[Fork Me on GitHub](https://github.com/hadoopecosystemtable/hadoopecosystemtable.github.io)

**The Hadoop Ecosystem Table**

**This page is a summary to keep the track of Hadoop related projects, focused on FLOSS environment.**

|  |  |  |
| --- | --- | --- |
| **Distributed Filesystem** | | |
| Apache HDFS | The Hadoop Distributed File System (HDFS) offers a way to store large files across multiple machines. Hadoop and HDFS was derived from Google File System (GFS) paper. Prior to Hadoop 2.0.0, the NameNode was a single point of failure (SPOF) in an HDFS cluster. With Zookeeper the HDFS High Availability feature addresses this problem by providing the option of running two redundant NameNodes in the same cluster in an Active/Passive configuration with a hot standby. | [1. hadoop.apache.org](http://hadoop.apache.org/)  [2. Google FileSystem - GFS Paper](http://research.google.com/archive/gfs.html)  [3. Cloudera Why HDFS](http://blog.cloudera.com/blog/2012/07/why-we-build-our-platform-on-hdfs/)  [4. Hortonworks Why HDFS](http://hortonworks.com/blog/thinking-about-the-hdfs-vs-other-storage-technologies/) |
| Red Hat GlusterFS | GlusterFS is a scale-out network-attached storage file system. GlusterFS was developed originally by Gluster, Inc., then by Red Hat, Inc., after their purchase of Gluster in 2011. In June 2012, Red Hat Storage Server was announced as a commercially-supported integration of GlusterFS with Red Hat Enterprise Linux. Gluster File System, known now as Red Hat Storage Server. | [1. www.gluster.org](http://www.gluster.org/)  [2. Red Hat Hadoop Plugin](http://www.redhat.com/about/news/archive/2013/10/red-hat-contributes-apache-hadoop-plug-in-to-the-gluster-community) |
| Quantcast File System QFS | QFS is an open-source distributed file system software package for large-scale MapReduce or other batch-processing workloads. It was designed as an alternative to Apache Hadoop’s HDFS, intended to deliver better performance and cost-efficiency for large-scale processing clusters. It is written in C++ and has fixed-footprint memory management. QFS uses Reed-Solomon error correction as method for assuring reliable access to data. Reed–Solomon coding is very widely used in mass storage systems to correct the burst errors associated with media defects. Rather than storing three full versions of each file like HDFS, resulting in the need for three times more storage, QFS only needs 1.5x the raw capacity because it stripes data across nine different disk drives. | [1. QFS site](https://www.quantcast.com/engineering/qfs/)  [2. GitHub QFS](https://github.com/quantcast/qfs)  [3. HADOOP-8885](https://issues.apache.org/jira/browse/HADOOP-8885) |
| Ceph Filesystem | Ceph is a free software storage platform designed to present object, block, and file storage from a single distributed computer cluster. Ceph's main goals are to be completely distributed without a single point of failure, scalable to the exabyte level, and freely-available. The data is replicated, making it fault tolerant. | [1. Ceph Filesystem site](http://ceph.com/ceph-storage/file-system/)  [2. Ceph and Hadoop](http://ceph.com/docs/next/cephfs/hadoop/)  [3. HADOOP-6253](https://issues.apache.org/jira/browse/HADOOP-6253) |
| Lustre file system | The Lustre filesystem is a high-performance distributed filesystem intended for larger network and high-availability environments. Traditionally, Lustre is configured to manage remote data storage disk devices within a Storage Area Network (SAN), which is two or more remotely attached disk devices communicating via a Small Computer System Interface (SCSI) protocol. This includes Fibre Channel, Fibre Channel over Ethernet (FCoE), Serial Attached SCSI (SAS) and even iSCSI. With Hadoop HDFS the software needs a dedicated cluster of computers on which to run. But folks who run high performance computing clusters for other purposes often don't run HDFS, which leaves them with a bunch of computing power, tasks that could almost certainly benefit from a bit of map reduce and no way to put that power to work running Hadoop. Intel's noticed this and, in version 2.5 of its Hadoop distribution that it quietly released last week, has added support for Lustre: the Intel® HPC Distribution for Apache Hadoop\* Software, a new product that combines Intel Distribution for Apache Hadoop software with Intel® Enterprise Edition for Lustre software. This is the only distribution of Apache Hadoop that is integrated with Lustre, the parallel file system used by many of the world's fastest supercomputers | [1. wiki.lustre.org/](http://wiki.lustre.org/)  [2. Hadoop with Lustre](http://wiki.lustre.org/index.php/Running_Hadoop_with_Lustre)  [3. Intel HPC Hadoop](http://hadoop.intel.com/products/distribution) |
| Alluxio | Alluxio, the world’s first memory-centric virtual distributed storage system, unifies data access and bridges computation frameworks and underlying storage systems. Applications only need to connect with Alluxio to access data stored in any underlying storage systems. Additionally, Alluxio’s memory-centric architecture enables data access orders of magnitude faster than existing solutions.  In big data ecosystem, Alluxio lies between computation frameworks or jobs, such as Apache Spark, Apache MapReduce, or Apache Flink, and various kinds of storage systems, such as Amazon S3, OpenStack Swift, GlusterFS, HDFS, Ceph, or OSS. Alluxio brings significant performance improvement to the stack; for example, Baidu uses Alluxio to improve their data analytics performance by 30 times. Beyond performance, Alluxio bridges new workloads with data stored in traditional storage systems. Users can run Alluxio using its standalone cluster mode, for example on Amazon EC2, or launch Alluxio with Apache Mesos or Apache Yarn.  Alluxio is Hadoop compatible. This means that existing Spark and MapReduce programs can run on top of Alluxio without any code changes. The project is open source (Apache License 2.0) and is deployed at multiple companies. It is one of the fastest growing open source projects. With less than three years open source history, Alluxio has attracted more than 160 contributors from over 50 institutions, including Alibaba, Alluxio, Baidu, CMU, IBM, Intel, NJU, Red Hat, UC Berkeley, and Yahoo. The project is the storage layer of the Berkeley Data Analytics Stack (BDAS) and also part of the Fedora distribution. | [1. Alluxio site](http://www.alluxio.org/) |
| GridGain | GridGain is open source project licensed under Apache 2.0. One of the main pieces of this platform is the In-Memory Apache Hadoop Accelerator which aims to accelerate HDFS and Map/Reduce by bringing both, data and computations into memory. This work is done with the GGFS - Hadoop compliant in-memory file system. For I/O intensive jobs GridGain GGFS offers performance close to 100x faster than standard HDFS. Paraphrasing Dmitriy Setrakyan from GridGain Systems talking about GGFS regarding Tachyon:   * GGFS allows read-through and write-through to/from underlying HDFS or any other Hadoop compliant file system with zero code change. Essentially GGFS entirely removes ETL step from integration. * GGFS has ability to pick and choose what folders stay in memory, what folders stay on disc, and what folders get synchronized with underlying (HD)FS either synchronously or asynchronously. * GridGain is working on adding native MapReduce component which will provide native complete Hadoop integration without changes in API, like Spark currently forces you to do. Essentially GridGain MR+GGFS will allow to bring Hadoop completely or partially in-memory in Plug-n-Play fashion without any API changes. | [1. GridGain site](http://www.gridgain.org/) |
| XtreemFS | XtreemFS is a general purpose storage system and covers most storage needs in a single deployment. It is open-source, requires no special hardware or kernel modules, and can be mounted on Linux, Windows and OS X. XtreemFS runs distributed and offers resilience through replication. XtreemFS Volumes can be accessed through a FUSE component, that offers normal file interaction with POSIX like semantics. Furthermore an implementation of Hadoops FileSystem interface is included which makes XtreemFS available for use with Hadoop, Flink and Spark out of the box. XtreemFS is licensed under the New BSD license. The XtreemFS project is developed by Zuse Institute Berlin. The development of the project is funded by the European Commission since 2006 under Grant Agreements No. FP6-033576, FP7-ICT-257438, and FP7-318521, as well as the German projects MoSGrid, "First We Take Berlin", FFMK, GeoMultiSens, and BBDC. | [1. XtreemFS site](http://www.xtreemfs.org/) [2. Flink on XtreemFS](https://github.com/xtreemfs/xtreemfs/wiki/Apache-Flink-with-XtreemFS) [. Spark XtreemFS](https://github.com/xtreemfs/xtreemfs/wiki/Apache-Spark-with-XtreemFS) |
| **Distributed Programming** | | |
| Apache MapReduce | MapReduce is a programming model for processing large data sets with a parallel, distributed algorithm on a cluster. Apache MapReduce was derived from Google MapReduce: Simplified Data Processing on Large Clusters paper. The current Apache MapReduce version is built over Apache YARN Framework. YARN stands for “Yet-Another-Resource-Negotiator”. It is a new framework that facilitates writing arbitrary distributed processing frameworks and applications. YARN’s execution model is more generic than the earlier MapReduce implementation. YARN can run applications that do not follow the MapReduce model, unlike the original Apache Hadoop MapReduce (also called MR1). Hadoop YARN is an attempt to take Apache Hadoop beyond MapReduce for data-processing. | [1. Apache MapReduce](http://wiki.apache.org/hadoop/MapReduce/)  [2. Google MapReduce paper](http://research.google.com/archive/mapreduce.html)  [3. Writing YARN applications](http://hadoop.apache.org/docs/stable/hadoop-yarn/hadoop-yarn-site/WritingYarnApplications.html) |
