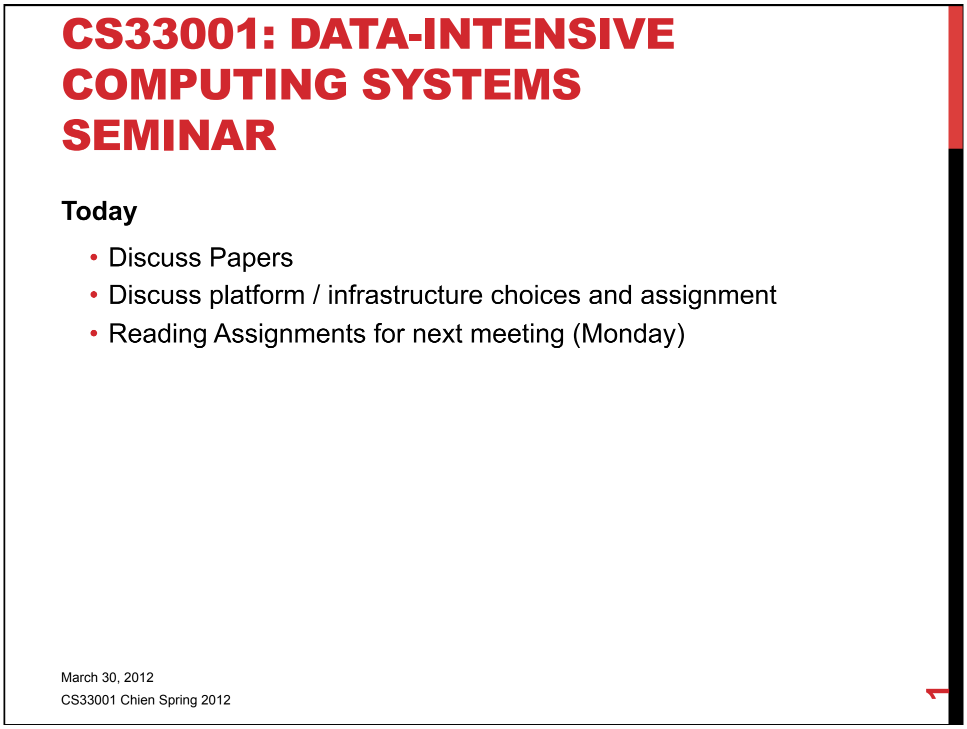


CS33001: DATA-INTENSIVE COMPUTING SYSTEMS SEMINAR

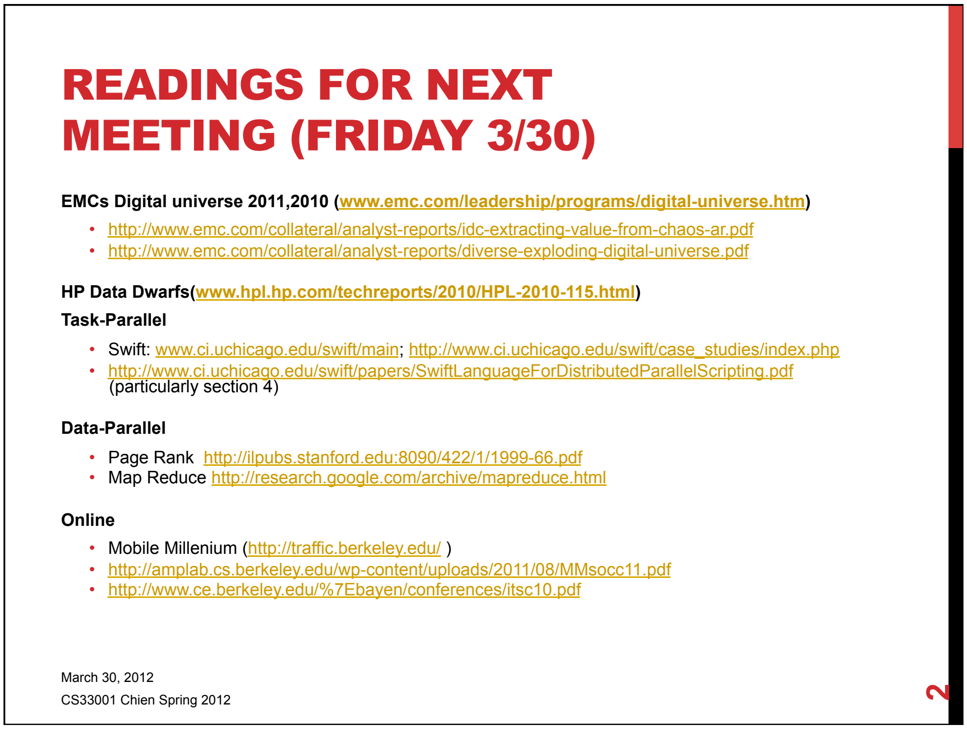
Today

- Discuss Papers
- Discuss platform / infrastructure choices and assignment
- Reading Assignments for next meeting (Monday)

March 30, 2012

CS33001 Chien Spring 2012

1



READINGS FOR NEXT MEETING (FRIDAY 3/30)

EMCs Digital universe 2011,2010 (www.emc.com/leadership/programs/digital-universe.htm)

- http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf
- http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

HP Data Dwarfs(www.hpl.hp.com/techreports/2010/HPL-2010-115.html)

