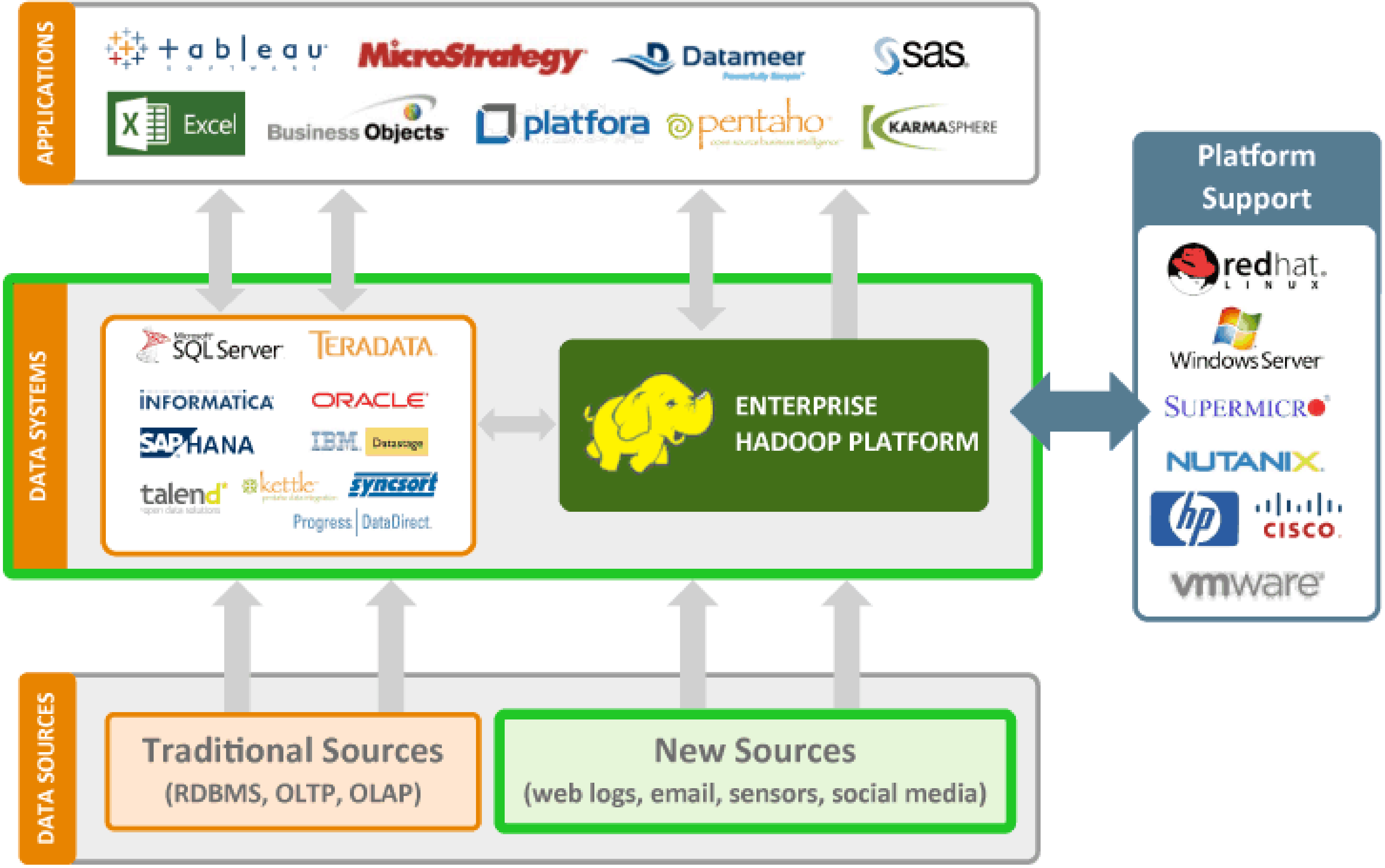
Hadoop…….

- Hadoop was created by computer scientists Doug Cutting and Mike Cafarella in 2006 to support distribution for the Nutch search engine.
- It was inspired by Google's MapReduce, a software framework in which an application is broken down into numerous small parts. Any of these parts, which are also called fragments or blocks, can be run on any node in the cluster.
- After years of development within the open source community, Hadoop 1.0 became publically available in November 2012 as part of the Apache project sponsored by the Apache Software Foundation.



