The Hadoop Ecosystem:

So much free stuff!

After this video you will be able to..

- Differentiate the major layers in the Hadoop ecosystem

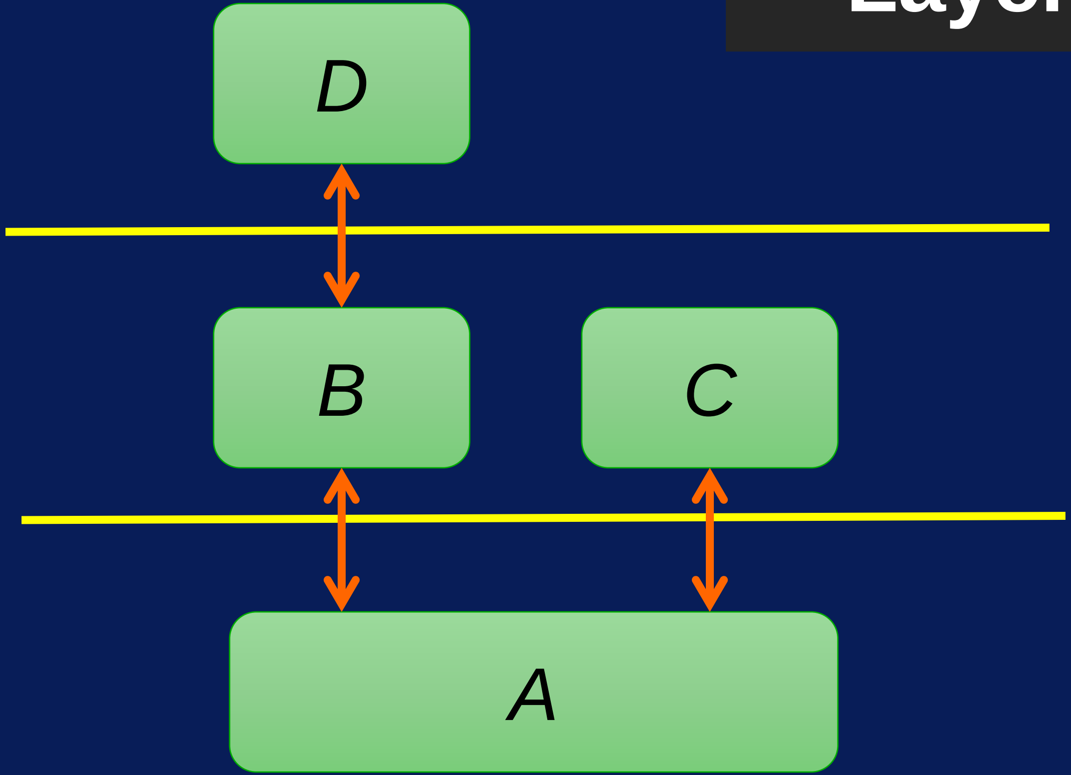
- Recognize key tools of the Hadoop ecosystem including HDFS, YARN, and MapReduce

