PoweredBy

Powered by Apache Hadoop

This page documents an alphabetical list of institutions that are using Apache Hadoop for educational or production uses. Companies that offer services on or based around Hadoop are listed in Commercial Support. Please include details about your cluster hardware and size. Entries without this may be mistaken for spam references and deleted._ _

To add entries you need write permission to the wiki, which you can get by subscribing to the common-dev@hadoop.apache.org mailing list and asking for permissions on the wiki account username you've registered yourself as. If you are using Apache Hadoop in production you ought to consider getting involved in the development process anyway, by filing bugs, testing beta releases, reviewing the code and turning your notes into shared documentation. Your participation in this process will ensure your needs get met.

_

- Powered by Apache Hadoop
- A
- B
- C
- D
- E
- F
- G
- H
- I
- J
- K
- L
- M
- N
- O
- P
- Q
- R
- S
- T
- U
- V
- W
- X
- Y
- Z

_

A

- *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*
- *Accela Communications*
  - *We use an Apache Hadoop cluster to rollup registration and view data each night.*
  - *Our cluster has 10 1U servers, with 4 cores, 4GB ram and 3 drives*
  - *Each night, we run 112 Hadoop jobs*
  - *It is roughly 4X faster to export the transaction tables from each of our reporting databases, transfer the data to the cluster, perform the rollups, then import back into the databases than to perform the same rollups in the database.*
- *Adobe*
  - *We use Apache Hadoop and Apache HBase in several areas from social services to structured data storage and processing for internal use.*
  - *We currently have about 30 nodes running HDFS, Hadoop and HBase in clusters ranging from 5 to 14 nodes on both production and development. We plan a deployment on an 80 nodes cluster.*
  - *We constantly write data to Apache HBase and run MapReduce jobs to process then store it back to Apache HBase or external systems.*
  - *Our production cluster has been running since Oct 2008.*
- *adyard*
  - *We use Apache Flume, Apache Hadoop and PApache ig for log storage and report generation as well as ad-Targeting.*
  - *We currently have 12 nodes running HDFS and Pig and plan to add more from time to time.*
  - *50% of our recommender system is pure Pig because of it's ease of use.*
  - *Some of our more deeply-integrated tasks are using the streaming API and ruby as well as the excellent Wukong-Library.*
- *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 Apache 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.*
- *Aguja- E-Commerce Data analysis*
    - *We use hadoop, pig and hbase to analyze search log, product view data, and analyze all of our logs*
    - *3 node cluster with 48 cores in total, 4GB RAM and 1 TB storage each.*
- *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 Apache Hadoop for variety of things ranging from ETL style processing and statistics generation to running advanced algorithms for doing behavioral analysis and targeting.*
    - *The cluster that we use for mainly behavioral analysis and targeting has 150 machines, Intel Xeon, dual processors, dual core, each with 16GB Ram and 800 GB hard-disk.*
- *ARA.COM.TR - Ara Com Tr - Turkey's first and only search engine*
    - *We build Ara.com.tr search engine using the Python tools.*
    - *We use Apache Hadoop for analytics.*
    - *We handle about 400TB per month*
    - *Our clusters vary from 10 to 100 nodes*
- *Archive.is*
    - *HDFS, Apache Accumulo, Scala*
    - *Currently 3 nodes (16Gb RAM, 6Tb storage)*
- *Atbrox*
    - *We use Hadoop for information extraction & search, and data analysis consulting*
    - *Cluster: we primarily use Amazon's Elastic MapReduce*
- *ATXcursions*
    - *Two applications that are side products/projects of a local tour company: 1. Sentiment analysis of review websites and social media data. Targeting the tourism industry. 2. Marketing tool that analyzes the most valuable/useful reviewers from sites like Tripadvisor and Yelp as well as social media. Lets marketers and business owners find community members most relevant to their businesses.*
    - *Using Apache Hadoop, HDFS, Hive, and HBase.*
    - *3 node cluster, 4 cores, 4GB RAM.*

B

- *BabaCar*
    - *4 nodes cluster (32 cores, 1TB).*
    - *We use Apache Hadoop for searching and analysis of millions of rental bookings.*
- *Basenfasten*
    - *Experimental installation - various TB storage for logs and digital assets*
    - *Currently 4 nodes cluster*
    - *Using hadoop for log analysis/data mining/machine learning*
- *Benipal Technologies - Big Data. Search. AI.*
    - *35 Node Cluster*
    - _We have been running our cluster with no downtime for over 2 ½ years and have successfully handled over 75 Million files on a 64 GB Namenode with 50 TB cluster storage._
    - *We are heavy MapReduce and Apache HBase users and use Apache Hadoop with Apache HBase for semi-supervised Machine Learning, AI R&D, Image Processing & Analysis, and Apache Lucene index sharding using katta.*
- *Beebler*
    - *14 node cluster (each node has: 2 dual core CPUs, 2TB storage, 8GB RAM)*
    - *We use Apache Hadoop for matching dating profiles*
- *Bixo Labs - Elastic web mining*
    - *The Bixolabs elastic web mining platform uses Hadoop + Cascading to quickly build scalable web mining applications.*
    - *We're doing a 200M page/5TB crawl as part of the public terabyte dataset project.*
    - *This runs as a 20 machine Elastic MapReduce cluster.*
- *BrainPad - Data mining and analysis*
    - *We use Apache Hadoop to summarize of user's tracking data.*
    - *And use analyzing.*
- *Brilig - Cooperative data marketplace for online advertising*
    - *We use Apache Hadoop/MapReduce and Apache Hive for data management, analysis, log aggregation, reporting, ETL into Apache Hive, and loading data into distributed K/V stores*
    - *Our primary cluster is 10 nodes, each member has 2x4 Cores, 24 GB RAM, 6 x 1TB SATA.*
    - *We also use AWS EMR clusters for additional reporting capacity on 10 TB of data stored in S3. We usually use m1.xlarge, 60 - 100 nodes.*
- *Brockmann Consult GmbH - Environmental informatics and Geoinformation services*
    - *We use Apache Hadoop to develop the Calvalus system - parallel processing of large amounts of satellite data.*
    - *Focus on generation, analysis and validation of environmental Earth Observation data products.*
    - *Our cluster is a rack with 20 nodes (4 cores, 8 GB RAM each),*
    - *112 TB diskspace total.*

