A List of Big Data Technologies

Mai Hai Thanh

KIWI@ETRI

Source of this useful summary: http://blog.andreamostosi.name/big-data/

Frameworks

- **Apache Hadoop**: framework for distributed processing. Integrates MapReduce (parallel processing), YARN (job scheduling) and HDFS (distributed file system)

Distributed Programming

- **AddThis Hydra**: distributed data processing and storage system originally developed at AddThis
- **Akela**: Mozilla's utility library for Hadoop, HBase, Pig, etc.
- **Amazon Lambda**: a compute service that runs your code in response to events and automatically manages the compute resources for you
- **AMPLab SIMR**: run Spark on Hadoop MapReduce v1
- **AMPLab Succinct**: Enabling Queries on Compressed Data
- **Apache Crunch**: a simple Java API for tasks like joining and data aggregation that are tedious to implement on plain MapReduce
- **Apache DataFu**: collection of user-defined functions for Hadoop and Pig developed by LinkedIn
- **Apache Flink**: high-performance runtime, and automatic program optimization
- **Apache Gora**: framework for in-memory data model and persistence
- **Apache Hama**: BSP (Bulk Synchronous Parallel) computing framework
- **Apache MapReduce**: programming model for processing large data sets with a parallel, distributed algorithm on a cluster
- **Apache Pig**: high level language to express data analysis programs for Hadoop
- **Apache S4**: framework for stream processing, implementation of S4
- **Apache Spark**: framework for in-memory cluster computing
- **Apache Spark Streaming**: framework for stream processing, part of Spark
- **Apache Storm**: framework for stream processing by Twitter also on YARN
- **Apache Tez**: application framework for executing a complex DAG (directed acyclic graph) of tasks, built on YARN
- **Apache Twill**: abstraction over YARN that reduces the complexity of developing distributed applications
- **Cascalog**: data processing and querying library

- **Cheetah**: High Performance, Custom Data Warehouse on Top of MapReduce
- **Concurrent Cascading**: framework for data management/analytics on Hadoop
- **Damballa Parkour**: MapReduce library for Clojure
- **Datasalt Pangool**: alternative MapReduce paradigm
- **DataTorrent StrAM**: real-time engine is designed to enable distributed, asynchronous, real time in-memory big-data computations in as unblocked a way as possible, with minimal overhead and impact on performance
- **DistributedR**: scalable high-performance platform for the R language
- **Drools**: a Business Rules Management System (BRMS) solution
- **eBay Oink**: REST based interface for PIG execution
- **Esper**: a highly scalable, memory-efficient, in-memory computing, SQL-standard, minimal latency, real-time streaming-capable Big Data processing engine for historical data
- **Facebook Corona**: Hadoop enhancement which removes single point of failure
- **Facebook Peregrine**: Map Reduce framework
- **Facebook Scuba**: distributed in-memory datastore
- **GearPump**: a lightweight real-time big data streaming engine
- **Geotrellis**: geographic data processing engine for high performance applications
- **GetStream Stream Framework**: a Python library, which allows you to build newsfeed and notification systems using Cassandra and/or Redis
- **GIS Tools for Hadoop**: Big Data Spatial Analytics for the Hadoop Framework
- **Google Dataflow**: create data pipelines to help themæingest, transform and analyze data
- **Google MapReduce**: map reduce framework
- **Google MillWheel**: fault tolerant stream processing framework
- **GraphLab Dato**: fast, scalable engine of GraphLab Create, a Python library
- **Hazelcast**: In-Memory Data Grid
- **HParser**: data parsing transformation environment optimized for Hadoop
- **IBM Streams**: advanced analytic platform that allows user-developed applications to quickly ingest, analyze and correlate information as it arrives from thousands of real-time sources
- **JAQL**: declarative programming language for working with structured, semi-structured and unstructured data
- **Kite**: is a set of libraries, tools, examples, and documentation focused on making it easier to build systems on top of the Hadoop ecosystem
- **Kryo**: Java serialization and cloning: fast, efficient, automatic
- **LinkedIn Cubert**: a fast and efficient batch computation engine for complex analysis and reporting of massive datasets on Hadoop
- **Lipstick**: Pig workflow visualization tool
- **Metamarkers Druid**: framework for real-time analysis of large datasets
- **Microsoft Azure Stream Analytics**: an event processing engine that helps uncover real-time insights from devices, sensors, infrastructure, applications and data
- **Microsoft Orleans**: a straightforward approach to building distributed high-scale computing applications
- **Microsoft Trill**: a high-performance in-memory incremental analytics engine

