Scalable Twitter Analytics

A Hadoop framework for spatiotemporal social media analytics

Ben Liebersohn

Many researchers are interested in finding better ways of exploring boundaries between physical spaces and digital spaces. Social media users continue the trend towards sharing more information regarding their activities, views, and interests, and even share information relevant to crime monitoring, civil unrest, and human mobility patterns.

With a scalable social media event detection model, we can better extract the relevant interaction between cyber-events and real-world phenomena. Because over 300 Million Tweets are published annually, sifting through such an enormous database has proved non-trivial. Through database size reduction, this Hadoop tool makes it possible to examine a reduced size database, making more computationally intensive datamining possible.

1. Preliminaries

   • Twitter gives a bounding box of four sets of GPS coordinates, whose bounding box (w, h) encapsulates the specified place. Larger bounding boxes are less useful.
   • My Hadoop code, written in Java, begins by taking user specified time boundaries. In most of my experiments, I selected data by month. The existing HDFS structure I am using makes this task trivial.

2. Results

   • Once logged into a Hadoop node on ROGER, we instruct Hadoop on how to run the job.
   • supply the location of the data in HDFS (ADDRESS="$HOME/twitter_analysis")
   • set the upper bound of reducers (MAXREDUCERS=64)
   • set the time range DATE by month (DATE=201601 is December, 2016)
   • Next, we can begin extracting Tweets. First, let’s go over some pseudocode

3. Conclusions

   • Once logged into a Hadoop node on ROGER, we instruct Hadoop on how to run the job.
   • supply the location of the data in HDFS (ADDRESS="$HOME/twitter_analysis")
   • set the upper bound of reducers (MAXREDUCERS=64)
   • set the time range DATE by month (DATE=201601 is December, 2016)
   • Next, we can begin extracting Tweets. First, let’s go over some pseudocode
   • For each file i containing Tweets
   • Input IN=\$DATE[i], and OUT=tweets/$DATE[i]
   • create generic filesystem fi and recursively delete directory
   • Now we can run a mapreduce on the code
   • Instruct Hadoop where our bounded code is (hadoop jar \$HADOOP\$PROCESS.jar, \$IN, \$OUT, \$MAXREDUCERS, bounding GPS coordinates, "true" for messages in our domain

   • Last, we recursively delete the fi \$IN (hadoop fs -rmr $IN)

   • Once logged into a Hadoop node on ROGER, we instruct Hadoop on how to run the job.
   • supply the location of the data in HDFS (ADDRESS="$HOME/twitter_analysis")
   • set the upper bound of reducers (MAXREDUCERS=64)
   • set the time range DATE by month (DATE=201601 is December, 2016)
   • Next, we can begin extracting Tweets. First, let’s go over some pseudocode
   • For each file i containing Tweets
   • Input IN=\$DATE[i], and OUT=tweets/$DATE[i]
   • create generic filesystem fi and recursively delete directory
   • Now we can run a mapreduce on the code
   • Instruct Hadoop where our bounded code is (hadoop jar \$HADOOP\$PROCESS.jar, \$IN, \$OUT, \$MAXREDUCERS, bounding GPS coordinates, "true" for messages in our domain

   • Last, we recursively delete the fi \$IN (hadoop fs -rmr $IN)

   • Once logged into a Hadoop node on ROGER, we instruct Hadoop on how to run the job.
   • supply the location of the data in HDFS (ADDRESS="$HOME/twitter_analysis")
   • set the upper bound of reducers (MAXREDUCERS=64)
   • set the time range DATE by month (DATE=201601 is December, 2016)
   • Next, we can begin extracting Tweets. First, let’s go over some pseudocode
   • For each file i containing Tweets
   • Input IN=\$DATE[i], and OUT=tweets/$DATE[i]
   • create generic filesystem fi and recursively delete directory
   • Now we can run a mapreduce on the code
   • Instruct Hadoop where our bounded code is (hadoop jar \$HADOOP\$PROCESS.jar, \$IN, \$OUT, \$MAXREDUCERS, bounding GPS coordinates, "true" for messages in our domain

   • Last, we recursively delete the fi \$IN (hadoop fs -rmr $IN)