Big Data and Analytics
Hadoop Ecosystem

Dr. Abzetdin Adamov
School of Information Technology and Engineering
ADA University
http://site.ada.qu.edu.az/~aadamov

Previously Covered Topics

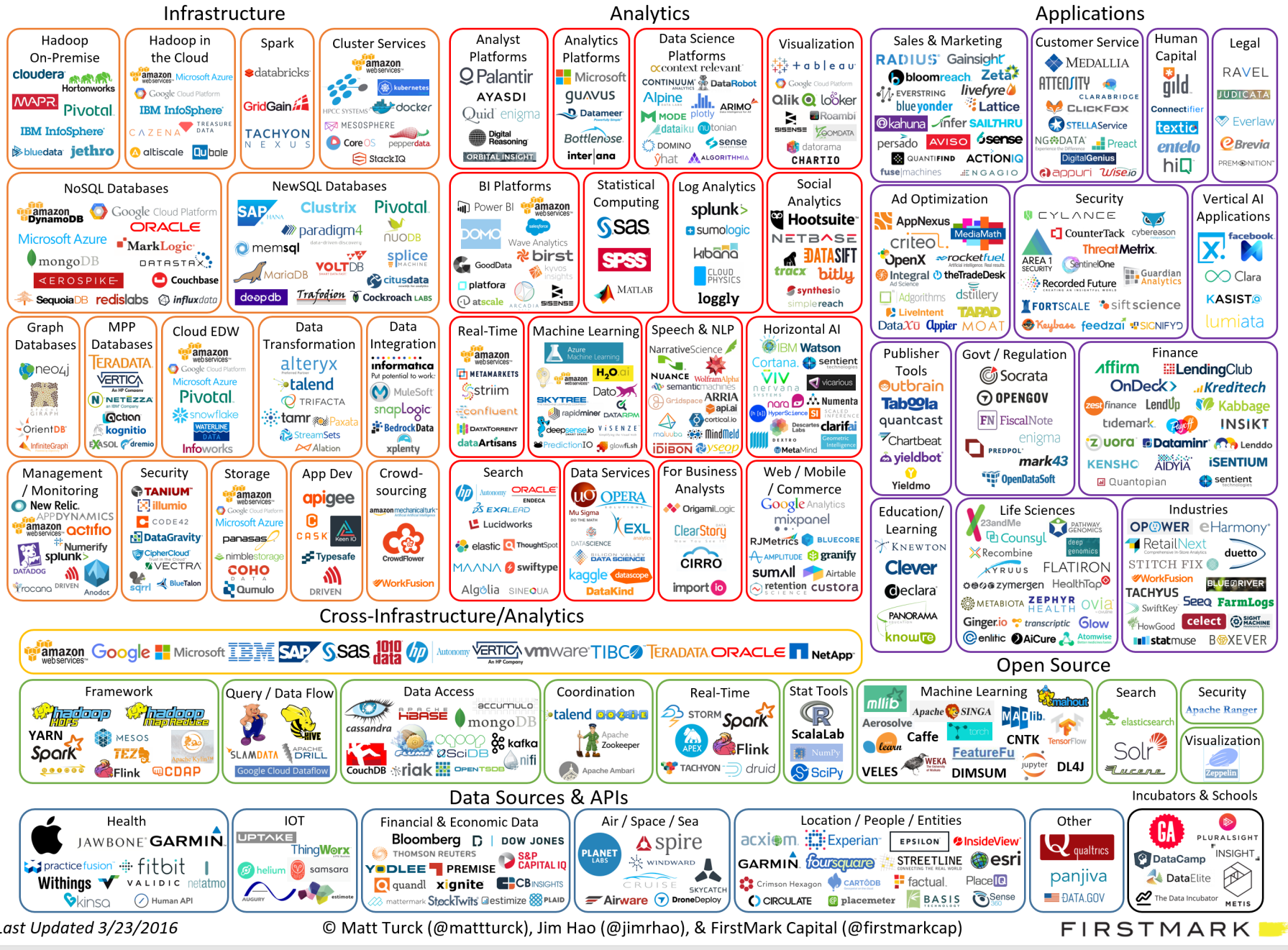
- Key differences of Traditional and Big Data Architecture
- Transferring Computation Power against Transferring Data
- Schema on Read vs Schema on Write
- Hadoop Core – Storage: HDFS Architecture
- Hadoop Core – Processing: MapReduce Architecture

Objectives

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

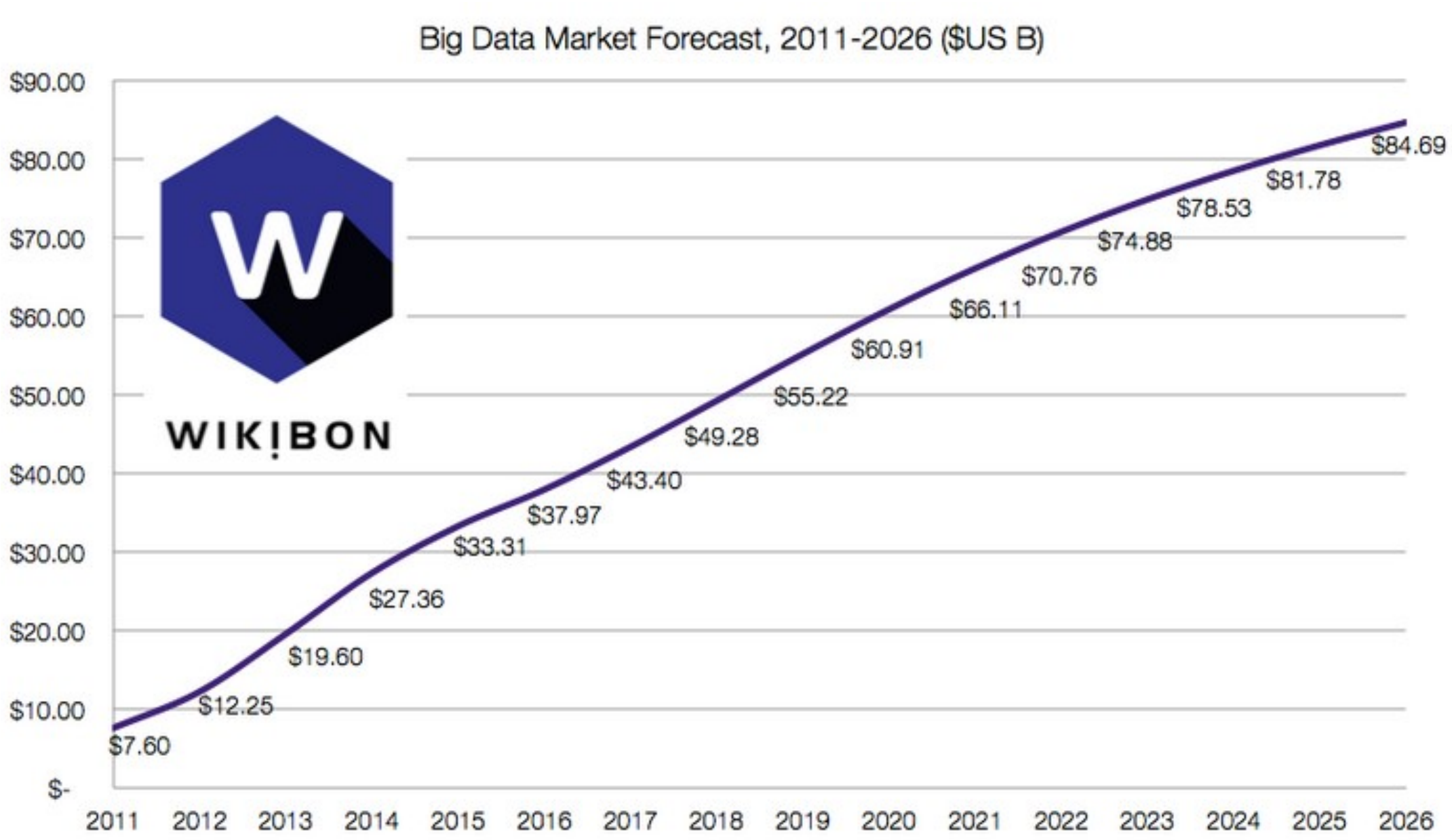

HADOOP
ECOSYSTEM
Big Data Handling
Before the advent of Hadoop, storage and processing of big data was a big challenge. But now that Hadoop is available, companies have realized the business impact of Big Data and how understanding this data will drive the growth.

Big Data Landscape 2016 (Version 3.0)

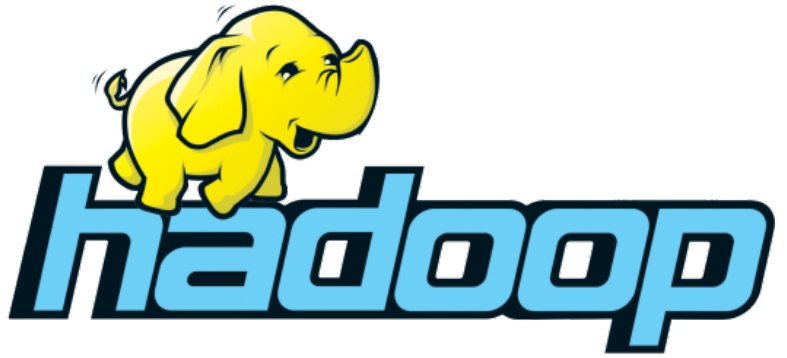


Last Updated 3/23/2016

FIRSTMARK


Big Data Market Forecast, 2011-2026 ($US B)
$90.00
$80.00
$70.00
$60.00
$50.00
$40.00
$30.00
$20.00
$10.00
$-
WIKIBON
$7.60
$12.25
$19.60
$27.36
$33.31
$37.97
$43.40
$49.28
$55.22
$60.91
$66.11
$70.76
$74.88
$78.53
$81.78
$84.69
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026