| Apache Pig | Pig provides an engine for executing data flows in parallel on Hadoop. It includes a language, Pig Latin, for expressing these data flows. Pig Latin includes operators for many of the traditional data operations (join, sort, filter, etc.), as well as the ability for users to develop their own functions for reading, processing, and writing data. Pig runs on Hadoop. It makes use of both the Hadoop Distributed File System, HDFS, and Hadoop’s processing system, MapReduce. Pig uses MapReduce to execute all of its data processing. It compiles the Pig Latin scripts that users write into a series of one or more MapReduce jobs that it then executes. Pig Latin looks different from many of the programming languages you have seen. There are no if statements or for loops in Pig Latin. This is because traditional procedural and object-oriented programming languages describe control flow, and data flow is a side effect of the program. Pig Latin instead focuses on data flow. | [1. pig.apache.org/](https://pig.apache.org/)  [2.Pig examples by Alan Gates](https://github.com/alanfgates/programmingpig) |
| JAQL | JAQL is a functional, declarative programming language designed especially for working with large volumes of structured, semi-structured and unstructured data. As its name implies, a primary use of JAQL is to handle data stored as JSON documents, but JAQL can work on various types of data. For example, it can support XML, comma-separated values (CSV) data and flat files. A "SQL within JAQL" capability lets programmers work with structured SQL data while employing a JSON data model that's less restrictive than its Structured Query Language counterparts. Specifically, Jaql allows you to select, join, group, and filter data that is stored in HDFS, much like a blend of Pig and Hive. Jaql’s query language was inspired by many programming and query languages, including Lisp, SQL, XQuery, and Pig.  JAQL was created by workers at IBM Research Labs in 2008 and released to open source. While it continues to be hosted as a project on Google Code, where a downloadable version is available under an Apache 2.0 license, the major development activity around JAQL has remained centered at IBM. The company offers the query language as part of the tools suite associated with InfoSphere BigInsights, its Hadoop platform. Working together with a workflow orchestrator, JAQL is used in BigInsights to exchange data between storage, processing and analytics jobs. It also provides links to external data and services, including relational databases and machine learning data. | [1. JAQL in Google Code](https://code.google.com/p/jaql/)  [2. What is Jaql? by IBM](http://www-01.ibm.com/software/data/infosphere/hadoop/jaql/) |
| Apache Spark | Data analytics cluster computing framework originally developed in the AMPLab at UC Berkeley. Spark fits into the Hadoop open-source community, building on top of the Hadoop Distributed File System (HDFS). However, Spark provides an easier to use alternative to Hadoop MapReduce and offers performance up to 10 times faster than previous generation systems like Hadoop MapReduce for certain applications. Spark is a framework for writing fast, distributed programs. Spark solves similar problems as Hadoop MapReduce does but with a fast in-memory approach and a clean functional style API. With its ability to integrate with Hadoop and inbuilt tools for interactive query analysis (Shark), large-scale graph processing and analysis (Bagel), and real-time analysis (Spark Streaming), it can be interactively used to quickly process and query big data sets. To make programming faster, Spark provides clean, concise APIs in Scala, Java and Python. You can also use Spark interactively from the Scala and Python shells to rapidly query big datasets. Spark is also the engine behind Shark, a fully Apache Hive-compatible data warehousing system that can run 100x faster than Hive. | [1. Apache Spark](http://spark.apache.org/)  [2. Mirror of Spark on Github](https://github.com/apache/spark)  [3. RDDs - Paper](http://www.cs.berkeley.edu/%7Ematei/papers/2012/nsdi_spark.pdf)  [4. Spark: Cluster Computing... - Paper](https://people.csail.mit.edu/matei/papers/2010/hotcloud_spark.pdf)  [Spark Research](http://spark.apache.org/research.html) |
| Apache Storm | Storm is a complex event processor (CEP) and distributed computation framework written predominantly in the Clojure programming language. Is a distributed real-time computation system for processing fast, large streams of data. Storm is an architecture based on master-workers paradigma. So a Storm cluster mainly consists of a master and worker nodes, with coordination done by Zookeeper.  Storm makes use of zeromq (0mq, zeromq), an advanced, embeddable networking library. It provides a message queue, but unlike message-oriented middleware (MOM), a 0MQ system can run without a dedicated message broker. The library is designed to have a familiar socket-style API. Originally created by Nathan Marz and team at BackType, the project was open sourced after being acquired by Twitter. Storm was initially developed and deployed at BackType in 2011. After 7 months of development BackType was acquired by Twitter in July 2011. Storm was open sourced in September 2011.  Hortonworks is developing a Storm-on-YARN version and plans finish the base-level integration in 2013 Q4. This is the plan from Hortonworks. Yahoo/Hortonworks also plans to move Storm-on-YARN code from github.com/yahoo/storm-yarn to be a subproject of Apache Storm project in the near future. Twitter has recently released a Hadoop-Storm Hybrid called “Summingbird.” Summingbird fuses the two frameworks into one, allowing for developers to use Storm for short-term processing and Hadoop for deep data dives,. a system that aims to mitigate the tradeoffs between batch processing and stream processing by combining them into a hybrid system. | [1. Storm Project/](http://storm-project.net/)  [2. Storm-on-YARN](https://hadoopecosystemtable.github.io/github.com/yahoo/storm-yarn) |
| Apache Flink | Apache Flink (formerly called Stratosphere) features powerful programming abstractions in Java and Scala, a high-performance runtime, and automatic program optimization. It has native support for iterations, incremental iterations, and programs consisting of large DAGs of operations. Flink is a data processing system and an alternative to Hadoop's MapReduce component. It comes with its own runtime, rather than building on top of MapReduce. As such, it can work completely independently of the Hadoop ecosystem. However, Flink can also access Hadoop's distributed file system (HDFS) to read and write data, and Hadoop's next-generation resource manager (YARN) to provision cluster resources. Since most Flink users are using Hadoop HDFS to store their data, it ships already the required libraries to access HDFS. | [1. Apache Flink incubator page](http://flink.incubator.apache.org/)  [2. Stratosphere site](http://stratosphere.eu/) |
| Apache Apex | Apache Apex is an enterprise grade Apache YARN based big data-in-motion platform that unifies stream processing as well as batch processing. It processes big data in-motion in a highly scalable, highly performant, fault tolerant, stateful, secure, distributed, and an easily operable way. It provides a simple API that enables users to write or re-use generic Java code, thereby lowering the expertise needed to write big data applications.  The Apache Apex platform is supplemented by Apache Apex-Malhar, which is a library of operators that implement common business logic functions needed by customers who want to quickly develop applications. These operators provide access to HDFS, S3, NFS, FTP, and other file systems; Kafka, ActiveMQ, RabbitMQ, JMS, and other message systems; MySql, Cassandra, MongoDB, Redis, HBase, CouchDB and other databases along with JDBC connectors. The library also includes a host of other common business logic patterns that help users to significantly reduce the time it takes to go into production. Ease of integration with all other big data technologies is one of the primary missions of Apache Apex-Malhar.  Apex, available on GitHub, is the core technology upon which DataTorrent's commercial offering, DataTorrent RTS 3, along with other technology such as a data ingestion tool called dtIngest, are based. | [1. Apache Apex from DataTorrent](https://www.datatorrent.com/apex/)  [2. Apache Apex main page](http://apex.incubator.apache.org/)  [3. Apache Apex Proposal](https://wiki.apache.org/incubator/ApexProposal) |
| Netflix PigPen | PigPen is map-reduce for Clojure which compiles to Apache Pig. Clojure is dialect of the Lisp programming language created by Rich Hickey, so is a functional general-purpose language, and runs on the Java Virtual Machine, Common Language Runtime, and JavaScript engines. In PigPen there are no special user defined functions (UDFs). Define Clojure functions, anonymously or named, and use them like you would in any Clojure program. This tool is open sourced by Netflix, Inc. the American provider of on-demand Internet streaming media. | [1. PigPen on GitHub](https://github.com/Netflix/PigPen) |
| AMPLab SIMR | Apache Spark was developed thinking in Apache YARN. However, up to now, it has been relatively hard to run Apache Spark on Hadoop MapReduce v1 clusters, i.e. clusters that do not have YARN installed. Typically, users would have to get permission to install Spark/Scala on some subset of the machines, a process that could be time consuming. SIMR allows anyone with access to a Hadoop MapReduce v1 cluster to run Spark out of the box. A user can run Spark directly on top of Hadoop MapReduce v1 without any administrative rights, and without having Spark or Scala installed on any of the nodes. | [1. SIMR on GitHub](http://databricks.github.io/simr/) |
