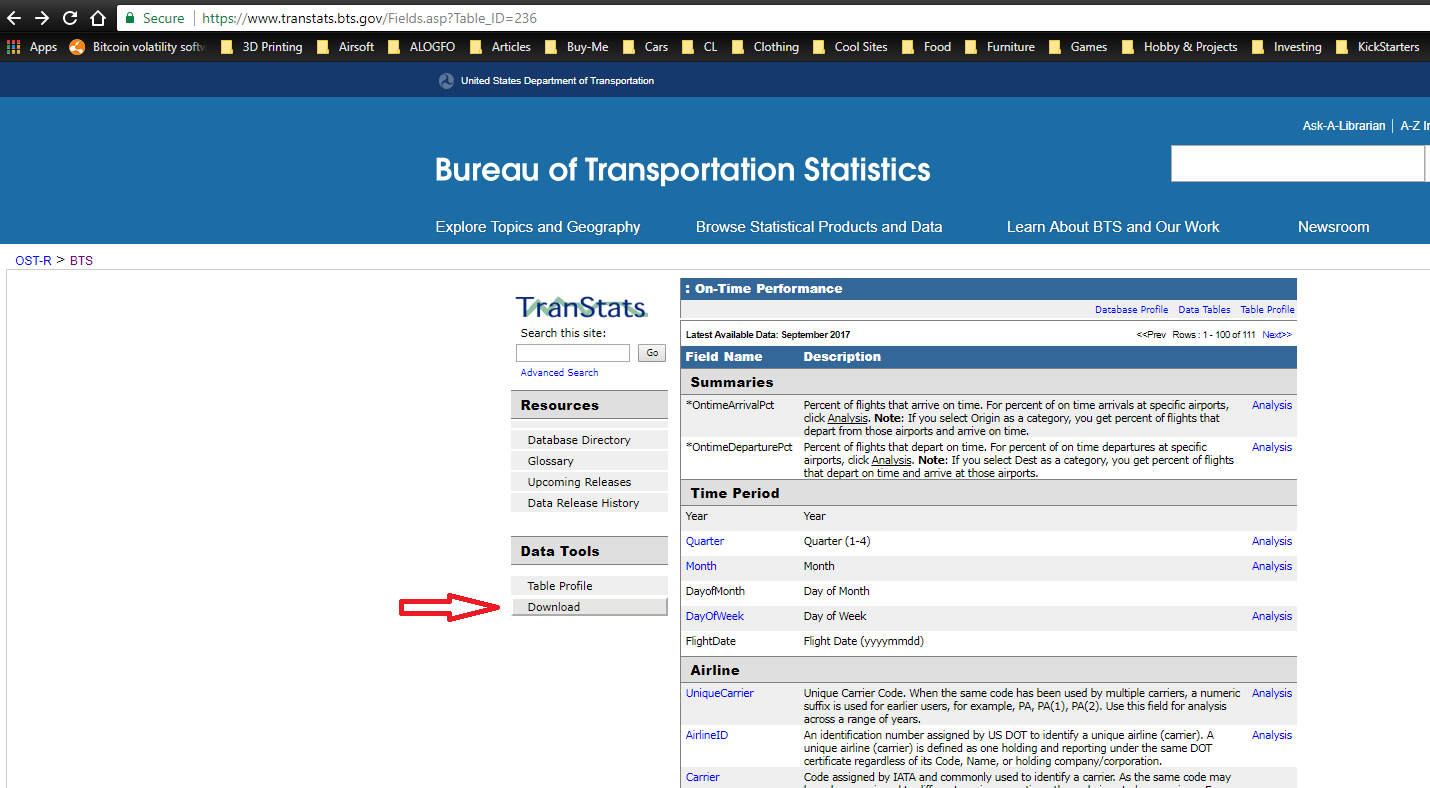
