Advanced Data Management (CMM524)

Solution to Laboratory #5: Introduction to HDFS

1. Aims

- To familiarise with the virtual machine environment used in the labs.
- To experience basic HDFS commands.

2. Outcomes

In completing this exercise, you should be able to:
- Use the CMM524 virtual machine environment in VMWare.
- Transfer files from/to between the local file system and HDFS.
- Access files stored in HDFS.

3. Running the CMM524 Virtual Machine

The CMM524 virtual machine is a Ubuntu Desktop with Hadoop, Pig and Spark installed. While it runs Hadoop on a single node rather than a multi-node cluster, the commands to use Hadoop and HDFS are the same.

Instructions to run the CMM524 virtual machine is in Lab 02, Section 3.

4. Using HDFS

4.1. Checking the Status of Hadoop

When the VM boots up, a number of daemon processes are started automatically.

To check the status of these processes:
- Open a terminal.
- Type the command:

```
jps
```

- How many processes can you see? What are their names?

In most cases if some processes are stopped, the easiest way to restart them is to reboot the VM.

In some rare situations you may want to stop or start all processes manually without rebooting. ***Only do this if you know what you are doing:***
- To stop all process: `stop-all.sh`
- To start all processes: `start-all.sh`

4.2. Exploring HDFS

The basic operation is to show the content of a HDFS directory.

- Open a Terminal in the VM.

- Show the HDFS root directory content with the following command:

```
hadoop fs -ls /
```

  - What folders/directories can you see under the root of HDFS?
    - You should see the following directory under root: `user`. In some cases there may be more, if some other directories have been created at the root of the virtual machine.

- List the content of your home directory by the following command:

```
hadoop fs -ls /user/training
```

  - Is there any file/folder in your home?
    - By default the directory is empty and thus there should be nothing returned by the command.
  - What is the equivalent command for showing the content of your home folder in HDFS?
    - The easy command is simply "`hadoop fs -ls`" as not specifying a location means your HDFS home directory.

4.3. Uploading Files to HDFS

One common task is in Hadoop to upload files from the local file system into HDFS.

- In your local Unix file system, change into the directory containing the sample data:

```
cd ~/training_materials/developer/data
```

- Show the content of the current directory in the local file system:

```
ls
```

  - Does the Unix `ls` command resemble the HDFS one?
    - The HDFS counterpart is "`hadoop fs -ls`" which very likely is inspired by the Unix `ls` command.
    - However, the Unix `ls` command looks at the "current directory" by default. HDFS does not have the "current directory" concept.
  - How many files/folders are in the current directory?
    - There are quite a number of files and folders in the directory. Your number may differ from mine as I have worked with the folder.
  - Can you see a `shakespeare.tar.gz` file in the current directory?
    - You should see a `shakespeare.tar.gz` file in the current (Unix) direcctory.

We will upload files contained in a GZIP compressed file. Before we can do this, we need to decompress the GZIP.

- Unzip `shakespeare.tar.gz` with the following:

```
tar zxvf shakespeare.tar.gz
```

- If you wonder that the above command does:
  - "`tar`" is a Unix program that extracts files from an archive.
  - "`zxvf`" are options to the "`tar`" command/program.
    - "`z`" means unzip the file before extraction.
    - "`x`" means extraction from the archive.
    - "`v`" means provides verbal information during processing.
    - "`f`" means the filename is the one that follows.
  - If you type "`man tar`" at the Unix command line, you can see the manual page of the `tar` command. Press "q" to quit.
- Show the current directory content again. What new folder is created?
  - The extraction creates a "`shakespeare`" directory.

We can now upload files in the `shakespeare` sub-folder (in the local file system) to HDFS.

- Type the following to upload the content of folder `shakespeare` into HDFS:

```
hadoop fs -put shakespeare /user/training/shakespeare
```

- Note the HDFS does not overwrite files by default. This is a safety feature that prevents you from accidentally destroying files that you already have. If you try to run the above command again, Hadoop will return errors.
- List the content of your HDFS home directory:

```
hadoop fs -ls /user/training
```

  - Can you see the uploaded folder/files?
    - You should see a "`shakespeare`" directory in HDFS, which is the result of your uploading.
- Does the following command gives the same result?

```
hadoop fs -ls
```

  - What is the different between specifying or not specifying "`/user/training`"?
    - The results are the same, as not mentioning the location in HDFS means the user's home directory.
- What command can you use to list the content of your `shakespeare` folder in HDFS?
  - `hadoop fs -ls shakespeare`

We also need another data file of web server log which is current compressed as `access_log.gz` in the local file system.

- Uncompress and view the weblog with the following command:

```
gunzip -c access_log.gz
```

- This command sends the uncompressed file content to the screen. It does not create any file.
- The web log is quite long and it will go on for a while. **Break the listing by pressing Ctrl-C.**

The "`hadoop fs -put`" command can also take input from the "*standard input*" and upload it directly to HDFS.

