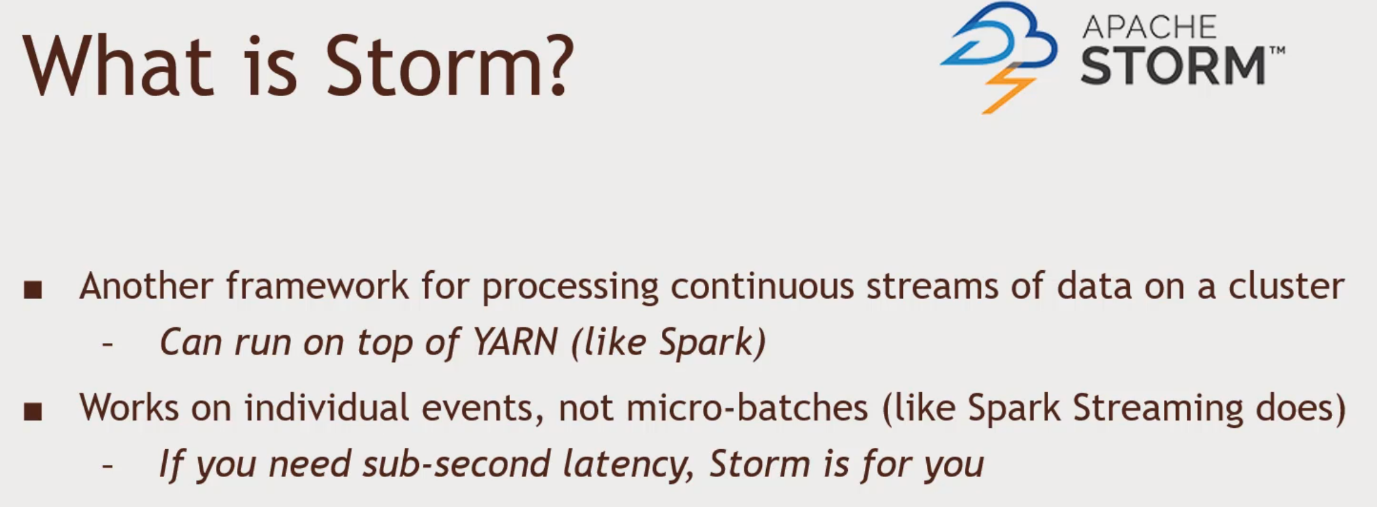
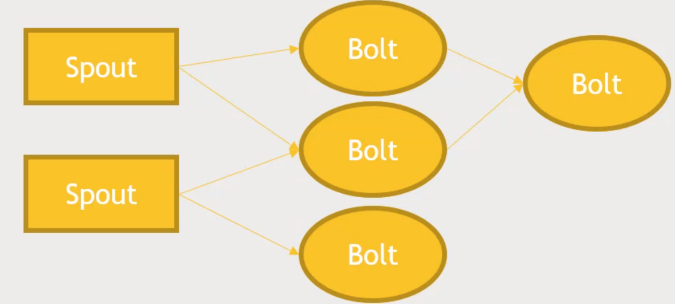
Apache Storm – Real-time stream processing

Storm works on individual events. Not batches.

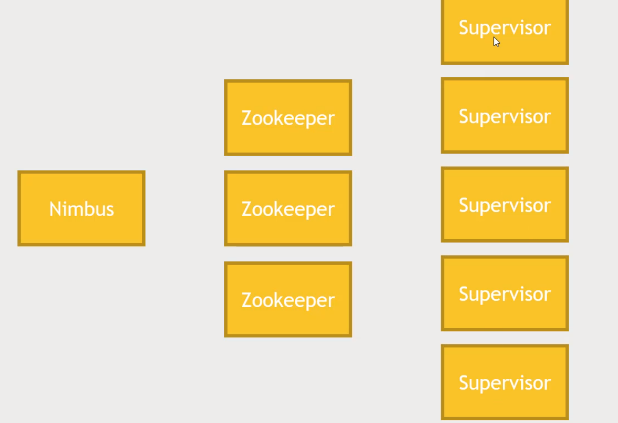
Storm terminology

* A stream consist of tuples that flow through..
