Lab 01: A Gentle Introduction to Hadoop

Lab Instructors:

Đỗ Trọng Lễ
dtle@selab.hcmus.edu.vn

Bùi Huỳnh Trung Nam
huynhtrungnam2001@gmail.com

Abstract

This lab aims to familiarize you with setting up a Hadoop cluster. Following this, you'll explore Hadoop by developing a MapReduce program in Java, informed by your study of the original paper on the MapReduce concept. Lastly, there are optional bonus assignments available for additional points.

Credit for this document goes to the aforementioned lab instructors and Mr. Doan Dinh Toan, the original creator of this content.

0 Preliminary

0.1 Reminder

The main objective of this course is to learn, and truly learn. You can discuss this with your classmates, but you need to take responsibility for your submission, which 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 should submit their results to a folder titled ***[GID]-[SID1]-[SID2]-[SID3]-[SID4]***, where "GID" represents your Group ID, and "SID" represents the Student IDs. Please ensure that the Student IDs are listed in ascending order. The expected folder structure is as follows:

```
[GID]-[SID1]-[SID2]-SID3]-SID4]
|--- src
|     |--- section01
|     |--- section02
|--- docs
|     |--- report.pdf
|     |--- images
|--- readme.md
```

- src is the folder for your source code. If the lab assignment is split into multiple sections, you must 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.
  - report.pdf is your report file in PDF format. The report must be written in English. The report must include the following items:
    1. Your team's result (How much work, in percent (%), have you finished in each section?)
    2. The answer to each section's tasks.
    3. Reflection on 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.
    4. References to your work.
  - images: If the lab assignment requires screenshots as proof, the images need to be stored in this folder if you inserted them in the report.
- Readme.md is the text 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 (1 point)

The report writing will take 2 points. In total, this assignment has 10 + 1 points. If you can achieve more than 10 points, the bonus will count towards 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 Apache Hadoop's official documentation [3]. When following the tutorial, the student needs to take screenshots of the installation and verify if Hadoop is installed correctly. The shell/terminal screenshots need to have your Student ID on them explicitly. To ensure the authenticity of the installation process, it is imperative that each screenshot includes a ***timestamp*** indicating when the action was performed. Additionally, the screenshots should display pertinent information such as your ***computer's hostname*** and ***user account*** details. Furthermore, it is advisable to have a ***document open on your computer, with your name, student ID, and the current date clearly visible,*** positioned alongside the terminal or the browser window. These measures are crucial for validating that the task was executed on your personal computer.

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 member must have a mutual understanding to help each other during this lab assignment.

2 Paper Reading

2.1 Requirements

The student needs to read the original paper of MapReduce [5] paper by Dean and Ghemawat 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 Word Count (1pt)

3.1.1 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.

3.1.2 Expected outputs

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

3.2 Bigrams count (1pt)

Enhance the WordCount code to include the capability of counting bigrams, which are pairs of adjacent words in the text. For instance, in the sentence "Enhance the WordCount code to include …" the bigrams would be "Enhance the", "the WordCount", "WordCount code", "code to", and so on

4 Bonus: Setting up Fully Distributed Mode

4.1 Hadoop Cluster Setup in Non-Secure Mode (0.5 pt)

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 Research about Security in Hadoop Set-up (0.5 pt)

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 secure? 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 3.3.6 – Hadoop Cluster Setup. https://hadoop.apache.org/docs/current/hadoopproject-dist/hadoop-common/ClusterSetup.html.

[2] Apache Hadoop 3.3.6 – Hadoop in Secure Mode. https://hadoop.apache.org/docs/current/hadoopproject-dist/hadoop-common/SecureMode.html.

[3] Apache Hadoop 3.3.6 – Hadoop: Setting up a Single Node Cluster. https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html.

[4] Apache Hadoop 3.3.6 – 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 (research.google) 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-securityarchitecture.