- **Netflix Aegisthus**: Bulk Data Pipeline out of Cassandra. implements a reader for the SSTable format and provides a map/reduce program to create a compacted snapshot of the data contained in a column family
- **Netflix Lipstick**: Pig Visualization framework
- **Netflix Mantis**: Event Stream Processing System
- **Netflix PigPen**: map-reduce for Clojure whiche compiles to Apache Pig
- **Netflix STAASH**: language-agnostic as well as storage-agnostic web interface for storing data into persistent storage systems
- **Netflix Surus**: a collection of tools for analysis in Pig and Hive
- **Netflix Zeno**: Netflix's In-Memory Data Propagation Framework
- **Nextflow**: Dataflow oriented toolkit for parallel and distributed computational pipelines
- **Nokia Disco**: MapReduce framework developed by Nokia
- **Parsely Streamparse**: streamparse lets you run Python code against real-time streams of data. It also integrates Python smoothly with Apache Storm.
- **PigPen**: PigPen is map-reduce for Clojure, or distributed Clojure. It compiles to Apache Pig, but you don't need to know much about Pig to use it
- **Pinterest Pinlater**: asynchronous job execution system
- **Pubnub**: Data stream network
- **Pydoop**: Python MapReduce and HDFS API for Hadoop
- **ScaleOut hServer**: fast, scalable in-memory data grid for Hadoop
- **SeqPig**: Simple and scalable scripting for large sequencing data set(ex: bioinfomation) in Hadoop
- **SigmoidAnalytics Spork**: Pig on Apache Spark
- **spark-dataflow**: allows users to execute dataflow pipelines with Spark
- **SpatialHadoop**: SpatialHadoop is a MapReduce extension to Apache Hadoop designed specially to work with spatial data.
- **Spring for Apache Hadoop**: unified configuration model and easy to use APIs for using HDFS, MapReduce, Pig, and Hive
- **SQLStream Blaze**: stream processing platform
- **Stratio Streaming**: the union of a real-time messaging bus with a complex event processing engine using Spark Streaming
- **Stratosphere**: general purpose cluster computing framework
- **Streamdrill**: usefull for counting activities of event streams over different time windows and finding the most active one
- **Sumo Logic**: cloud based analyzer for machine-generated data.
- **Teradata QueryGrid**: data-access layer that can orchestrate multiple modes of analysis across multiple databases plus Hadoop
- **TIBCO ActiveSpaces**: in-memory data grid
- **Tigon**: a distributed framework built on Apache HadoopTM and Apache HBaseTM for real-time, high-throughput, low-latency data processing and analytics applications
- **Torch**: Scientific computing for LuaJIT
- **Trident**: a high-level abstraction for doing realtime computing on top of Storm
- **Twitter Scalding**: Scala library for Map Reduce jobs, built on Cascading

- **Twitter Summingbird**: Streaming MapReduce with Scalding and Storm, by Twitter
- **Twitter TSAR**: TimeSeries AggregatoR by Twitter

Distributed Filesystem

- **Apache HDFS**: a way to store large files across multiple machines
- **BeeGFS**: formerly FhGFS, parallel distributed file system
- **Ceph Filesystem**: software storage platform designed
- **Disco DDFS**: distributed filesystem
- **Facebook Haystack**: object storage system
- **Google Colossus**: distributed filesystem (GFS2)
- **Google GFS**: distributed filesystem
- **Google Megastore**: scalable, highly available storage
- **GridGain**: GGFS, Hadoop compliant in-memory file system
- **HDSF-DU**: HDFS-DU is an interactive visualization of the Hadoop distributed file system.
- **Lustre file system**: high-performance distributed filesystem
- **Netflix S3mper**: library that provides an additional layer of consistency checking on top of Amazon's S3 index through use of a consistent, secondary index
- **Quantcast File System QFS**: open-source distributed file system
- **Red Hat GlusterFS**: scale-out network-attached storage file system
- **Tachyon**: reliable file sharing at memory speed across cluster frameworks

Key-Map Data Model

- **Actian Vector**: column-oriented analytic database
- **Apache Accumulo**: distribuited key/value store, built on Hadoop
- **Apache Cassandra**: column-oriented distribuited datastore, inspired by BigTable
- **Apache HBase**: column-oriented distribuited datastore, inspired by BigTable
- **Facebook HydraBase**: evolution of HBase made by Facebook
- **Google BigTable**: column-oriented distributed datastore
- **Google Cloud Datastore**: is a fully managed, schemaless database for storing non-relational data over BigTable
- **Hypertable**: column-oriented distribuited datastore, inspired by BigTable
- **InfiniDB**: is accessed through a MySQL interface and use massive parallel processing to parallelize queries
- **MapR-DB**: fast, scalable, and enterprise-ready in-Hadoop database architected to manage big data
- **Netflix Priam**: Co-Process for backup/recovery, Token Management, and Centralized Configuration management for Cassandra
- **OhmData C5**: improved version of HBase
- **Sqrrl**: NoSQL databases on top of Apache Accumulo
- **Tephra**: Transactions for HBase
- **Twitter Manhattan**: real-time, multi-tenant distributed database for Twitter scale

