Projects and Datasets

CS498CC
A. Yahoo Projects

- Visualization
- Scheduling
- Storage
- Streaming
- Workload Simulation
- Tools Performance
- Open Source – Software Engineering
B. Home Brew Projects

- Profiling using Performance Instrumentation
- Datastore Transfer
- Packet Data Analysis
- DNS data/query analysis
- Analyze a Data set --- Amazon, Twitter, Ebay, Spam, Flickr, Census, Wikipedia, Genomes...
  - Use Hadoop, Mahout, Hive, Hbase, Cassandra
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?
- Integrating the schedulers. Pluggable Capacity scheduler has hierarchical queues.
- VMWARE or similar network topology.
- In Hadoop. Plug in decides what machine you are in.
Storage

- Is there a place for SSD in hadoop? Intermediate data? Distributed Cache? Hot blocks?
- HBASE. Brian Abhishek spill to disk. Where to put the money Ram or SSD,
- Shark and Spark can perform 100x faster using RAM. When doesn’t fit in memory. Banks scared of SCALA.
Streaming

• - Scheduling/flow-control/rebalancing/fault-tolerance/realtime characteristics
• Storm on top of YARN. The scheduler assumptions are all for map/reduce.
• 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.
• Properly simulate the job. Does 1.0 Hadoop predict Hadoop 2.0
Tools, Performance

- Memcache D not a standard for caching. Sorting is a problem.
- Caching within M/R Jobs is essentially left to linux. Maybe there are some better alternatives (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 application logs being stored in HDFS, it becomes possible to write jobs which process these logs looking for error patterns.
•
Open Source-Software Engineering

• - 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.
• Debugging software engineers Use big data on hadoop.
B. Home Brew Projects

- Profiling using Performance Instrumentation
- Datastore Transfer
- Packet Data Analysis
- DNS data/query analysis
- Analyze a Data set --- Amazon, Twitter, Ebay, Spam, Flickr, Census, Wikipedia, Genomes…
  - Use Hadoop, Mahout, Hive, Hbase, Cassandra
Profiling

- Profiling user code.
- Use Machine learning applied to code performance counters
- Deduce pathological cases
- Create decision trees to predict improvements to user code.
Sentiment Analysis (AG)

- Using sentiment analysis to detect possible malware applications on Google's Play Store:
- Among other information we are currently collecting, twice every day, all the reviews of all the applications on Google's Play Store. We are looking to implement different sentiment analysis algorithms on top of Hadoop with the objective of classifying whether an application could be malware.
Datastore transfer (AG)

• Secure transactions between Hadoop and different types of structured datastores:
• Design a tool that reliably and securely transfers data between HDFS/Hive/HBase/Relational Database
Packet Data Analysis

• Examine packet streams entering University
• Searching for patterns and trends
• Machine Learning.
Examine DNS traffic for anomalies

- Machine learning
- Behavioral statistics
Amazon Datasets

- Annotated Human Genome Data provided by ENSEMBL
- The Ensembl project produces genome databases for human as well as almost 50 other species, and makes this information freely available.
- Various US Census Databases from The US Census Bureau
- UniGene provided by the National Center for Biotechnology Information
- A set of transcript sequences of well-characterized genes and hundreds of thousands of expressed sequence tags (EST) that provide an organized view of the transcriptome.
- Freebase Data Dump from Freebase.com
- A data dump of all the current facts and assertions in the Freebase system. Freebase is an open database of the world’s information, covering millions of topics in hundreds of categories. Drawing from large open data sets like Wikipedia, MusicBrainz, and the SEC archives, it contains structured information on many popular topics, including movies, music, people and locations – all reconciled and freely available.
Datasets

- http://aws.amazon.com/datasets
- http://databib.org
- http://datacite.org
- http://figshare.com
- http://linkeddata.org
- http://reddit.com/r/datasets
More Public Datasets

- http://crawdad.org/
- http://data.cityofchicago.org
- http://data.govloop.com
- http://data.gov.uk/
- http://data.medicare.gov
- http://data.gov.uk/
- http://data.seaSle.gov
- http://data.sfgov.org
- http://data.sunlightlabs.com
- https://datamarket.azure.com/
- http://developer.yahoo.com/geo/g...
- http://en.wikipedia.org/wiki/Wik...
- http://factfinder.census.gov/ser...
- http://gettingpastgo.socrata.com
- http://googleresearch.blogspot.c...
- http://books.google.com/ngrams/
- http://public.resource.org/
- http://snap.stanford.edu/data/in...
- https://wist.echo.nasa.gov/~wist...
- http://www2.jpl.nasa.gov/srtm
- http://www.archives.gov/research...
Collections

