Big Data Solutions Reference Glossary (14 pages)

Very brief descriptions and links are listed here to provide starting point references for the multitude of Big Data solutions. No endorsements implied. Collected by Bob Marcus

Accumulo - (Database, NoSQL, Key-Value) from Apache
http://accumulo.apache.org/

Acunu Analytics - (Analytics Tool) on top of Aster Data Platform based on Cassandra
http://www.acunu.com/acunu-analytics.html

Aerospike - (Database NoSQL Key-Value)
http://www.aerospike.com/

Alteryx - (Analytics Tool)
http://www.alteryx.com/

Ambari - (Hadoop Cluster Management) from Apache
http://incubator.apache.org/ambari/

Analytica - (Analytics Tool) from Lumina
http://www.lumina.com/why-analytica/

ArangoDB - (Database, NoSQL, Multi-model) Open source from Europe
http://www.arangodb.org/2012/03/07/avocadodbs-design-objectives

Aster - (Analytics) Combines SQL and Hadoop on top of Aster MPP Database
http://www.asterdata.com/

Asterix - (Database for unstructured data) built on top of Hyracks from UCI
http://asterix.ics.uci.edu/

Avro - (Data Serialization) from Apache
http://en.wikipedia.org/wiki/Apache_Avro

Ayasdi Iris - (Machine Learning) uses Topological Data Analysis
http://www.ayasdi.com/product/

Azkaban - (Process Scheduler) for Hadoop
http://bigdata.globant.com/?p=441

Azure Table Storage - (Database, NoSQL, Columnar) from Microsoft
http://msdn.microsoft.com/en-us/library/windowsazure/jj553018.aspx

Behemoth - (Large-scale document processing platform) from Apache
http://www.findbestopensource.com/product/behemoth

Berkeley DB - (Database)
http://www.oracle.com/technetwork/products/berkeleydb/overview/index.html

BigData Appliance - (Integrated Hardware and Software Architecture) from Oracle includes Cloudera, Oracle NoSQL ,Oracle R and Sun Servers
http://nosql.mypopescu.com/post/15729871938/comparing-hadoop-appliances-oracles-big-data

BigML - (Analytics tool) for business end-users
https://bigml.com/

BigQuery - (Query Tool) on top of Google Storage
https://cloud.google.com/products/big-query

BigSheets - (BI Tool) from IBM
http://www-01.ibm.com/software/ebusiness/jstart/downloads/BigSheetsOverview.pdf

BigTable - (Database, NOSQL. Column oriented) from Google
http://en.wikipedia.org/wiki/BigTable

Caffeine - (Search Engine) from Google use BigTable Indexing
http://googleblog.blogspot.com/2010/06/our-new-search-index-caffeine.html

Cascading - (Processing) Java APIs on top of Hadoop from Concurrent
http://www.cascading.org/

Cascalog - (Query Tool using Clojure) on top of Hadoop
http://nathanmarz.com/blog/introducing-cascalog-a-clojure-based-query-language-for-hado.html

Cassandra - (Database, NoSQL, Column oriented)
http://cassandra.apache.org/

Chukwa - (Monitoring Hadoop Clusters) from Apache
http://incubator.apache.org/chukwa/

Clojure - (Lisp-based Programming Language) compiles to JVM byte code
http://clojure.org/

Cloudant - (Distributed Database as a Service)
https://cloudant.com/

Cloudera - (Hadoop Distribution) including real-time queries
http://www.cloudera.com/content/cloudera/en/home.html

Clustrix - (NewSQL DB) runs on AWS
http://www.clustrix.com/

Coherence - (Data Grid/Caching) from Oracle
http://www.oracle.com/technetwork/middleware/coherence/overview/index.html

Colossus - (File System) Next Generation Google File System
http://www.highlyscalablesystems.com/3202/colossus-successor-to-google-file-system-gfs/

Continuity - (Data fabric layer) Interfaces to Hadoop Processing and data stores
http://www.continuuity.com/

Corona - (Hadoop Processing tool) used internally by Facebook and now open sourced
http://gigaom.com/2012/11/08/facebook-open-sources-corona-a-better-way-to-do-webscale-hadoop/

Couchbase - (Database, NoSQL, Document) with CouchDB and Membase capabilities
http://www.couchbase.com/

CouchDB - (Database, NoSQL, Document)
http://couchdb.apache.org/

Customer Experience Digital Data Acquisition - Possible future standards from W3C
http://www.w3.org/community/custexpdata/wiki/images/9/93/W3C_CustomerExperienceDigitalDataAcquisition_Draft_v0.5.pdf

