

---

## **Lab 01: A Gentle Introduction to Hadoop**

CSC14118 Introduction to Big Data 20KHMT1

The Girls

2023-02-17

# Contents

<b>1</b>	<b>Lab 01: A Gentle Introduction to Hadoop</b>	<b>2</b>
1.1	Setting up Single-node Hadoop Cluster . . . . .	2
1.1.1	20127011 <b>Le Tan Dat</b> . . . . .	2
1.1.2	20127458 <b>Dang Tien Dat</b> . . . . .	2
1.1.3	20127438 <b>Le Nguyen Nguyen Anh</b> . . . . .	4
1.1.4	20127627 <b>Nguyen Quoc Thang</b> . . . . .	4
1.2	Introduction to MapReduce . . . . .	4
1.3	Running a warm-up problem: Word Count . . . . .	5
1.4	Bonus . . . . .	5
1.4.1	Extended Word Count: Unhealthy relationships . . . . .	5
1.4.2	Setting up Fully Distributed Mode . . . . .	5
1.4.2.1	Hadoop Cluster Setup in Non-Secure Mode . . . . .	5
1.4.2.2	Research about Security in Hadoop Set-up . . . . .	5
1.5	References . . . . .	8

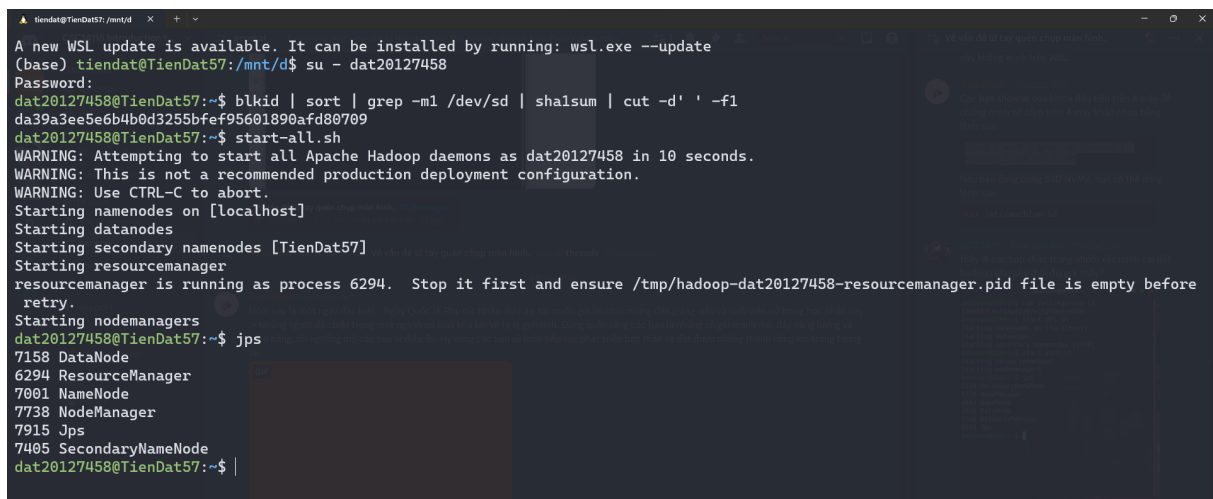
# 1 Lab 01: A Gentle Introduction to Hadoop

## 1.1 Setting up Single-node Hadoop Cluster

Verify hadoop installation for each member of the group

### 1.1.1 20127011 Le Tan Dat

### 1.1.2 20127458 Dang Tien Dat



```
A new WSL update is available. It can be installed by running: wsl.exe --update
(base) tiendat@TienDat57:/mnt/d$ su - dat20127458
Password:
dat20127458@TienDat57:~$ blkid | sort | grep -ml /dev/sd | shasum | cut -d' ' -f1
da39a3ee5e6b4b0d3255bfef95601890afd80709
dat20127458@TienDat57:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as dat20127458 in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [TienDat57]
Starting resourcemanager
resourcemanager is running as process 6294. Stop it first and ensure /tmp/hadoop-dat20127458-resourcemanager.pid file is empty before
retry.
Starting nodemanagers
dat20127458@TienDat57:~$ jps
7158 DataNode
6294 ResourceManager
7001 NameNode
7738 NodeManager
7915 Jps
7405 SecondaryNameNode
dat20127458@TienDat57:~$ |
```

Figure 1.1: 01-20127458

```
NguyenAnh20127438@master:~ (0.082s)
blkid | sort | grep -m1 /dev/sd | sha1sum | cut -d' ' -f1
15d8605d386c2871db39133d54906f37f44de0e9

NguyenAnh20127438@master:~ (0.281s)
jps
10582 SecondaryNameNode
10331 NameNode
11084 Jps

NguyenAnh20127438@master:~
```

**Figure 1.2:** 01-20127438

```
docker exec -it hadoop-namenode bash
root@nqthang_20127627:/# blkid | sort | grep -m1 /dev/sd | sha1sum | cut -d' ' -f1
da39a3ee5e6b4b0d3255bfef95601890afd80709
root@nqthang_20127627:/# jps
772 Jps
357 NameNode
root@nqthang_20127627:/#
```