| Facebook Corona | “The next version of Map-Reduce" from Facebook, based in own fork of Hadoop. The current Hadoop implementation of the MapReduce technique uses a single job tracker, which causes scaling issues for very large data sets. The Apache Hadoop developers have been creating their own next-generation MapReduce, called YARN, which Facebook engineers looked at but discounted because of the highly-customised nature of the company's deployment of Hadoop and HDFS. Corona, like YARN, spawns multiple job trackers (one for each job, in Corona's case). | [1. Corona on Github](https://github.com/facebookarchive/hadoop-20/tree/master/src/contrib/corona) |
| Apache Twill | Twill is an abstraction over Apache Hadoop® YARN that reduces the complexity of developing distributed applications, allowing developers to focus more on their business logic. Twill uses a simple thread-based model that Java programmers will find familiar. YARN can be viewed as a compute fabric of a cluster, which means YARN applications like Twill will run on any Hadoop 2 cluster. YARN is an open source application that allows the Hadoop cluster to turn into a collection of virtual machines. Weave, developed by Continuuity and initially housed on Github, is a complementary open source application that uses a programming model similar to Java threads, making it easy to write distributed applications. In order to remove a conflict with a similarly named project on Apache, called "Weaver," Weave's name changed to Twill when it moved to Apache incubation. Twill functions as a scaled-out proxy. Twill is a middleware layer in between YARN and any application on YARN. When you develop a Twill app, Twill handles APIs in YARN that resemble a multi-threaded application familiar to Java. It is very easy to build multi-processed distributed applications in Twill. | [1. Apache Twill Incubator](https://incubator.apache.org/projects/twill.html) |
| Damballa Parkour | Library for develop MapReduce programs using the LISP like language Clojure. Parkour aims to provide deep Clojure integration for Hadoop. Programs using Parkour are normal Clojure programs, using standard Clojure functions instead of new framework abstractions. Programs using Parkour are also full Hadoop programs, with complete access to absolutely everything possible in raw Java Hadoop MapReduce. | [1. Parkour GitHub Project](https://github.com/damballa/parkour) |
| Apache Hama | Apache Top-Level open source project, allowing you to do advanced analytics beyond MapReduce. Many data analysis techniques such as machine learning and graph algorithms require iterative computations, this is where Bulk Synchronous Parallel model can be more effective than "plain" MapReduce. | [1. Hama site](http://hama.apache.org/) |
| Datasalt Pangool | A new MapReduce paradigm. A new API for MR jobs, in higher level than Java. | [1.Pangool](http://pangool.net)  [2.GitHub Pangool](https://github.com/datasalt/pangool) |
| Apache Tez | Tez is a proposal to develop a generic application which can be used to process complex data-processing task DAGs and runs natively on Apache Hadoop YARN. Tez generalizes the MapReduce paradigm to a more powerful framework based on expressing computations as a dataflow graph. Tez is not meant directly for end-users – in fact it enables developers to build end-user applications with much better performance and flexibility. Hadoop has traditionally been a batch-processing platform for large amounts of data. However, there are a lot of use cases for near-real-time performance of query processing. There are also several workloads, such as Machine Learning, which do not fit will into the MapReduce paradigm. Tez helps Hadoop address these use cases. Tez framework constitutes part of Stinger initiative (a low latency based SQL type query interface for Hadoop based on Hive). | [1. Apache Tez Incubator](http://incubator.apache.org/projects/tez.html)  [2. Hortonworks Apache Tez page](http://hortonworks.com/hadoop/tez/) |
| Apache DataFu | DataFu provides a collection of Hadoop MapReduce jobs and functions in higher level languages based on it to perform data analysis. It provides functions for common statistics tasks (e.g. quantiles, sampling), PageRank, stream sessionization, and set and bag operations. DataFu also provides Hadoop jobs for incremental data processing in MapReduce. DataFu is a collection of Pig UDFs (including PageRank, sessionization, set operations, sampling, and much more) that were originally developed at LinkedIn. | [1. DataFu Apache Incubator](http://incubator.apache.org/projects/datafu.html) |
| Pydoop | Pydoop is a Python MapReduce and HDFS API for Hadoop, built upon the C++ Pipes and the C libhdfs APIs, that allows to write full-fledged MapReduce applications with HDFS access. Pydoop has several advantages over Hadoop’s built-in solutions for Python programming, i.e., Hadoop Streaming and Jython: being a CPython package, it allows you to access all standard library and third party modules, some of which may not be available. | [1. SF Pydoop site](http://pydoop.sourceforge.net/docs/)  [2. Pydoop GitHub Project](https://github.com/crs4/pydoop) |
| Kangaroo | Open-source project from Conductor for writing MapReduce jobs consuming data from Kafka. The introductory post explains Conductor’s use case—loading data from Kafka to HBase by way of a MapReduce job using the HFileOutputFormat. Unlike other solutions which are limited to a single InputSplit per Kafka partition, Kangaroo can launch multiple consumers at different offsets in the stream of a single partition for increased throughput and parallelism. | [1. Kangaroo Introduction](http://www.conductor.com/nightlight/data-stream-processing-bulk-kafka-hadoop/)  [2. Kangaroo GitHub Project](https://github.com/Conductor/kangaroo) |
| TinkerPop | Graph computing framework written in Java. Provides a core API that graph system vendors can implement. There are various types of graph systems including in-memory graph libraries, OLTP graph databases, and OLAP graph processors. Once the core interfaces are implemented, the underlying graph system can be queried using the graph traversal language Gremlin and processed with TinkerPop-enabled algorithms. For many, TinkerPop is seen as the JDBC of the graph computing community. | [1. Apache Tinkerpop Proposal](https://wiki.apache.org/incubator/TinkerPopProposal)  [2. TinkerPop site](http://www.tinkerpop.com/) |
| Pachyderm MapReduce | Pachyderm is a completely new MapReduce engine built on top Docker and CoreOS. In Pachyderm MapReduce (PMR) a job is an HTTP server inside a Docker container (a microservice). You give Pachyderm a Docker image and it will automatically distribute it throughout the cluster next to your data. Data is POSTed to the container over HTTP and the results are stored back in the file system. You can implement the web server in any language you want and pull in any library. Pachyderm also creates a DAG for all the jobs in the system and their dependencies and it automatically schedules the pipeline such that each job isn’t run until it’s dependencies have completed. Everything in Pachyderm “speaks in diffs” so it knows exactly which data has changed and which subsets of the pipeline need to be rerun. CoreOS is an open source lightweight operating system based on Chrome OS, actually CoreOS is a fork of Chrome OS. CoreOS provides only the minimal functionality required for deploying applications inside software containers, together with built-in mechanisms for service discovery and configuration sharing | [1. Pachyderm site](http://www.pachyderm.io/)  [2. Pachyderm introduction article](https://medium.com/pachyderm-data/lets-build-a-modern-hadoop-4fc160f8d74f) |
| Apache Beam | Apache Beam is an open source, unified model for defining and executing data-parallel processing pipelines, as well as a set of language-specific SDKs for constructing pipelines and runtime-specific Runners for executing them.  The model behind Beam evolved from a number of internal Google data processing projects, including MapReduce, FlumeJava, and Millwheel. This model was originally known as the “Dataflow Model” and first implemented as Google Cloud Dataflow, including a Java SDK on GitHub for writing pipelines and fully managed service for executing them on Google Cloud Platform.  In January 2016, Google and a number of partners submitted the Dataflow Programming Model and SDKs portion as an Apache Incubator Proposal, under the name Apache Beam (unified Batch + strEAM processing). | [1. Apache Beam Proposal](https://wiki.apache.org/incubator/BeamProposal)  [2. DataFlow Beam and Spark Comparasion](https://cloud.google.com/dataflow/blog/dataflow-beam-and-spark-comparison) |
| **NoSQL Databases** | | |
| **Column Data Model** | | |
| Apache HBase | Google BigTable Inspired. Non-relational distributed database. Ramdom, real-time r/w operations in column-oriented very large tables (BDDB: Big Data Data Base). It’s the backing system for MR jobs outputs. It’s the Hadoop database. It’s for backing Hadoop MapReduce jobs with Apache HBase tables | [1. Apache HBase Home](https://hbase.apache.org/)  [2. Mirror of HBase on Github](https://github.com/apache/hbase) |
| Apache Cassandra | Distributed Non-SQL DBMS, it’s a BDDB. MR can retrieve data from Cassandra. This BDDB can run without HDFS, or on-top of HDFS (DataStax fork of Cassandra). HBase and its required supporting systems are derived from what is known of the original Google BigTable and Google File System designs (as known from the Google File System paper Google published in 2003, and the BigTable paper published in 2006). Cassandra on the other hand is a recent open source fork of a standalone database system initially coded by Facebook, which while implementing the BigTable data model, uses a system inspired by Amazon’s Dynamo for storing data (in fact much of the initial development work on Cassandra was performed by two Dynamo engineers recruited to Facebook from Amazon). | [1. Apache HBase Home](http://cassandra.apache.org)  [2. Cassandra on GitHub](https://github.com/apache/cassandra)  [3. Training Resources](https://academy.datastax.com)  [4. Cassandra - Paper](https://www.cs.cornell.edu/projects/ladis2009/papers/lakshman-ladis2009.pdf) |
