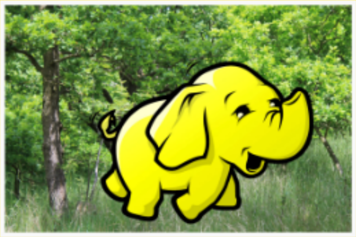
The Hadoop Ecosystem



Hadoop Overview

What is Hadoop?

"An open source **software platform** for **distributed storage** and **distributed processing** of **very large data sets** on **computer clusters** built from commodity hardware" - *Hortonworks*

Why Hadoop?

- Data's too darn big - terabytes per day
- Vertical scaling doesn't cut it
  - Disk seek times
  - Hardware failures
  - Processing times
- Horizontal scaling is linear
- Hadoop: It's not just for batch processing any more

Hadoop History

- Google published GFS and MapReduce papers in 2003-2004
- Yahoo! was building "Nutch," an open source web search engine at the same time
- Hadoop was primarily driven by Doug Cutting and Tom White in 2006
- It's been evolving ever since...

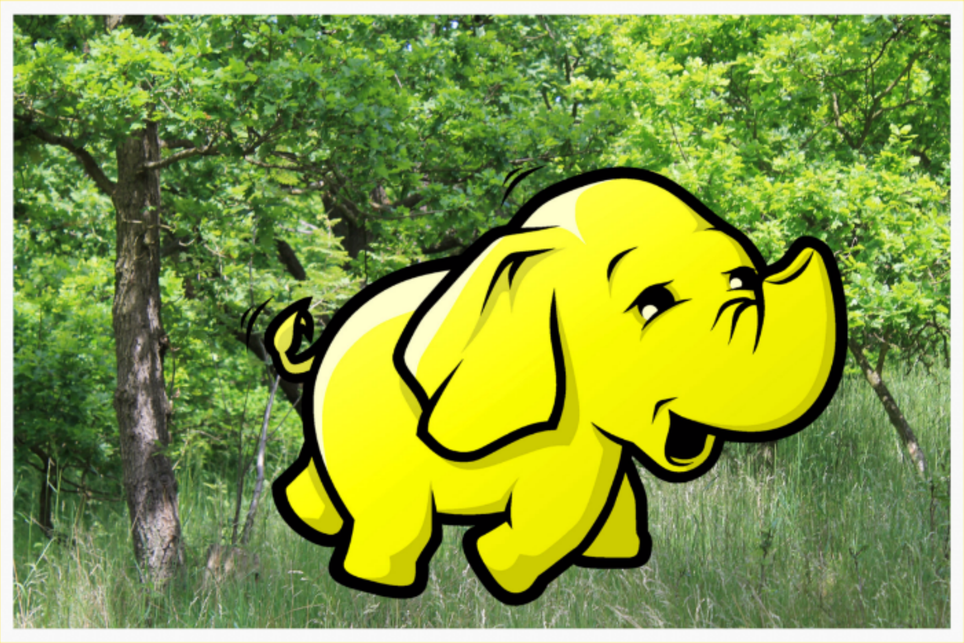
World of Hadoop

Query Engines

Core Hadoop Ecosystem

External Data Storage

The Hadoop Ecosystem



Hadoop Overview

What is Hadoop?

"an open source **software platform** for **distributed storage** and **distributed processing** of **very large data sets** on **computer clusters** built from commodity hardware" - *Hortonworks*

Why Hadoop?

- Data's too darn big - terabytes per day
- Vertical scaling doesn't cut it
  - Disk seek times
  - Hardware failures
  - Processing times
- Horizontal scaling is linear
- Hadoop: It's not just for batch processing anymore

Hadoop History

- Google published GFS and MapReduce papers in 2003-2004
- Yahoo! was building "Nutch," an open source web search engine at the same time
- Hadoop was primarily driven by Doug Cutting and Tom White in 2006
- It's been evolving ever since...

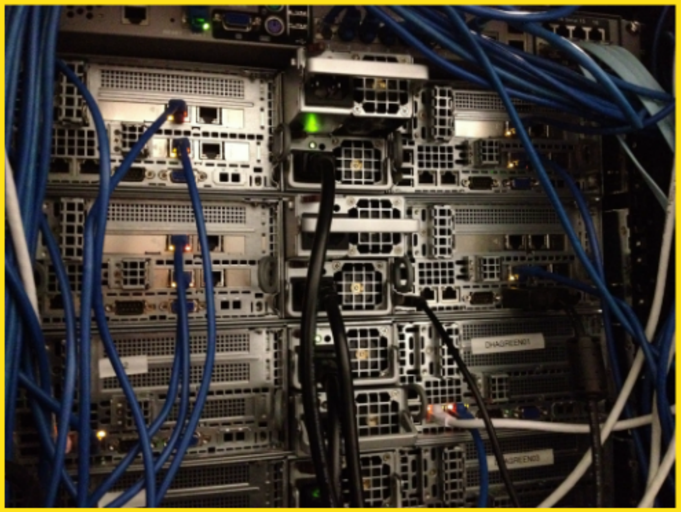
What is Hadoop?