Document Data Model

- **Actian Versant**: commercial object-oriented database management systems
- **Amazon SimpleDB**: a highly available and flexible non-relational data store that offloads the work of database administration
- **Clusterpoint**: a database software for high-speed storage and large-scale processing of XML and JSON data on clusters of commodity hardware
- **Crate Data**: is an open source massively scalable data store. It requires zero administration
- **Facebook Apollo**: Facebook's Paxos-like NoSQL database
- **jumboDB**: document oriented datastore over Hadoop
- **LinkedIn Espresso**: horizontally scalable document-oriented NoSQL data store
- **MarkLogic**: Schema-agnostic Enterprise NoSQL database technology
- **Microsoft DocumentDB**: fully-managed, highly-scalable, NoSQL document database service
- **MongoDB**: Document-oriented database system
- **RavenDB**: A transactional, open-source Document Database
- **RethinkDB**: document database that supports queries like table joins and group by
- **Terrastore**: a modern document store which provides advanced scalability and elasticity features without sacrificing consistency
- **TokuMX**: High-Performance MongoDB Distribution

Key-value Data Model

- **Aerospike**: NoSQL flash-optimized, in-memory. Open source and "Server code in 'C' (not Java or Erlang) precisely tuned to avoid context switching and memory copies.
- **Amazon DynamoDB**: distributed key/value store, implementation of Dynamo paper
- **Couchbase ForestDB**: Fast Key-Value Storage Engine Based on Hierarchical B+-Tree Trie
- **Edis**: is a protocol-compatible Server replacement for Redis
- **ElephantDB**: Distributed database specialized in exporting data from Hadoop
- **EventStore**: distributed time series database
- **Exasolution**: an in-memory, column-oriented, relational database management system
- **HyperDex**: next generation key-value store
- **KAI**: a distributed key-value datastore
- **LinkedIn Krati**: is a simple persistent data store with very low latency and high throughput
- **Linkedin Voldemort**: distributed key/value storage system
- **MemcacheDB**: a distributed key-value storage system designed for persistent
- **Netflix Dynomite**: thin Dynamo-based replication for cached data
- **Oracle NoSQL Database**: distributed key-value database by Oracle Corporation
- **RAMCloud**: storage system that provides large-scale low-latency storage by keeping all data in DRAM all the time and aggregating the main memories of thousands of servers
- **Redis**: in memory key value datastore
- **Redis Cluster**: distributed implementation of Redis
- **Redis Sentinel**: system designed to help managing Redis instances
- **Riak**: a decentralized datastore

- **Scalaris**: a distributed transactional key-value store
- **Storehaus**: library to work with asynchronous key value stores, by Twitter
- **Tarantool**: an efficient NoSQL database and a Lua application server
- **TreodeDB**: key-value store that's replicated and sharded and provides atomic multirow writes
- **Yahoo Sherpa**: hosted, distributed and geographically replicated key-valueÊcloud storage platform

Graph Data Model

- **Apache Giraph**: implementation of Pregel, based on Hadoop
- **Apache Spark Bagel**: implementation of Pregel, part of Spark
- **ArangoDB**: multi model distribuited database
- **Facebook TAO**: TAO is the distributed data store that is widely used at facebook to store and serve the social graph
- **Faunus**: Hadoop-based graph analytics engine for analyzing graphs represented across a multi-machine compute cluster
- **Google Cayley**: open-source graph database
- **Google Pregel**: graph processing framework
- **GraphLab PowerGraph**: a core C++ GraphLab API and a collection of high-performance machine learning and data mining toolkits built on top of the GraphLab API
- **GraphX**: resilient Distributed Graph System on Spark
- **Gremlin**: graph traversal Language
- **HyperGraphDB**: general purpose, open-source data storage mechanism based on a powerful knowledge management formalism known as directed hypergraphs
- **InfiniteGraph**: distributed graph database
- **Infovore**: RDF-centric Map/Reduce framework
- **Intel GraphBuilder**: tools to construct large-scale graphs on top of Hadoop
- **MapGraph**: Massively Parallel Graph processing on GPUs
- **Neo4j**: graph database writting entirely in Java
- **OrientDB**: document and graph database
- **Phoebus**: framework for large scale graph processing
- **Pinterest Zen**: Pinterest's Graph Storage Service
- **Sparksee**: scalable high-performance graph database
- **Stardog**: graph database: search, query, reasoning, and constraints in a lightweight, pure Java system
- **Titan**: distributed graph database, built over Cassandra
- **Twitter FlockDB**: distribuited graph database

NewSQL Databases

- **Actian Ingres**: commercially supported, open-source SQL relational database management system
- **BayesDB**: statistic oriented SQL database
- **Cockroach**: Scalable, Geo-Replicated, Transactional Datastore

- Datomic: distributed database designed to enable scalable, flexible and intelligent applications
- FoundationDB: distributed database, inspired by F1
- Google F1: distributed SQL database built on Spanner
- Google Spanner: globally distributed semi-relational database
- H-Store: is an experimental main-memory, parallel database management system that is optimized for on-line transaction processing (OLTP) applications
- HandlerSocket: NoSQL plugin for MySQL/MariaDB
- IBM DB2: object-relational database management system
- InfiniSQL: infinity scalable RDBMS
- MemSQL: in memory SQL database witho optimized columnar storage on flash
- NuoDB: SQL/ACID compliant distributed database
- Oracle Database: object-relational database management system
- Oracle TimesTen in-Memory Database: in-memory, relational database management system with persistence and recoverability
- Pivotal GemFire XD: Low-latency, in-memory, distributed SQL data store. Provides SQL interface to in-memory table data, persistable in HDFS
- SAP HANA: is an in-memory, column-oriented, relational database management system
- Segment SQL: Track your customer data to Amazon Redshift
- SenseiDB: distributed, realtime, semi-structured database
- Sky: database used for flexible, high performance analysis of behavioral data
- SymmetricDS: open source software for both file and database synchronization
- Teradata Database: complete relational database management system
- VoltDB: in-memory NewSQL database