Data Tamer - (Data integration and curation tools) from MIT
http://www.cidrdb.org/cidr2013/Papers/CIDR13_Paper28.pdf

Datameer - (Analytics) built on top of Hadoop
http://www.datameer.com/

Datastax - (Integration) Built on Cassandra, Solr, Hadoop
http://www.datastax.com/

DeepDB - (Database) Supports SQL and NoSQL APIs
http://deep.is/a-new-approach-to-information-theory/

Dremel - (Query Tool) interactive for columnar DBs from Google
http://research.google.com/pubs/pub36632.html

Drill - (Query Tool) interactive for columnar DBs from Apache
http://en.wikipedia.org/wiki/Apache_Drill

Dryad - (Parallel execution environment) from Microsoft Research
http://research.microsoft.com/en-us/projects/dryad/default.aspx

Dynamo DB - (Database, NoSQL, Key-Value)
http://aws.amazon.com/dynamodb/

Elastic MapReduce - (Cloud-based MapReduce) from Amazon
http://aws.amazon.com/elasticmapreduce/

ElasticSearch - (Search Engine) on top of Apache Lucerne
http://www.elasticsearch.org/

Enterprise Control Language (ECL) - (Data Processing Language) from HPPC
http://hpccsystems.com/download/docs/ecl-language-reference

Erasure Codes - (Alternate to file replication for availability) Replicates fragments.
http://oceanstore.cs.berkeley.edu/publications/papers/pdf/erasure_iptps.pdf

EUDAT(European Data Infrastructure) - European Collaborative Initiative
http://www.eudat.eu/

eXtreme Scale - (Data Grid/Caching) from IBM
http://www-03.ibm.com/software/products/us/en/websphere-extreme-scale/

F1 - (Combines relational and Hadoop processing) from Google built on Google Spanner http://research.google.com/pubs/pub38125.html

Falcon - (Data processing and management platform) from Apache
http://wiki.apache.org/incubator/FalconProposal

Flume - (Data Collection, Aggregation, Movement)
http://flume.apache.org/

FlumeJava - (Java Library) Supports development and running data parallel pipelines
http://pages.cs.wisc.edu/~akella/CS838/F12/838-CloudPapers/FlumeJava.pdf

Fusion-io - (SSD Storage Platform) can be used with HBase
http://www.fusionio.com/company/

GemFire - (Data Grid/Caching) from VMware
https://www.vmware.com/products/application-platform/vfabric-gemfire/overview.html

Gensonix - (NoSQL database) from Scientel
http://scientel.com/platform.html

Gephi - (Visualization Tool) for Graphs
https://gephi.org/features/

Gigaspaces - (Data Grid/Caching)
http://www.gigaspaces.com/

Giraph - (Graph Processing) from Apache
http://giraph.apache.org/

Google Refine - (Data Cleansing)
http://code.google.com/p/google-refine/

Google Storage - (Database, NoSQL, Key-Value)
https://developers.google.com/storage/

Graphbase - (Database, NoSQL, Graphical)
http://graphbase.net/

Greenplum - (MPP Database. Analytic Tools, Hadoop)
http://www.greenplum.com/

Grunt Shell - (Interactive Shell for Apache Pig)
http://pig.apache.org/docs/r0.7.0/setup.html#Grunt+Shell

Guavas - (Stream Analytics)
http://www.guavus-new.com/solutions/

H20 - (Math toolkit) runs on topof Hadoop
http://0xdata.com/h2o/

HBase - (Database, NoSQL, Column oriented)
http://en.wikipedia.org/wiki/HBase

Hadapt - (Combined SQL Layer and Hadoop)
http://hadapt.com/

Hadoop Distributed File System - (Distributed File System) from Apache
http://hadoop.apache.org/docs/stable/hdfs_design.html

Hama - (Processing Framework) Uses BSP model on top of HDFS
http://hama.apache.org/

Hana - (Database, NewSQL) from SAP
http://en.wikipedia.org/wiki/SAP_HANA

Haven - (Analytics Package) from HP
http://www.itworldcanada.com/news/hp-unveils-haven-for-big-data/147217

HAWQ - (SQL Interface to Hadoop) from Greenplum and Pivotal
http://www.greenplum.com/blog/dive-in/hawq-the-new-benchmark-for-sql-on-hadoop

HCatalog - (Table and Storage Management) for Hadoop data
http://incubator.apache.org/hcatalog/

HDF5- (A data model, library, and file format for storing/managing large complex data)
http://www.hdfgroup.org/HDF5/

High Performance Computing Cluster (HPCC) - (Big Data Processing Platform)
http://hpccsystems.com/why-hpcc