"an open source **software platform** for **distributed storage** and **distributed processing** of **very large data sets** on **computer clusters** built from commodity hardware" - *Hortonworks*

Hadoop History

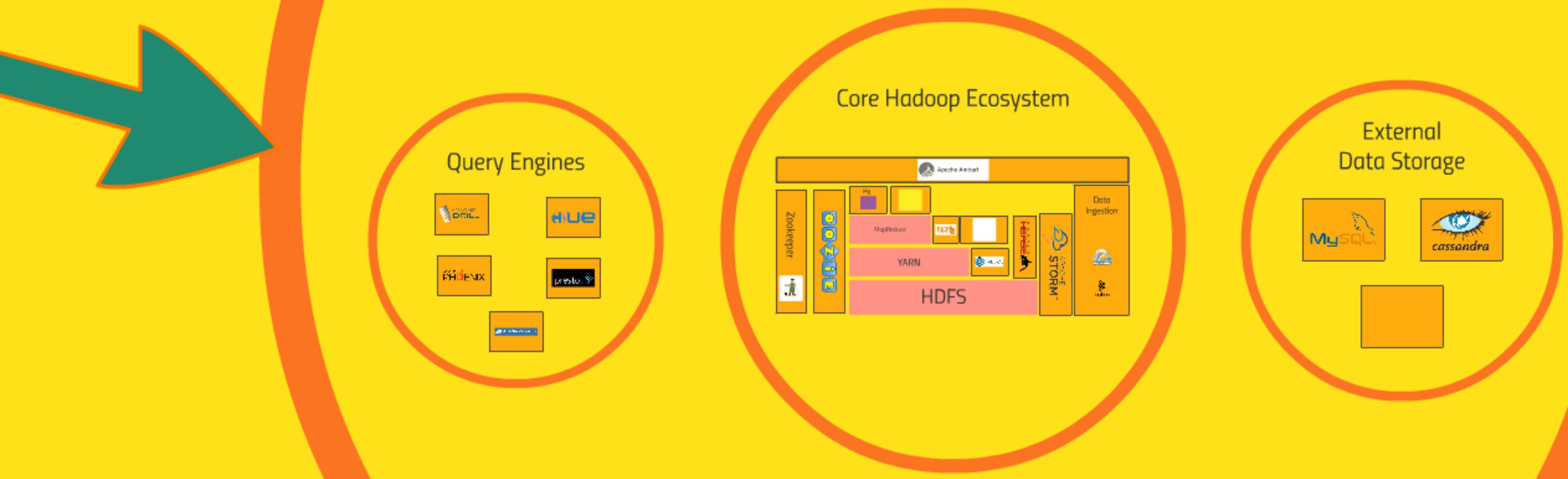
- Google published GFS and MapReduce papers in 2003-2004
- Yahoo! was building "Nutch," an open source web search engine at the same time
- Hadoop was primarily driven by Doug Cutting and Tom White in 2006
- It's been evolving ever since...

Why Hadoop?

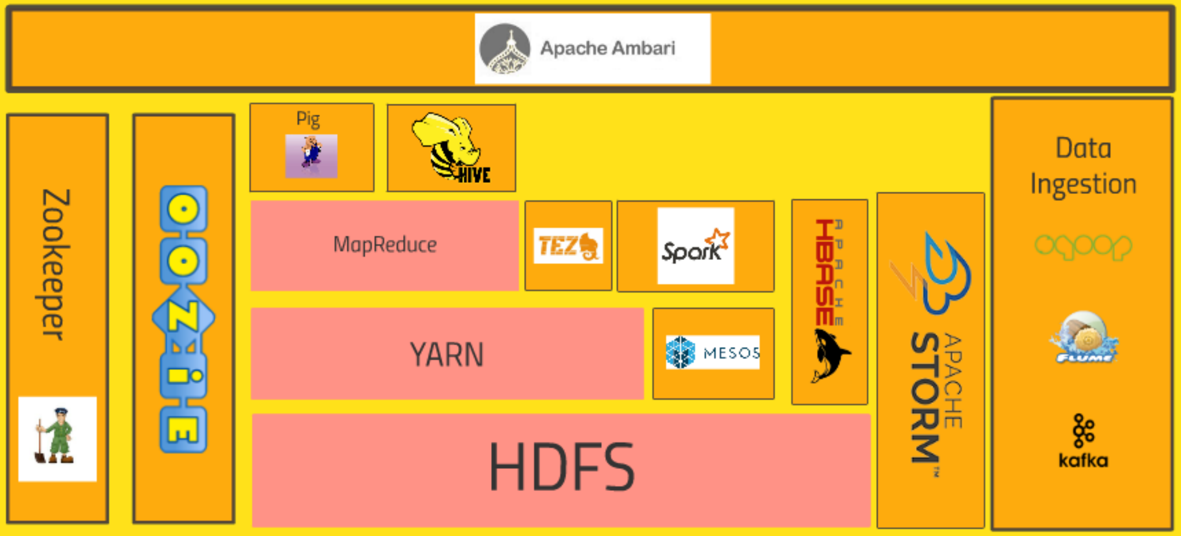


- Data's too darn big - terabytes per day
- Vertical scaling doesn't cut it
  - Disk seek times
  - Hardware failures
  - Processing times
- Horizontal scaling is linear
- Hadoop: It's not just for batch processing anymore