Columnar Databases

- Amazon RedShift: data warehouse service, based on PostgreSQL
- C-Store: column oriented DBMS
- Google BigQuery: framework for interactive analysis, implementation of Dremel
- Google Dremel: framework for interactive analysis, implementation of Dremel
- MonetDB: column store database
- Parquet: columnar storage format for Hadoop
- Pivotal Greenplum: purpose-built, dedicated analytic data warehouse
- Vertica: is designed to manage large, fast-growing volumes of data and provide very fast query performance when used for data warehouses

Time-Series Databases

- Cube: uses MongoDB to store time series data
- Etsy StatsD: simple daemon for easy stats aggregation
- InfluxDB: distributed time series database
- Kairosdb: similar to OpenTSDB but allows for Cassandra
- OpenTSDB: distributed time series database on top of HBase
- Prometheus: an open-source service monitoring system and time series database

- **Square Cube**: system for collecting timestamped events and deriving metrics
- **TempoIQ**: Cloud-based sensor analytics

SQL-like processing

- **Actian SQL for Hadoop**: high performance interactive SQL access to all Hadoop data
- **AMPLAB Shark**: data warehouse system for Spark
- **Apache Drill**: framework for interactive analysis, inspired by Dremel
- **Apache HCatalog**: table and storage management layer for Hadoop
- **Apache Hive**: SQL-like data warehouse system for Hadoop
- **Apache Optiq**: framework that allows efficient translation of queries involving heterogeneous and federated data
- **Apache Phoenix**: SQL skin over HBase
- **BlinkDB**: massively parallel, approximate query engine
- **Cloudera Impala**: framework for interactive analysis, Inspired by Dremel
- **Concurrent Lingual**: SQL-like query language for Cascading
- **Datasalt Splout SQL**: full SQL query engine for big datasets
- **eBay Kylin**: Distributed Analytics Engine from eBay Inc. that provides SQL interface and multi-dimensional analysis (OLAP) on Hadoop supporting extremely large datasets
- **Facebook PrestoDB**: distributed SQL query engine
- **Hadapt**: a native implementation of SQL for the Apache Hadoop open-source project
- **JethroData**: index-based SQL engine for Hadoop
- **Metanautix Quest**: data compute engine
- **Pivotal HAWQ**: SQL-like data warehouse system for Hadoop
- **RainstorDB**: database for storing petabyte-scale volumes of structured and semi-structured data
- **Spark Catalyst**: is a Query Optimization Framework for Spark and Shark
- **SparkSQL**: Manipulating Structured Data Using Spark
- **Splice Machine**: a full-featured SQL-on-Hadoop RDBMS with ACID transactions
- **Stinger**: interactive query for Hive
- **Tajo**: distributed data warehouse system on Hadoop
- **Trafodion**: enterprise-class SQL-on-HBase solution targeting big data transactional or operational workloads

Integrated Development Environments

- **R-Studio**: IDE for R

Data Ingestion

- **Amazon Kinesis**: real-time processing of streaming data at massive scale
- **Apache BookKeeper**: a distributed logging service called BookKeeper and a distributed publish/subscribe system built on top of BookKeeper called Hedwig
- **Apache Chukwa**: data collection system
- **Apache Flume**: service to manage large amount of log data
- **Apache Samza**: stream processing framework, based on Kafla and YARN

- Apache Sqoop: tool to transfer data between Hadoop and a structured datastore
- Apache UIMA: Unstructured Information Management applications are software systems that analyze large volumes of unstructured information in order to discover knowledge that is relevant to an end user
- Cloudera Morphlines: framework that help ETL to Solr, HBase and HDFS
- Facebook Scribe: streamed log data aggregator
- Fluentd: tool to collect events and logs
- Google Photon: geographically distributed system for joining multiple continuously flowing streams of data in real-time with high scalability and low latency
- Heka: open source stream processing software system
- HIHO: framework for connecting disparate data sources with Hadoop
- LinkedIn Camus: Kafka to HDFS pipeline. It is a mapreduce job that does distributed data loads out of Kafka
- LinkedIn Databus: stream of change capture events for a database
- LinkedIn Gobblin: a framework for Solving Big Data Ingestion Problem
- LinkedIn Kamikaze: utility package for compressing sorted integer arrays
- Linkedin Lumos: bridge from OLTP to OLAP for use it on Hadoop
- LinkedIn White Elephant: log aggregator and dashboard
- Logstash: a tool for managing events and logs
- Netflix Suro: data pipeline service for collecting, aggregating, and dispatching large volume of application events including log data based on Chukwa
- Pinterest Secor: is a service implementing Kafka log persistance
- Record Breaker: Automatic structure for your text-formatted data
- TIBCO Enterprise Message Service: standards-based messaging middleware
- Twitter Zipkin: distributed tracing system that helps us gather timing data for all the disparate services at Twitter
- Vibe Data Stream: streaming data collection for real-time Big Data analytics

