

Copyright © 2014 Tata Consultancy Services Limited



- Hadoop Distribution Platforms
- Hadoop Distribution Platforms Comparison

- These are packaged software products that aim to ease deployment and management of Hadoop clusters compared with simply downloading the various Apache code bases and trying to cobble together a system. Presently, Cloudera, Hortonworks, MapR offer their own Hadoop distributions.
- Many Hadoop distributions integrate with various data warehouses, databases and other data-management products, with the goal of moving data between Hadoop clusters and other environments so each might process or query data stored in the other.

Distribution Platforms

Companies offering commercial implementations or support for Hadoop

- ✓ Apache
- ✓ Cloudera
- ✓ Hortonworks
- ✓ MapR



cloudera®



Apache

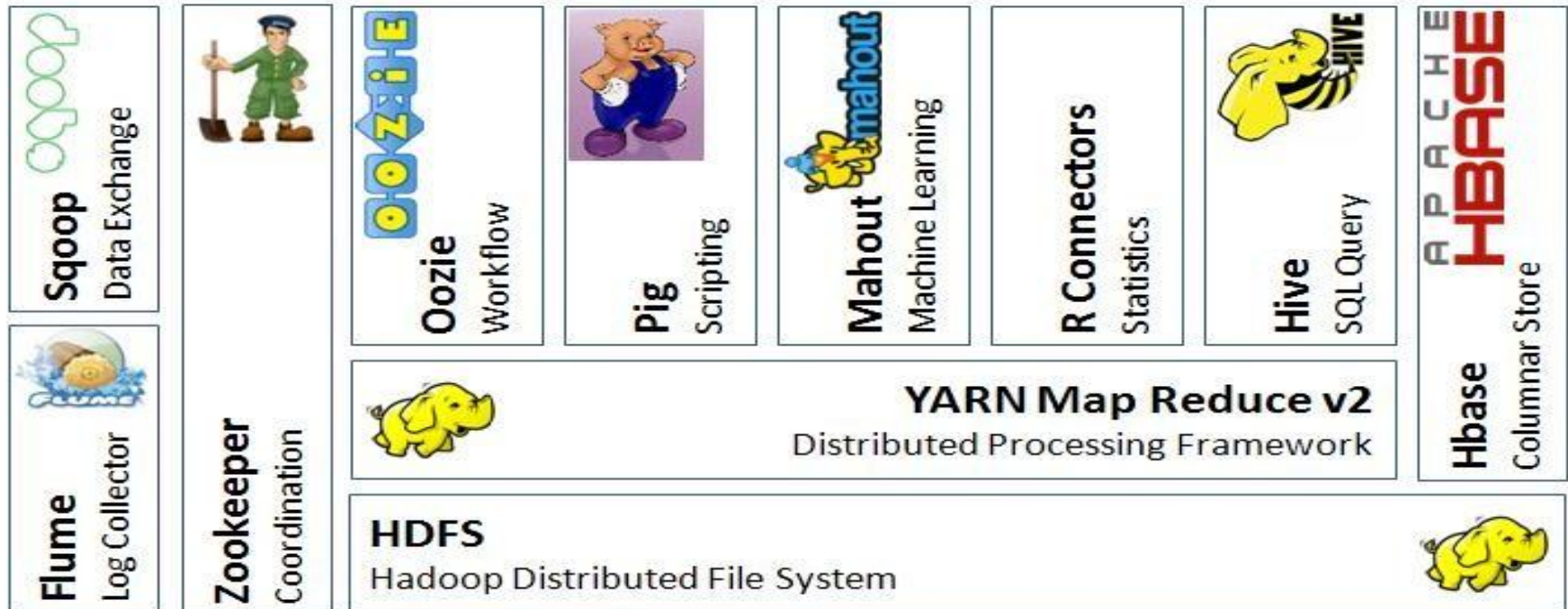
- Apache Hadoop is a software framework that supports data-intensive distributed applications under a free license
- It enables applications to work with thousands of nodes and petabytes of data
- Hadoop was inspired by Google's MapReduce and Google File System (GFS) papers
- Hadoop is a top-level Apache project being built and used by a global community of contributors, written in the Java programming language.
- Hadoop components:

Hive, Oozie, Pig, Zookeeper, Avro, Flume, HBase, Sqoop, Mahout, Whirr

Distribution Platforms - Apache



Apache Hadoop Ecosystem




Cloudera

- Offers enterprises a powerful new data platform built on the popular Apache Hadoop open-source software package
- Free edition: CDH
 - Integrated, tested distribution of Apache Hadoop
- Enterprise edition: Cloudera enterprise
 - Adds management software layer over CDH
- Hadoop components:
 - Hive, Oozie, Pig, Zookeeper, Avro, Flume, HBase, Sqoop, Mahout
- Serves a wide range of customers including retail, government, financial service, healthcare, life sciences, digital media, advertising, networking and telephony enterprises.