World of Hadoop



Core Hadoop Ecosystem



Zookeeper

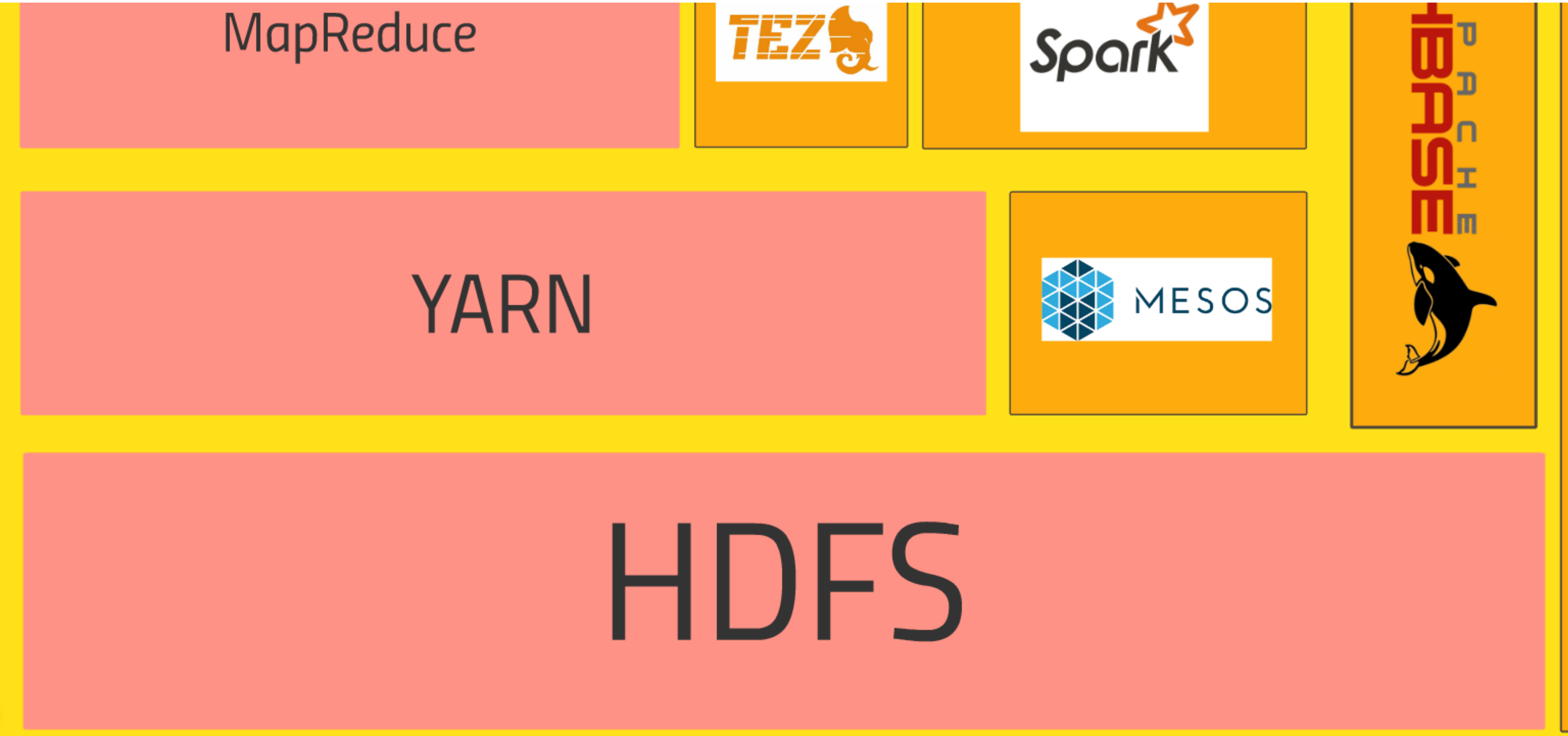
Pig

MapReduce

YARN

HDFS

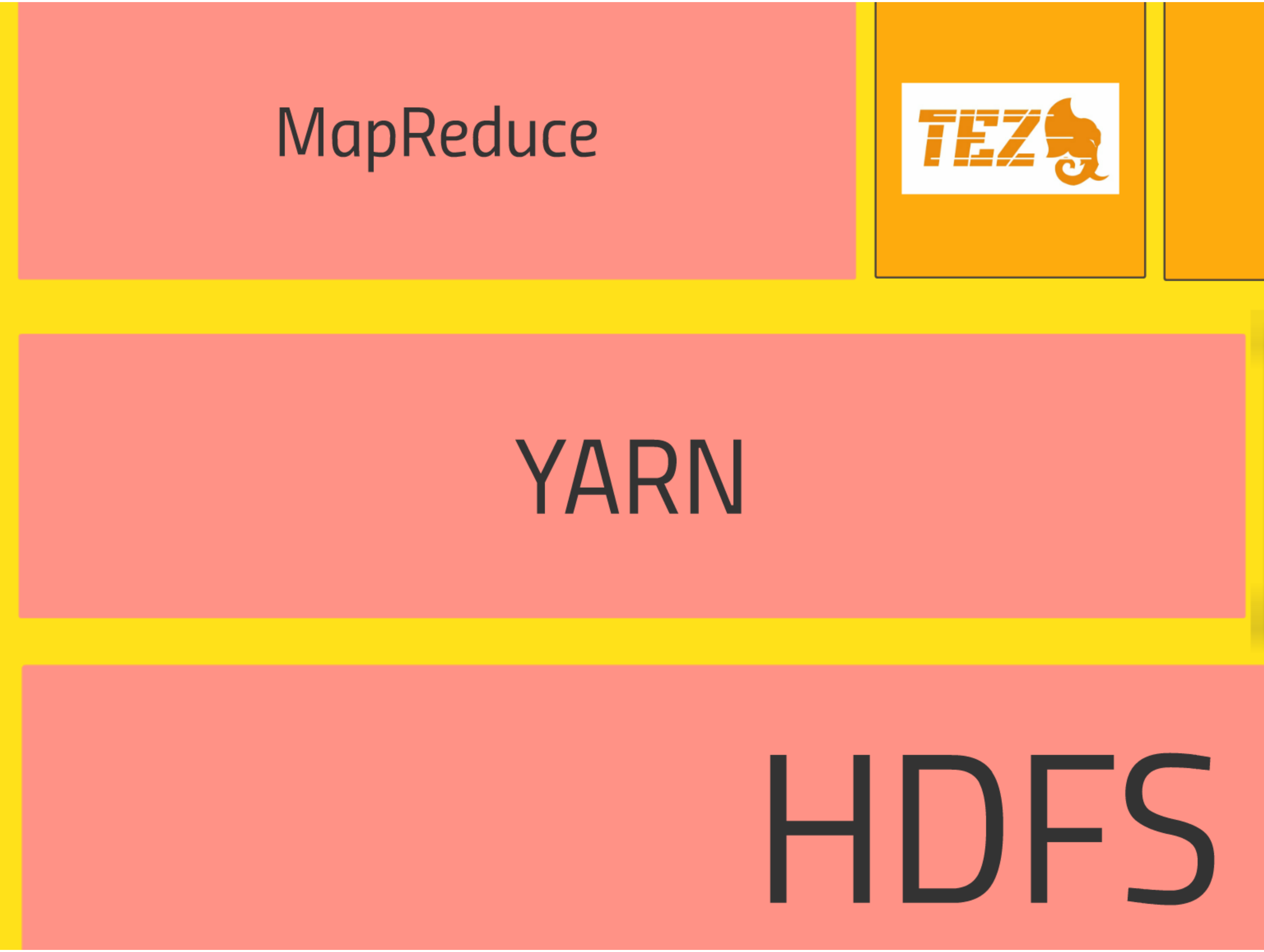
Data Ingestion



MapReduce

YARN

HDFS

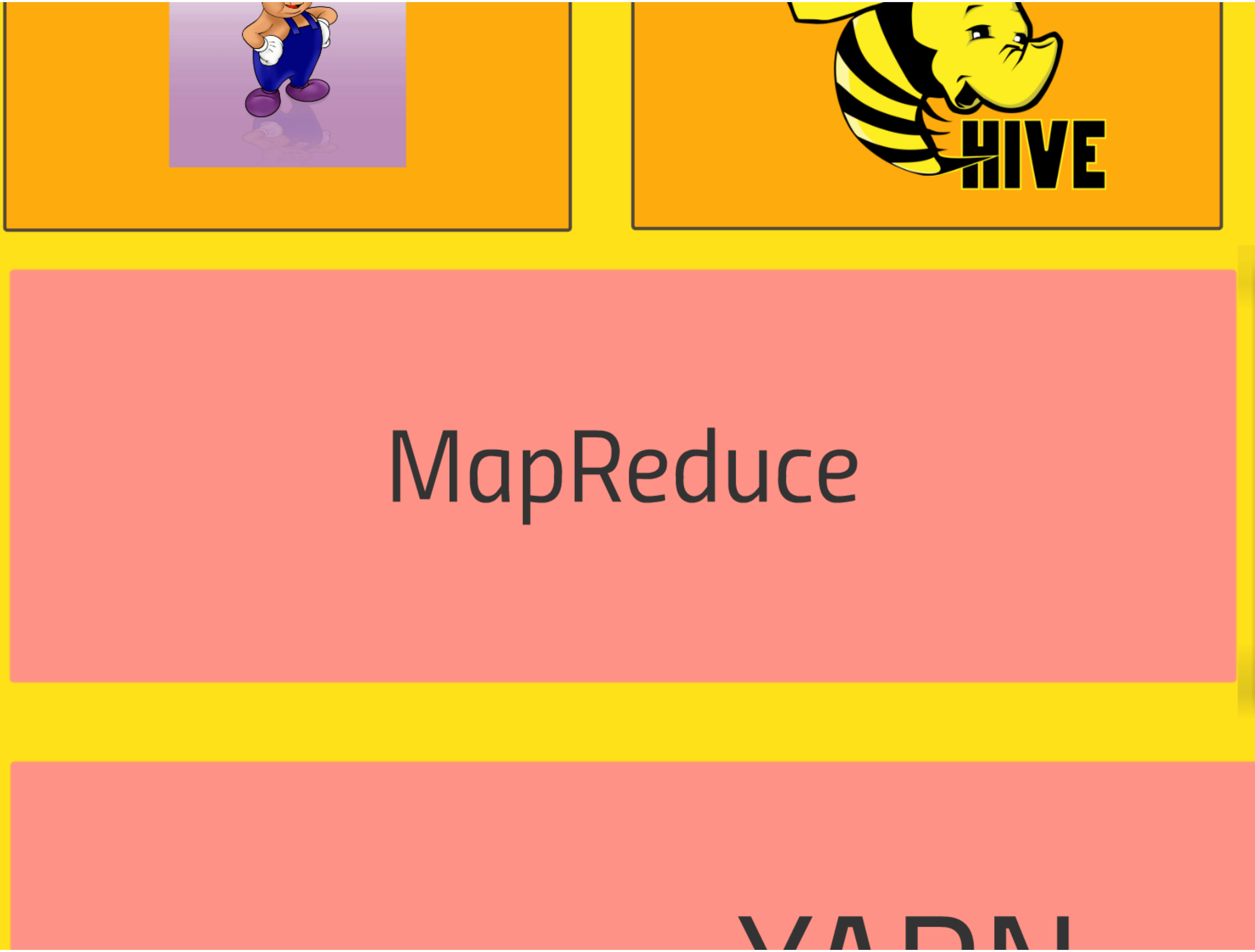


MapReduce