**Figure 1.3:** 01-20127627

### 1.1.3 20127438 Le Nguyen Nguyen Anh

### 1.1.4 20127627 Nguyen Quoc Thang

## 1.2 Introduction to MapReduce

**This a section we will answer the following questions: 1. How do the input keys-values, the intermediate keys-values, and the output keys-values relate? Answer:** In a MapReduce job, the input keys-value (represent the input data that needs to be processed) are processed by a map function to produce intermediate key-value pairs (the value represent the data that is associated with each key). These intermediate key-value pairs are then sorted by key and passed on to the reduce function, which groups the values associated with each intermediate key and produces the final output key-value pairs.

2. **How does MapReduce deal with node failures? Answer:** MapReduce handles the fault node in a fault tolerant manner. When a node fails during execution, the tasks running on the node are automatically rescheduled to run on other nodes in the cluster. There are 2 mechanisms in MapReduce to handle node errors including: Speculative Execution and Task Tracking as follow.

- **Speculative Execution:** MapReduce can launch duplicate copies of a task on different nodes to ensure that at least one copy of the task completes successfully. If one of the nodes fails or is slow to complete its task, the duplicate copy can take over and complete the work.
- **Task Tracking:** MapReduce tracks the progress of each task and can detect when a task is taking too long to complete. If a task is taking too long, MapReduce can launch a duplicate copy of the task on a different node. If the duplicate copy completes successfully, the original task is killed.

3. **What is the meaning and implication of locality? What does it use? Answer:**

- **Meaning of locality:** Locality in Hadoop refers to the ability to process data on the same node or machine where the data is stored, in order to avoid data transmission between nodes in the network.
- **Implication of locality:** Locality is an important feature of Hadoop to optimize data processing performance by minimizing the time it takes to transmit data over the network.
- **Used for:** Locality is used in Hadoop to optimize data processing performance by ensuring that data processing tasks are performed on the same node where the data is stored. This helps to minimize the time it takes to transmit data over the network and improve data processing performance in Hadoop.

**4. Which problem is addressed by introducing a combiner function to the MapReduce model?**

**Answer:** The introduction of a combiner function to the MapReduce model addresses the problem of excessive data shuffling and network traffic during the Reduce phase. The combiner function is used to reduce the amount of data that needs to be transferred between the Map and Reduce tasks in a MapReduce job. The combiner function is executed on the output of the Map task on each node before the data is transferred to the Reduce task. The combiner function is optional and is only used if it reduces the amount of data that needs to be transferred between the Map and Reduce tasks.

## 1.3 Running a warm-up problem: Word Count

### 1.4 Bonus

#### 1.4.1 Extended Word Count: Unhealthy relationships

#### 1.4.2 Setting up Fully Distributed Mode

##### 1.4.2.1 Hadoop Cluster Setup in Non-Secure Mode

This section includes the machine id image of each machine

##### 1.4.2.2 Research about Security in Hadoop Set-up

This a section we will answer the following questions:

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

**Answer:** Yes my Hadoop is secured. Because I have to use the password to access the Hadoop. If I don't use the password, I can't access the Hadoop. So, I think my Hadoop is secured.

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

**Answer:** I think the authentication is better than authorization and encryption. Because the authentication is the first step to access the Hadoop. If I don't have the authentication, I can't access the Hadoop. So, I think the authentication is better than authorization and encryption.

Insert table example:

```

NguyenAnh20127438@master:~ (0.082s)
blkid | sort | grep -m1 /dev/sd | sha1sum | cut -d' ' -f1
15d86e5d386c2871db39133d54906f37f44de0e9

NguyenAnh20127438@master:~ (0.281s)
jps
10592 SecondaryNameNode
10331 NameNode
11084 Jps

NguyenAnh20127438@master:~
(NaDS) ~/Project (0.025s)
clear

NguyenAnh20127438@worker-0:~ (0.12s)
blkid | sort | grep -m1 /dev/sd | sha1sum | cut -d' ' -f1
15d86e5d386c2871db39133d54906f37f44de0e9

NguyenAnh20127438@worker-0:~ (0.32s)
jps
10032 NodeManager
10460 Jps
9869 DataNode

NguyenAnh20127438@worker-1:~ (0.14s)
blkid | sort | grep -m1 /dev/sd | sha1sum | cut -d' ' -f1
15d86e5d386c2871db39133d54906f37f44de0e9

NguyenAnh20127438@worker-1:~ (0.30s)
jps
9921 DataNode
10483 Jps
10086 NodeManager

NguyenAnh20127438@worker-2:~ (0.141s)
blkid | sort | grep -m1 /dev/sd | sha1sum | cut -d' ' -f1
15d86e5d386c2871db39133d54906f37f44de0e9

NguyenAnh20127438@worker-2:~ (0.364s)
jps
10097 NodeManager
9928 DataNode
10383 Jps

NguyenAnh20127438@worker-3:~ (0.136s)
blkid | sort | grep -m1 /dev/sd | sha1sum | cut -d' ' -f1
15d86e5d386c2871db39133d54906f37f44de0e9

NguyenAnh20127438@worker-3:~ (0.307s)
jps
9881 DataNode
10046 NodeManager
10334 Jps

