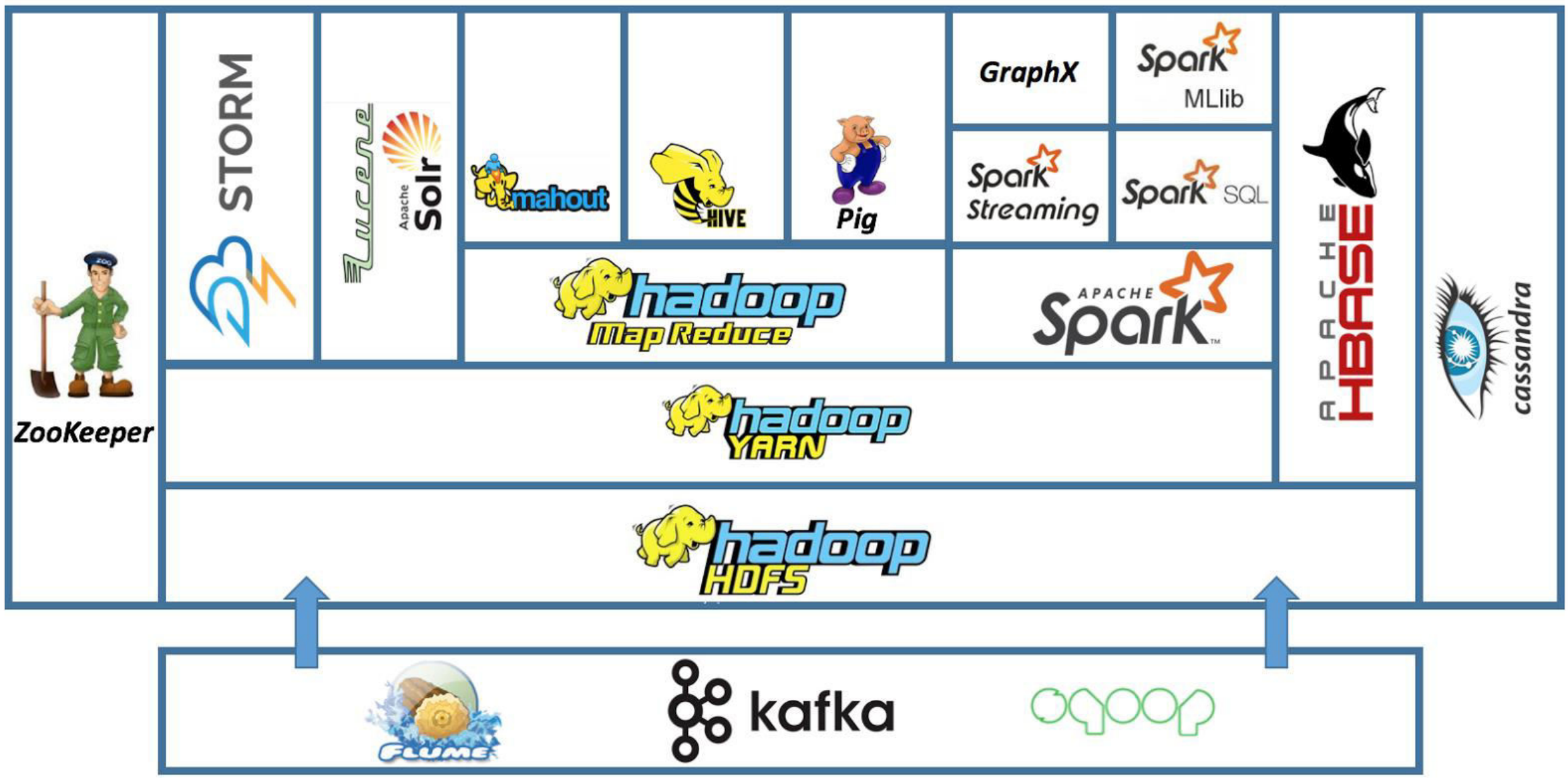
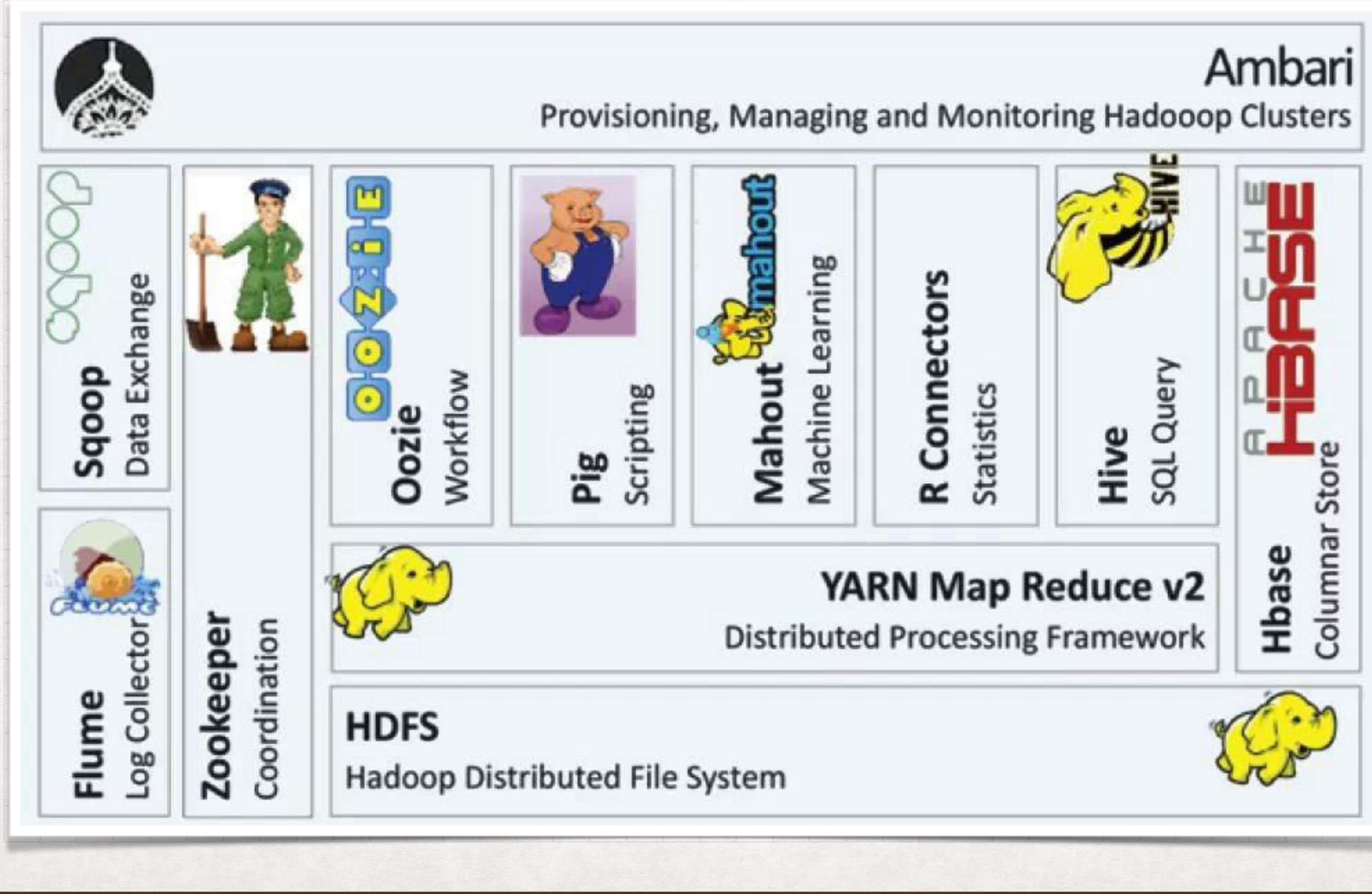
HADOOP ECOSYSTEM
BASICS