Distribution Platforms – Cloudera

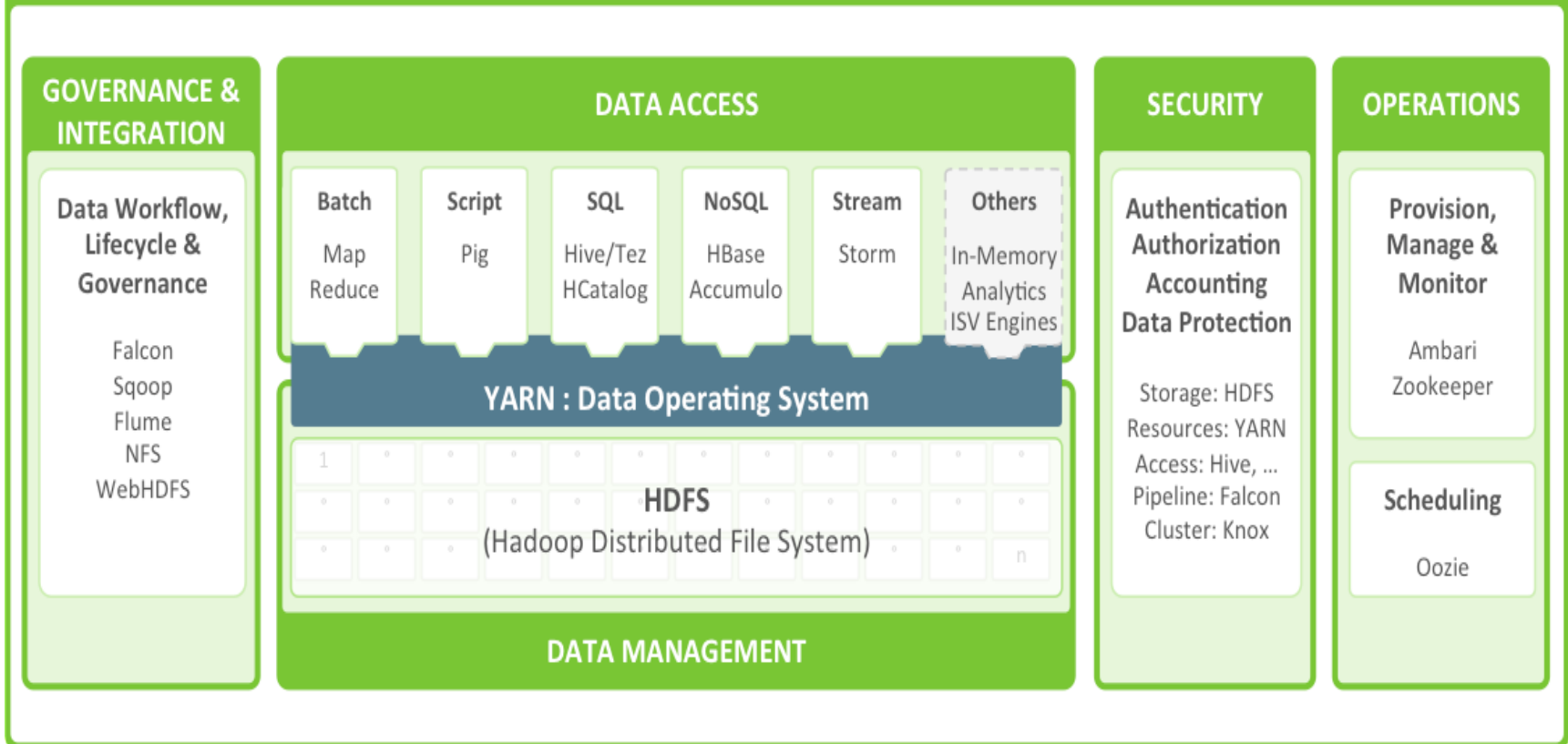
Cloudera's Distribution for Hadoop

UI Framework		Hue	SDK		Hue SDK	
Workflow		Oozie	Scheduling		Oozie	
			Metadata		Hive	
Data Integration	Languages, Compilers			Fast read/write access	HBase	
	Pig/ Hive					
Flume, Sqoop						
Coordination						Zookeeper