C

- *Caree.rs*

    - *Hardware: 15 nodes*
    - *We use Apache Hadoop to process company and job data and run Machine learning algorithms for our recommendation engine.*
- *CDU now!*
    - *We use Apache Hadoop for our internal searching, filtering and indexing*
- *Charleston*
    - *Hardware: 15 nodes*
    - *We use Apache Hadoop to process company and job data and run Machine learning algorithms for our recommendation engine.*
- *Cloudspace*
    - *Used on client projects and internal log reporting/parsing systems designed to scale to infinity and beyond.*
    - *Client project: Amazon S3-backed, web-wide analytics platform*
    - *Internal: cross-architecture event log aggregation & processing*
- *Contextweb - Ad Exchange*
    - *We use Hadoop to store ad serving logs and use it as a source for ad optimizations, analytics, reporting and machine learning.*
    - *Currently we have a 50 machine cluster with 400 cores and about 140TB raw storage. Each (commodity) node has 8 cores and 16GB of RAM.*
- *Cooliris - Cooliris transforms your browser into a lightning fast, cinematic way to browse photos and videos, both online and on your hard drive.*
    - *We have a 15-node Hadoop cluster where each machine has 8 cores, 8 GB ram, and 3-4 TB of storage.*
    - *We use Hadoop for all of our analytics, and we use Pig to allow PMs and non-engineers the freedom to query the data in an ad-hoc manner.*
- *Cornell University Web Lab*
    - *Generating web graphs on 100 nodes (dual 2.4GHz Xeon Processor, 2 GB RAM, 72GB Hard Drive)*
- *Criteo - Criteo is a global leader in online performance advertising*
    - *Criteo R&D uses Hadoop as a consolidated platform for storage, analytics and back-end processing, including Machine Learning algorithms*
    - *Two production clusters, each with a corresponding pre-production and an experimental cluster*
    - *More than 58,000 cores in 3,000 machines*
    - *Our biggest cluster: 2,000 machines (24 cores, 90TB disk storage and 256GB RAM)*
        - *Growth to 3,000 machines by end 2017*
    - *We run a mix of jobs and workflows based on a variety of frameworks:*
        - *Hive*
        - *Cascading/Scalding*
        - *Mapreduce*
        - *Spark*
        - *Tez*
        - *Flink*
        - *Mono (.Net)*
        - *Custom Yarn applications (for Machine Learning)*
- *CRS4*
    - *Hadoop deployed dynamically on subsets of a 400-node cluster*
        - *node: two quad-core 2.83GHz Xeons, 16 GB RAM, two 250GB HDDs*
        - *most deployments use our high-performance GPFS (3.8PB, 15GB/s random r/w)*
    - *Computational biology applications*
- *crowdmedia*
    - *Crowdmedia has a 5 Node Hadoop cluster for statistical analysis*
    - *We use Apache Hadoop to analyse trends on Facebook and other social networks*

D

- *Datagraph*
    - *We use Apache Apache adoop for batch-processing large RDF datasets, in particular for indexing RDF data.*
    - *We also use Apache Hadoop for executing long-running offline SPARQL queries for clients.*
    - *We use Amazon S3 and Apache Cassandra to store input RDF datasets and output files.*
    - *We've developed RDFgrid, a Ruby framework for map/reduce-based processing of RDF data.*
    - *We primarily use Ruby, RDF.rb and RDFgrid to process RDF data with Apache Hadoop Streaming.*
    - *We primarily run Apache Hadoop jobs on Amazon Elastic MapReduce, with cluster sizes of 1 to 20 nodes depending on the size of the dataset (hundreds of millions to billions of RDF statements).*
- *Dataium*
    - *We use a combination of Apache Pig and Java based Map/Reduce jobs to sort, aggregate and help make sense of large amounts of data.*
- *Deepdyve*
    - *Elastic cluster with 5-80 nodes*
    - *We use Hadoop to create our indexes of deep web content and to provide a high availability and high bandwidth storage service for index shards for our search cluster.*
- *DEMyC-DICIFO-Universidad Autónoma Chapingo*
    - We use an Apache Hadoop cluster to research, teach and service.
    - *Our cluster has 7 servers, each one with 8 cores, 12GB RAM and 1 drive of 1 TB.*
    - Using Hadoop for data analysis, data visualization, searching, data mining, operations research, machine learning and statistical analysis.
- *Detektei Berlin*
    - *We are using Hadoop in our data mining and multimedia/internet research groups.*
    - *3 node cluster with 48 cores in total, 4GB RAM and 1 TB storage each.*
- *Detikcom - Indonesia's largest news portal*
    - *We use Apache Hadoop, Apache Pig and Apache HBase to analyze search log, generate Most View News, generate top wordcloud, and analyze all of our logs*
    - *Currently We use 9 nodes*
- *devdaily.com*

