Lab 01: A Gentle Introduction to Hadoop

Instructor: Doan Dinh Toan
Department of Computer Science,
Faculty of Information Technology
University of Science - VNU-HCM
toandd.i81@gmail.com

Abstract

This lab will get you up and running with the setting up of a Hadoop cluster. Then, you will experiment with Hadoop by creating a MapReduce Hadoop program with Java after reading the original paper on the MapReduce idea. Finally, there are some extra assignments for bonus points.

0 Preliminary

0.1 Reminder

The main objective of this course is to learn, and truly learn. You can discuss this with your classmate, but you need to take responsibility for your submission, which actually depends on your understanding of this course. For any kind of cheating and plagiarism, students will be graded 0 marks for the whole course.

0.2 Submission guideline

Each team submits its result to a folder named `teamABC`, with `ABC` being the team's name. The folder structure is as follows:

```
teamABC
├── src
│   ├── section01
│   └── section02...
├── docs
│   ├── report.md
│   ├── report.pdf
│   └── images
└── readme.md
```

- `src` is the folder for your source code. If the lab assignment is split into multiple sections, you have to save your script in a separate folder, corresponding to the given lab assignment.
- `docs` is the folder for your documents, including the work report and images associated with your report. If the lab assignment requires screenshots as proof, the images need to be stored in this folder if you inserted them in the report.
  - `report.md` is your report file in MarkDown format. The report must be written in English. This assignment will come with a template folder that already has a report template (you can use my OSCP template or create your own). If you are not familiar with MarkDown, see this cheat sheet. The report must include the following items:
    * Your team's result (How much work, in percent (%), have you finished in each section?)
    * The answer to each section's tasks.

    - Reflection of your team. (Does your journey to the deadline have any bugs? How have you overcome it? What have you learned after this process? If you cannot overcome the bugs, describe where the bottlenecks are in your work.
    - References to your work.
  - `report.pdf` is the PDF file of your report, converted from the MarkDown file mentioned above.
- `readme.md` is the file that introduces your team and this lab assignment, this file should include the following basic information:
  1. Information about the course, the assignment, and notes to the instructors (if any).
  2. Information about your team (Student ID, full name of each member).

0.3 Rubrics

This lab assignment is divided into four parts, mentioned in the next sections.

1. Setting up SNC - Single Node Cluster (4 points)
   If every member of your team has set up an SNC successfully, the team gets 4 points. Otherwise,
   - *If that team is a four-member team, you will lose 1 point per failed member. The total points of Section 4 will be reduced to 1 point.*
   - *If that team is a three-member team, you will lose 1 point for your bad teamwork and 1 point per failed member. The total points of Section 4 will not be reduced.*
2. Introduction to MapReduce (2 points)
3. Running a warm-up problem: Word Count (2 points)
4. Bonus (2 points)

The report writing will take 2 points. In total, this assignment has 10 + 2 points. If you can achieve more than 10 points, the bonus will count towards to the next lab, but it will be decreased by half.

1 Setting up Single-node Hadoop Cluster

1.1 Requirements

Work as a team to install a single node Hadoop cluster by following the tutorial from the Apache Hadoop's official documentations [3]. When following the tutorial, the student needs to take screenshots of the installation and verifies if Hadoop is installed correctly. The shell/terminal screenshots need to have your Student ID on them explicitly. Here are some recommendations:

- Change your shell prompt (known as environment variable `PS1` on Linux) to your Student ID. For those who are not familiar with Linux, try this tutorial to understand what `PS1` is and how to change it, Note that this tutorial is using `/bin/sh` as the default shell; if you are using e.g. `bash`, `zsh` as the default shell instead (which many of you probably are), refer to your shell's documentations on how to properly set environment variables.
- Create a new user with elevated privilege (aka "sudo user") with your Student ID as username. Read this article to know how to do that safely. If you are using Linux as your main operating system, carefully add administrator privileges to the user. I have seen a lot of cases where the "sudo" privilege is indiscriminately granted to a bunch of users, leaving the main account locked out and the machine unable to log in normally. The only way to fix that error is to reinstall your OS after logging into a non-GUI version of your OS and making a backup.

1.2 Expected outputs

- Students can install a Hadoop cluster/instance on their own device. This cluster/instance would be used in the next lab and the midterm exam.
- To incorporate a high team spirit, each team members must have a mutual understanding to help each other during this lab assignment.