Hive - (Data warehouse structure on top of Hadoop)
http://en.wikipedia.org/wiki/Apache_Hive

HiveQL - (SQL-like interface on Hadoop File System)
https://www.inkling.com/read/hadoop-definitive-guide-tom-white-3rd/chapter-12/hiveql

Hortonworks - (Extensions of Hadoop)
http://hortonworks.com/

HStreaming - (Real time analytics on top of Hadoop)
http://www.hstreaming.com/

Hue - (UI and Web applications for Cloudera Hadoop) from Cloudera
http://www.cloudera.com/content/cloudera-content/cloudera-docs/CDH4/4.2.0/Hue-2-User-Guide/hue2.html

Hypertable - (Database, NoSQL, Key-Value) open source runs on multiple file systems
http://hypertable.org/

Hyracks - **(**Parallel Data Processing) from UCI runs Pregelix (Pregel API), Hiivesterix (Hive API), Algebrix (algebra), Iterative Map-Reduce Update (IMRU) and Hyracks jobs
www.ics.uci.edu/~rares/pub/icde11-borkar.pdf

Impala - (Ad hoc query capability for Hadoop) from Cloudera
http://blog.cloudera.com/blog/2012/10/cloudera-impala-real-time-queries-in-apache-hadoop-for-real/

Indexed Database API - (Possible future standards for NoSqL queries) from W3C
http://www.w3.org/TR/2013/CR-IndexedDB-20130704/

InfiniDB - (Scale-up analytic database)
http://infinidb.org/

Infochimps - (Big Data Storage and Analytics in the Cloud)
http://www.infochimps.com/

Infosphere Big Insights - (Analytic) from IBM
http://www-01.ibm.com/software/data/infosphere/biginsights/

InnoDB - (Default storage engine for MYSQL)
http://en.wikipedia.org/wiki/InnoDB

iRODS - (integrated Rule-Oriented Data System)open source from iRODS Consortium
http://www.irods.org

Jaql = (Query Language for Hadoop) from IBM
http://www-01.ibm.com/software/data/infosphere/hadoop/jaql/

JSONIQ - (Query language for JSON)
http://www.jsoniq.org/

JustOneDB - (Relational DB) with flexible capabilities for Big Data
http://www.justonedb.com/

Kafka - (Publish-and-subscribe for data) from Apache
http://kafka.apache.org/

Karmasphere - (Analytics)
http://www.karmasphere.com/

Knox - (Secure gateway to Hadoop) from Apache
http://knox.incubator.apache.org/

LightFlow - (Data fabric middleware) from Lightwolf Technologies
http://www.lightwolftech.com/index.php?page=product-overview

Lingual - (SQL queries to Hadoop)
http://www.cascading.org/lingual/

Lucidworks - (Search built on Solr/Lucene) and an associated Big Data Platform
http://www.lucidworks.com/

Knowledge Graph - (Graphical data store) from Google
http://en.wikipedia.org/wiki/Knowledge_Graph

Mahout - (Machine Learning Toolkit) built on Apache Hadoop
http://en.wikipedia.org/wiki/Knowledge_Graph

MapD - (Massive Parallel Database) Open Source on top of GPUs
http://istc-bigdata.org/index.php/mapd-a-way-to-map-big-data-faster/

MapReduce - (Processing algorithm)
http://en.wikipedia.org/wiki/MapReduce

MapR - (MapReduce extensions) built on NFS
http://en.wikipedia.org/wiki/Knowledge_Graph

MarkLogic - (Database, NoSQL, Document) interfaced with Hadoop
http://www.marklogic.com/

Memcached - (Data Grid/Caching)
http://en.wikipedia.org/wiki/Memcached

MemSQL - (In memory analytics database)
http://www.memsql.com/

MongoDB - (Database, NoSQL, Document) from 10gen
http://www.mongodb.org/

mrjob - (Workflow) for Hadoop from Yelp
http://bighadoop.wordpress.com/2012/04/15/yelps-mrjob-a-python-package-for-hadoop-jobs/

MRQL - (Query Language) supports Map-Reduce and BSP processing
http://code.google.com/p/mrql/

Muppet - (Stream Processing) MapUpdate implementation
http://arxiv.org/pdf/1208.4175.pdf

MySql - (Database Relational)
http://www.mysql.com/

Namenode - Directory service for Hadoop
http://wiki.apache.org/hadoop/NameNode

NCDS (National Consortium for Data Sciences) - US Collaborative Data Initiative
http://data2discovery.org/