    - *We use Apache Hadoop and Apache Nutch to research data on programming-related websites, such as looking for current trends, story originators, and related information.*
    - *We're currently using three nodes, with each node having two cores, 4GB RAM, and 1TB storage. We'll expand these once we settle on our related technologies (Scala, Apache Pig, Apache HBase, other).*
- *DropFire*
    - *We generate Pig Latin scripts that describe structural and semantic conversions between data contexts*
    - *We use Apache Hadoop to execute these scripts for production-level deployments*
    - *Eliminates the need for explicit data and schema mappings during database integration*

E

- *EBay*
    - *532 nodes cluster (8 * 532 cores, 5.3PB).*
    - *Heavy usage of Java MapReduce, Apache Pig, Apache Hive, Apache HBase*
    - *Using it for Search optimization and Research.*
- *eCircle*
    - *two 60 nodes cluster each >1000 cores, total 5T Ram, 1PB*
    - *mostly Apache HBase, some M/R*
    - *marketing data handling*
- *Enet, 'Eleftherotypia' newspaper, Greece*
    - *Experimental installation - storage for logs and digital assets*
    - *Currently 5 nodes cluster*
    - *Using Apache Hadoop for log analysis/data mining/machine learning*
- *Enormo*
    - *4 nodes cluster (32 cores, 1TB).*
    - *We use Apache Hadoop to filter and index our listings, removing exact duplicates and grouping similar ones.*
    - *We plan to use Apache Pig very shortly to produce statistics.*
- *ESPOL University (Escuela Superior Politécnica del Litoral) in Guayaquil, Ecuador*
    - *4 nodes proof-of-concept cluster.*
    - *We use Apache Hadoop in a Data-Intensive Computing capstone course. The course projects cover topics like information retrieval, machine learning, social network analysis, business intelligence, and network security.*
    - *The students use on-demand clusters launched using Amazon's EC2 and EMR services, thanks to its AWS in Education program.*
- *ETH Zurich Systems Group*
    - *We are using Apache 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*
    - *Facial similarity and recognition across large datasets.*
    - *Image content based advertising and auto-tagging for social media.*
    - *Image based video copyright protection.*
- *Explore.To Yellow Pages - Explore To Yellow Pages*
    - *We use Apache Hadoop for our internal search, filtering and indexing*
    - *Elastic cluster with 5-80 nodes*

F

- *Facebook*
    - *We use Apache Hadoop to store copies of internal log and dimension data sources and use it as a source for reporting/analytics and machine learning.*
    - *Currently we have 2 major clusters:*
        - *A 1100-machine cluster with 8800 cores and about 12 PB raw storage.*
        - *A 300-machine cluster with 2400 cores and about 3 PB raw storage.*
        - *Each (commodity) node has 8 cores and 12 TB of storage.*
        - *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.*
- *FollowNews*
    - *We use Hadoop for storing logs, news analysis, tag analysis.*
- *FOX Audience Network*
    - *40 machine cluster (8 cores/machine, 2TB/machine storage)*
    - *70 machine cluster (8 cores/machine, 3TB/machine storage)*
    - *30 machine cluster (8 cores/machine, 4TB/machine storage)*
    - *Use for log analysis, data mining and machine learning*
- *Forward3D*
    - *5 machine cluster (8 cores/machine, 5TB/machine storage)*
    - *Existing 19 virtual machine cluster (2 cores/machine 30TB storage)*
    - *Predominantly Apache Hive and Streaming API based jobs (~20,000 jobs a week) using our Ruby library, or see the canonical WordCount example.*
    - *Daily batch ETL with a slightly modified clojure-hadoop*
    - *Log analysis*
    - *Data mining*
    - *Machine learning*
- *FQuotes*
    - *We use Hadoop for analyzing quotes, quote authors and quote topics.*
- *Freestylers - Image retrieval engine*

    - *We, the Japanese company Freestylers, use Hadoop to build the image processing environment for image-based product recommendation system mainly on Amazon EC2, from April 2009.*
    - *Our Apache Hadoop environment produces the original database for fast access from our web application.*
    - *We also uses Hadoop to analyzing similarities of user's behavior.*