APPLICATIONS

DATA SYSTEMS

ENTERPRISE HADOOP PLATFORM

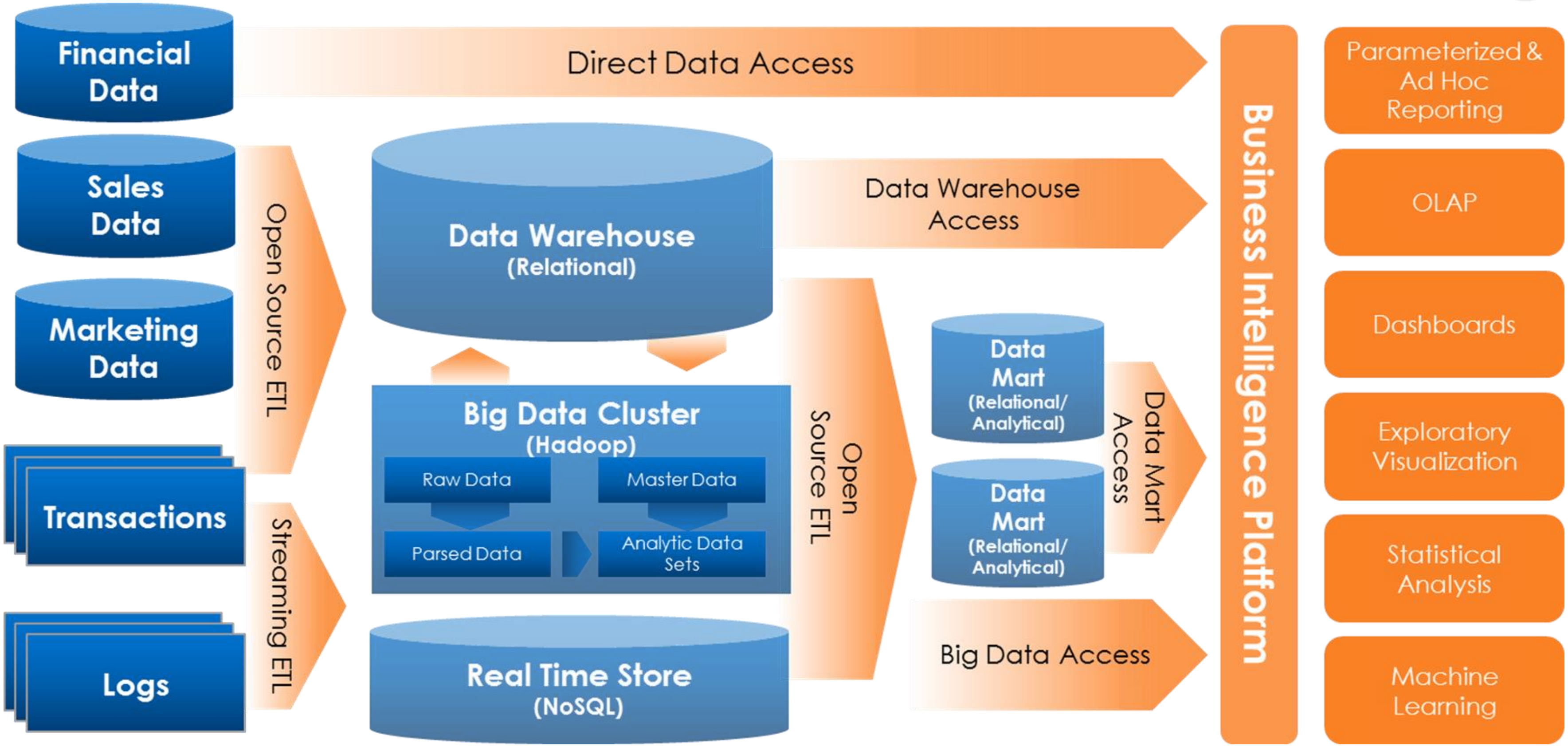
DATA SOURCES

Traditional Sources (RDBMS, OLTP, OLAP)

New Sources (web logs, email, sensors, social media)

Platform Support

hadoop



Financial Data

Sales Data

Marketing Data

Transactions

Logs

Direct Data Access

Open Source ETL

Streaming ETL

Data Warehouse (Relational)

Big Data Cluster (Hadoop)

Raw Data

Master Data

Parsed Data

Analytic Data Sets

Real Time Store (NoSQL)

Data Warehouse Access

Open Source ETL

Data Mart (Relational/ Analytical)

Data Mart (Relational/ Analytical)