Task-Parallel

- Swift: www.ci.uchicago.edu/swift/main; http://www.ci.uchicago.edu/swift/case_studies/index.php
- http://www.ci.uchicago.edu/swift/papers/SwiftLanguageForDistributedParallelScripting.pdf (particularly section 4)

Data-Parallel

- Page Rank http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf
- Map Reduce http://research.google.com/archive/mapreduce.html

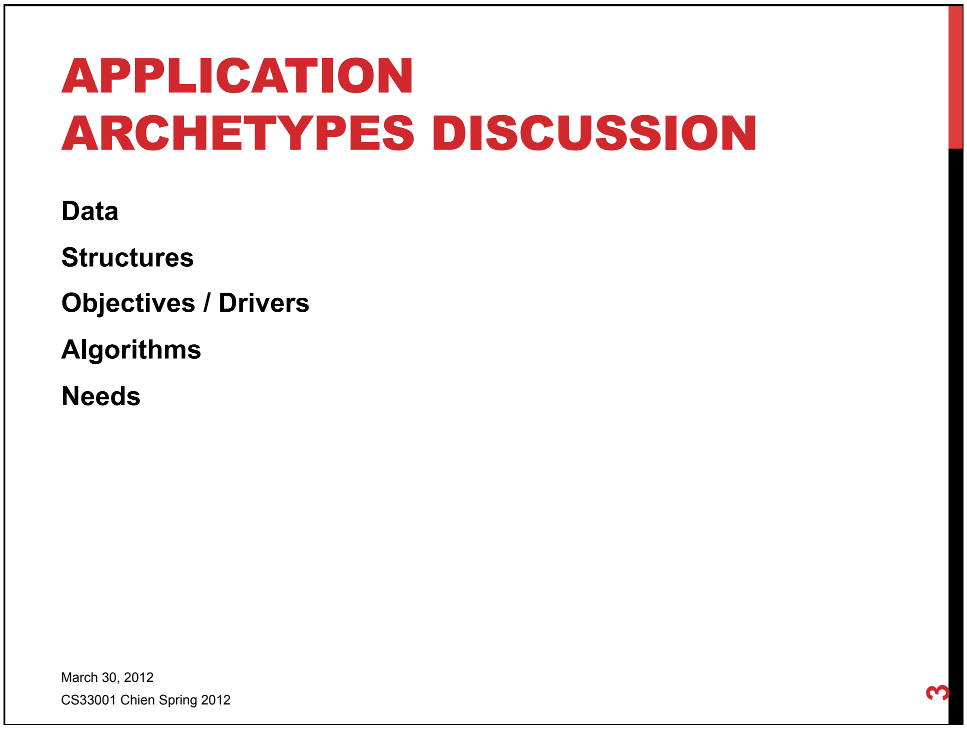
Online

- Mobile Millenium (http://traffic.berkeley.edu/)
- http://amplab.cs.berkeley.edu/wp-content/uploads/2011/08/MMsocc11.pdf
- http://www.ce.berkeley.edu/%7Ebayen/conferences/itsc10.pdf