Hadoop Ecosystem Components



Companies building on top of Hadoop

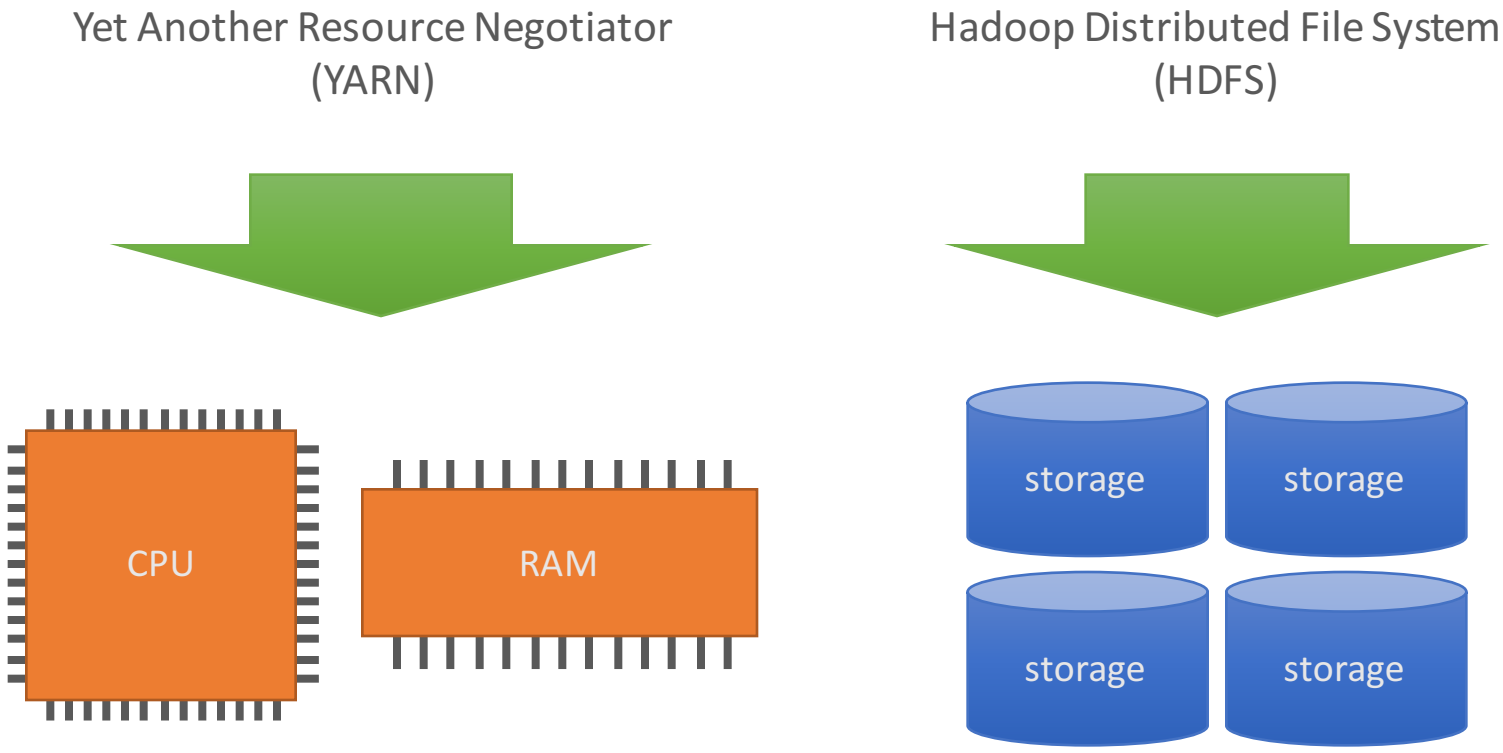
- Amazon Web Services
- Cloudera
- Hortonworks
- IBM
- Intel
- MapR Technologies
- Microsoft
- Pivotal Software
- Teradata

Powered by Apache Hadoop

- https://wiki.apache.org/hadoop/PoweredBy

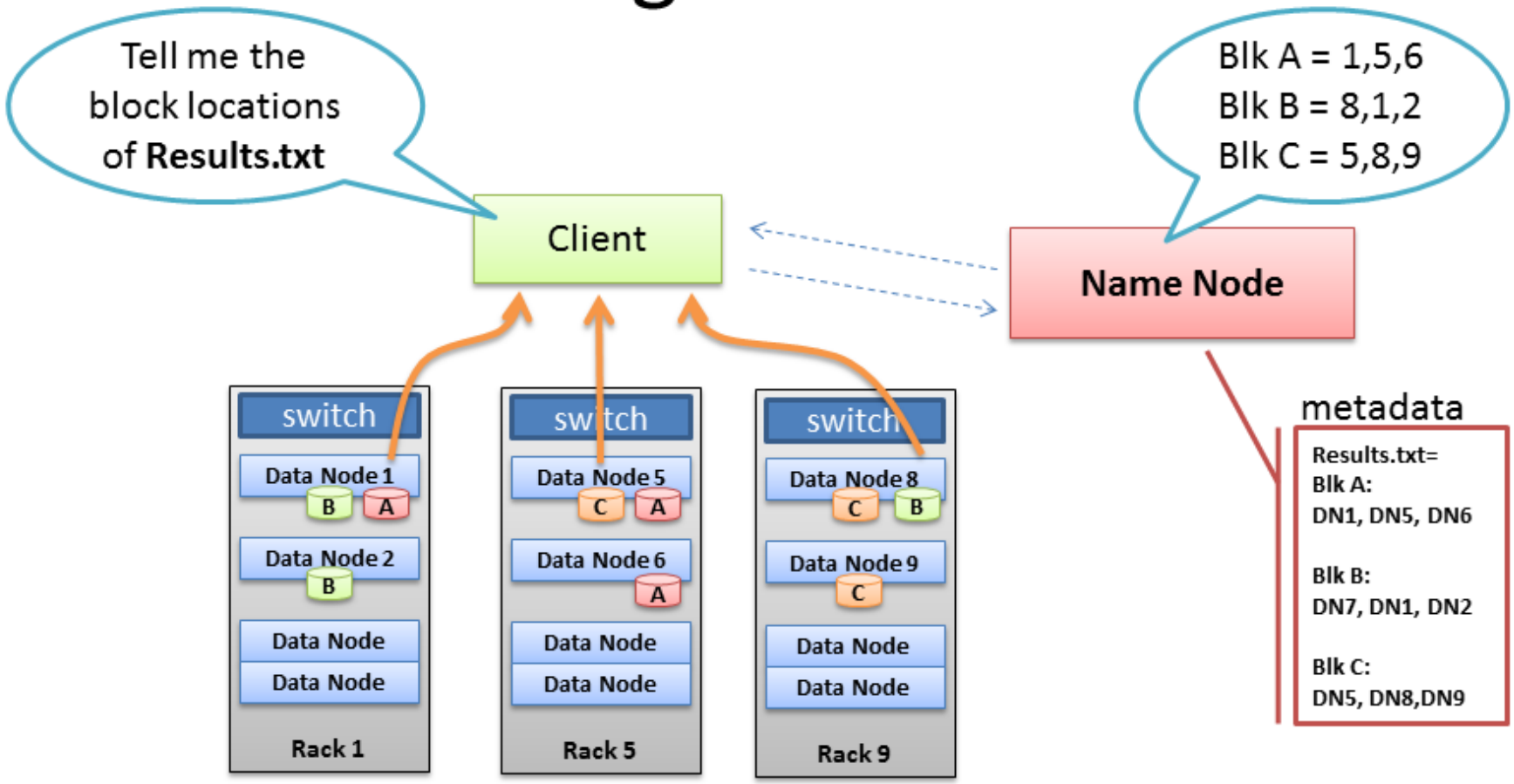
- Thousands companies and organizations with Hadoop Cluster size from several to hundreds thousands nodes (40.000 at Yahoo)

Hadoop Core = Storage + Compute



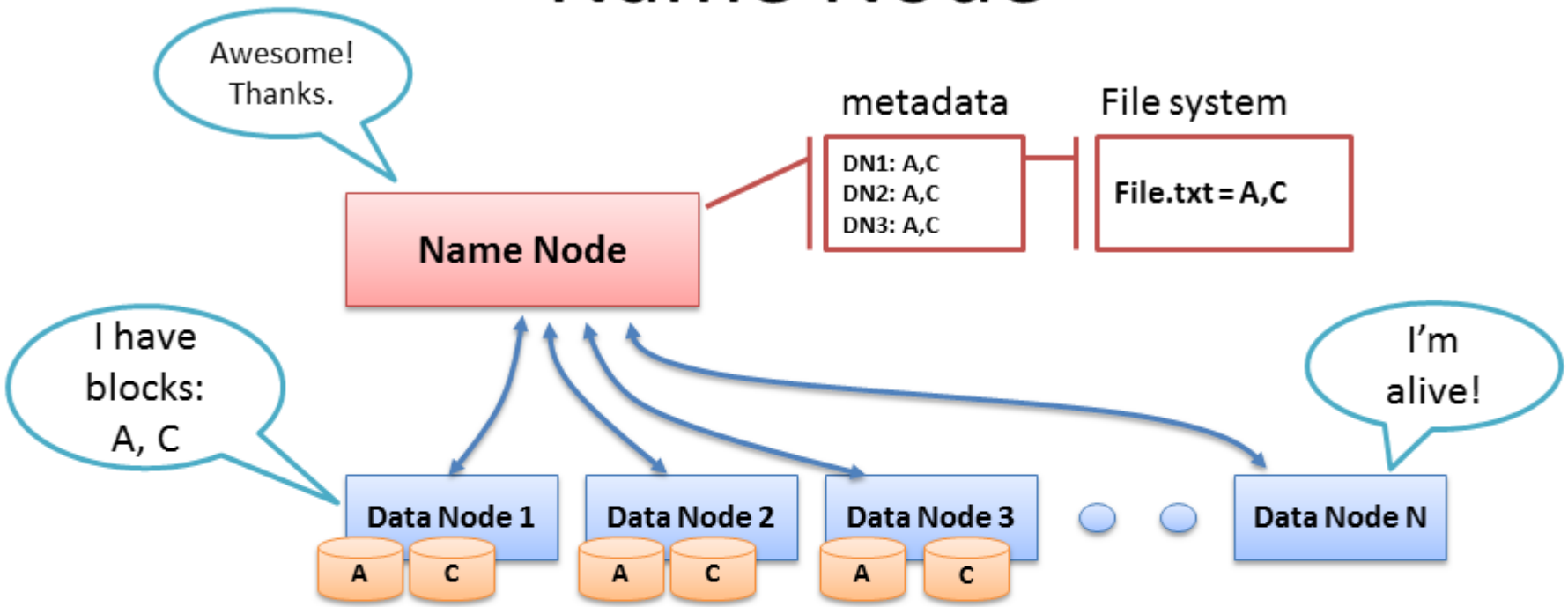
Hadoop 2.0 vs Hadoop 1.0

Client reading files from HDFS



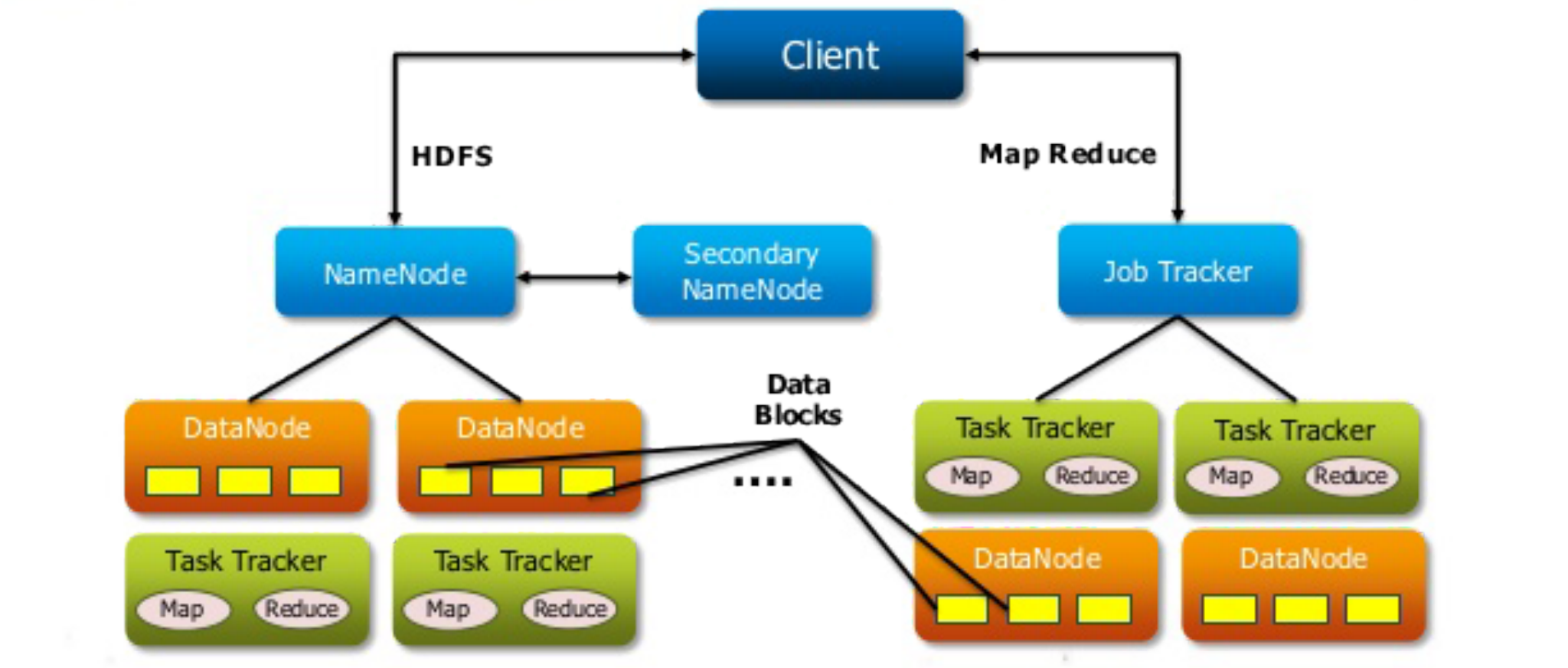
- Client receives Data Node list for each block
- Client picks first Data Node for each block
- Client reads blocks sequentially