Data Mart Access

Big Data Access

Business Intelligence Platform

Parameterized & Ad Hoc Reporting

OLAP

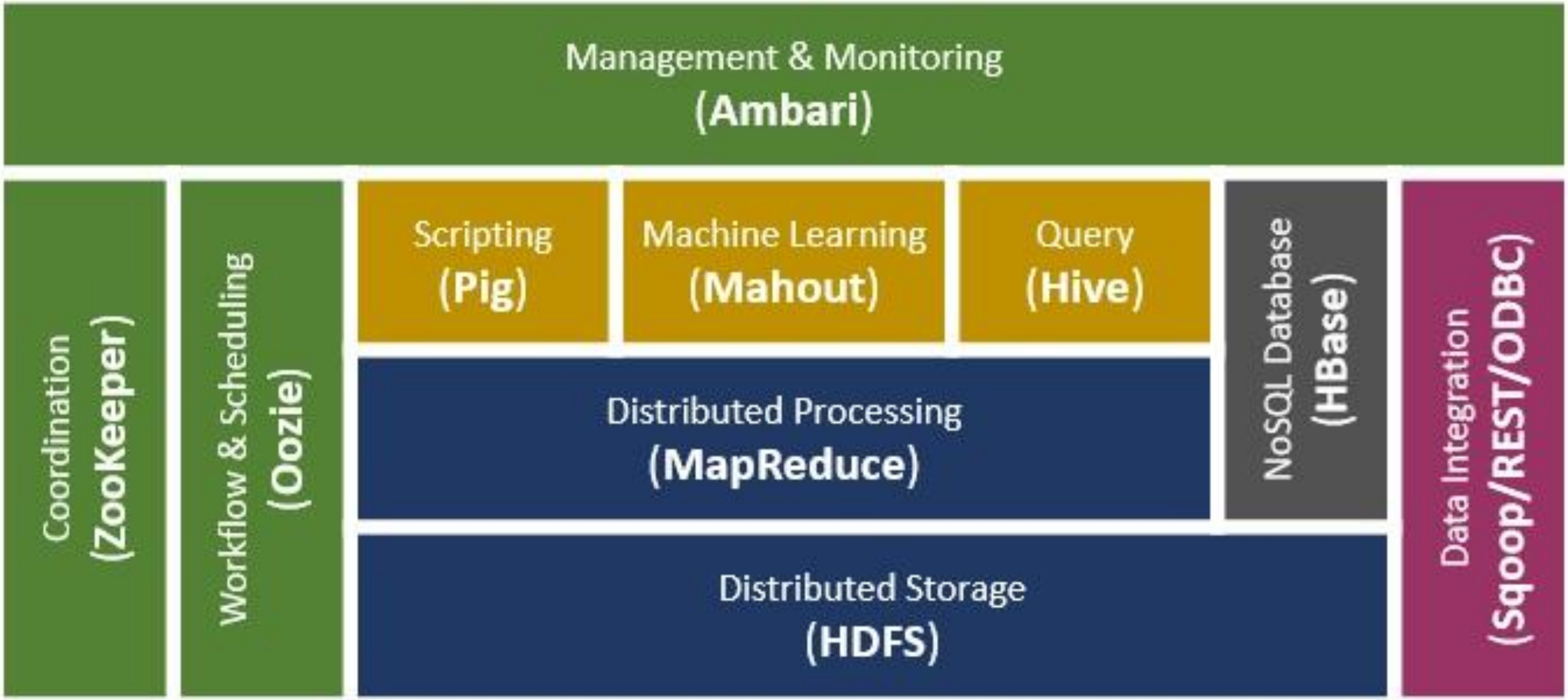
Dashboards

Exploratory Visualization

Statistical Analysis

Machine Learning

Apache Hadoop Ecosystem



Management & Monitoring
(Ambari)

Coordination
(ZooKeeper)

Workflow & Scheduling
(Oozie)

Scripting
(Pig)

Machine Learning
(Mahout)

Query
(Hive)

NoSQL Database
(HBase)

Data Integration
(Sqoop/REST/ODBC)

Distributed Processing
(MapReduce)