March 30, 2012

CS33001 Chien Spring 2012

2



APPLICATION ARCHETYPES DISCUSSION

Data

Structures

Objectives / Drivers

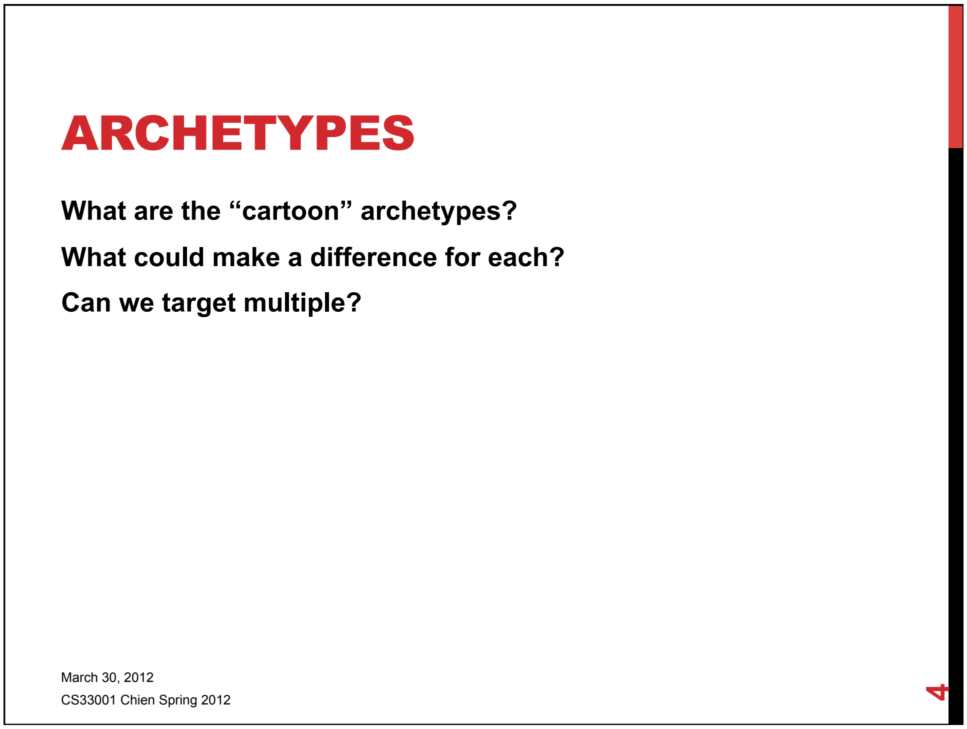
Algorithms

Needs

March 30, 2012

CS33001 Chien Spring 2012

3



ARCHETYPES

What are the “cartoon” archetypes?

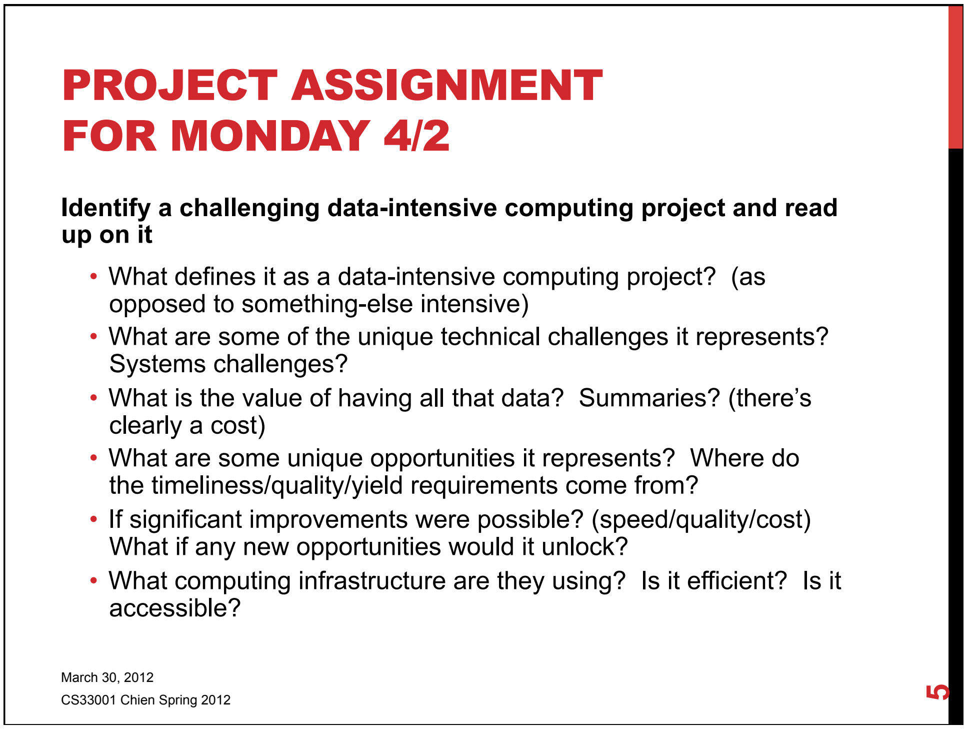
What could make a difference for each?