Name Node

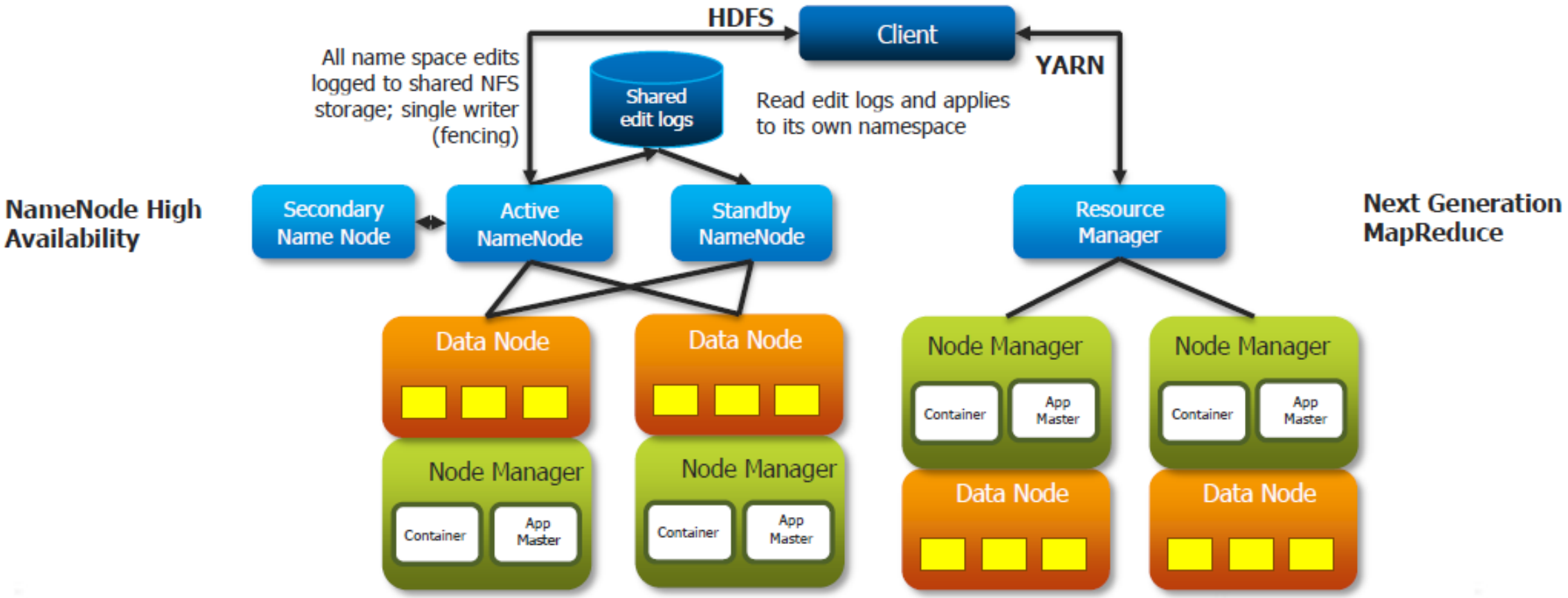


- Data Node sends Heartbeats
- Every 10th heartbeat is a Block report
- Name Node builds metadata from Block reports
- TCP – every 3 seconds
- If Name Node is down, HDFS is down

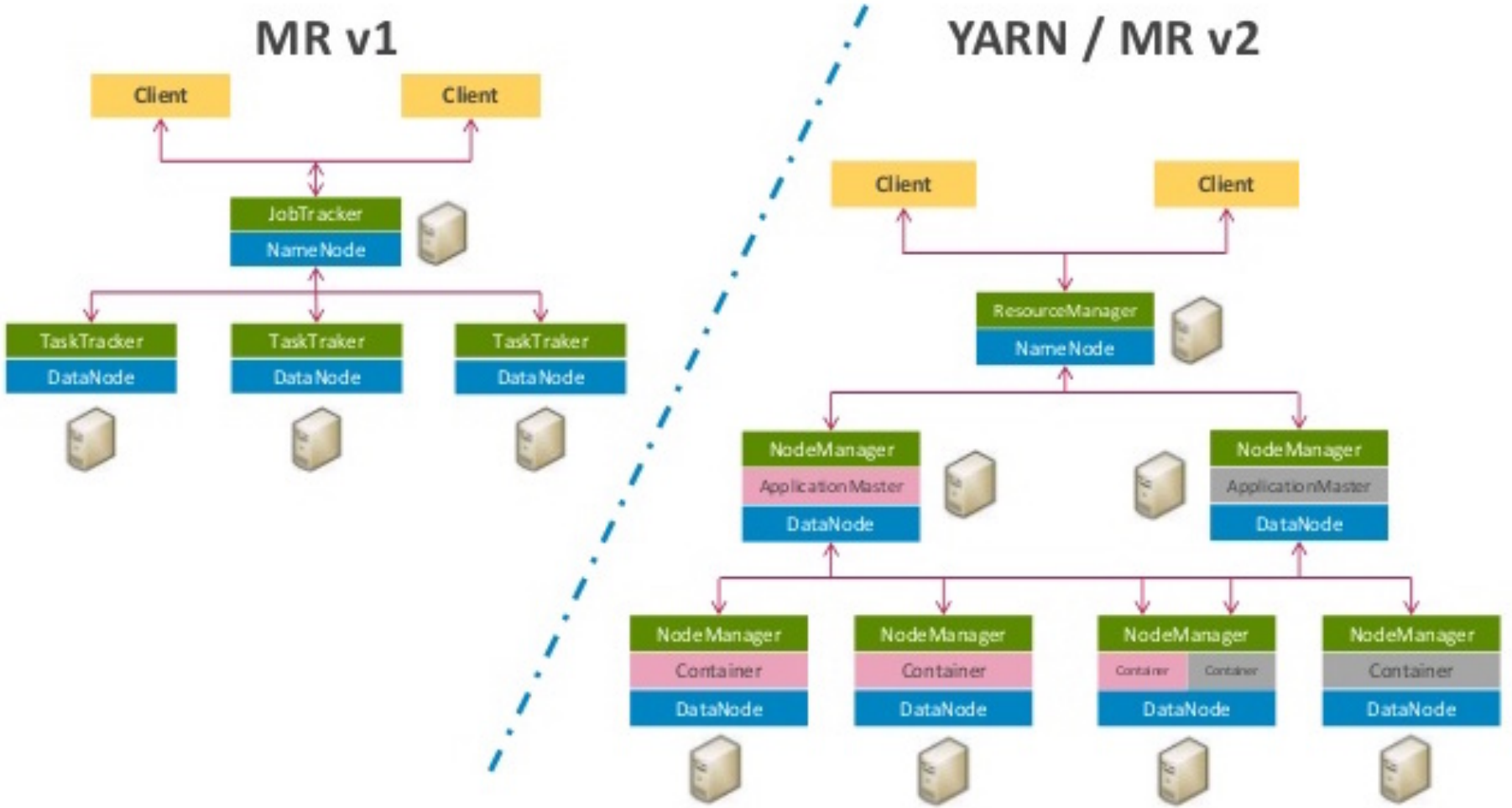
Hadoop 1.0 Bottlenecks: HDFS / MapReduce