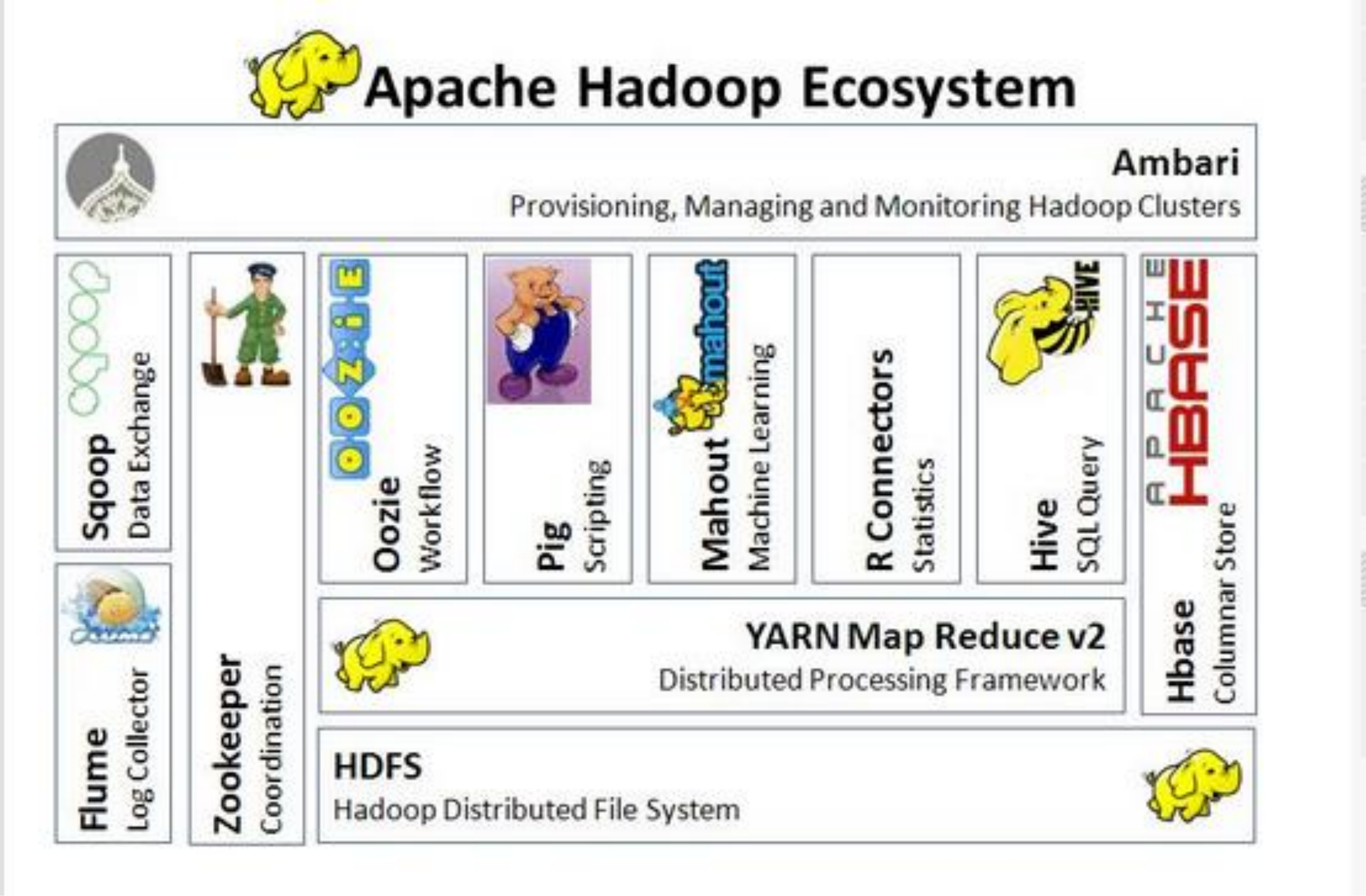
Distributed Storage
(HDFS)

Hadoop.....

- **Flume**. A tool used to collect, aggregate and move huge amounts of streaming data into HDFS.
- **HBase**. An open source, nonrelational, distributed database;
- **Hive**. A data warehouse that provides data summarization, query and analysis;
- **Cloudera Impala**. A massively parallel processing database for Hadoop, originally created by the software company Cloudera, but now released as open source software;
- **Oozie**. A server-based workflow scheduling system to manage Hadoop jobs;

Hadoop…..

- **Apache Phoenix**. An open source, massively parallel processing, relational database engine for Hadoop that is based on Apache HBase;
- **Pig**. A high-level platform for creating programs that run on Hadoop;
- **Sqoop**. A tool to transfer bulk data between Hadoop and structured data stores, such as relational databases;
- **Spark**. A fast engine for big data processing capable of streaming and supporting SQL, machine learning and graph processing;
- **Apache Storm**. An open source data processing system; and
- **ZooKeeper**. An open source configuration, synchronization and naming registry service for large distributed systems.



