CSE 4/587
Data Intensive Computing

Dr. Eric Mikida
epmikida@buffalo.edu
208 Capen Hall

Dr. Shamshad Parvin
shamsadp@buffalo.edu
313 Davis Hall

Midterm Review #2

Announcements

- Midterm is Wednesday @ 11:00-11:50 AM
- In Cooke 121 AND Alumni 97
  - If your UB email begins with a-j, you will be in Alumni 97
  - If your UB email begins with k-z you will be in Cooke 121

Midterm Format and Logistics

- [a-j] report to Alumni 97, [k-z] report to Cooke 121
  - Seating will be randomized within each exam room
- Handwritten, Closed Book
  - You must have a pen/pencil, and your UB ID card
  - You may also bring a non-graphing calculator
  - All other bags/electronics/etc will be placed at the front of the room during the exam
- 3-4 short answer questions with a mix of theory and application

Midterm Review

Potential Topics:

1. HDFS Architecture and Protocol
2. MapReduce Fundamentals/Word Count
3. Word Co-Occurrence
4. NGS in MapReduce

Hadoop and HDFS Architecture [Lec 14,16]

- Understand and discuss the evolution of Hadoop from 1.0 to 3.0.
- Understand the basics of the HDFS architecture, the different components involved, and their roles and responsibilities.
  - What are name nodes vs data nodes?
  - What tasks are handled by name nodes and data nodes?
  - What are the JobTracker and TaskTracker?
- Understand and discuss block replication and its importance
  - How is fault tolerance achieved in HDFS?
  - How are blocks replicated?
  - What kind of failures might occur?

MapReduce Basics [Lec 17-22]

- Understand and discuss the roles of the different types of MapReduce tasks that are part of a MapReduce Job.
- Understand the type of data that MapReduce deals with
  - Understand and discuss how this data flows throughout a MapReduce job and how it is split up over mappers and/or reducers.
- Be able to read/write MapReduce pseudocode.
- Understand what intermediate data is and the role it plays in MapReduce performance
- Understand common bottlenecks in MapReduce and techniques for dealing with them

Word Co-Occurrence [Lec 21]

- Understand and discuss the relevance of word co-occurrence.
- Understand and discuss the basic matrix formulation of the problem.
- Understand and discuss the two different MapReduce formulations (pairs and stripes), and the pros and cons of each formulation.
- Understand and discuss the difference between absolute and relative co-occurrence.
- Understand and discuss the modifications needed to compute relative co-occurrence with both the pairs and stripes method.

NGS in MapReduce [Lec 22]

- Understand the basics of k-mer counting in MapReduce
  - What unique problems does it face w.r.t. intermediate data?
- Understand the basics of the optimization discussed in the paper
  - Which types of optimizations provided the largest benefits and why?
  - What is a spill? How does the tradeoff between larger and smaller blocks affect performance?
  - How was I/O cost reduced in the paper?
  - How was network cost reduced in the paper?