Message-oriented middleware

- ActiveMQ: open source messaging and Integration Patterns server
- Amazon Simple Queue Service: fast, reliable, scalable, fully managed queue service
- Apache Kafka: distributed publish-subscribe messaging system
- Apache Qpid: messaging tools that speak AMQP and support many languages and platforms
- Apollo: ActiveMQ’s next generation of messaging
- Beanstalkd: simple, fast work queue
- Bit.ly NSQ: realtime distributed message processing at scale
- Celery: Distributed Task Queue
- Crossroads I/O: library for building scalable and high performance distributed applications
- Darner: simple, lightweight message queue

- **Facebook Iris**: a totally ordered queue of messaging updates with separate pointers into the queue indicating the last update sent to your Messenger app and the traditional storage tier
- **Gearman**: Job Server
- **Google Cloud Pub/Sub**: reliable, many-to-many, asynchronous messaging hosted on Google's infrastructure
- **HornetQ**: open source project to build a multi-protocol, embeddable, very high performance, clustered, asynchronous messaging system
- **IronMQ**: easy-to-use highly available message queuing service
- **Kestrel**: distributed message queue system
- **Marconi**: queuing and notification service made by and for OpenStack, but not only for it
- **RabbitMQ**: Robust messaging for applications
- **RestMQ**: message queue which uses HTTP as transport, JSON to format a minimalist protocol and is organized as REST resources
- **RQ**: simple Python library for queueing jobs and processing them in the background with workers
- **Sidekiq**: Simple, efficient background processing for Ruby
- **ZeroMQ**: The Intelligent Transport Layer

Service Programming

- **Akka Toolkit**: runtime for distributed, and fault tolerant event-driven applications on the JVM
- **Apache Avro**: data serialization system
- **Apache Curator**: Java libaries for Apache ZooKeeper
- **Apache Karaf**: OSGi runtime that runs on top of any OSGi framework
- **Apache Thrift**: framework to build binary protocols
- **Apache Zookeeper**: centralized service for process management
- **Google Chubby**: a lock service for loosely-coupled distributed systems
- **Linkedin Norbert**: cluster manager
- **MPICH**: high performance and widely portable implementation of the Message Passing Interface (MPI) standard
- **OpenMPI**: message passing framework
- **Serf**: decentralized solution for service discovery and orchestration
- **Spotify Luigi**: a Python package for building complex pipelines of batch jobs. It handles dependency resolution, workflow management, visualization, handling failures, command line integration, and much more
- **Spring XD**: distributed and extensible system for data ingestion, real time analytics, batch processing, and data export
- **Twitter Elephant Bird**: libraries for working with LZOP-compressed data
- **Twitter Finagle**: asynchronous network stack for the JVM

Scheduling

- **Apache Aurora**: is a service scheduler that runs on top of Apache Mesos

- Apache Falcon: data management framework
- Apache Oozie: workflow job scheduler
- Chronos: distributed and fault-tolerant scheduler
- Linkedin Azkaban: batch workflow job scheduler
- Pinterest Pinball: customizable platform for creating workflow managers
- Sparrow: scheduling platform

Machine Learning

- Apache Mahout: machine learning library for Hadoop
- Ayasdi Core: tool for topological data analysis
- brain: Neural networks in JavaScript
- Cloudera Oryx: real-time large-scale machine learning
- Concurrent Pattern: machine learning library for Cascading
- convnetjs: Deep Learning in Javascript. Train Convolutional Neural Networks (or ordinary ones) in your browser
- cuDNN: GPU-accelerated library of primitives for deep neural networks
- Decider: Flexible and Extensible Machine Learning in Ruby
- etcML: text classification with machine learning
- Etsy Conjecture: scalable Machine Learning in Scalding
- fbcunn: Deep Learning CUDA Extensions from Facebook AI Research
- Google Sibyl: System for Large Scale Machine Learning at Google
- H2O: statistical, machine learning and math runtime for Hadoop
- IBM Watson: cognitive computing system
- LinkedIn ml-ease: ADMM based large scale logistic regression
- MLbase: distributed machine learning libraries for the BDAS stack
- MLPNeuralNet: Fast multilayer perceptron neural network library for iOS and Mac OS X
- nupic: Numenta Platform for Intelligent Computing: a brain-inspired machine intelligence platform, and biologically accurate neural network based on cortical learning algorithms
- PredictionIO: machine learning server buit on Hadoop, Mahout and Cascading
- scikit-learn: scikit-learn: machine learning in Python
- Spark MLlib: a Spark implementation of some common machine learning (ML) functionality
- Sparkling Water: combine H2OÕs Machine Learning capabilities with the power of the Spark platform
- Theano: Python package for deep learning that can utilize NVIDIA's CUDA toolkit to run on the GPU
- Thunder: Large-scale analysis of neural data
- Vahara: Machine learning and natural language processing with Apache Pig
- Viv: global platform that enables developers to plug into and create an intelligent, conversational interface to anything
- Vowpal Wabbit: learning system sponsored by Microsoft and Yahoo!
- WEKA: suite of machine learning software
- Wit: Natural Language for the Internet of Things

