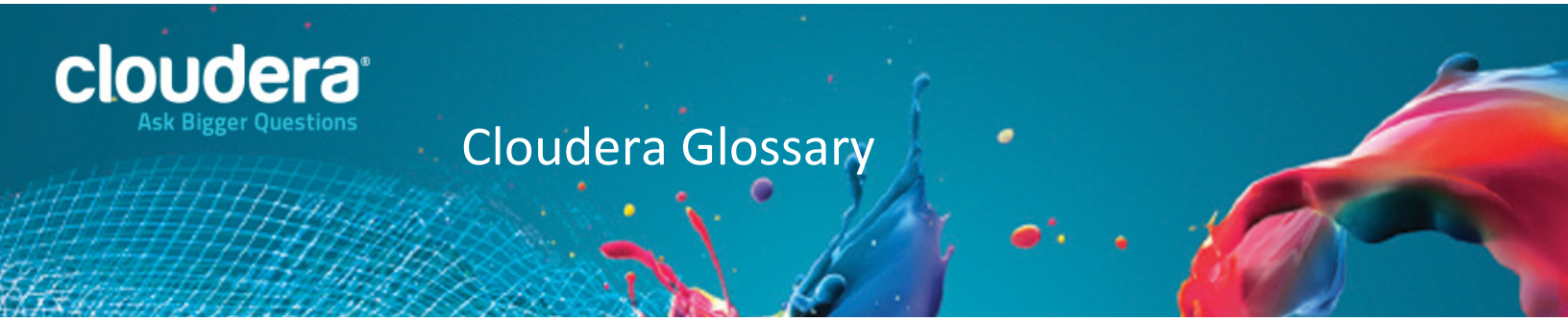

Cloudera Glossary

Cloudera, Inc.
220 Portage Avenue
Palo Alto, CA 94306
info@cloudera.com
US: 1-888-789-1488
Intl: 1-650-362-0488
www.cloudera.com

Important Notice

Date: November 20, 2012

Contents

This is a reference list of terms that arise frequently in discussions of Cloudera products and services. Additional information is available from a number of resources.

A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | R | S | T | W | Y | Z

Apache

See Apache Software Foundation.

Apache Avro

A serialization system for storing and transmitting data over a network. Avro supports rich data structures, a compact binary encoding, and a container file for sequences of Avro data (often referred to as "Avro data files"). Avro is language-independent and several language bindings are available for it, including Java, C, C++, Python, and Ruby. All components in CDH that produce or consume files support Avro data files as a file format.

Avro provides functionality similar to systems such as Apache Thrift and Protocol Buffers.

Apache Bigtop

A project to develop the packaging and interoperability testing of the Apache Hadoop ecosystem projects.

Apache Flume

A distributed, reliable, and available system for efficiently collecting, aggregating, and moving large amounts of text or streaming data from many different sources to a centralized data store.

Apache Giraph

A large-scale, fault-tolerant, graph processing framework that runs on Apache Hadoop.

Apache Hadoop

A free, open source software framework that supports data-intensive distributed applications. The core components of Apache Hadoop are the Hadoop Distributed File System and the MapReduce processing framework. The term is also used for an ecosystem of projects related to Hadoop that fall under the umbrella of infrastructure for distributed computing and large-scale data processing.

Apache HBase

A scalable, distributed, column-oriented data store. It provides real-time read/write random access to very large datasets hosted on HDFS.

Apache Hive

A data warehouse system for Hadoop that facilitates summarization and the analysis of large datasets stored in HDFS using an SQL-like language called HiveQL.

Apache Incubator

Apache Software Foundation gateway for open source projects that are aiming to become "top-level" Apache projects. Projects that are incubating are open source but may or may not become Apache projects.

Apache Mahout

A machine-learning library for Hadoop. It enables you to build machine-learning libraries that are scalable to large datasets, thus simplifying the task of building intelligent applications. The main use cases supported by Mahout are:

- Recommendation mining - identifies things users will like based on past preferences; for example, online shopping recommendations.
- Clustering - groups similar items; for example, documents that address similar topics
- Classification - learns what members of existing categories have in common then uses that information to categorize new items.
- Frequent item-set mining - takes a set of item-groups (such as items in a query session or shopping cart content) and identifies items that usually appear together.

Apache Maven

A software project management tool. Based on the concept of a project object model, Maven can manage a project's build, reporting, and documentation. CDH artifacts are available in the Cloudera Maven repository.

Apache Oozie

A workflow and coordination service to orchestrate data ingest, store, transform, and analysis actions.

Apache Pig

A dataflow language and parallel execution framework that is built on top of MapReduce. Internally, a compiler translates Pig statements into a directed acyclic graph of MapReduce jobs, which are submitted to Hadoop for execution.