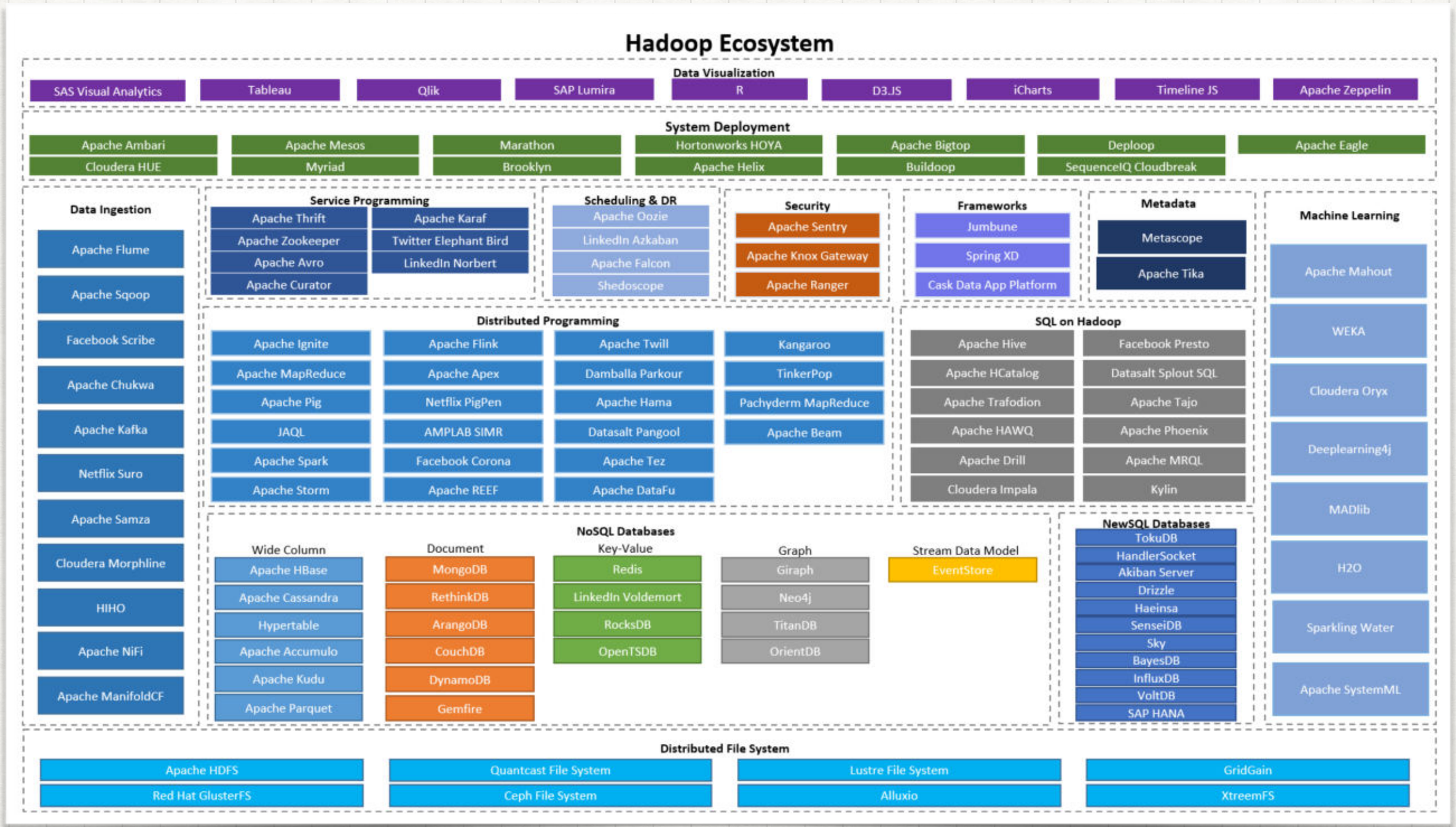
HADOOP HISTORY

- 2002 – запуск проекта Nutch
- 2003 – публикация с описанием GFS
- 2004 – создание NDFS (Nutch Distributed File System)
- 2004 – публикация Google и MapReduce
- 2005 – реализация MR в Nutch
- 2006 – выделение проекта Hadoop
- 2008 – выход Hadoop в лидеры ASF (Apache Software Foundation)
- 2008 - Apache Pig
- 2010 - выделение проекта HBase (ноябрь - Facebook messages)
- 2012 - Apache Impala
- 2012 - Apache Scoop
- 2012 - Apache Accumulo
- 2013 - Apache Kafka
- 2013 - Apache YARN
- 2013 - Apache Hive
- 2013 - Apache Spark (начало работы в 2009)