Hadoop 2.0 Architechture

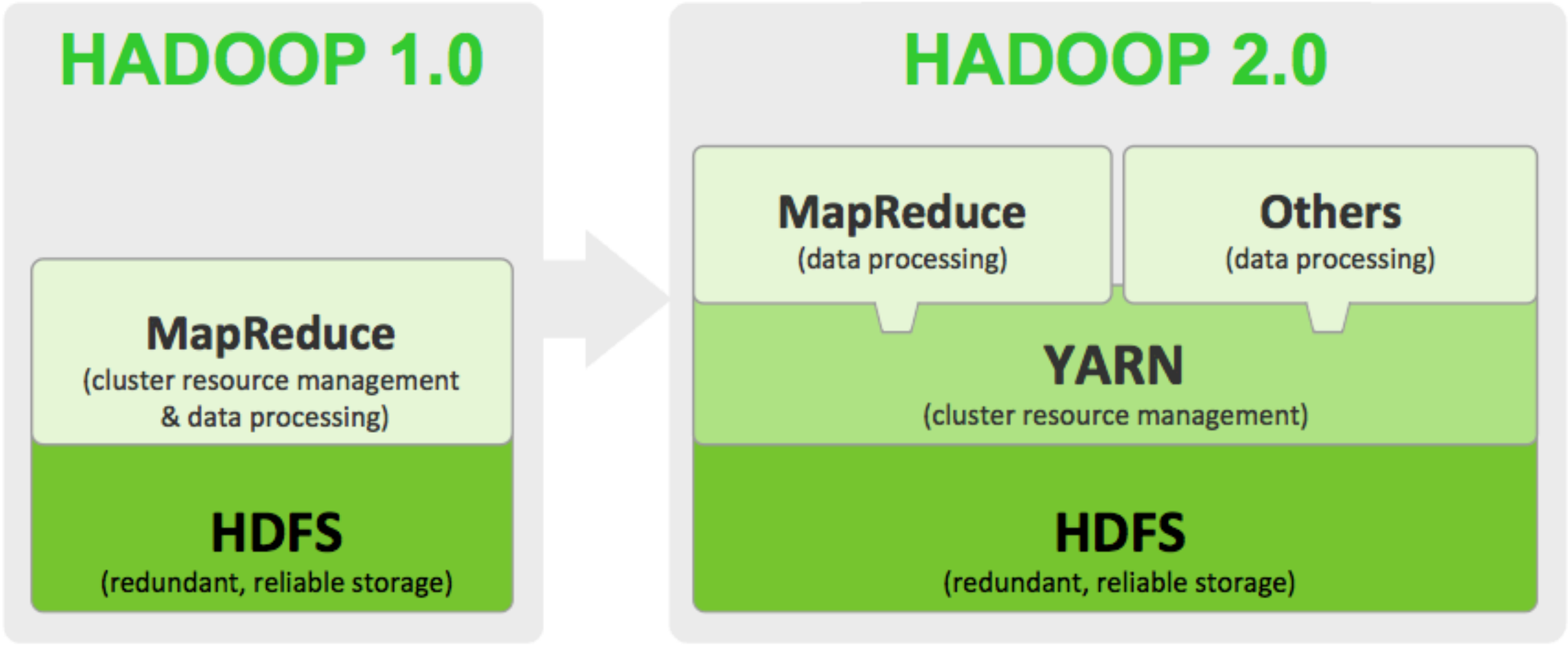


YARN/MR v2 vs MR v1 Architecture



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

Hadoop 2.0 vs Hadoop 1.0 – Processing



The Hadoop Ecosystem

TEZ

Apache Ambari
http://incubator.apache.org/ambari

Spark

APACHE FALCON

sqoop

Apache orc

HIVE

OOZIE

APACHE KNOX

Hadoop

Apache Solr

APACHE HBASE

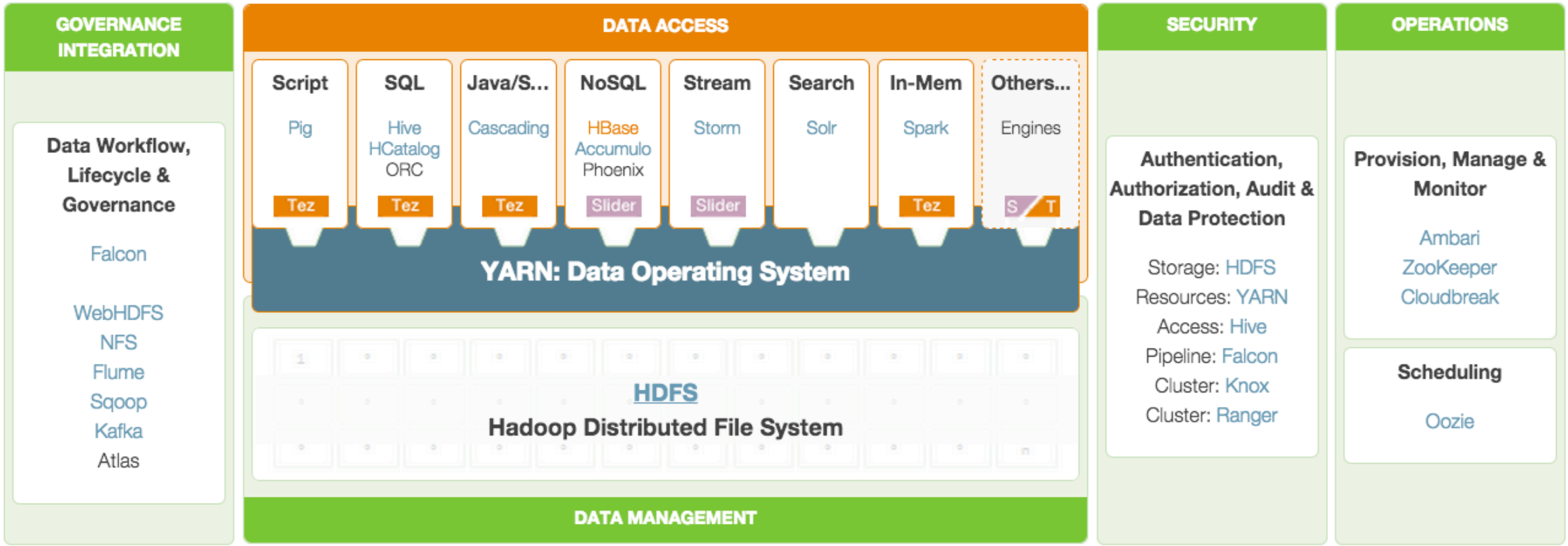
STORM

accumulo

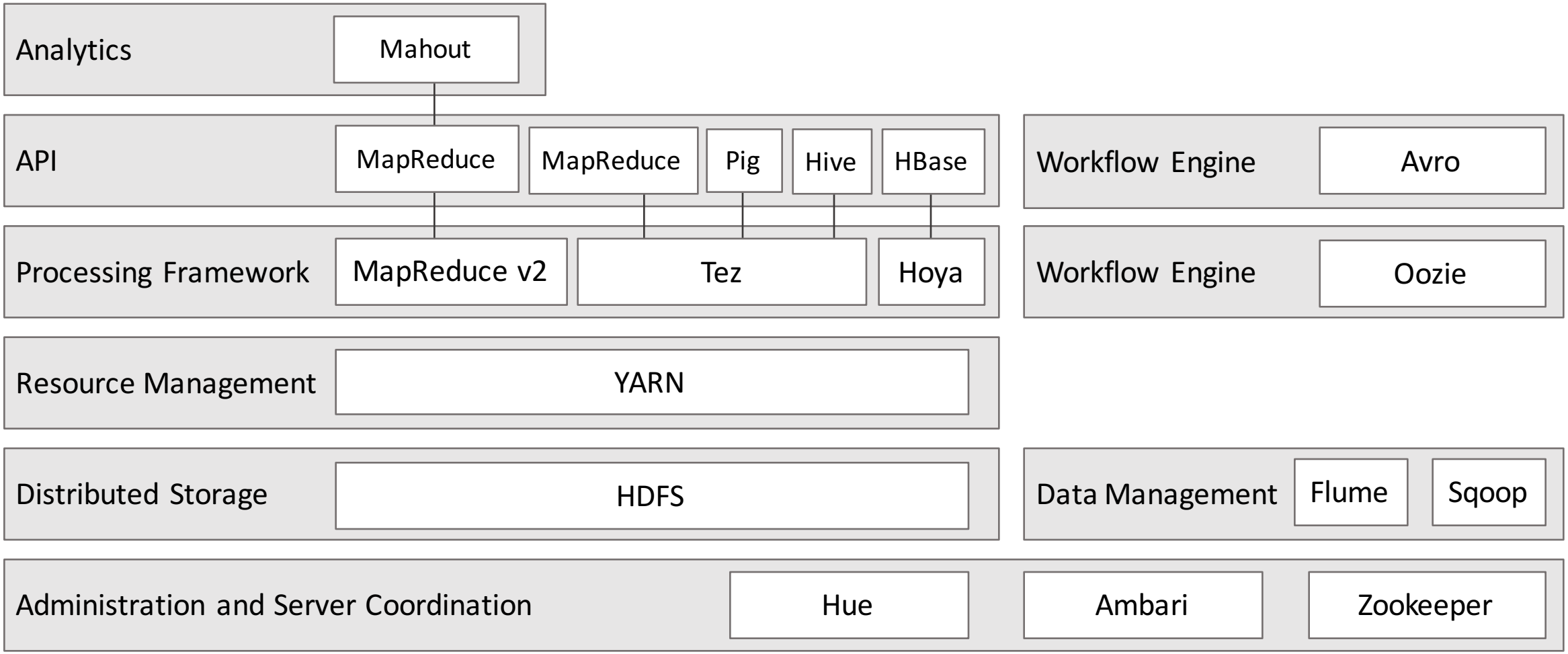
kafka

FLUME

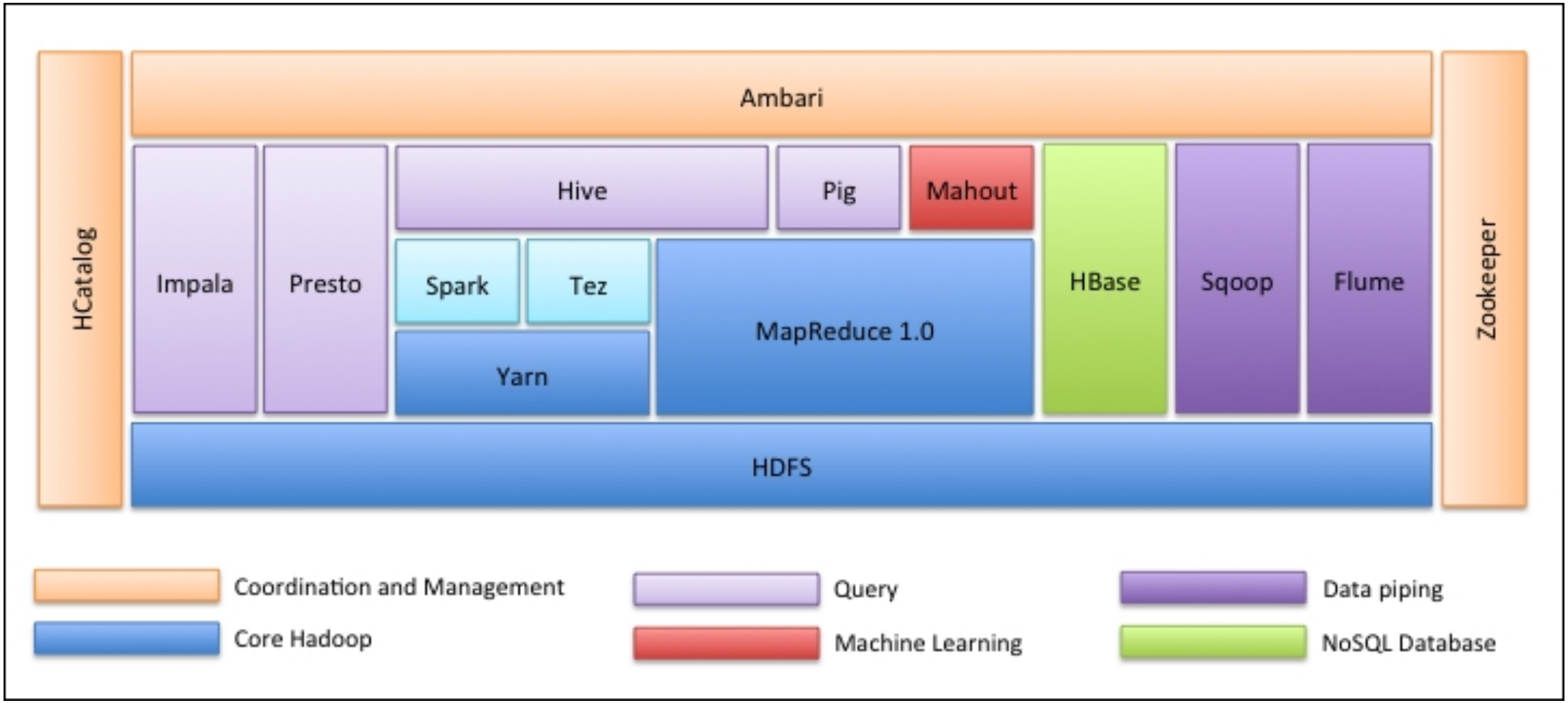