2 Introduction to MapReduce

2.1 Requirements

This exercise is adapted from a random book on Cognitive Science [7]. The student needs to read the original paper of MapReduce [5] and then answer the following questions:

1. How do the input keys-values, the intermediate keys-values, and the output keys-values relate?
2. How does MapReduce deal with node failures?
3. What is the meaning and implication of locality? What does it use?
4. Which problem is addressed by introducing a combiner function to the MapReduce model?

2.2 Expected outputs

- Students can research new concepts to master how to express scientific concepts and understanding.

3 Running a warm-up problem: Word Count

3.1 Requirements

Follow the tutorial to get the Example `WordCount v1.0` [4]. Students need to compile the code to a JAR file, then run them in the installed Hadoop cluster/instance. Take screenshots of each step with a short explanation in the report.

3.2 Expected outputs

- Students can verify their Hadoop cluster/instance is set up correctly and get used to run a MapReduce code in Hadoop.

4 Bonus

4.1 Extended Word Count: Unhealthy relationships (0.5 points)

Create a file named `Unhealthy_relationship.java` and save it in the `src` folder if you do this task. Each team needs to contribute one test case in the `src\input.txt` and `src\output.txt` which should be different from my test cases.

Our GenZ have an old quote "Trà đổ vào sữa hay sữa đổ vào trà đều như nhau, thế anh đổ em sao em không đổ anh?". According to science, you should you must pour the milk into the cup first before the tea to keep the milk's protein structure and prevent it from unwanted transformations due to the high temperature of the tea. But in practice, The Royal Butler of Buckingham said that tea should go in first. So "sữa" and "trà" is a quite complicated relationship. Let us take a simpler relationship, $\mathcal{R}$, for which $a\mathcal{R}b \neq b\mathcal{R}a$, $a \in A, b \in B, \mathbb{R} \subseteq A \times B$.

- A set of $\Delta_a = \{\forall u \in \Omega \ \texttt{if} \ \ a\mathcal{R}u\}$
- A set of $\Gamma_a = \{\forall u \in \Omega \ \texttt{if} \ \ u\mathcal{R}a\}$
- A score function $Z_a = |\Delta_a| - |\Gamma_a|$:
  - If $Z_a > 0$: node a is labelled "pos".
  - If $Z_a == 0$: node a is labelled "eq".
  - If $Z_a < 0$: node a is labelled "neg".

Given a list of relationships as $a\mathcal{R}b$, print the label of each node as output.

4.1.1 Input

A list of lines, each line has two words separated by a space. For example, "a b" indicates the relationship $a\mathcal{R}b$.

4.1.2 Output

A list of lines, each line has two elements n l, in which n is the name of the node and l is the label of the node.

4.1.3 Example

Input	Output
A B B C A C D E	A pos B eq C neg D pos E neg

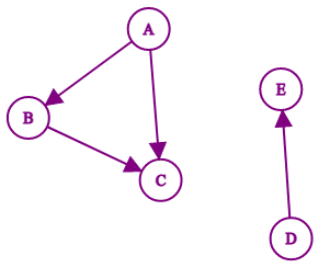

Figure 1: Visualization

4.2 Setting up Fully Distributed Mode (1.5 points)

4.2.1 Hadoop Cluster Setup in Non-Secure Mode (1 point)

Students follow the tutorial to set up Fully Distributed Mode [1] on at least 2 physical devices. Students should take screenshots for each step, using the same requirements in section 1.

4.2.2 Research about Security in Hadoop Set-up (0.5 points)

Students must finish the task of installing Fully Distributed Mode before doing this task. Read the documents about setting up Hadoop in "Secure Mode" [2, 6] and answer the following questions:

- Is your Hadoop secured? Give a short explanation if your answer is yes. Otherwise, give some examples of risks to your system.

- From your perspective, which method is better when securing your HDFS: authentication, authorization, or encryption? Give an explanation about your choices.

References

[1] Apache Hadoop 3.3.4 – Hadoop Cluster Setup. https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/ClusterSetup.html.

[2] Apache Hadoop 3.3.4 – Hadoop in Secure Mode. https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SecureMode.html.

[3] Apache Hadoop 3.3.4 – Hadoop: Setting up a Single Node Cluster. https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html.

[4] Apache Hadoop 3.3.4 – MapReduce Tutorial. https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Example:_WordCount_v1.0.

[5] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. In *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, pages 137–150, San Francisco, CA, 2004.

[6] Owen O'Malley. Hadoop Security Architecture. https://www.slideshare.net/oom65/hadoop-security-architecture.

[7] K Umamaheswari and V Priya. *Computational techniques for text summarization based on cognitive intelligence*. Taylor & Francis, London, England, 2023.