TEZ

YARN

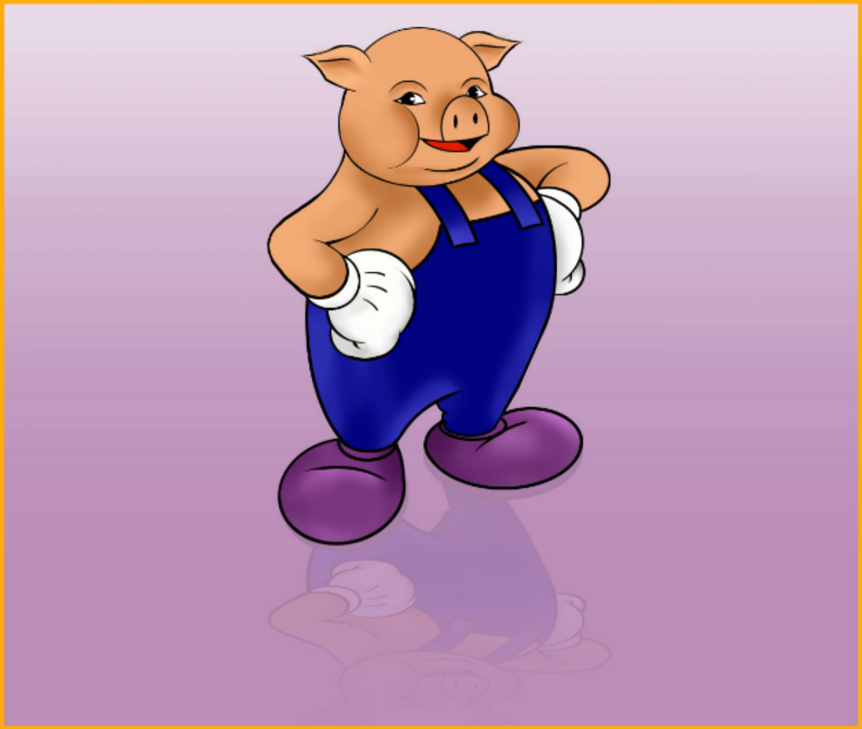
HDFS



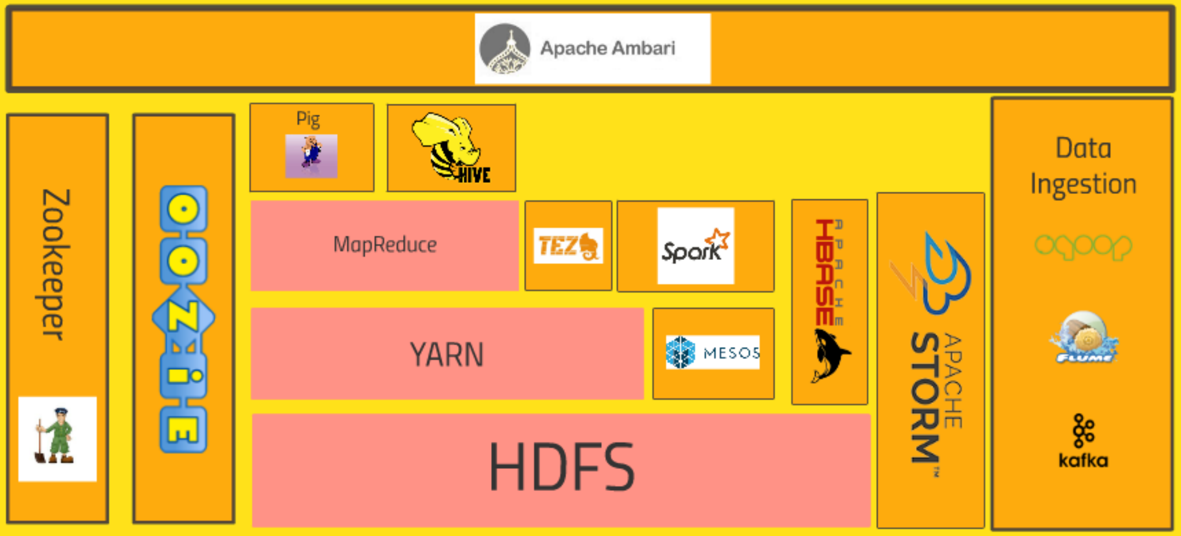
MapReduce

YARN

Pig



Core Hadoop Ecosystem



Zookeeper

Pig

MapReduce

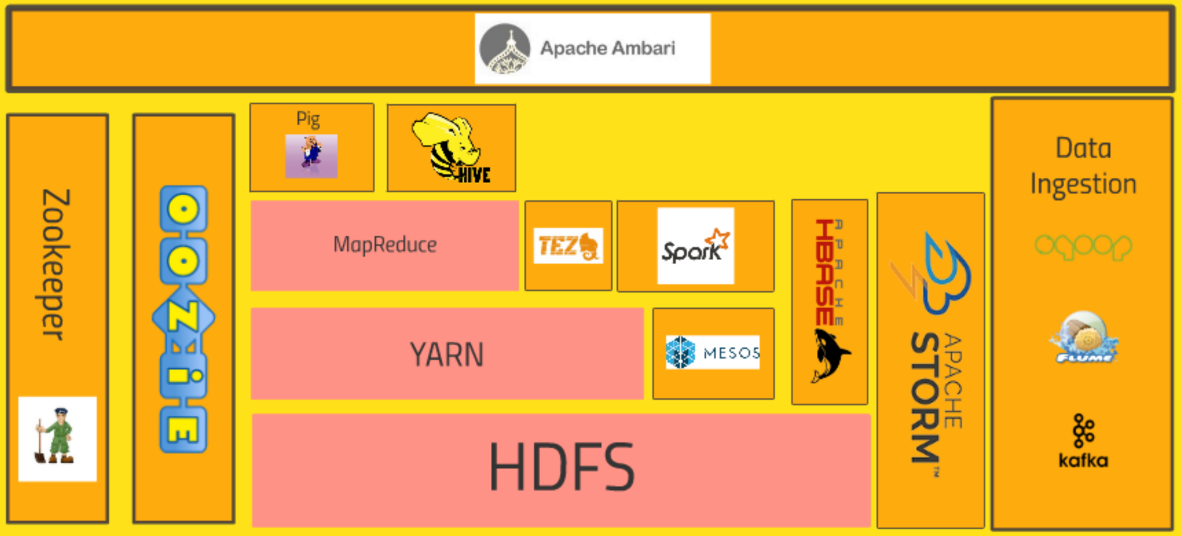
YARN

HDFS

Data Ingestion

