

# Big Data and Analytics

## Hadoop Ecosystem

Dr. Abzетdin Adamov

School of Information Technology and Engineering

ADA University

<http://site.ada.qu.edu.az/~aadamov>

# Previously Covered Topics

- Key differences of Traditional and Big Data Architecture
- Transferring Computation Power against Transferring Data
- Schema on Read vs Schema on Write
- Hadoop Core – Storage: HDFS Architecture
- Hadoop Core – Processing: MapReduce Architecture

# Objectives

- Vagrant + Provisioning + VirtualBox = Repeatable Multi WMs
- Hadoop 2.0 vs Hadoop 1.0
- Hadoop Ecosystem Components Classification
- Hadoop Ecosystem Components Key Features



# HADOOP ECOSYSTEM

## Big Data Handling

---

Before the advent of Hadoop, storage and processing of big data was a big challenge. But now that Hadoop is available, companies have realized the business impact of Big Data and how understanding this data will drive the growth.

# Big Data Landscape 2016 (Version 3.0)

## Infrastructure

**Hadoop On-Premise**  
 cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, bluedata, jethro

**Hadoop in the Cloud**  
 amazon web services, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, altiscale, baile

**Spark**  
 databricks, GridGain, TACHYON NEXUS

**Cluster Services**  
 amazon web services, Kubernetes, HPCC SYSTEMS, docker, MESOSPHERE, Core OS, pepperdata, StackIQ

## Analytics

**Analyst Platforms**  
 Palantir, AYASDI, Quid, enigma, Digital Reasoning, ORBITALINSIGHT

**Analytics Platforms**  
 Microsoft, GUAVUS, Datameer, Bottlenose, inter|ana

**Data Science Platforms**  
 context relevant, CONTINUUM, DataRobot, Alpine, MODE, plotly, ARIMO, dataiku, tonian, DOMINO, sense, yhat, ALGORITHMIA

**Visualization**  
 tableau, Google Cloud Platform, Qlik, looker, Roambi, SENSE, QONDATA, datarama, CHARTIO

## Applications

**Sales & Marketing**  
 RADIUS, Gainsight, bloomreach, Zeta, EVERSTRING, livefyre, blueyonder, Lattice, kahuna, infer, SAILTHRU, persado, AVISO, sense, QUANTIFIND, ACTIONIQ, fuse:machines, JEN GAGIO

**Customer Service**  
 MEDALLIA, ATTENSIY, CLARABRIDGE, CLICKFOX, STELLAService, NGDATA, Preact, DigitalGenius, appuri, Wise.io

**Human Capital**  
 gild, Connectifier, textio, entelo, hiQ

**Legal**  
 RAVEL, JUDICATA, Everlaw, Brevia, PREMONITION

**NoSQL Databases**  
 amazon DynamoDB, Google Cloud Platform, Microsoft Azure, ORACLE, mongoDB, MarkLogic, DATSTAX, Couchbase, SequoiaDB, redislabs, influxdata

**NewSQL Databases**  
 SAP HANA, Clustring, Pivotal, paradigm4, nuODB, memsql, VOLTDB, spice, MarioDB, citusdata, deapdb, Trafodion, Cockroach LABS

**BI Platforms**  
 Power BI, amazon web services, DOMO, Wave Analytics, GoodData, birst, KIVOS insights, platforma, atscale, ARCADIA, SISENSE

**Statistical Computing**  
 sas, SPSS, MATLAB

**Log Analytics**  
 splunk, sumologic, kibana, CLOUD PHYSICS, loggly

**Social Analytics**  
 Hootsuite, NETBASE, DATASIFT, tracx, bitly, synthosia, simplereach

**Ad Optimization**  
 AppNexus, MediaMath, critico, rocketfuel, Integral, theTradeDesk, Adgorithms, dstillery, Livelihood, TAPAD, DataXu, Uppier, MOAT

**Security**  
 CYLANCE, CounterTack, cybereason, ThreatMetrix, AREA 1 SECURITY, SentinelOne, Recorded Future, Guardian Analytics, FORTSCALE, siftscience, Keybase, feedzai, SICINFYD

**Vertical AI Applications**  
 facebook, Clara, KASIST, lumiata

**Graph Databases**  
 neo4j, OrientDB, InfiniteGraph

**MPP Databases**  
 TERADATA, VERTICA, Netezza, Microsoft Azure, Pivotal, snowflake, cognitio, kognitio, EXASOL, dremio, Infoworks

**Cloud EDW**  
 amazon web services, Microsoft Azure, Pivotal, snowflake, cognitio, kognitio, EXASOL, dremio, Infoworks

**Data Transformation**  
 alteryx, talend, TRIFACTA, tamr, StreamSets, Alation

**Data Integration**  
 informatica, MuleSoft, snaplogic, Bedrock Data, xplenty

**Real-Time**  
 amazon web services, METAMARKETS, stream, confluent, DATATORRENT, dataArtisans

**Machine Learning**  
 Azure Machine Learning, H2O, SKYTREE, rapidminer, DATASIPW, deepinsight, VIZENZE, PredictionIO, glowfish, IDIBON, YSEOP

**Speech & NLP**  
 NarrativeScience, NUANCE, Wavema Alpha, semantic machines, ARRIA, nora, apiai, cortico.io, maluba, MindMeld, IDIBON, YSEOP

**Horizontal AI**  
 IBM Watson, Cortana, sentiment, viv, nermana, nora, Numenta, (i)Q HyperScience, SI, Disruptor Labs, clarifai, DEXTR0, MetaMind

**Publisher Tools**  
 outbrain, Taboola, quantcast, Chartbeat, yieldbot, Yieldmo

**Govt / Regulation**  
 Socrata, OPENGOV, FN FiscalNote, enigma, PREPOL, mark43, OpenDataSoft

**Finance**  
 affirm, LendingClub, OnDeck, Kreditech, zest finance, LendUp, Kabbage, tdemark, INSIKT, uora, Dataminr, Lenddo, KENSHC, AIDYA, ISENTIUM, Quantopian, sentiment

**Management / Monitoring**  
 New Relic, APPDYNAMICS, amazon web services, actifio, Numerify, splunk, DATADOG, Trocana, DRIVEN, Anodot

**Security**  
 TANIUM, Illumio, CODE42, Amazon Web Services, CIPHERCLOUD, VECTRA, sqrrl, BlueTalon

**Storage**  
 amazon web services, Microsoft Azure, panasas, nimblestorage, COHO DATA, Qumulo

**App Dev**  
 apigee, CASK, KENIC, Typesafe, DRIVEN

**Crowd-sourcing**  
 amazon mechanicalturk, CrowdPower, WorkFusion

**Search**  
 hp, ORACLE, ENDECA, EXALEAD, Lucidworks, elastic, ThoughtSpot, MAANA, swifttype, Algolia, SINEQUA

**Data Services**  
 UC OPERA, Mu Sigma, EXL, DATA SCIENCE, kaggle, datascope, DataKind

**For Business Analysts**  
 OrigamiLogic, ClearStory, CIRRO, import ib

**Web / Mobile / Commerce / Analytics**  
 Google Analytics, mixpanel, RJMetrics, BLUECORE, AMPLITUDE, granify, sumal, Airtable, retention, custora

**Education / Learning**  
 KNEWTON, Clever, ceclara, PANORAMA, knowre

**Life Sciences**  
 23andMe, Counsyl, RECOMBINE, KYRUUS, FLATIRON, zymogen, HealthTap, METABIOTA, ZEPHYR HEALTH, ovig, Gingerio, transcriptic, Glow, enlitic, AiCure, Atomwise

**Industries**  
 OP@WER, eHarmony, RetailNext, STITCH FIX, WorkFusion, BLUE RIVER, TACHYUS, SwiftKey, Seeq, FarmLogs, HowGood, elect, BIGHT MACHINE, statmuse, BOEYER

## Cross-Infrastructure/Analytics

amazon web services, Google, Microsoft, IBM, SAP, Sas, data, hp, Autonomy, VERTICA, vmware, TIBCO, TERADATA, ORACLE, NetApp

**Framework**  
 hadoop HDFS, YARN, Spark, MESOS, TEZ, Flink, CDAP

**Query / Data Flow**  
 SLAMDATA, APACHE DRILL, Google Cloud Dataflow, HIVE, CouchDB, riak, OPENTSDB, nifi

**Data Access**  
 accumulo, mongoDB, cassandra, kafka, CouchDB, riak, OPENTSDB, nifi

**Coordination**  
 Apache Ambari, Apache Zookeeper

**Real-Time**  
 STORM, Spark, APEX, Flink, TACHYON, druid

**Stat Tools**  
 R, ScalaLab, Numpy, SciPy

**Machine Learning**  
 milib, Aerosolve, Caffe, WEKA, DIMSUM, jupyter, DL4J

**Search**  
 elasticsearch, Solr, Lucene

**Security**  
 Apache Ranger

**Visualization**  
 Zepplin

## Data Sources & APIs

**Health**  
 Apple, JAWBONE, GARMIN, practicefusion, fitbit, Withings, VALIDIC, netatmo, kinsa, Human API

**IOT**  
 UPTAKE, ThingWorx, helium, samsara, ABBYURY, estimote

**Financial & Economic Data**  
 Bloomberg, DOW JONES, THOMSON REUTERS, YODLEE, PREMISE, S&P CAPITAL IQ, quandl, xignite, CBNSIGHTS, mattermark, Stocktwits, estimize, PLAID

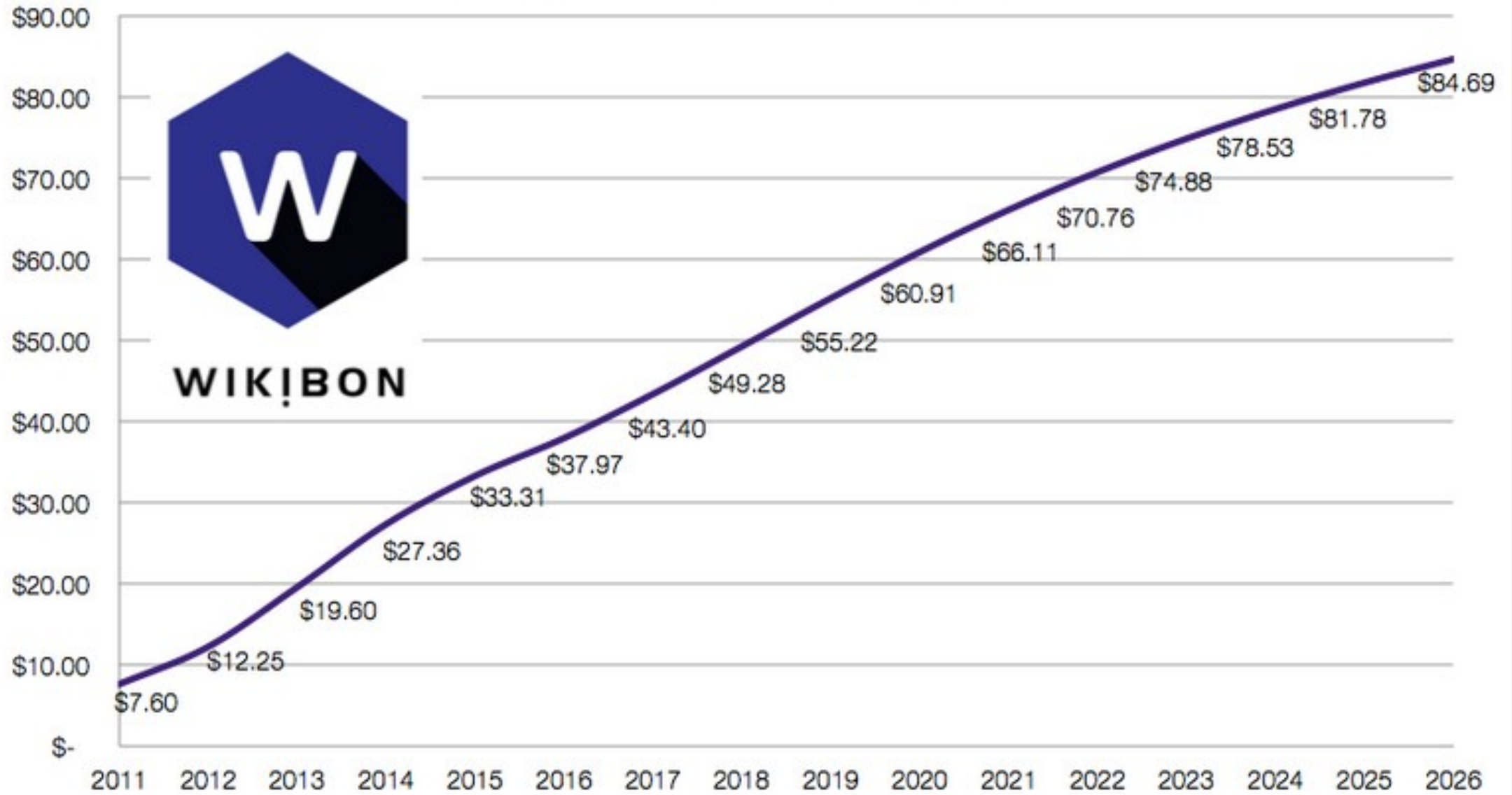
**Air / Space / Sea**  
 PLANET LABS, spire, WINDWARD, CRUISE, Airware, DroneDeploy, SKYWATCH

