Mass Street Analytics

Hands On: Introduction to the Hadoop Ecosystem

Lab Directions

Lab 1: Get data Into HDFS with Ambari

1. Go to files view in Ambari.
2. In the maria_dev directory, create a new directory called data.
3. Enter data directory and click upload. You can only upload one file at a time. Find and upload, geolocation.csv and trucks.csv.
4. Step back out to the data directory and make sure you have full control on the folder permissions.

Lab 2: Get data into Hive

1. Go to Ambari dashboard.
2. Go into the Hive service.
3. Click Configs.
4. Check to see if Tez is the execution engine.
5. Go to Files View.
6. Load the .tsv files into the maria_dev data directory.
7. Go to Hive View.
8. Open the exercise file “load data into hive.sql”. Run the create table scripts one at a time.
9. Click mydatabase and look at the table schema.
10. Run load data scripts.
11. Click mydatabase. Click the icon next to the table name to view top 100 records.
12. LOAD DATA is an ETL operation. Go back and check the data directory to see that the files are gone.
13. Using the CREATE VIEW script, save that for later use.
14. Go to save queries and execute the script you just saved.
15. Run the view.
16. Run the join query and see the results using the things you’ve learned so far.
17. Download the results of the Join query by using save results.

Lab 3: Big Data BI with Zeppelin

1. Go to Ambari dashboard and check the Zeppelin Service.
2. From Quick Links, open the Zeppelin UI.
3. Import ClickstreamAnalytics.json. (Hint: It’s not drag and drop.)
4. Run the first cell with a query in it.
5. Click settings so you can see what goes into the viz.
6. Drag state to the Values field.
7. Play around with the various graphs.
8. Run the next query.
9. Make sure age is in keys, gender_cd is in Groups, category is in values as a count.
10. Play around with the last cell and see what you can find.

Lab 4: Your turn to play

1. Using what you’ve learned so far, us the data to explore HDFS, Hive, and Zeppelin.
2. Try to combine the two files using the files view.
3. Using your existing SQL skills see what you can find using Hive view.
4. Hint: For table creation follow this guide on datatypes: https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Types