- **Wolfram Alpha**: computational knowledge engine
- **YHat ScienceOps**: platform for deploying, managing, and scaling predictive models in production applications

Benchmarking

- **Apache Hadoop Benchmarking**: micro-benchmarks for testing Hadoop performances
- **Berkeley SWIM Benchmark**: real-world big data workload benchmark
- **Big-Bench**: Big Bench Workload Development
- **Hive-benchmarks**: some benchmarking queries for Apache Hive
- **Hive-testbench**: Testbench for experimenting with Apache Hive at any data scale.
- **Intel HiBench**: a Hadoop benchmark suite
- **Mesosaurus**: Mesos task load simulator framework for (cluster and Mesos) performance analysis
- **Netflix Inviso**: performance focused Big Data tool
- **PUMA Benchmarking**: benchmark suite for MapReduce applications
- **Yahoo Gridmix3**: Hadoop cluster benchmarking from Yahoo engineer team

Security

- **Apache Knox Gateway**: single point of secure access for Hadoop clusters
- **Apache Ranger**: framework to enable, monitor and manage comprehensive data security across the Hadoop platform (formerly called Apache Argus)
- **Apache Sentry**: security module for data stored in Hadoop
- **PacketPig**: Open Source Big Data Security Analytics
- **Voltage SecureData**: data protection framework

System Deployment

- **Ankush**: A big data cluster management tool that creates and manages clusters of different technologies.
- **Apache Ambari**: operational framework for Hadoop mangement
- **Apache Bigtop**: system deployment framework for the Hadoop ecosystem
- **Apache Helix**: cluster management framework
- **Apache Mesos**: cluster manager
- **Apache Slider**: is a YARN application to deploy existing distributed applications on YARN
- **Apache Whirr**: set of libraries for running cloud services
- **Apache YARN**: Cluster manager
- **Brooklyn**: library that simplifies application deployment and management
- **Buildoop**: Similar to Apache BigTop based on Groovy language
- **Cloudera Director**: a comprehensive data management platform with the flexibility and power to evolve with your business
- **Cloudera HUE**: web application for interacting with Hadoop
- **Deimos**: Mesos containerizer hooks for Docker
- **Developp**: tool for provisioning, managing and monitoring Apache Hadoop
- **Etsy Sahale**: Visualizing Cascading Workflows at Etsy

- **Facebook Autoscale**: the load balancer will concentrate workload to a server until it has at least a medium-level workload
- **Facebook Prism**: multi datacenters replication system
- **Ganglia Monitoring System**: scalable distributed monitoring system for high-performance computing systems such as clusters and Grids
- **Genie**: Genie provides REST-ful APIs to run Hadoop, Hive and Pig jobs, and to manage multiple Hadoop resources and perform job submissions across them.
- **Google Borg**: job scheduling and monitoring system
- **Google Omega**: job scheduling and monitoring system
- **Hannibal**: Hannibal is tool to help monitor and maintain HBase-Clusters that are configured for manual splitting.
- **Hortonworks HOYA**: application that can deploy HBase cluster on YARN
- **Jumbune**: Jumbune is an open-source product built for analyzing Hadoop cluster and MapReduce jobs.
- **Marathon**: Mesos framework for long-running services
- **Minotaur**: scripts/recipes/configs to spin up VPC-based infrastructure in AWS from scratch and deploy labs to it
- **Myriad**: a mesos framework designed for scaling YARN clusters on Mesos. Myriad can expand or shrink one or more YARN clusters in response to events as per configured rules and policies.
- **Neflix SimianArmy**: a suite of tools for keeping your cloud operating in top form
- **Tumblr Collins**: Infrastructure management for engineers
- **Tumblr Genesis**: a tool for data center automation

Container Manager

- **Amazon EC2 Container Service**: a highly scalable, high performance container management service that supports Docker containers
- **Docker**: an open platform for developers and sysadmins to build, ship, and run distributed applications
- **Fig**: fast, isolated development environments using Docker
- **Google Container Engine**: Run Docker containers on Google Cloud Platform, powered by Kubernetes
- **Kubernetes**: open source implementation of container cluster management
- **Rocket**: an alternative to the Docker runtime, designed for server environments with the most rigorous security and production requirements

Applications

- **Adobe Spindle**: Next-generation web analytics processing with Scala, Spark, and Parquet
- **Apache Kiji**: framework to collect and analyze data in real-time, based on HBase
- **Apache Nutch**: open source web crawler
- **Apache OODT**: capturing, processing and sharing of data for NASA's scientific archives
- **Apache Tika**: content analysis toolkit
- **Domino**: Run, scale, share, and deploy models Ñ without any infrastructure.