Yahoo created Hadoop in 2005

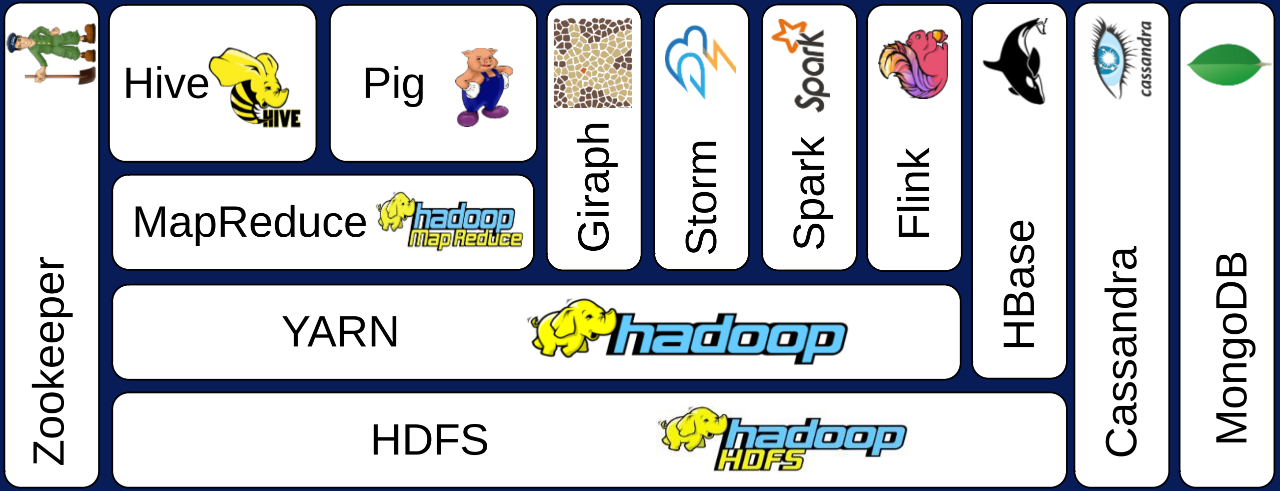
More Big Data frameworks released

Now there’s over a 100!

