

# HDP Overview: Apache Hadoop Essentials

Rev 3.0



## What is Apache Hadoop?

The Apache Hadoop project describes the technology as a software framework that:

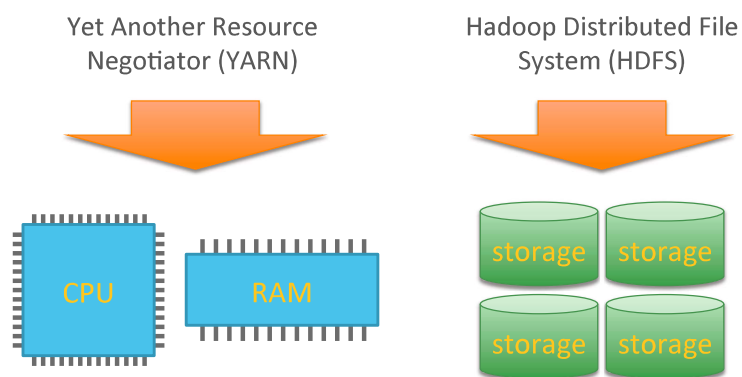
- ◆ Allows for the distributed processing of large data sets across clusters of computers using simple programming models
- ◆ Is designed to scale up from single servers to thousands of machines, each offering local computation and storage
- ◆ Does not rely on hardware to deliver high-availability, but rather the library itself is designed to detect and handle failures at the application layer
- ◆ Delivers a highly-available service on top of a cluster of computers, each of which may be prone to failures