- **Eclipse BIRT**: Eclipse-based reporting system
- **Eventhub**: open source event analytics platform
- **HIPI Library**: API for performing image processing tasks on Hadoop's MapReduce
- **Hunk**: Splunk analytics for Hadoop
- **MADlib**: data-processing library of an RDBMS to analyze data
- **PivotalR**: R on Pivotal HD / HAWQ and PostgreSQL
- **Qubole**: auto-scaling Hadoop cluster, built-in data connectors
- **Sense**: Cloud Platform for Data Science and Big Data Analytics
- **Snowplow**: enterprise-strength web and event analytics, powered by Hadoop, Kinesis, Redshift and Postgres
- **SparkR**: R frontend for Spark
- **Splunk**: analyzer for machine-generated date
- **Talend**: unified open source environment for YARN, Hadoop, HBASE, Hive, HCatalog & Pig

Search engine and framework

- **Apache Blur**: a search engine capable of querying massive amounts of structured data at incredible speeds
- **Apache Lucene**: Search engine library
- **Apache Solr**: Search platform for Apache Lucene
- **ElasticSearch**: Search and analytics engine based on Apache Lucene
- **Elasticsearch Hadoop**: Elasticsearch real-time search and analytics natively integrated with Hadoop. Supports Map/Reduce, Cascading, Apache Hive and Apache Pig.
- **Enigma.io**: Freemium robust web application for exploring, filtering, analyzing, searching and exporting massive datasets scraped from across the Web
- **Facebook Unicorn**: social graph search platform
- **Google Caffeine**: continuous indexing system
- **Google Percolator**: continuous indexing system
- **TeraGoogle**: large search index
- **Haeinsa**: linearly scalable multi-row, multi-table transaction library for HBase based on Percolator
- **HBase Coprocessor**: implementation of Percolator, part of HBase
- **hIndex**: Secondary Index for HBase
- **SF1R Search Engine**: distributed search engine written in c++
- **Lily HBase Indexer**: quickly and easily search for any content stored in HBase
- **LinkedIn Bobo**: is a Faceted Search implementation written purely in Java, an extension to Apache Lucene
- **LinkedIn Cleo**: is a flexible software library for enabling rapid development of partial, out-of-order and real-time typeahead search
- **LinkedIn Galene**: search architecture at LinkedIn
- **LinkedIn Zoie**: is a realtime search/indexing system written in Java
- **Sphnix Search Server**: fulltext search engine

MySQL forks and evolutions

- **Amazon Aurora**: a MySQL-compatible, relational database engine that combines the speed and availability of high-end commercial databases with the simplicity and cost-effectiveness of open source databases
- **Amazon RDS**: MySQL databases in Amazon's cloud
- **Drizzle**: evolution of MySQL 6.0
- **Google Cloud SQL**: MySQL databases in Google's cloud
- **HiveDB**: an open source framework for horizontally partitioning MySQL systems
- **MariaDB**: enhanced, drop-in replacement for MySQL
- **MariaDB Galera**: a synchronous multi-master cluster for MariaDB
- **MySQL Cluster**: MySQL implementation using NDB Cluster storage engine providing shared-nothing clustering and auto-sharding
- **Percona Server**: enhanced, drop-in replacement for MySQL
- **ProxySQL**: High Performance Proxy for MySQL
- **TokuDB**: TokuDB is a storage engine for MySQL and MariaDB
- **WebScaleSQL**: is a collaboration among engineers from several companies that face similar challenges in running MySQL at scale
- **Youtube Vitess**: provides servers and tools which facilitate scaling of MySQL databases for large scale web services

PostgreSQL forks and evolutions

- **HadoopDB**: hybrid of MapReduce and DBMS
- **IBM Netezza**: high-performance data warehouse appliances
- **Postgres-XL**: Scalable Open Source PostgreSQL-based Database Cluster
- **RecDB**: Open Source Recommendation Engine Built Entirely Inside PostgreSQL
- **Stado**: open source MPP database system solely targeted at data warehousing and data mart applications
- **Yahoo Everest**: multi-peta-byte database / MPP derived by PostgreSQL

Memcached forks and evolutions

- **Box Tron**: proxy to memcached servers
- **Facebook McDipper**: key/value cache for flash storage
- **Facebook Mcrouter**: a memcached protocol router for scaling memcached deployments
- **Facebook Memcached**: fork of Memcache
- **Twemproxy**: A fast, light-weight proxy for memcached and redis
- **Twitter Fatcache**: key/value cache for flash storage
- **Twitter Twemcache**: fork of Memcache

Embedded Databases

- **Actian PSQL**: ACID-compliant DBMS developed by Pervasive Software, optimized for embedding in applications
- **BerkeleyDB**: a software library that provides a high-performance embedded database for key/value data