```

Figure 1.4: Machine 20127438

```

root@nqthang_20127627:~# blkid | sort | grep -m1 /dev/sd | sha1sum | cut -d' ' -f1
da39a3ee5e6b4b0d3255bfef95601890afd80709
root@nqthang_20127627:~# jps
772 Jps
357 NameNode
root@nqthang_20127627:~#

root@b83862a2d201:~# blkid | sort | grep -m1 /dev/sd | sha1sum | cut -d' ' -f1
da39a3ee5e6b4b0d3255bfef95601890afd80709
root@b83862a2d201:~# jps
458 NodeManager
875 Jps
root@b83862a2d201:~#

root@2c63d72ce520:~# blkid | sort | grep -m1 /dev/sd | sha1sum | cut -d' ' -f1
da39a3ee5e6b4b0d3255bfef95601890afd80709
root@2c63d72ce520:~# jps
888 Jps
348 ResourceManager
root@2c63d72ce520:~#

root@a46f41421b751:~# blkid | sort | grep -m1 /dev/sd | sha1sum | cut -d' ' -f1
da39a3ee5e6b4b0d3255bfef95601890afd80709
root@a46f41421b751:~# jps
465 ApplicationHistoryServer
814 Jps
root@a46f41421b751:~#

root@4efd6abcc12:~# blkid | sort | grep -m1 /dev/sd | sha1sum | cut -d' ' -f1
da39a3ee5e6b4b0d3255bfef95601890afd80709
root@4efd6abcc12:~# jps
787 Jps
357 DataNode
root@4efd6abcc12:~#

```

Figure 1.5: Machine 20127627

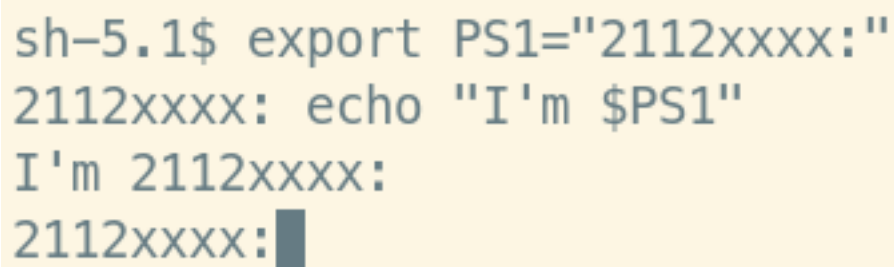
Server IP Address	Ports Open
192.168.1.1	<b>TCP:</b> 21,22,25,80,443
192.168.1.2	<b>TCP:</b> 22,55,90,8080,80
192.168.1.3	<b>TCP:</b> 1433,3389 <b>UDP:</b> 1434,161

Code example:

```
print("Hello")
```

```
cat ~/.bashrc
```

Screenshot example:



```
sh-5.1$ export PS1="2112xxxx:"  
2112xxxx: echo "I'm $PS1"  
I'm 2112xxxx:  
2112xxxx: █
```

**Figure 1.6:** Proof of change your shell prompt's name



Screenshot example:



**Figure 1.7:** ImgPlaceholder

Reference examples:

Some text in which I cite an author.<sup>1</sup>

More text. Another citation.<sup>2</sup>

What is this? Yet *another* citation?<sup>3</sup>

## 1.5 References

- Three Cloudera version of WordCount problem:
  - [https://docs.cloudera.com/documentation/other/tutorial/CDH5/topics/ht\\_wordcount1.html](https://docs.cloudera.com/documentation/other/tutorial/CDH5/topics/ht_wordcount1.html)
  - [https://docs.cloudera.com/documentation/other/tutorial/CDH5/topics/ht\\_wordcount2.html](https://docs.cloudera.com/documentation/other/tutorial/CDH5/topics/ht_wordcount2.html)
  - [https://docs.cloudera.com/documentation/other/tutorial/CDH5/topics/ht\\_wordcount3.html](https://docs.cloudera.com/documentation/other/tutorial/CDH5/topics/ht_wordcount3.html)
- Book: MapReduce Design Patterns [Donald Miner, Adam Shook, 2012]
- All of StackOverflow link related.

---

<sup>1</sup>So Chris Krycho, “Not Exactly a Millennium,” [chriskrycho.com](http://v4.chriskycho.com/2015/not-exactly-a-millennium.html), July 2015, <http://v4.chriskycho.com/2015/not-exactly-a-millennium.html> (accessed July 25, 2015)

<sup>2</sup>Contra Krycho, 15, who has everything *quite* wrong.

<sup>3</sup>ibid