Apache Software Foundation (ASF)

A non-profit corporation that supports various open source software products, including Apache Hadoop and related projects on which Cloudera products are based. Apache projects are developed by teams of collaborators and protected by an ASF license that provides legal protection to volunteers who work on Apache products and protect the Apache brand name.

Apache projects are characterized by a collaborative, consensus-based development process and an open and pragmatic software license. Each project is managed by a self-selected team of technical experts who are active contributors to the project.

Cloudera employees are major contributors to many Apache projects.

Apache Sqoop

A tool for efficiently transfering bulk data between Hadoop and external structured data stores such as relational databases. Using JDBC to interface with the database, Sqoop imports the contents of tables into HDFS, Apache Hive, and Apache HBase; it then generates Java classes that enable users to interpret the table's schema. Sqoop can also extract data from Hadoop storage and export records from HDFS to external structured datastores such as relational databases and enterprise data warehouses.

Apache Thrift

An interface definition language, runtime library, and a code generation engine to build services that can be invoked from many languages. Thrift can be used for serialization and RPC, but within Hadoop is mainly used for RPC.

Apache Whirr

A set of libraries for running applications on cloud services. Whirr can be used to run CDH clusters on services such as Amazon Elastic Compute Cloud (Amazon EC2); a working cluster starts immediately when the appropriate command is issued – it is not necessary to install the CDH packages in the cloud or do any configuration first. This is ideal for running temporary Hadoop clusters as proof-of-concept or training exercises. The cluster and all its data can be destroyed with a single command when it is no longer needed.

Apache Zookeeper

A centralized service for maintaining configuration information, naming, and providing distributed synchronization and group services.

Authentication

The function of confirming the identity of a person or software program.

Authorization

The function of specifying access rights to resources.

Avro

See Apache Avro.

Beeswax

A Hue application that enables you to perform queries on Apache Hive. You can create Hive tables, load data, run, and manage Hive queries.

Big Data

Data sets whose input/output velocity, variety of data structure, and volume is beyond the beyond the capabilities of systems which were designed assuming smaller data sets to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are an expanding target, currently ranging from terabytes to many petabytes of data in a single data set.

BigTable

A compressed, high performance, column-oriented data base built on Google File System (GFS). The BigTable design was the inspiration for Apache HBase, but the implementation, unlike other Google projects such as Protocol Buffers, is proprietary.

BigTop

See Apache BigTop.

CDH

Cloudera's Apache Hadoop distribution containing core Hadoop and the following related projects: Apache Avro, DataFu, Apache Flume, Fuse-DFS, Apache HBase, Apache Hive, Hue, Apache Mahout, Apache MRv1, Apache Oozie, Apache Pig, Apache Sqoop, Apache Whirr, Apache ZooKeeper.

CDH is free, 100% open source, and is licensed under the Apache 2.0 license. CDH is supported on many Linux distributions.

Cloudera Enterprise Free

A package of software (Cloudera Manager Free Edition and CDH) offered by Cloudera that enables data-driven enterprises to evaluate Apache Hadoop.

Cloudera Enterprise Core

A package of software (Cloudera Manager Enterprise Edition and CDH) and services offered by Cloudera that enables data-driven enterprises to run production Apache Hadoop environments cost effectively and with repeatable success.

Cloudera Enterprise RTD

An optional subscription module that can be added to Cloudera Enterprise Core. Cloudera Enterprise RTD offers management tools and technical support for Apache HBase.

Cloudera Enterprise RTQ

An optional subscription module that can be added to Cloudera Enterprise Core. Cloudera Enterprise RTQ offers management tools and technical support for Cloudera Impala.

Cloudera Impala

A set of services that enables real-time querying of data stored in HDFS or Apache HBase. It supports the same metadata, HiveQL query language, ODBC driver, and user interface (Beeswax) as Apache Hive. To avoid latency, Impala circumvents MapReduce to directly access the data through a specialized distributed query engine that is similar to those found in commercial parallel RDBMS.

Cloudera Manager

An end-to-end management application for CDH and Cloudera Impala. Cloudera Manager enables administrators to easily and effectively provision, monitor, and manage Hadoop clusters and CDH

installations. Cloudera Manager is available in two editions: Cloudera Manager Free Edition and Cloudera Manager Enterprise Edition.

Cluster, Hadoop

A set of computers or racks of computers that contains an HDFS file system and runs MapReduce and other processes on that data. A pseudo-distributed cluster is a CDH installation run on a single machine and useful for demonstrations and individual study.

