

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

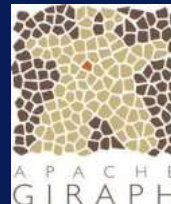


More Big Data frameworks released

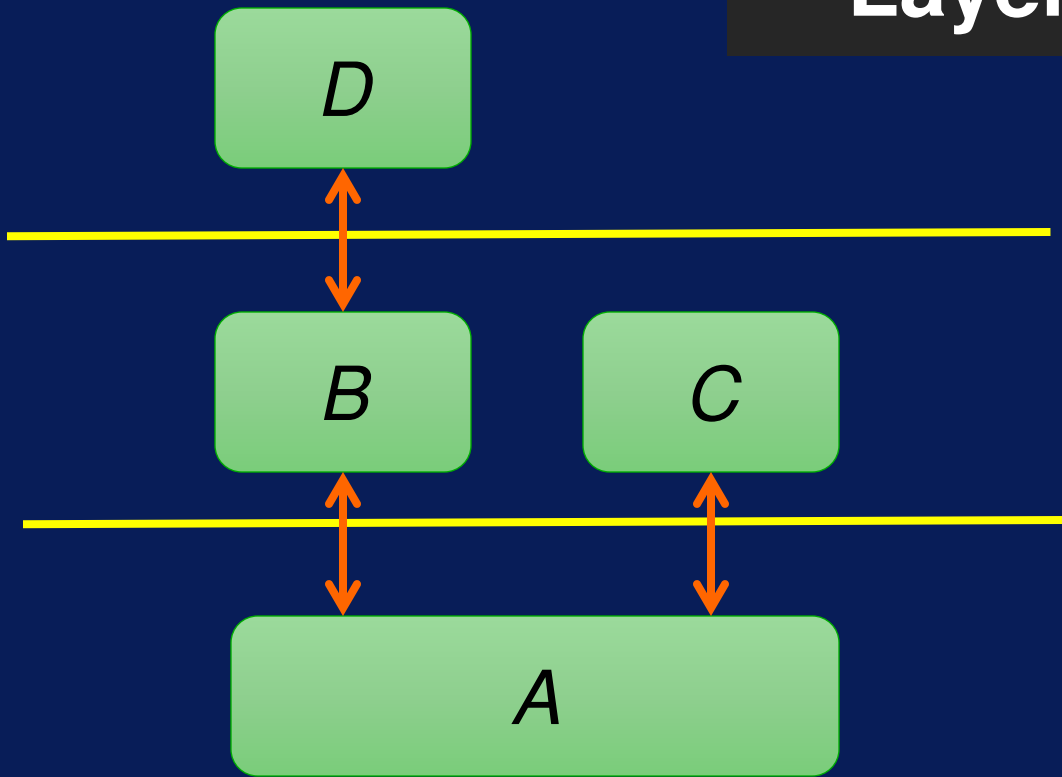




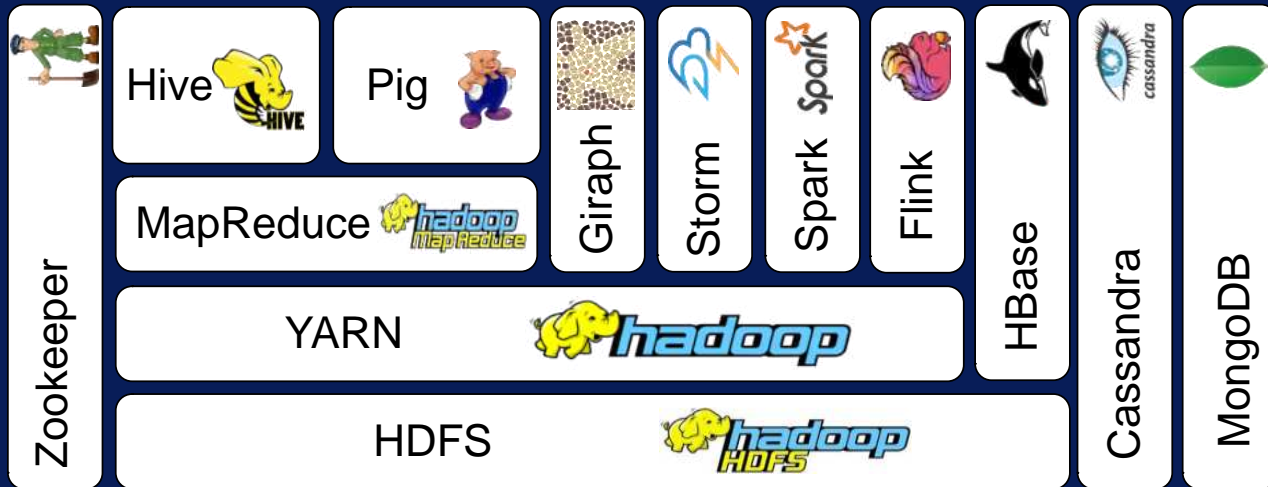
Now there's over a 100!