G

- *GBIF (Global Biodiversity Information Facility) - nonprofit organization that focuses on making scientific data on biodiversity available via the Internet*
    - *18 nodes running a mix of Apache Hadoop and Apache HBase*
    - *Apache Hive ad hoc queries against our biodiversity data*
    - *Regular Apache Oozie workflows to process biodiversity data for publishing*
    - *All work is Open source (e.g. Oozie workflow, Ganglia)*
- *GIS.FCU*
    - *Feng Chia University*
    - *3 machine cluster (4 cores, 1TB/machine)*
    - *storage for sensor data*
- *Google*
    - *University Initiative to Address Internet-Scale Computing Challenges*
- *Gruter. Corp.*
    - *30 machine cluster (4 cores, 1TB~2TB/machine storage)*
    - *storage for blog data and web documents*
    - *used for data indexing by MapReduce*
    - *link analyzing and Machine Learning by MapReduce*
- *Gewinnspiele*
    - *6 node cluster (each node has: 4 dual core CPUs, 1,5TB storage, 4GB RAM, RedHat OS)*
    - *Using Apache Hadoop for our high speed data mining applications in corporation with Twilight*
- *GumGum*
    - *9 node cluster (Amazon EC2 c1.xlarge)*
    - *Nightly MapReduce jobs on Amazon Elastic MapReduce process data stored in S3*
    - *MapReduce jobs written in Groovy use Apache Hadoop Java APIs*
    - *Image and advertising analytics*
- *Greece.com*
    - *Using Apache Hadoop for analyzing data for millions of images, log analysis, data mining*

H

- *Hadoop Korean User Group, a Korean Local Community Team Page.*
    - *50 node cluster In the Korea university network environment.*
    - *Pentium 4 PC, HDFS 4TB Storage*
- *Used for development projects*
    - *Retrieving and Analyzing Biomedical Knowledge*
    - *Latent Semantic Analysis, Collaborative Filtering*
- *Helprace, a customer service software.*
    - *3 node cluster (4 cores, 32 GB of RAM each).*
    - *Hadoop for search engine analtyics and internal searching and data filtering*
- *Hotels & Accommodation*
    - *3 machine cluster (4 cores/machine, 2TB/machine)*
    - *Apache Hadoop for data for search and aggregation*
    - *Apache HBase hosting*
- *Hulu*
    - *13 machine cluster (8 cores/machine, 4TB/machine)*
    - *Log storage and analysis*
    - *Apache HBase hosting*
- *Hundeshagen*
    - *6 node cluster (each node has: 4 dual core CPUs, 1,5TB storage, 4GB RAM, RedHat Enterprise Linux)*
    - *Using Apache Hadoop for our high speed data mining applications in corporation with Online Scheidung*
- *Hadoop Taiwan User Group*
- *Hipotecas y euribor*
    - *Evolución del euribor y valor actual*
    - *Simulador de hipotecas en crisis económica*
- *Hosting Habitat*
    - *We use a customised version of Apache Hadoop and Apache Nutch in a currently experimental 6 node/Dual Core cluster environment.*
    - *What we crawl are our clients Websites and from the information we gather. We fingerprint old and non updated software packages in that shared hosting environment. We can then inform our clients that they have old and non updated software running after matching a signature to a Database. With that information we know which sites would require patching as a free and courtesy service to protect the majority of users. Without the technologies of Nutch and Hadoop this would be a far harder to accomplish task.*

I

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

- *ICCS*
  - *We are using Apache Hadoop and Apache 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 targetting 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.*
- *IMVU*
  - *We use Apache Hadoop to analyze our virtual economy*
  - *We also use Apache Hive to access our trove of operational data to inform product development decisions around improving user experience and retention as well as meeting revenue targets*
  - *Our data is stored in Amazon S3 and pulled into our clusters of up to 4 m1.large EC2 instances. Our total data volume is on the order of 5Tb*
- *Infolinks*
  - *We use Apache Hadoop to analyze production logs and to provide various statistics on our In-Text advertising network.*
  - *We also use Apache Hadoop/Apache HBase to process user interactions with advertisements and to optimize ad selection.*
- *Information Sciences Institute (ISI)*
  - *Used Apache Hadoop and 18 nodes/52 cores to plot the entire internet.*
- *Infochimps*
  - *30 node AWS EC2 cluster (varying instance size, currently EBS-backed) managed by Chef & Poolparty running Apache Hadoop 0.20.2 +228, Apache Pig 0.5.0+30, Azkaban 0.04, Wukong*
  - *Used for ETL & data analysis on terascale datasets, especially social network data.*
- *Inmobi*
  - Running Apache Hadoop on around 700 nodes (16800 cores, 5+ PB) in 6 Data Centers for ETL, Analytics, Data Science and Machine Learning
- *Iterend*
  - *using 10 node HDFS cluster to store and process retrieved data on.*
- *iNews*
  - *Using Hadoop for crawling, data analysis, log analysis.*

J

- *Joost*
  - *Session analysis and report generation*
- *Journey Dynamics*
  - *Using Apache Hadoop MapReduce to analyse billions of lines of GPS data to create TrafficSpeeds, our accurate traffic speed forecast product.*

K

- *Kalooga - Kalooga is a discovery service for image galleries.*
  - *Uses Apache Hadoop, Apache HBase, Apache Chukwa and Apache Pig on a 20-node cluster for crawling, analysis and events processing.*
- *Katta - Katta serves large Lucene indexes in a grid environment.*
  - *Uses Apache Hadoop FileSytem, RPC and IO*
- *Korrelate - Korrelate correlates online media to offline purchases.*
  - *Use Apache Sqoop to get data out of our MPP database into Apache HBase*
  - *Use Apache HBase and Apache Pig to process events, summarize event data for reporting, and generate reports on online to offline correlations.*
  - *When our transition is complete mid-2014, we will be processing billions of events a month through HBase and have a total data size on the order of 5 TB*
- *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 Apache Hadoop and Apache Nutch.*