Can we target multiple?

March 30, 2012

CS33001 Chien Spring 2012

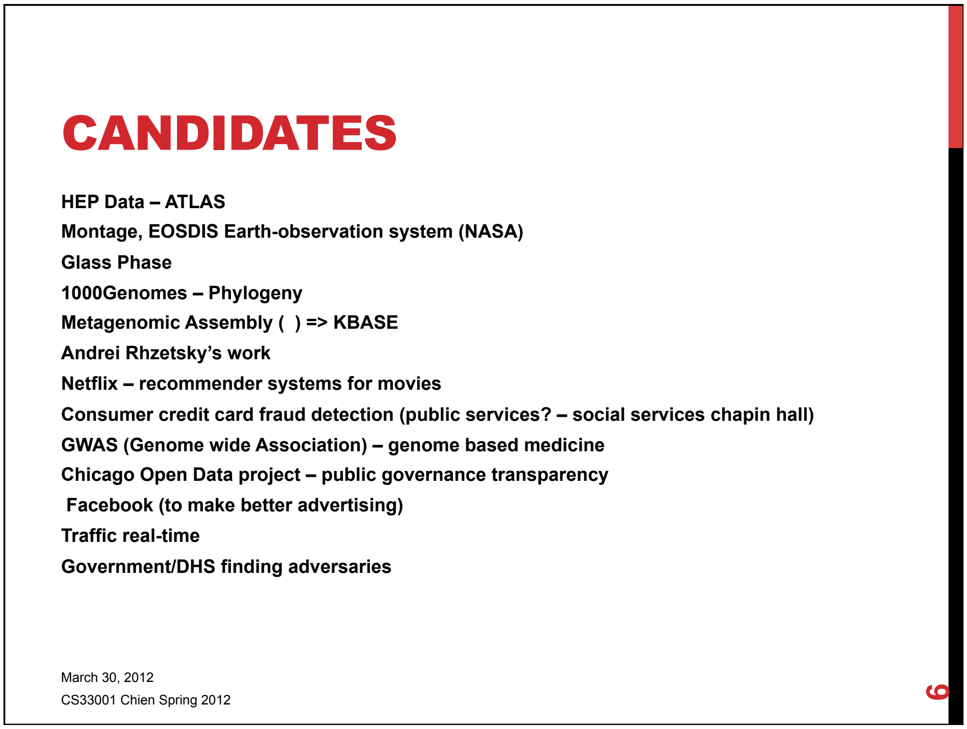
4



PROJECT ASSIGNMENT FOR MONDAY 4/2

Identify a challenging data-intensive computing project and read up on it

- What defines it as a data-intensive computing project? (as opposed to something-else intensive)
- What are some of the unique technical challenges it represents? Systems challenges?
- What is the value of having all that data? Summaries? (there's clearly a cost)
- What are some unique opportunities it represents? Where do the timeliness/quality/yield requirements come from?
- If significant improvements were possible? (speed/quality/cost) What if any new opportunities would it unlock?
- What computing infrastructure are they using? Is it efficient? Is it accessible?