HADOOP WORLD 1.0



HADOOP WORLD 2017



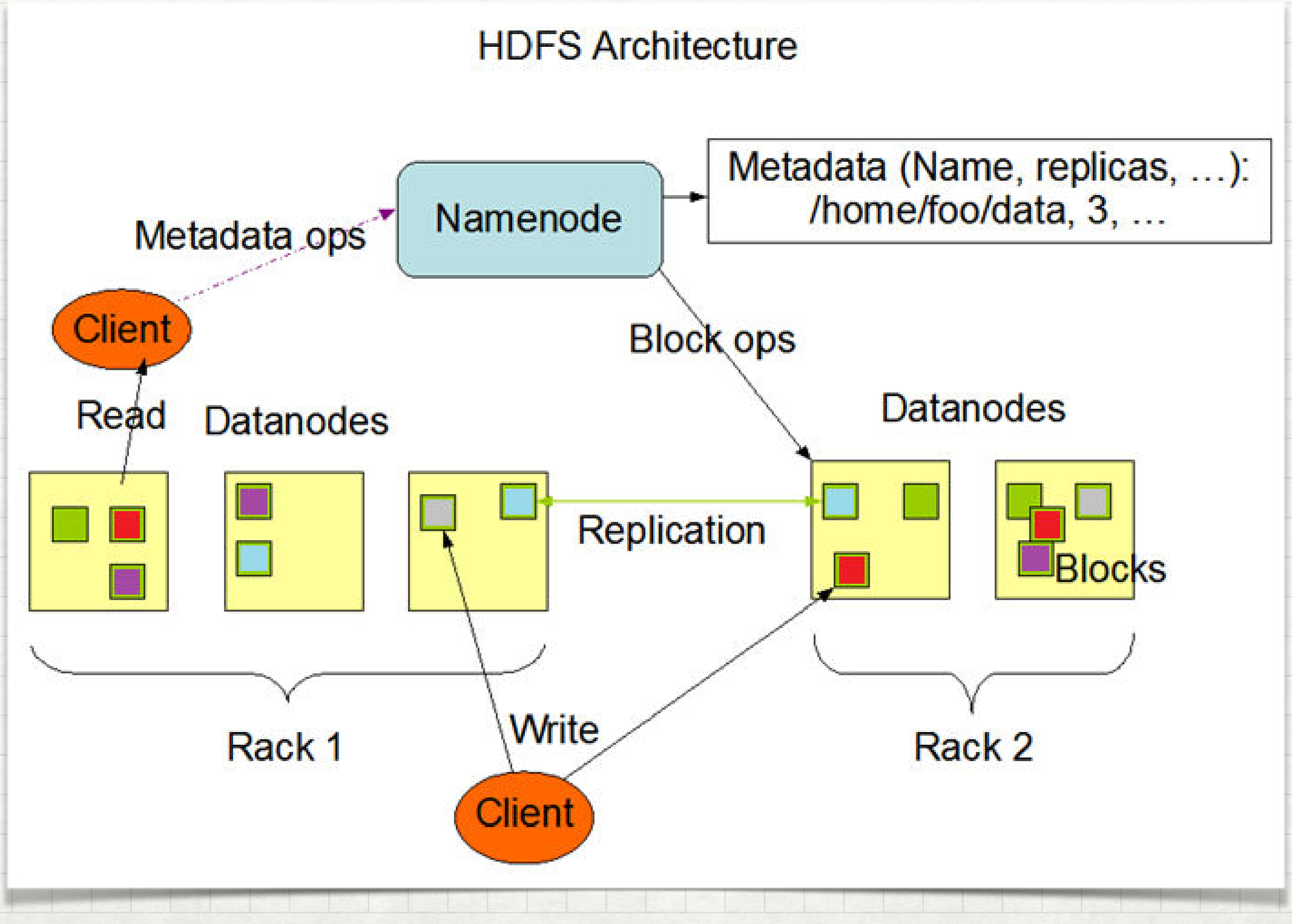
HADOOP BASICS

1.Hadoop Distributed File System (HDFS):

