Advanced Data Management (CMM524)

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?

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

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

- Is there any file/folder in your home?
- Is there a shorter command to show the content of your home folder in HDFS?

4.3. Uploading Files to HDFS

One common task in Hadoop is 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?
- How many files/folders are in the current directory?
- Can you see a `shakespeare.tar.gz` file in the current directory?

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
```

- Show the current directory content again. What new folder is created?

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
```

- List the content of your HDFS home directory:

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

    - Can you see the uploaded folder/files?
- Does the following command gives the same result?

```
hadoop fs -ls
```

    - What is the different between specifying or not specifying "`/user/training`"?
- What command can you use to list the content of your `shakespeare` folder in HDFS?

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
```

    - 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.
    - Look at the "`hadoop fs -put`" command above. Does it specify where the input file is? What does the "`-`" mean?
- Verify that the log file is uploaded to HDFS.
    - What command should you use?

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
```

- Create a sub-folder `testlog` in your HDFS home.
    - What is the command to use?
- 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.

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
```

- 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
```

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 ~
```

- 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
```

4.5. Other Commands

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

- Show all commands by typing:

```
hadoop fs
```

4.6. Summary

Complete the following quick summary of HDFS and Unix commands:

- Show directory content in HDFS: _____
- Upload a file to HDFS: _____
- Download a file from HDFS: _____
- View a file's content in HDFS: _____
- Remove a file in HDFS: _____
- Create a directory in HDFS: _____
- Remove a directory in HDFS: _____

- Show directory content in Unix: _____
- Change current directory in Unix: _____
- View a file's content in Unix: _____
- Remove a file in Unix: _____
- Create a directory in Unix: _____
- Remove a directory in Unix: _____