CANDIDATES

HEP Data – ATLAS

Montage, EOSDIS Earth-observation system (NASA)

Glass Phase

1000Genomes – Phylogeny

Metagenomic Assembly ( ) => KBASE

Andrei Rhzetsky's work

Netflix – recommender systems for movies

Consumer credit card fraud detection (public services? – social services chapin hall)

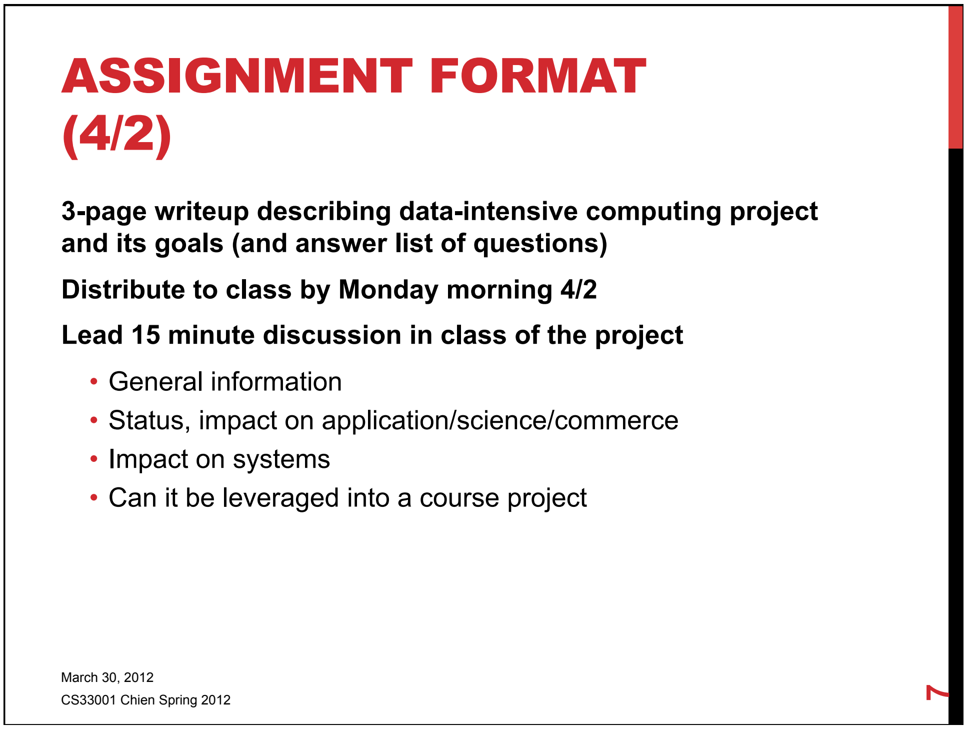
GWAS (Genome wide Association) – genome based medicine

Chicago Open Data project – public governance transparency

Facebook (to make better advertising)

Traffic real-time

Government/DHS finding adversaries



ASSIGNMENT FORMAT (4/2)

3-page writeup describing data-intensive computing project and its goals (and answer list of questions)

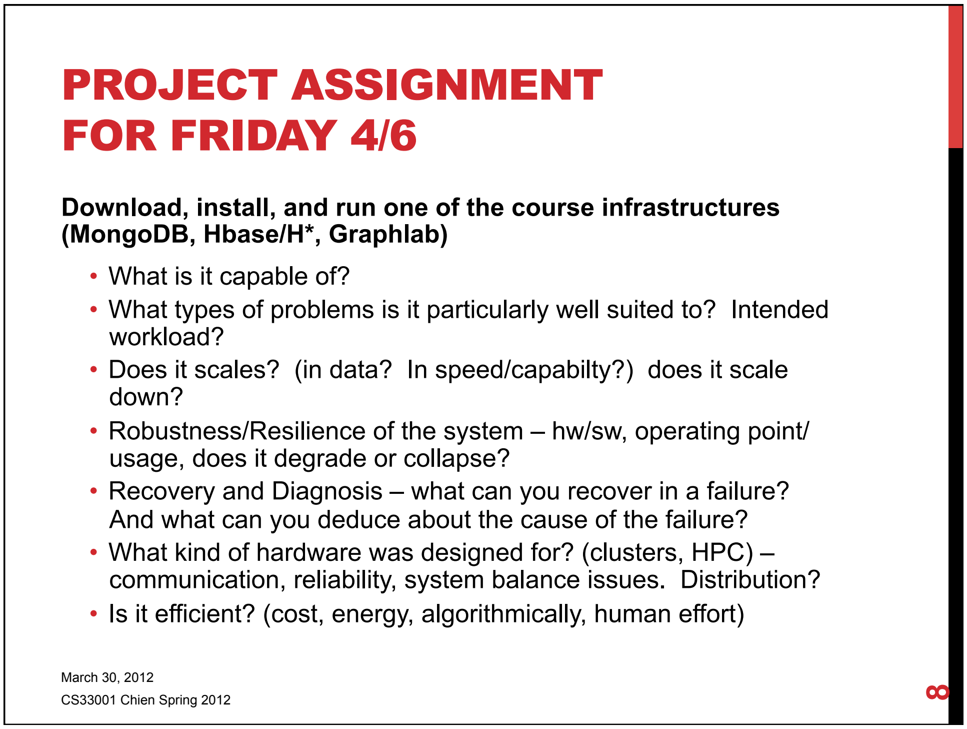
Distribute to class by Monday morning 4/2