**Location / People / Entities**  
 acxiom, Experian, EPSON, InsideView, GARMIN, foursquare, STREETLINE, Crimon Hexagon, CARTODB, factual, PlacIQ, CIRCULATE, placemeter, BASIS, Sense

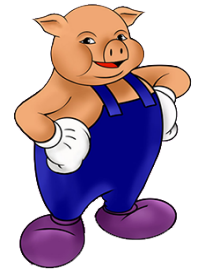
**Other**  
 qualtrics, panjiva, DATA.GOV

**Incubators & Schools**  
 GA, PLURALSIGHT, INSIGHT, DataCamp, DataElite, The Data Incubator, METIS

Big Data Market Forecast, 2011-2026 (\$US B)



# Hadoop Ecosystem Components



# Companies building on top of Hadoop

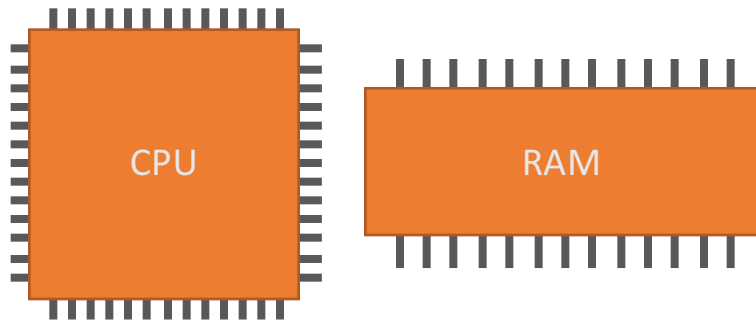
- Amazon Web Services
- Cloudera
- Hortonworks
- IBM
- Intel
- MapR Technologies
- Microsoft
- Pivotal Software
- Teradata

# Powered by Apache Hadoop

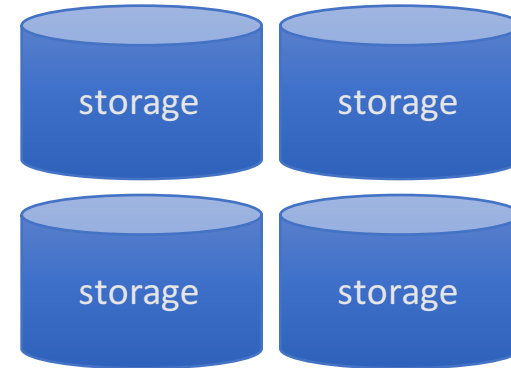
- <https://wiki.apache.org/hadoop/PoweredBy>
- Thousands companies and organizations with Hadoop Cluster size from several to hundreds thousands nodes (40.000 at Yahoo)

# Hadoop Core = Storage + Compute

Yet Another Resource Negotiator  
(YARN)

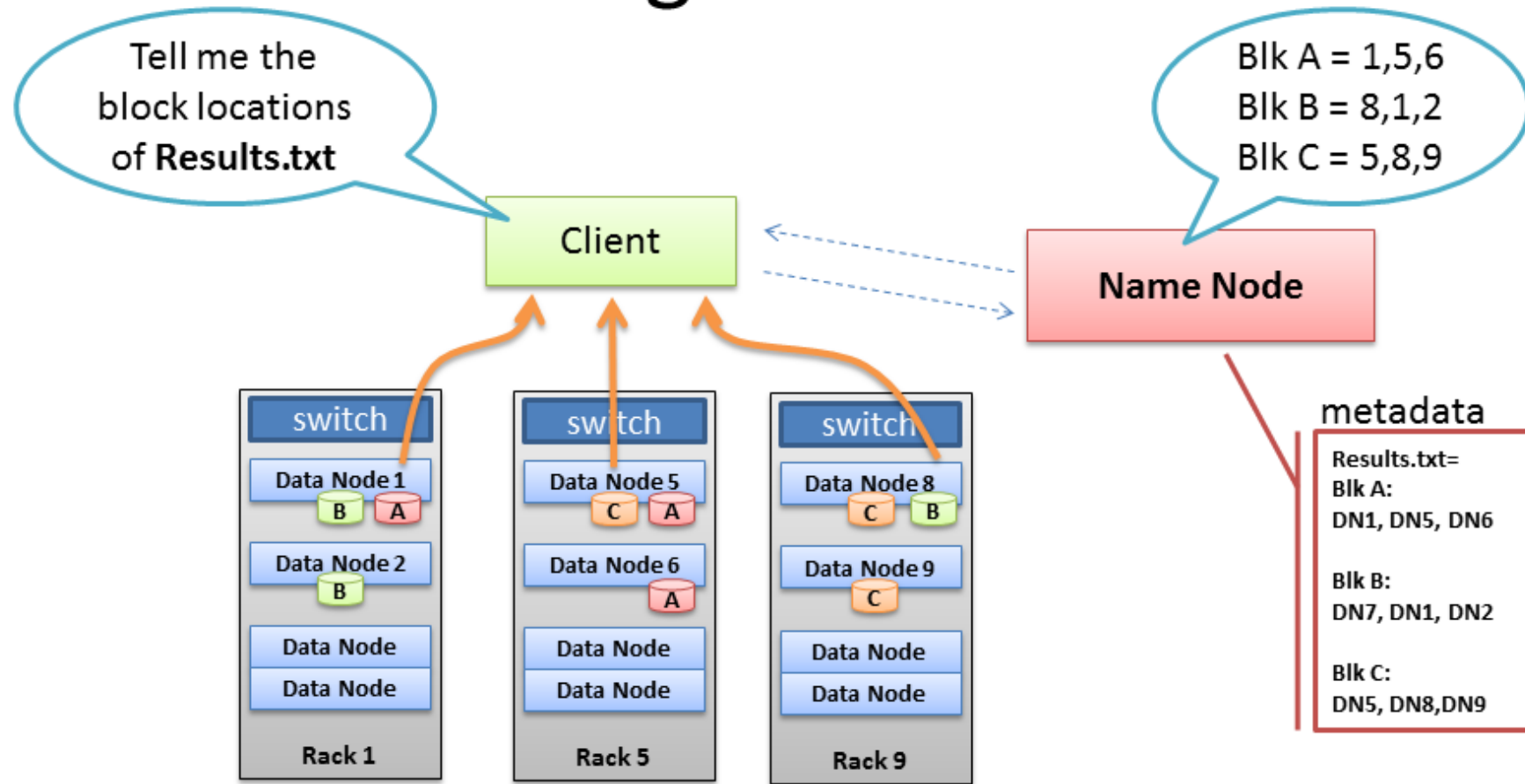


Hadoop Distributed File System  
(HDFS)