| Hypertable | Database system inspired by publications on the design of Google's BigTable. The project is based on experience of engineers who were solving large-scale data-intensive tasks for many years. Hypertable runs on top of a distributed file system such as the Apache Hadoop DFS, GlusterFS, or the Kosmos File System (KFS). It is written almost entirely in C++. Sposored by Baidu the Chinese search engine. | TODO |
| Apache Accumulo | Distributed key/value store is a robust, scalable, high performance data storage and retrieval system. Apache Accumulo is based on Google's BigTable design and is built on top of Apache Hadoop, Zookeeper, and Thrift. Accumulo is software created by the NSA with security features. | [1. Apache Accumulo Home](https://accumulo.apache.org/) |
| Apache Kudu | Distributed, columnar, relational data store optimized for analytical use cases requiring very fast reads with competitive write speeds.   * Relational data model (tables) with strongly-typed columns and a fast, online alter table operation. * Scale-out and sharded with support for partitioning based on key ranges and/or hashing. * Fault-tolerant and consistent due to its implementation of Raft consensus. * Supported by Apache Impala and Apache Drill, enabling fast SQL reads and writes through those systems. * Integrates with MapReduce and Spark. * Additionally provides "NoSQL" APIs in Java, Python, and C++. | [1. Apache Kudu Home](http://getkudu.io/) [2. Kudu on Github](http://github.com/cloudera/kudu) [3. Kudu technical whitepaper (pdf)](http://getkudu.io/kudu.pdf) |
| **Document Data Model** | | |
| MongoDB | Document-oriented database system. It is part of the NoSQL family of database systems. Instead of storing data in tables as is done in a "classical" relational database, MongoDB stores structured data as JSON-like documents | [1. Mongodb site](http://www.mongodb.org/) |
| RethinkDB | RethinkDB is built to store JSON documents, and scale to multiple machines with very little effort. It has a pleasant query language that supports really useful queries like table joins and group by, and is easy to setup and learn. | [1. RethinkDB site](http://www.rethinkdb.com/) |
| ArangoDB | An open-source database with a flexible data model for documents, graphs, and key-values. Build high performance applications using a convenient sql-like query language or JavaScript extensions. | [1. ArangoDB site](https://www.arangodb.org/) |
| **Stream Data Model** | | |
| EventStore | An open-source, functional database with support for Complex Event Processing. It provides a persistence engine for applications using event-sourcing, or for storing time-series data. Event Store is written in C#, C++ for the server which runs on Mono or the .NET CLR, on Linux or Windows. Applications using Event Store can be written in JavaScript. Event sourcing (ES) is a way of persisting your application's state by storing the history that determines the current state of your application. | [1. EventStore site](http://geteventstore.com/) |
| **Key-value Data Model** | | |
| Redis DataBase | Redis is an open-source, networked, in-memory, data structures store with optional durability. It is written in ANSI C. In its outer layer, the Redis data model is a dictionary which maps keys to values. One of the main differences between Redis and other structured storage systems is that Redis supports not only strings, but also abstract data types. Sponsored by Redis Labs. It’s BSD licensed. | [1. Redis site](http://redis.io/)  [2. Redis Labs site](http://redislabs.com/) |
| Linkedin Voldemort | Distributed data store that is designed as a key-value store used by LinkedIn for high-scalability storage. | [1. Voldemort site](http://www.project-voldemort.com/voldemort/) |
| RocksDB | RocksDB is an embeddable persistent key-value store for fast storage. RocksDB can also be the foundation for a client-server database but our current focus is on embedded workloads. | [1. RocksDB site](http://rocksdb.org/) |
| OpenTSDB | OpenTSDB is a distributed, scalable Time Series Database (TSDB) written on top of HBase. OpenTSDB was written to address a common need: store, index and serve metrics collected from computer systems (network gear, operating systems, applications) at a large scale, and make this data easily accessible and graphable. | [1. OpenTSDB site](http://opentsdb.net/) |
| **Graph Data Model** | | |
| ArangoDB | An open-source database with a flexible data model for documents, graphs, and key-values. Build high performance applications using a convenient sql-like query language or JavaScript extensions. | [1. ArangoDB site](https://www.arangodb.org/) |
| Neo4j | An open-source graph database writting entirely in Java. It is an embedded, disk-based, fully transactional Java persistence engine that stores data structured in graphs rather than in tables. | [1. Neo4j site](http://www.neo4j.org/) |
| TitanDB | TitanDB is a highly scalable graph database optimized for storing and querying large graphs with billions of vertices and edges distributed across a multi-machine cluster. Titan is a transactional database that can support thousands of concurrent users. | [1. Titan site](http://thinkaurelius.github.io/titan/) |
| **NewSQL Databases** | | |
| TokuDB | TokuDB is a storage engine for MySQL and MariaDB that is specifically designed for high performance on write-intensive workloads. It achieves this via Fractal Tree indexing. TokuDB is a scalable, ACID and MVCC compliant storage engine. TokuDB is one of the technologies that enable Big Data in MySQL. | TODO |
| HandlerSocket | HandlerSocket is a NoSQL plugin for MySQL/MariaDB (the storage engine of MySQL). It works as a daemon inside the mysqld process, accepting TCP connections, and executing requests from clients. HandlerSocket does not support SQL queries. Instead, it supports simple CRUD operations on tables. HandlerSocket can be much faster than mysqld/libmysql in some cases because it has lower CPU, disk, and network overhead. | TODO |
| Akiban Server | Akiban Server is an open source database that brings document stores and relational databases together. Developers get powerful document access alongside surprisingly powerful SQL. | TODO |
| Drizzle | Drizzle is a re-designed version of the MySQL v6.0 codebase and is designed around a central concept of having a microkernel architecture. Features such as the query cache and authentication system are now plugins to the database, which follow the general theme of "pluggable storage engines" that were introduced in MySQL 5.1. It supports PAM, LDAP, and HTTP AUTH for authentication via plugins it ships. Via its plugin system it currently supports logging to files, syslog, and remote services such as RabbitMQ and Gearman. Drizzle is an ACID-compliant relational database that supports transactions via an MVCC design | TODO |
| Haeinsa | Haeinsa is linearly scalable multi-row, multi-table transaction library for HBase. Use Haeinsa if you need strong ACID semantics on your HBase cluster. Is based on Google Perlocator concept. | TODO |
| SenseiDB | Open-source, distributed, realtime, semi-structured database. Some Features: Full-text search, Fast realtime updates, Structured and faceted search, BQL: SQL-like query language, Fast key-value lookup, High performance under concurrent heavy update and query volumes, Hadoop integration | [1. SenseiDB site](http://senseidb.com/) |
| Sky | Sky is an open source database used for flexible, high performance analysis of behavioral data. For certain kinds of data such as clickstream data and log data, it can be several orders of magnitude faster than traditional approaches such as SQL databases or Hadoop. | [1. SkyDB site](http://skydb.io/) |
| BayesDB | BayesDB, a Bayesian database table, lets users query the probable implications of their tabular data as easily as an SQL database lets them query the data itself. Using the built-in Bayesian Query Language (BQL), users with no statistics training can solve basic data science problems, such as detecting predictive relationships between variables, inferring missing values, simulating probable observations, and identifying statistically similar database entries. | [1. BayesDB site](http://probcomp.csail.mit.edu/bayesdb/index.html) |
| InfluxDB | InfluxDB is an open source distributed time series database with no external dependencies. It's useful for recording metrics, events, and performing analytics. It has a built-in HTTP API so you don't have to write any server side code to get up and running. InfluxDB is designed to be scalable, simple to install and manage, and fast to get data in and out. It aims to answer queries in real-time. That means every data point is indexed as it comes in and is immediately available in queries that should return under 100ms. | [1. InfluxDB site](http://influxdb.org/) |
| **SQL-on-Hadoop** | | |
| Apache Hive | Data Warehouse infrastructure developed by Facebook. Data summarization, query, and analysis. It’s provides SQL-like language (not SQL92 compliant): HiveQL. | [1. Apache HIVE site](http://hive.apache.org/)  [2. Apache HIVE GitHub Project](https://github.com/apache/hive) |
| Apache HCatalog | HCatalog’s table abstraction presents users with a relational view of data in the Hadoop Distributed File System (HDFS) and ensures that users need not worry about where or in what format their data is stored. Right now HCatalog is part of Hive. Only old versions are separated for download. | TODO |
| Apache Trafodion | Apache Trafodion is a webscale SQL-on-Hadoop solution enabling enterprise-class transactional and operational workloads on HBase. Trafodion is a native MPP ANSI SQL database engine that builds on the scalability, elasticity and flexibility of HDFS and HBase, extending these to provide guaranteed transactional integrity for all workloads including multi-column, multi-row, multi-table, and multi-server updates. | [1. Apache Trafodion website](http://trafodion.incubator.apache.org)  [2. Apache Trafodion wiki](https://cwiki.apache.org/confluence/display/TRAFODION/Apache+Trafodion+Home)  [3. Apache Trafodion GitHub Project](https://github.com/apache/incubator-trafodion) |