L

- *Language, Interaction and Computation Laboratory (Clic - CIMeC)*
  - *Hardware: 10 nodes, each node has 8 core and 8GB of RAM*
  - *Studying verbal and non-verbal communication.*
- *Last.fm*
  - *100 nodes*
  - *Dual quad-core Xeon L5520 @ 2.27GHz & L5630 @ 2.13GHz , 24GB RAM, 8TB(4x2TB)/node storage.*
  - *Used for charts calculation, royalty reporting, log analysis, A/B testing, dataset merging*

    - *Also used for large scale audio feature analysis over millions of tracks*
- *Lineberger Comprehensive Cancer Center - Bioinformatics Group*
    - *This is the cancer center at UNC Chapel Hill. We are using Apache Hadoop/Apache HBase for databasing and analyzing Next Generation Sequencing (NGS) data produced for the Cancer Genome Atlas (TCGA) project and other groups. This development is based on the SeqWare open source project which includes SeqWare Query Engine, a database and web service built on top of Apache HBase that stores sequence data types. Our prototype cluster includes:*
        - *8 dual quad core nodes running CentOS*
        - *total of 48TB of HDFS storage*
        - *HBase & Hadoop version 0.20*
- *Legolas Media*
- *LinkedIn*
    - _We have multiple grids divided up based upon purpose.
      **_ Hardware: _
      ***_ ~800 Westmere-based HP SL 170x, with 2x4 cores, 24GB RAM, 6x2TB SATA _
      ***_ ~1900 Westmere-based SuperMicro X8DTT-H, with 2x6 cores, 24GB RAM, 6x2TB SATA _
      ***_ ~1400 Sandy Bridge-based SuperMicro with 2x6 cores, 32GB RAM, 6x2TB SATA _
      **_ Software: _
      ***_ RHEL 6.3 _
      ***_ Sun JDK 1.6.0_32 _
      ***_ Apache Hadoop 0.20.2+patches and Apache Hadoop 1.0.4+patches _
      ***_ Pig 0.10 + DataFu _
      ***_ Azkaban and Azkaban 2 for scheduling _
      ***_ Apache Hive, Apache Avro, Apache Kafka, and other bits and pieces... _
      **_ We use these things for discovering People You May Know and other fun facts. _
      *_ LiveBet _
      **_ We use Hadoop for storing logs, odds analysis, markets analysis. _
      *_ Lookery _
      **_ We use Hadoop to process clickstream and demographic data in order to create web analytic reports. _
      **_ Our cluster runs across Amazon's EC2 infrastructure and makes use of the streaming module to use Python for most operations. _
      *_ Lotame _
      **_ Using Apache Hadoop and Apache HBase for storage, log analysis, and pattern discovery/analysis. _

M

*_ Markt24 _
**_ We use Apache Hadoop to filter user behaviour, recommendations and trends from externals sites _
**_ Using zkpython to connect with Apache Zookeeper_
**_ Used EC2, no using many small machines (8GB Ram, 4 cores, 1TB) _
*_ MicroCode _
**_ 18 node cluster (Quad-Core Intel Xeon, 1TB/node storage) _
**_ Financial data for search and aggregation _
**_ Customer Relation Management data for search and aggregation _
*_ Media 6 Degrees _
**_ 20 node cluster (dual quad cores, 16GB, 6TB) _
**_ Used log processing, data analysis and machine learning. _
**_ Focus is on social graph analysis and ad optimization. _
**_ Use a mix of Java, Pig and Hive. _
*_ Medical Side Fx _
**_ Use Apache Hadoop to analyze FDA AERS(Adverse Events Reporting System) data and present an easy way to search and query side effects of medicines_
**_ Apache Lucene is used for indexing and searching. _
*_ MeMo News - Online and Social Media Monitoring _
**_ we use Apache Hadoop _
***_ as platform for distributed crawling _
***_ to store and process unstructured data, such as news and social media (Apache Hadoop, Apache Pig, MapRed and Apache HBase) _
***_ log file aggregation and processing (Apache Flume) _
*_ Mercadolibre.com _
**_ 20 nodes cluster (12 * 20 cores, 32GB, 53.3TB) _
**_ Customers log on on-line apps _
**_ Operations log processing _
**_ Use java, Apache Pig, Apache Hive, Apache Oozie _
*_ MobileAnalytic.TV _
**_ We use Hadoop to develop MapReduce algorithms: _
***_ Information retrieval and analytics _
***_ Machine generated content - documents, text, audio, & video _
***_ Natural Language Processing _
**_ Project portfolio includes: * Natural Language Processing _
***_ Mobile Social Network Hacking _
***_ Web Crawlers/Page scrapping _
***_ Text to Speech _
***_ Machine generated Audio & Video with remuxing _
***_ Automatic PDF creation & IR _
***_ 2 node cluster (Windows Vista/CYGWIN, & CentOS) for developing MapReduce programs. _
*_ Moesif API Insights _
**_ We use Hadoop for ETL and processing time series event data for alerts/notifications along with visualizations for frontend._
**_ 2 master nodes and 6 data nodes running on Azure using HDInsight_
*_ MyLife _