Neo4j - (Database, NoSQL, Graphical)
http://www.neo4j.org/

Netezza - (Database Appliance)
http://www-01.ibm.com/software/data/netezza/

NuoDB - (MPP Database)
http://www.nuodb.com/

Oozie - (Workflow Scheduler for Hadoop) from Apache
http://oozie.apache.org/

Oracle NoSQL - (Database, Key-Value)
http://www.oracle.com/technetwork/products/nosqldb/overview/index.html

ORC (Optimized Row Columnar) Files - File Format for Hive data in HDFS
http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.0.0.2/ds_Hive/orcfile.html

Parquet - (Columnar file format for Hadoop) from Cloudera
http://blog.cloudera.com/blog/2013/03/introducing-parquet-columnar-storage-for-apache-hadoop/

ParStream - (Real-Time Processing) claims high-performance
http://www.parstream.com/product/technology/

Pattern - (Machine Learning on top of Cascading)
http://www.cascading.org/pattern/

Pentaho - (Analytic tools)
http://www.pentaho.com/

Percolater - (Data Processing) from Google
http://research.google.com/pubs/pub36726.html

Pig - (Procedural framework on top of Hadoop)
http://pig.apache.org/

Pig Latin - (Interface language for Pig procedures)
http://pig.apache.org/docs/r0.7.0/piglatin_ref1.html

Pivotal - (New company utilizing VMware and EMC technologies)
http://www.gopivotal.com/

Platfora - (In memory caching for BI on top of Hadoop)
http://www.platfora.com/

Postgres - (Database Relational)
http://www.postgresql.org/

Precog - (Analytics Tool) for JSON data
http://precog.com/

Pregel - (Graph Processing) used by Google
http://kowshik.github.io/JPregel/pregel_paper.pdf

Presto - (SQL Query for HDFS) from Facebook
http://www.datanami.com/datanami/2013-06-07/big_data_big_five.html

Protocol Buffers - (Serialization) from Google)
http://en.wikipedia.org/wiki/Protocol_Buffers

Protovis - (Visualization)
http://mbostock.github.io/protovis/

PureData - (Database Products) from IBM
http://www-01.ibm.com/software/data/puredata/

R - (Data Analysis Tool)
http://en.wikipedia.org/wiki/R_%28programming_language%29

Rainstor - (Combines Hadoop and Relational Processing)
http://rainstor.com/

RCFile (Record Columnar File) - File format optimized for HDFS data warehouses
http://en.wikipedia.org/wiki/RCFile

Redis - (Database, NoSQL, Key-Value)
http://redis.io/

Redshift - (Database Relational) Amazon
http://aws.amazon.com/redshift/

Resilient Distributed Datasets - (Fauklt tolerant in memory data sharing)
http://www.cs.berkeley.edu/~matei/papers/2011/tr_spark.pdf

Riak - (Database, NoSQL,Key-Value with built-in MapReduce) from Basho
http://basho.com/riak/

Roxie - (Query processing cluster) from HPCC
http://hpccsystems.com/FAQ/what-roxie

RushAnalytics - (Analytics) from Pervasive
http://bigdata.pervasive.com/Products/Big-Data-Analytics-RushAnalytics.aspx

S3 - (Database, NoSQL,Key-Value)
http://en.wikipedia.org/wiki/R_%28programming_language%29

S4 - (Stream Processing) from Apache
http://incubator.apache.org/s4/

Sawzall - (Query Language for Map-Reduce) from Google
http://en.wikipedia.org/wiki/Sawzall_%28programming_language%29

Scala - (Programming Language) Combines functional and imperative programming
http://www.scala-lang.org/

Scalding - (Scala layer on top of Cascading)
http://polyglotprogramming.com/papers/ScaldingForHadoop.pdf

Scalebase - (Scalable Front-end to distributed Relational Databases)
http://www.scalebase.com/

Scaleout Stateserver - (In memory data grid) from Scaleout Software
http://www.scaleoutsoftware.com/

Scaleout hServer - (In memory data grid integrated with Hadoop) from Scaleout
http://www.scaleoutsoftware.com/products/scaleout-hserver

SciDB - (Database, NoSQL, Arrays)
http://www.scidb.org/

scikit learn - (Machine Learning Toolkit) in Python
http://scikit-learn.org/stable/

Scribe - (Server for Aggregating Log Data) originally from Facebook may be inactive
http://en.wikipedia.org/wiki/Scribe_%28log_server%29

SequenceFiles - (File format) Binary key-value pairs
http://wiki.apache.org/hadoop/SequenceFile

Sentry - (Find Grained Security for Hadoop) from Cloudera
http://cloudera.com/content/cloudera/en/campaign/introducing-sentry.html

