


Assignment-2: MapReduce Programming

- Due Apr 2 by 11:59pm
- Points 10

In this assignment, you will write MapReduce code that counts the bigrams appeared in a text file, using the two design patterns taught in class: pairs and stripes.

To start, accept [this](https://classroom.github.com/a/42zFQW5y)  (<https://classroom.github.com/a/42zFQW5y>) assignment invitation and follow the instructions in `README.md` carefully.

Environment Setup

Considering that everyone's local environments may differ, potentially leading to various compatibility issues. We recommend launching an AWS instance (Instance Type: `t2.large`, OS: `Ubuntu 22.04`, Security Group: Allow all TCP connections) to configure the Hadoop environment. After setting up the instance, run two off-the-shelf scripts ([install_docker.sh](https://canvas.ust.hk/courses/56448/files/8775005?wrap=1) (<https://canvas.ust.hk/courses/56448/files/8775005?wrap=1>), [start_hadoop.sh](https://canvas.ust.hk/courses/56448/files/8775005/download?download_frd=1) (https://canvas.ust.hk/courses/56448/files/8775005/download?download_frd=1), [start_hadoop.sh](https://canvas.ust.hk/courses/56448/files/8789045?wrap=1) (<https://canvas.ust.hk/courses/56448/files/8789045?wrap=1>), [download?download_frd=1](https://canvas.ust.hk/courses/56448/files/8789045/download?download_frd=1)) to streamline the process.

```
$ bash install_docker.sh
```

```
$ sudo sh -c su
```

```
$ bash start_hadoop.sh
```

In a nutshell, the above commands will install Docker on the server and launch a Hadoop container (named `cloudera_quickstart`). The host's `/home/ubuntu` directory will be mapped to `/host` inside the container. You can clone the assignment repository to the host and execute commands within the container.


Use the `docker ps -a` command to verify if the Hadoop container is running properly.

To enter the container and run Hadoop-related commands, use `docker exec -it cloudera_quickstart bash` command.

You can access the Hadoop-related services by querying `<Public IPv4 DNS>:<port>`.

```
Port Service
2181 Zookeeper
7077 Spark Master
7180 Cloudera Manager
8020 HDFS
8022 SSH
8042 YARN Node Manager
8080 HBase Rest
8088 YARN Resource Manager
8888 Hue
8983 Solr
9090 HBase Thrift
11000 Oozie
16000 Sqoop Metastore
19888 MapReduce Job History
50030 MapReduce Job Tracker
50070 HDFS Rest Namenode
50075 HDFS Rest Datanode
60010 HBase Master
60030 HBase Region
```

For remote code development, we recommended using [VSCode](https://code.visualstudio.com)  (<https://code.visualstudio.com>) (w/ an extension "[Remote - SSH](https://marketplace.visualstudio.com/items?itemName=ms-vscode-remote.remote-ssh)"  (<https://marketplace.visualstudio.com/items?itemName=ms-vscode-remote.remote-ssh>)).

If you are unfamiliar with environment setup, you can refer to the steps demonstrated in [this demo video](https://hkust.zoom.us/rec/play/V5xzbinG2eEYn16ver0_iVOqnnepzsoJixWDQvgN2tU1LK_Ehq1mQdhTNxXoQvgHDB8TAToXBtQ_NtuD.TOC1OIP2hThkS9Ey2canPlayFromShare=true&from=share_recording_detail&continueMode=true&componentName=rec-play&originRequestUrl=https%3A%2F%2Fhkust.zoom.us%2Frec%2Fshare%2FFRAflAjNKRR1TA2AaUKoibycxNtRkK4TTN8STd-ZngkQLhsVxogq_t31oVpFRNuJ.Pm7E6y9r1spXJ4Vg)  (https://hkust.zoom.us/rec/play/V5xzbinG2eEYn16ver0_iVOqnnepzsoJixWDQvgN2tU1LK_Ehq1mQdhTNxXoQvgHDB8TAToXBtQ_NtuD.TOC1OIP2hThkS9Ey2canPlayFromShare=true&from=share_recording_detail&continueMode=true&componentName=rec-play&originRequestUrl=https%3A%2F%2Fhkust.zoom.us%2Frec%2Fshare%2FFRAflAjNKRR1TA2AaUKoibycxNtRkK4TTN8STd-ZngkQLhsVxogq_t31oVpFRNuJ.Pm7E6y9r1spXJ4Vg) for guidance.

Logging

It's important to note that the logs printed to the standard output by the tasks running on the nodes are not directly displayed in the terminal, but rather **saved to files**. A typical example of how to view the logs for a MapReduce program is as follows:

1. When you start the MapReduce program, you can find an application ID (e.g., `application_1710728478025_0002`).
2. You can locate the logs for that specific application in the `/usr/local/hadoop/logs/userlogs/` directory. Each MapReduce job will have several different log files, typically three: `stdout`, `stderr`, and `syslog`. The `stdout` file contains the logs printed to the standard output, such as those from `System.out.println()`. The `syslog` file contains the logs printed using `log4j`, and this is usually the primary reference for debugging.