Hortonworks Hadoop Distribution



Classification of Hadoop Ecosystem Components



Classification of Hadoop Ecosystem Components



Hadoop Ecosystem Components

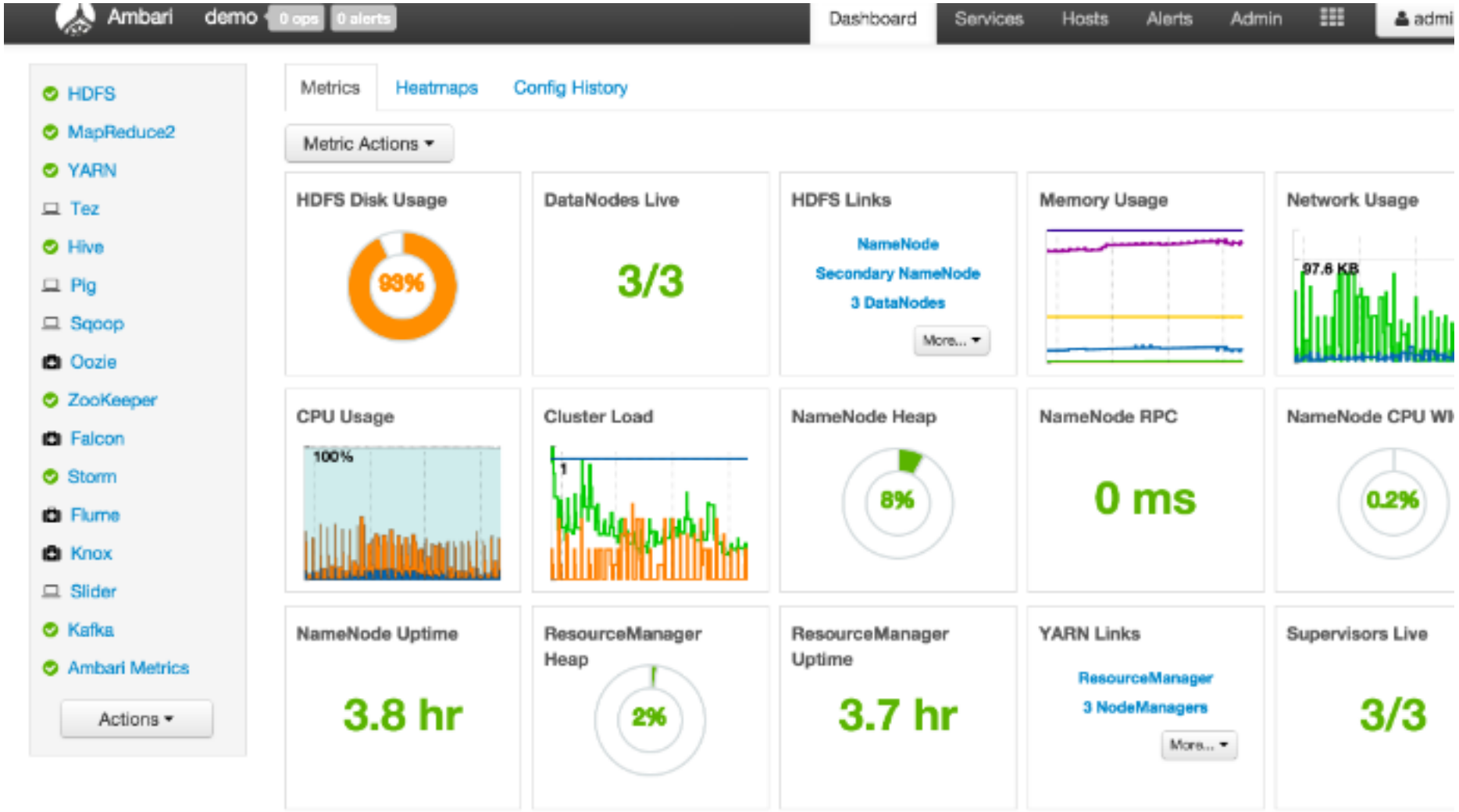
Data Management Frameworks

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

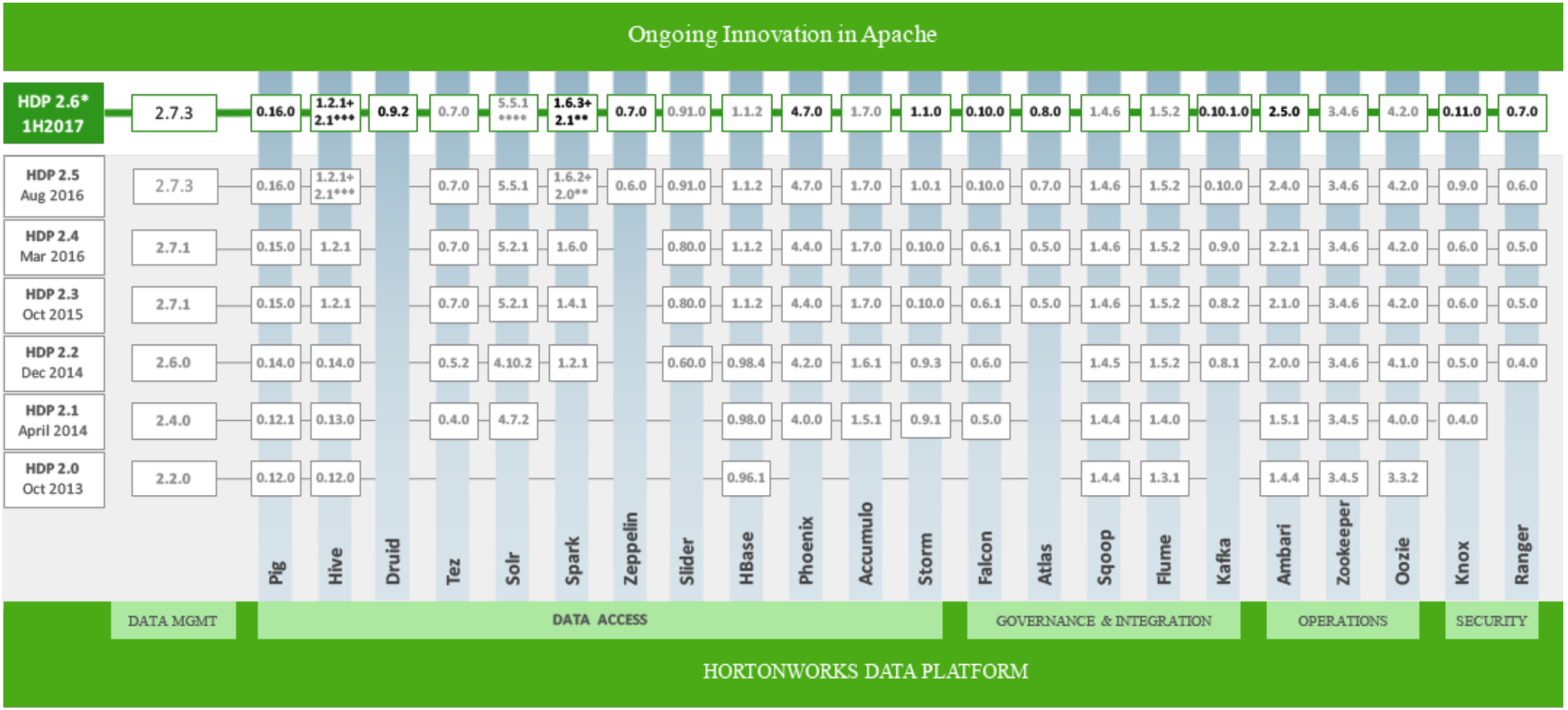
Governance and Integration Frameworks

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

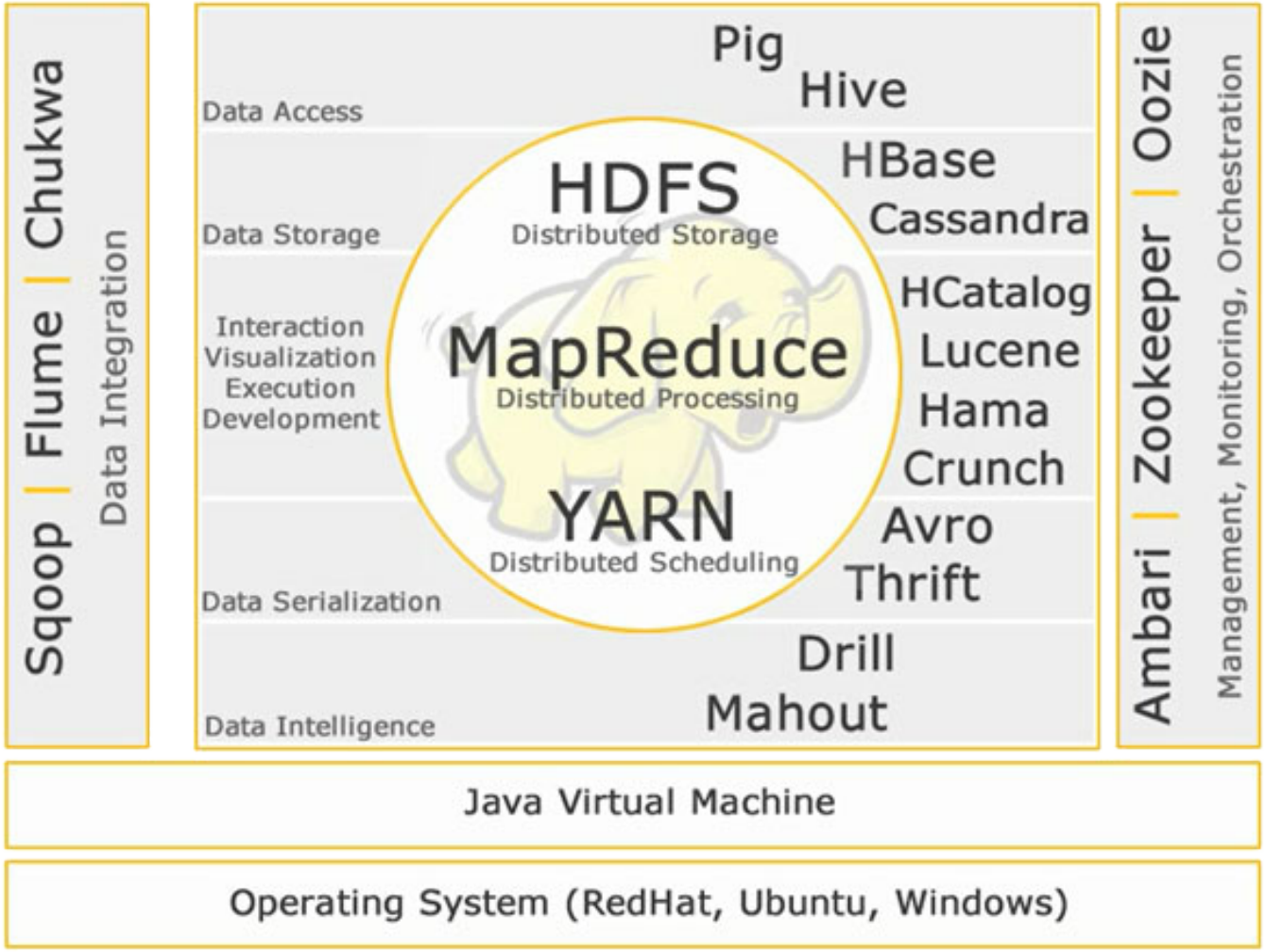
Ecosystem Component Versions