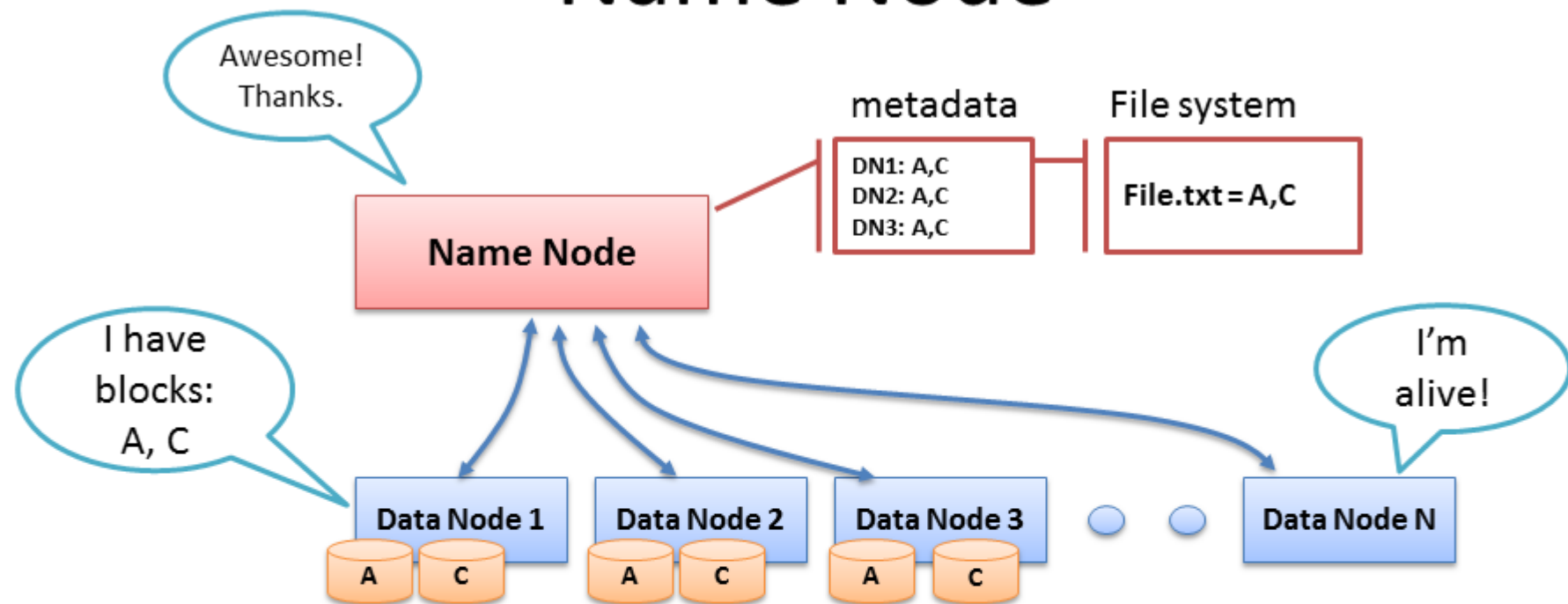
# Hadoop 2.0 vs Hadoop 1.0

# Client reading files from HDFS



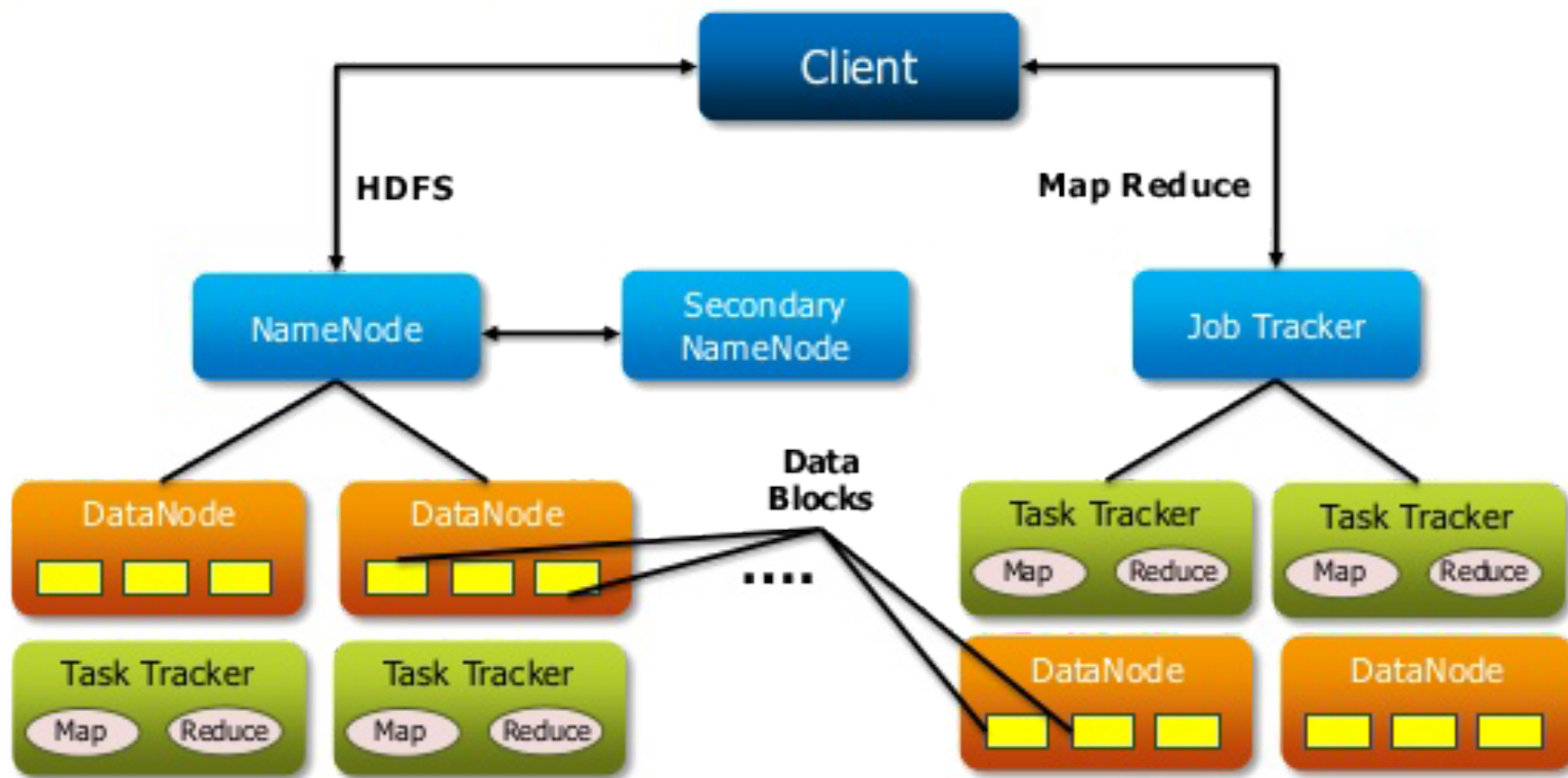
- Client receives Data Node list for each block
- Client picks first Data Node for each block
- Client reads blocks sequentially

# Name Node

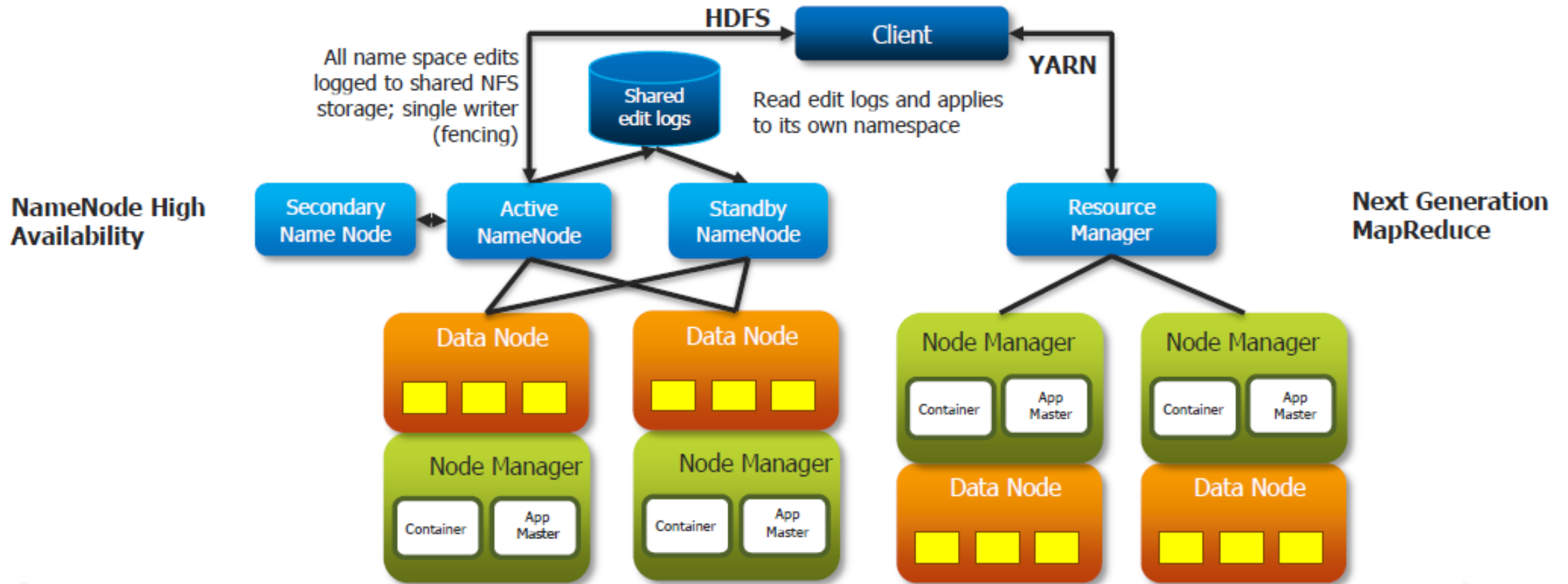


- Data Node sends Heartbeats
- Every 10<sup>th</sup> heartbeat is a Block report
- Name Node builds metadata from Block reports
- TCP – every 3 seconds
- If Name Node is down, HDFS is down

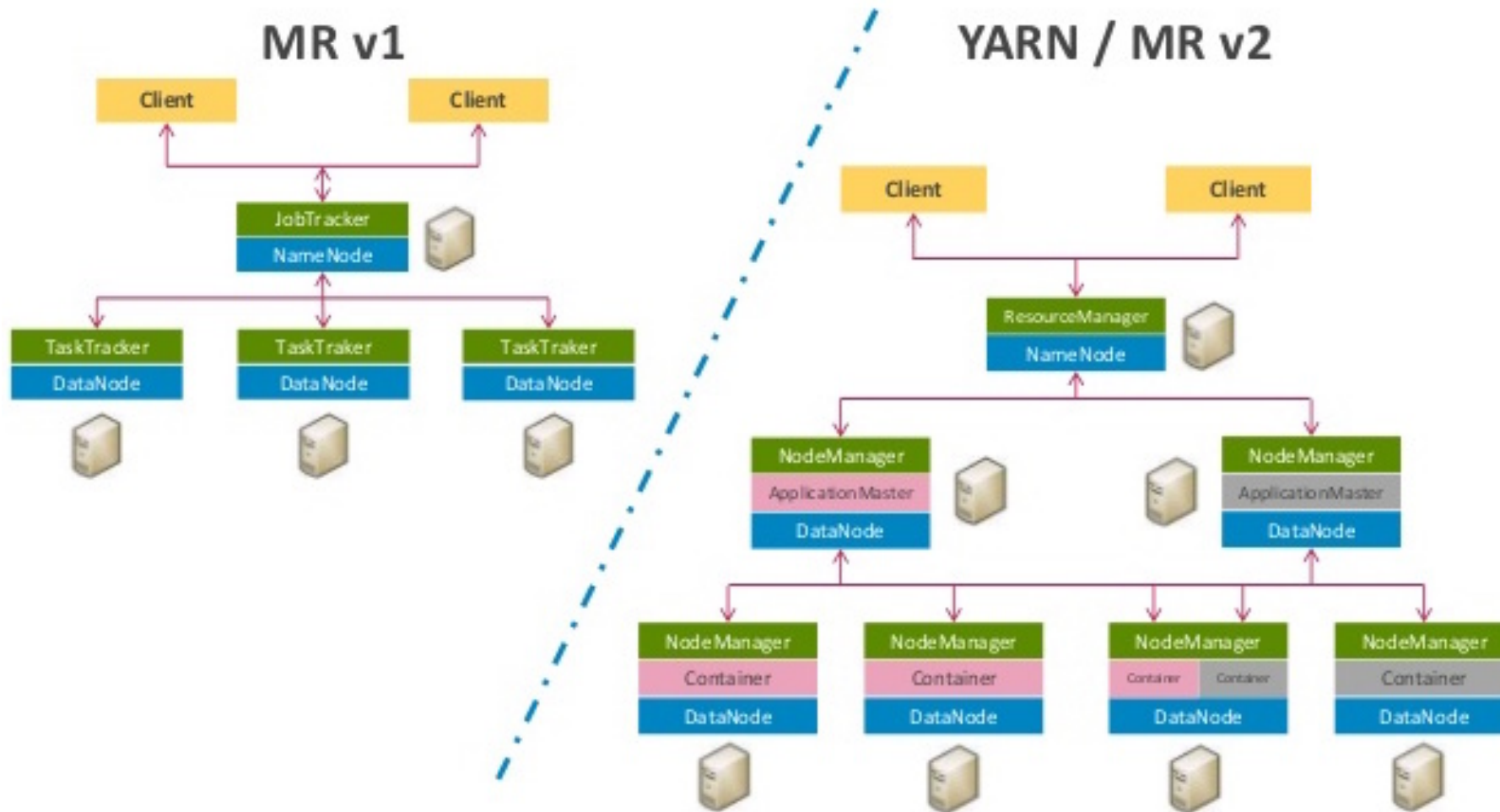
# Hadoop 1.0 Bottlenecks: HDFS / MapReduce