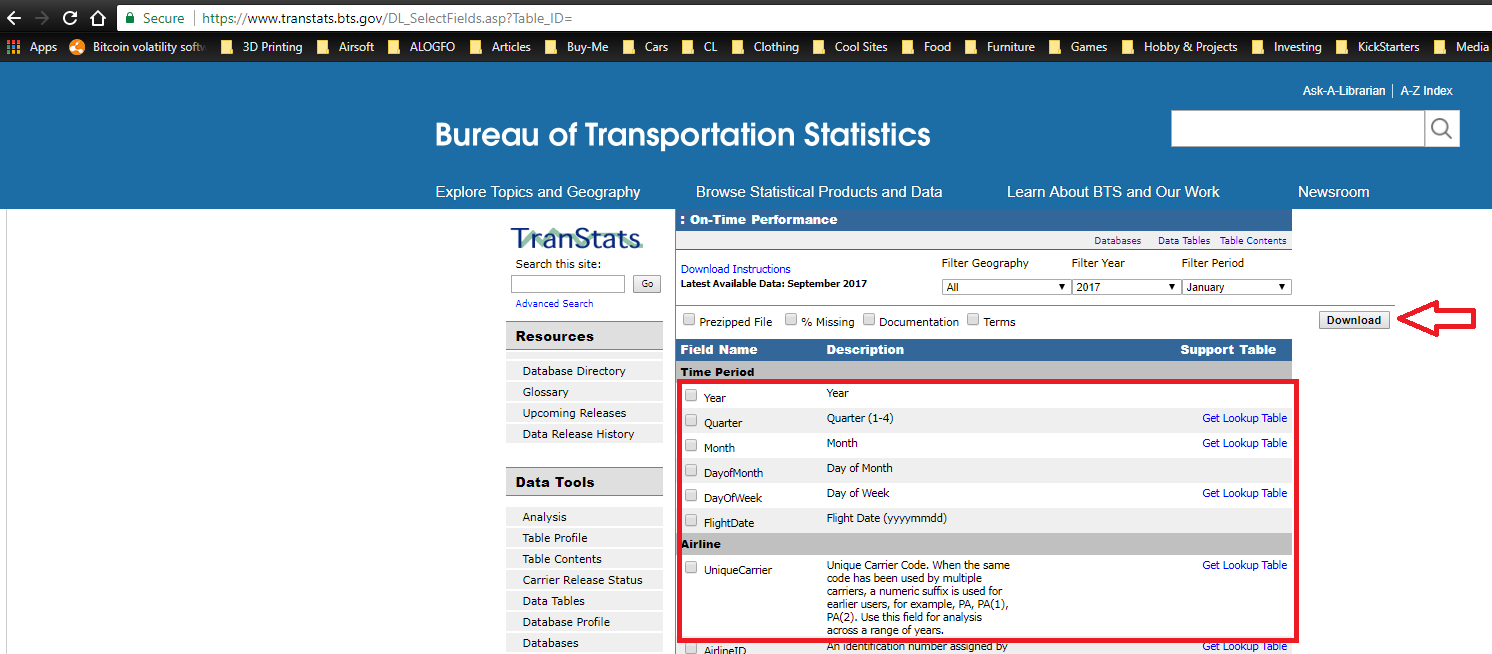
