VIET NAM NATIONAL UNIVERSITY HO CHI MINH CITY
UNIVERSITY OF SCIENCE

INTRODUCTION TO BIG DATA ANALYSIS

Lab instructors

Vũ Công Thành | vcthanh.work@gmail.com
Huỳnh Lâm Hải Đăng | danghailamhuynh@gmail.com

Lab 01: Introduction to Hadoop Ecosystem

1. Setup a Hadoop cluster

1.1. Requirements

- Install a Hadoop cluster with a pseudo-distributed configuration.
- Follow the instructions to perform some of common HDFS commands.
- All steps of Hadoop installation must be screenshotted and put into a file named *Report.pdf*.
- You're required to run a jar file, take the output of that jar file and save it to a file named <*StudentID*>_*verification.txt*.

1.2. Expected output

- Know how to setup a Hadoop cluster by yourself. It will be used in the next labs.
- Know how to assign permissions to files on HDFS.

1.3. Instructions

- Install Hadoop, you can choose one of the following methods to install a Hadoop cluster:
    - For Ubuntu users: Follow the official tutorial from Apache to setup a Hadoop cluster in pseudo-distributed mode. Link
    - Docker-based: Link
    - Install Ubuntu VM and install Hadoop inside it: Link (VirtualBox) or Link (VMWare)
    - WSL: Link
- Create a folder with path **/hcmus** on HDFS.
- Create a user named *khtn_<StudentID>*.
- Create a subfolder at **/hcmus/<StudentID>.**
- Upload a file into **/hcmus/<StudentID>.**
- Chmod 744 **/hcmus/<StudentID>** and set the owner of that subfolder to the user named *khtn_<StudentID>*.
- Run the attached JAR file named **hadoop-test.jar**:

```
java -jar /path/to/jar/file.jar <YOUR_HDFS_PORT> /hcmus/<StudentID>
```

- Take the file named <StudentID>_verification.txt which is generated after running the JAR file, include it in your submission.

2. Warm up with Word Count

2.1. Requirements

- Write a program to count the number of words starting with letters a, f, j, g, h, c, m, u, s (case insensitive) in the attached dictionary file named **words.txt**.
- Export the list of the starting letters (in the list above) and those number of words in TSV format. Example:

```
a	1
f	3
j	5
```

2.2. Expected output

- Get familiar with MapReduce through the most basic problem: Word Count.

3. Submission Guideline

File structure of StudentID.zip

```
StudentID
|--- docs
|    |--- Report.pdf
|    |--- <StudentID>_verfication.txt
|    |    ...
|--- src
|    |--- WordCount
|    |--- | results.txt
|    |--- |   ...
|--- README (if needed)
```

You must follow the file structure above and compress it into a ZIP file named <StudentID>.zip