# Hadoop 2.0 Architecture

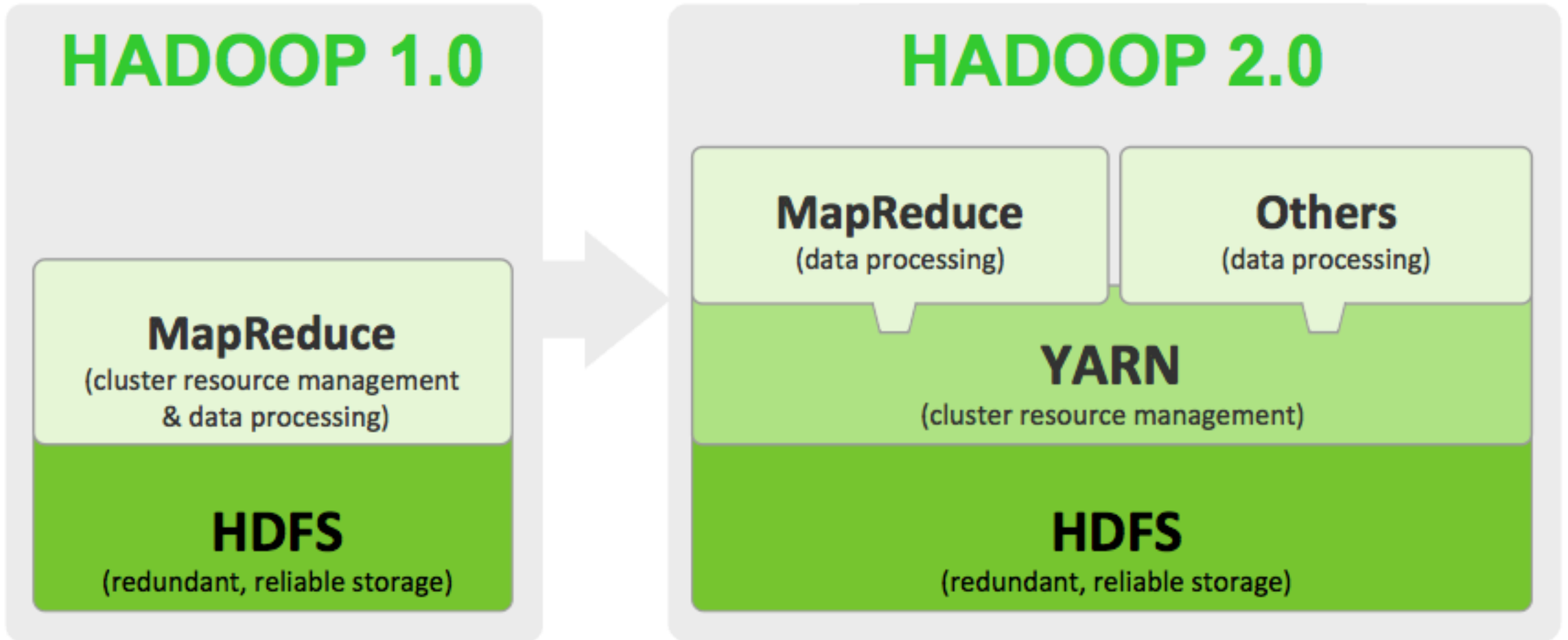


# YARN/MR v2 vs MR v1 Architecture



- **YARN** : Yet Another Resource Negotiator
- **MR** : MapReduce

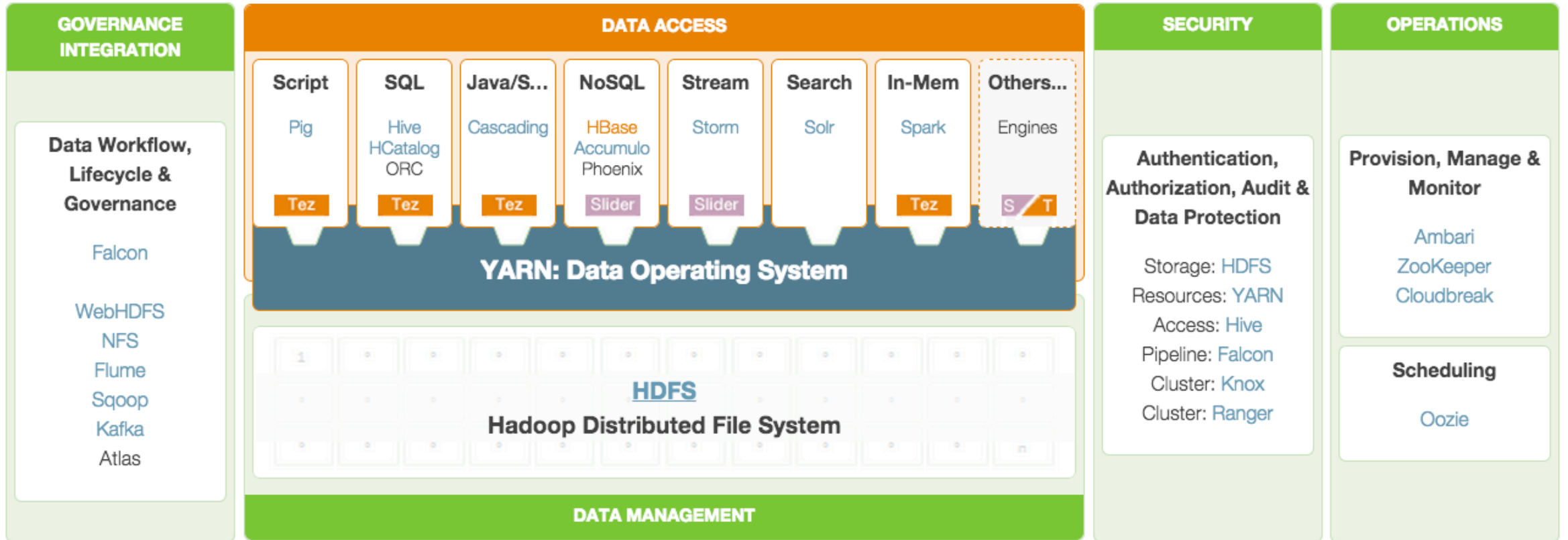
# Hadoop 2.0 vs Hadoop 1.0 – Processing



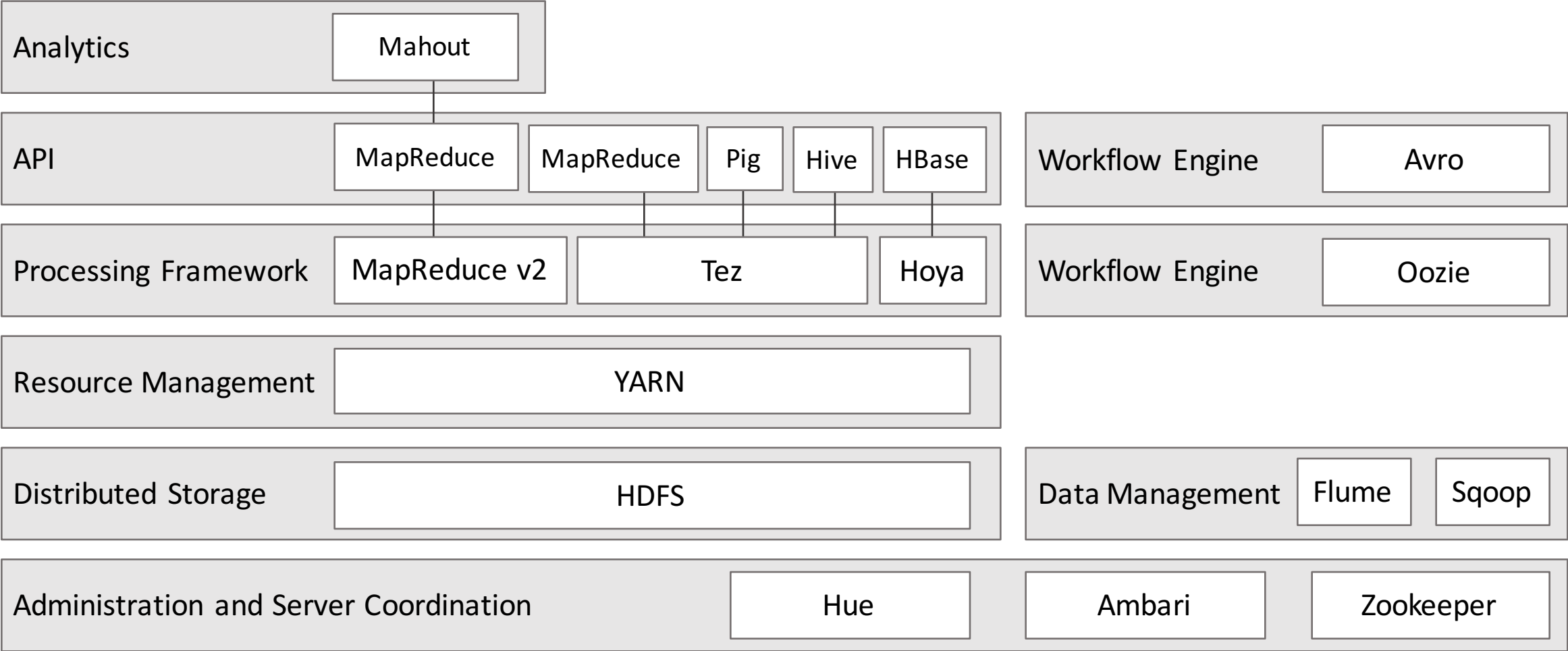
# The Hadoop Ecosystem

The image displays a variety of logos for projects within the Hadoop ecosystem. At the top center is the Hadoop logo, featuring a blue elephant silhouette inside a dark blue square frame with the word "Hadoop" written below it. To the left of the Hadoop logo is the Hive logo, which depicts a yellow and black striped bee with the word "HIVE" in bold black letters. Below the Hadoop logo is the HBase logo, consisting of the word "HBASE" in large, bold, red letters. To the right of the Hadoop logo is the Oozie logo, which features the word "OOZIE" in a stylized, colorful font. Above the Hadoop logo is the Tez logo, showing the word "TEZ" in orange letters with a stylized orange flame or bird-like shape. To the right of the Tez logo is the Apache Ambari logo, which includes a circular icon with a white building-like structure and the text "Apache Ambari" and the URL "http://incubator.apache.org/ambari". To the right of the Ambari logo is the Spark logo, featuring the word "Spark" in a bold, black font with an orange starburst shape above the letter "k". Below the Ambari logo is the Apache Falcon logo, which includes a blue falcon head icon and the text "APACHE FALCON". To the right of the Falcon logo is the Solr logo, which features a colorful sunburst icon and the text "Apache Solr". Below the Solr logo is the Flume logo, which includes a circular icon with a brown gear and blue water splashes, and the word "FLUME" in blue letters. To the left of the Hadoop logo is the Pig logo, which features a cartoon pig character in blue overalls and purple shoes. Below the Pig logo is the Apache ORC logo, which includes a green circular icon with a white house-like shape and the text "Apache ORC". To the left of the ORC logo is the Apache Knox logo, which features the word "KNOX" in large, green, block letters with a colorful feather below it. Below the Knox logo is the Storm logo, which features a blue circular icon with a white "S" shape and the word "STORM" in bold, black letters. To the left of the Storm logo is the Accumulo logo, which features a grid of white lines and the word "ACCUMULO" in black letters. Below the Accumulo logo is the Kafka logo, which features a circular icon with three white circles and the word "kafka" in lowercase black letters. To the right of the Kafka logo is a cartoon character of a zookeeper in a green uniform and a blue cap with "ZOO" written on it, holding a brown shovel. The entire collection of logos is arranged in a roughly circular pattern around the central Hadoop logo.