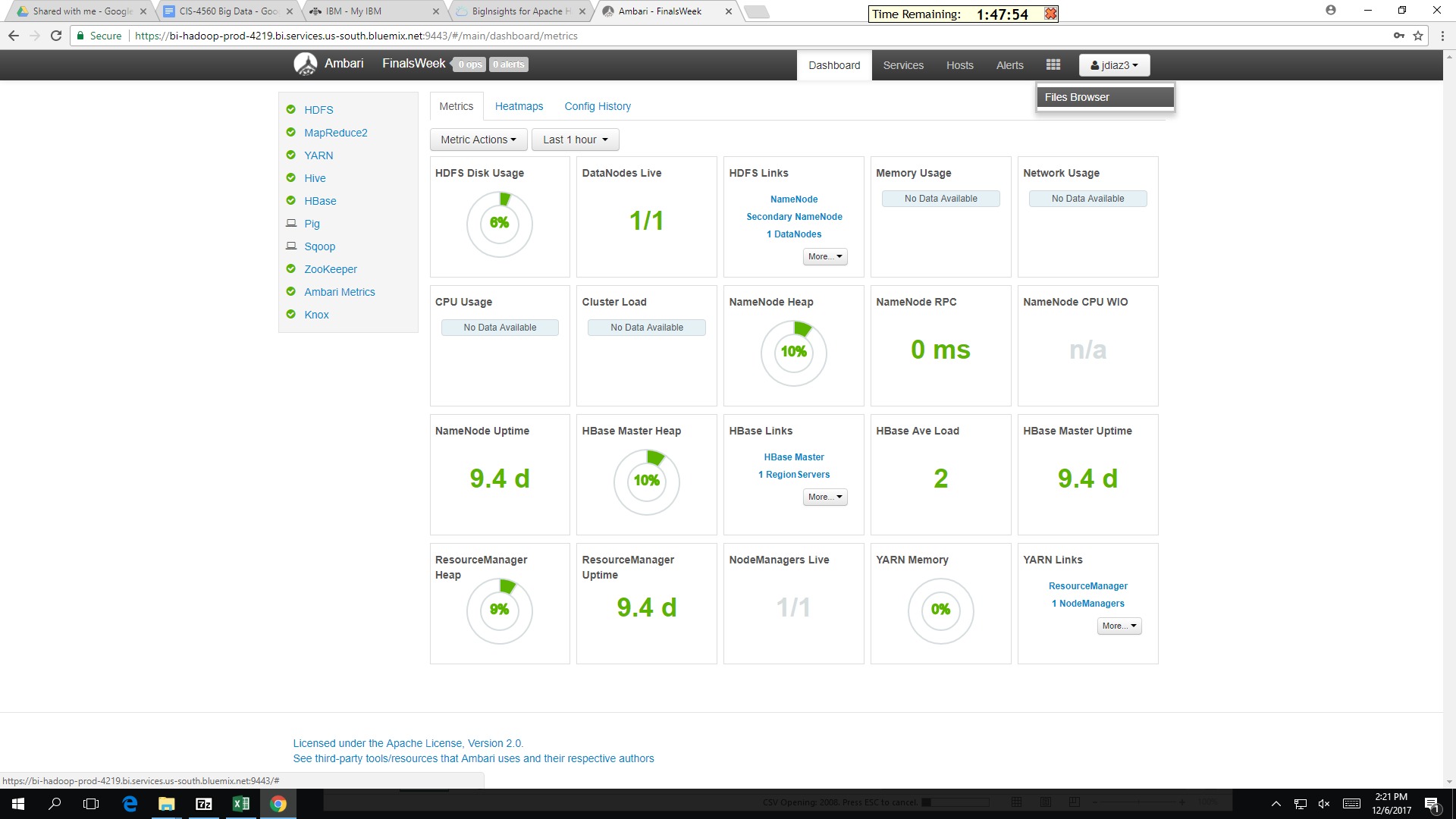
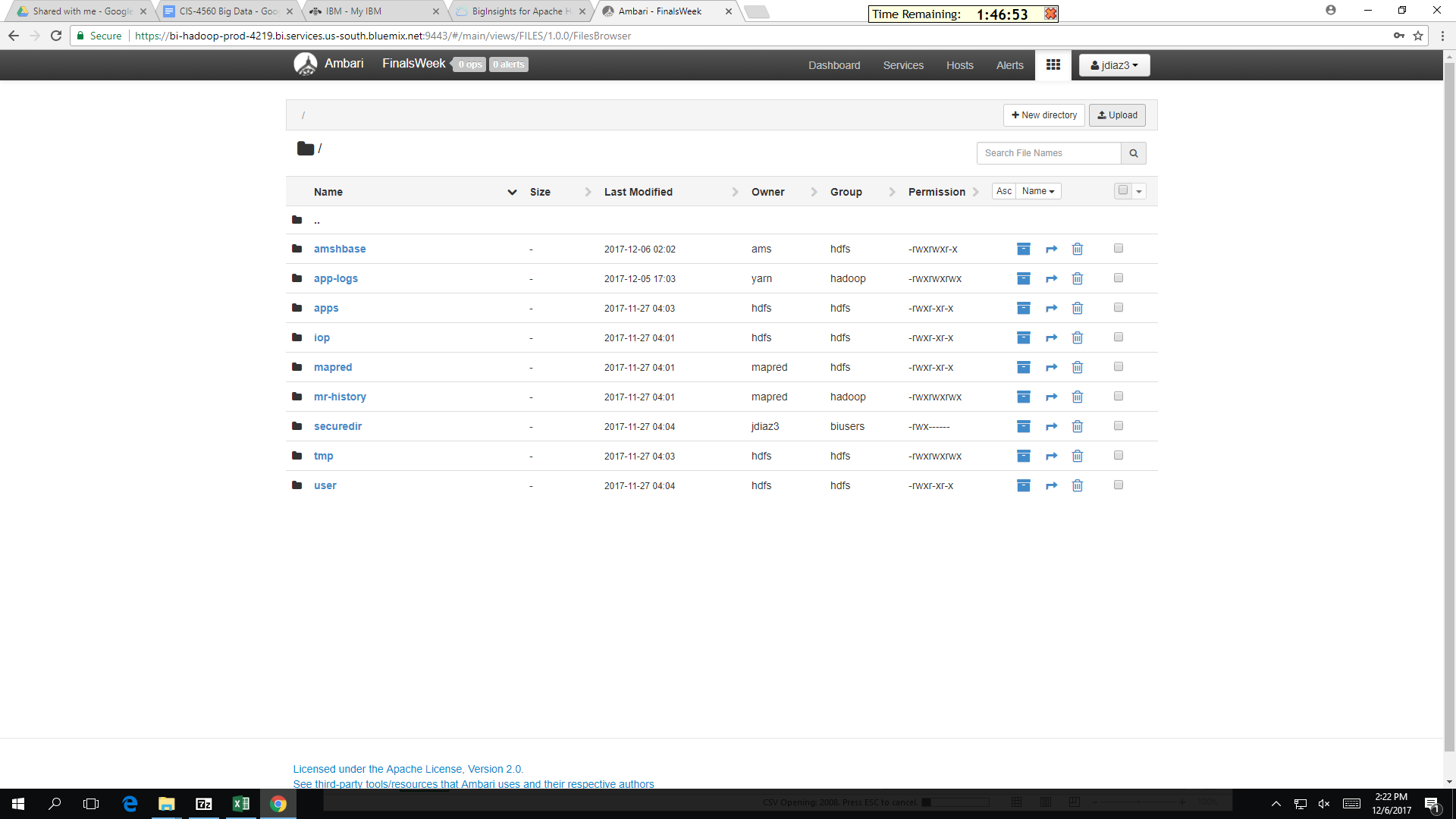
CIS-4560 Term Project Tutorial

