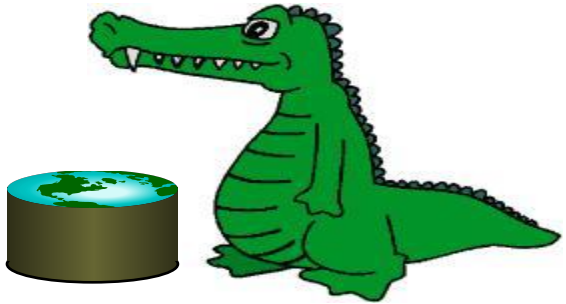
Lab 3: MapReduce and AWS (JAVA)

Xiaofeng Zhou



Goal

- Learn how to program using MapReduce with Java.
- Learn how to use AWS to run your MR job.
- Learn how to generate an adjacency graph from a large wikilink dataset using MapReduce on AWS.

Input & Output

You program should take two arguments, an input and an output, your output layout should be like this:

output/ --- output folder, the second argument to your Jar file.

graph/ -- containing the part-r-xxxxx files of the actual output, will be examined by TA.

temp/ -- containing intermediate results, will be ignored by TA.

And use HDFS API instead of Java File API for file operations.

Submission format

Firstname_Lastname/ -- *Your firstname and lastname as shown on Canvas*
src/ -- *source code only, no deps or project folder.*
extract.jar -- *your jar file should be named exactly as shown here.*
report.pdf -- *your report.*
Compress your folder as a zip file and submit it.

Failure to comply with the input & output requirement or submission format will result in zero point for lab 5. A regrade will be needed at minimum 30% penalty.

Frequently Asked Questions

- Hadoop on Windows
- AWS educational account
- Downloading XML file
- How to use HDFS api
- Memory issue on AWS

Q/A session

Quiz 5

If you have wifi connection issues, go to lab rooms to finish the quiz. The quiz will be available for 25 mins.