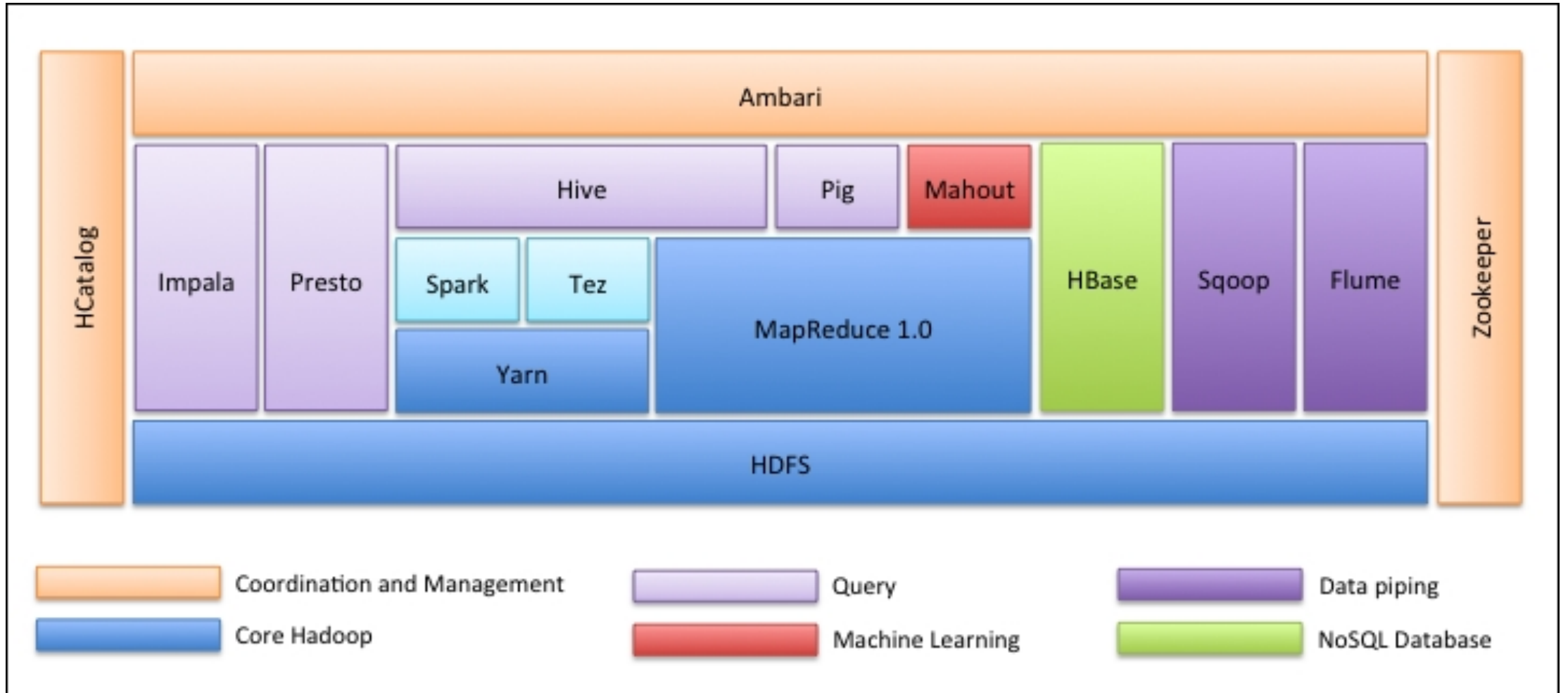
# Hortonworks Hadoop Distribution



# Classification of Hadoop Ecosystem Components



# Classification of Hadoop Ecosystem Components



# Hadoop Ecosystem Components

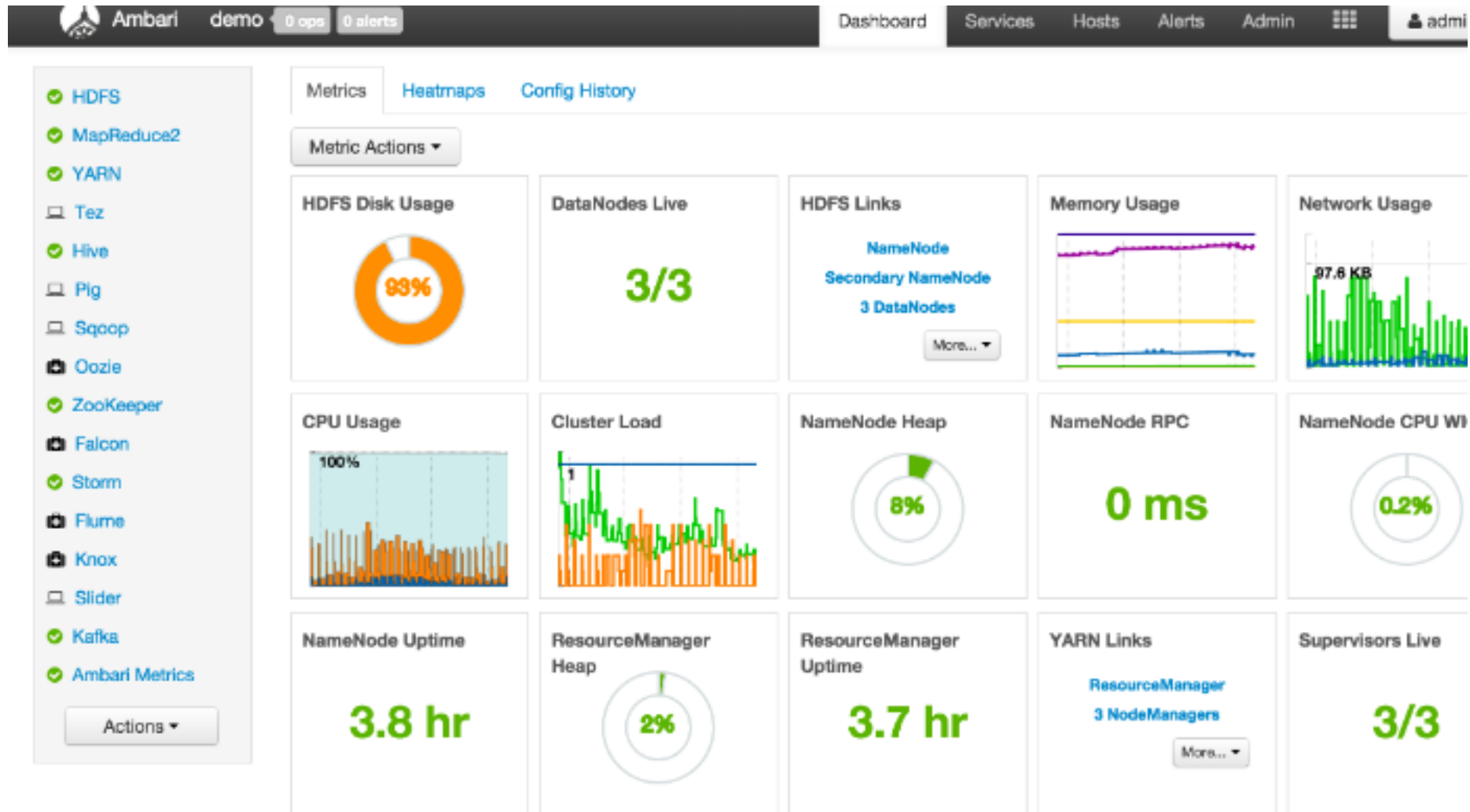
# Data Management Frameworks

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

# Ambari WEB UI (REST)



# 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 Version	Pig	Hive	Druid	Tez	Solr	Spark	Zeppelin	Slider	HBase	Phoenix	Accumulo	Storm	Falcon	Atlas	Sqoop	Flume	Kafka	Ambari	Zookeeper	Oozie	Knox	Ranger	
<b>HDP 2.6*</b> 1H2017	2.7.3	0.16.0	1.2.1+ 2.1****	0.9.2	0.7.0	5.5.1 ****	1.6.3+ 2.1**	0.7.0	0.91.0	1.1.2	4.7.0	1.7.0	1.1.0	0.10.0	0.8.0	1.4.6	1.5.2	0.10.1.0	2.5.0	3.4.6	4.2.0	0.11.0	0.7.0
HDP 2.5 Aug 2016	2.7.3	0.16.0	1.2.1+ 2.1****	0.7.0	5.5.1	1.6.2+ 2.0**	0.6.0	0.91.0	1.1.2	4.7.0	1.7.0	1.0.1	0.10.0	0.7.0	1.4.6	1.5.2	0.10.0	2.4.0	3.4.6	4.2.0	0.9.0	0.6.0	
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	3.4.6	4.2.0	0.6.0	0.5.0	
HDP 2.3 Oct 2015	2.7.1	0.15.0	1.2.1	0.7.0	5.2.1	1.4.1		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	0.8.2	2.1.0	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		0.60.0	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 April 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		
HDP 2.0 Oct 2013	2.2.0	0.12.0	0.12.0						0.96.1						1.4.4	1.3.1		1.4.4	3.4.5	3.3.2			

DATA MGMT

DATA ACCESS

GOVERNANCE & INTEGRATION

OPERATIONS

SECURITY

HORTONWORKS DATA PLATFORM

\* HDP 2.6 – Shows current Apache branches being used. Final component version subject to change based on Apache release process.

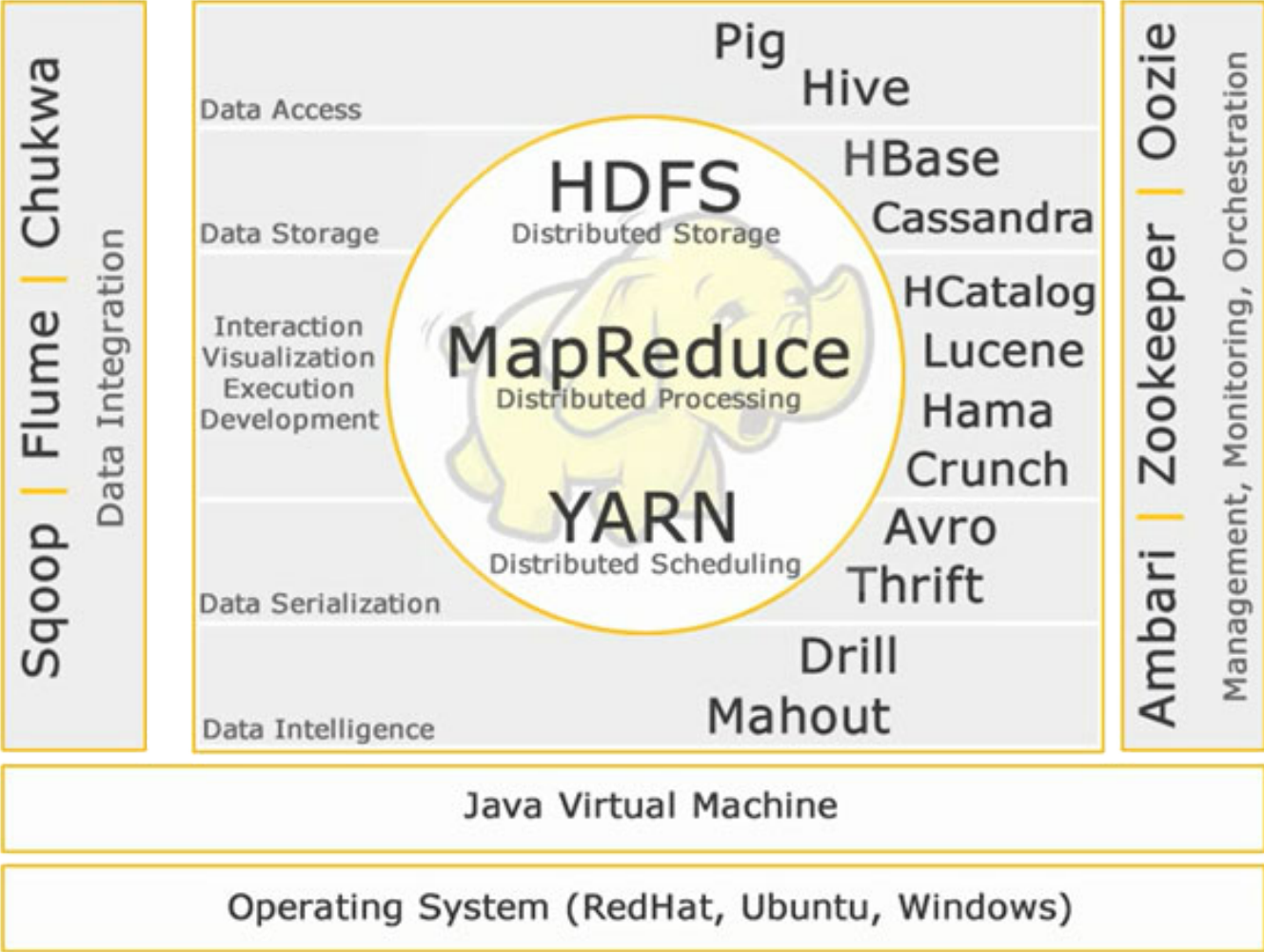
\*\* Spark 1.6.3+ Spark 2.1 – HDP 2.6 supports both Spark 1.6.3 and Spark 2.1 as GA.

\*\*\* Hive 2.1 is GA within HDP 2.6.

\*\*\*\* Apache Solr is available as an add-on product HDP Search.