** 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.*

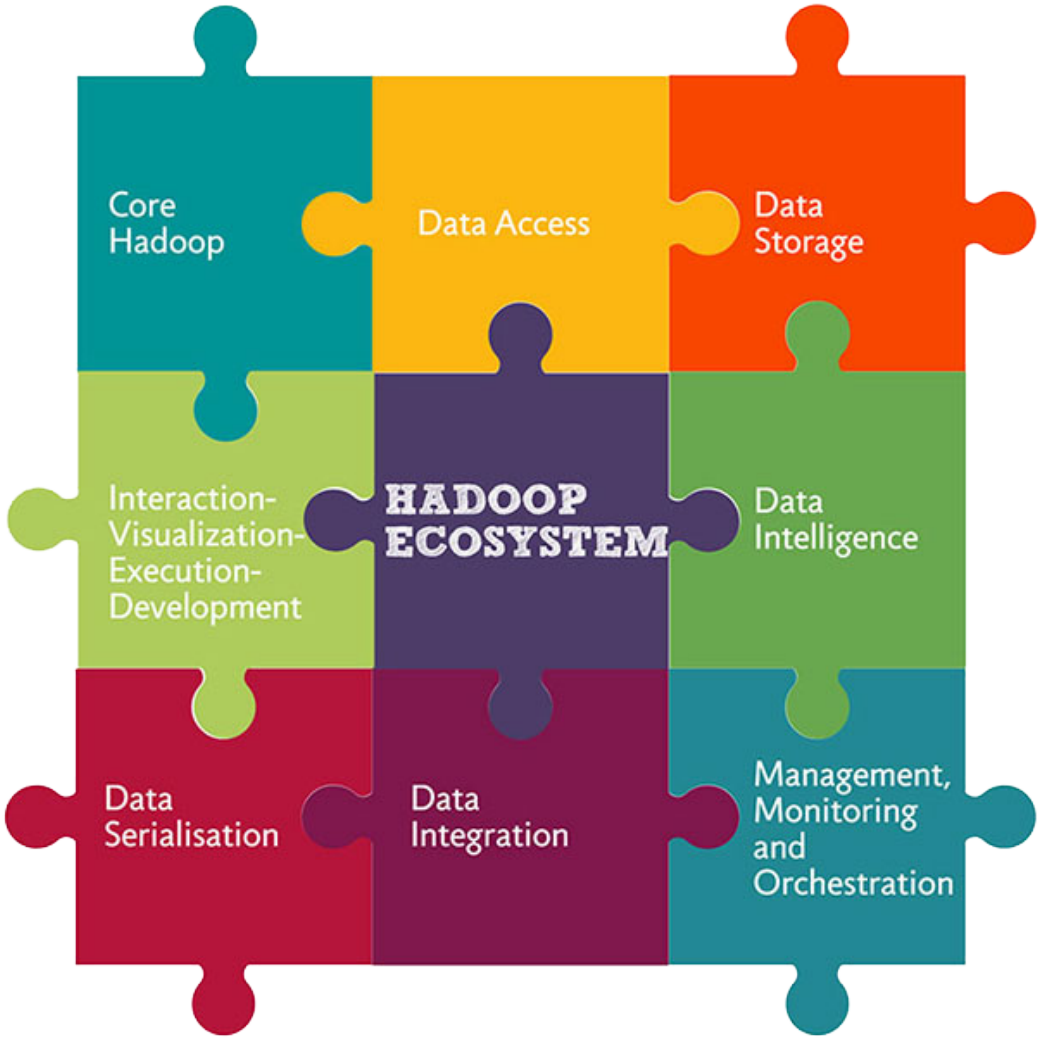
Hadoop Ecosystem Components' Key Features

HADOOP ECOSYSTEM COMPONENTS



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

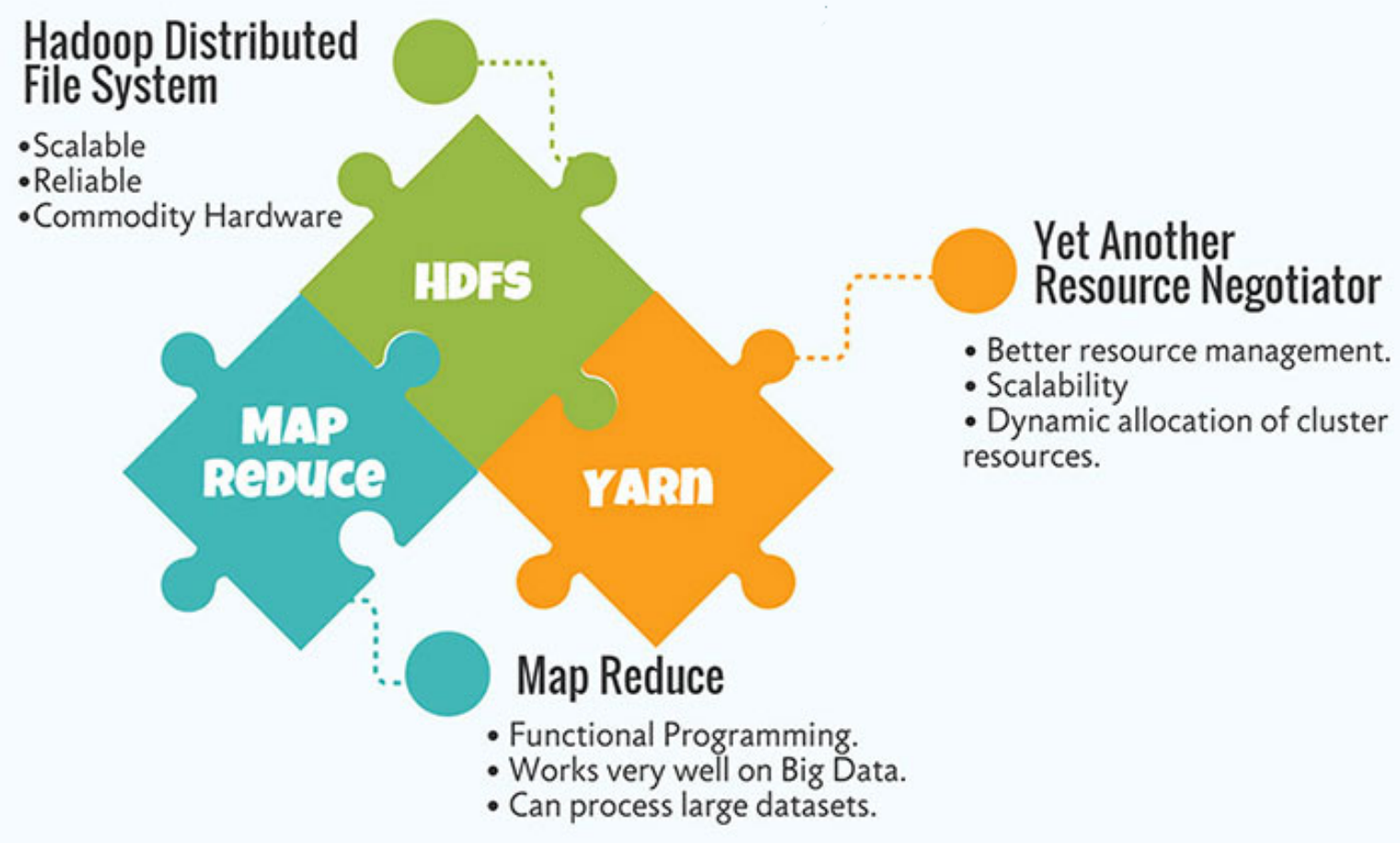
Classification of the Hadoop Ecosystem Components



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

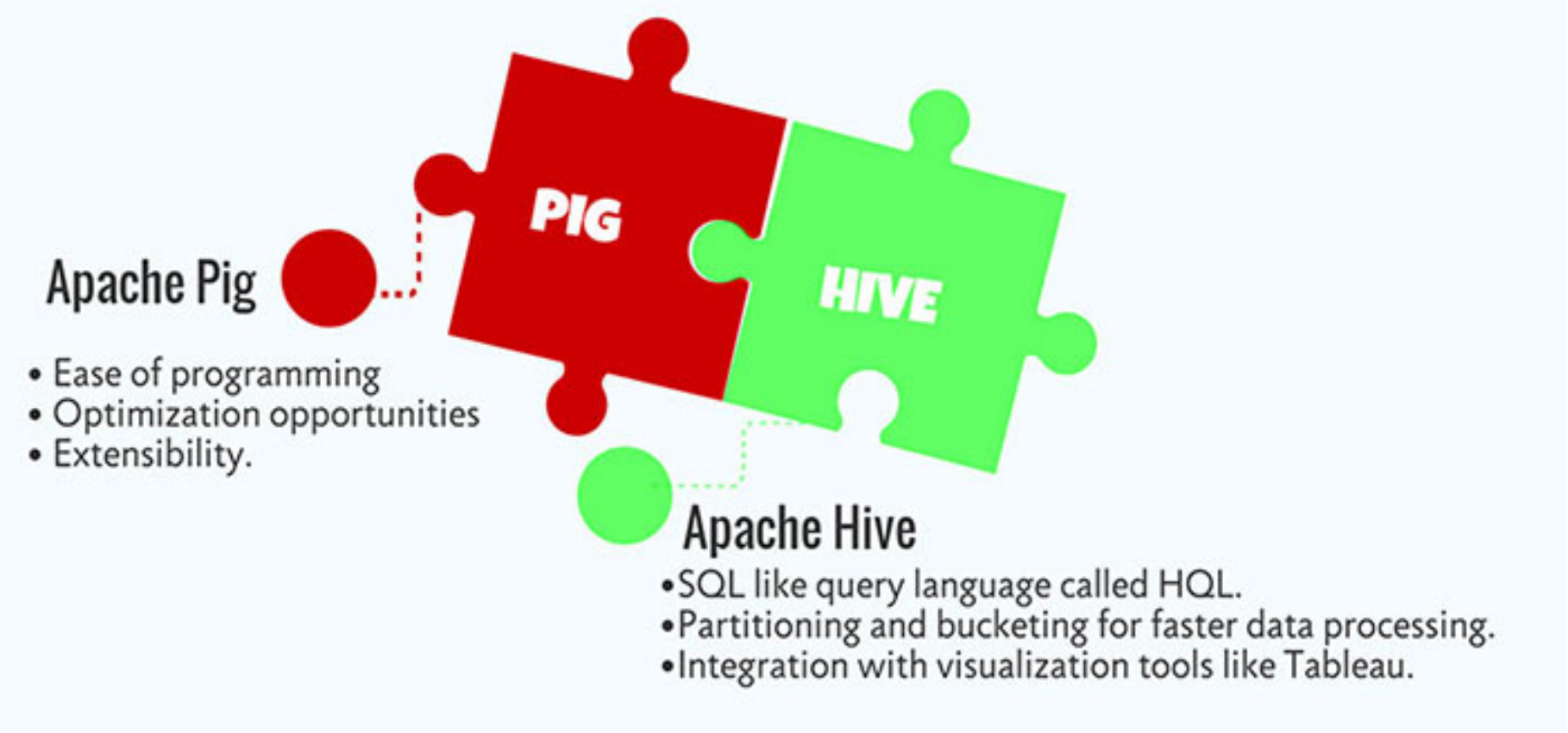


Hadoop Distributed File System (HDFS)
Stands for: managing big data sets with High Volume, Velocity and Variety.