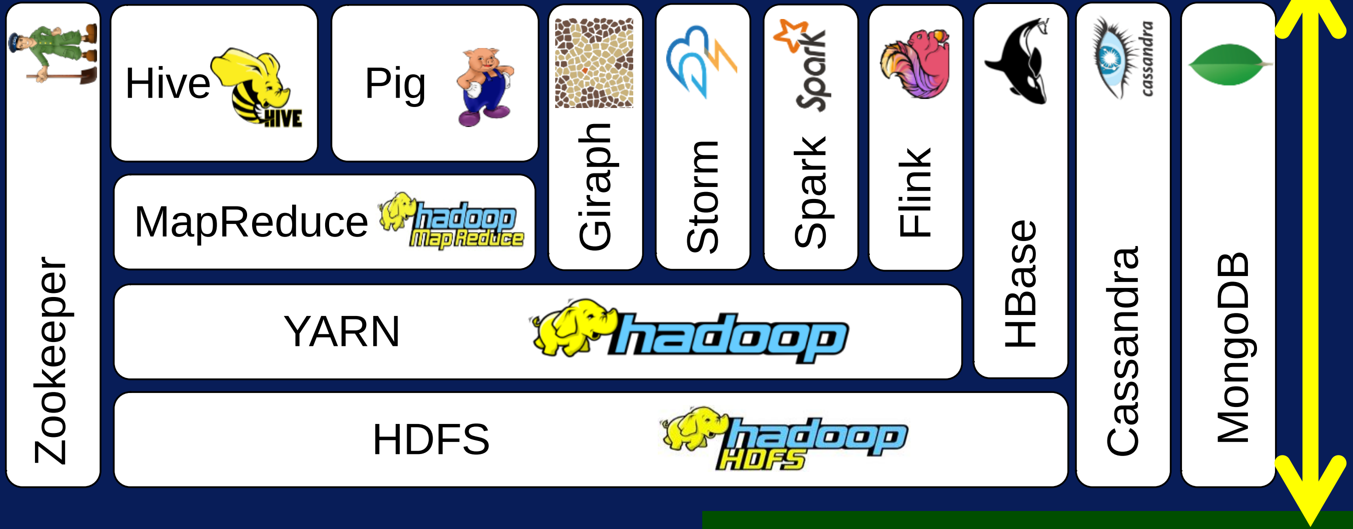
Layer Diagram



One possible layer diagram for Hadoop



One possible layer diagram for Hadoop



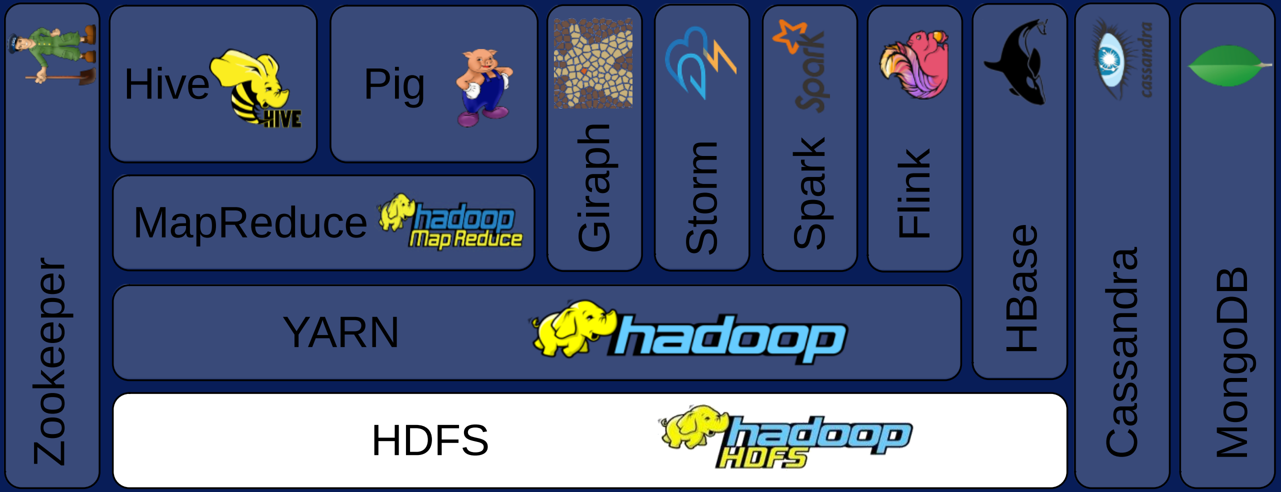
Higher levels: Interactivity

Lower levels: Storage and scheduling

Distributed file system as foundation

Scalable storage

Fault tolerance



Zookeeper