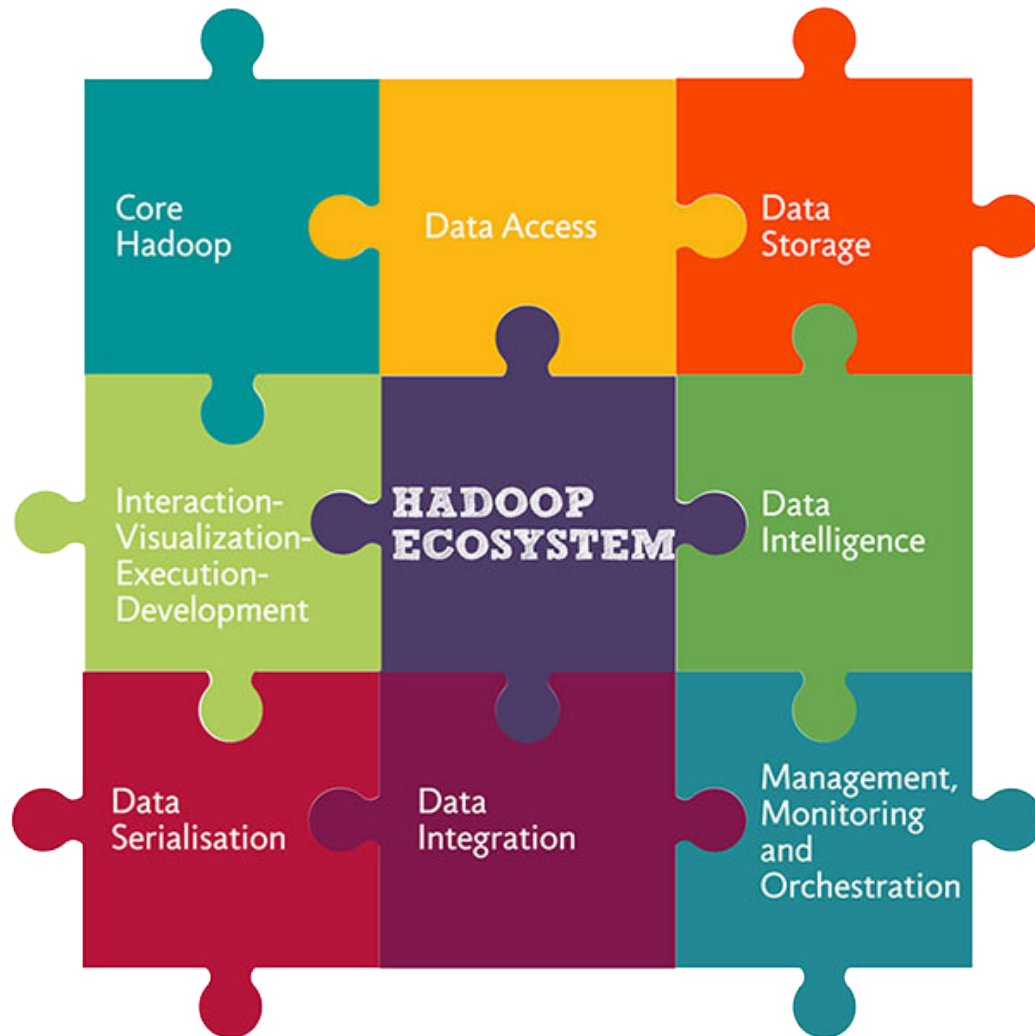
# Hadoop Ecosystem Components' Key Features

# HADOOP ECOSYSTEM COMPONENTS



Its important to understand the components in Hadoop Ecosystem to build right solutions for a given business problem.

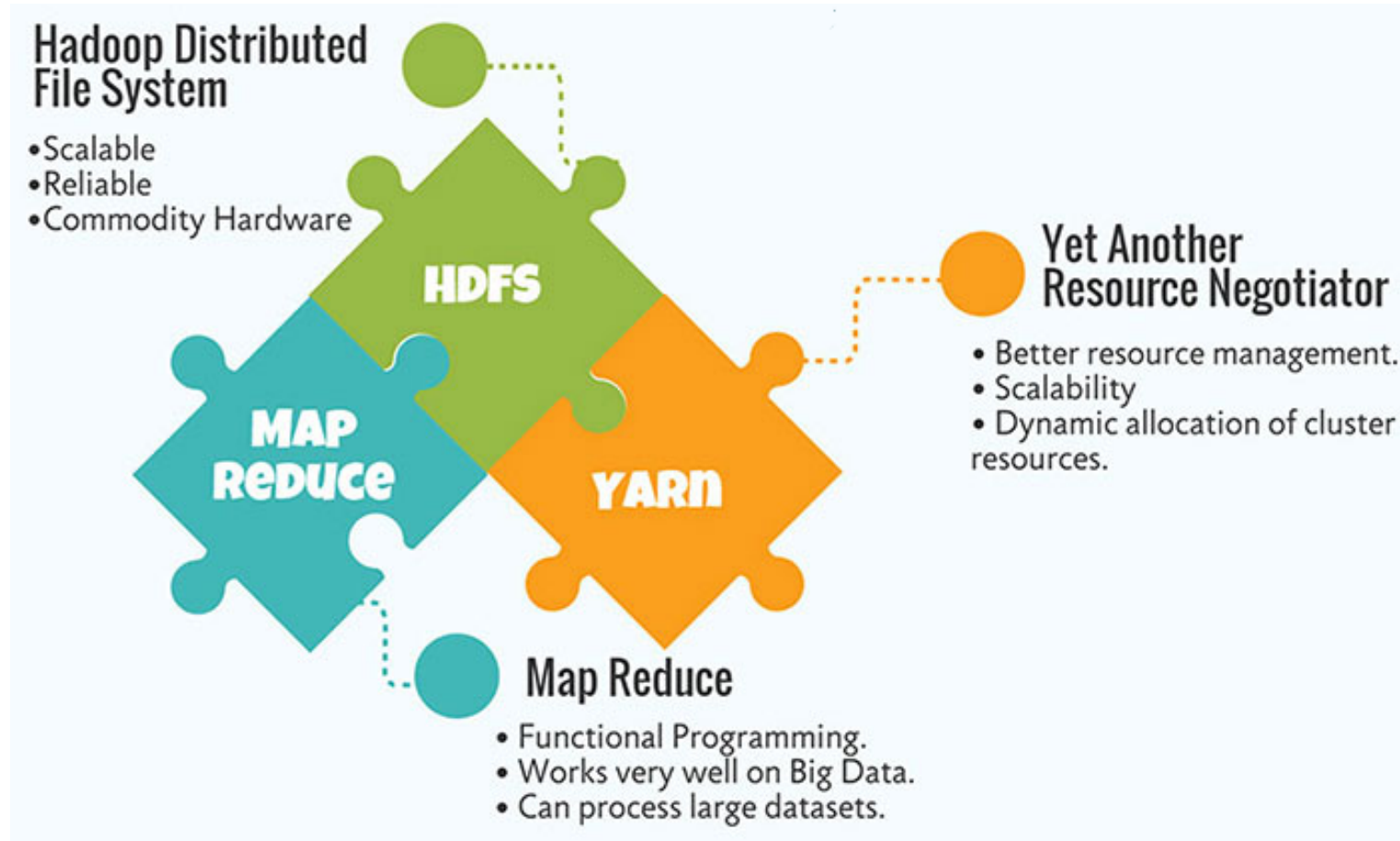
# Classification of the Hadoop Ecosystem Components



Hadoop is straight answer for processing Big Data.

Hadoop Ecosystem has a combination of technologies which proficient advantage in solving Data-oriented business problem.

# CORE HADOOP



## **Hadoop Distributed File System (HDFS)**

Stands for: managing big data sets with High Volume, Velocity and Variety.

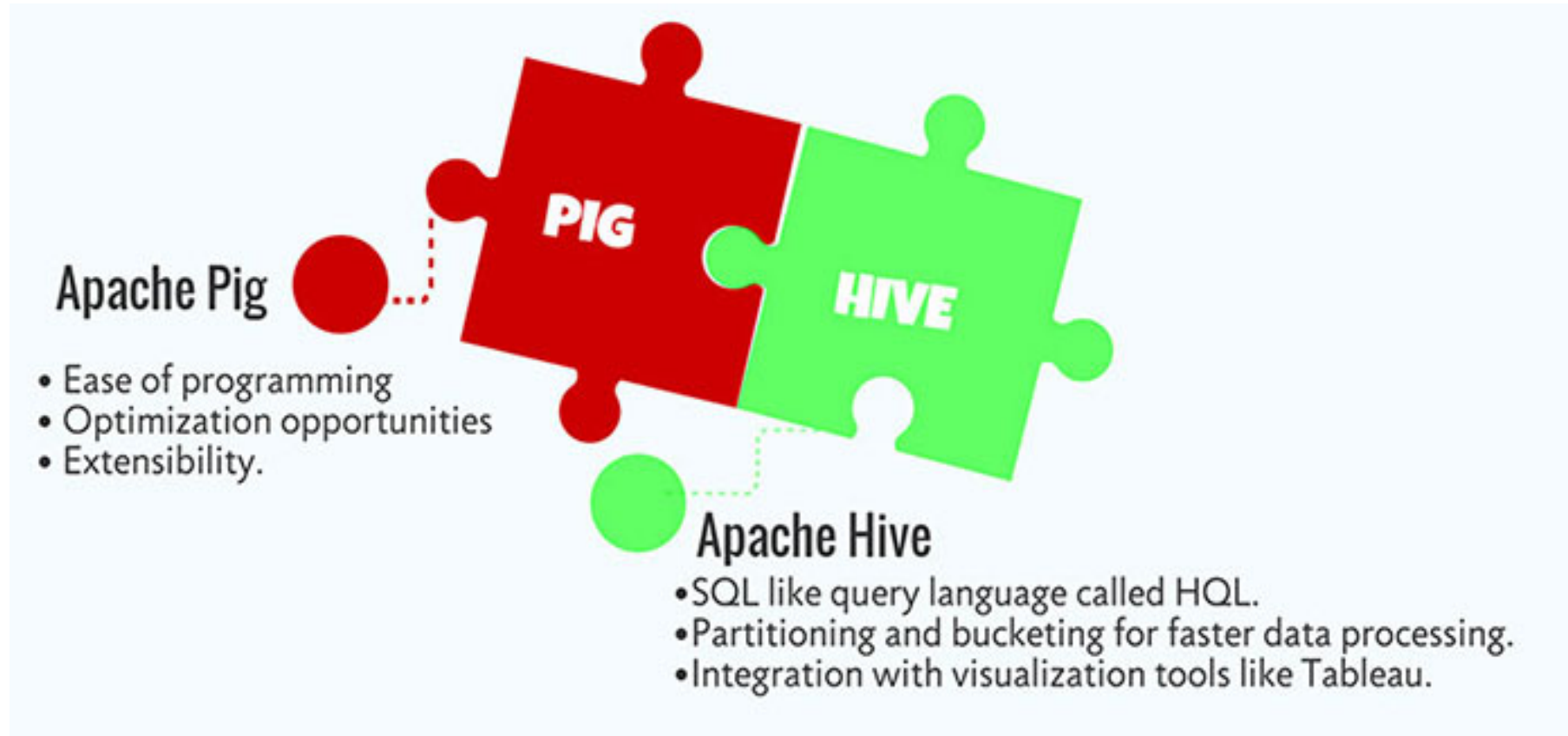
## **Map Reduce**

Stands for: processing high volume distributed data

## **Yet Another Resource Negotiator (YARN)**

Stands for: resource management, job scheduling and monitoring

# DATA ACCESS



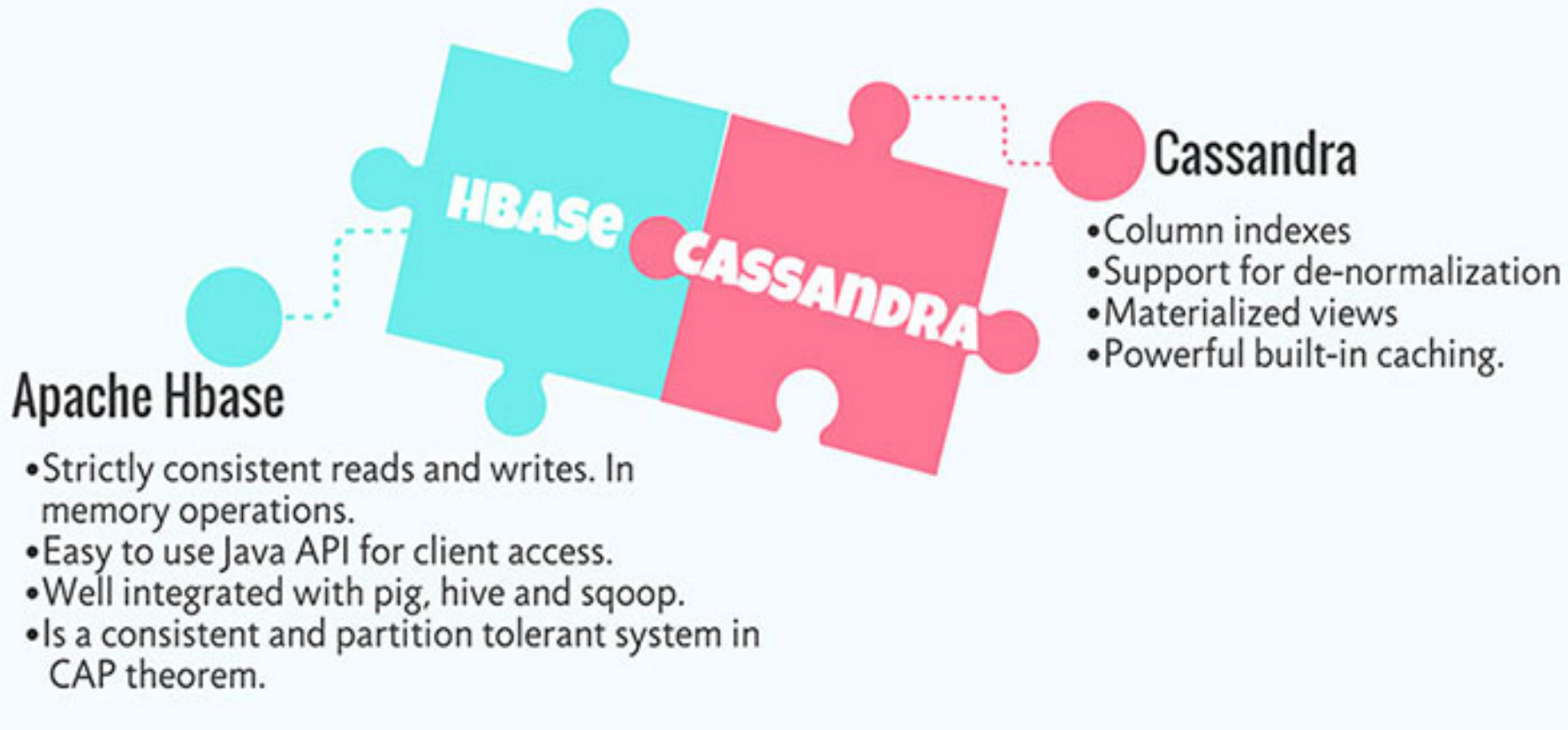
## **Apache Pig**

Stands for: high level language built on top of MapReduce for analyzing large datasets and for Data Flow.