| Apache HAWQ | Apache HAWQ is a Hadoop native SQL query engine that combines key technological advantages of MPP database evolved from Greenplum Database, with the scalability and convenience of Hadoop. | [1. Apache HAWQ site](http://hawq.incubator.apache.org/)  [2. HAWQ GitHub Project](https://github.com/apache/incubator-hawq) |
| Apache Drill | Drill is the open source version of Google's Dremel system which is available as an infrastructure service called Google BigQuery. In recent years open source systems have emerged to address the need for scalable batch processing (Apache Hadoop) and stream processing (Storm, Apache S4). Apache Hadoop, originally inspired by Google's internal MapReduce system, is used by thousands of organizations processing large-scale datasets. Apache Hadoop is designed to achieve very high throughput, but is not designed to achieve the sub-second latency needed for interactive data analysis and exploration. Drill, inspired by Google's internal Dremel system, is intended to address this need | [1. Apache Incubator Drill](http://incubator.apache.org/drill/) |
| Cloudera Impala | The Apache-licensed Impala project brings scalable parallel database technology to Hadoop, enabling users to issue low-latency SQL queries to data stored in HDFS and Apache HBase without requiring data movement or transformation. It's a Google Dremel clone (Big Query google). | [1. Cloudera Impala site](http://www.cloudera.com/content/cloudera/en/products-and-services/cdh/impala.html)  [2. Impala GitHub Project](https://github.com/cloudera/impala) |
| Facebook Presto | Facebook has open sourced Presto, a SQL engine it says is on average 10 times faster than Hive for running queries across large data sets stored in Hadoop and elsewhere. | [1. Presto site](http://prestodb.io/) |
| Datasalt Splout SQL | Splout allows serving an arbitrarily big dataset with high QPS rates and at the same time provides full SQL query syntax. | TODO |
| Apache Tajo | Apache Tajo is a robust big data relational and distributed data warehouse system for Apache Hadoop. Tajo is designed for low-latency and scalable ad-hoc queries, online aggregation, and ETL (extract-transform-load process) on large-data sets stored on HDFS (Hadoop Distributed File System) and other data sources. By supporting SQL standards and leveraging advanced database techniques, Tajo allows direct control of distributed execution and data flow across a variety of query evaluation strategies and optimization opportunities. For reference, the Apache Software Foundation announced Tajo as a Top-Level Project in April 2014. | [1. Apache Tajo site](http://tajo.apache.org/) |
| Apache Phoenix | Apache Phoenix is a SQL skin over HBase delivered as a client-embedded JDBC driver targeting low latency queries over HBase data. Apache Phoenix takes your SQL query, compiles it into a series of HBase scans, and orchestrates the running of those scans to produce regular JDBC result sets. The table metadata is stored in an HBase table and versioned, such that snapshot queries over prior versions will automatically use the correct schema. Direct use of the HBase API, along with coprocessors and custom filters, results in performance on the order of milliseconds for small queries, or seconds for tens of millions of rows. | [1. Apache Phoenix site](http://phoenix.incubator.apache.org/index.html) |
| Apache MRQL | MRQL is a query processing and optimization system for large-scale, distributed data analysis, built on top of Apache Hadoop, Hama, and Spark. MRQL (pronounced miracle) is a query processing and optimization system for large-scale, distributed data analysis. MRQL (the MapReduce Query Language) is an SQL-like query language for large-scale data analysis on a cluster of computers. The MRQL query processing system can evaluate MRQL queries in three modes:   * in Map-Reduce mode using Apache Hadoop, * in BSP mode (Bulk Synchronous Parallel mode) using Apache Hama, and * in Spark mode using Apache Spark. * in Flink mode using Apache Flink. | [1. Apache Incubator MRQL site](http://mrql.incubator.apache.org/) |
| Kylin | Kylin is an open source Distributed Analytics Engine from eBay Inc. that provides SQL interface and multi-dimensional analysis (OLAP) on Hadoop supporting extremely large datasets | [1. Kylin project site](http://www.kylin.io/) |
| **Data Ingestion** | | |
| Apache Flume | Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It has a simple and flexible architecture based on streaming data flows. It is robust and fault tolerant with tunable reliability mechanisms and many failover and recovery mechanisms. It uses a simple extensible data model that allows for online analytic application. | [1. Apache Flume project site](http://flume.apache.org/) |
| Apache Sqoop | System for bulk data transfer between HDFS and structured datastores as RDBMS. Like Flume but from HDFS to RDBMS. | [1. Apache Sqoop project site](http://sqoop.apache.org/) |
| Facebook Scribe | Log agregator in real-time. It’s a Apache Thrift Service. | TODO |
| Apache Chukwa | Large scale log aggregator, and analytics. | TODO |
| Apache Kafka | Distributed publish-subscribe system for processing large amounts of streaming data. Kafka is a Message Queue developed by LinkedIn that persists messages to disk in a very performant manner. Because messages are persisted, it has the interesting ability for clients to rewind a stream and consume the messages again. Another upside of the disk persistence is that bulk importing the data into HDFS for offline analysis can be done very quickly and efficiently. Storm, developed by BackType (which was acquired by Twitter a year ago), is more about transforming a stream of messages into new streams. | [1. Apache Kafka](http://kafka.apache.org/)  [2. GitHub source code](https://github.com/apache/kafka/) |
| Netflix Suro | Suro has its roots in Apache Chukwa, which was initially adopted by Netflix. Is a log agregattor like Storm, Samza. | TODO |
| Apache Samza | Apache Samza is a distributed stream processing framework. It uses Apache Kafka for messaging, and Apache Hadoop YARN to provide fault tolerance, processor isolation, security, and resource management. Developed by http://www.linkedin.com/in/jaykreps Linkedin. | TODO |
| Cloudera Morphline | Cloudera Morphlines is a new open source framework that reduces the time and skills necessary to integrate, build, and change Hadoop processing applications that extract, transform, and load data into Apache Solr, Apache HBase, HDFS, enterprise data warehouses, or analytic online dashboards. | TODO |
| HIHO | This project is a framework for connecting disparate data sources with the Apache Hadoop system, making them interoperable. HIHO connects Hadoop with multiple RDBMS and file systems, so that data can be loaded to Hadoop and unloaded from Hadoop | TODO |
| Apache NiFi | Apache NiFi is a dataflow system that is currently under incubation at the Apache Software Foundation. NiFi is based on the concepts of flow-based programming and is highly configurable. NiFi uses a component based extension model to rapidly add capabilities to complex dataflows. Out of the box NiFi has several extensions for dealing with file-based dataflows such as FTP, SFTP, and HTTP integration as well as integration with HDFS. One of NiFi’s unique features is a rich, web-based interface for designing, controlling, and monitoring a dataflow. | [1. Apache NiFi](http://nifi.apache.org/index.html) |
| Apache ManifoldCF | Apache ManifoldCF provides a framework for connecting source content repositories like file systems, DB, CMIS, SharePoint, FileNet ... to target repositories or indexes, such as Apache Solr or ElasticSearch. It's a kind of crawler for multi-content repositories, supporting a lot of sources and multi-format conversion for indexing by means of Apache Tika Content Extractor transformation filter. | [1. Apache ManifoldCF](http://manifoldcf.apache.org/) |
| **Service Programming** | | |
| Apache Thrift | A cross-language RPC framework for service creations. It’s the service base for Facebook technologies (the original Thrift contributor). Thrift provides a framework for developing and accessing remote services. It allows developers to create services that can be consumed by any application that is written in a language that there are Thrift bindings for. Thrift manages serialization of data to and from a service, as well as the protocol that describes a method invocation, response, etc. Instead of writing all the RPC code -- you can just get straight to your service logic. Thrift uses TCP and so a given service is bound to a particular port. | [1. Apache Thrift](http://thrift.apache.org/) |
| Apache Zookeeper | It’s a coordination service that gives you the tools you need to write correct distributed applications. ZooKeeper was developed at Yahoo! Research. Several Hadoop projects are already using ZooKeeper to coordinate the cluster and provide highly-available distributed services. Perhaps most famous of those are Apache HBase, Storm, Kafka. ZooKeeper is an application library with two principal implementations of the APIs—Java and C—and a service component implemented in Java that runs on an ensemble of dedicated servers. Zookeeper is for building distributed systems, simplifies the development process, making it more agile and enabling more robust implementations. Back in 2006, Google published a paper on "Chubby", a distributed lock service which gained wide adoption within their data centers. Zookeeper, not surprisingly, is a close clone of Chubby designed to fulfill many of the same roles for HDFS and other Hadoop infrastructure. | [1. Apache Zookeeper](http://zookeeper.apache.org/)  [2. Google Chubby paper](http://research.google.com/archive/chubby.html) |