1. Go to https://www.transtats.bts.gov/DL\_SelectFields.asp?Table\_ID= and click the download button shown.
2. Select all the fields and date ranges for the data you wish to work with, then click download.

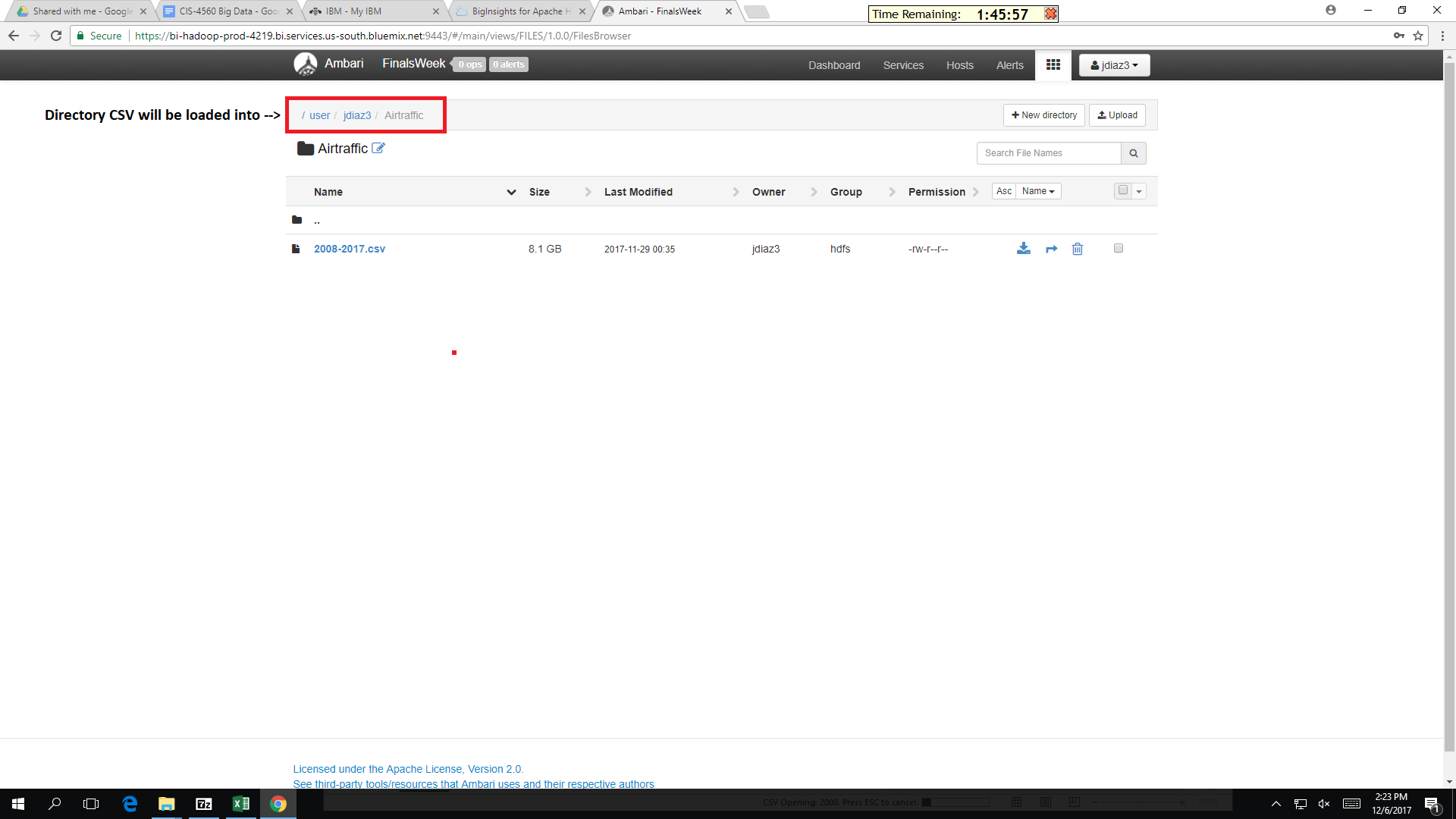
1. Next you must upload your data into HDFS. If you are hosting your data on a server you can simply use cURL or wGET. But for our example we will use the Ambari browser window. Login to BigInsights and click the button shown in the top right to open the Ambari file browser.

