CSE 4/587
Data Intensive Computing

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

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

Final Revew-2

Final Exam Logistics

- The final is Monday 5/15/23 in **Cooke 121** from 11:45PM to 2:45PM
  - **EVERYONE IS IN COOKE 121**
  - If you have any official conflicts (2 exams at same time, or 3 same day) please let me know ASAP
- Seating is randomized
- All bags/electronics must be placed at the front of the room
- **What to Bring:**
  - Pen/Pencil
  - UB ID Card
  - A non-graphing calculator (if you like, not required)

Final Exam Review

Potential Topics:

1. Data Science Overview
2. Models/Algorithms (Linear Regression, Classifiers, K-Means)
3. HDFS Architecture and Protocol
4. MapReduce Fundamentals
5. MapReduce Applications
6. **Page Rank**
7. **Spark/Spark Streaming/Hive**
8. **Cloud**
9. **Ethics in DIC**

} Most of the depth will be focused on these topics

Graph Analysis and page rank [Lec 24-26]

- Understand the concepts of various types of graph representation
- Understand how we can model the internet using graph
- Understand why it is important to model the internet using graph
- Understand the concepts of page rank and its importance

Graph Analysis and page rank [Lec 24-26]

- The concepts of the link analysis algorithm
- Definition of page rank with example
- Be Able to find out page rank using the recursive formulation
- Understand the concepts of the flow model.
- Be Able to find out the page rank vector using the flow model
- Understand the methodology of finding rank vector using power iteration

Graph Analysis and page rank [Lec 24-26]

- Find out the Eigenvector from the Matrix Formulations for the Page rank
- Understand Google formulation
- Understand the problem of Page rank (dead ends and spider traps)
- Be able to solve problems using a teleport
- Why and how is teleport able to solve the page rank problem
- How we can handle graph problems in MapReduce
- What problem we should consider when processing graphs in MapReduce
- How can we find page rang in MapReduce

Spark Streaming [Lec 30-31]

- Understand the motivation for Spark streaming
- Why do we need a separate framework for Streaming data?
- Why MR is not a solution for big streaming data?
- Definition of spark streaming and various data sources of spark streaming
- Understand how spark streaming works and the programming model.
- Understand the concepts of Dstream.
- Difference between Stateless and stateful operations
- Understand the fault-tolerant features of Spark Streaming

HIVE [Lec 32]

- Understand the basic concept of HIVE and its principles.
- What are the advantages of HIVE and where should we use HIVE
- Limitations of HIVE
- Understand the concepts of Hbase