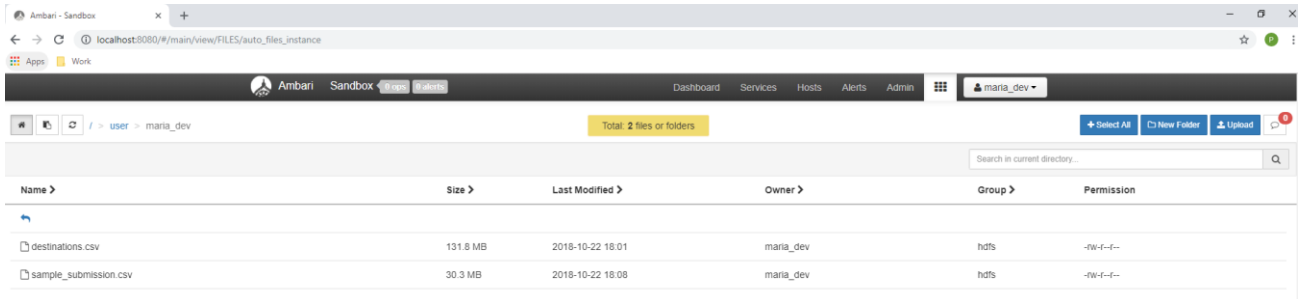


# Hadoop homework.

After trying several times to upload all the 4 files through Ambari UI, I found out that it's impossible to upload very big files through it, so, I was only able to upload the smaller ones of the 4:



Name	Size	Last Modified	Owner	Group	Permission
destinations.csv	131.8 MB	2018-10-22 18:01	maria_dev	hdfs	-rw-r--r--
sample_submission.csv	30.3 MB	2018-10-22 18:08	maria_dev	hdfs	-rw-r--r--

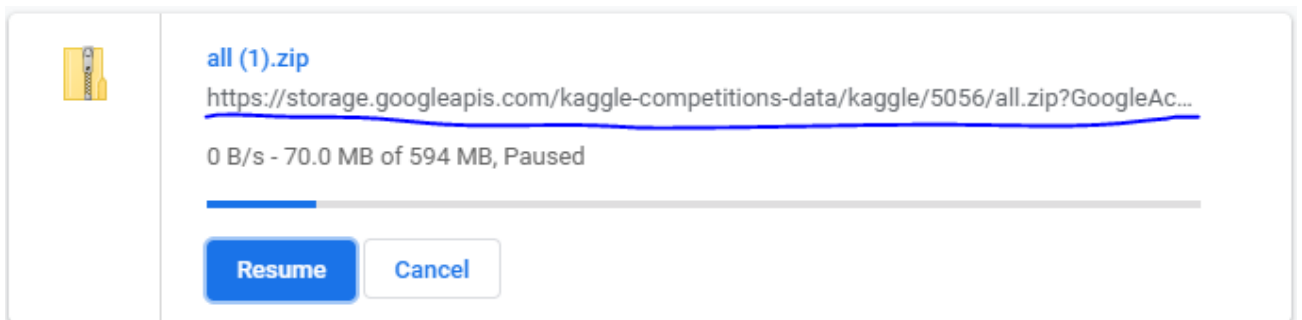
As it turns out, command line should be used to upload large files.

In order to solve this task, let's create a folder, e.g. "tempJunk":

```
[root@sandbox-hdp /]# mkdir tempJunk
[root@sandbox-hdp /]# ls tempJunk/
[root@sandbox-hdp /]#
```

Here we'll wget our files into the local system and then upload them into HDFS.

