PoweredBy

Applications and organizations using Hadoop include (alphabetically):

- **A9.com - Amazon**
  - We build Amazon's product search indices using the streaming API and pre-existing C++, Perl, and Python tools.
  - We process millions of sessions daily for analytics, using both the Java and streaming APIs.
  - Our clusters vary from 1 to 100 nodes.

- **Able Grape - Vertical search engine for trustworthy wine information**
  - We have one of the world's smaller hadoop clusters (2 nodes @ 8 CPUs/node)
  - Hadoop and Nutch used to analyze and index textual information

- **Adknowledge - Ad network**
  - Hadoop used to build the recommender system for behavioral targeting, plus other clickstream analytics
  - We handle 500MM clickstream events per day
  - Our clusters vary from 50 to 200 nodes, mostly on EC2.
  - Investigating use of R clusters atop Hadoop for statistical analysis and modeling at scale.

- **Alibaba**
  - A 15-node cluster dedicated to processing sorts of business data dumped out of database and joining them together. These data will then be fed into iSearch, our vertical search engine.
  - Each node has 8 cores, 16G RAM and 1.4T storage.

- **AOL**
  - We use hadoop for variety of things ranging from ETL style processing and statistics generation to running advanced algorithms for doing behavioral analysis and targeting.
  - Our cluster size is 50 machines, Intel Xeon, dual processors, dual core, each with 16GB Ram and 800 GB hard-disk giving us a total of 37 TB HDFS capacity.

- **backdocsearch.com - search engine for chiropractic information, local chiropractors, products and schools**

- **Baidu - the leading Chinese language search engine**
  - Hadoop used to analyze the log of search and do some mining work on
web page database
  o We handle about 200TB per week
  o Our clusters vary from 10 to 500 nodes
  o Hypertable is also supported by Baidu

- Cascading - Cascading is a feature rich API for defining and executing complex and fault tolerant data processing workflows on a Hadoop cluster.
- CbIR - Image retrieval engine
  o We Japanese company CbIR (Content-based Information Retrieval) use Hadoop to build the image processing environment for image-based product recommendation system mainly on Amazon EC2, from April 2008.
  o Our Hadoop environment produces the original database for fast access from our web application.
  o We also use Hadoop to analyzing similarities of user's behavior.

- Cloudera, Inc - Cloudera provides commercial support and professional training for Hadoop.
  o We provide our support customers a stable/tested/optimized release of Hadoop.
  o We are developing improved deployment, monitoring and integration tools.
  o We're writing "Hadoop: The Definitive Guide" (Tom White/O'Reilly)

- Contextweb - ADSDAQ Ad Exchange
  o We use Hadoop to store ad serving log and use it as a source for Ad optimizations/Analytics/reporting/machine learning.
  o Currently we have a 23 machine cluster with 184 cores and about 35TB raw storage. Each (commodity) node has 8 cores, 8GB RAM and 1.7 TB of storage.

- Cornell University Web Lab
  o Generating web graphs on 100 nodes (dual 2.4GHz Xeon Processor, 2 GB RAM, 72GB Hard Drive)

- Enormo
  o 4 nodes cluster (32 cores, 1TB).
  o We use Hadoop to filter and index our listings, removing exact duplicates and grouping similar ones.
  o We plan to use Pig very shortly to produce statistics.

- ETH Zurich Systems Group
We are using Hadoop in a course that we are currently teaching: "Massively Parallel Data Analysis with MapReduce". The course projects are based on real use-cases from biological data analysis.

Cluster hardware: 16 x (Quad-core Intel Xeon, 8GB RAM, 1.5 TB Hard-Disk)

- Eyealike - Visual Media Search Platform
  o Facial similarity and recognition across large datasets.
  o Image content based advertising and auto-tagging for social media.
  o Image based video copyright protection.

- Facebook
  o We use Hadoop to store copies of internal log and dimension data sources and use it as a source for reporting/analytics and machine learning.
  o Currently have a 600 machine cluster with 4800 cores and about 2 PB raw storage. Each (commodity) node has 8 cores and 4 TB of storage.
  o We are heavy users of both streaming as well as the Java apis. We have built a higher level data warehousing framework using these features called Hive (see the http://hadoop.apache.org/hive/). We have also developed a FUSE implementation over hdfs.

