CS 686: Special Topics in Big Data

Storm

Lecture 20

MapReduce Hints

- I'll keep adding useful commands and tips to the assignment page
- Check out the setup() and cleanup() methods that you can override in both Map and Reduce
  - Cleanup() – great for aggregating results
- You can also create your own input/output formats – we'll talk a bit about this next class
- It's okay if your MR job gets you 99% of the way to the answer and you need a little post-processing to get the final result
- We'll have another Hadoop lab on Monday to go over a few more concepts

Tweets

- “Streams are passed. Storm is fast.”
- “Tweet storm is Hadoop realtime”
- “This bird is headed straight towards (Thunder)bolts and (water)spouts!”
- “It’s basically real-time MapReduce for data streams”
- “Fail fast and restart fast: Storm can guarantee every data stream be processed at least once in a real time manner”
- “First ever “Storm” which never destroys anything but helps in building things”

Discussion

- Discuss your opinions on:
  - Most innovative/interesting part of the approach
  - Potential downsides
  - What parts were unclear?
  - What would you change about the paper **or** system?
- Sketch out your answers (**1** page) and then we'll merge for the class discussion
- **We'll present really quickly around @ 12:25ish**