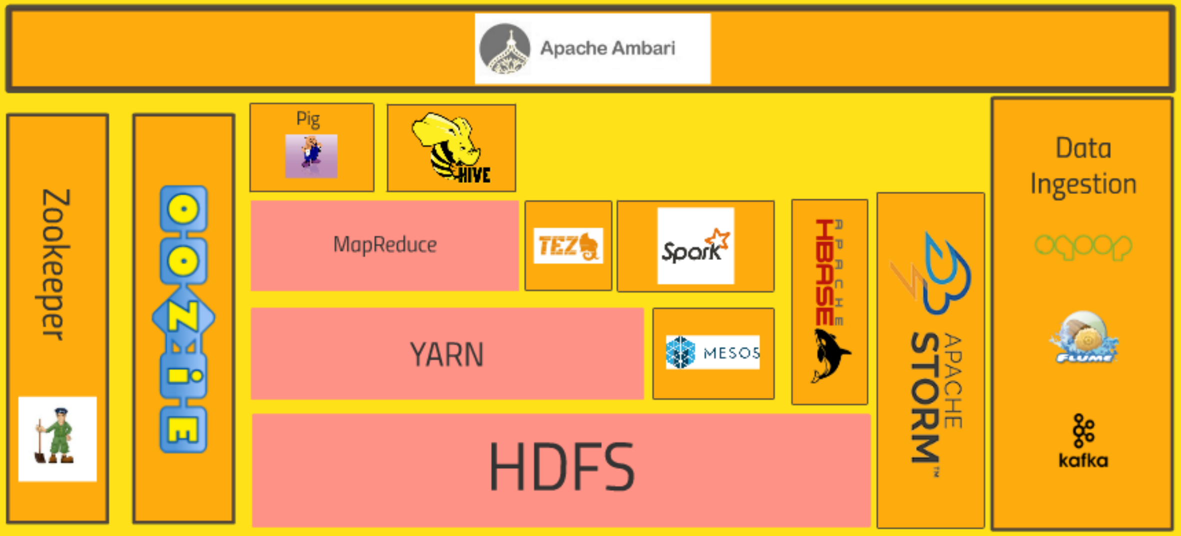
HIVE

Core Hadoop Ecosystem



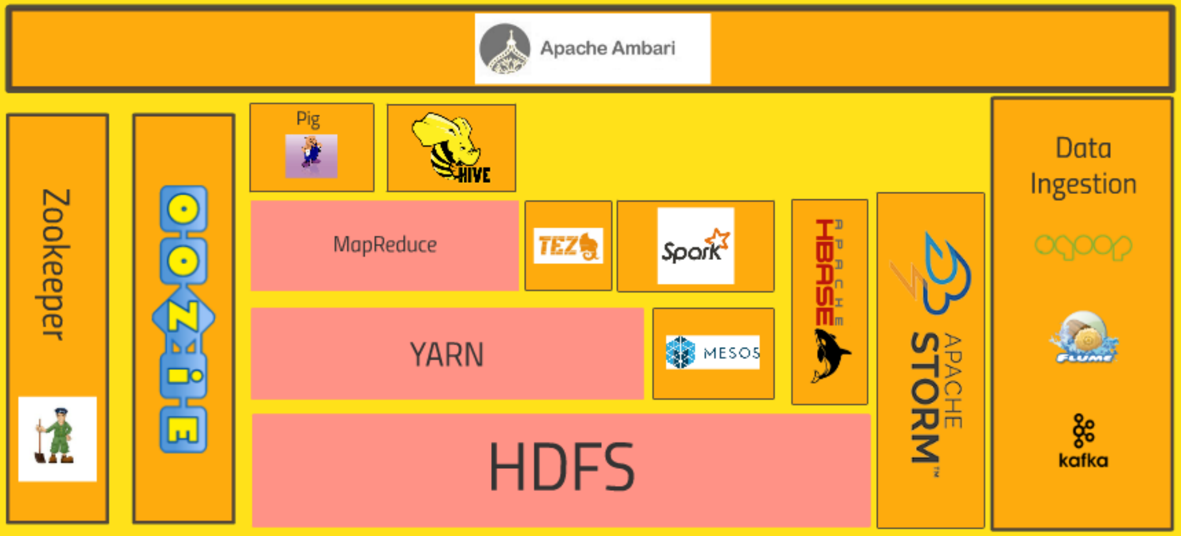
Apache Ambari

Core Hadoop Ecosystem



MESOS

Core Hadoop Ecosystem



Apache Ambari

Zookeeper

Pig

MapReduce

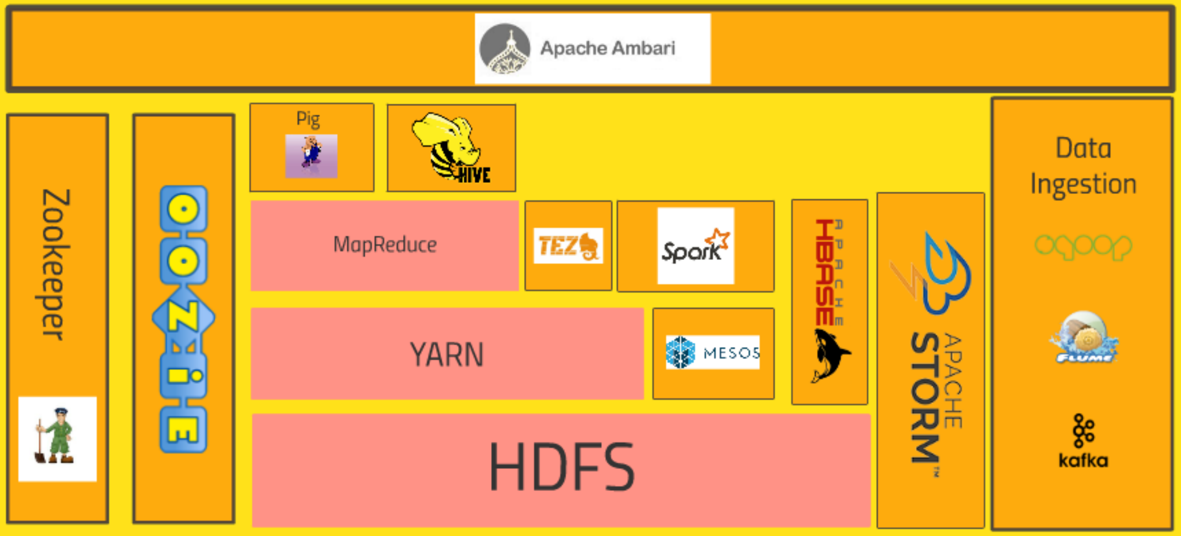