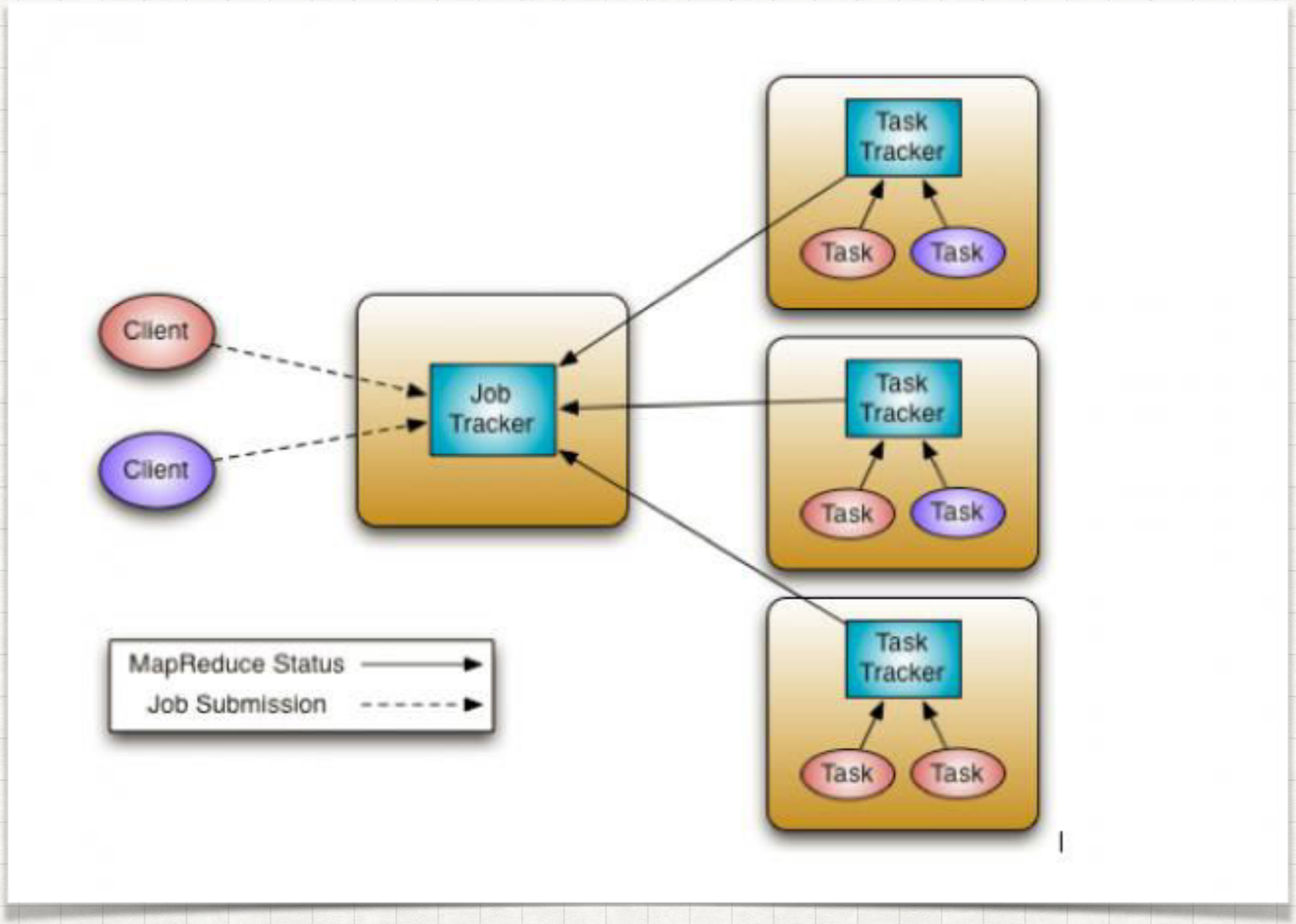
2.Hadoop YARN

3.Hadoop MapReduce

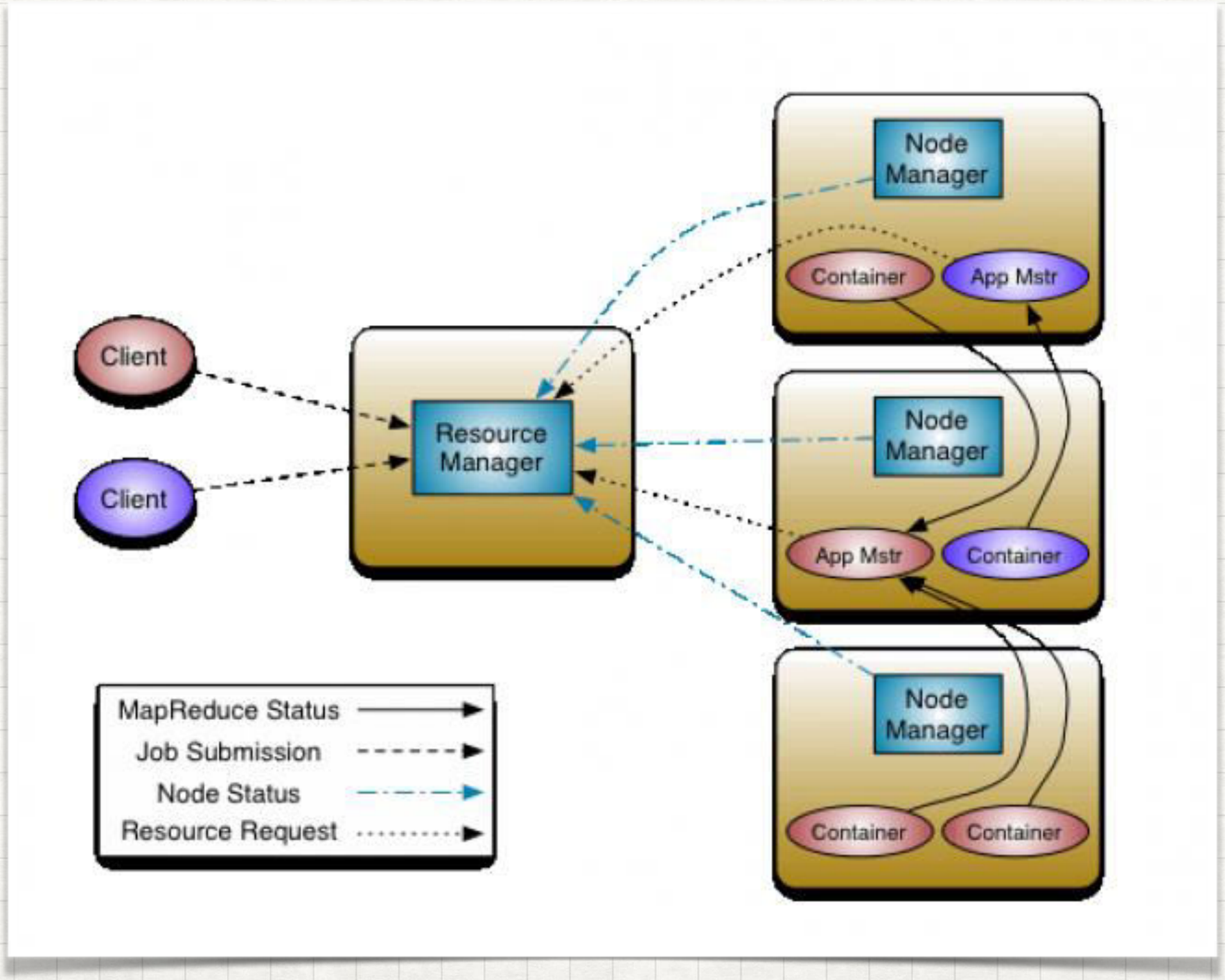
HDFS BASICS


HDFS Architecture
Namenode
Metadata (Name, replicas, …):
/home/foo/data, 3, …
Metadata ops
Client
Block ops
Read
Datanodes
Datanodes
Replication
Blocks
Rack 1
Write
Client
Rack 2