Hortonworks

- Focused on accelerating the development and adoption of Apache Hadoop and its ecosystem
- Making Hadoop more robust and easier to install, manage and use
- Hortonworks also provide support and training for Apache Hadoop.
- Hadoop components:
 - Hive, Pig, Zookeeper, HBase, Ambari

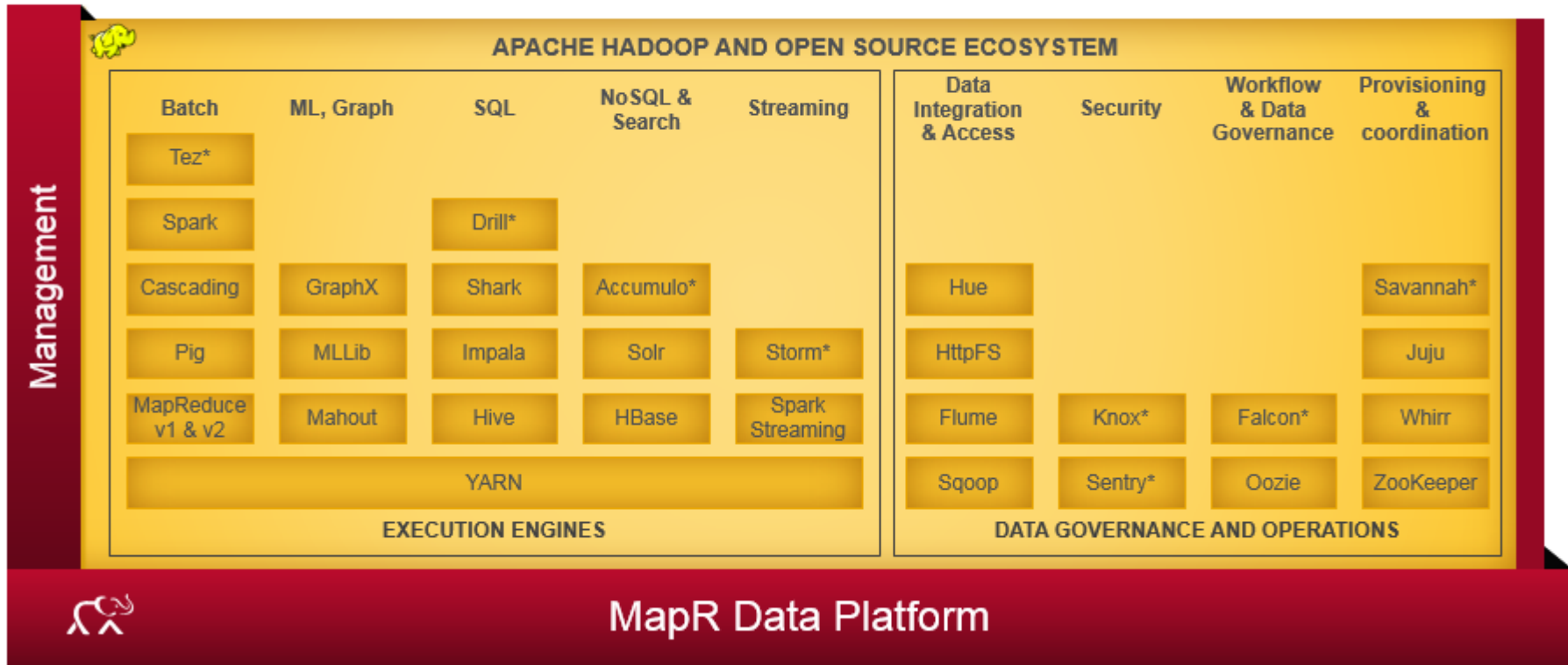
Hortonworks Data Platform



MapR

- MapR is a complete, industry-standard, Hadoop distribution with key improvements.
- MapR Hadoop includes the full family of Hadoop ecosystem components, such as HBase, Hive, Pig, and Flume, all of which have been tested together on specific platforms.
- The filesystem used on MapR clusters. MapR-FS is written in C/C++ and replaces the host operating system's file system, resulting in higher performance compared to HDFS, which runs in Java.
- A core-differentiating component of the MapR Distribution including Apache™ Hadoop® is the MapR File System, also known as MapR-FS.