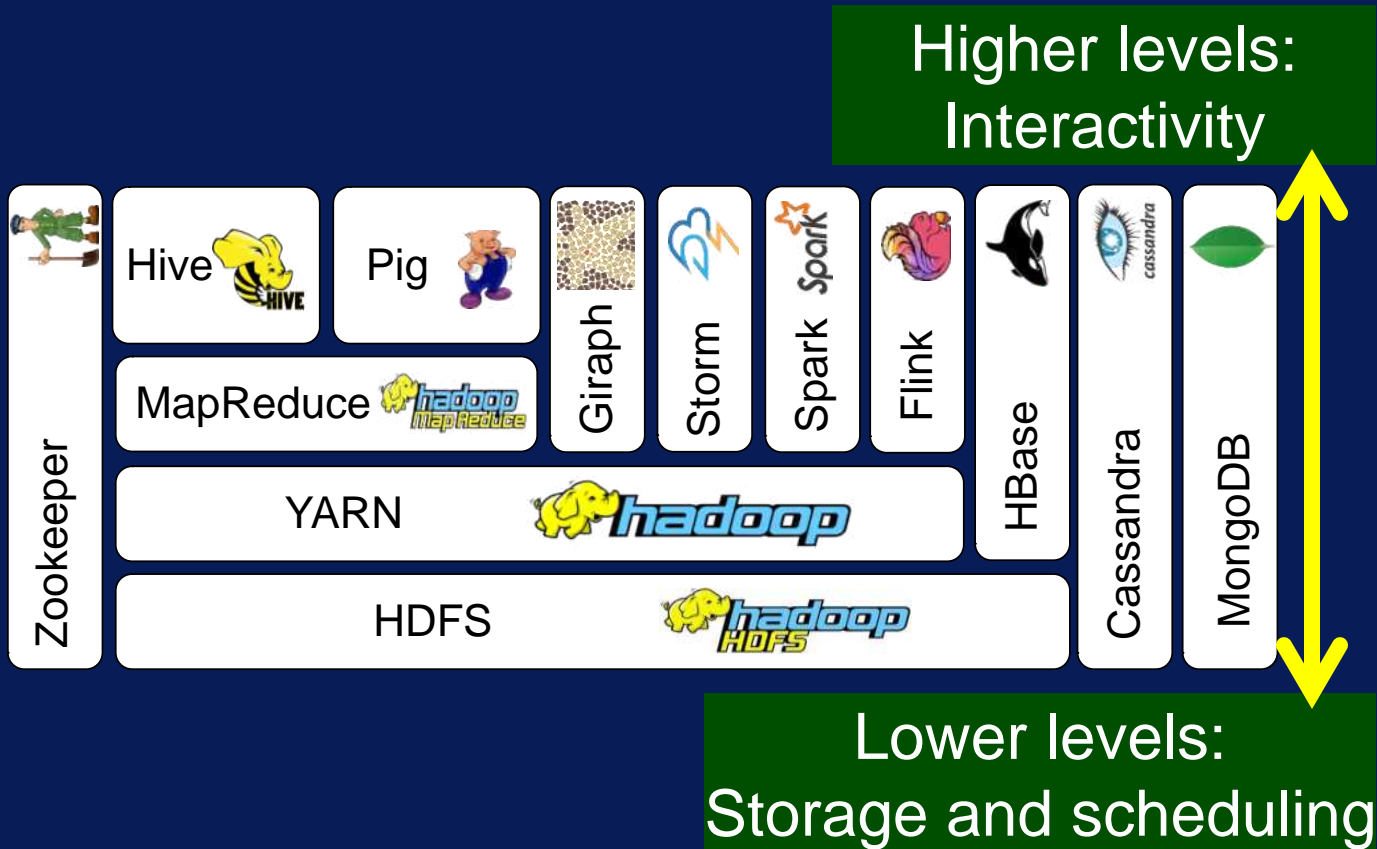
Layer Diagram



One possible layer diagram for Hadoop



One possible layer diagram for Hadoop



Distributed file system as foundation

Scalable storage

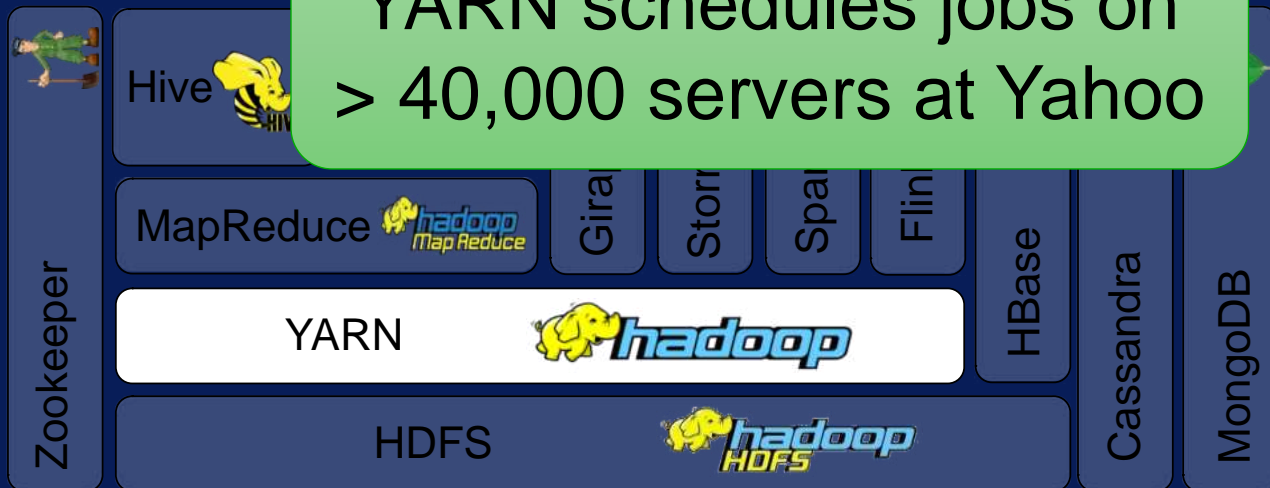
Fault tolerance



Flexible scheduling and resource management



YARN schedules jobs on
> 40,000 servers at Yahoo



Simplified programming model

Map → apply()

Reduce → summarize()



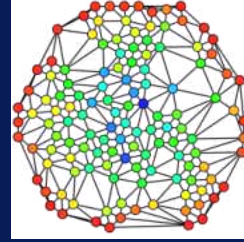
Higher-level programming models

Pig = dataflow scripting

Hive = SQL-like queries



Specialized models for graph processing



Giraph used by Facebook
to analyze social graphs



Real-time and in-memory processing



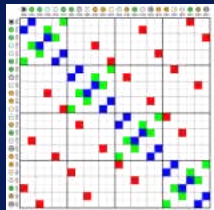
In-memory → 100x faster
for some tasks



NoSQL for non-files

Key-values

Sparse tables



HBase used for Facebook's
Messaging Platform

Zookeeper for management

Synchronization

Configuration

High-availability



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



Growing number of open-source tools