Intro to Big Data - Lab 1 (Hadoop Shell Commands)

What we will cover in this lab:

- **Getting familiar with the Hadoop Core (HDFS, Distributed File-System)**

Hadoop File System (FS) Shell Commands

https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html

Preamble:

- On your local machine, start the Oracle VM VirtualBox; press Start on the upper menu of the Oracle VM VirtualBox Manager.
- Launch the command line (from the upper menu bar)
- Basic information:
  The HDFS is completely separated from your local file system. You cannot access it directly through a path within the normal linux file-system hierarchy, but have to use the hadoop fs tool for this purpose.
  Places within HDFS are given as a URL starting with hdfs://. You can drop this prefix when working with the hadoop fs tools, they assume it at the right places.
  Your home is at hdfs://user/[username] (hdfs://user/cloudera in your case), it is not accessible though the normal operating system methods. You need to use hadoop fs command to access and transfer files.
  The hadoop command always uses your home as base path if you supply a relative path or no path.

Your tasks:

- Find about your home directory:

```
hadoop fs -ls
```

- To show the contents of the root of HDFS, try:

```
hadoop fs –ls /
```

- Find out what else the Hadoop fs command can do:

```
hadoop fs -help
```

- To show the contents of the linux home directory (cloudera):

```
ls
```

- Create a file named “sample.txt” into the linux home directory:

```
cat > sample.txt
```

  Use ctrl+D to save and exit the file.

- Copy the file to the HDFS home directory:

```
hadoop fs -put sample.txt
```

- Show the contents of the HDFS home directory:

```
hadoop fs -ls
```

- Create a directory named “lab1” into the HDFS home directory:

```
hadoop fs –mkdir lab1
```

- Move the file “sample.txt” from the HDFS home directory to the “lab1” directory:

```
hadoop fs –mv sample.txt lab1
```

- Display the contents of the “lab1” directory:

```
hadoop fs –ls lab1
```

- Display the contents of the file “sample.txt” inside the “lab1” directory:

```
hadoop fs –cat lab1/sample.txt
```

- Remove the file “sample.txt” from the linux home directory:

```
rm –r sample.txt
```

- Show access rights:

```
Hadoop fs –ls –d lab1
```

Assignment:

1. Read the input from stdin and write to the HDFS home directory in “file1.txt” (using **put**).

2. Create a new directory “src” and populate it with two text files “file1.txt” and “file2.txt”. Concatenate “file1” and “file2” into the text file “output.txt” (outside of the directory “src”) using **getmerge**.

   What do you observe, where is “output.txt” created? Display the contents of “output.txt”.

3. Copy the file “output.txt” between the local file system and hdfs using **get**.

4. Display the number of directories, files and bytes under “lab1” using **count**.

5. Print all files under the “src” directory using **find**.

6. In a file “textFile.txt” print all .txt files that begin with “file” using **find**.