* Spouts that are sources of stream data (Kafka,Twitter,etc.)
* Bolts that process stream data as it’s received
  + Transform, aggregate, write to databases/ HDFS
* A topology is a graph of spouts and bolts that process your stream

Storm architecture

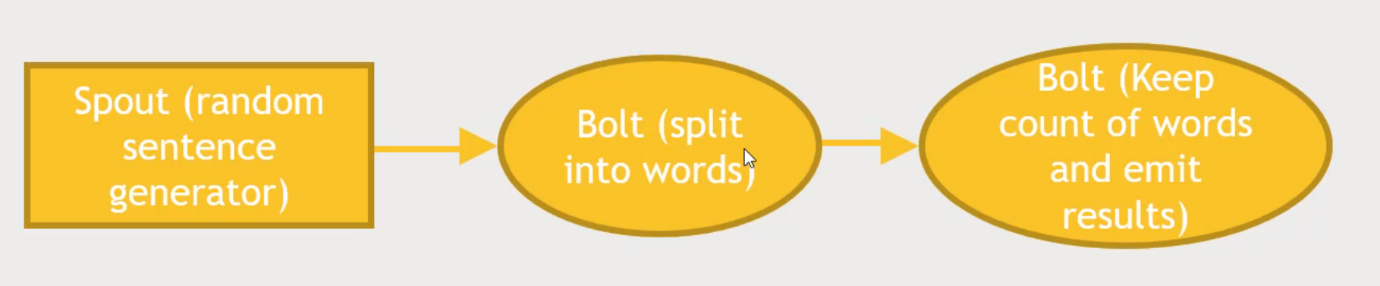
Nimbus Node Server which is a Job Tracker -Single point of failure but not that of a big deal