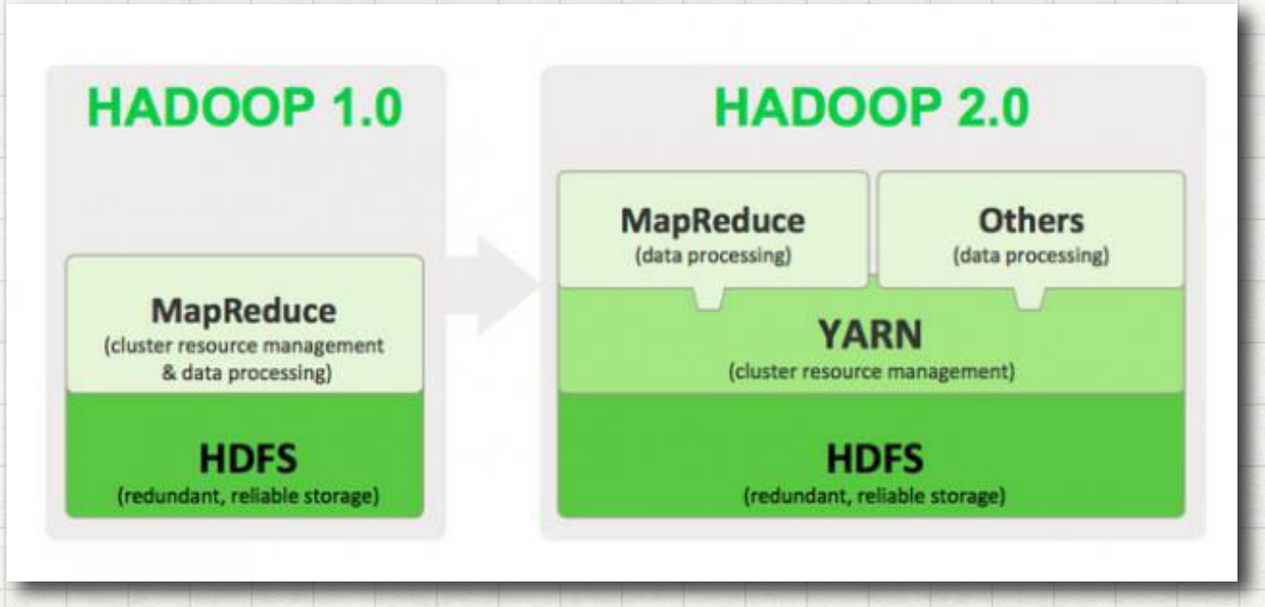
MR 1

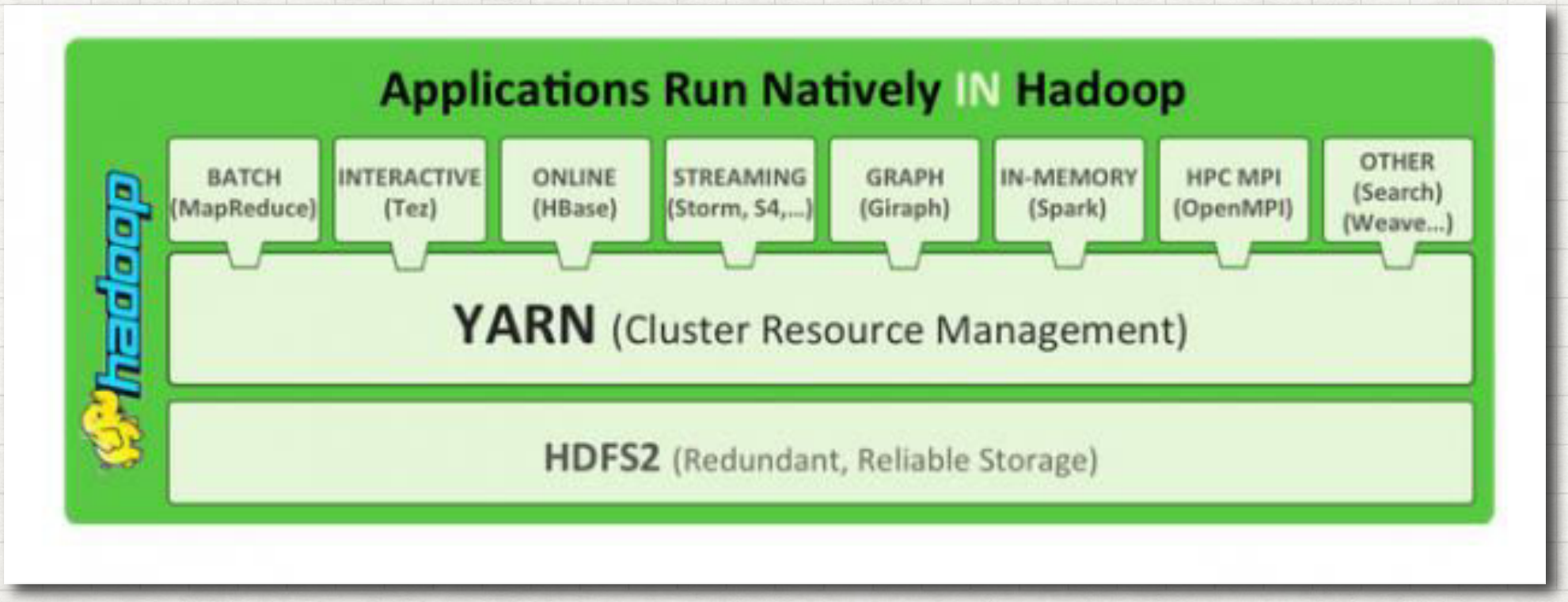

Client
Client
Job Tracker
Task Tracker
Task
Task
Task Tracker
Task
Task
Task Tracker
Task
Task
MapReduce Status
Job Submission

YARN (MR 2)