Compression

A mechanism to reduce the size of a file so it takes up less disk space for storage and consumes less network bandwidth when transferred. The common compression tools used with Apache Hadoop include gzip, bzip2, Snappy, and LZO.

Connector

Usually refers to software for connecting external systems with Apache Hadoop. Some connectors work with Apache Sqoop to enable efficient data transfer between an external system and Hadoop. Other connectors translate ODBC driver calls from business intelligence systems into HiveQL queries.

The JDBC drivers supported by Cloudera Manager are also referred to as connectors.

Crunch

Java library that can be used to write, test, and run MapReduce pipelines. See Crunch.

Data science

A discipline that builds on techniques and theories from many fields, including mathematics, statistics, and computer science with the goal of extracting meaning from data and creating data products.

DataFu

A collection of Apache Pig user-defined functions (UDFs) for statistical analysis.

DataNode

See Hadoop Distributed File System.

Data store

A repository of a set of integrated information objects. Data stores include repositories like databases and flat files.

Distributed computing

A field of computer science that studies distributed systems.

Distributed system

A system composed of multiple autonomous computers that communicate through a computer network.

Extract, Load, Transform (ELT)

See Extract, Transform, Load (ETL).

Extract, Transform, Load (ETL)

A process that involves extracting data from sources, transforming the data to fit operational needs, and loading the data into the end target, typically a database or data warehouse.

Fault tolerant design

A design that enables a system to continue operation, possibly at a reduced level rather than failing completely, when some part of the system fails.

Flume

See Apache Flume.

Fuse-DFS

A service that allows HDFS to be mounted on Linux and accessed using standard filesystem tools.

Giraph

See Apache Giraph.

HA

See High availability.

Hadoop

See Apache Hadoop.

Hadoop Distributed File System (HDFS)

A user space filesystem designed for storing very large files with streaming data access patterns, running on clusters of industry-standard machines. HDFS defines two components:

- NameNode - maintains the namespace tree for HDFS and a mapping of file blocks to DataNodes where the data is stored. A simple HDFS cluster can have only one primary NameNode, supported by a secondary NameNode that periodically compresses the NameNode edits log file that contains a list of HDFS metadata modifications. This reduces the amount of disk space consumed by the log file on the NameNode, which also reduces the restart time for the primary NameNode. A High Availability cluster contains two NameNodes, one of which acts as hot standby.
- DataNode - stores data in a Hadoop cluster and is the name of the daemon that manages the data. File data is replicated on multiple DataNodes for reliability and so that localized computation can be executed near the data.

Hadoop User Group (HUG)

A club focused on the use of Hadoop technology.

Hadoop World

An industry conference for Apache Hadoop users, contributors, administrators, and application developers.

Hue

A platform for building custom GUI applications for CDH services and a tool containing the following built-in applications: an application for submitting MapReduce, Java, and streaming jobs, Apache Pig and Apache HBase shells, Beeswax, Oozie application editor, scheduler, and submitter, an HDFS file manager, and a MapReduce job browser.

HBase

See Apache HBase.

HBaseCon

An industry conference for Apache HBase users, contributors, administrators, and application developers.

HDFS

See Hadoop Distributed File System.

High availability (HA)

A system and implementation design to keep a service available at all times in the face of failures, without regard to its performance.

Hive

See Apache Hive.

HiveServer

A server process that supports clients that connect to Hive over an Apache Thrift connection.

HiveServer2

A server process that supports clients that connect to Hive over a network connection. These clients may be native command-line editors or applications and tools that use an ODBC or JDBC driver.

HiveQL

A query language for Hadoop that uses a syntax that is similar to standard SQL to execute MapReduce jobs on HDFS. HiveQL does not support all SQL functionality. Transactions and materialized views are not supported and support for indexes and subquery is limited. It supports features that are not part of standard SQL, such as multitable, including multitable inserts, and create table as select.

Internally, a compiler translates HiveQL statement into a directed acyclic graph of MapReduce jobs, which are submitted to Hadoop for execution. Beeswax, which is included in Hue, provides a graphical front-end for HiveQL queries.

Impala

See Cloudera Impala.

Incubator

See Apache Incubator.

Index

A data structure that improves the speed of data retrieval operations on a database table at the cost of slower writes and increased storage space. Indices can be created using one or more columns of a database table, providing the basis for both rapid random lookups and efficient access of ordered records.

JDBC driver

A client-side adapter that implements the JDBC Java programming language API for accessing relational database management systems.

JobTracker

See MapReduce v1 (MRv1).

Kerberos

A computer network authentication protocol that works on the basis of "tickets" to allow nodes communicating over an insecure network to prove their identity to one another in a secure manner. See the CDH4 Security Guide for information on which components support Kerberos.