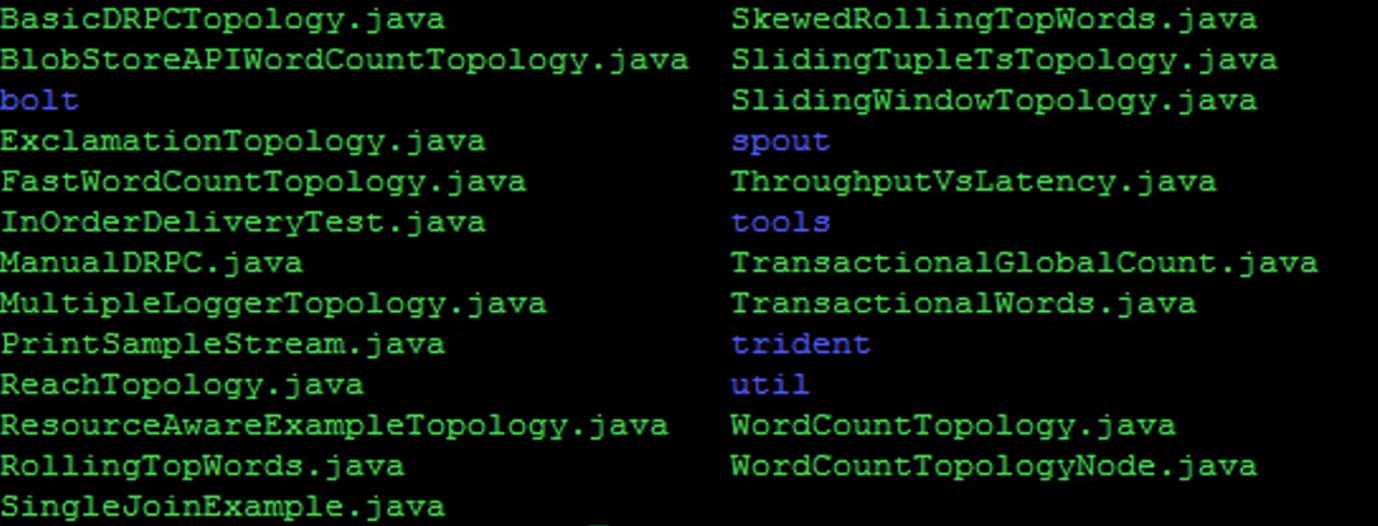
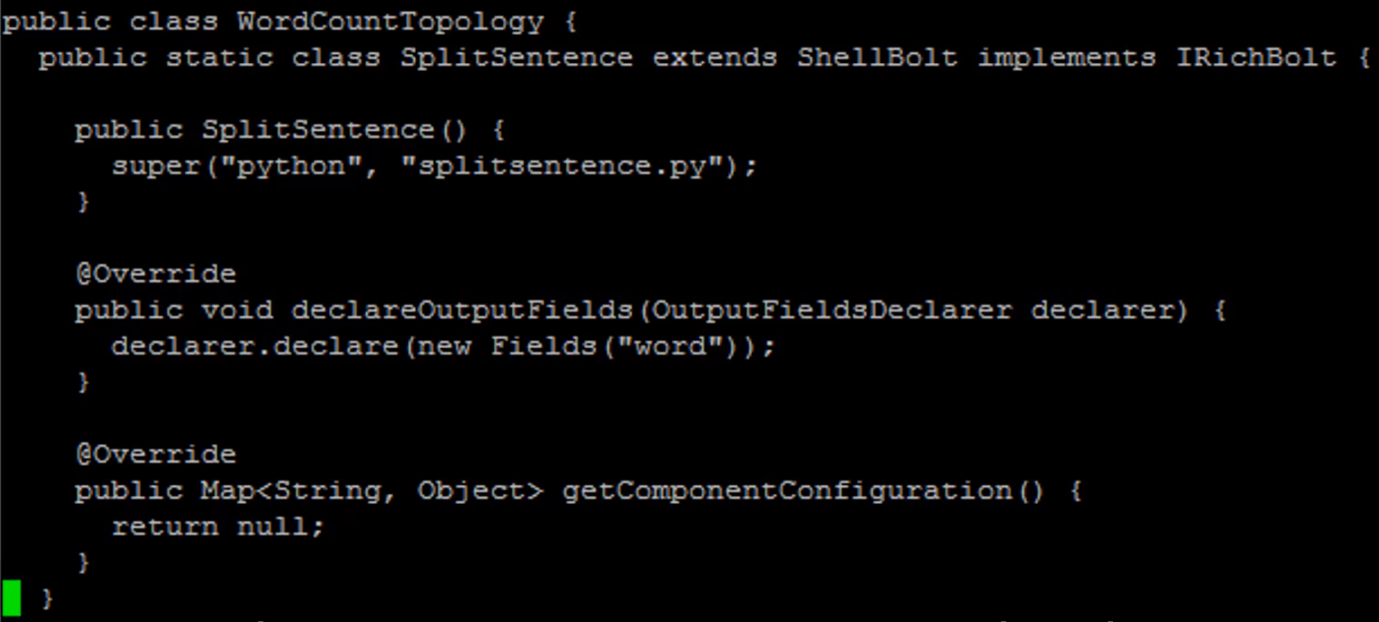
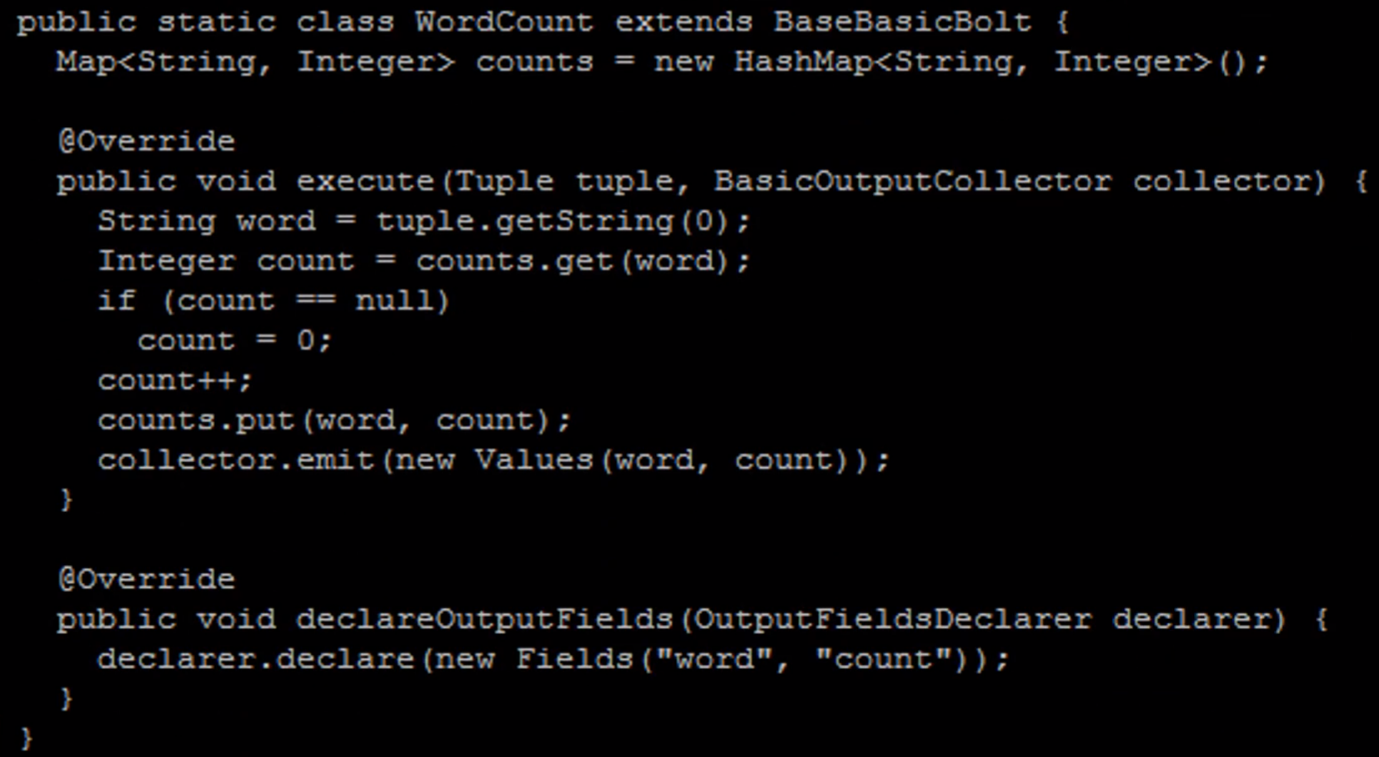
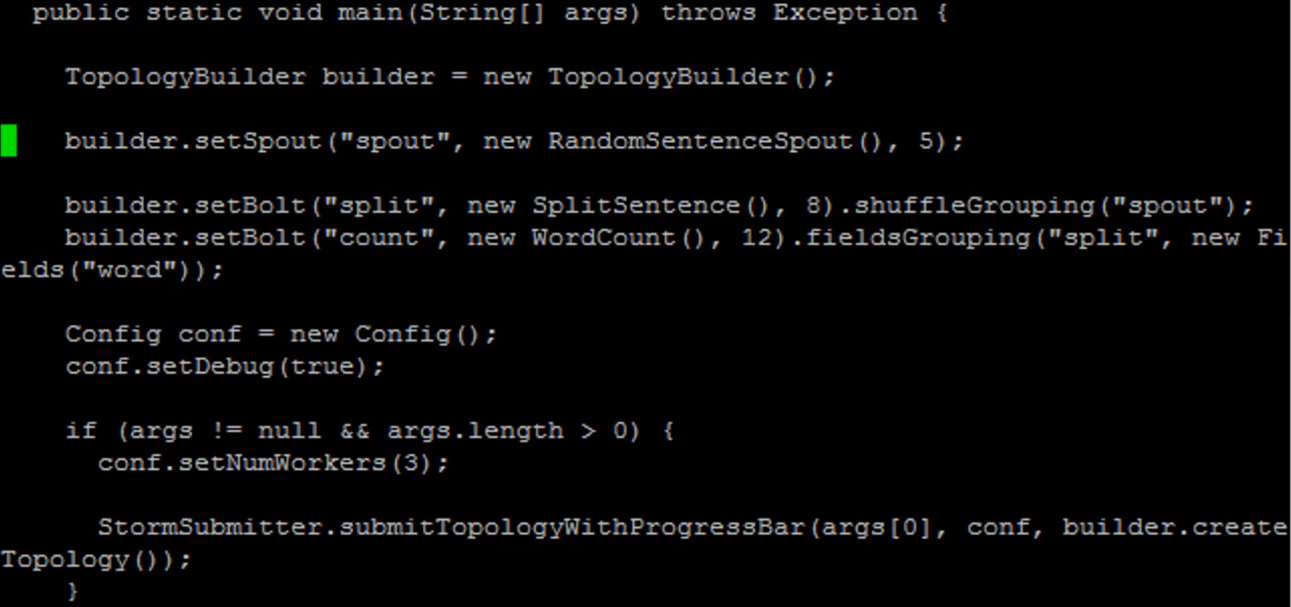
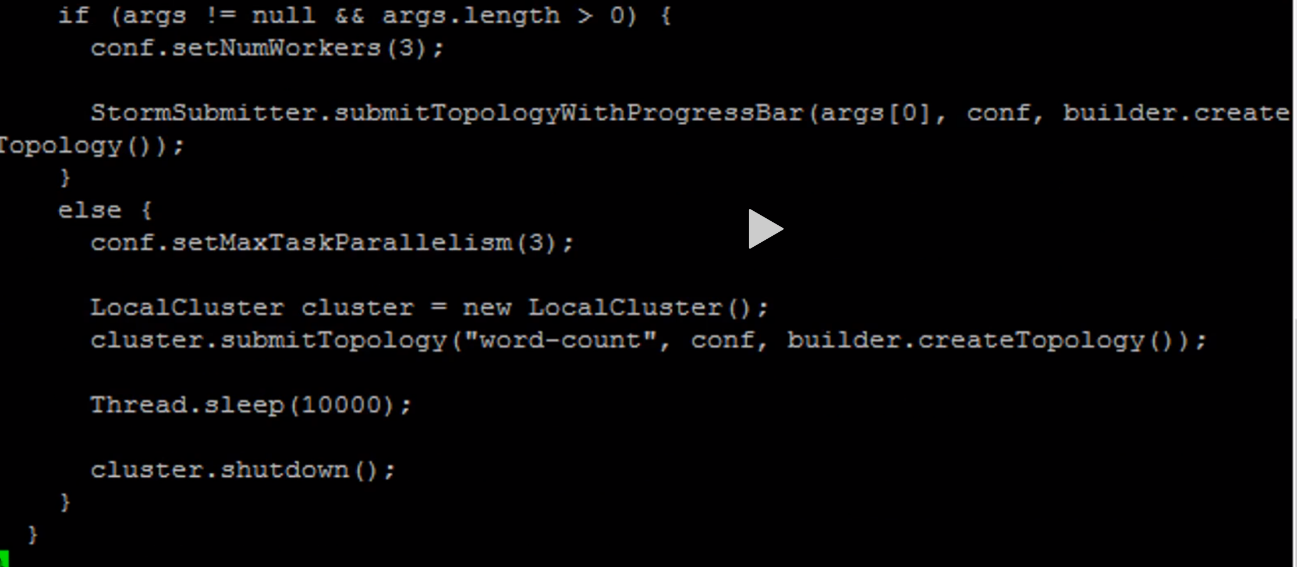
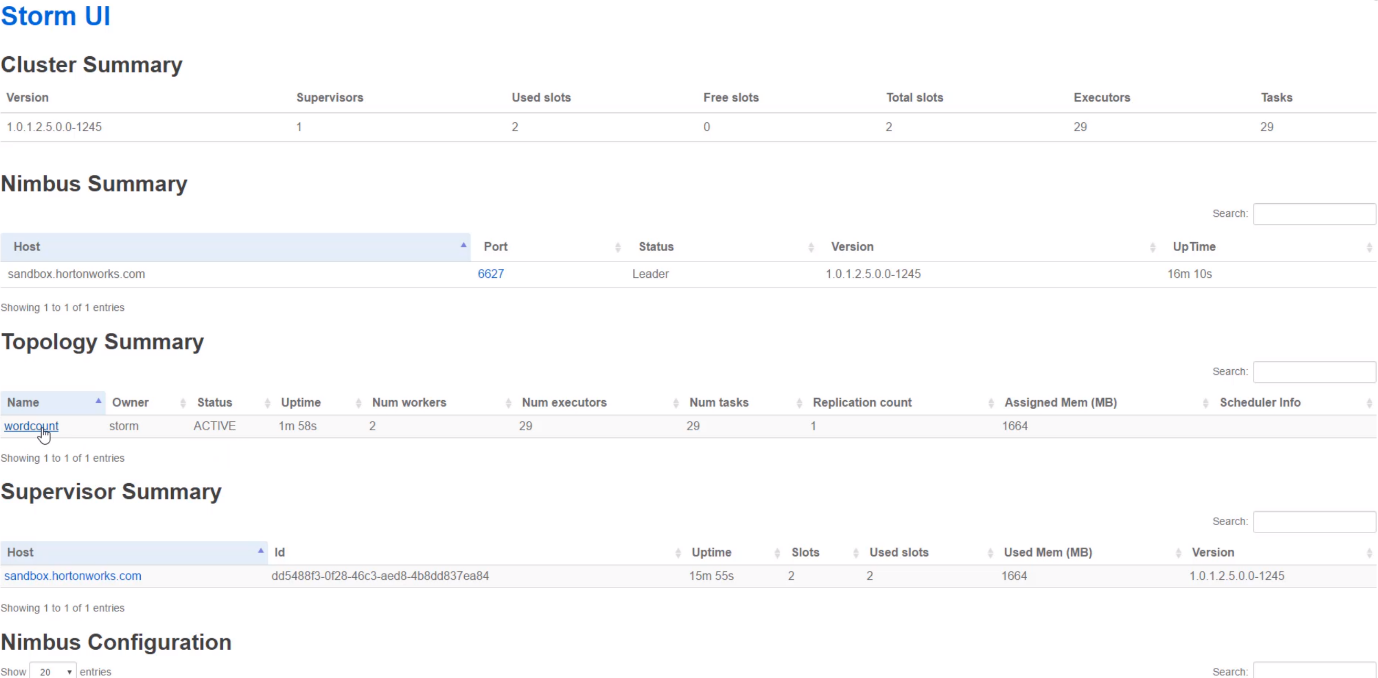
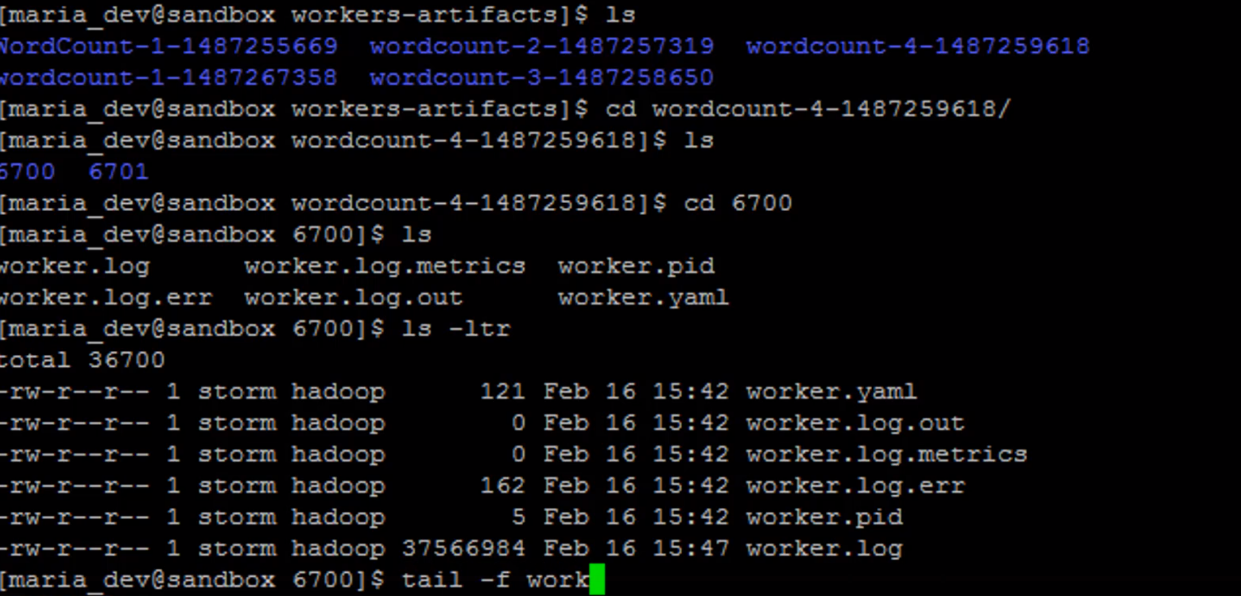
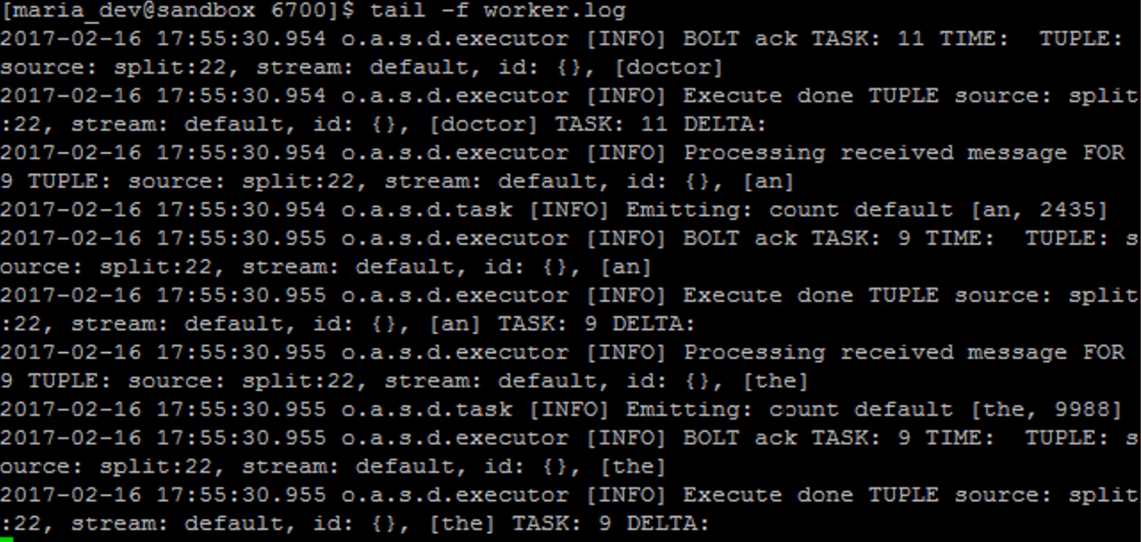
Zookeeper -

Supervisor - where work actually happens