Hive

Pig

MapReduce

Giraph

Storm

Spark

Flink

HBase

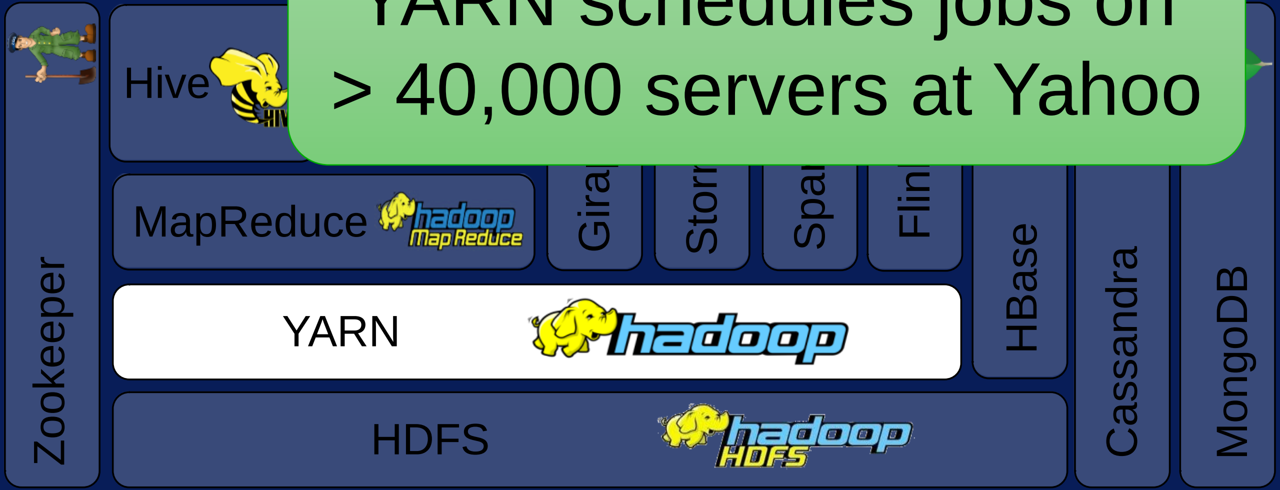
Cassandra

MongoDB

YARN

HDFS

Flexible scheduling and resource management

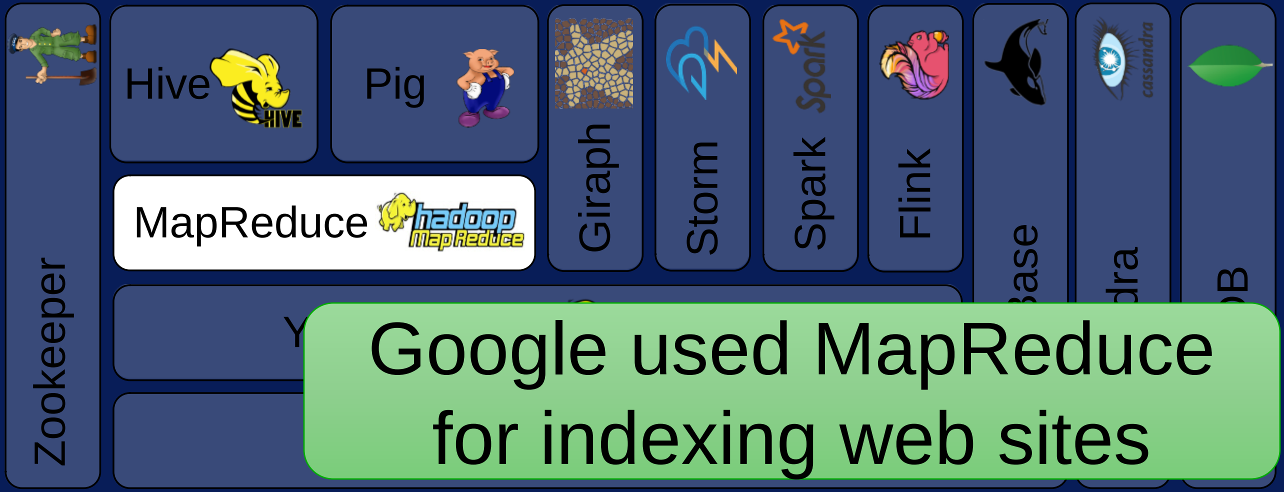


YARN schedules jobs on > 40,000 servers at Yahoo

Simplified programming model

Map → apply()

Reduce → summarize()