Latency

A measure of time delay experienced in a system.

Linux

A Unix-like computer operating system assembled under the model of free, open source software development and distribution. Linux is a leading operating system on servers, mainframe computers, supercomputers, and embedded systems such as mobile phones, tablets, network routers, televisions, and video game consoles. The major distributors of enterprise Linux are CentOS, Debian, Red Hat, SuSE, and Ubuntu.

LZO

A free, open source compression library. LZO compression provides a good balance between data size and speed of compression. The LZO compression algorithm is the most efficient of the codecs, using very little CPU. Its compression ratios are not as good as others, but its compression is still significant compared to the uncompressed file sizes. Further, unlike some other formats, LZO-compressed files are splittable, enabling MapReduce to process splits in parallel.

LZO is published under the GNU General Public License and so is not included in CDH but can be used with CDH components; the Cloudera public Git repository hosts the hadoop-lzo package that provides a version of LZO that can be used with CDH.

Mahout

See Apache Mahout.

MapReduce

A distributed processing framework for processing and generating large data sets and an implementation that runs on large clusters of industry-standard machines.

The processing model defines two types of functions: a map function that processes a key-value pair to generate a set of intermediate key-value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key.

A MapReduce job partitions the input data set into independent chunks which are processed by the map functions in a parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce functions. Typically both the input and the output of the job are stored in a distributed filesystem.

The implementation provides an API for configuring and submitting jobs and job scheduling and management services, a library of search, sort, index, inverted index, and word co-occurrence algorithms, and the runtime. The runtime system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication.

MapReduce v1 (MRv1)

The runtime framework on which MapReduce jobs execute. It defines two daemons:

- Job tracker - coordinates running MapReduce jobs and provides resource management and job life-cycle management. In YARN, those functions are performed by two separate components.
- Task tracker - runs the tasks that the MapReduce jobs have been split into.

MapReduce v2 (MRv2)

See YARN.

Maven

See Apache Maven.

NameNode

See Hadoop Distributed File System.

ODBC driver

A client-side adapter that implements a standard C programming language API for accessing relational database management systems.

Oozie

See Apache Oozie.

Petabyte

10^{15} bytes. 1,000 terabytes or 1,000,000 gigabytes.

Pig

See Apache Pig.

RegionServer

In Apache HBase, applications store data into labeled tables, which are partitioned horizontally into regions. RegionServer is responsible for managing one or more regions.

Relational database management system (RDBMS)

A database management system based on the relational model. In the relational model, all data is represented in terms of tuples, grouped into relations. Most implementations of the relational model use the SQL data definition and query language.

Serialization

The process of converting a data structure or object state into a format that can be stored (for example, in a file or memory buffer, or transmitted across a network connection). Deserialization is the process of converting it back to the original state later in the same or another computer environment. See Apache Avro and Apache Thrift.

Snappy

A compression library. Snappy aims for very high speeds and reasonable compression rather than maximum compression or compatibility with other compression libraries. Snappy is provided in the Hadoop package along with the other native libraries (such as native gzip compression).

SQL

A declarative programming language designed for managing data in relational database management systems. Originally based upon relational algebra and tuple relational calculus, its scope includes data insert, query, update and delete, schema creation and modification, and data access control.

Sqoop

See Apache Sqoop.

TaskTracker

See MapReduce v1 (MRv1).

Terabyte

10^{12} bytes. 1,000 gigabytes.

Thrift

See Apache Thrift.

Whirr

See Apache Whirr.

YARN (Yet Another Resource Negotiator)

A general architecture for running distributed applications. YARN specifies the following components:

- Resource manager - manages the global assignment of compute resources to applications.
- Application master - manages the life cycle of applications
- Node manager - launches and monitors the compute containers on machines in the cluster

The application master negotiates with the resource manager for cluster resources – described in terms of a number of containers, each with a certain memory limit – and then runs application-specific processes in those containers. The containers are overseen by node managers running on cluster nodes, which ensure that the application does not use more resources than it has been allocated.

MapReduce v2 (MRv2) is implemented as a YARN application.

Zookeeper

See Apache Zookeeper.

Resources

Documentation

- Cloudera product documentation
- Cloudera web mirror
- Apache Hadoop documentation – The official documentation for Apache Hadoop technologies.

Books

- *Hadoop: The Definitive Guide* – A detailed overview of the Hadoop technologies, with an emphasis on information required to develop applications on Hadoop.
- *Hadoop Operations* – Provides in-depth information on running Hadoop in production, from planning, installing, and configuring the system, to providing ongoing maintenance.
- *HBase: The Definitive Guide*
- *Programming Pig*
- *Programming Hive*