**_ 18 node cluster (Quad-Core AMD Opteron 2347, 1TB/node storage) _
**_ Powers data for search and aggregation _
*_ Mail.gr - we use HDFS for hosting our users' mailboxes ._

N

*_ NAVTEQ Media Solutions _
**_ We use Apache Hadoop/Apache Mahout to process user interactions with advertisements to optimize ad selection. _
*_ Neptune _
**_ Another Apache 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 Apache Pig for reporting, analytics, Cascading for machine learning, and on a proprietary JavaScript API for ad-hoc queries _
**_ We use commodity hardware, with 8 cores and 16 GB of RAM per machine _

O

*_ Openstat _
**_ Hadoop is used to run a customizable web analytics log analysis and reporting _
**_ 50-node production workflow cluster (dual quad-core Xeons, 16GB of RAM, 4-6 HDDs) and a couple of smaller clusters for individual analytics purposes _
**_ About 500 mln of events processed daily, 15 bln monthly _
**_ Cluster generates about 25 GB of reports daily _
**_ Technologies used: Cascading, Janino _
*_ optivo - Email marketing software _
**_ We use Apache Hadoop to aggregate and analyse email campaigns and user interactions. _
**_ Development is based on the github repository. _

P

*_ Papertrail - Hosted syslog and app log management _
**_ Hosted syslog and app log service can feed customer logs into Apache Hadoop for their analysis (usually with Hive) _
**_ Most customers load gzipped TSVs from S3 (which are uploaded nightly) into Amazon Elastic MapReduce _
*_ PARC - Used Hadoop to analyze Wikipedia conflicts paper. _
*_ PCPhase - A Japanese mobile integration company _
**_ Using Apache Hadoop/Apache HBase in conjunction with Apache Cassandra to analyze log and generate reports for a large mobile web site. _
**_ 4 nodes in a private cloud with 4 cores, 4G RAM & 500G storage each. _
*_ Performable - Web Analytics Software _
**_ We use Apache Hadoop to process web clickstream, marketing, CRM, & email data in order to create multi-channel analytic reports. _
**_ Our cluster runs on Amazon's EC2 webservice and makes use of Python for most of our codebase. _
*_ Pharm2Phork Project - Agricultural Traceability _
**_ Using Hadoop on EC2 to process observation messages generated by RFID/Barcode readers as items move through supply chain. _
**_ Analysis of BPEL-generated log files for monitoring and tuning of workflow processes. _
*_ Powerset / Microsoft - Natural Language Search _
**_ up to 400 instances on Amazon EC2 _
**_ data storage in Amazon S3 _
**_ Microsoft is now contributing to Apache HBase (announcement). _
*_ Pressflip - Personalized Persistent Search _
**_ Using Apache 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 _
*_ Pronux _
**_ 4 nodes cluster (32 cores, 1TB). _
**_ We use Apache Hadoop for searching and analysis of millions of bookkeeping postings _
**_ Also used as a proof of concept cluster for a cloud based ERP system _
*_ PokerTableStats _
**_ 2 nodes cluster (16 cores, 500GB). _
**_ We use Apache Hadoop for analyzing poker players game history and generating gameplay related players statistics _
*_ Portabilité _
**_ 50 node cluster in a colocated site. _
**_ Also used as a proof of concept cluster for a cloud based ERP system. _
*_ PSG Tech, Coimbatore, India _
**_ Multiple alignment of protein sequences helps to determine evolutionary linkages and to predict molecular structures. The dynamic nature of the algorithm coupled with data and compute parallelism of Hadoop data grids improves the accuracy and speed of sequence alignment. Parallelism at the sequence and block level reduces the time complexity of MSA problems. The scalable nature of Hadoop makes it apt to solve large scale alignment

problems. _
**_ Our cluster size varies from 5 to 10 nodes. Cluster nodes vary from 2950 Quad Core Rack Server, with 2x6MB Cache and 4 x 500 GB SATA Hard Drive to E7200 / E7400 processors with 4 GB RAM and 160 GB HDD. _

Q

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

R

*_ Rackspace _
**_ 30 node cluster (Dual-Core, 4-8GB RAM, 1.5TB/node storage) _
**_ Parses and indexes logs from email hosting system for search: http://blog.racklabs.com/?p=66 _
*_ Rakuten - Japan's online shopping mall _
**_ 69 node cluster _
**_ We use Apache Hadoop to analyze logs and mine data for recommender system and so on. _
*_ Rapleaf _
**_ 80 node cluster (each node has: 2 quad core CPUs, 4TB storage, 16GB 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 _
*_ Recruit _
**_ Hardware: 50 nodes (2*4cpu 2TB*4 disk 16GB RAM each) _
**_ We use Apache Hive to analyze logs and mine data for recommendation. _
*_ reisevision _
**_ We use Apache Hadoop for our internal search _
*_ Redpoll _
**_ Hardware: 35 nodes (2*4cpu 10TB disk 16GB RAM each) _
**_ We intend to parallelize some traditional classification, clustering algorithms like Naive Bayes, K-Means, EM so that can deal with large-scale data sets. _
*_ Resu.me _
**_ Hardware: 5 nodes _
**_ We use Apache Hadoop to process user resume data and run algorithms for our recommendation engine. _
*_ RightNow Technologies - Powering Great Experiences _
**_ 16 node cluster (each node has: 2 quad core CPUs, 6TB storage, 24GB RAM) _
**_ We use Apache Hadoop for log and usage analysis _
**_ We predominantly leverage Hive and HUE for data access _
*_ Rodacino_
**_ We use Apache Hadoop for crawling news sites and log analysis._
**_ We also use Apache Cassandra as our back end and Apache Lucene for searching capabilities_
*_ Rovi Corporation_
**_ We use Apache Hadoop, Apache Pig and map/reduce to process extracted SQL data to generate JSON objects that are stored in MongoDB and served through our web services _
**_ We have two clusters with a total of 40 nodes with 24 cores at 2.4GHz and 128GB RAM _
**_ Each night we process over 160 pig scripts and 50 map/reduce jobs that process over 600GB of data _
*_ Rubbellose _