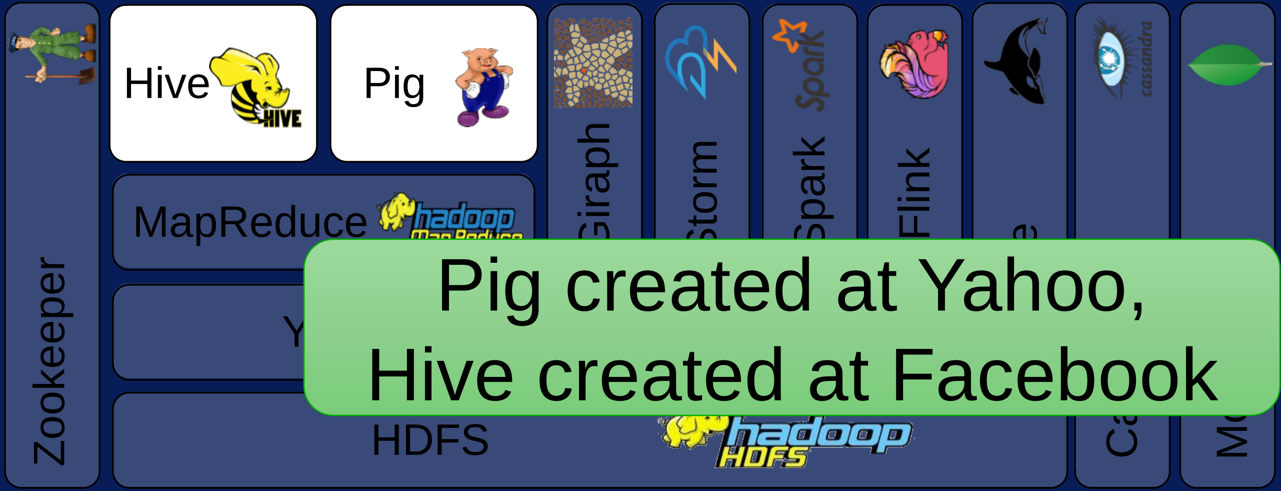


Google used MapReduce for indexing web sites

Higher-level programming models

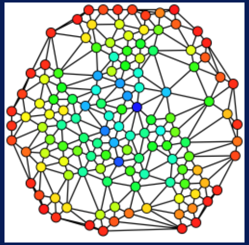
Pig = dataflow scripting

Hive = SQL-like queries

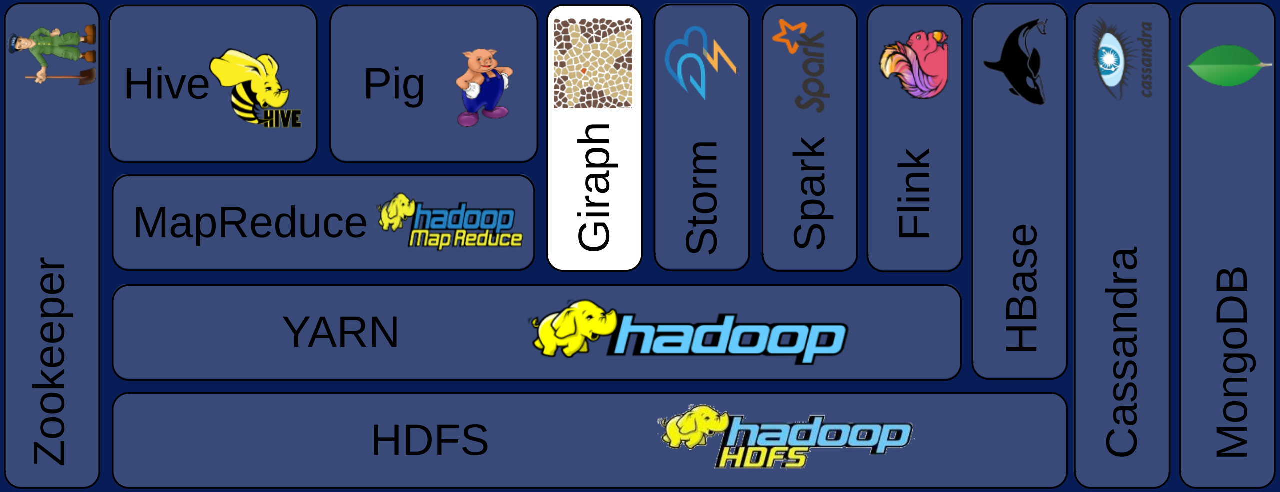


Pig created at Yahoo,
Hive created at Facebook

Specialized models for graph processing



Giraph used by Facebook to analyze social graphs



Hive

Pig

MapReduce

Giraph

Storm

Spark

Flink

HBase

Cassandra

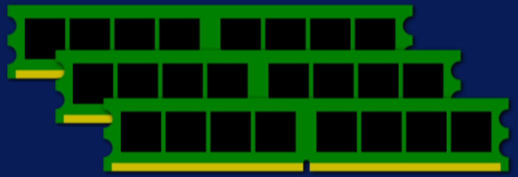
MongoDB

YARN

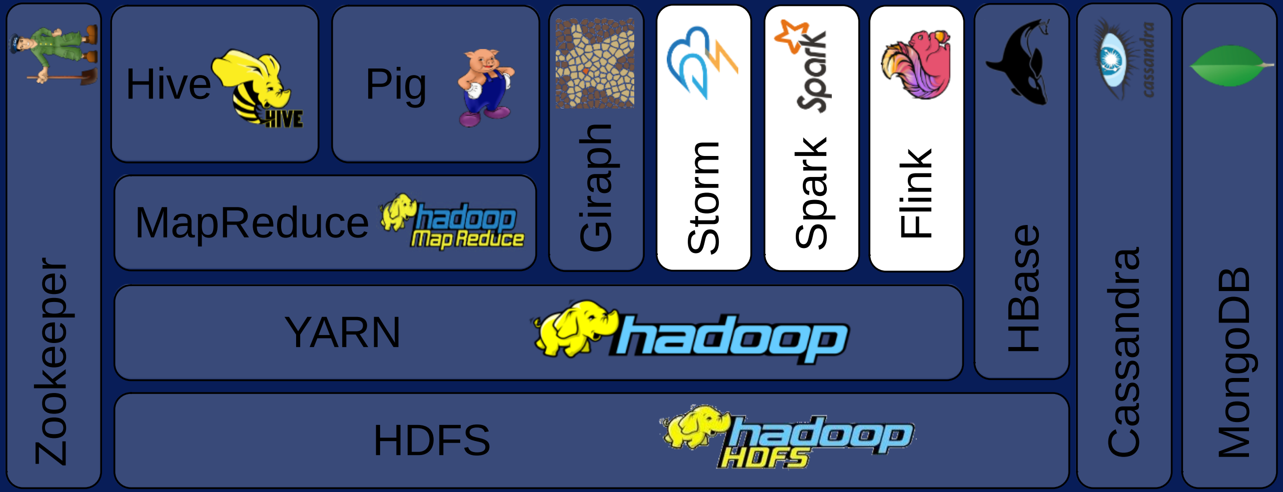
HDFS

Zookeeper