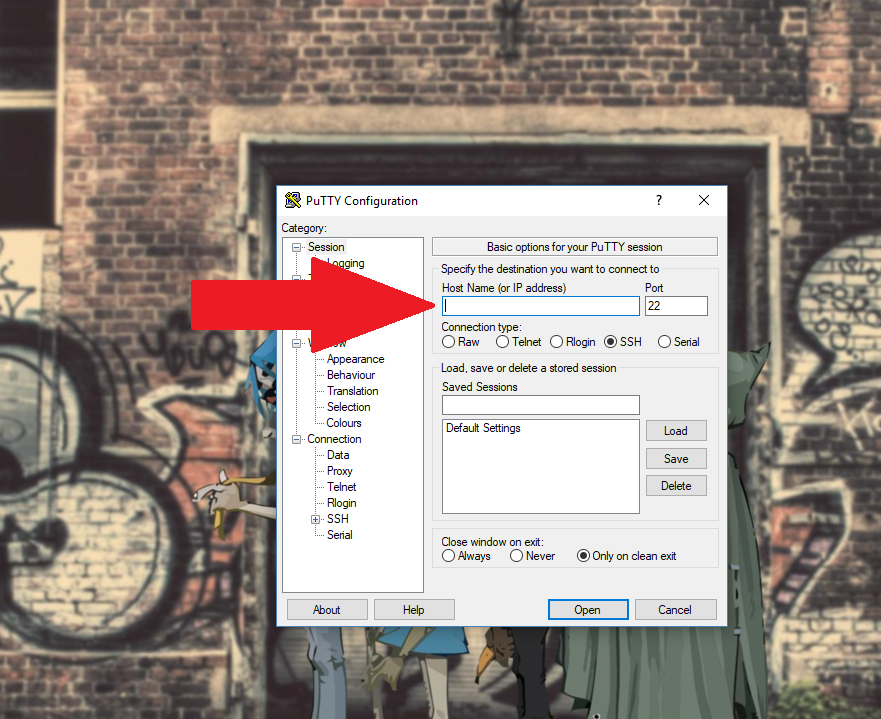
i

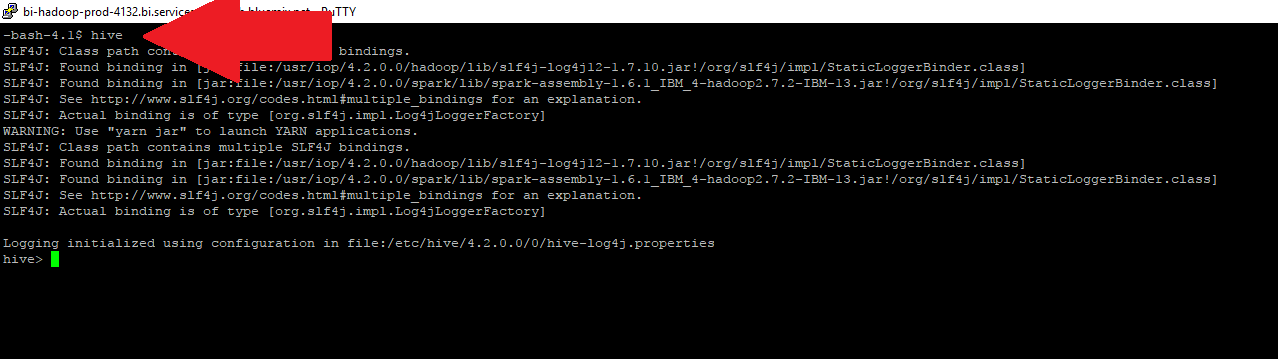
4. Navigate the the directory you want to upload your data to, most likely /user/<username>/<data directory>/. Then click the upload button in the top right corner.



5. Example of which directory the CSVs will uploaded to, in order to use them in a Hive External table.



6. Once the data CSV has been uploaded to HDFS through Ambari File Browser, we can login to our BigInsights Cluster through SSH. Input the URL to your cluster in your SSH client. If using PuTTY it should look like the picture below.

7. After logging in, type Hive into the command line to enable the Hive CLI, that will allow you to create the tables necessary.

8. Next input the following query to create a table for the data you uploaded to HDFS. Replace <username> with your user name and <data directory> with the name of directory you uploaded to.

**CREATE EXTERNAL TABLE IF NOT EXISTS AirtrafficData**

**(**

**year INT,month INT,day\_of\_month INT,day\_of\_week INT,unique\_carrier STRING,tail\_num STRING,fl\_num STRING,origin\_airport\_id STRING,dest\_airport\_id STRING,dest STRING,crs\_dep\_time INT,dep\_time INT,dep\_delay INT,taxi\_out INT,taxi\_in INT,crs\_arr\_time INT,arr\_time INT,arr\_delay INT,cancelled INT,cancellation\_code CHAR(1),diverted INT,crs\_elapsed\_time INT,actual\_elapsed\_time INT,air\_time INT,carrier\_delay INT,weather\_delay INT,nas\_delay INT,security\_delay INT,late\_aircraft\_delay INT**