Shark - (Complex Analytics Platform) on top of Spark
https://amplab.cs.berkeley.edu/projects/shark-making-apache-hive-run-at-interactive-speeds/

Simba - (ODBC SQL Driver for Hive)
http://www.simba.com/Apache-Hadoop-Hive-ODBC-Driver-SQL-Connector.htm

SimpleDB - (Database, NoSQL, Document) from Amazon
http://aws.amazon.com/simpledb/

Skytree - (Analytics Server) provides machine learning capabilities
http://www.skytree.net/

Solr/Lucene - (Search Engine) from Apache
http://lucene.apache.org/solr/

Spotfire - (Stream processing tool) from TIBCO
http://spotfire.tibco.com/

Spanner - (Database, NewSQL) from Google
http://en.wikipedia.org/wiki/Spanner_%28database%29

Spark - (In memory cluster computing system)
http://spark-project.org/

Splunk - (Machine Data Analytics)
http://www.splunk.com/

Spring Data - (Data access tool for Hadoop and NoSQL) in Spring Framework
http://www.springsource.org/spring-data

SQLite - (Software library supporting server-less relational database)
http://www.sqlite.org/

SQLstream - (Streaming data analysis products)
http://www.sqlstream.com/

Sqoop - (Data movement) from Apache
http://en.wikipedia.org/wiki/Sqoop

Sqrrl - (Security and Analytics on top of Apache Accumulo)
http://www.sqrrl.com/

Stinger - (Optimized Hive Queries) from Hortonworks
http://hortonworks.com/blog/100x-faster-hive/

StorageHandler - (Storage driver) for Apache Hive to external data stores
https://cwiki.apache.org/confluence/display/Hive/StorageHandlers

Storm - (Stream Processing)
http://www.drdobbs.com/open-source/easy-real-time-big-data-analysis-using-s/240143874

Sumo Logic - (Log Analytics)
http://www.sumologic.com/

Tableau - (Visualization)
http://www.tableausoftware.com/

Tachyon - (File system) from Berkeley
http://strata.oreilly.com/2013/04/tachyon-open-source-distributed-fault-tolerant-in-memory-file-system.html

Talend - (Data Integration Product)
http://www.talend.com

Templeton - (REST interface to HCatalog and Hadoop interfaces) from Apache
http://people.apache.org/~thejas/templeton_doc_v1/

TempoDB - (Database, NoSQL, Time Series)
https://tempo-db.com/

Teradata Active EDW - (Database, Relational)
http://www.teradata.com/Active-Enterprise-Data-Warehouse/

Terracotta - (In memory data management)
http://terracotta.org/

Terraswarm - (Data Acquisition) Sensor Integration
http://www.terraswarm.org/

Tez - (Execution engine on top of Yarn) from Apache
http://hortonworks.com/blog/introducing-tez-faster-hadoop-processing/

Thor - (Filesystem and Processing Cluster) from HPCC Systems
http://hpccsystems.com/FAQ/what-thor

Thrift - (Framework for cross-language development)
http://thrift.apache.org/

Tika - (Toolkit for extracting metadata and content from documents) from Apache
http://tika.apache.org/

Tinkerpop - (Graph Database and Toolkit)
http://thrift.apache.org/

UIMA - (Unstructured Information Management Architecture) from OASIS, and Apache
http://en.wikipedia.org/wiki/UIMA

Vertica - (Database Relational)
http://www.vertica.com/

Voldemort - (Database, NoSQL, Key- Value)
http://www.project-voldemort.com/voldemort/

VoltDB - (Database NewSQL)
http://voltdb.com/

vSMP (Virtual SMP on a Cluster) Versatile SMP from ScaleMP
http://www.scalemp.com/media-hub/resources/white-papers/

Watson from IBM - (Analytic Framework)
http://www-03.ibm.com/innovation/us/watson/

WebHDFS - (REST API for Hadoop)
http://hadoop.apache.org/docs/r1.0.4/webhdfs.html

WEKA - (Machine Learning Toolkit) in Java
http://en.wikipedia.org/wiki/Weka_%28machine_learning%29

Wibidata - (Components for building Big Data applications)
http://www.wibidata.com/

YarcData - (Graph Analytics for Big Data)
http://www.yarcdata.com/

Yarn - (Next Generation Hadoop) from Apache
http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html

Yottamine - (Machine Learning Toolkit) Cloud-based
http://yottamine.com/

Zettaset Orchestrator - (Management and Security for Hadoop)
http://www.zettaset.com/platform.php

ZooKeeper - (Distributed Computing Management)
http://zookeeper.apache.org/