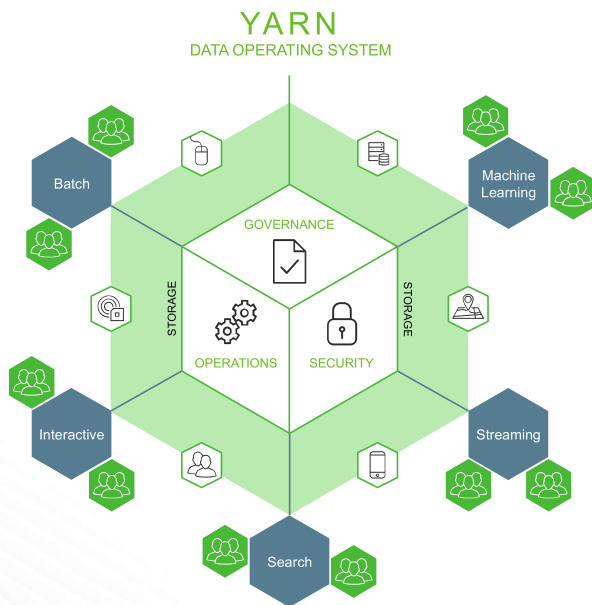
Source: <http://hadoop.apache.org>



## Hadoop Core = Storage + Compute



## Centralized Platform with YARN-Based Architecture



### Centralized Platform

for operations, governance and security

### Diverse Applications

run simultaneously on a single cluster

### Maximum Data Ingest

including existing and new sources, regardless of raw format

### Shared Big Data Assets

across business groups, functions and users



## Offering You the Most Flexibility

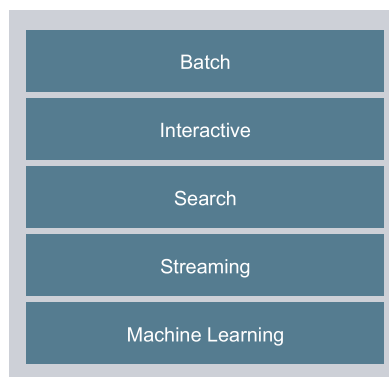
### ANY DATA

Existing and new datasets



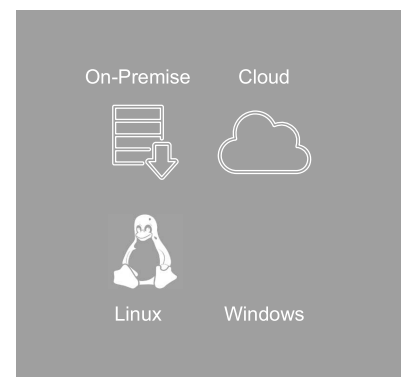
### ANY APPLICATION

Multiple engines for data analysis



### ANYWHERE

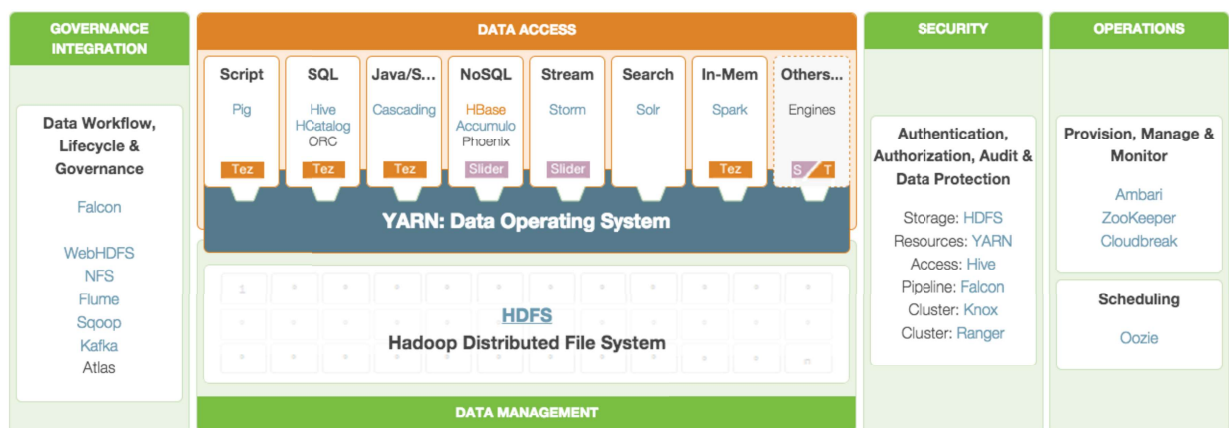
Complete range of deployment options



## The Hadoop Ecosystem



# Hortonworks Hadoop Distribution



## Data Management Frameworks

Framework	Description
Hadoop Distributed File System (HDFS)	A Java-based, distributed file system that provides scalable, reliable, high-throughput access to application data stored across commodity servers
Yet Another Resource Negotiator (YARN)	A framework for cluster resource management and job scheduling





## Operations Frameworks

Framework	Description
Ambari	A Web-based framework for provisioning, managing, and monitoring Hadoop clusters
ZooKeeper	A high-performance coordination service for distributed applications
Cloudbreak	A tool for provisioning and managing Hadoop clusters in the cloud
Oozie	A server-based workflow engine used to execute Hadoop jobs



## Data Access Frameworks

Framework	Description
Pig	A high-level platform for extracting, transforming, or analyzing large datasets
Hive	A data warehouse infrastructure that supports ad hoc SQL queries
HCatalog	A table information, schema, and metadata management layer supporting Hive, Pig, MapReduce, and Tez processing
Cascading	An application development framework for building data applications, abstracting the details of complex MapReduce programming
HBase	A scalable, distributed NoSQL database that supports structured data storage for large tables
Phoenix	A client-side SQL layer over HBase that provides low-latency access to HBase data
Accumulo	A low-latency, large table data storage and retrieval system with cell-level security
Storm	A distributed computation system for processing continuous streams of real-time data
Solr	A distributed search platform capable of indexing petabytes of data
Spark	A fast, general purpose processing engine use to build and run sophisticated SQL, streaming, machine learning, or graphics applications



## Governance and Integration Frameworks

Framework	Description
Falcon	A data governance tool providing workflow orchestration, data lifecycle management, and data replication services.
WebHDFS	A REST API that uses the standard HTTP verbs to access, operate, and manage HDFS
HDFS NFS Gateway	A gateway that enables access to HDFS as an NFS mounted file system
Flume	A distributed, reliable, and highly-available service that efficiently collects, aggregates, and moves streaming data
Sqoop	A set of tools for importing and exporting data between Hadoop and RDBM systems
Kafka	A fast, scalable, durable, and fault-tolerant publish-subscribe messaging system
Atlas	A scalable and extensible set of core governance services enabling enterprises to meet compliance and data integration requirements



## Security Frameworks

Framework	Description
HDFS	A storage management service providing file and directory permissions, even more granular file and directory access control lists, and transparent data encryption
YARN	A resource management service with access control lists controlling access to compute resources and YARN administrative functions
Hive	A data warehouse infrastructure service providing granular access controls to table columns and rows
Falcon	A data governance tool providing access control lists that limit who may submit Hadoop jobs
Knox	A gateway providing perimeter security to a Hadoop cluster
Ranger	A centralized security framework offering fine-grained policy controls for HDFS, Hive, HBase, Knox, Storm, Kafka, and Solr