**)**

**ROW FORMAT DELIMITED**

**FIELDS TERMINATED BY ','**

**STORED AS TEXTFILE**

**location '/user/<username>/<data directory>';**

**CREATE EXTERNAL TABLE IF NOT EXISTS IDs**

**(**

**Code STRING,**

**City STRING,**

**Location STRING,**

**Name String**

**)**

**COMMENT 'Airport Codes to Cities/Airports'**

**ROW FORMAT DELIMITED**

**FIELDS TERMINATED BY ','**

**STORED AS TEXTFILE**

**location '/user/<username>/<data directory>';**

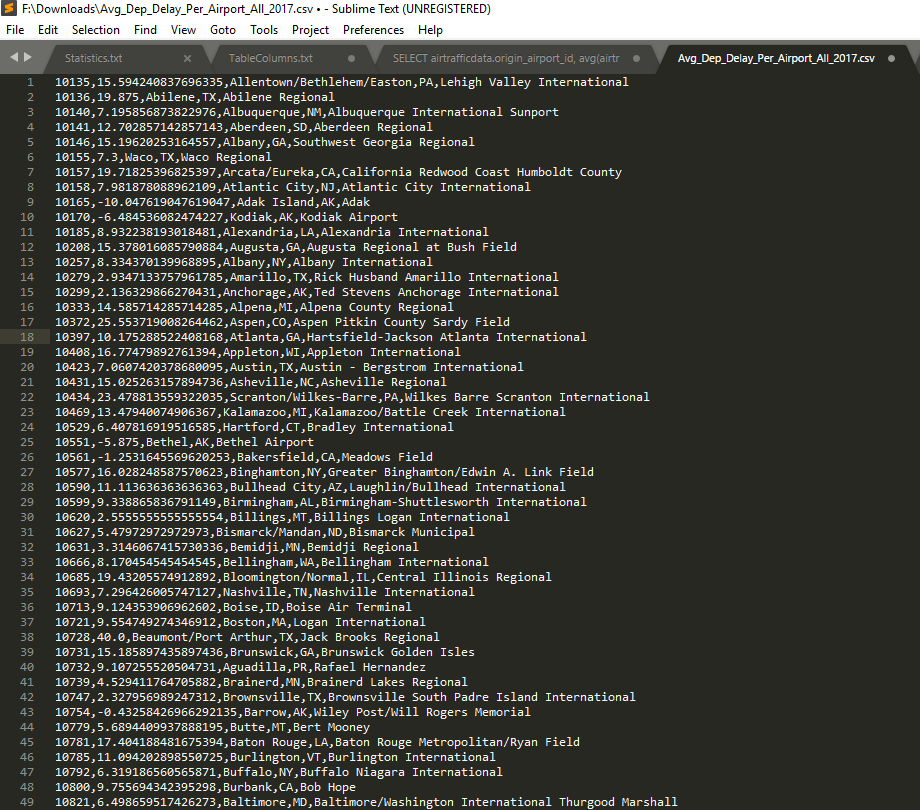
9. To verify the data is in place, using this query will display the average departure delay per airport in 2017.

**SELECT airtrafficdata.origin\_airport\_id, avg(airtrafficdata.dep\_delay), ids.city, ids.location, ids.name FROM airtrafficdata JOIN ids WHERE airtrafficdata.origin\_airport\_id = ids.code GROUP BY airtrafficdata.origin\_airport\_id, ids.city, ids.location, ids.name;**