Developing Storm applications

* Usually done with java
  + Although bolts may be directed through scripts in other languages
* Storm Core
  + The lower-level API for Storm
  + ‘at-least-once’ schematics
* Trident
  + High Level API for Storm
  + ‘Exactly once’ schematics
* Storm runs your applications ‘forever’ once submitted -until you explicitly stop them

Practice

1. Get your Hortonworks Docker Sandbox we will be using the sample that comes with storm
2. Going into Ambari and login as Admin
   1. Storm is preinstalled and select the service and start it.
   2. Need to start Kafka service because example have it.
3. Putty SSH connection & login
   1. cd /usr/hdp/current/storm-client
   2. cd contrib/storm-starter/src/jvm/org/apache/storm/starter
   3. ls
   4. Storm program that listen to spout of incoming random sentences and count each individual words that appear. SplitSentence bolt. Output of the bolt into a new field ‘Word’ ->
   5. New bolt called WordCount takes the tuple of words, for each word and update the map, that map words to count and update the hash map and emit the signal.
   6. Main function that put it all together
4. Dependencies declaration and running
   1. storm jar /usr/hdp/current/storm-client/contrib/storm-starter/storm-starter-topologies-\*.jar org.apache.storm.starter.WordCountTopology wordcount (submitting a job into storm)
5. Look into the storm UI using the IP address 127.0.0.1:8744/index.html
6. cd /usr/hdp/current/storm-client/logs/workers-artifacts/
   1. ls
   2. tail -f worker.log