types. If the source file does not exist, the program should signal an error. If the target file already exists, it should report that and fail. If the target file cannot be created, for any reason, this should also be reported. Test your program on your local machine and make sure that it works correctly. Your program should use the FileSystem methods as described in class such as FileSystem#open and FileSystem#create so that it can work seamlessly with the local file system and HFDS. Install HDFS in pseudo-distributed mode on your development machine to test the HDFS functionality correctly.

Use the following three tasks to measure the performance of the file system and compare the performance of the LocalFileSystem to the DistributedFileSystem.

1. The total time for copying the 2GB file provided in the instructions below from the local file system to the local file system.
2. The total time for copying the 2GB file from the local file system to HDFS.
3. The total time for copying the 2GB file from HDFS to the local file system.

Submission instructions

- The assignment is due on Thursday, 11/10/2020, at 11:59 PM Pacific Time.
- The Java class should be named HDFSUpload in the package edu.ucr.cs.cs226.<ucrnetid> where <ucrnetid> is replaced with your ID all in lower-case letters. For example, the Java class could be named 'edu.ucr.cs.cs226.eldawy.HDFSUpload'
- The test file can be accessed on the following link: https://drive.google.com/file/d/0B1jY75xGiy7eR3VpNC1XMzB5cWs/view
  Notice that the file is compressed. Make sure to decompress it first.
- Please upload your answer in a single ZIP file named 'cs226-asg1-<ucrnetid>.zip' where <ucrnetid> is replaced with your ID. The ZIP file should contain a directory named 'HDFSUpload' which contains the full directory structure as generated by Maven with your implementation. Please remove any binary files and keep only the HDFSUpload.java file and pom.xml file. You can optionally include a README file for compilation instructions and a LICENSE file.
- Failing to follow the instructions above might result in losing some points.