- The following command unzips the weblog ZIP file and uploads it to HDFS in 1 step **(Note: In 1 single line!)**:

```
gunzip -c access_log.gz | hadoop fs -put - weblog/access_log
```

- The above uses the Unix "*pipe*" which connects the output of 1 command to the input of the next command.
  - The "pipe" (i.e. "`|`") above connects the output of the first command to the input of the next command.
    - The first command "`gunzip -c access_log.gz`" decompresses `access_log.gz` and sends the output to "*standard output*" (i.e. the screen). This output is now piped through to the next command.
  - Look at the "`hadoop fs -put`" command above. Does it specify where the input file is? What does the "`-`" mean?
    - In a normal "`hadoop fs -put`" command, "`-`" specifies the Unix file/directory which will be uploaded to HDFS. However, when the `hadoop` program sees the "`-`", it knows to get its input from "standard input" rather than a file called "-" in the Unix file system. It is thus getting the file/data to upload through the pipe which connects to `gunzip`'s output.
    - Unix pipe is a convenient mechanism which allows multiple commands to be chained together.
- Verify that the log file is uploaded to HDFS.
  - What command should you use?
    - `hadoop fs -ls weblog`

The web log file is quite large (~50MB). We will create a smaller version of its first 5000 lines.

- The following command pipes the decompressed output to the Unix `head` command and takes only the first 5000 lines:

```
gunzip -c access_log.gz | head -n 5000
```

  - The first command alone in the above line decompresses `access_log.gz` to the standard output (i.e. the screen). However, with the pipe, the output is not sent to the screen but to the standard input of the `head` program which outputs the first 5000 lines to the screen.
- Create a sub-folder `testlog` in your HDFS home.
  - What is the command to use?

        - `hadoop fs -mkdir testlog`
        - Notice that by saying "`testlog`", you mean a sub-directory under your HDFS home.
- Take the first 5000 lines from the web log file and upload it to the HDFS file "`testlog/test_access_log`".
    - What is the command to use? (Hint: Consider piping the put of the "`head`" command to "`hadoop fs -put`" which reads from standard input.
        - `gunzip -c access_log.gz | head -n 5000 | hadoop fs -put - testlog/test_access_log`
        - This is an extension of the `gunzip+head` example above. You extend it by adding another pipe that connect the output of `head` (i.e. the 1st 5000 line of the decompressed log) to `hadoop` which is told to read from its standard input and uploads it to the HDFS file `testlog/test_access_log`.

4.4. Viewing & Manipulating Files in HDFS

You can view the content of a HDFS file by sending it to the console.

- Type the following command to show the content of the `shakespeare/histories` file in HDFS:

```
hadoop fs -cat shakespeare/histories
```

    - Note: The file can be quite long. **Use Ctrl-C to break it if needed.**
- If you only want to see the last 50 lines of the file, you can do so by piping the output to the Unix `tail` command:

```
hadoop fs -cat shakespeare/histories | tail -n 50
```

    - `tail` is similar to `head`, while 1 returns the last n lines and the other returns the first n lines. Both can be used as filters.
- Unix commands/filters like `head`, `tail`, `more` and `less` can help you to view content of a file in a controlled manner. To get the manual of a command, type "*man <command name>*" at the Unix prompt. e.g. "`man less`".

Files in HDFS can be manipulated just like their Unix counterparts.

- The glossary file is not strictly a work of Shakespeare. Remove it with the following command:

```
hadoop fs -rm shakespeare/glossary
```

    - If you check HDFS after the removal, `glossary` should be gone.

Files in HDFS can be also downloaded back to the local file system.

- To download a file from HDFS to the local file system:

```
hadoop fs -get shakespeare/poems ~/shakepoems.txt
```

- You can check the content of your home directory in the local file system by the following Unix command:

```
ls ~
```

  - You should see the file `shakepoems.txt` in your Unix home directory.
- Once a file is in the local file system, you can use ordinary Unix commands to manipulate it. For example, to see its content using the `less` command/filter:

```
less ~/shakepoems.txt
```

  - `less` is a convenient program/filter which allows you to scroll through a long file when it is shown on the screen.
    - When `less` is given a filename (like the above), it displays the specified file.
    - When `less` is not given any filename, it expects to read data from the standard input. So you use use a pipe to connect another program's output to `less`.

4.5. Other Commands

FsShell (i.e. `hadoop fs`) supports many other commands for HDFS.

- Show all commands by typing:

```
hadoop fs
```

  - The `hadoop` program is the gateway to Hadoop's features. The sub-command "`hadoop fs`" accesses functions in HDFS. You can see many other HDFS functions available with the "`haddop fs`" command.

4.6. Summary

Complete the following quick summary of HDFS and Unix commands:

- Show directory content in HDFS: `hadoop fs -ls`
- Upload a file to HDFS: `hadoop fs -put`
- Download a file from HDFS: `hadoop fs -get`
- View a file's content in HDFS: `hadoop fs -cat`
- Remove a file in HDFS: `hadoop fs -rm`
- Create a directory in HDFS: `hadoop fs -mkdir`
- Remove a directory in HDFS: `hadoop fs -rmdir`

- Show directory content in Unix: `ls`

- Change current directory in Unix: `cd`
- View a file's content in Unix: `cat`
- Remove a file in Unix: `rm`
- Create a directory in Unix: `mkdir`
- Remove a directory in Unix: `rmdir`