Real-time and in-memory processing



In-memory → 100x faster for some tasks



Zookeeper

Hive

Pig

MapReduce

Giraph

Storm

Spark

Flink

YARN

HDFS

HBase

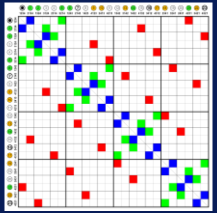
Cassandra

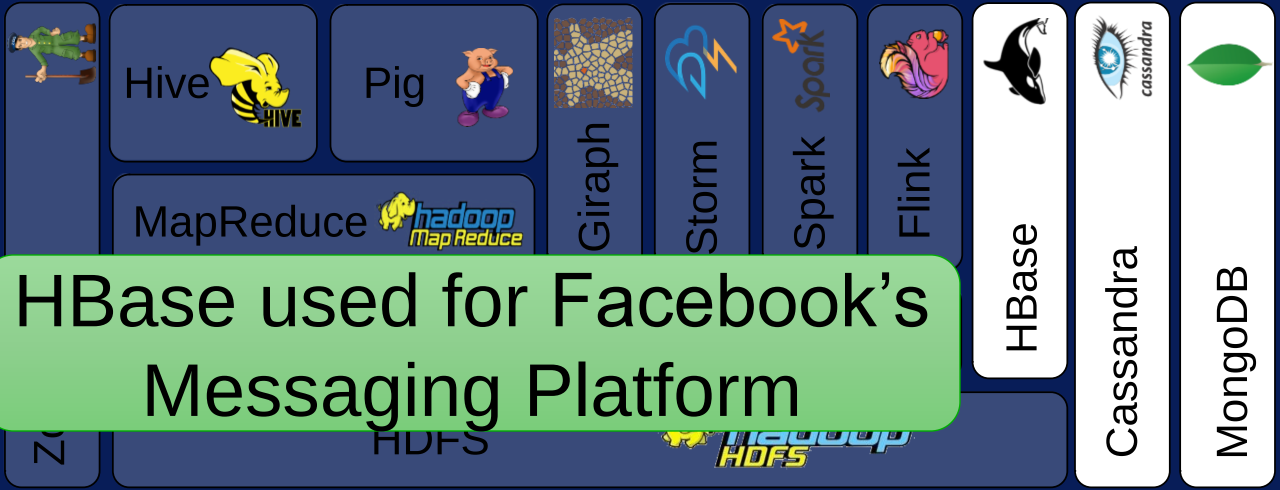
MongoDB

NoSQL for non-files

Key-values

Sparse tables





Hive

Pig

MapReduce

Giraph

Storm

Spark

Flink

HBase

Cassandra

MongoDB

HDFS

Z

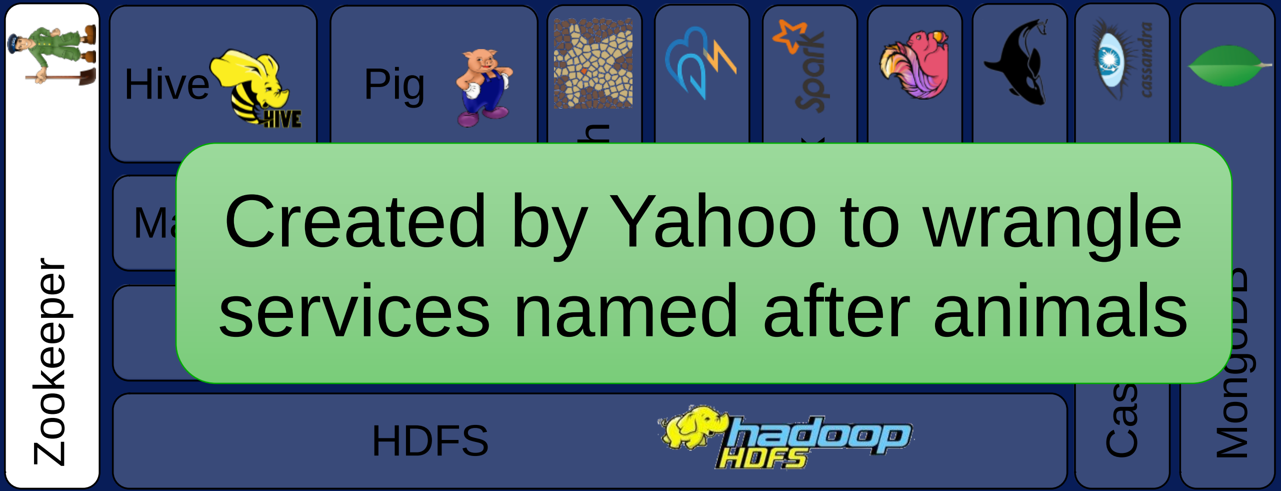
HBase used for Facebook’s Messaging Platform

Zookeeper for management

Synchronization

Configuration

High-availability