Apache Hadoop Ecosystem

Ambari
Provisioning, Managing and Monitoring Hadoop Clusters

Sqoop
Data Exchange

Flume
Log Collector

Zookeeper
Coordination

Oozie
Workflow

Pig
Scripting

Mahout
Machine Learning

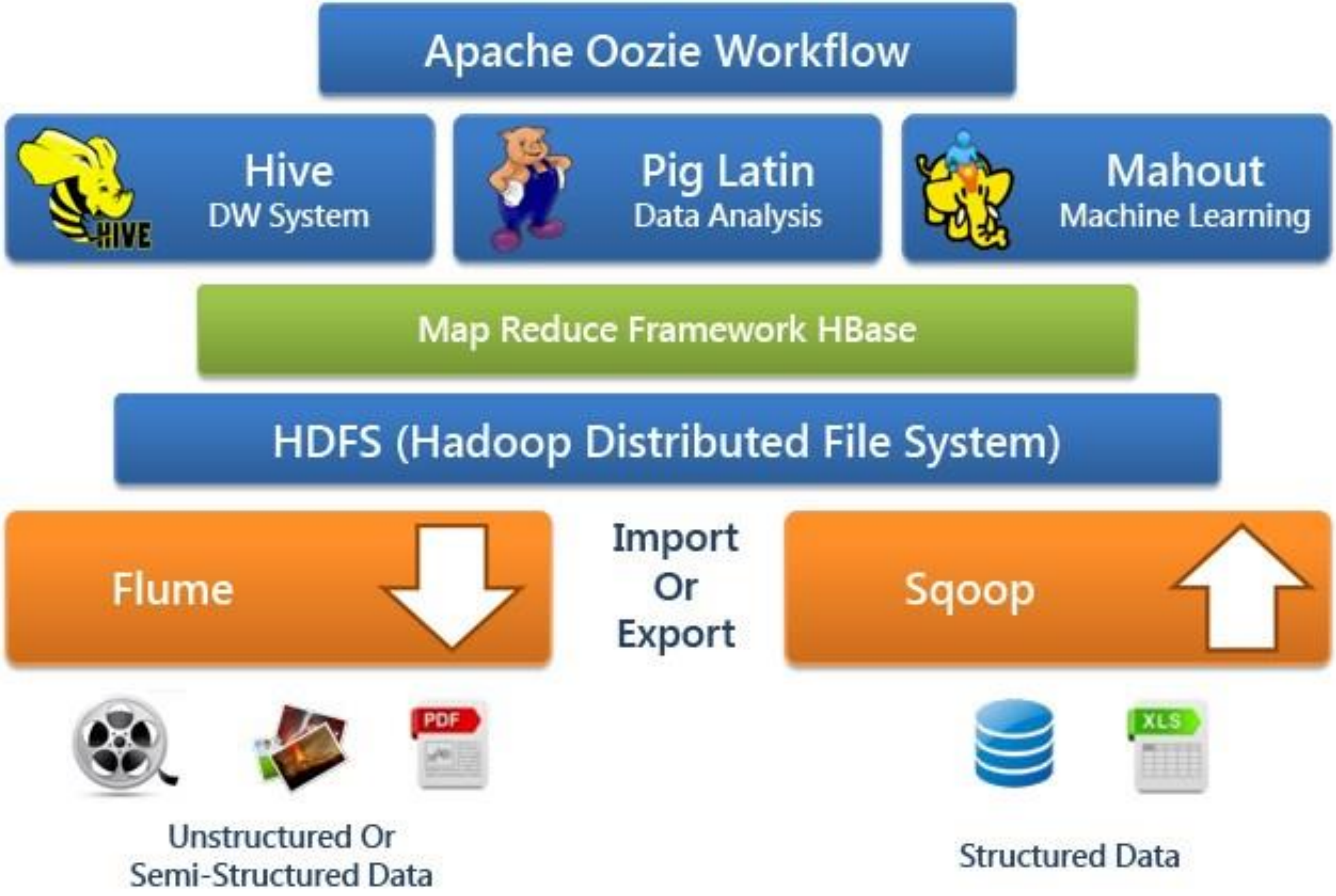
R Connectors
Statistics

Hive
SQL Query

Hbase
Columnar Store

YARN Map Reduce v2
Distributed Processing Framework

HDFS
Hadoop Distributed File System



Apache Oozie Workflow

Hive
DW System

Pig Latin
Data Analysis

Mahout
Machine Learning

Map Reduce Framework HBase

HDFS (Hadoop Distributed File System)

