cloudera

CDH PROJECTS AND SPECIFICATIONS

Proven, Enterprise-Ready, 100% Open Source Hadoop Distribution

CDH is the world's most complete and popular distribution combining Apache Hadoop, Spark, and Kafka for the enterprise, and powers Cloudera's modern platform for machine learning and analytics optimized for the cloud. Testing, packaging, and integration simplify the build-out of your machine learning and analytics deployment.

PROJECT	DESCRIPTION	BENEFIT
Apache Avro	A serialization system for storing and transmitting data over a network	Obtain better performance of Hadoop workloads and other workloads
Apache Flume	Distributed framework for collecting and aggregating log and event data and streaming it into HDFS or HBase in real-time	Load data into HDFS or HBase and make it available for operations immediately
Apache Hadoop	Reliable, scalable distributed storage and computing	Single solution for big data processing and management that improves operational efficiency and creates competitive advantage
Fuse-DFS	Module for mounting HDFS as a traditional file system	Read and write data to HDFS via standard network protocols like NFS
HDFS	Hadoop Distributed File System — scalable, distributed, fault tolerant data storage	Store any type and quantity of data in native format; make it available for processing, exploration, and analysis
MapReduce 2 (YARN)	The next generation of MapReduce framework	Improve scalability, resource management, and extensibility — give more granular control to developers; facilitate the use of additional processing frameworks
Apache HBase	Scalable record and table storage with real-time read/write access	Serve data to many users through fast, random read/write access
Apache HCatalog	A table and storage management service for data stored in Hadoop	Share schema and data types across data processing tools such as MapReduce, Hive, and Pig
Apache Hive	Metadata repository with SQL-like interface and ODBC/JDBC drivers for connecting BI applications to Hadoop	Make Hadoop more accessible to DBAs and analysts; define data structures through metadata and allow SQL queries to be executed
Hue	Browser-based desktop interface for Hadoop	Interact with Hadoop via an easy-to-use visual interface
Apache Impala	MPP Analytic SQL query engine for Hadoop	Enable BI analytics and data discovery via SQL and BI tools
Apache Kafka	Apache Kafka is publish-subscribe messaging rethought as a distributed commit log	Flexible, highly efficient central data bus for a large environment
Kite	A collection of Apache-licensed libraries, tools, and examples to simplify Hadoop application development	Expedite time to production by shortening development cycles
Apache Kudu	Storage engine that combines fast inserts/updates and efficient columnar scans	Enable fast analytics on changing data to power real-time analytic workloads
Apache Oozie	Workflow engine to coordinate Hadoop activities	Combine and schedule multiple jobs to form a complete, end-to-end data workflow
Apache Parquet	Apache-licensed, column-oriented file format	Achieve faster query times
Apache Pig	High-level data flow language for processing data stored in Hadoop	Create complex data processing pipelines using a high level procedural language
Apache Solr	Free text, fuzzy matching, and faceted search engine	Powerful, proven, robust search engine
Apache Sentry	Module that provides fine-grained, role-based authorization for Impala, Hive, Search, and HDFS	Ensure appropriate access permissions for users and groups

Apache Spark	Fast and general data processing engine that supports cyclic data flow and in-memory computing	Run programs up to 100x times faster than Hadoop MapReduce in memory
Apache Sqoop	Data transport engine for integrating Hadoop with relational databases	Copy data between Hadoop and relational database systems

PROJECT	CDH 5.16	CDH 6.2
Apache Avro	1.7.6	1.8.2
Apache Flume	1.7.0	1.9.0
Apache Hadoop	2.6.0	3.0.0
FUSE-DFS	2.6.0	3.0.0
HDFS	2.6.0	3.0.0
MapReduce 2 (YARN)	2.6.0	3.0.0
Apache HBase	1.2.0	2.1.2
Apache HCatalog	Included with Apache Hive	Included with Apache Hive
Apache Hive	1.1.0	2.1.1
Hue	4.2.0	4.3.0
Apache Impala	2.12.0	3.2.0
Apache Kafka	1.0.1	2.1.0
Apache Kudu	1.7.0	1.9.0
Kite	1.0.0	1.0.0
Apache Oozie	4.1.0	5.1.0
Apache Parquet	1.5.0	1.9.0
Apache Pig	0.12.0	0.17.0
Apache Solr	4.10.3	7.4
Apache Sentry	1.5.1	2.1.0
Apache Spark	1.6, 2.0, 2.1, 2.2, 2.3	2.4.0
Apache Sqoop	1.4.6	1.4.7
Apache Zookeeper	3.4.5	3.4.5

Cloudera, Inc. 395 Page Mill Road Palo Alto, CA 94306 USA cloudera.com