| Apache Avro | Apache Avro is a framework for modeling, serializing and making Remote Procedure Calls (RPC). Avro data is described by a schema, and one interesting feature is that the schema is stored in the same file as the data it describes, so files are self-describing. Avro does not require code generation. This framework can compete with other similar tools like: Apache Thrift, Google Protocol Buffers, ZeroC ICE, and so on. | [1. Apache Avro](http://avro.apache.org/) |
| Apache Curator | Curator is a set of Java libraries that make using Apache ZooKeeper much easier. | TODO |
| Apache karaf | Apache Karaf is an OSGi runtime that runs on top of any OSGi framework and provides you a set of services, a powerful provisioning concept, an extensible shell and more. | TODO |
| Twitter Elephant Bird | Elephant Bird is a project that provides utilities (libraries) for working with LZOP-compressed data. It also provides a container format that supports working with Protocol Buffers, Thrift in MapReduce, Writables, Pig LoadFuncs, Hive SerDe, HBase miscellanea. This open source library is massively used in Twitter. | [1. Elephant Bird GitHub](https://github.com/kevinweil/elephant-bird) |
| Linkedin Norbert | Norbert is a library that provides easy cluster management and workload distribution. With Norbert, you can quickly distribute a simple client/server architecture to create a highly scalable architecture capable of handling heavy traffic. Implemented in Scala, Norbert wraps ZooKeeper, Netty and uses Protocol Buffers for transport to make it easy to build a cluster aware application. A Java API is provided and pluggable load balancing strategies are supported with round robin and consistent hash strategies provided out of the box. | [1. Linedin Project](http://data.linkedin.com/opensource/norbert)  [2. GitHub source code](https://github.com/rhavyn/norbert) |
| **Scheduling** | | |
| Apache Oozie | Workflow scheduler system for MR jobs using DAGs (Direct Acyclical Graphs). Oozie Coordinator can trigger jobs by time (frequency) and data availability | [1. Apache Oozie](http://oozie.apache.org/)  [2. GitHub source code](https://github.com/apache/oozie) |
| Linkedin Azkaban | Hadoop workflow management. A batch job scheduler can be seen as a combination of the cron and make Unix utilities combined with a friendly UI. | TODO |
| Apache Falcon | Apache Falcon is a data management framework for simplifying data lifecycle management and processing pipelines on Apache Hadoop. It enables users to configure, manage and orchestrate data motion, pipeline processing, disaster recovery, and data retention workflows. Instead of hard-coding complex data lifecycle capabilities, Hadoop applications can now rely on the well-tested Apache Falcon framework for these functions. Falcon’s simplification of data management is quite useful to anyone building apps on Hadoop. Data Management on Hadoop encompasses data motion, process orchestration, lifecycle management, data discovery, etc. among other concerns that are beyond ETL. Falcon is a new data processing and management platform for Hadoop that solves this problem and creates additional opportunities by building on existing components within the Hadoop ecosystem (ex. Apache Oozie, Apache Hadoop DistCp etc.) without reinventing the wheel. | TODO |
| Schedoscope | Schedoscope is a new open-source project providing a scheduling framework for painfree agile development, testing, (re)loading, and monitoring of your datahub, lake, or whatever you choose to call your Hadoop data warehouse these days. Datasets (including dependencies) are defined using a scala DSL, which can embed MapReduce jobs, Pig scripts, Hive queries or Oozie workflows to build the dataset. The tool includes a test framework to verify logic and a command line utility to load and reload data. | [GitHub source code](https://github.com/ottogroup/schedoscope) |
| **Machine Learning** | | |
| Apache Mahout | Machine learning library and math library, on top of MapReduce. | TODO |
| WEKA | Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is free software available under the GNU General Public License. | TODO |
| Cloudera Oryx | The Oryx open source project provides simple, real-time large-scale machine learning / predictive analytics infrastructure. It implements a few classes of algorithm commonly used in business applications: collaborative filtering / recommendation, classification / regression, and clustering. | [1. Oryx at GitHub](https://github.com/cloudera/oryx)  [2. Cloudera forum for Machine Learning](https://community.cloudera.com/t5/Data-Science-and-Machine/bd-p/Mahout) |
| Deeplearning4j | The Deeplearning4j open-source project is the most widely used deep-learning framework for the JVM. DL4J includes deep neural nets such as recurrent neural networks, Long Short Term Memory Networks (LSTMs), convolutional neural networks, various autoencoders and feedforward neural networks such as restricted Boltzmann machines and deep-belief networks. It also has natural language-processing algorithms such as word2vec, doc2vec, GloVe and TF-IDF. All Deeplearning4j networks run distributed on multiple CPUs and GPUs. They work as Hadoop jobs, and integrate with Spark on the slace level for host-thread orchestration. Deeplearning4j's neural networks are applied to use cases such as fraud and anomaly detection, recommender systems, and predictive maintenance. | [1. Deeplearning4j Website](http://deeplearning4j.org/)  [2. Gitter Community for Deeplearning4j](https://gitter.im/deeplearning4j/deeplearning4j) |
| MADlib | The MADlib project leverages the data-processing capabilities of an RDBMS to analyze data. The aim of this project is the integration of statistical data analysis into databases. The MADlib project is self-described as the Big Data Machine Learning in SQL for Data Scientists. The MADlib software project began the following year as a collaboration between researchers at UC Berkeley and engineers and data scientists at EMC/Greenplum (now Pivotal) | [1. MADlib Community](http://madlib.net/community/) |
| H2O | H2O is a statistical, machine learning and math runtime tool for bigdata analysis. Developed by the predictive analytics company H2O.ai, H2O has established a leadership in the ML scene together with R and Databricks’ Spark. According to the team,  H2O is the world’s fastest in-memory platform for machine learning and predictive analytics on big data. It is designed to help users scale machine learning, math, and statistics over large datasets.  In addition to H2O’s point and click Web-UI, its REST API allows easy integration into various clients. This means explorative analysis of data can be done in a typical fashion in R, Python, and Scala; and entire workflows can be written up as automated scripts. | [1. H2O at GitHub](https://github.com/h2oai/h2o-dev)  [2. H2O Blog](http://h2o.ai/blog) |
| Sparkling Water | Sparkling Water combines two open source technologies: Apache Spark and H2O - a machine learning engine.  It makes H2O’s library of Advanced Algorithms including Deep Learning, GLM, GBM, KMeans, PCA, and Random Forest accessible from Spark workflows. Spark users are provided with the options to select the best features from either platforms to meet their Machine Learning needs.  Users can combine Sparks’ RDD API and Spark MLLib with H2O’s machine learning algorithms, or use H2O independent of Spark in the model building process and post-process the results in Spark.  Sparkling Water provides a transparent integration of H2O’s framework and data structures into Spark’s RDD-based environment by sharing the same execution space as well as providing a RDD-like API for H2O data structures. | [1. Sparkling Water at GitHub](https://github.com/h2oai/sparkling-water)  [2. Sparkling Water Examples](https://github.com/h2oai/sparkling-water/tree/master/examples) |
| Apache SystemML | Apache SystemML was open sourced by IBM and it's pretty related with Apache Spark. If you thinking in Apache Spark as the analytics operating system for any application that taps into huge volumes of streaming data. MLLib, the machine learning library for Spark, provides developers with a rich set of machine learning algorithms. And SystemML enables developers to translate those algorithms so they can easily digest different kinds of data and to run on different kinds of computers.  SystemML allows a developer to write a single machine learning algorithm and automatically scale it up using Spark or Hadoop.  SystemML scales for big data analytics with high performance optimizer technology, and empowers users to write customized machine learning algorithms using simple, domain-specific language (DSL) without learning complicated distributed programming. It is an extensible complement framework of Spark MLlib. | [1. Apache SystemML](http://systemml.apache.org)  [2. Apache Proposal](https://wiki.apache.org/incubator/SystemML) |
| **Benchmarking and QA Tools** | | |
| Apache Hadoop Benchmarking | There are two main JAR files in Apache Hadoop for benchmarking. This JAR are micro-benchmarks for testing particular parts of the infrastructure, for instance TestDFSIO analyzes the disk system, TeraSort evaluates MapReduce tasks, WordCount measures cluster performance, etc. Micro-Benchmarks are packaged in the tests and exmaples JAR files, and you can get a list of them, with descriptions, by invoking the JAR file with no arguments. With regards Apache Hadoop 2.2.0 stable version we have available the following JAR files for test, examples and benchmarking. The Hadoop micro-benchmarks, are bundled in this JAR files: hadoop-mapreduce-examples-2.2.0.jar, hadoop-mapreduce-client-jobclient-2.2.0-tests.jar. | [1. MAPREDUCE-3561 umbrella ticket to track all the issues related to performance](https://issues.apache.org/jira/browse/MAPREDUCE-3561) |
| Yahoo Gridmix3 | Hadoop cluster benchmarking from Yahoo engineer team. | TODO |
