PoweredBy

Applications and organizations using Pig include (alphabetically):

- AOL
  - AOL has multiple clusters from a few nodes to several hundred nodes.
  - We use Hadoop for analytics and batch data processing for various applications.
  - Hadoop is used by MapQuest, Ad, Search, Truveo, and Media groups.
  - All of our jobs are written in Pig or native map reduce.

- ChaCha
  - We use Hadoop/Pig/etc. to data mine our SMS, web, and mobile app (iOS, Android) traffic.
  - The most commonly-run jobs do things like try to identify trending topics in questions asked, as well as identify patterns in questions that we need humans to answer so that we can prioritize those patterns/topics/questions for automation development.
  - Almost all of our jobs are written in Pig.

- 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.<
>

- Dataium
  - We use Pig to sort and prep our data before it is handed off to our Java Map/Reduce jobs.

- DropFire
  - We generate Pig Latin scripts that describe structural and semantic conversions between data contexts
  - We use Hadoop to execute these scripts for production-level deployments
  - Eliminates the need for explicit data and schema mappings during database integration

- LinkedIn
  - 3x30 Nehalem-based node grids, with 2x4 cores, 16GB RAM, 8x1TB storage using ZFS in a JBOD configuration.
  - We use Hadoop and Pig for discovering People You May Know and other fun facts.

- Mendeley
  - We are creating a platform for researchers to collaborate and share their research online
  - We moved all our catalogue stats and analysis to HBase and Pig
  - We are using Scribe in combination with Pig for all our server, application and user log processing.
  - Pig helps our business analytics, user experience evaluation, feature feedback and more out of these logs.
  - You can find more on how we use Pig and HBase on these slides: http://www.slideshare.net/danharvey/hbase-at-mendeley

- Mortar Data
  - We provide an open-source development framework and Hadoop Platform-as-a-Service
  - Our service is powered by Pig, which we run on private, ephemeral clusters in Amazon Web Services

- Ning
  - We use Hadoop to store and process our log file
  - We rely on Apache Pig for reporting, analytics, Cascading for machine learning, and on a proprietary [/hadoop/JavaScript] API for ad-hoc queries
  - We use commodity hardware, with 8 cores and 16 GB of RAM per machine

- Nokia | Ovi
  - We use Pig for exploring unstructured datasets coming from logs, database dumps, data feeds, etc.
  - Several data pipelines that go into building product datasets and for further analysis use Pig tied together with Oozie to other jobs
  - We have multiple Hadoop clusters, some for R&D and some for production jobs
  - In R&D we run on very commodity hardware: 8-core, 16GB RAM, 4x 1TB disk per data node
- PayPal
  - We use Pig to analyze transaction data in order to prevent fraud.
  - We are the main contributors to the Pig-Eclipse project.

- Realweb - Internet Advertising company based in Russia.
  - We are using Pig over Hadoop to compute statistics on banner views, clicks, user behavior on target websites after click, etc.
  - We've chosen Cloudera Hadoop (http://www.cloudera.com/hadoop/) packages on Ubuntu servers 10.04. Each machine has 2/4 cores, 4 GB ram, and 1 TB of storage.
  - All jobs are written using Pig language and only few user defined functions were needed to achieve our needs.

- Salesforce.com
  - We have multiple clusters in production, a 10 node and 20 node development clusters
  - Hadoop (native Java MapReduce) is used for Search and Recommendations
  - We are using Apache Pig for log processing and Search, and to generate usage reports for several products and features at SFDC
  - Pig makes it easy to develop custom UDFs. We developed our own library containing UDFs and loaders and are actively

contributing back to the community
    - The goal is to allow Hadoop/Pig to be used across Data Warehouse, Analytics and other teams making it easier for folks outside engineering to use data

- SARA Computing and Networking Services
    - We provide a Hadoop service for scientific computing in The Netherlands
    - Pig is being used by a number of scientists for fast exploration of large datasets
    - Sciences extensively using Pig include Information Retrieval and Natural Language Processing
    - Read more on our use of Hadoop in this presentation
    - Read about selected use cases on Hadoop in this blogpost

- Stanford University WebBase Project
    - The WebBase project has crawled and archived fixed sets of Web sites very regularly for several years. These time series site snapshots include government, and a number of topic-specific destinations. We use Pig to power an emerging interface to these archives for social scientists. The goal is to support these scientists in analyzing the archives for their research. Pig and Hadoop are used for the underlying processing and indexing.

- Twitter
    - We use Pig extensively to process usage logs, mine tweet data, and more.
    - We have maintain Elephant Bird, a set of libraries for working with Pig, LZO compression, protocol buffers, and more.
    - More details can be seen in this presentation: http://www.slideshare.net/kevinweil/nosql-at-twitter-nosql-eu-2010

- Tynt
    - We use Hadoop to assemble web publishers' summaries of what users are copying from their websites, and to analyze user engagement on the web.
    - We use Pig and custom Java map-reduce code, as well as chukwa.
    - We have 94 nodes (752 cores) in our clusters, as of July 2010, but the number grows regularly.

- WhitePages
    - We use Pig to clean, merge, and filter multi-billion record data sets for our People and Business Search applications.
    - We use Pig to analyze daily search and web logs to produce Key Performance Indicators for our most critical web services.

- Yahoo!
    - More than 100,000 CPUs in >25,000 computers running Hadoop
    - Our biggest cluster: 4000 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 Hadoop on larger clusters
    - Our Blog - Learn more about how we use Hadoop.
    - >40% of Hadoop Jobs within Yahoo are Pig jobs.