The link for download can be seen through google chrome:



```
[root@sandbox-hdp tempJunk]# wget --content-disposition -O test.zip 'https://storage.googleapis.com/kaggle-competitions-data/kaggle/5056/all.zip?GoogleAccessId=web-data@kaggle-161607.iam.gserviceaccount.com&Expires=1540478929&Signature=PmHWIYqvS0Wkmv5p0MXeraHBn%2FIYbwXFRub2d8e9DrxLhbI%2Fm2sFY1j3DUx4qSKioPh2Nr6xcLQHRB2UygWYCUFpFEfhXdURx5eKLFk8omtwVC%2FIhZObFQUh2DJMplamGzLKw8%2BoB0b4MejNt7bKZ5cMHKW9XtV5q4AH75Pd1WnkCOQbudeNAA049STLBBkrymHE8anrxRYudo0nSVFTMfEC3jff8KbW13JXq1D1FwmySvLsVdBCOAu%2F7Z6RzBN9gQwyV79oe3mjnP1tY8U3uTMuqDqQa1OBfF8CrOPJmCAPD1RZGgrwPJQqjLTXBqEKAwmk2%2B%2B4ZGfMcg7jjBhA%3D%3D'
```

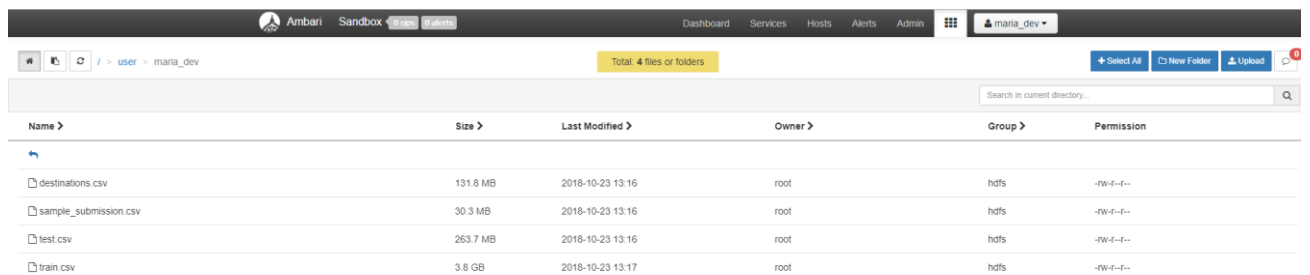
```
[root@sandbox-hdp tempJunk]# unzip test.zip
Archive:  test.zip
  inflating: destinations.csv.gz
  inflating: test.csv.gz
  inflating: sample_submission.csv.gz
  inflating: train.csv.gz
[root@sandbox-hdp tempJunk]#
```

Next, we can gunzip our .gz files and get csvs.

Next, in order to put our files into hdfs, we need to execute `hadoop fs -copyFromLocal file1 file2 ... fileN /path/to/folder`

```
[root@sandbox-hdp tempJunk]# ls
destinations.csv sample_submission.csv test.csv test.zip train.csv
[root@sandbox-hdp tempJunk]# hadoop fs -copyFromLocal sample_submission.csv test.csv destinations.csv train.csv /user/
maria_dev
[root@sandbox-hdp tempJunk]# hadoop fs -ls /user/maria_dev
Found 4 items
-rw-r--r-- 1 root hdfs 138159416 2018-10-23 10:16 /user/maria_dev/destinations.csv
-rw-r--r-- 1 root hdfs 31756066 2018-10-23 10:16 /user/maria_dev/sample_submission.csv
-rw-r--r-- 1 root hdfs 276554476 2018-10-23 10:16 /user/maria_dev/test.csv
-rw-r--r-- 1 root hdfs 4070445781 2018-10-23 10:17 /user/maria_dev/train.csv
[root@sandbox-hdp tempJunk]#
```

In the picture above we can see that our files are in HDFS. We can also see this in ambari UI:



Name	Size	Last Modified	Owner	Group	Permission
destinations.csv	131.8 MB	2018-10-23 13:16	root	hdfs	-rw-r--r--
sample_submission.csv	30.3 MB	2018-10-23 13:16	root	hdfs	-rw-r--r--
test.csv	263.7 MB	2018-10-23 13:16	root	hdfs	-rw-r--r--
train.csv	3.8 GB	2018-10-23 13:17	root	hdfs	-rw-r--r--