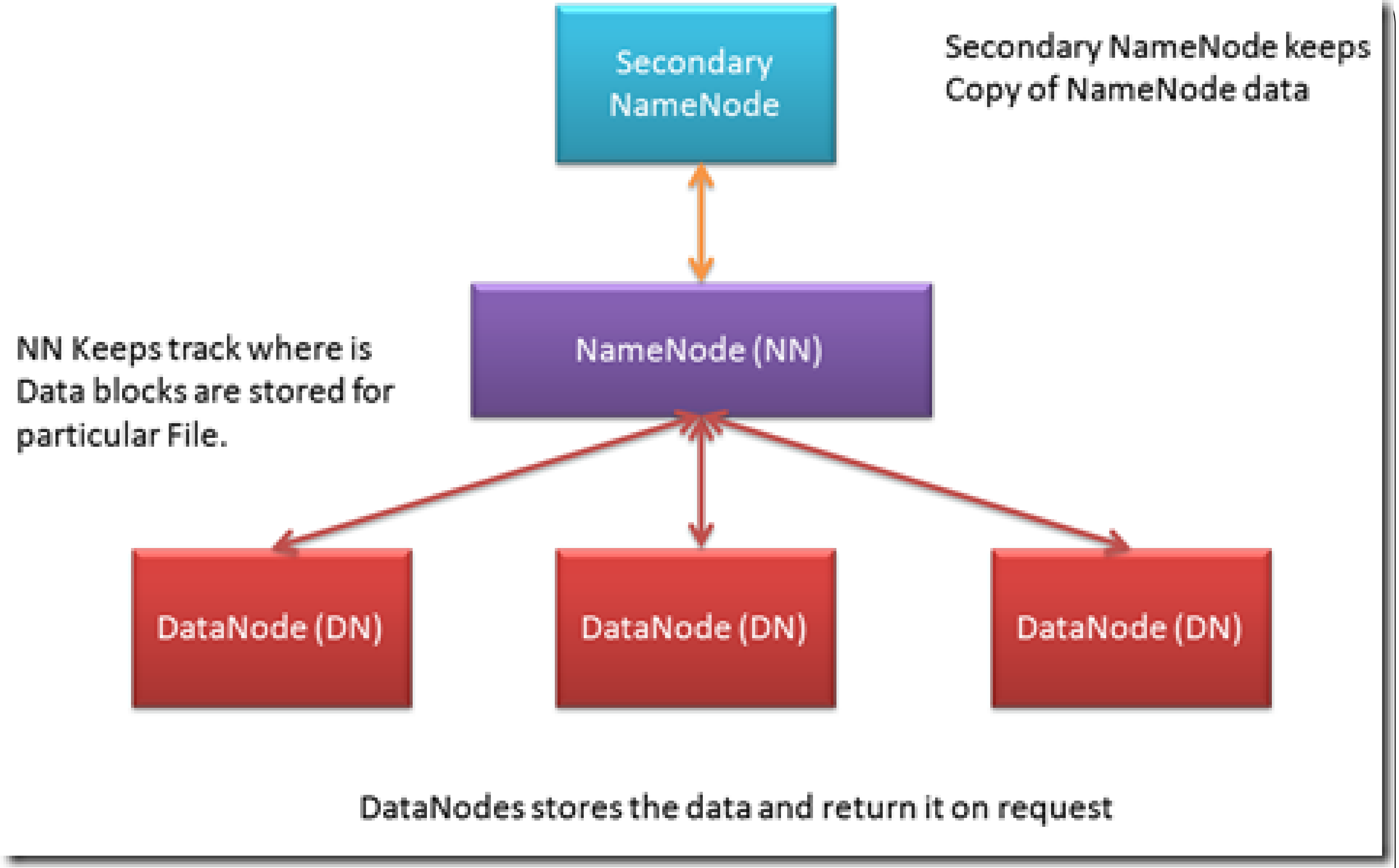
Flume

Import
Or
Export

Sqoop

Unstructured Or
Semi-Structured Data

Structured Data

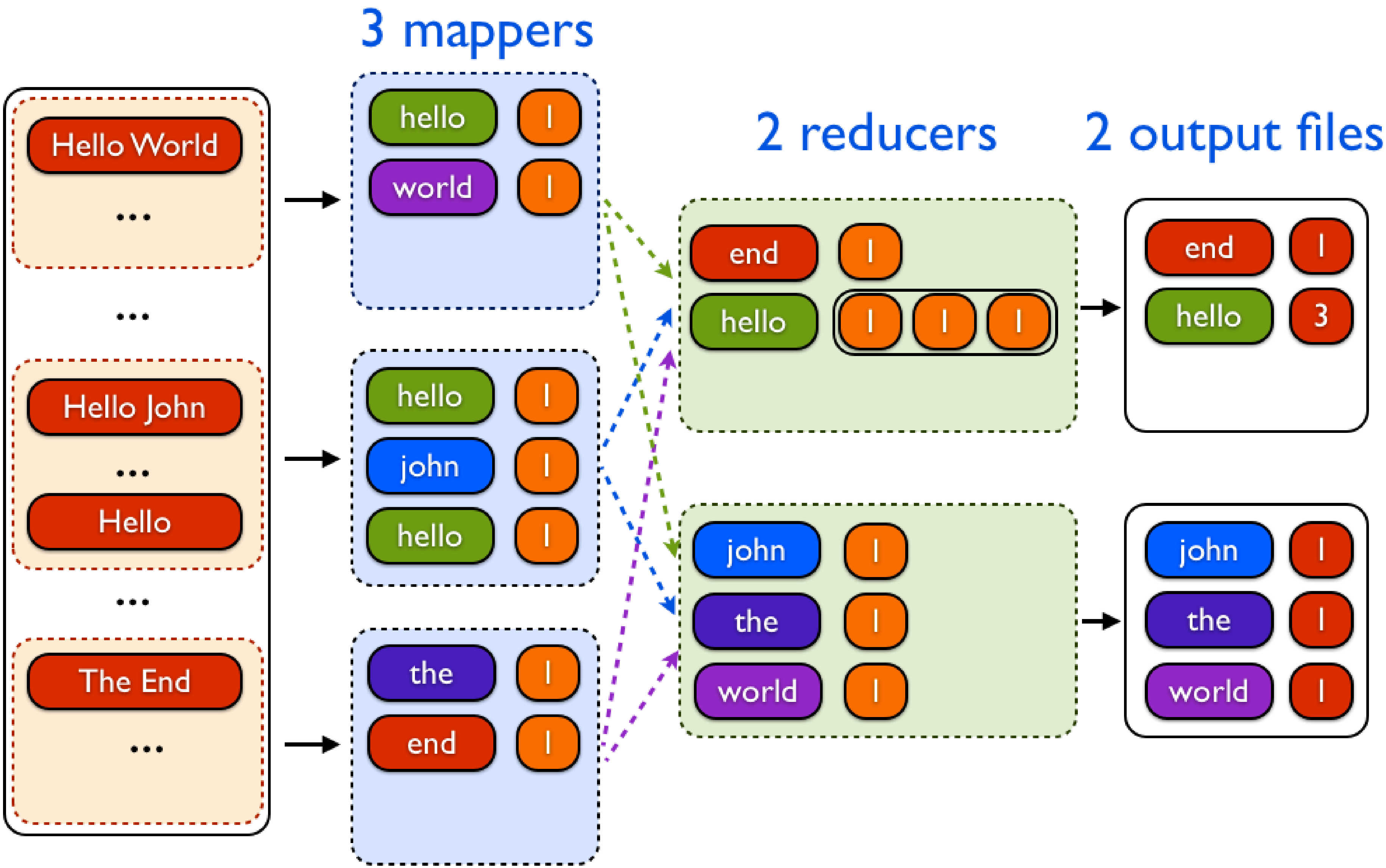


