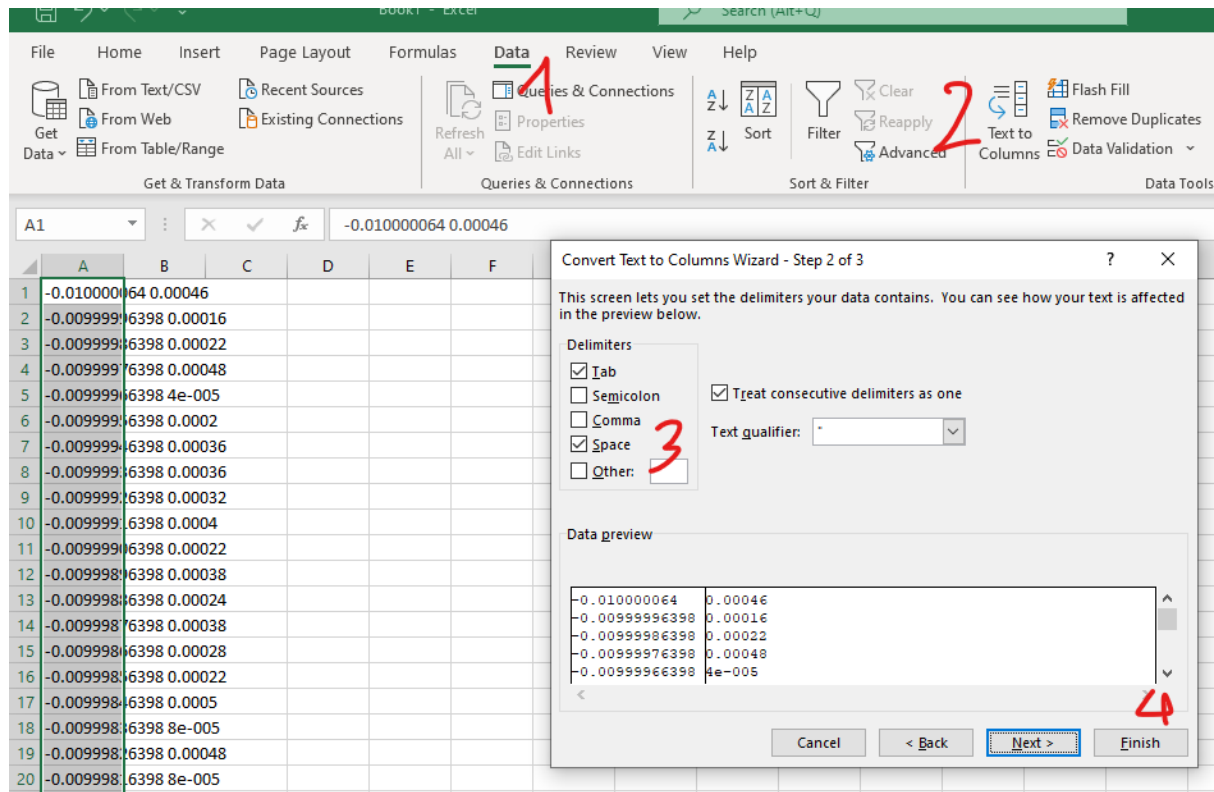


Name : Ha Thanh Nga  
Student's ID : 11202700

## Hadoop Mini Project

Firstly I used Excel to extract the data I want to aggregate



and got result

	A	B	C
1	-0.01	0.00046	
2	-0.01	0.00016	
3	-0.01	0.00022	
4	-0.01	0.00048	
5	-0.01	4.00E-05	
6	-0.01	0.0002	
7	-0.01	0.00036	
8	-0.01	0.00036	
9	-0.01	0.00032	
10	-0.01	0.0004	
11	-0.01	0.00022	
12	-0.01	0.00038	
13	-0.01	0.00024	
14	-0.01	0.00038	
15	-0.01	0.00028	
16	-0.01	0.00022	
17	-0.01	0.0005	
18	-0.01	8.00E-05	
19	-0.01	0.00048	
20	-0.01	8.00E-05	

After that I pasted the second column to new file (cksensor.txt)

cksensor - Note

File Edit Format

```

0.00046
0.00016
0.00022
0.00048
4.00E-05
0.0002
0.00036
0.00036
0.00032
0.0004
0.00022
0.00038
0.00024
0.00038
0.00028
0.00022
0.0005
8.00E-05
0.00048
8.00E-05
0.0005
8.00E-05
0.0004
0.00036