• bigdata data sets for download (twitter)
• [www.infochimps.com/tags/bigdata](www.infochimps.com/tags/bigdata)
• Research quality datasets (spam, Flickr, Yahoo, computer science bibliography) [https://bitly.com/bundles/hmason/1](https://bitly.com/bundles/hmason/1)
http://www.kdnuggets.com/datasets/

- KDD Cup center, with all data, tasks, and results.
- UCI KDD Database Repository for large datasets used in machine learning and knowledge discovery research.
- UCI Machine Learning Repository.
- Bioassay data, described in Virtual screening of bioassay data, by Amanda Schierz, J. of Cheminformatics, with 21 Bioassay datasets (Active / Inactive compounds) available for download.
- Bitly 1.usa.gov data, anonymized clicks on gov links.
- Canada Open Data, pilot project with many government and geospatial datasets.
- Causality Workbench data repository.
- Data.gov.uk, publicly available data from UK (also London datastore.)
- Data.gov/Education, central guide for education data resources including high-value data sets, data visualization tools, resources for the classroom, applications created from open data and more.
- DataMarket, visualize the world's economy, societies, nature, and industries, with 100 million time series from UN, World Bank, Eurostat and other important data providers.
- Datamob, public data put to good use.
- DataFerrett, a data mining tool that accesses and manipulates TheDataWeb, a collection of many on-line US Government datasets.
- Delve, Data for Evaluating Learning in Valid Experiments
- EconData, thousands of economic time series, produced by a number of US Government agencies.
- Enron Email Dataset, data from about 150 users, mostly senior management of Enron.
- Europeana Data, contains open metadata on 20 million texts, images, videos and sounds gathered by Europeana - the trusted and comprehensive resource for European cultural heritage content.
- FEDSTATS, a comprehensive source of US statistics and more
- FIMI repository for frequent itemset mining, implementations and datasets.
- Financial Data Finder at OSU, a large catalog of financial data sets.
- GEO (GEO Gene Expression Omnibus), a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval.
- GeoDa Center, geographical and spatial data.
- Google ngrams datasets, text from millions of books scanned by Google.
- Grain Market Research, financial data including stocks, futures, etc.
- Hilary Mason research-quality Big Data sets collection - many text and image datasets.
- ICWSM-2009 dataset contains 44 million blog posts made between August 1st and October 1st, 2008.
- Infochimps, an open catalog and marketplace for data. You can share, sell, curate, and download data about anything and everything.
- Investor Links, includes financial data.
- Kevin Chai list of datasets, for text, SNA, and other fields.
- KONECT, the Koblenz Network Collection, with large network datasets of all types in order to perform research in the area of network mining.
- Linking Open Data project, at making data freely available to everyone.
- Million Song Dataset
- MIT Cancer Genomics gene expression datasets and publications, from MIT Whitehead Center for Genome Research.
- ML Data, the data repository of the EU Pascal2 networks.
- NASDAQ Data Store, provides access to market data.
- National Government Statistical Web Sites, data, reports, statistical yearbooks, press releases, and more from about 70 web sites, including countries from Africa, Europe, Asia, and Latin America.
- National Space Science Data Center (NSIDC), NASA data sets from planetary exploration, space and solar physics, life sciences, astrophysics, and more.
- OpenData from Socrata, access to over 10,000 datasets including business, education, government, and fun.
- Peter Skomoroch dataset Bookmarks
- PubGene(TM) Gene Database and Tools, genomic-related publications database
- Quandl, a collaboratively curated portal to millions of financial and economic time-series datasets.
- qunb, a platform to find and visualize quantitative data.
- Robert Schiller data on housing, stock market, and more from his book Irrational Exuberance.
- SMD: Stanford Microarray Database, stores raw and normalized data from microarray experiments.
- Jerry Smith dataset collection, with Finance, Government, Machine Learning, Science, and other data.
- SourceForge.net Research Data, includes historic and status statistics on approximately 100,000 projects and over 1 million registered users' activities at the project management web site.
- StatLib, CMU Datasets Archive.
- STATOO Datasets part 1 and STATOO Datasets part 2
- Time Series Data Library
- Visual Analytics Benchmark Repository.
- UCR Time Series Data Archive, offering datasets, papers, links, and code.
- United States Census Bureau.
- Wikiposit, a (virtual) amalgamation of (mostly financial) data from many different sites, allowing users to merge data from different sources
- Wolfram Alpha disease and patient level data.
- Yahoo Sandbox datasets, Language, Graph, Ratings, Advertising and Marketing, Competition
- Yelp Academic Dataset, all the data and reviews of the 250 closest businesses for 30 universities for students and academics to explore and research.