## **Apache Hive**

Stands for: high level query language and data warehouse infrastructure built on top of Hadoop for providing data summarization, query and analysis.

# DATA STORAGE



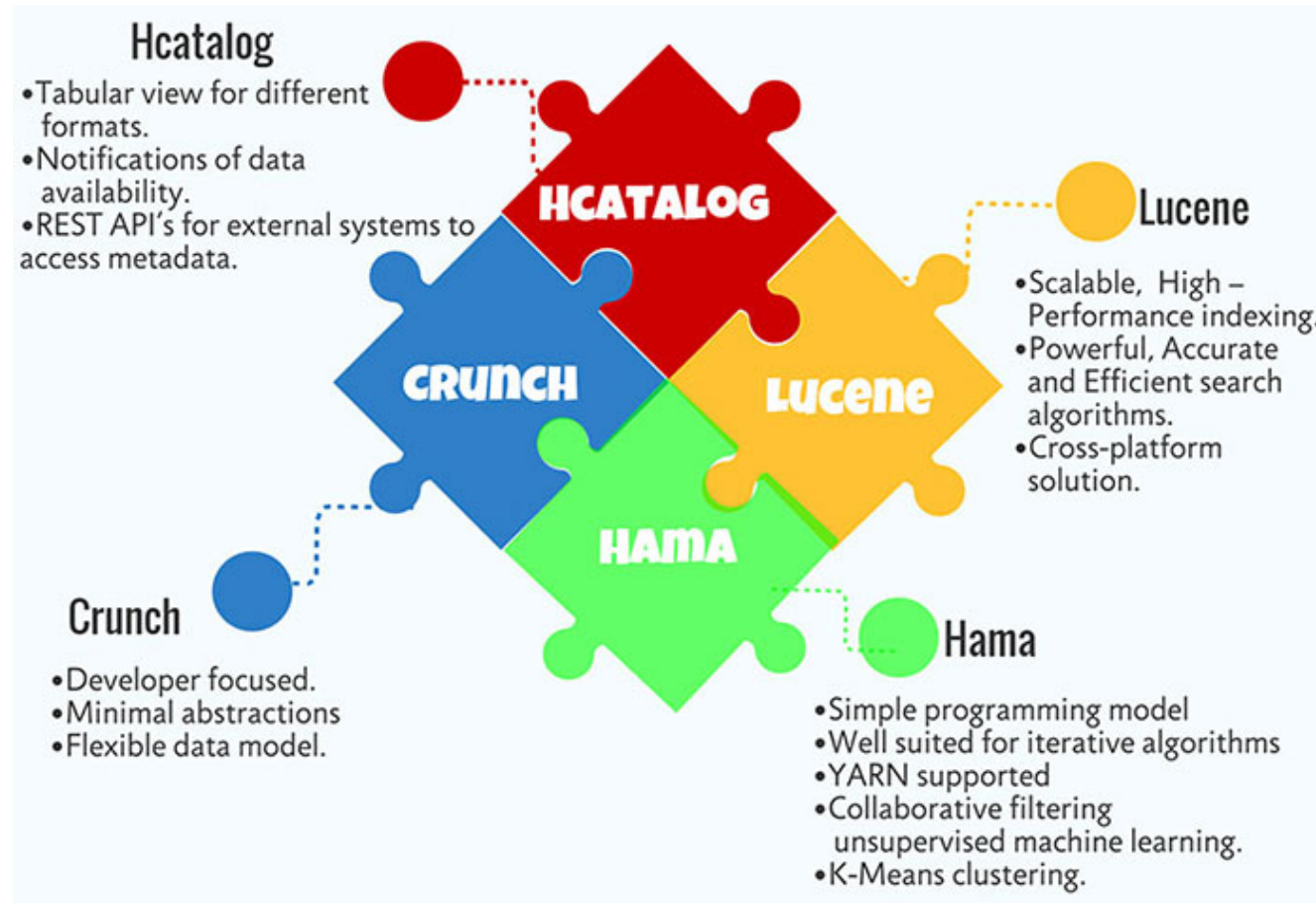
## **Apache HBase**

Stands for: NoSQL database built for hosting large tables with billions of rows and millions of columns on top of Hadoop.

## **Cassandra**

Stands for: NoSQL database based on key-value model designed for linear scalability and high availability.

# INTERACTION-VISUALIZATION-DEVELOPMENT



## Hcatalog

Stands for: providing integration of Hive metadata for other Hadoop applications like Pig, MapReduce and others.

## Lucene

Stands for: high-performance, full-featured text search engine library written entirely in Java.

## Hama

Stands for: distributed framework based on Bulk Synchronous Parallel(BSP) computing for massive scientific computations like matrix, graph and network algorithms.

## Crunch

Stands for: writing, testing and running MapReduce pipelines.

# DATA INTELLIGENCE



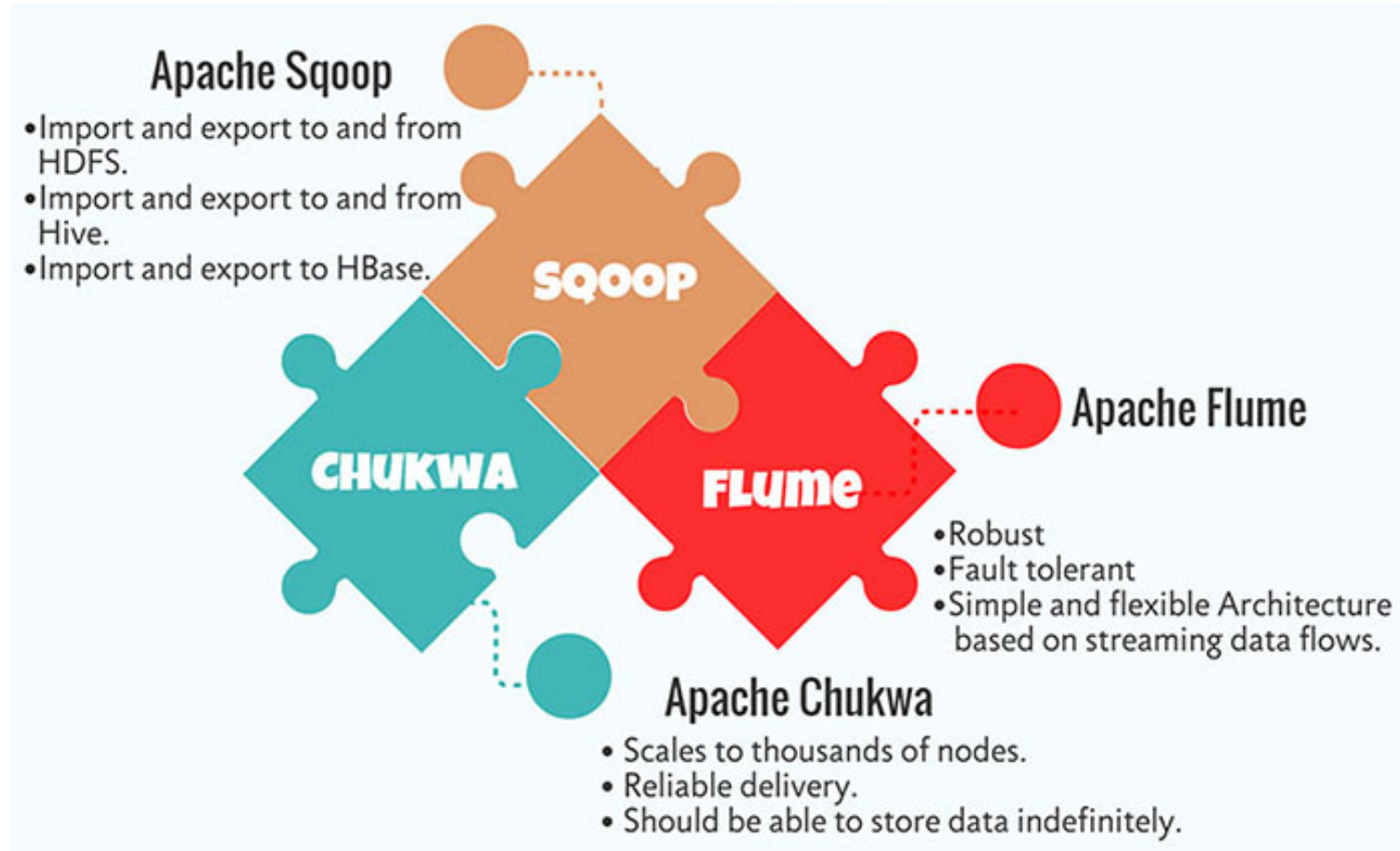
## Apache Drill

Stands for: low latency SQL query engine for Hadoop and NoSQL.

## Apache Mahout

Stands for: scalable machine learning library designed for building predictive analytics on Big Data. Mahout now has implementations apache spark for faster in memory computing.

# DATA INTEGRATION



## **Apache Sqoop**

Stands for: low latency SQL query engine for Hadoop and NoSQL.