- **eXtreme DB**: in-memory database combines exceptional performance, reliability and developer efficiency in a proven real-time embedded database engine
- **FairCom c-treeACE**: a cross-platform database engine
- **Google Firebase**: a powerful API to store and sync data in realtime
- **HamsterDB**: transactional key-value database
- **HanoiDB**: Erlang LSM BTree Storage
- **LevelDB**: a fast key-value storage library written at Google that provides an ordered mapping from string keys to string values
- **LMDB**: ultra-fast, ultra-compact key-value embedded data store developed by Symas
- **RocksDB**: embeddable persistent key-value store for fast storage based on LevelDB
- **TokioCabinet**: a library of routines for managing a database

Business Intelligence

- **ActivePivot**: Java In-Memory OLAP cube stored in columns, with clearly decoupled pre/post processing
- **Adatao**: business intelligence and data science platform
- **Apama analytics**: platform for streaming analytics and intelligent automated action
- **Atigeo xPatterns**: data analytics platform
- **BIME Analytics**: business intelligence platform in the cloud
- **Chartio**: lean business intelligence platform to visualize and explore your data
- **Datapine**: self-service business intelligence tool in the cloud
- **Jaspersoft**: powerful business intelligence suite
- **Jedox Palo**: customisable Business Intelligence platform
- **Lavastorm Analytics**: used for audit analytics, revenue assurance, fraud management, and customer experience management
- **LinkedIn GoSpeed**: provides RUM data processing, visualization, monitoring, and analyses data daily, hourly, or on a near real-time basis
- **Map-D**: GPU in-memory database, big data analysis and visualization platform
- **Microsoft**: business intelligence software and platform
- **Microstrategy**: software platforms for business intelligence, mobile intelligence, and network applications
- **Pentaho**: business intelligence platform
- **Qlik**: business intelligence and analytics platform
- **SpagoBI**: open source business intelligence platform
- **Spotfire**: business intelligence platform
- **Tableau**: business intelligence platform
- **Teradata Aster**: Big Data Analytics
- **Tessera**: Environment for Deep Analysis of Large Complex Data
- **Zeppelin**: open source data analysis environment on top of Hadoop.
- **Zoomdata**: Big Data Analytics

Data Analysis

- **LinkedIn Pinot**: a distributed system that supports columnar indexes with the ability to add new types of indexes
- **Myria**: scalable Analytics-as-a-Service platform based on relational algebra
- **Pinalytics**: Pinterestâ��s data analytics engine
- **Zillabyte**: an API for distributed data computation. Scale with your data.

Data Warehouse

- **Google Mesa**: highly scalable analytic data warehousing system
- **IBM BigInsights**: data processing, warehousing and analytics
- **IBM dashDB**: Data Warehousing and Analysis Needs, all in the Cloud
- **Microsoft Cosmos**: Microsoft's internal BigData analysis platform

Data Visualization

- **Arbor**: graph visualization library using web workers and jQuery
- **C3**: D3-based reusable chart library
- **CartoDB**: open-source or freemium hosting for geospatial databases with powerful front-end editing capabilities and a robust API
- **Chart.js**: open source HTML5 Charts visualizations
- **Chartist.js**: another open source HTML5 Charts visualization
- **Crossfilter**: avaScript library for exploring large multivariate datasets in the browser. Works well with dc.js and d3.js
- **Cubism**: JavaScript library for time series visualization
- **Cytoscape**: JavaScript library for visualizing complex networks
- **D3**: javaScript library for manipulating documents
- **DC.js**: Dimensional charting built to work natively with crossfilter rendered using d3.js. Excellent for connecting charts/additional metadata to hover events in D3
- **Envisionjs**: dynamic HTML5 visualization
- **FnordMetric ChartSQL**: allows you to write SQL queries that return charts instead of tables. The charts are rendered as SVG vector graphics.
- **Freeboard**: open source real-time dashboard builder for IOT and other web mashups
- **Gephi**: An award-winning open-source platform for visualizing and manipulating large graphs and network connections
- **Google Charts**: simple charting API
- **Grafana**: open source, feature rich metrics dashboard and graph editor for Graphite, InfluxDB & OpenTSDB
- **Graphite**: scalable Realtime Graphing
- **Highcharts**: simple and flexible charting API
- **IPython**: provides a rich architecture for interactive computing
- **Keylines**: toolkit for visualizing the networks in your data
- **Kibana**: visualize logs and time-stamped data
- **Matplotlib**: plotting with Python
- **NVD3**: chart components for d3.js
- **Peity**: Progressive SVG bar, line and pie charts

- **Plot.ly**: Easy-to-use web service that allows for rapid creation of complex charts, from heatmaps to histograms. Upload data to create and style charts with Plotly's online spreadsheet. Fork others' plots.
- **Recline**: simple but powerful library for building data applications in pure Javascript and HTML
- **Redash**: open-source platform to query and visualize data
- **Sigma.js**: JavaScript library dedicated to graph drawing
- **Square Cubism.js**: a plugin for visualizing time series. Use Cubism to construct better realtime dashboards
- **Vega**: a visualization grammar

Internet of Things

- **2lemetry**: Platform for Internet of things
- **Evrything**: Making products smart
- **ThingWorx**: Rapid development and connection of intelligent systems