Distribution Platforms - MapR



Hadoop Distribution Platforms Comparison

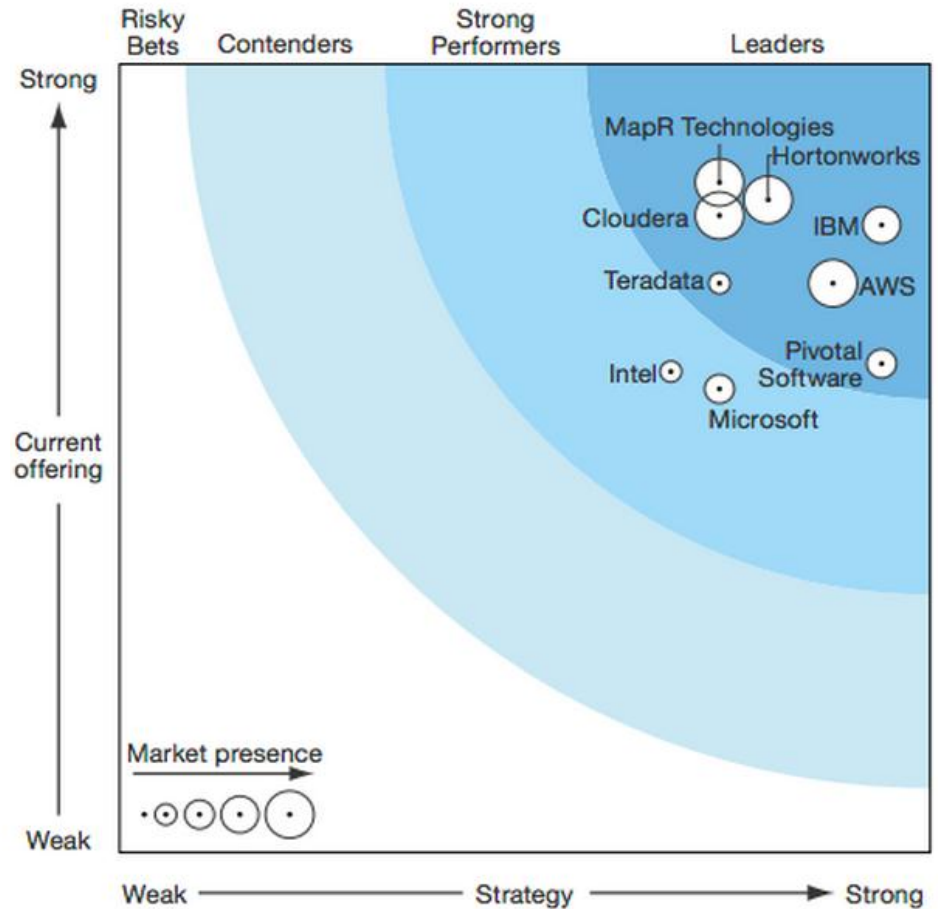
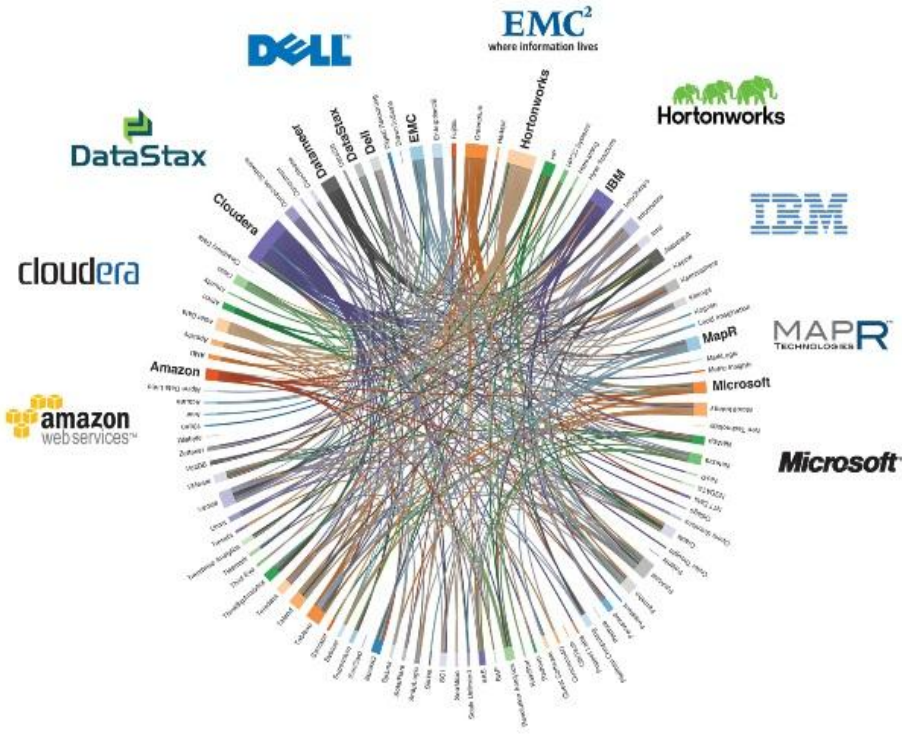
Hadoop Distribution Platforms - Comparison