YARN

HDFS

Data Ingestion

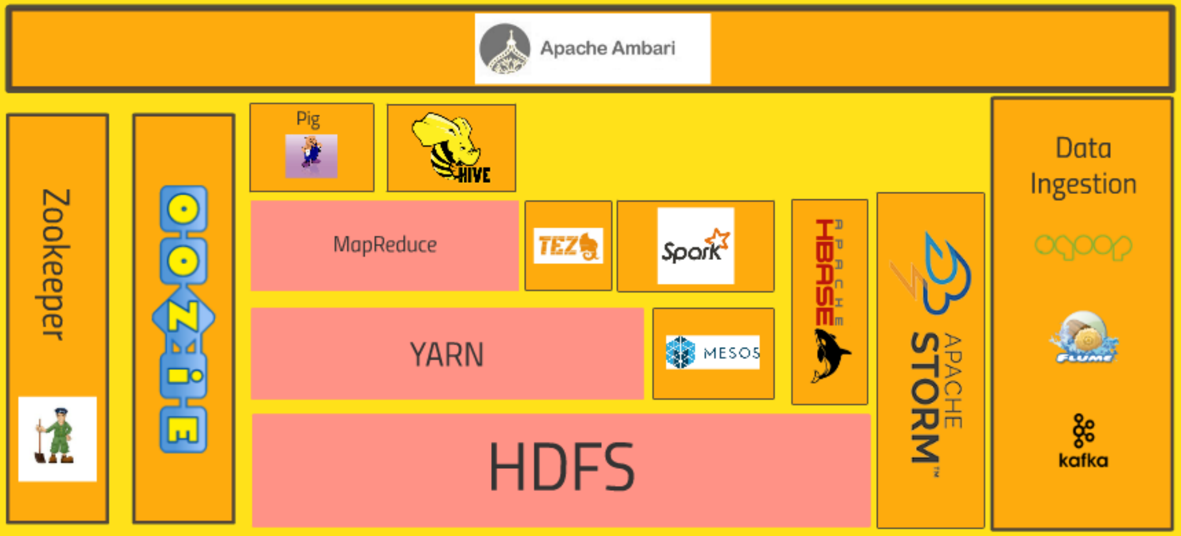
Spark

Core Hadoop Ecosystem



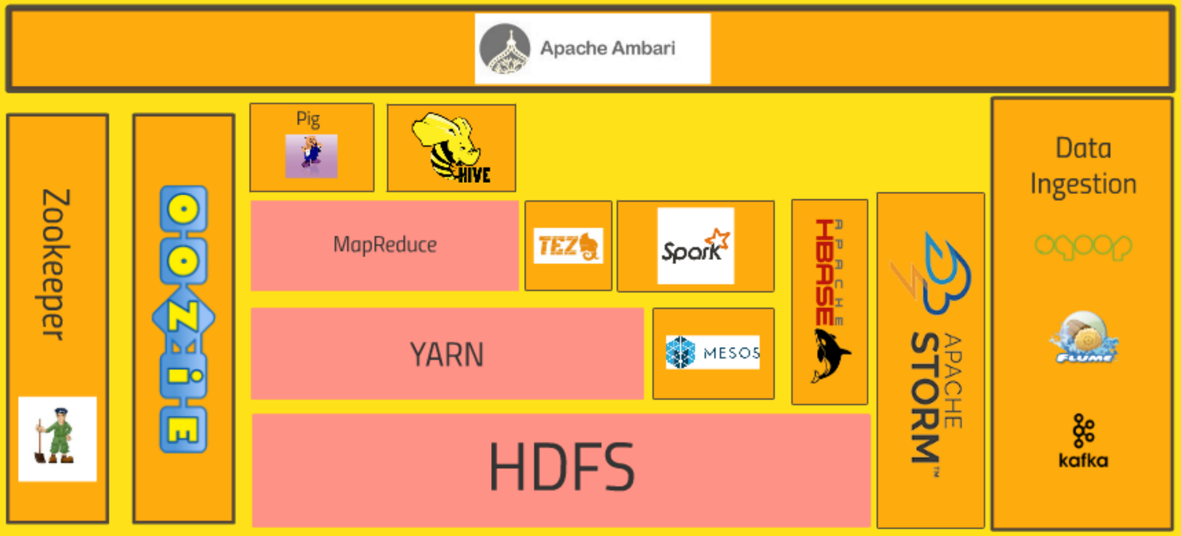
TEZ

Core Hadoop Ecosystem



APACHE
HBASE

Core Hadoop Ecosystem



Zookeeper

Pig

MapReduce

YARN

HDFS