Secondary NameNode keeps Copy of NameNode data

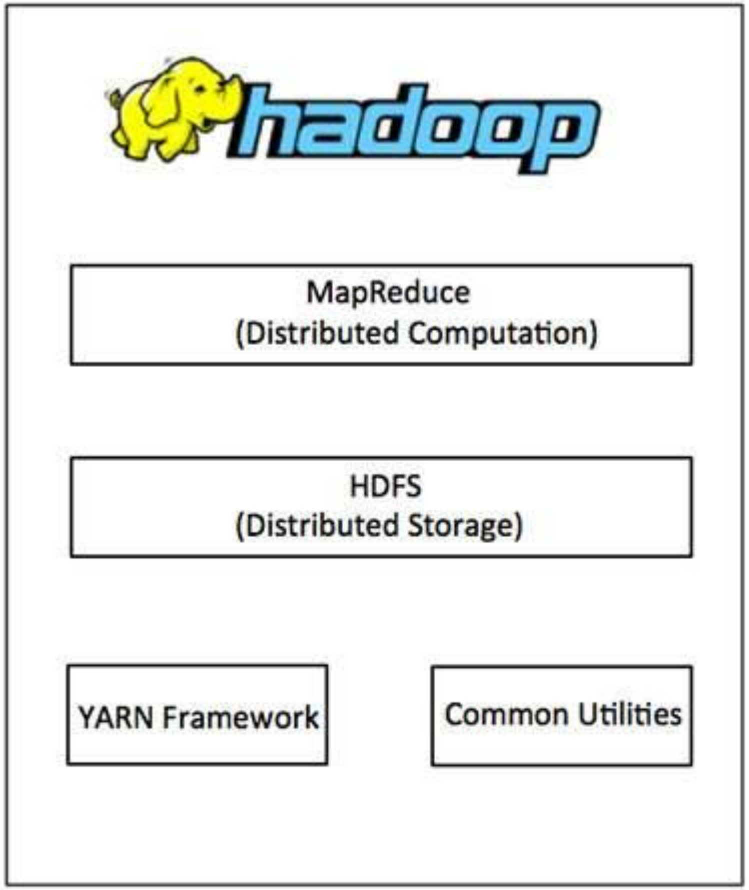
NN Keeps track where is Data blocks are stored for particular File.

