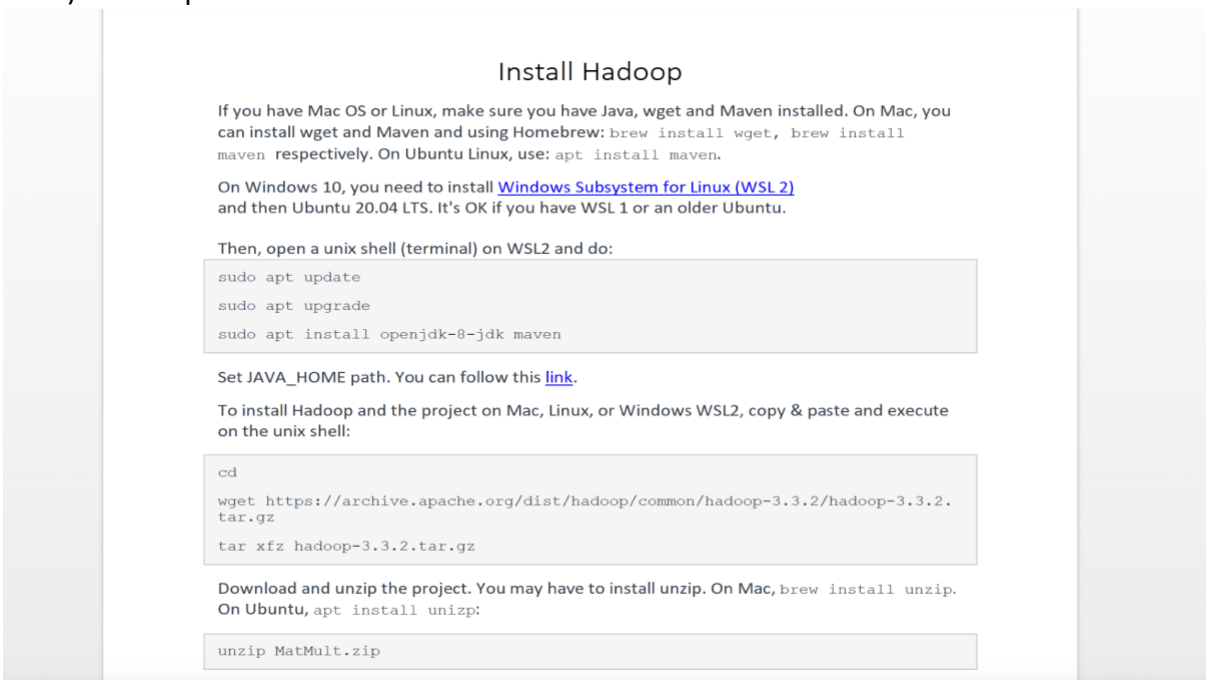
Hadoop and spark installation document

1) Hadoop installation



Install Hadoop

If you have Mac OS or Linux, make sure you have Java, wget and Maven installed. On Mac, you can install wget and Maven and using Homebrew: `brew install wget`, `brew install maven` respectively. On Ubuntu Linux, use: `apt install maven`.

On Windows 10, you need to install Windows Subsystem for Linux (WSL 2) and then Ubuntu 20.04 LTS. It's OK if you have WSL 1 or an older Ubuntu.

Then, open a unix shell (terminal) on WSL2 and do:

```
sudo apt update
sudo apt upgrade
sudo apt install openjdk-8-jdk maven
```