- - We use AWS EMR with Cascading to create personalization and recommendation job flows

S

*_ SARA, Netherlands _
**_ SARA has initiated a Proof-of-Concept project to evaluate the Hadoop software stack for scientific use. _
*_ 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. _
*_ Sematext International _
**_ We use Hadoop to store and analyze large amounts search and performance data for our Search Analytics and Scalable Performance Monitoring services. _
*_ SLC Security Services LLC _
**_ 18 node cluster (each node has: 4 dual core CPUs, 1TB storage, 4GB RAM, RedHat OS) _
**_ We use Hadoop for our high speed data mining applications _
*_ Sling Media _
**_ We have a core analytics group that is using a 10-Node cluster running RedHat OS _
**_ Hadoop is used as an infrastructure to run MapReduce (MR) algorithms on a number of raw data _
**_ Raw data ingest happens hourly. Raw data comes from hardware and software systems out in the field _
**_ Ingested and processed data is stored into a relational DB and rolled up using Hive/Pig _
**_ Plan to implement Mahout to build recommendation engine _
*_ 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 _
*_ Spadac.com _
**_ We are developing the MrGeo (Map/Reduce Geospatial) application to allow our users to bring cloud computing to geospatial processing. _
**_ We use Apache HDFS and MapReduce to store, process, and index geospatial imagery and vector data. _
**_ MrGeo is soon to be open sourced as well. _
*_ Specific Media _
**_ We use Apache Hadoop for log aggregation, reporting and analysis _
**_ Two Apache Hadoop clusters, all nodes 16 cores, 32 GB RAM _
**_ Cluster 1: 27 nodes (total 432 cores, 544GB RAM, 280TB storage) _
**_ Cluster 2: 111 nodes (total 1776 cores, 3552GB RAM, 1.1PB storage) _
**_ We contribute to Hadoop and related projects where possible, see http://code.google.com/p/bigstreams/ and http://code.google.com/p/hadoop-gpl-packing/ _
*_ Spotify _
**_ We use Apache Hadoop for content generation, data aggregation, reporting, analysis (see more: The Evolution of Hadoop at Spotify - Through Failures and Pain) and even for generating music recommendations (How Apache Drives Music Recommendations At Spotify_
**_ 1650 node cluster : 43,000 virtualized cores, ~70TB RAM, ~65 PB storage (read more about our Hadoop issues while growing fast: Hadoop Adventures At Spotify)_
**_ +20,000 daily Hadoop jobs (scheduled by Luigi, our open-sourced job orchestrator - code and video)_
*_ Stampede Data Solutions (Stampedehost.com) _
**_ Hosted Apache Hadoop data warehouse solution provider _
*_ Sthenica _
**_ We use Apache Hadoop for sentiment analysis/social media monitoring and personalized marketing _
**_ Using 3 node cluster in a visualized environment with a 4th node for SQL reporting _
*_ StumbleUpon (StumbleUpon.com) _
**_ We use Apache HBase to store our recommendation information and to run other operations. We have HBase committers on staff. _