Map Reduce
Stands for: processing high volume distributed data

Yet Another Resource Negotiator (YARN)
Stands for: resource management, job scheduling and monitoring

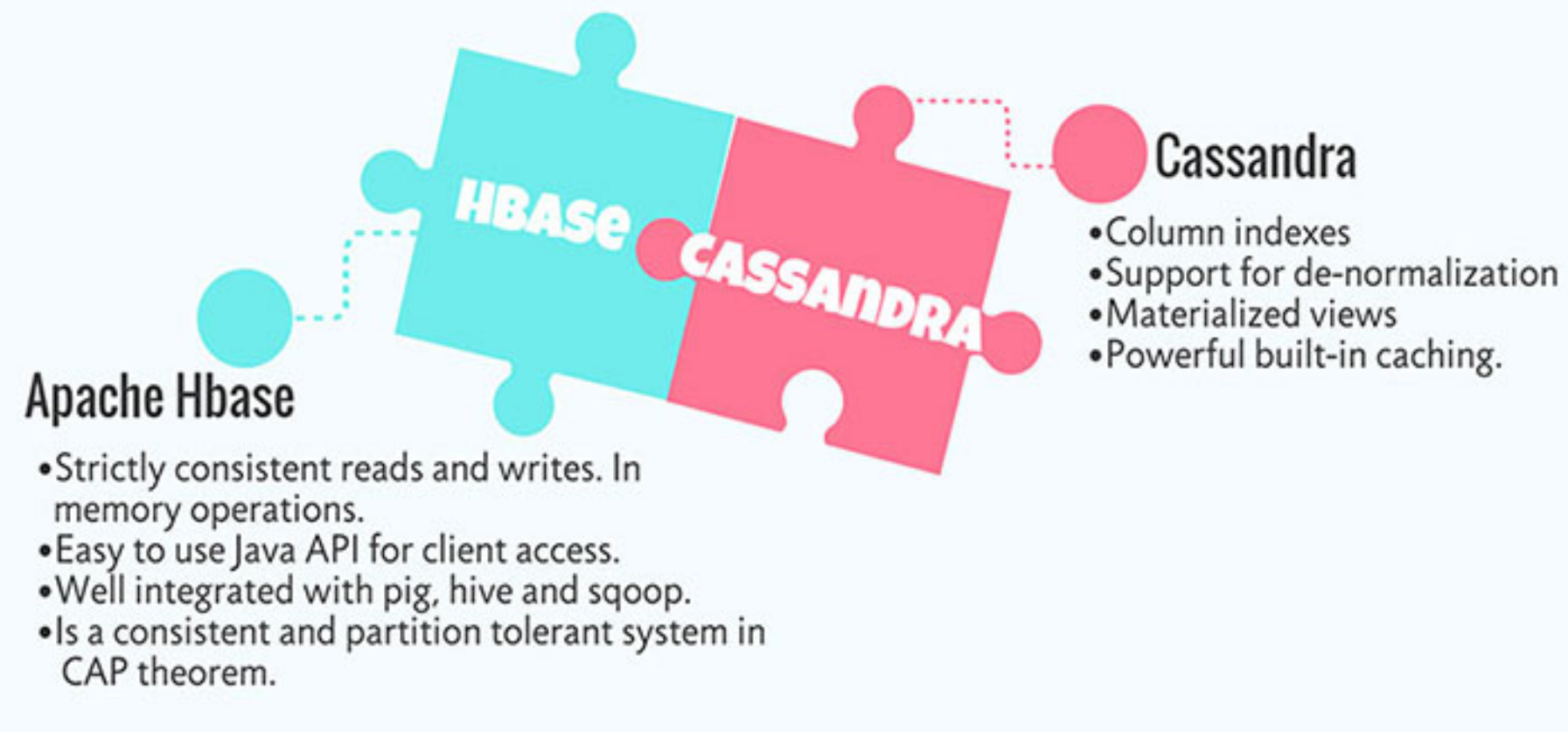
DATA ACCESS



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.

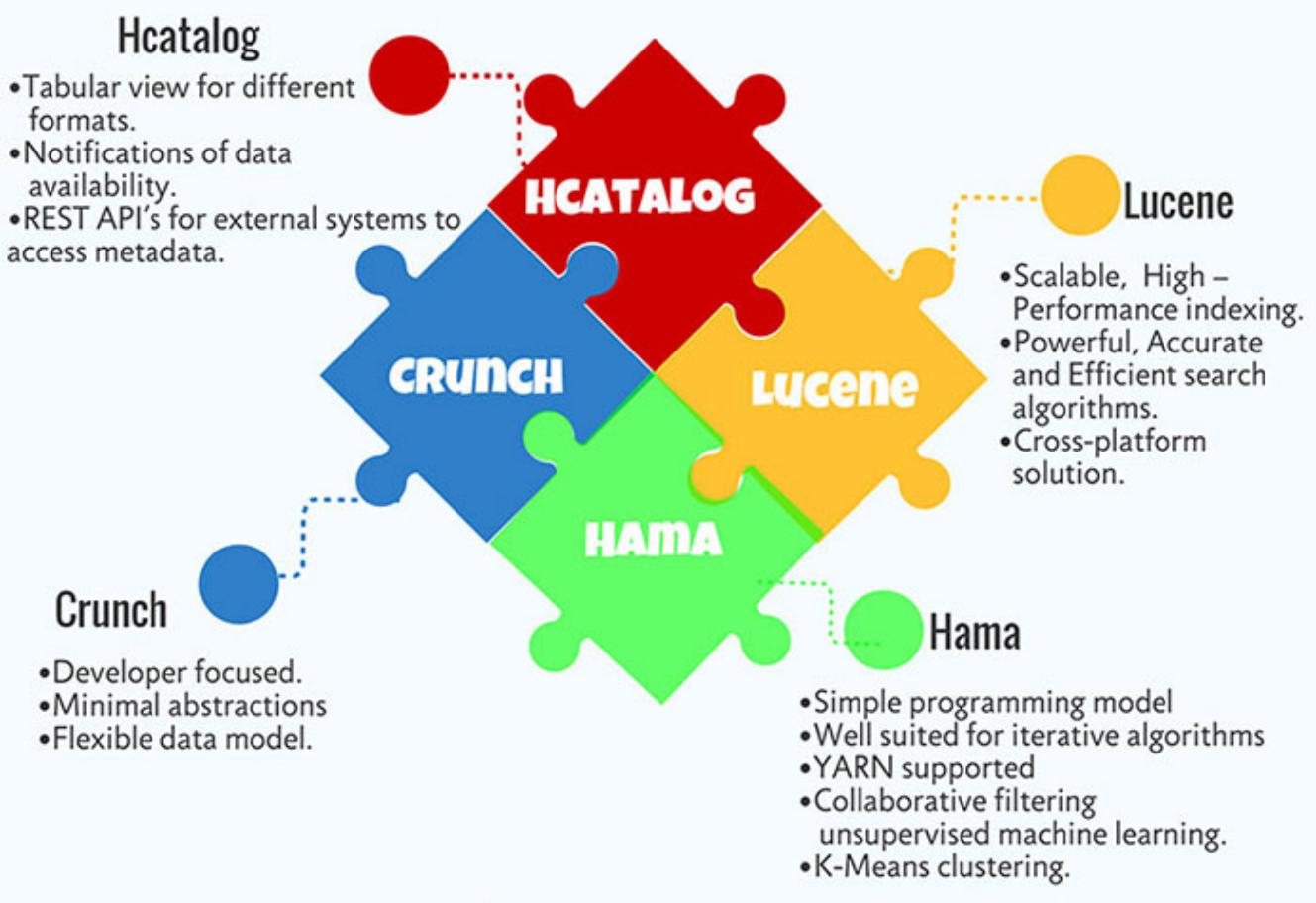
DATA STORAGE



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

Casandra
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.

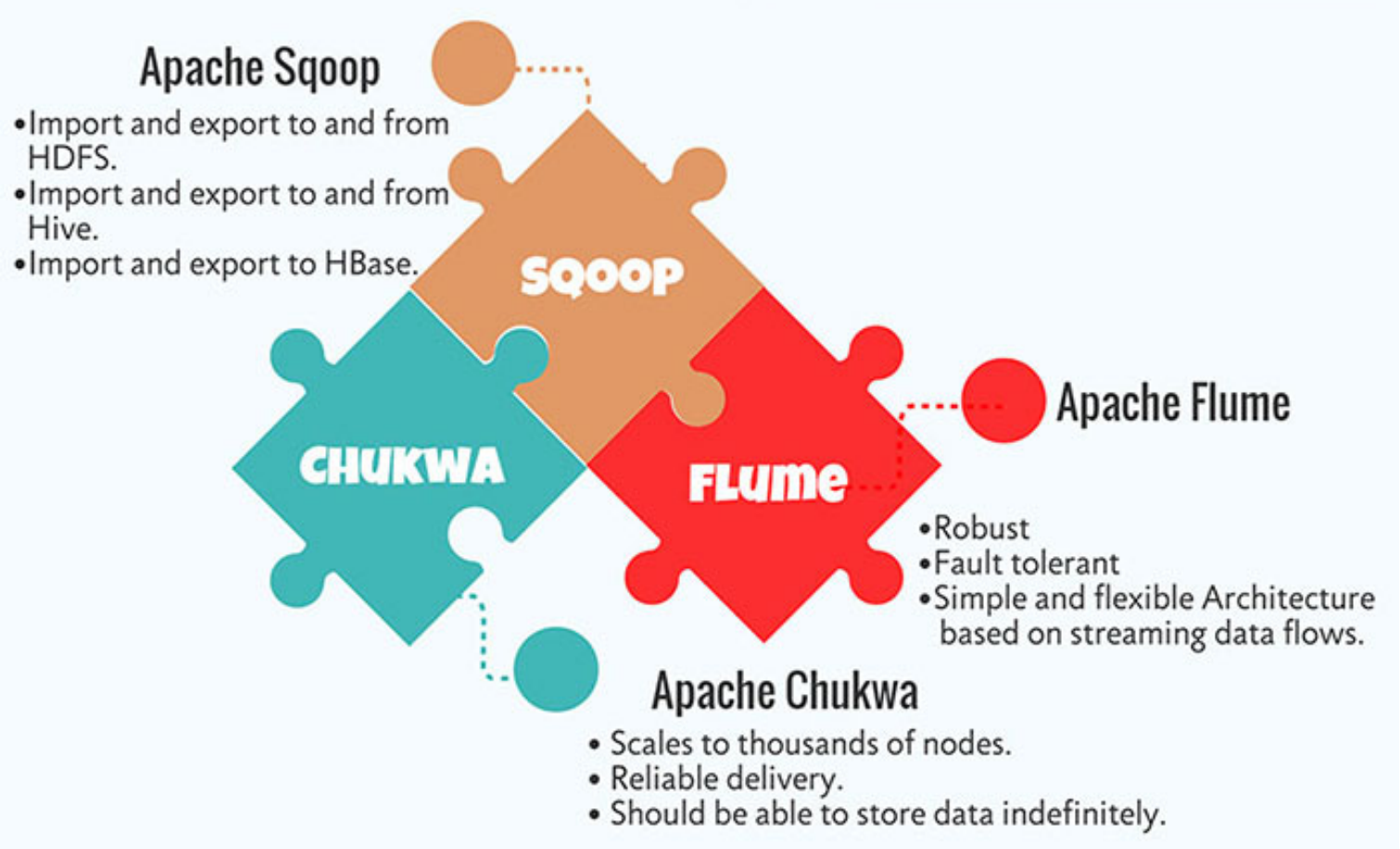
Crunch
Stands for: writing, testing and running MapReduce pipelines.