| PUMA Benchmarking | Benchmark suite which represents a broad range of MapReduce applications exhibiting application characteristics with high/low computation and high/low shuffle volumes. There are a total of 13 benchmarks, out of which Tera-Sort, Word-Count, and Grep are from Hadoop distribution. The rest of the benchmarks were developed in-house and are currently not part of the Hadoop distribution. The three benchmarks from Hadoop distribution are also slightly modified to take number of reduce tasks as input from the user and generate final time completion statistics of jobs. | [1. MAPREDUCE-5116](https://issues.apache.org/jira/browse/MAPREDUCE-5116)  [2. Faraz Ahmad researcher](https://sites.google.com/site/farazahmad/)  [3. PUMA Docs](https://sites.google.com/site/farazahmad/pumabenchmarks) |
| Berkeley SWIM Benchmark | The SWIM benchmark (Statistical Workload Injector for MapReduce), is a benchmark representing a real-world big data workload developed by University of California at Berkley in close cooperation with Facebook. This test provides rigorous measurements of the performance of MapReduce systems comprised of real industry workloads.. | [1. GitHub SWIN](https://github.com/SWIMProjectUCB/SWIM/wiki) |
| Intel HiBench | HiBench is a Hadoop benchmark suite. | TODO |
| Apache Yetus | To help maintain consistency over a large and disconnected set of committers, automated patch testing was added to Hadoop’s development process. This automated patch testing (now included as part of Apache Yetus) works as follows: when a patch is uploaded to the bug tracking system an automated process downloads the patch, performs some static analysis, and runs the unit tests. These results are posted back to the bug tracker and alerts notify interested parties about the state of the patch.  However The Apache Yetus project addresses much more than the traditional patch testing, it's a better approach including a massive rewrite of the patch testing facility used in Hadoop. | [1. Altiscale Blog Entry](https://www.altiscale.com/blog/apache-yetus-faster-more-reliable-software-development/)  [2. Apache Yetus Proposal](https://wiki.apache.org/incubator/YetusProposal)  [3. Apache Yetus Project site](https://yetus.apache.org/) |
| **Security** | | |
| Apache Sentry | Sentry is the next step in enterprise-grade big data security and delivers fine-grained authorization to data stored in Apache Hadoop. An independent security module that integrates with open source SQL query engines Apache Hive and Cloudera Impala, Sentry delivers advanced authorization controls to enable multi-user applications and cross-functional processes for enterprise data sets. Sentry was a Cloudera development. | TODO |
| Apache Knox Gateway | System that provides a single point of secure access for Apache Hadoop clusters. The goal is to simplify Hadoop security for both users (i.e. who access the cluster data and execute jobs) and operators (i.e. who control access and manage the cluster). The Gateway runs as a server (or cluster of servers) that serve one or more Hadoop clusters. | [1. Apache Knox](http://knox.apache.org/)  [2. Apache Knox Gateway Hortonworks web](http://hortonworks.com/hadoop/knox-gateway/) |
| Apache Ranger | Apache Argus Ranger (formerly called Apache Argus or HDP Advanced Security) delivers comprehensive approach to central security policy administration across the core enterprise security requirements of authentication, authorization, accounting and data protection. It extends baseline features for coordinated enforcement across Hadoop workloads from batch, interactive SQL and real–time and leverages the extensible architecture to apply policies consistently against additional Hadoop ecosystem components (beyond HDFS, Hive, and HBase) including Storm, Solr, Spark, and more. | [1. Apache Ranger](http://ranger.apache.org/)  [2. Apache Ranger Hortonworks web](http://hortonworks.com/hadoop/ranger/) |
| **Metadata Management** | | |
| Metascope | Metascope is a metadata management and data discovery tool which serves as an add-on to Schedoscope. Metascope is able to collect technical, operational and business metadata from your Hadoop Datahub and provides them easy to search and navigate via a portal. | [GitHub source code](https://github.com/ottogroup/metascope) |
| **System Deployment** | | |
| Apache Ambari | Intuitive, easy-to-use Hadoop management web UI backed by its RESTful APIs. Apache Ambari was donated by Hortonworks team to the ASF. It's a powerful and nice interface for Hadoop and other typical applications from the Hadoop ecosystem. Apache Ambari is under a heavy development, and it will incorporate new features in a near future. For example Ambari is able to deploy a complete Hadoop system from scratch, however is not possible use this GUI in a Hadoop system that is already running. The ability to provisioning the operating system could be a good addition, however probably is not in the roadmap.. | [1. Apache Ambari](http://ambari.apache.org/) |
| Cloudera HUE | Web application for interacting with Apache Hadoop. It's not a deploment tool, is an open-source Web interface that supports Apache Hadoop and its ecosystem, licensed under the Apache v2 license. HUE is used for Hadoop and its ecosystem user operations. For example HUE offers editors for Hive, Impala, Oozie, Pig, notebooks for Spark, Solr Search dashboards, HDFS, YARN, HBase browsers.. | [1. HUE home page](http://gethue.com/) |
| Apache Whirr | Apache Whirr is a set of libraries for running cloud services. It allows you to use simple commands to boot clusters of distributed systems for testing and experimentation. Apache Whirr makes booting clusters easy. | TODO |
| Apache Mesos | Mesos is a cluster manager that provides resource sharing and isolation across cluster applications. Like HTCondor, SGE or Troque can do it. However Mesos is hadoop centred design | TODO |
| Myriad | Myriad is 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. | [1. Myriad Github](https://github.com/mesos/myriad) |
| Marathon | Marathon is a Mesos framework for long-running services. Given that you have Mesos running as the kernel for your datacenter, Marathon is the init or upstart daemon. | TODO |
| Brooklyn | Brooklyn is a library that simplifies application deployment and management. For deployment, it is designed to tie in with other tools, giving single-click deploy and adding the concepts of manageable clusters and fabrics: Many common software entities available out-of-the-box. Integrates with Apache Whirr -- and thereby Chef and Puppet -- to deploy well-known services such as Hadoop and elasticsearch (or use POBS, plain-old-bash-scripts) Use PaaS's such as OpenShift, alongside self-built clusters, for maximum flexibility | TODO |
| Hortonworks HOYA | HOYA is defined as “running HBase On YARN”. The Hoya tool is a Java tool, and is currently CLI driven. It takes in a cluster specification – in terms of the number of regionservers, the location of HBASE\_HOME, the ZooKeeper quorum hosts, the configuration that the new HBase cluster instance should use and so on. So HOYA is for HBase deployment using a tool developed on top of YARN. Once the cluster has been started, the cluster can be made to grow or shrink using the Hoya commands. The cluster can also be stopped and later resumed. Hoya implements the functionality through YARN APIs and HBase’s shell scripts. The goal of the prototype was to have minimal code changes and as of this writing, it has required zero code changes in HBase. | [1. Hortonworks Blog](http://hortonworks.com/blog/introducing-hoya-hbase-on-yarn/) |
| Apache Helix | Apache Helix is a generic cluster management framework used for the automatic management of partitioned, replicated and distributed resources hosted on a cluster of nodes. Originally developed by Linkedin, now is in an incubator project at Apache. Helix is developed on top of Zookeeper for coordination tasks. | [1. Apache Helix](http://helix.apache.org/) |
| Apache Bigtop | Bigtop was originally developed and released as an open source packaging infrastructure by Cloudera. BigTop is used for some vendors to build their own distributions based on Apache Hadoop (CDH, Pivotal HD, Intel's distribution), however Apache Bigtop does many more tasks, like continuous integration testing (with Jenkins, maven, ...) and is useful for packaging (RPM and DEB), deployment with Puppet, and so on. BigTop also features vagrant recipes for spinning up "n-node" hadoop clusters, and the bigpetstore blueprint application which demonstrates construction of a full stack hadoop app with ETL, machine learning, and dataset generation. Apache Bigtop could be considered as a community effort with a main focus: put all bits of the Hadoop ecosystem as a whole, rather than individual projects. | [1. Apache Bigtop.](http://bigtop.apache.org/) |
| Buildoop | Buildoop is an open source project licensed under Apache License 2.0, based on Apache BigTop idea. Buildoop is a collaboration project that provides templates and tools to help you create custom Linux-based systems based on Hadoop ecosystem. The project is built from scrach using Groovy language, and is not based on a mixture of tools like BigTop does (Makefile, Gradle, Groovy, Maven), probably is easier to programming than BigTop, and the desing is focused in the basic ideas behind the buildroot Yocto Project. The project is in early stages of development right now. | [1. Hadoop Ecosystem Builder.](http://buildoop.github.io/) |
| Deploop | Deploop is a tool for provisioning, managing and monitoring Apache Hadoop clusters focused in the Lambda Architecture. LA is a generic design based on the concepts of Twitter engineer Nathan Marz. This generic architecture was designed addressing common requirements for big data. The Deploop system is in ongoing development, in alpha phases of maturity. The system is setup on top of highly scalable techologies like Puppet and MCollective. | [1. The Hadoop Deploy System.](http://deploop.github.io/) |