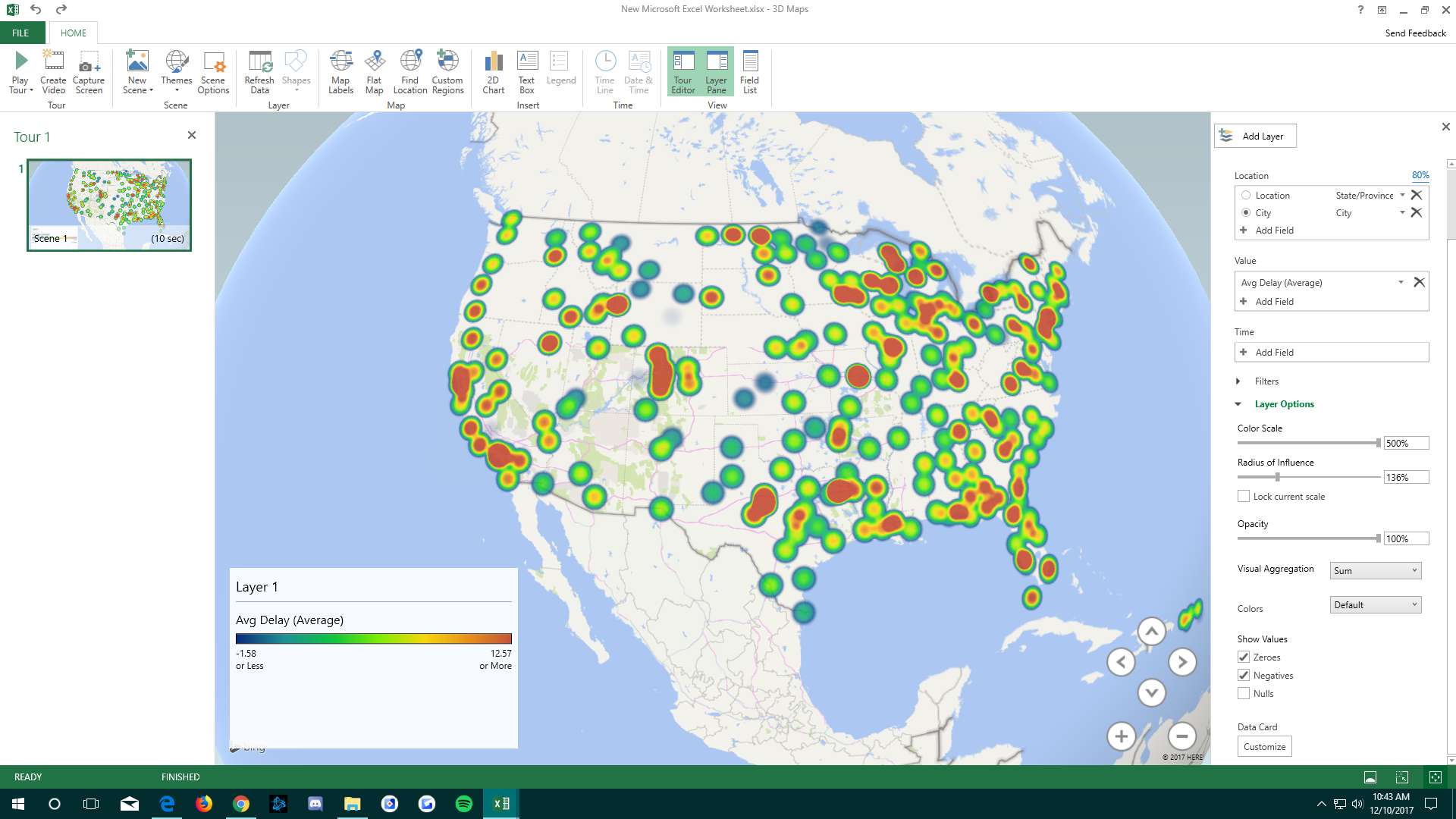
You can get CSV output in with this version

**INSERT OVERWRITE DIRECTORY '/user/<username>/<data directory>';' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE SELECT airtrafficdata.origin\_airport\_id, avg(airtrafficdata.dep\_delay), ids.city, ids.location, ids.name FROM airtrafficdata JOIN ids WHERE airtrafficdata.origin\_airport\_id = ids.code GROUP BY airtrafficdata.origin\_airport\_id, ids.city, ids.location, ids.name;**

10. Output should look like this once downloaded from Ambari.



11. Excel Visualization of average departure delay for airports in the US



**REFERENCES**

* Data URL: <https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=>
* GIthub URL: <https://github.com/diazjordi/CIS-4560-AirTraffic-Data>
* Reference Materials:
* <https://cwiki.apache.org/confluence/display/Hive/LanguageManual>
* <https://www.ibm.com/support/knowledgecenter/v1/sso/login?redirect_url=https%3A%2F%2Fwww.ibm.com%2Fsupport%2Fknowledgecenter%2Fen%2FSSPT3X_4.2.0%2Fcom.ibm.swg.im.infosphere.biginsights.welcome.doc%2Fdoc%2Fwelcome.html>
* <https://docs.hortonworks.com/HDPDocuments/Ambari/Ambari-2.2.2.0/index.html>