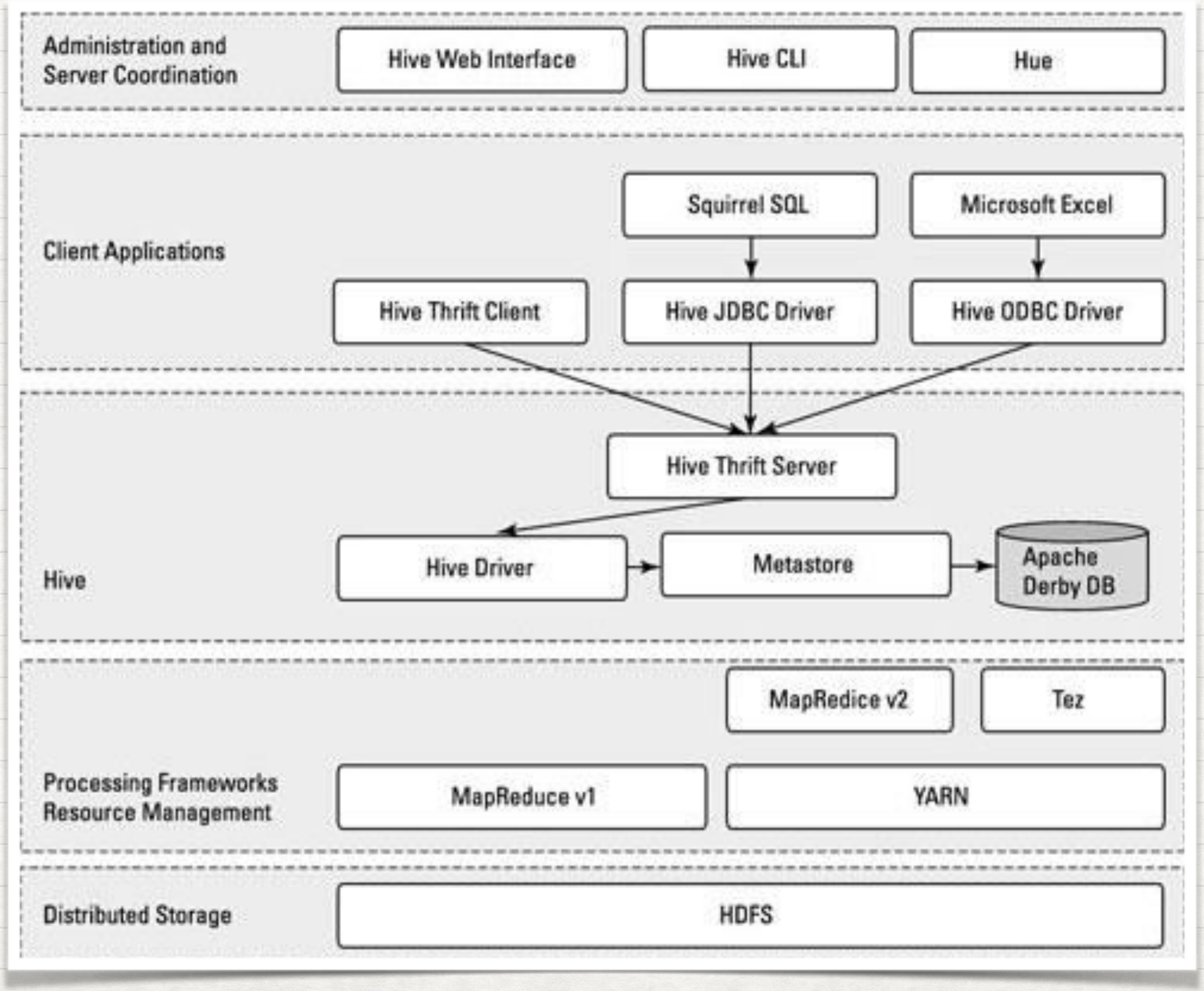

Client
Client
Resource Manager
Node Manager
Container
App Mstr
Node Manager
App Mstr
Container
Node Manager
Container
Container
MapReduce Status
Job Submission
Node Status
Resource Request

YARN

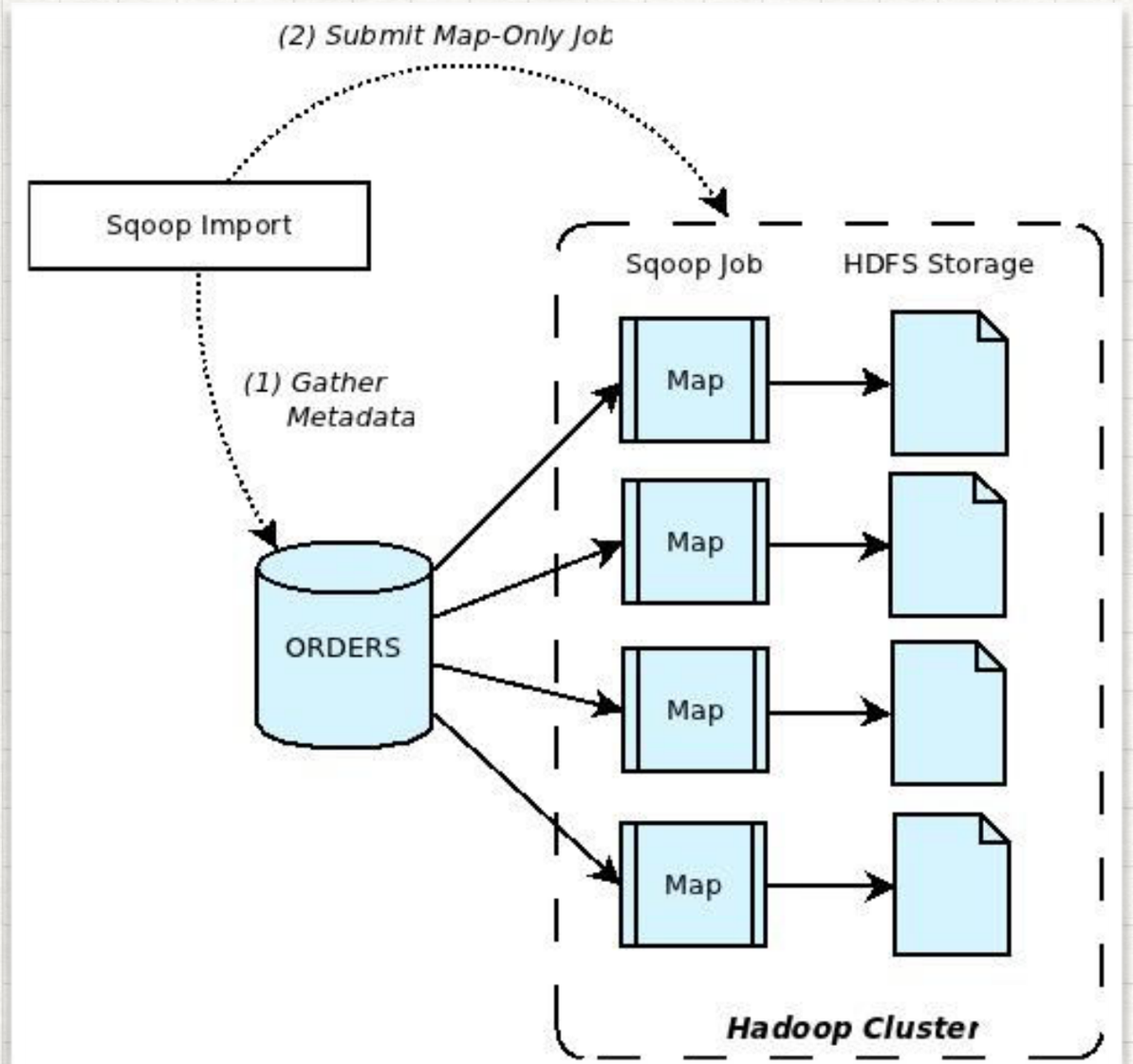

HADOOP 1.0
MapReduce
(cluster resource management & data processing)
HDFS
(redundant, reliable storage)
HADOOP 2.0
MapReduce
(data processing)
Others
(data processing)
YARN
(cluster resource management)
HDFS
(redundant, reliable storage)