| SequenceIQ Cloudbreak | Cloudbreak is an effective way to start and run multiple instances and versions of Hadoop clusters in the cloud, Docker containers or bare metal. It is a cloud and infrastructure agnostic and cost effictive Hadoop As-a-Service platform API. Provides automatic scaling, secure multi tenancy and full cloud lifecycle management.  Cloudbreak leverages the cloud infrastructure platforms to create host instances, uses Docker technology to deploy the requisite containers cloud-agnostically, and uses Apache Ambari (via Ambari Blueprints) to install and manage a Hortonworks cluster. This is a tool within the HDP ecosystem. | [1. GitHub project.](https://github.com/sequenceiq/cloudbreak)  [2. Cloudbreak introduction.](http://sequenceiq.com/cloudbreak-docs/latest/#introduction)  [3. Cloudbreak in Hortonworks.](http://hortonworks.com/hadoop/cloudbreak/) |
| **Applications** | | |
| Apache Nutch | Highly extensible and scalable open source web crawler software project. A search engine based on Lucene: A Web crawler is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing. Web crawlers can copy all the pages they visit for later processing by a search engine that indexes the downloaded pages so that users can search them much more quickly. | TODO |
| Sphnix Search Server | Sphinx lets you either batch index and search data stored in an SQL database, NoSQL storage, or just files quickly and easily — or index and search data on the fly, working with Sphinx pretty much as with a database server. | TODO |
| Apache OODT | OODT was originally developed at NASA Jet Propulsion Laboratory to support capturing, processing and sharing of data for NASA's scientific archives | TODO |
| HIPI Library | HIPI is a library for Hadoop's MapReduce framework that provides an API for performing image processing tasks in a distributed computing environment. | TODO |
| PivotalR | PivotalR is a package that enables users of R, the most popular open source statistical programming language and environment to interact with the Pivotal (Greenplum) Database as well as Pivotal HD / HAWQ and the open-source database PostgreSQL for Big Data analytics. R is a programming language and data analysis software: you do data analysis in R by writing scripts and functions in the R programming language. R is a complete, interactive, object-oriented language: designed by statisticians, for statisticians. The language provides objects, operators and functions that make the process of exploring, modeling, and visualizing data a natural one. | [1. PivotalR on GitHub](https://github.com/gopivotal/PivotalR) |
| **Development Frameworks** | | |
| Jumbune | Jumbune is an open source product that sits on top of any Hadoop distribution and assists in development and administration of MapReduce solutions. The objective of the product is to assist analytical solution providers to port fault free applications on production Hadoop environments. Jumbune supports all active major branches of Apache Hadoop namely 1.x, 2.x, 0.23.x and commercial MapR, HDP 2.x and CDH 5.x distributions of Hadoop. It has the ability to work well with both Yarn and non-Yarn versions of Hadoop. It has four major modules MapReduce Debugger, HDFS Data Validator, On-demand cluster monitor and MapReduce job profiler. Jumbune can be deployed on any remote user machine and uses a lightweight agent on the NameNode of the cluster to relay relevant information to and fro. | [1. Jumbune](https://jumbune.org)  [2. Jumbune GitHub Project](https://github.com/impetus-opensource/jumbune)  [3. Jumbune JIRA page](http://jumbune.org/jira/secure/Dashboard.jspa) |
| Spring XD | Spring XD (Xtreme Data) is a evolution of Spring Java application development framework to help Big Data Applications by Pivotal. SpringSource was the company created by the founders of the Spring Framework. SpringSource was purchased by VMware where it was maintained for some time as a separate division within VMware. Later VMware, and its parent company EMC Corporation, formally created a joint venture called Pivotal. Spring XD is more than development framework library, is a distributed, and extensible system for data ingestion, real time analytics, batch processing, and data export. It could be considered as alternative to Apache Flume/Sqoop/Oozie in some scenarios. Spring XD is part of Pivotal Spring for Apache Hadoop (SHDP). SHDP, integrated with Spring, Spring Batch and Spring Data are part of the Spring IO Platform as foundational libraries. Building on top of, and extending this foundation, the Spring IO platform provides Spring XD as big data runtime. Spring for Apache Hadoop (SHDP) aims to help simplify the development of Hadoop based applications by providing a consistent configuration and API across a wide range of Hadoop ecosystem projects such as Pig, Hive, and Cascading in addition to providing extensions to Spring Batch for orchestrating Hadoop based workflows. | [1. Spring XD on GitHub](https://github.com/spring-projects/spring-xd) |
| Cask Data Application Platform | Cask Data Application Platform is an open source application development platform for the Hadoop ecosystem that provides developers with data and application virtualization to accelerate application development, address a range of real-time and batch use cases, and deploy applications into production. The deployment is made by Cask Coopr, an open source template-based cluster management solution that provisions, manages, and scales clusters for multi-tiered application stacks on public and private clouds. Another component is Tigon, a distributed framework built on Apache Hadoop and Apache HBase for real-time, high-throughput, low-latency data processing and analytics applications. | [1. Cask Site](http://cask.co/) |
| **Categorize Pending ...** | | |
| Twitter Summingbird | A system that aims to mitigate the tradeoffs between batch processing and stream processing by combining them into a hybrid system. In the case of Twitter, Hadoop handles batch processing, Storm handles stream processing, and the hybrid system is called Summingbird. | TODO |
| Apache Kiji | Build Real-time Big Data Applications on Apache HBase. | TODO |
| S4 Yahoo | S4 is a general-purpose, distributed, scalable, fault-tolerant, pluggable platform that allows programmers to easily develop applications for processing continuous unbounded streams of data. | TODO |
| Metamarkers Druid | Realtime analytical data store. | TODO |
| Concurrent Cascading | Application framework for Java developers to simply develop robust Data Analytics and Data Management applications on Apache Hadoop. | TODO |
| Concurrent Lingual | Open source project enabling fast and simple Big Data application development on Apache Hadoop. project that delivers ANSI-standard SQL technology to easily build new and integrate existing applications onto Hadoop | TODO |
| Concurrent Pattern | Machine Learning for Cascading on Apache Hadoop through an API, and standards based PMML | TODO |
| Apache Giraph | Apache Giraph is an iterative graph processing system built for high scalability. For example, it is currently used at Facebook to analyze the social graph formed by users and their connections. Giraph originated as the open-source counterpart to Pregel, the graph processing architecture developed at Google | TODO |
| Talend | Talend is an open source software vendor that provides data integration, data management, enterprise application integration and big data software and solutions. | TODO |
| Akka Toolkit | Akka is an open-source toolkit and runtime simplifying the construction of concurrent applications on the Java platform. | TODO |
| Eclipse BIRT | BIRT is an open source Eclipse-based reporting system that integrates with your Java/Java EE application to produce compelling reports. | TODO |
| Spango BI | SpagoBI is an Open Source Business Intelligence suite, belonging to the free/open source SpagoWorld initiative, founded and supported by Engineering Group. It offers a large range of analytical functions, a highly functional semantic layer often absent in other open source platforms and projects, and a respectable set of advanced data visualization features including geospatial analytics | TODO |
| Jedox Palo | Palo Suite combines all core applications — OLAP Server, Palo Web, Palo ETL Server and Palo for Excel — into one comprehensive and customisable Business Intelligence platform. The platform is completely based on Open Source products representing a high-end Business Intelligence solution which is available entirely free of any license fees. | TODO |
| Twitter Finagle | Finagle is an asynchronous network stack for the JVM that you can use to build asynchronous Remote Procedure Call (RPC) clients and servers in Java, Scala, or any JVM-hosted language. | TODO |
| Intel GraphBuilder | Library which provides tools to construct large-scale graphs on top of Apache Hadoop | TODO |
| Apache Tika | Toolkit detects and extracts metadata and structured text content from various documents using existing parser libraries. | TODO |
| Apache Zeppelin | Zeppelin is a modern web-based tool for the data scientists to collaborate over large-scale data exploration and visualization projects. It is a notebook style interpreter that enable collaborative analysis sessions sharing between users. Zeppelin is independent of the execution framework itself. Current version runs on top of Apache Spark but it has pluggable interpreter APIs to support other data processing systems. More execution frameworks could be added at a later date i.e Apache Flink, Crunch as well as SQL-like backends such as Hive, Tajo, MRQL. | [1. Apache Zeppelin site](https://zeppelin.incubator.apache.org/) |

Published with [GitHub Pages](http://pages.github.com) by [Javi Roman](http://es.linkedin.com/in/javiroman/), and [contributors](https://github.com/hadoopecosystemtable/hadoopecosystemtable.github.io/graphs/contributors)

<https://hadoopecosystemtable.github.io/>