```

## Using the code

The screenshot shows an IDE with a Maven `pom.xml` file open. The `properties` section is visible, showing `<maven.compiler.target>1.8</maven.compiler.target>` and `<project.build.sourceEncoding>UTF-8</project.build.sourceEncoding>`. Below the editor, the terminal window displays the output of a Hadoop command-line execution. The output shows the Hadoop command-line option parsing not performed, and the application running in uber mode. The final output is a list of file format counters.

```

File Output Format Counters
Bytes Written=11557482
PS C:\Users\temp\IdeaProjects\MLprojects> hadoop jar target\MLProjects-1.0-SNAPSHOT.jar org.example.ba62.Main /input/cksensor.txt /output1
2023-10-29 23:42:55,092 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
2023-10-29 23:42:56,156 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
2023-10-29 23:42:58,350 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your ap
plication with ToolRunner to remedy this.
2023-10-29 23:42:58,397 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/temp/.staging/job_1698594228133_0002
2023-10-29 23:42:59,386 INFO mapred.FileInputFormat: Total input files to process : 1
2023-10-29 23:42:59,822 INFO mapreduce.JobSubmitter: number of splits:2
2023-10-29 23:43:00,619 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1698594228133_0002
2023-10-29 23:43:00,624 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-10-29 23:43:01,470 INFO conf.Configuration: resource-types.xml not found
2023-10-29 23:43:01,471 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-10-29 23:43:02,069 INFO impl.YarnClientImpl: Submitted application application_1698594228133_0002
2023-10-29 23:43:02,219 INFO mapreduce.Job: The url to track the job: http://thanhnga:8088/proxy/application_1698594228133_0002/
2023-10-29 23:43:02,223 INFO mapreduce.Job: Running job: job_1698594228133_0002
2023-10-29 23:43:42,576 INFO mapreduce.Job: Job job_1698594228133_0002 running in uber mode : false
2023-10-29 23:43:42,585 INFO mapreduce.Job: map 0% reduce 0%
2023-10-29 23:44:15,658 INFO mapreduce.Job: map 100% reduce 0%
2023-10-29 23:44:29,654 INFO mapreduce.Job: map 100% reduce 100%

```

Here is my result

## Browse Directory

/

Go!

Show

25

entries

Search:

<input type="checkbox"/>	<div><div></div></div> Permission	<div><div></div></div> Owner	<div><div></div></div> Group	<div><div></div></div> Size	<div><div></div></div> Last Modified	<div><div></div></div> Replication	<div><div></div></div> Block Size	<div><div></div></div> Name
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">temp</a>	<a href="#">supergroup</a>	0 B	Oct 29 23:56	<a href="#">0</a>	0 B	<div><div></div>input</div>
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">temp</a>	<a href="#">supergroup</a>	0 B	Oct 29 23:11	<a href="#">0</a>	0 B	<div><div></div>output</div>
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">temp</a>	<a href="#">supergroup</a>	0 B	Oct 29 23:44	<a href="#">0</a>	0 B	<div><div></div>output1</div>
<input type="checkbox"/>	<a href="#">drwxr-xr-x</a>	<a href="#">temp</a>	<a href="#">supergroup</a>	0 B	Oct 29 23:57	<a href="#">0</a>	0 B	<div><div></div>output2</div>
<input type="checkbox"/>	<a href="#">drwx-----</a>	<a href="#">temp</a>	<a href="#">supergroup</a>	0 B	Oct 29 23:09	<a href="#">0</a>	0 B	<div><div></div>tmp</div>

Can we skip to the good part

(The table about the frequency of the right column in the initial file(sensor.txt) )

The screenshot shows the Hadoop web interface with a modal window open for the file 'output2'. The modal displays block information and file contents.

**Block information**

- Block ID: 1073741853
- Block Pool ID: BP-838130265-192.168.20.1-1698593962309
- Generation Stamp: 1029
- Size: 15577
- Availability:
  - 192.168.20.1

**File contents**

```
-0.00002 763
-0.00004 572
-0.00006 428
-0.00008 390
-0.0001 1942
-0.00012 2042
-0.00014 2061
-0.00016 2007
```