Lead 15 minute discussion in class of the project

- General information
- Status, impact on application/science/commerce
- Impact on systems
- Can it be leveraged into a course project

March 30, 2012
CS33001 Chien Spring 2012

7



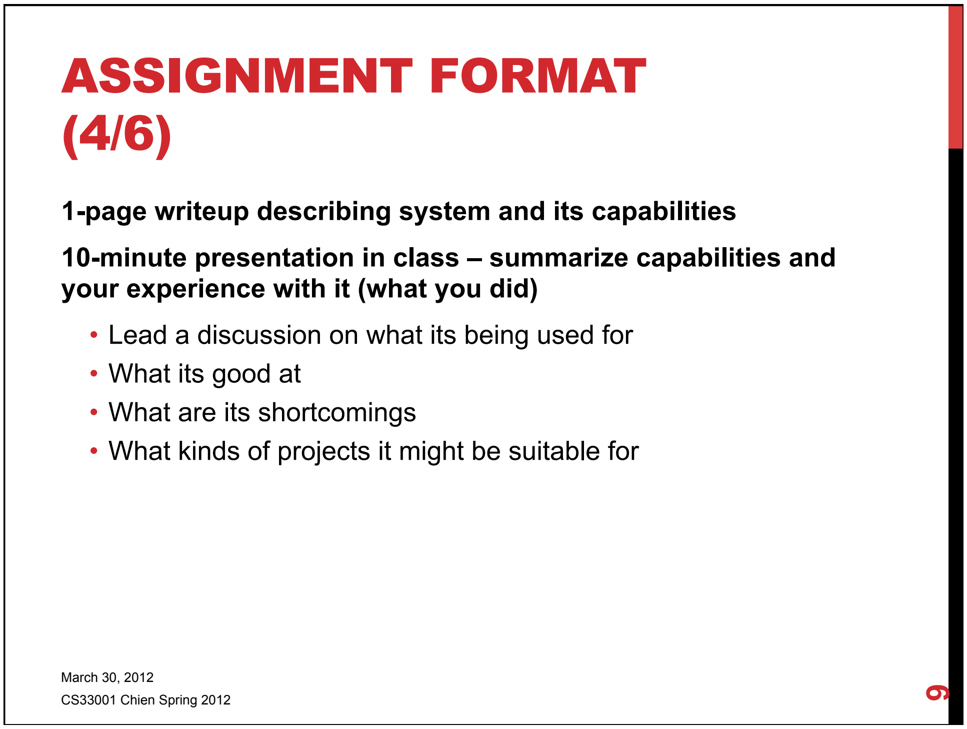
PROJECT ASSIGNMENT FOR FRIDAY 4/6

Download, install, and run one of the course infrastructures (MongoDB, Hbase/H*, Graphlab)

- What is it capable of?
- What types of problems is it particularly well suited to? Intended workload?
- Does it scales? (in data? In speed/capabilty?) does it scale down?
- Robustness/Resilience of the system – hw/sw, operating point/ usage, does it degrade or collapse?
- Recovery and Diagnosis – what can you recover in a failure? And what can you deduce about the cause of the failure?
- What kind of hardware was designed for? (clusters, HPC) – communication, reliability, system balance issues. Distribution?
- Is it efficient? (cost, energy, algorithmically, human effort)

March 30, 2012
CS33001 Chien Spring 2012

8



ASSIGNMENT FORMAT (4/6)