DATA INELLIGENCE

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

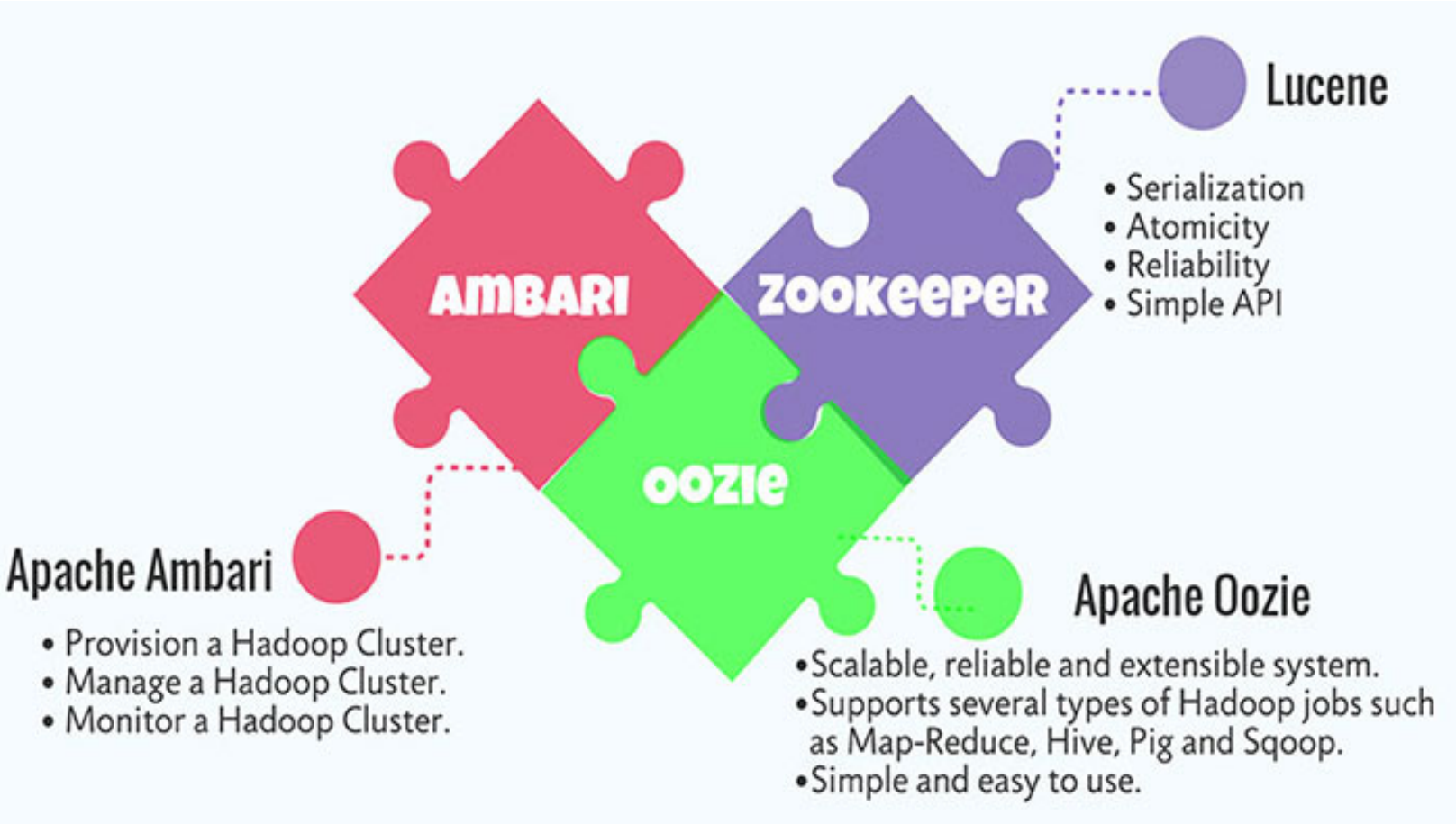


Apache Sqoop
Stands for: low latency SQL query engine for Hadoop and NoSQL.

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

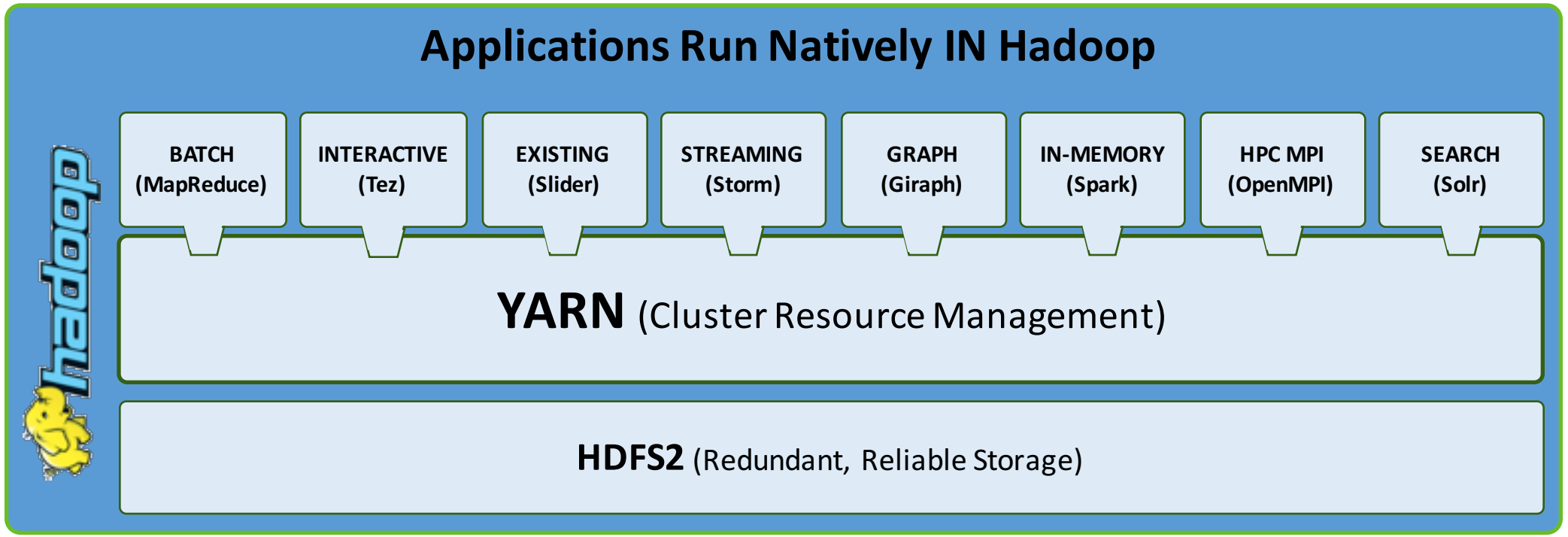
- Predict customer churn
- Predict equipment failure
- Customer behavior analysis

Oil & Gas

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

YARN as a Data Operating System

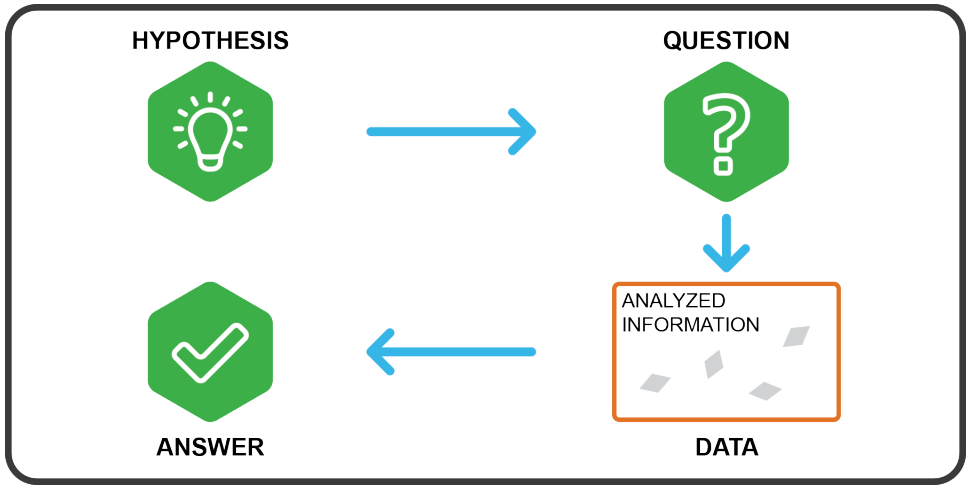
Applications now run “in” Hadoop,
instead of “on” Hadoop.



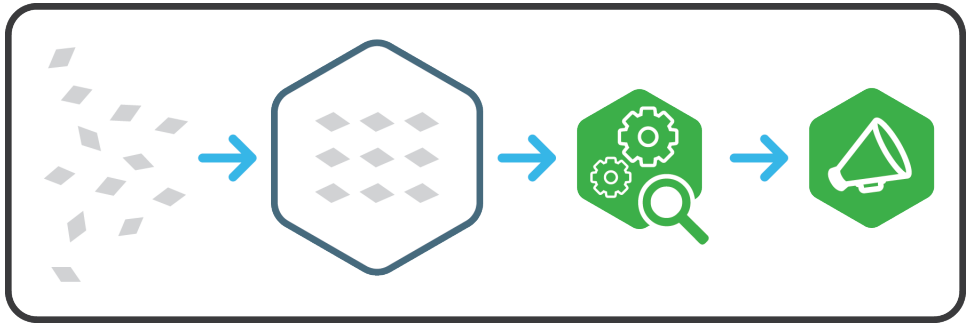
Modern Data Applications approach to Insights

Traditional Analytics

Structured & Repeatable
Structure built to store data



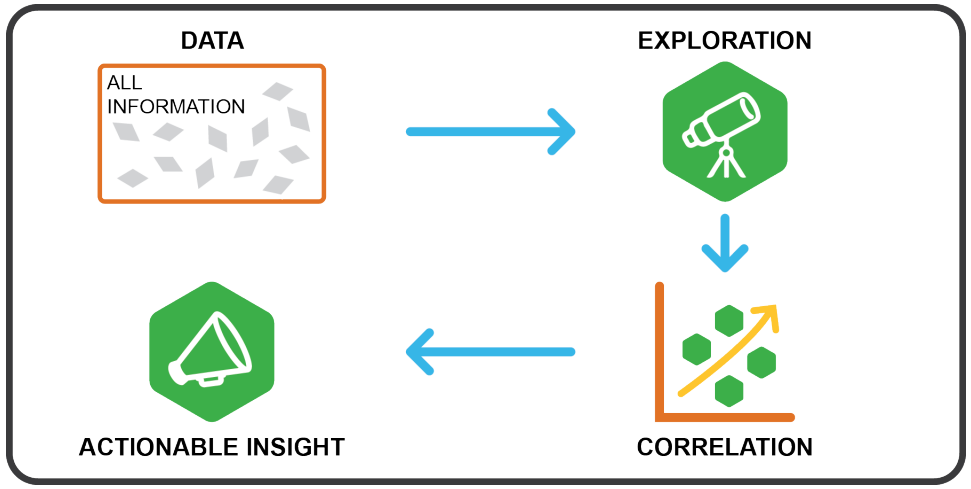
Start with hypothesis
Test against selected data



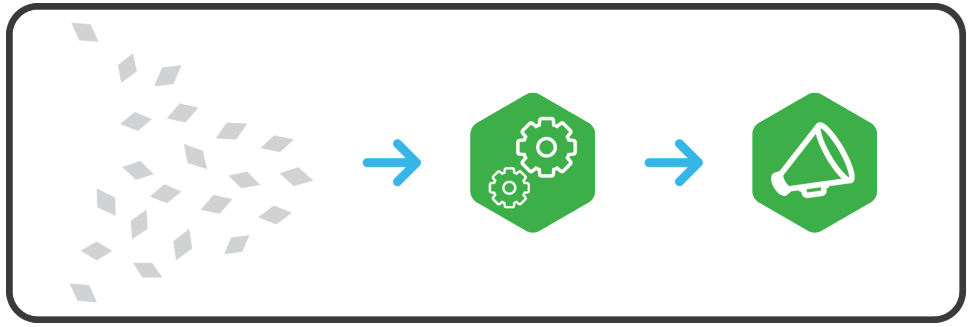
Analyze after landing…

Next Generation Analytics

Iterative & Exploratory
Data is the structure



Data leads the way
Explore *all* data, identify correlations



Analyze in motion…

Q & A

Abzetdin Adamov, Assoc Prof.
Email me at: aadamov@ada.edu.az
Follow me at: @
Link to me at: www.linkedin.com/in/adamov
Visit my blog at: aadamov.wordpress.com