Component (name and version)	Solution title	Hortonworks HDP	Cloudera CDH4.3	MapR M3 v3.0
FILE SYSTEM		HDFS 1.2.0	HDFS 2.0.0	MapR-FS
• non-Hadoop access		NFSv3	Fuse-DFS v2.0.0	Direct Access NFS
• Web access	REST HTTP API	WebHDFS	HttpFS	*
MAPREDUCE		1.2.0	0.20.2	**
• software abstraction layer	Cascading	x	x	2.1
NON-RELATIONAL DATABASE	Apache HBase	0.94.6.1	0.94.6	0.92.2
METADATA SERVICES	Apache HCatalog	***Hive	0.5.0	0.4.0
SCRIPTING PLATFORM	Apache Pig	0.11	0.11.0	0.10.0
• data analysis framework	DataFu	x	0.0.4	x
DATA ACCESS AND QUERY	Apache Hive	0.11.0	0.10.0	0.9.0
WORKFLOW SCHEDULER	Apache Oozie	3.3.2	3.3.2	3.2.0
CLUSTER COORDINATION	Apache Zookeeper	3.4.5	3.4.5	3.4(?)
BULK DATA TRANSFER BE- TWEEN RELATION- AL DATABASES AND HADOOP	Apache Sqoop	1.4.3	1.4.3	1.4.2

Hadoop Distribution Platforms - Comparison


Component (name and version)	Solution title	Hortonworks HDP	Cloudera CDH4.3	MapR M3 v3.0
DISTRIBUTED LOG MANAGEMENT SERVICES	Apache Flume	1.3.1	1.3.0	1.2.0
MACHINE LEARN- ING AND DATA ANALYSIS	Apache Mahout	0.7.0	0.7	0.7
HADOOP UI	Hue	2.2.0	2.3.0	-
• data integration service	Talend Open Stu- dio for Big Data	5.3	x	x
CLOUD SERVICES	Whirr	x	0.8.2	0.7.0
PARALLEL QUERY EXECUTION ENGINE		Tez (Stinger)	Impala	****
FULL-TEXT SEARCH	Search		0.1.5	
ADMINISTRATION		Apache Ambari	Cloudera Manager	MapR Control System
• installation		Apache Ambari	Cloudera Manager	-
• monitoring		Ganglia	x	x
		Nagios	x	x
• fine-grained authorization	Sentry		1.1	
NON-MAPREDUCE TASKS	YARN	2.0.4	2.0.0	-

Hadoop Distributors



- hadoop.apache.org/
- www.cloudera.com/hadoop/
- hortonworks.com/
- <http://gigaom.com/cloud/>
- <https://www.mapr.com/forrester-wave-hadoop-distribution-comparison-and-benchmark-report>

Thank You



TATA

Promise what we deliver.
Deliver what we promise. That's
certainty

Critical situations. Ruthless competition. Unforgiving customers. Thankfully you can be absolutely sure of your IT solutions with Tata Consultancy Services (TCS). As one of the world's fastest growing technology and business solutions providers, TCS has built a reputation of delivery excellence based on world class IT solutions that are on time, within budget and consistently deliver superior quality. So it comes as no surprise that we pioneered the concept of the Global Network Delivery Model, Developed Innovation Labs and Solution Accelerators. Achieving a level of delivery excellence that provides greater value to our customers and is the industry benchmark. Enabling our clients to experience certainty.

TATA CONSULTANCY SERVICES
Experience certainty.
IT Services • Business Solutions • Outsourcing

To learn how your business can experience certainty, visit www.tcs.com