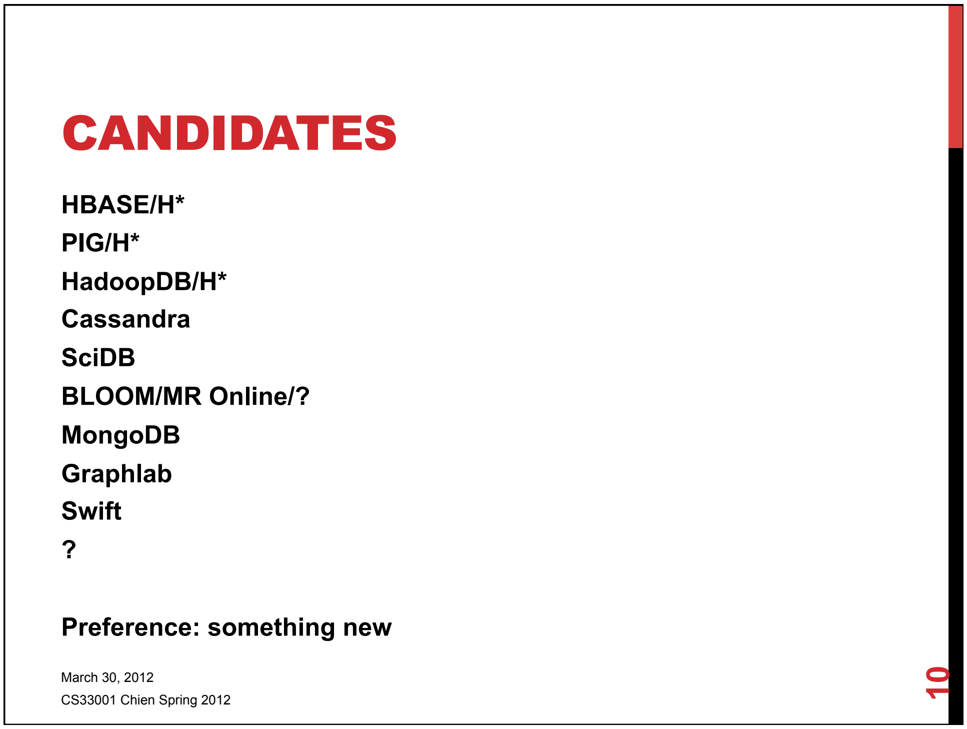
1-page writeup describing system and its capabilities

10-minute presentation in class – summarize capabilities and your experience with it (what you did)

- Lead a discussion on what its being used for
- What its good at
- What are its shortcomings
- What kinds of projects it might be suitable for