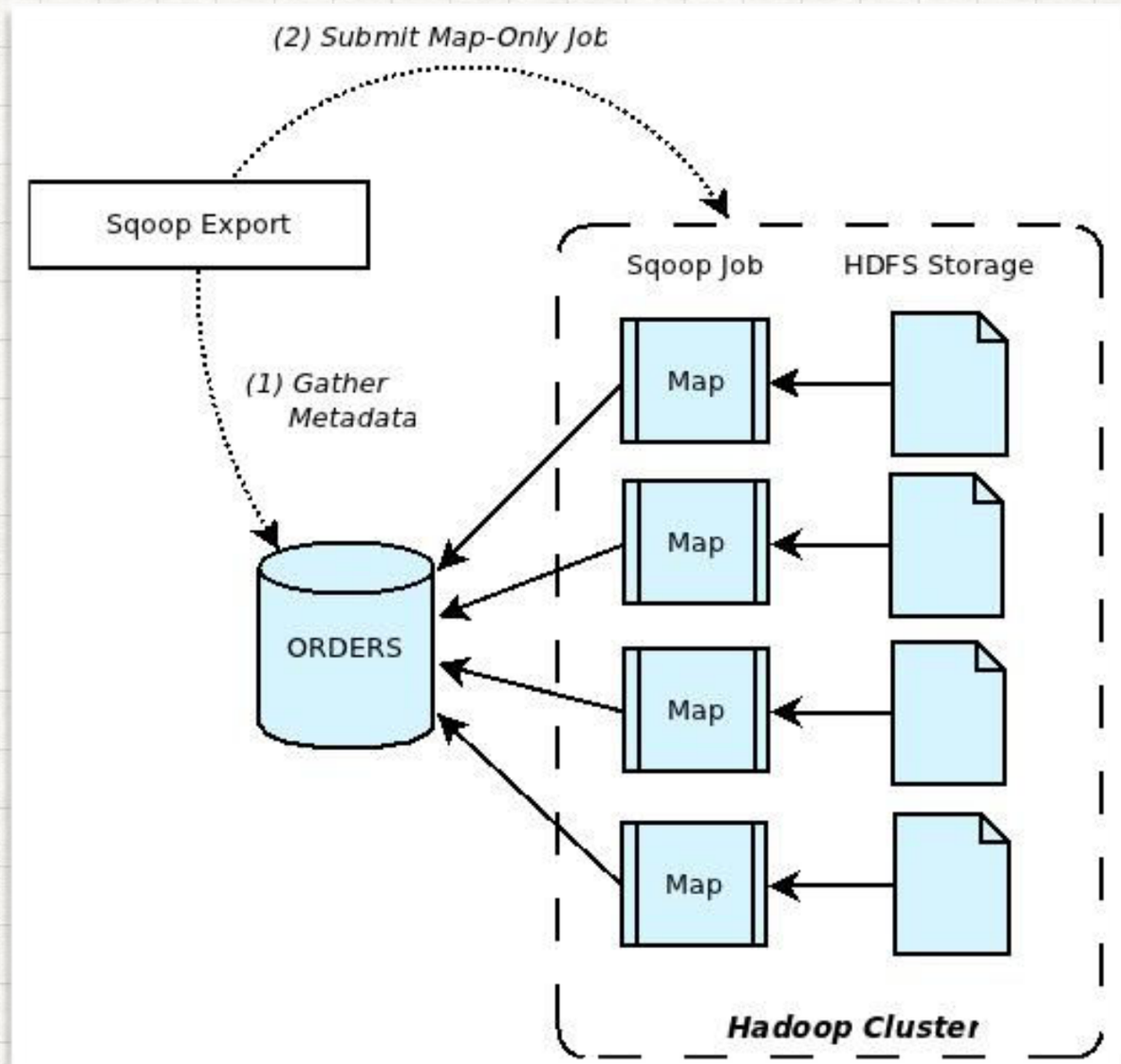

Applications Run Natively IN Hadoop
BATCH (MapReduce)
INTERACTIVE (Tez)
ONLINE (HBase)
STREAMING (Storm, S4,...)
GRAPH (Giraph)
IN-MEMORY (Spark)
HPC MPI (OpenMPI)
OTHER (Search) (Weave...)
YARN (Cluster Resource Management)
HDFS2 (Redundant, Reliable Storage)

