Project Ideas

Here are some potential project ideas for students taking the class for 4 units

See also Project and Dataset slides pptpdf

Project Ideas from Yahoo
Visualization/sonification of big data systems -

- Data flow
- Bottleneck identification
- Error frequencies

Scheduling work-
What are the attributes of a task which most impact scheduling decisions (data locality, balanced node load?, balanced network?, balanced disk load?)

- In-depth analysis of the interactions between Map/Reduce and Linux. Is the page cache effective? Is the linux scheduler doing us favors or is it actually getting in the way?

Storage- Is there a place for SSD in hadoop? Intermediate data? Distributed Cache? Hot blocks?


Streaming

- Scheduling/flow-control/rebalancing/fault-tolerance/realtime characteristics
- low latency adhoc queries
- Spark/Shark, Hive, Impala, Storm Trident
- Survey of the space

Workload simulation

- Explore/add more realism to gridmix/other simulation schemes.

Tools/Performance

- Caching within M/R Jobs is essentially left to linux. Maybe there are some better alterntatives (e.g. https://issues.apache.org/jira/browse/HADOOP-8705)

- Profiling jobs that spans thousands of nodes and making sense of the data.

- Debugging large map reduce jobs. What does it mean to set a breakpoing in a single map task?

- Job log analysis - With applicaiton logs being stored in HDFS, it becomes possible to write jobs which process these logs looking for error patterns.

Open Source

- Root cause analysis of Hadoop defects. Defect hot-spots. Visualization of such.

- Jira metrics - trend analysis, etc. This has been done on other large projects (e.g. linux kernel) but as far as I know similar techniques haven't been applied to hadoop.