Created by Yahoo to wrangle services named after animals

All these tools are open-source

All these tools are open-source

Large community
for support

All these tools are open-source

Large community
for support

Download separately
or part of pre-built image

All these tools are open-source

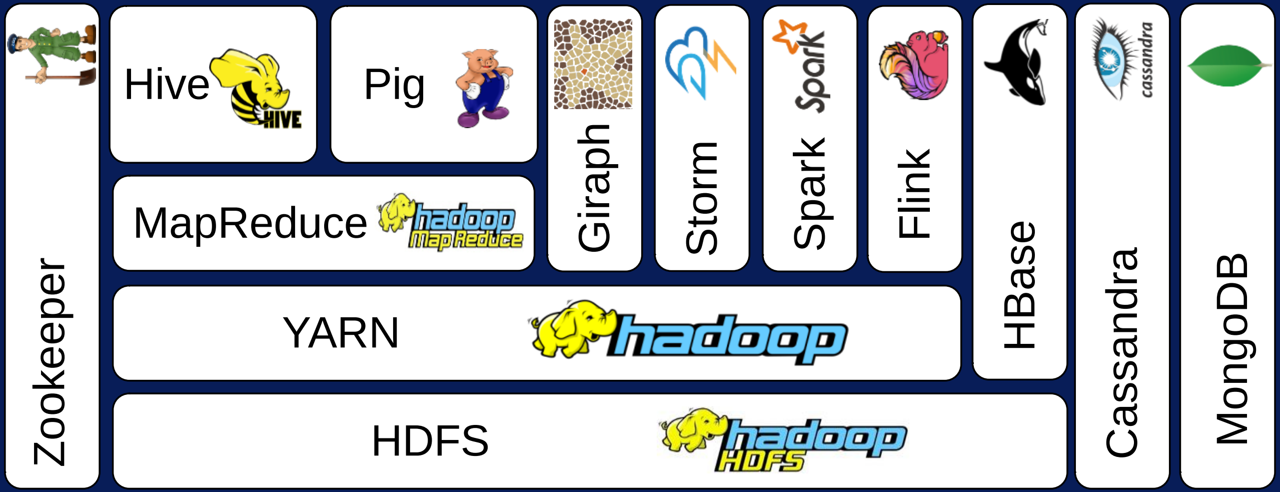
Large community
for support

Download separately
or part of pre-built image

cloudera®

MAPR®

Hortonworks



Zookeeper

Hive

Pig

MapReduce

Giraph

Storm

Spark

Flink

HBase

Cassandra

MongoDB

YARN

HDFS

Growing number of open-source tools