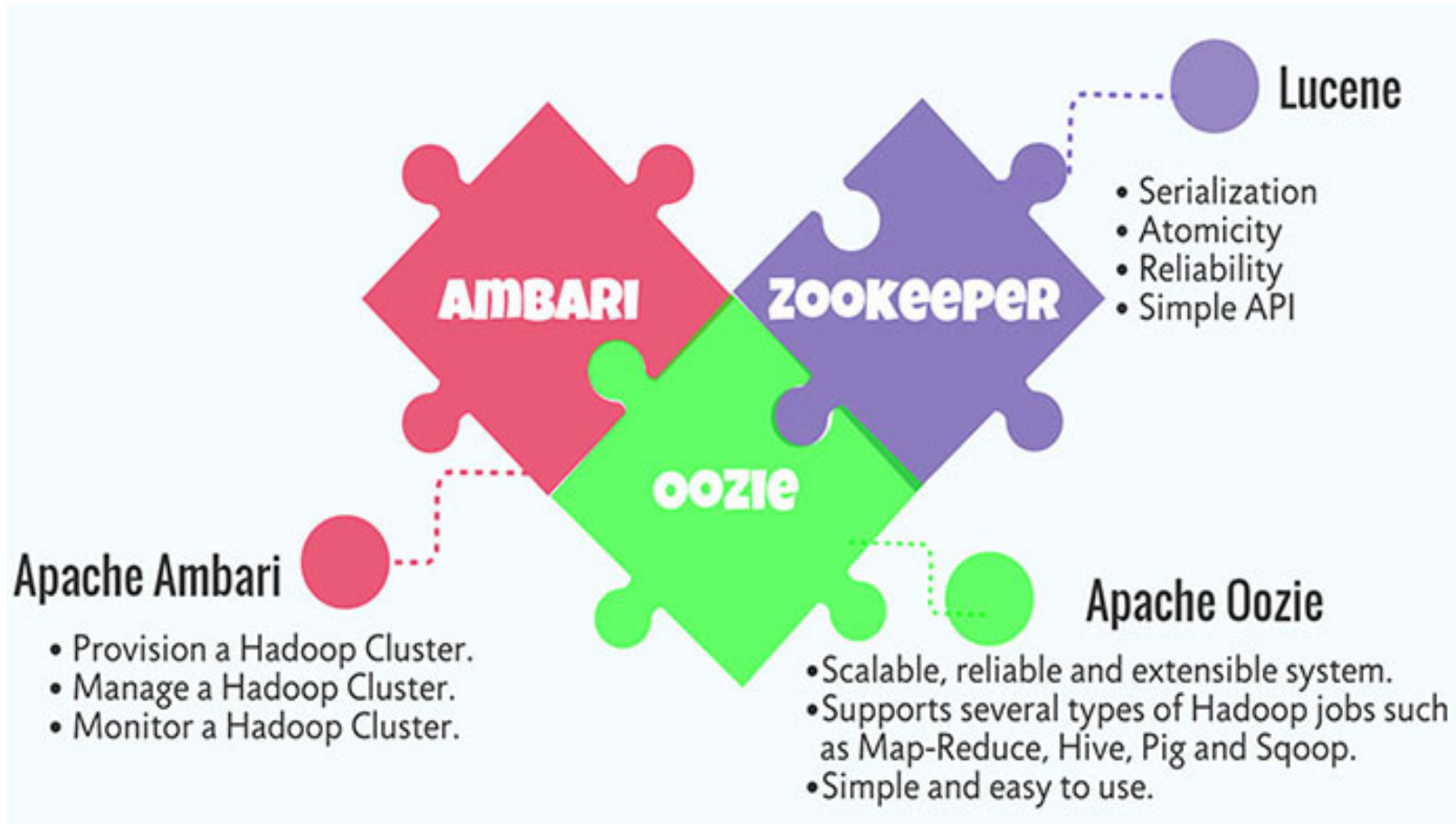
## **Apache Flume**

Stands for: distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.

## **Apache Chukwa**

Stands for: scalable log collector used for monitoring large distributed files systems.

# MANAGEMENT, MONITORING and ORCHESTRATION



## Apache Ambari

Stands for: simplifying Hadoop management by providing an interface for provisioning, managing and monitoring Apache Hadoop Clusters.

## Apache Zookeeper

Stands for: maintaining configuration information naming, providing distributed synchronization, and providing group services.

## Apache Oozie

Stands for: scheduling workflow to manage Apache Hadoop jobs.

# Where Can We Use Machine Learning (Data Science)



## Healthcare

- Predict diagnosis
- Prioritize screenings
- Reduce re-admittance rates



## Financial services

- Fraud Detection/prevention
- Predict underwriting risk
- New account risk screens



## Public Sector

- Analyze public sentiment
- Optimize resource allocation
- Law enforcement & security



## Retail

- Product recommendation
- Inventory management
- Price optimization



## Telco/mobile

- Predict customer churn
- Predict equipment failure
- Customer behavior analysis

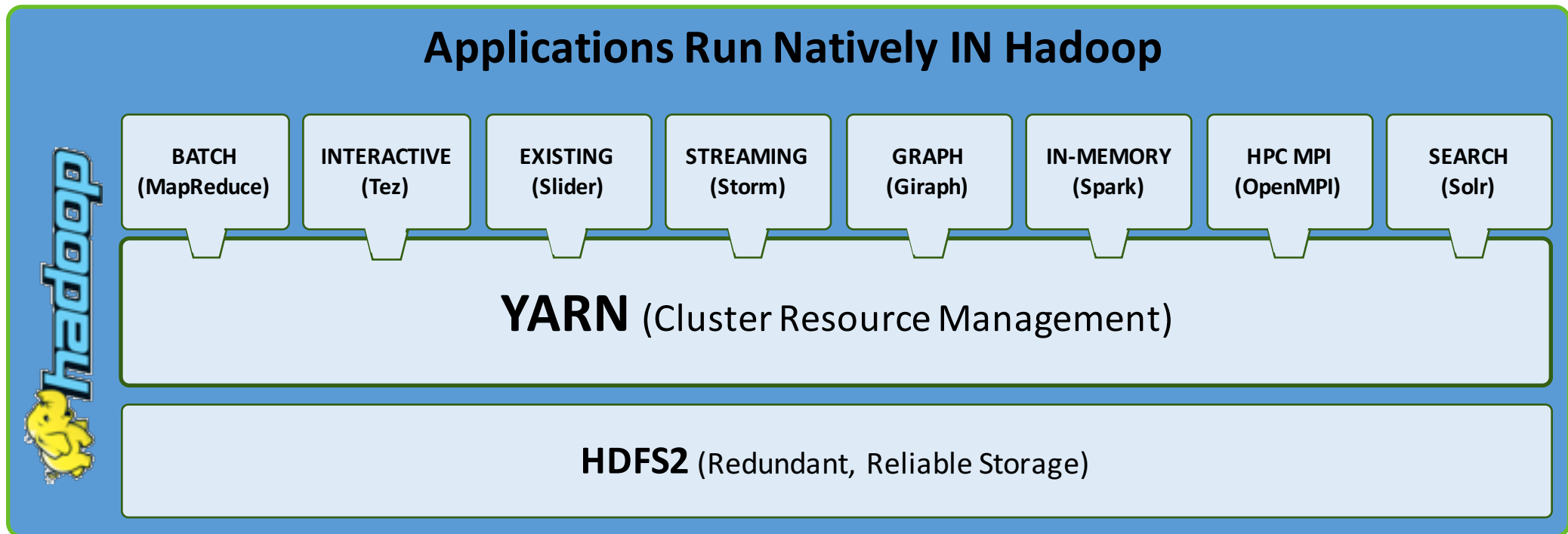


## Oil & Gas

- Predictive maintenance
- Seismic data management
- Predict well production levels

# YARN as a Data Operating System

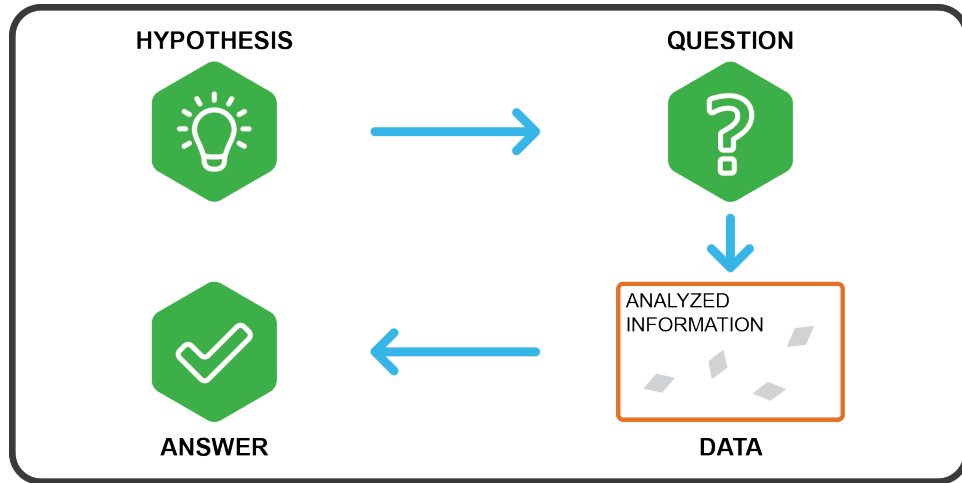
Applications now run “in” Hadoop,  
instead of “on” Hadoop.



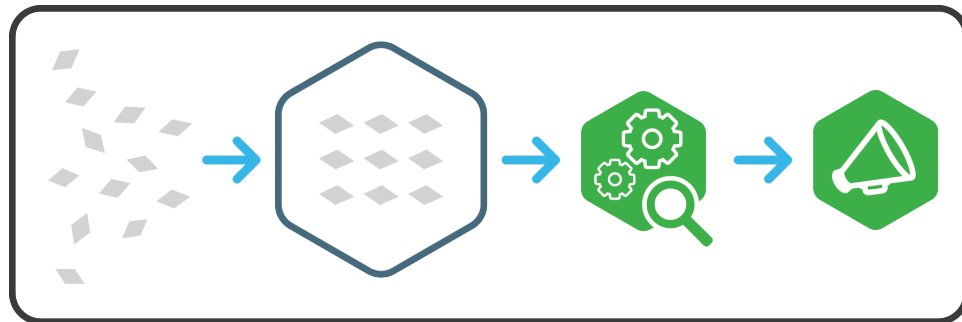
# Modern Data Applications approach to Insights

## Traditional Analytics

Structured & Repeatable  
Structure built to store data



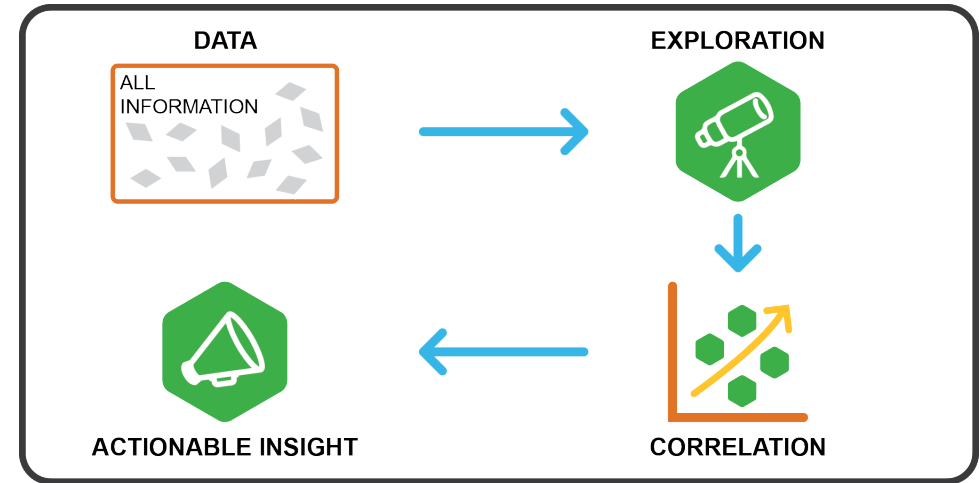
Start with hypothesis  
Test against selected data



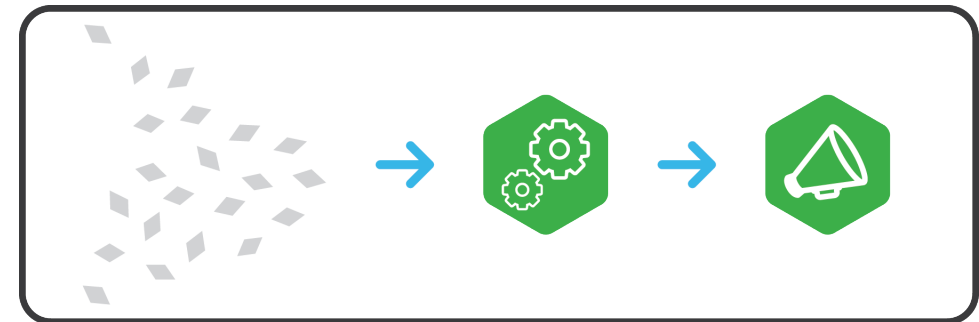
Analyze after landing...

## Next Generation Analytics

Iterative & Exploratory  
Data is the structure



Data leads the way  
Explore all data, identify correlations



Analyze in motion...



Q & A



Abzeturdin Adamov, Assoc Prof.

Email me at: [aadamov@ada.edu.az](mailto:aadamov@ada.edu.az)

Follow me at: @

Link to me at: [www.linkedin.com/in/adamov](http://www.linkedin.com/in/adamov)

Visit my blog at: [aadamov.wordpress.com](http://aadamov.wordpress.com)