T

*_ 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 Apache 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 _
*_ Technical analysis and Stock Research _
**_ Generating stock analysis on 23 nodes (dual 2.4GHz Xeon, 2 GB RAM, 36GB Hard Drive) _
*_ Tegatai _
**_ Collection and analysis of Log, Threat, Risk Data and other Security Information on 32 nodes (8-Core Opteron 6128 CPU, 32 GB RAM, 12 TB Storage per node) _
*_ Telefonica Research _
**_ We use Apache Hadoop in our data mining and user modeling, multimedia, and internet research groups. _
**_ 6 node cluster with 96 total cores, 8GB RAM and 2 TB storage per machine. _
*_ Telenav _
**_ 60-Node cluster for our Location-Based Content Processing including machine learning algorithms for Statistical Categorization, Deduping, Aggregation & Curation (Hardware: 2.5 GHz Quad-core Xeon, 4GB RAM, 13TB HDFS storage). _
**_ Private cloud for rapid server-farm setup for stage and test environments.(Using Elastic N-Node cluster) _
**_ Public cloud for exploratory projects that require rapid servers for scalability and computing surges (Using Elastic N-Node cluster) _
*_ Tepgo- E-Commerce Data analysis _
**_ We use Apache Hadoop, Apache Pig and Apache HBase to analyze search log, product view data, and analyze usage logs _
**_ 3 node cluster with 48 cores in total, 4GB RAM and 1 TB storage each. _
*_ Tianya _
**_ We use Apache Hadoop for log analysis. _
*_ TubeMogul _
**_ We use Apache Hadoop HDFS, Map/Reduce, Apache Hive and Apache HBase _
**_ We manage over 300 TB of HDFS data across four Amazon EC2 Availability Zone _
*_ tufee _
**_ We use Apache Hadoop for searching and indexing _
*_ Twitter _
**_ We use Apache Hadoop to store and process tweets, log files, and many other types of data generated across Twitter. We store all data as compressed LZO files. _
**_ We use both Scala and Java to access Hadoop's MapReduce APIs _
**_ We use Apache Pig heavily for both scheduled and ad-hoc jobs, due to its ability to accomplish a lot with few statements. _
**_ We employ committers on Apache Pig, Apache Avro, Apache Hive, and Apache Cassandra, and contribute much of our internal Hadoop work to opensource (see hadoop-lzo) _
**_ For more on our use of Apache Hadoop, see the following presentations: Hadoop and Pig at Twitter and Protocol Buffers and Hadoop at Twitter _
*_ Tynt _
**_ We use Apache Hadoop to assemble web publishers' summaries of what users are copying from their websites, and to analyze user engagement on the web. _
**_ We use Apache Pig and custom Java map-reduce code, as well as Apache Chukwa. _
**_ We have 94 nodes (752 cores) in our clusters, as of July 2010, but the number grows regularly. _

U

*_ Universidad Distrital Francisco Jose de Caldas (Grupo GICOGE/Grupo Linux UD GLUD/Grupo GIGA) _
**_ 5 node low-profile cluster. We use Hadoop to support the research project: Territorial Intelligence System of Bogota City. _
*_ University of Freiburg - Databases and Information Systems _
**_ 10 nodes cluster (Xeon Dual Core 3.16GHz, 4GB RAM, 3TB/node storage). _
**_ Our goal is to develop techniques for the Semantic Web that take advantage of MapReduce (Hadoop) and its scaling-behavior to keep up with the growing proliferation of semantic data. _
**_ RDFPath is an expressive RDF path language for querying large RDF graphs with MapReduce. _
**_ PigSPARQL is a translation from SPARQL to Pig Latin allowing to execute SPARQL queries on large RDF graphs with MapReduce. _
*_ 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, Holland Computing Center _
**_ We currently run one medium-sized Hadoop cluster (1.6PB) 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 Apache Hadoop. _
*_ University of Twente, Database Group _
**_ We run a 16 node cluster (dual-core Xeon E3110 64 bit processors with 6MB cache, 8GB main memory, 1TB disk) as of December 2008. We teach MapReduce and use Apache Hadoop in our computer science master's program, and for information retrieval research. For more information, see: http://mirex.sourceforge.net/_

V

*_ Veoh _
**_ We use a small Apache Hadoop cluster to reduce usage data for internal metrics, for search indexing and for recommendation data. _
*_ Bygga hus _
**_ We use an Apache Hadoop cluster to for search and indexing for our projects. _
*_ 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 Apache 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 Apache Hadoop for filtering and indexing listing, processing log analysis, and for recommendation data. _

W

*_ Web Alliance _
**_ We use Apache Hadoop for our internal search engine optimization (SEO) tools. It allows us to store, index, search data in a much faster way. _
**_ We also use it for logs analysis and trends prediction. '

- *Webmaster Site*
  - *We use Apache Hadoop for our webmaster tools. It allows us to store, index, search data in a much fast way. We also use it for logs analysis and trends prediction.*
  - *4 node cluster (each node has: 4 core AMD CPUs, 2TB storage, 32GB RAM)*
  - *We use Apache Hadoop to process log data and perform on-demand analytics as well*
- *WorldLingo*
  - *Hardware: 44 servers (each server has: 2 dual core CPUs, 2TB storage, 8GB RAM)*
  - *Each server runs Xen with one Apache Hadoop/Apache HBase instance and another instance with web or application servers, giving us 88 usable virtual machines.*
  - *We run two separate Apache Hadoop/Apache HBase clusters with 22 nodes each.*
  - *Apache Hadoop is primarily used to run Apache HBase and Map/Reduce jobs scanning over the Apache HBase tables to perform specific tasks.*
  - *Apache HBase is used as a scalable and fast storage back end for millions of documents.*
  - *Currently we store 12 million documents with a target of 450 million in the near future.*

X

Y

- *Yahoo!*
  - *More than 100,000 CPUs in >40,000 computers running Hadoop*
  - *Our biggest cluster: 4500 nodes (2*4cpu boxes w 4*1TB disk & 16GB RAM)*
    - *Used to support research for Ad Systems and Web Search*
    - *Also used to do scaling tests to support development of Apache Hadoop on larger clusters*

    - *Our Blog - Learn more about how we use Apache Hadoop.*
    - *>60% of Hadoop Jobs within Yahoo are Apache Pig jobs.*
- *YMC AG*
    - *operating a Cloudera cluster for media monitoring purpose*
    - *offering technical and operative consulting for the Apache Hadoop stack + ecosystem*
    - *editor of Hannibal, a open-source tool to visualize Apache Apache HBase regions sizes & splits that helps running HBase in production*

Z

- *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*