DataNodes stores the data and return it on request

Hadoop MapReduce



Hadoop Architecture



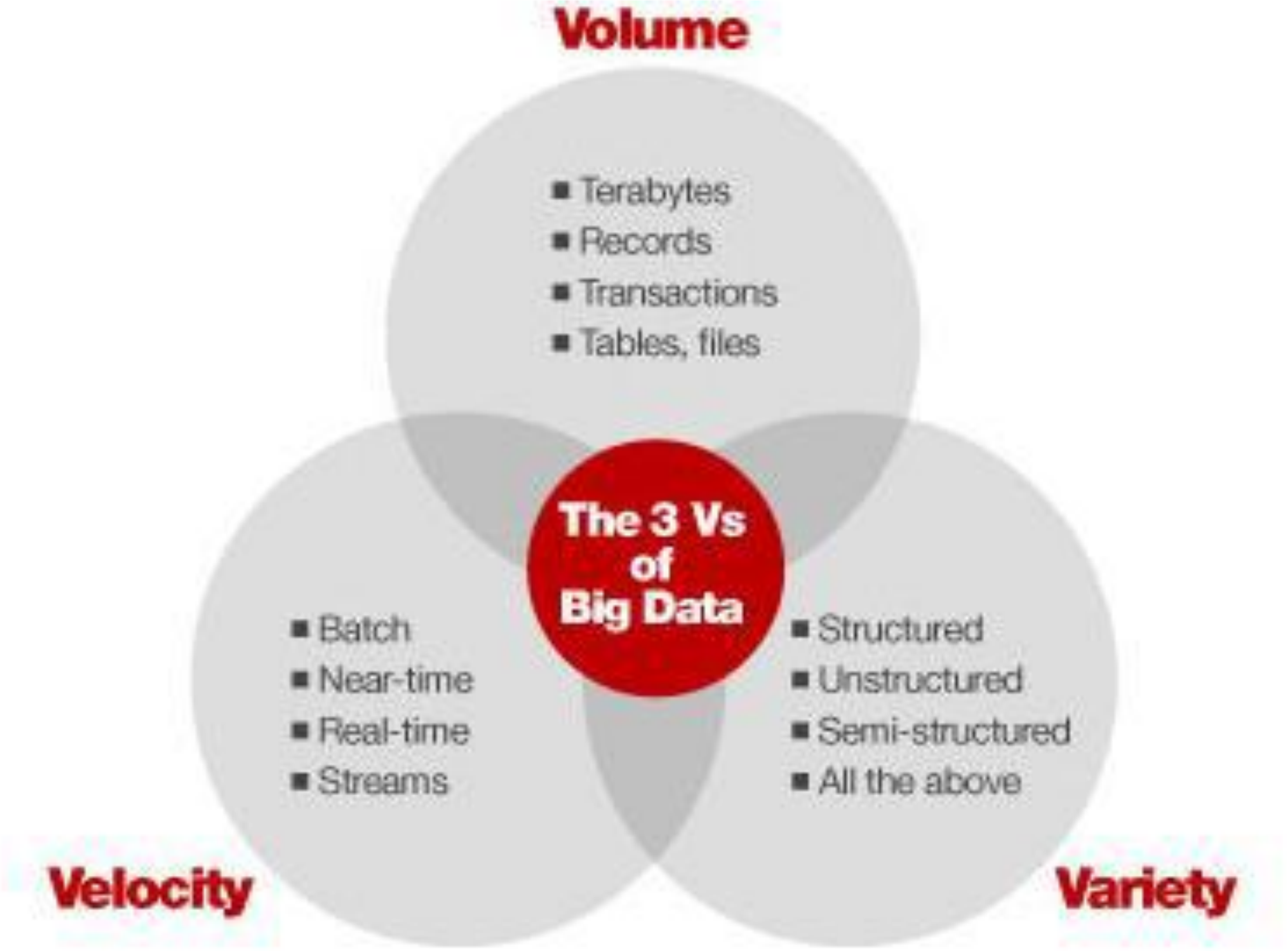
Hadoop Architecture

Hadoop framework includes following four modules:

- **Hadoop Common:** These are Java libraries and utilities required by other Hadoop modules. These libraries provides filesystem and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.
- **Hadoop YARN:** This is a framework for job scheduling and cluster resource management.
- **Hadoop Distributed File System (HDFS™):** A distributed file system that provides high-throughput access to application data.
- **Hadoop MapReduce:** This is YARN-based system for parallel processing of large data sets.

Hadoop Architecture

- Since 2012, the term "Hadoop" often refers not just to the base modules mentioned above but also to the collection of additional software packages that can be installed on top of or alongside Hadoop, such as ***Apache Pig, Apache Hive, Apache HBase, Apache Spark etc.***



The 3 Vs of Big Data

Volume

- Terabytes
- Records
- Transactions
- Tables, files

Velocity

- Batch
- Near-time
- Real-time
- Streams

Variety

- Structured
- Unstructured
- Semi-structured
- All the above