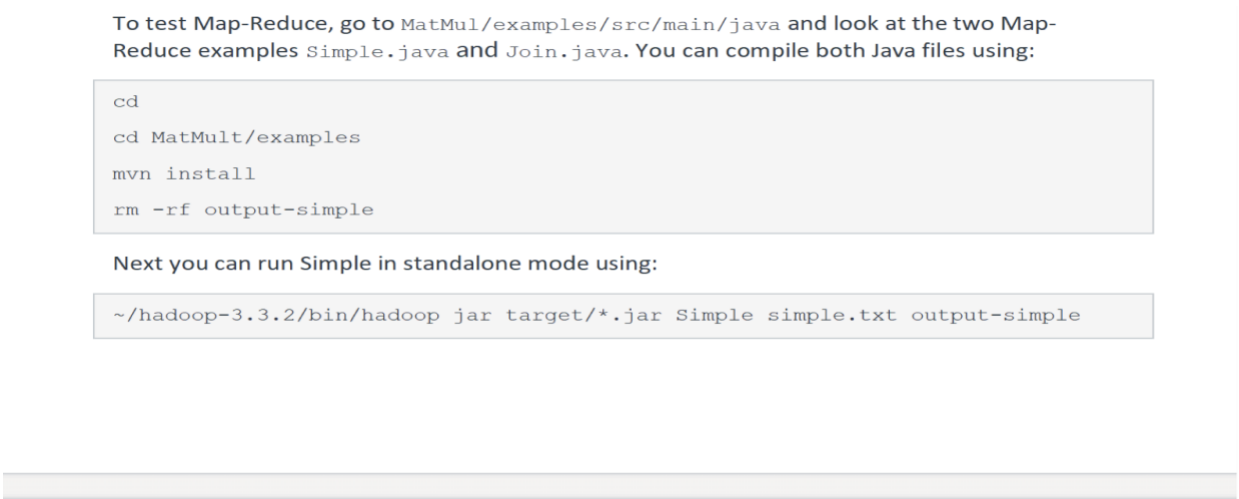
Set JAVA_HOME path. You can follow this link.

To install Hadoop and the project on Mac, Linux, or Windows WSL2, copy & paste and execute on the unix shell:

```
cd
wget https://archive.apache.org/dist/hadoop/common/hadoop-3.3.2/hadoop-3.3.2.tar.gz
tar xfz hadoop-3.3.2.tar.gz
```

Download and unzip the project. You may have to install unzip. On Mac, `brew install unzip`. On Ubuntu, `apt install unizp`:

```
unzip MatMult.zip
```

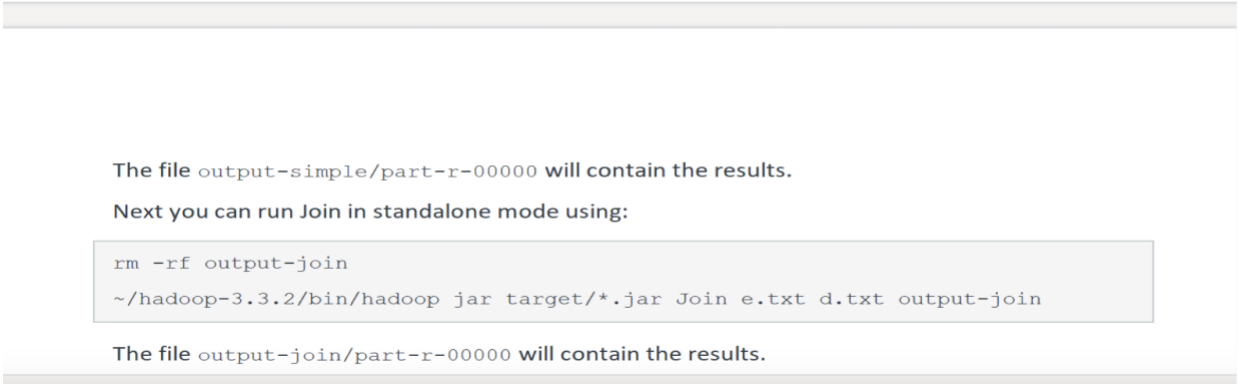


To test Map-Reduce, go to `MatMul/examples/src/main/java` and look at the two Map-Reduce examples `Simple.java` and `Join.java`. You can compile both Java files using:

```
cd
cd MatMult/examples
mvn install
rm -rf output-simple
```

Next you can run Simple in standalone mode using:

```
~/hadoop-3.3.2/bin/hadoop jar target/*.jar Simple simple.txt output-simple
```



The file `output-simple/part-r-00000` will contain the results.

Next you can run Join in standalone mode using:

```
rm -rf output-join
~/hadoop-3.3.2/bin/hadoop jar target/*.jar Join e.txt d.txt output-join
```

The file `output-join/part-r-00000` will contain the results.

2) Spark installation

Install Spark

To install Spark and the project on your laptop:

```
cd

wget https://archive.apache.org/dist/spark/spark-3.1.2/spark-3.1.2-bin-hadoop
3.2.tgz

tar xfz spark-3.1.2-bin-hadoop3.2.tgz
```

Untar, compile and run (example):

```
unzip SparkMatMul.zip

cd MatMul/examples

mvn install

rm -rf output

~/spark-3.1.2-bin-hadoop3.2/bin/spark-submit --class JoinSpark target/cse6331
-spark-examples-0.1.jar e.txt d.txt output
```