HIVE BASICS

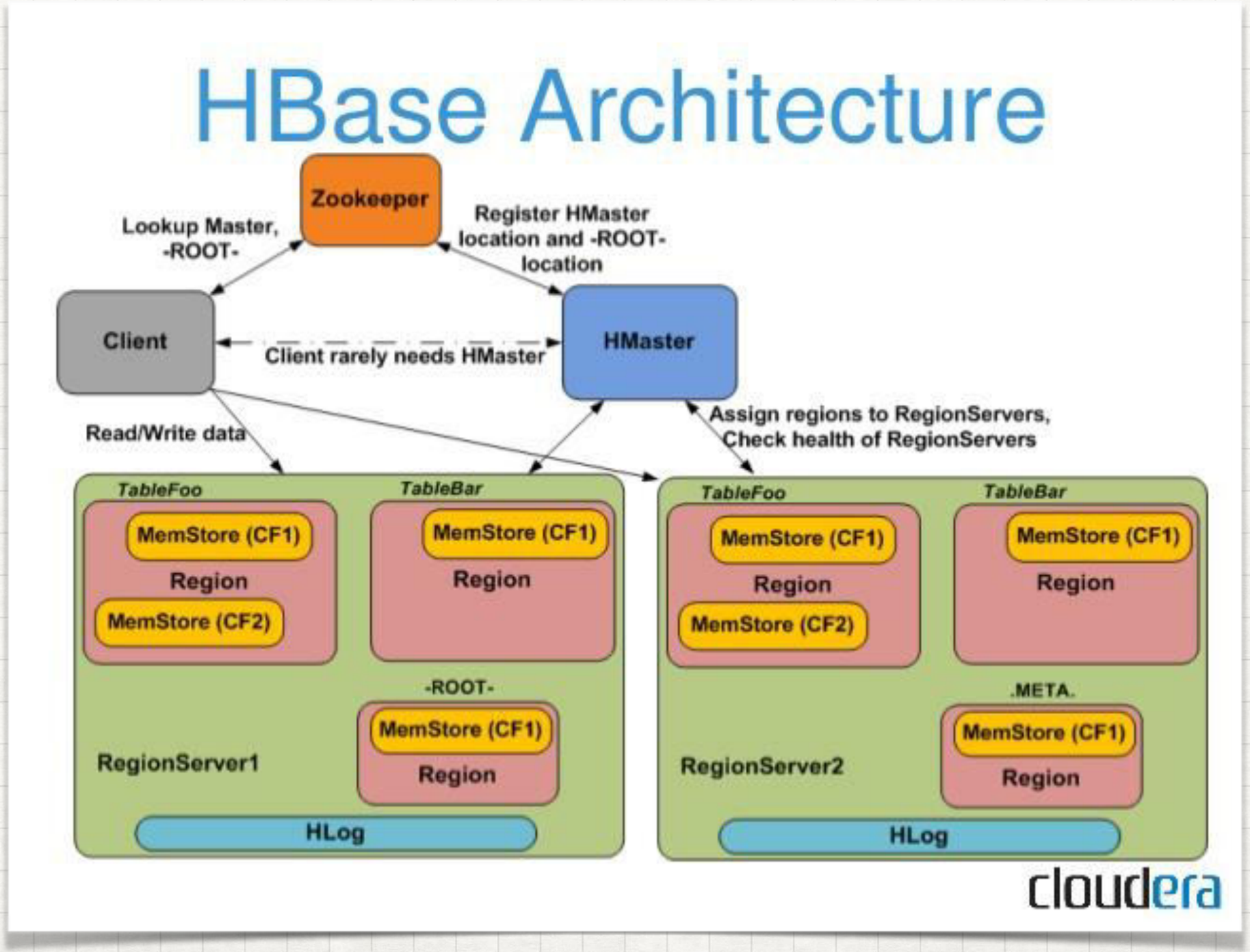
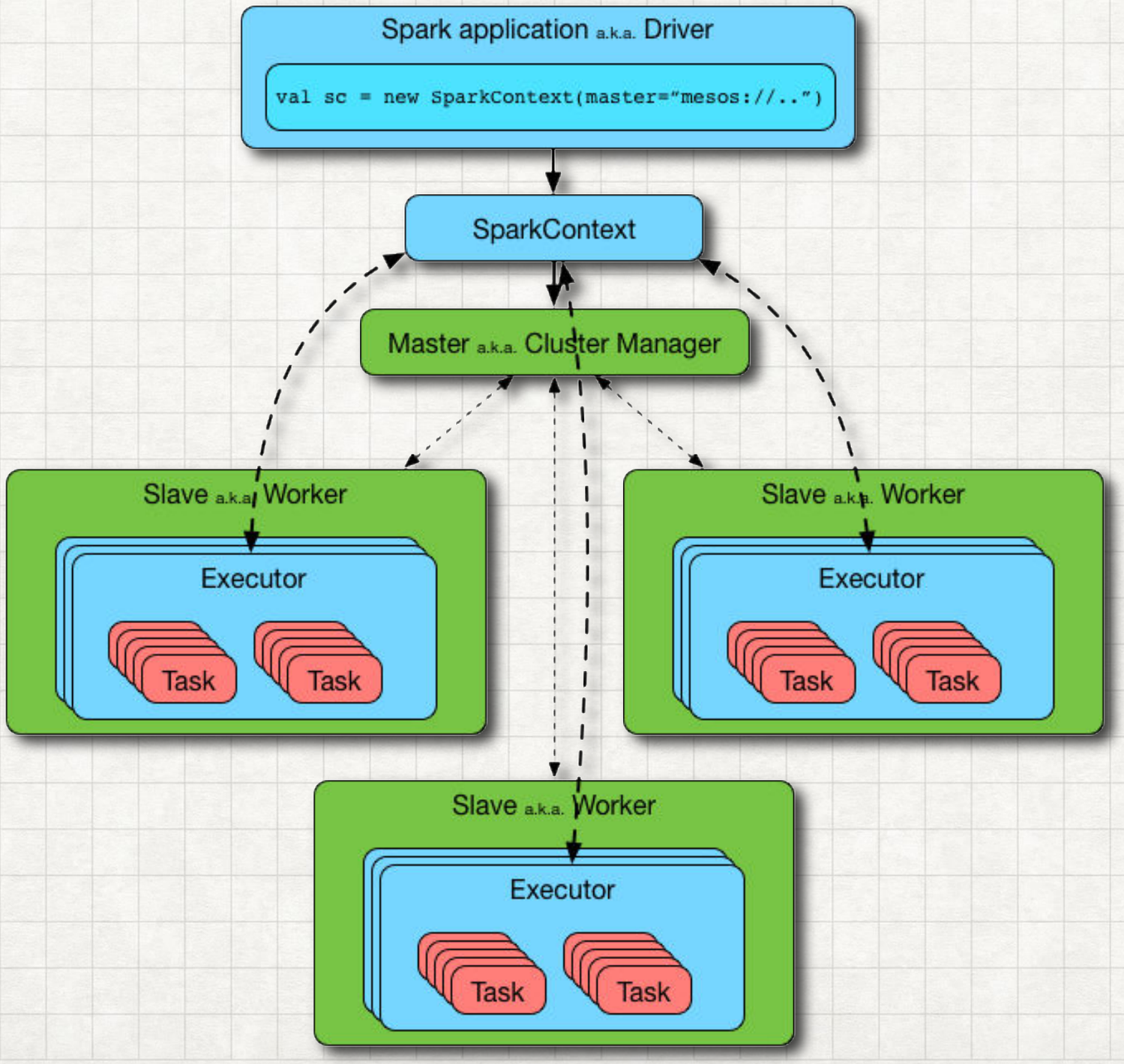

Administration and Server Coordination
Hive Web Interface
Hive CLI
Hue
Client Applications
Squirrel SQL
Microsoft Excel
Hive Thrift Client
Hive JDBC Driver
Hive ODBC Driver
Hive Thrift Server
Hive
Hive Driver
Metastore
Apache Derby DB
MapRedice v2
Tez
Processing Frameworks Resource Management
MapReduce v1
YARN
Distributed Storage
HDFS

SQOOP BASICS





HBASE BASICS



SPARK BASICS