March 30, 2012
CS33001 Chien Spring 2012

9



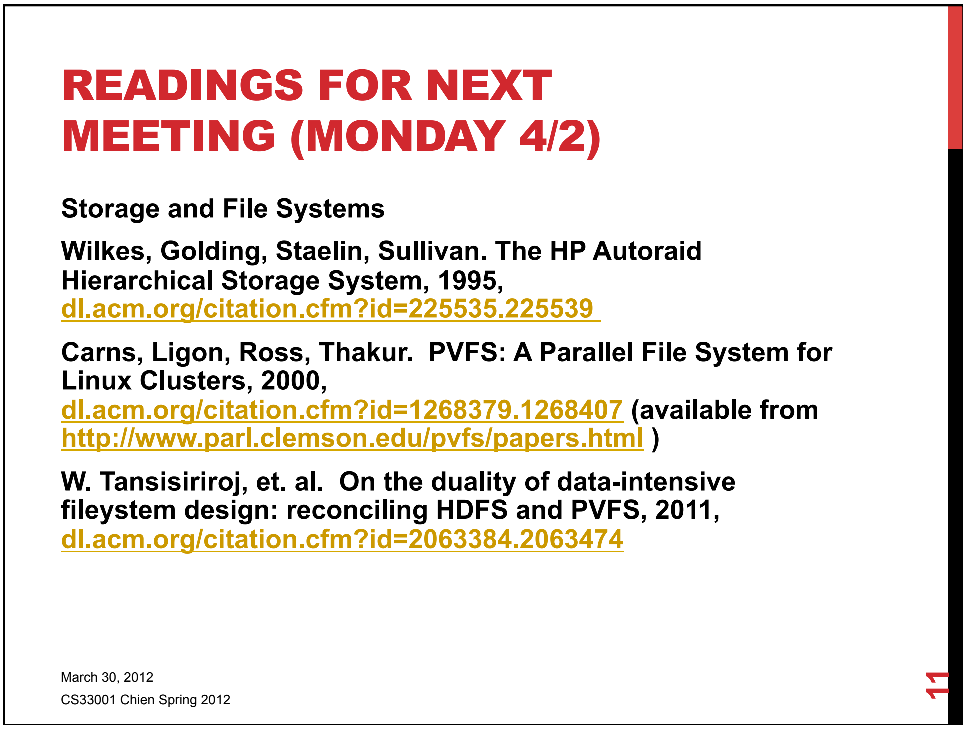
CANDIDATES

HBASE/H*
PIG/H*
HadoopDB/H*
Cassandra
SciDB
BLOOM/MR Online/?
MongoDB
Graphlab
Swift
?

Preference: something new

March 30, 2012
CS33001 Chien Spring 2012

10



READINGS FOR NEXT MEETING (MONDAY 4/2)

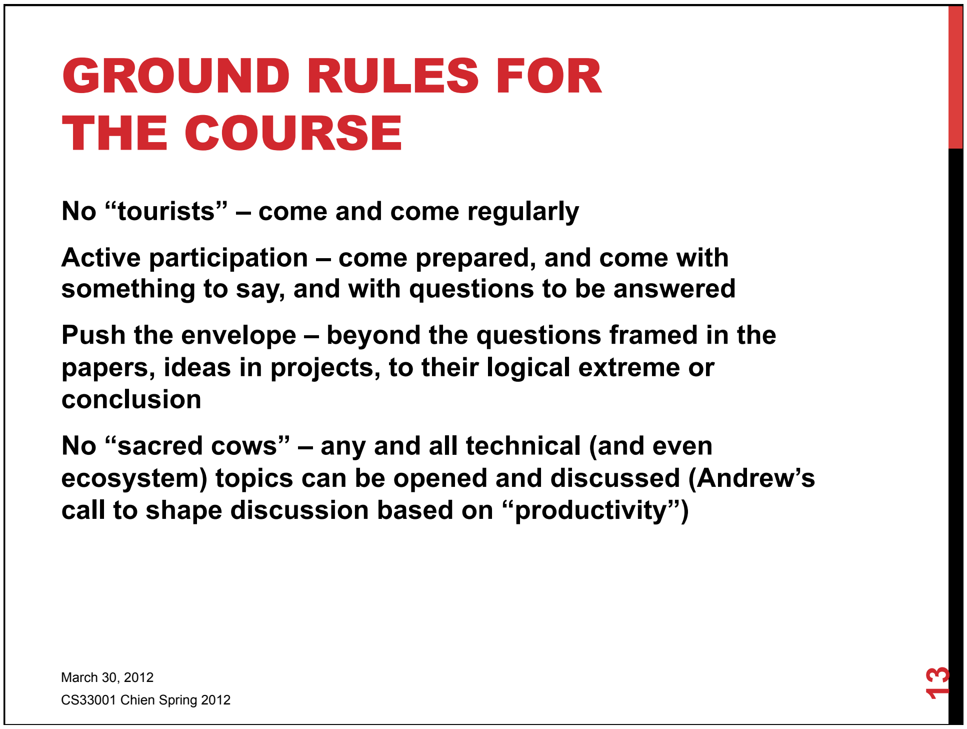
Storage and File Systems

Wilkes, Golding, Staelin, Sullivan. The HP Autoraid Hierarchical Storage System, 1995, dl.acm.org/citation.cfm?id=225535.225539

Carns, Ligon, Ross, Thakur. PVFS: A Parallel File System for Linux Clusters, 2000, dl.acm.org/citation.cfm?id=1268379.1268407 (available from http://www.parl.clemson.edu/pvfs/papers.html)

W. Tansisiriroj, et. al. On the duality of data-intensive fileystem design: reconciling HDFS and PVFS, 2011, dl.acm.org/citation.cfm?id=2063384.2063474

BACKUP



GROUND RULES FOR THE COURSE

No “tourists” – come and come regularly

Active participation – come prepared, and come with something to say, and with questions to be answered

Push the envelope – beyond the questions framed in the papers, ideas in projects, to their logical extreme or conclusion

No “sacred cows” – any and all technical (and even ecosystem) topics can be opened and discussed (Andrew’s call to shape discussion based on “productivity”)

March 30, 2012
CS33001 Chien Spring 2012

13