- FOX Interactive Media
  o 3 X 20 machine cluster (8 cores/machine, 2TB/machine storage)
  o 10 machine cluster (8 cores/machine, 1TB/machine storage)
  o Use for log analysis, data mining and machine learning

- Hadoop Korean User Group, a Korean Local Community Team Page.
  o 50 node cluster In the Korea university network environment.
    - Pentium 4 PC, HDFS 4TB Storage
  o Used for development projects
    - Retrieving and Analyzing Biomedical Knowledge
    - Latent Semantic Analysis, Collaborative Filtering

- Google
  o University Initiative to Address Internet-Scale Computing Challenges

- Gruter. Corp.
  o 30 machine cluster (4 cores, 1TB~2TB/machine storage)
  o storage for blog data and web documents
  o used for data indexing by MapReduce
- link analyzing and Machine Learning by MapReduce

- Hadoop Taiwan User Group

- IBM
  - Blue Cloud Computing Clusters
  - University Initiative to Address Internet-Scale Computing Challenges

- ICCS
  - We are using Hadoop and Nutch to crawl Blog posts and later process them. Hadoop is also beginning to be used in our teaching and general research activities on natural language processing and machine learning.

- IIIT, Hyderabad
  - We use hadoop for Information Retrieval and Extraction research projects. Also working on map-reduce scheduling research for multi-job environments.
  - Our cluster sizes vary from 10 to 30 nodes, depending on the jobs. Heterogenous nodes with most being Quad 6600s, 4GB RAM and 1TB disk per node. Also some nodes with dual core and single core configurations.

- ImageShack
  - From TechCrunch:
    Rather than put ads in or around the images it hosts, Levin is working on harnessing all the data his service generates about content consumption (perhaps to better target advertising on ImageShack or to syndicate that targeting data to ad networks). Like Google and Yahoo, he is deploying the open-source Hadoop software to create a massive distributed supercomputer, but he is using it to analyze all the data he is collecting.

- Information Sciences Institute (ISI)
  - Used Hadoop and 18 nodes/52 cores to plot the entire internet.

- Iterend
  - using 10 node hdfs cluster to store and process retrieved data.

- Joost
  - Session analysis and report generation

- Katta - Katta serves large Lucene indexes in a grid environment.
  - Uses Hadoop FileSytetm, RPC and IO

- Koubei.com Large local community and local search at China.
Using Hadoop to process apache log, analyzing user's action and click flow and the links click with any specified page in site and more. Using Hadoop to process whole price data user input with map/reduce.

- **Krugle**
  - Source code search engine uses Hadoop and Nutch.

- **Last.fm**
  - 50 nodes (dual xeon LV 2GHz, 4GB RAM, 1TB/node storage and dual xeon L5320 1.86GHz, 8GB RAM, 3TB/node storage).
  - Used for charts calculation, log analysis, A/B testing

- **Lookery**
  - We use Hadoop to process clickstream and demographic data in order to create web analytic reports.
  - Our cluster runs across Amazon's EC2 webservice and makes use of the streaming module to use Python for most operations.

- **Lotame**
  - Using Hadoop and Hbase for storage, log analysis, and pattern discovery/analysis.

- **Mailtrust**
  - 12 node cluster (Dual-Core AMD Opteron 1212, 4-8GB RAM, 1.5TB/node storage)
    - Parses and indexes mail logs for search

- **Mahout**

  Another Apache project using Hadoop to build scalable machine learning algorithms like canopy clustering, k-means and many more to come (naive bayes classifiers, others)

- **Neptune**
  - Another Bigtable cloning project using Hadoop to store large structured data set.
  - 200 nodes (each node has: 2 dual core CPUs, 2TB storage, 4GB RAM)

- **NetSeer**
  - Up to 1000 instances on Amazon EC2
  - Data storage in Amazon S3
  - 50 node cluster in Coloc
  - Used for crawling, processing, serving and log analysis

- **The New York Times**
  - Large scale image conversions
- Used EC2 to run hadoop on a large virtual cluster

  - Ning
    - We use Hadoop to store and process our log files
    - We rely on the Java API for our MR jobs for reporting, analytics, and machine learning, and on a proprietary JavaScript API for ad-hoc queries
    - We use commodity hardware, with 4 cores and 4 GB of RAM per machine

  - Nutch - flexible web search engine software

  - PARC - Used Hadoop to analyze Wikipedia conflicts paper.

  - Powerset / Microsoft - Natural Language Search
    - up to 400 instances on Amazon EC2
    - data storage in Amazon S3
    - Microsoft is now contributing to HBase, a Hadoop subproject (announcement).