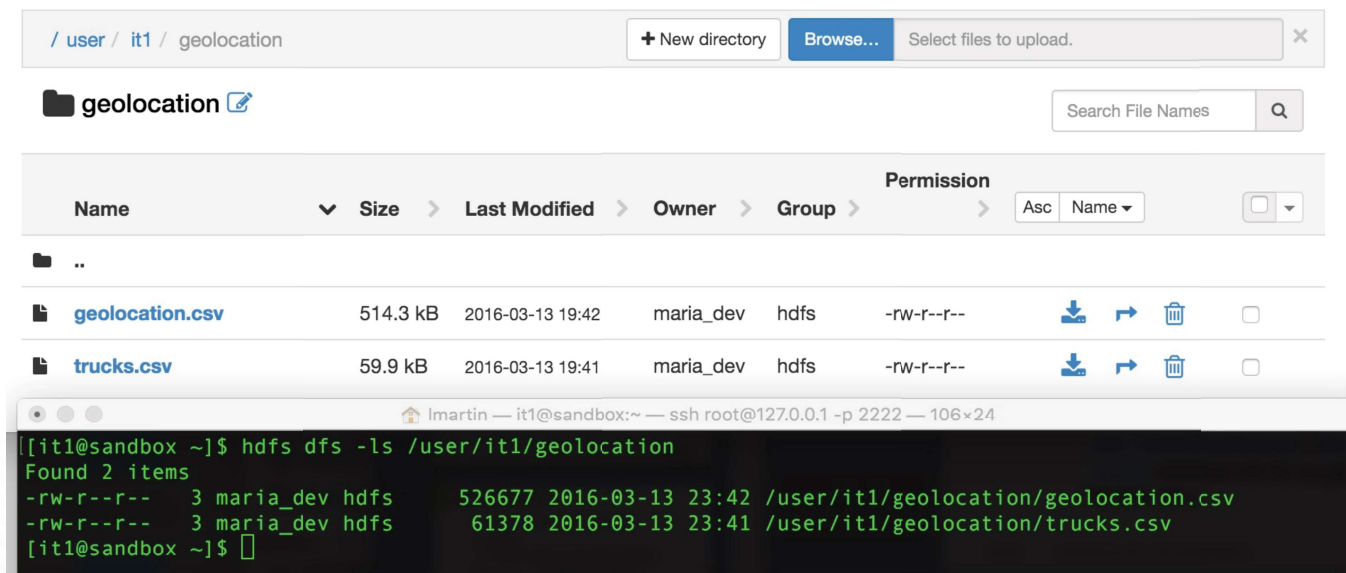


## Ecosystem Component Versions

Ongoing Innovation in Apache																						
HDP 2.4 Mar. 2016	2.7.1	0.15.0	1.2.1	0.7.0	5.2.1	1.6.0	0.80.0	1.1.2	4.4.0	1.7.0	0.10.0	0.6.1	0.5.0	1.4.6	1.5.2	0.9.0	2.2.1	1.0.0	3.4.6	4.2.0	0.6.0	0.5.0
HDP 2.3 Jul. 2015	2.7.1	0.15.0	1.2.1	0.7.0	5.2.1	1.3.1	0.80.0	1.1.1	4.4.0	1.7.0	0.10.0	0.6.1	0.5.0	1.4.6	1.5.2	0.8.2	2.1.0	1.0.0	3.4.6	4.2.0	0.6.0	0.5.0
HDP 2.2 Dec. 2014	2.6.0	0.14.0	0.14.0	0.5.2	4.10.2	1.2.1	0.80.0	0.98.4	4.2.0	1.6.1	0.9.3	0.6.0		1.4.5	1.5.2	0.8.1	2.0.0		3.4.6	4.1.0	0.5.0	0.4.0
HDP 2.1 Apr. 2014	2.4.0	0.12.1	0.13.0	0.4.0	4.7.2			0.98.0	4.0.0	1.5.1	0.9.1	0.5.0		1.4.4	1.4.0		1.5.1		3.4.5	4.0.0	0.4.0	
	Hadoop & YARN	Pig	Hive	Tez	Solr	Spark	Slider	HBase	Phoenix	Accumulo	Storm	Falcon	Atlas	Sqoop	Flume	Kafka	Ambari	Cloudbreak	Zookeeper	Oozie	Knox	Ranger
	DATA MGMT	DATA ACCESS					GOVERNANCE & INTEGRATION					OPERATIONS					SECURITY					
HORTONWORKS DATA PLATFORM																						



## It Looks Like a File System



The screenshot shows a web-based file browser interface. At the top, the breadcrumb path is `/ user / it1 / geolocation`. There are buttons for `+ New directory`, `Browse...`, and a text input for `Select files to upload.`. Below the path, the directory name `geolocation` is displayed with a search icon. A search bar labeled `Search File Names` is also present. The main content area shows a table of files and directories.

Name	Size	Last Modified	Owner	Group	Permission	
..						
<a href="#">geolocation.csv</a>	514.3 kB	2016-03-13 19:42	maria_dev	hdfs	-rw-r--r--	<input type="checkbox"/>
<a href="#">trucks.csv</a>	59.9 kB	2016-03-13 19:41	maria_dev	hdfs	-rw-r--r--	<input type="checkbox"/>

Below the file browser, a terminal window is shown with the following command and output:

```
[it1@sandbox ~]$ hdfs dfs -ls /user/it1/geolocation
Found 2 items
-rw-r--r--  3 maria_dev hdfs      526677 2016-03-13 23:42 /user/it1/geolocation/geolocation.csv
-rw-r--r--  3 maria_dev hdfs      61378 2016-03-13 23:41 /user/it1/geolocation/trucks.csv
[it1@sandbox ~]$
```



## Data Input Options