  - Pressflip - Personalized Persistent Search
    - Using Hadoop on EC2 to process documents from a continuous web crawl and distributed training of support vector machines
    - Using HDFS for large archival data storage

  - Quantcast
    - 3000 cores, 3500TB. 1PB+ processing each day.
    - Hadoop scheduler with fully custom data path / sorter
    - Significant contributions to KFS filesystem

  - Rapleaf
    - 16 node cluster (each node has: 2 dual core CPUs, 2TB storage, 4GB RAM)
    - We use hadoop to process data relating to people on the web
    - We also involved with Cascading to help simplify how our data flows through various processing stages

  - Search Wikia
    - A project to help develop open source social search tools. We run a 125 node hadoop cluster.

  - SEDNS - Security Enhanced DNS Group
    - We are gathering world wide DNS data in order to discover content distribution networks and configuration issues utilizing Hadoop DFS and MapRed.
- Socialmedia.com
  - 14 node cluster (each node has: 2 dual core CPUs, 2TB storage, 8GB RAM)
  - We use hadoop to process log data and perform on-demand analytics

- Taragana - Web 2.0 Product development and outsourcing services
  - We are using 16 consumer grade computers to create the cluster, connected by 100 Mbps network.
  - Used for testing ideas for blog and other data mining.

- The Lydia News Analysis Project - Stony Brook University
  - We are using Hadoop on 17-node and 103-node clusters of dual-core nodes to process and extract statistics from over 1000 U.S. daily newspapers as well as historical archives of the New York Times and other sources.

- Tailsweep - Ad network for blogs and social media
  - 8 node cluster (Xeon Quad Core 2.4GHz, 8GB RAM, 500GB/node Raid 1 storage)
  - Used as a proof of concept cluster
  - Handling i.e. data mining and blog crawling

- University of Glasgow - Terrier Team
  - 30 nodes cluster (Xeon Quad Core 2.4GHz, 4GB RAM, 1TB/node storage). We use Hadoop to facilitate information retrieval research & experimentation, particularly for TREC, using the Terrier IR platform. The open source release of Terrier includes large-scale distributed indexing using Hadoop Map Reduce.

- University of Maryland
  - We are one of six universities participating in IBM/Google's academic cloud computing initiative. Ongoing research and teaching efforts include projects in machine translation, language modeling, bioinformatics, email analysis, and image processing.

- University of Nebraska Lincoln, Research Computing Facility
  - We currently run one medium-sized Hadoop cluster (200TB) to store and serve up physics data for the computing portion of the Compact Muon Solenoid (CMS) experiment. This requires a filesystem which can download data at multiple Gbps and process data at an even higher rate locally. Additionally, several of our students are involved in research projects on Hadoop.

- Veoh
  - We use a small Hadoop cluster to reduce usage data for internal
metrics, for search indexing and for recommendation data.

- **Visible Measures Corporation** uses Hadoop as a component in our Scalable Data Pipeline, which ultimately powers VisibleSuite and other products. We use Hadoop to aggregate, store, and analyze data related to in-stream viewing behavior of Internet video audiences. Our current grid contains more than 128 CPU cores and in excess of 100 terabytes of storage, and we plan to grow that substantially during 2008.

- **VK Solutions**
  - We use a small Hadoop cluster in the scope of our general research activities at VK Labs to get a faster data access from web applications.
  - We also use Hadoop for filtering and indexing listing, processing log analysis, and for recommendation data.

- **WorldLingo**
  - Hardware: 44 servers (each server has: 2 dual core CPUs, 2TB storage, 8GB RAM)
  - Each server runs Xen with one Hadoop/HBase instance and another instance with web or application servers, giving us 88 usable virtual machines.
  - We run two separate Hadoop/HBase clusters with 22 nodes each.
  - Hadoop is primarily used to run HBase and Map/Reduce jobs scanning over the HBase tables to perform specific tasks.
  - HBase is used as a scalable and fast storage back end for millions of documents.
  - Currently we store 12million documents with a target of 450million in the near future.

- **Yahoo!**
  - More than 100,000 CPUs in ~20,000 computers running Hadoop
  - Our biggest cluster: 2000 nodes (2*4cpu boxes w 4TB disk each)
    - Used to support research for Ad Systems and Web Search
    - Also used to do scaling tests to support development of Hadoop on larger clusters
  - Our Blog - Learn more about how we use Hadoop.

- **Zvents**
  - 10 node cluster (Dual-Core AMD Opteron 2210, 4GB RAM, 1TB/node storage)
  - Run Naive Bayes classifiers in parallel over crawl data to discover event information
- Redpoll
  - Hardware: 35 nodes (2*4cpu 10TB disk 16GB RAM each)
  - We intent to parallelize some traditional classification, clustering algorithms like Naive Bayes, K-Means, EM so that can deal with large-scale data sets.

*When applicable, please include details about your cluster hardware and size.*

last edited 2009-02-09 05:46:19 by ZhuGuanyin