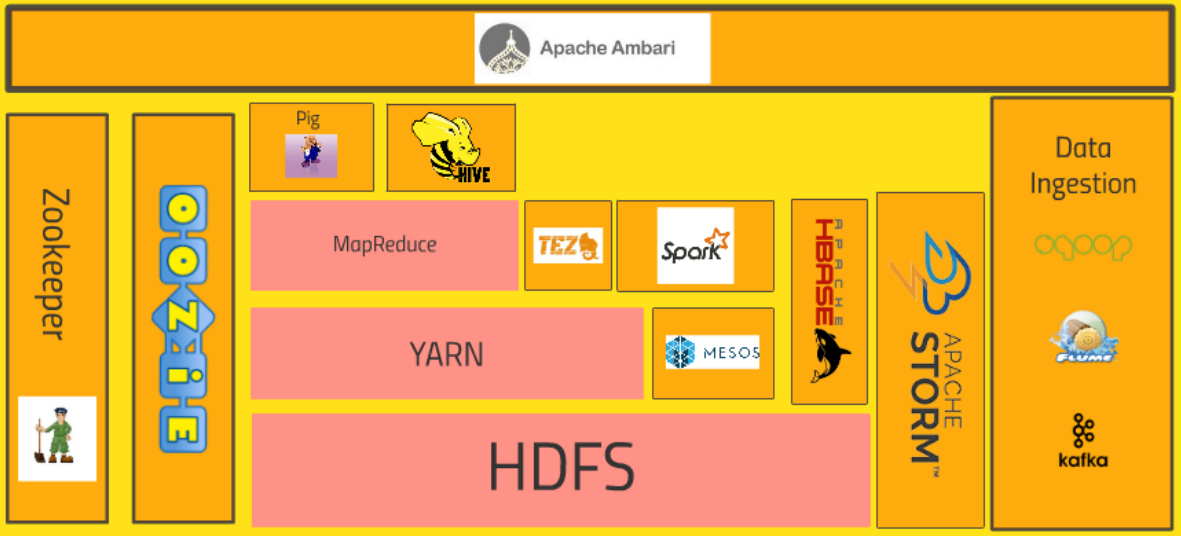
Data Ingestion

APACHE

STORM™

PACHE

Core Hadoop Ecosystem



Zookeeper

OOZiE

Pig

MapReduce

YAR

Pig

M

OOZiE

Zookeeper

Data Ingestion

sqoop

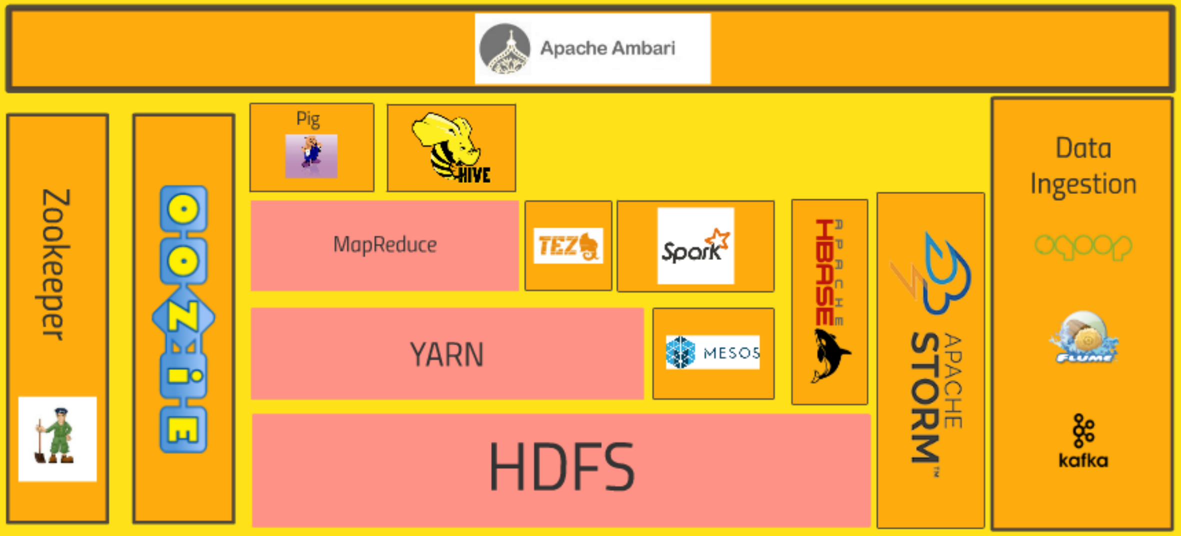
FLUME

kafka

APACHE STORM™

APACHE HBASE

Core Hadoop Ecosystem



Zookeeper

Pig

MapReduce

YARN

HDFS

Data Ingestion

External Data Storage

MySQL®

cassandra

mongoDB®

External Data Storage

Query Engines

APACHE DRILL

HUE

APACHE PHOENIX

presto

Apache Zeppelin

APACHE DRILL

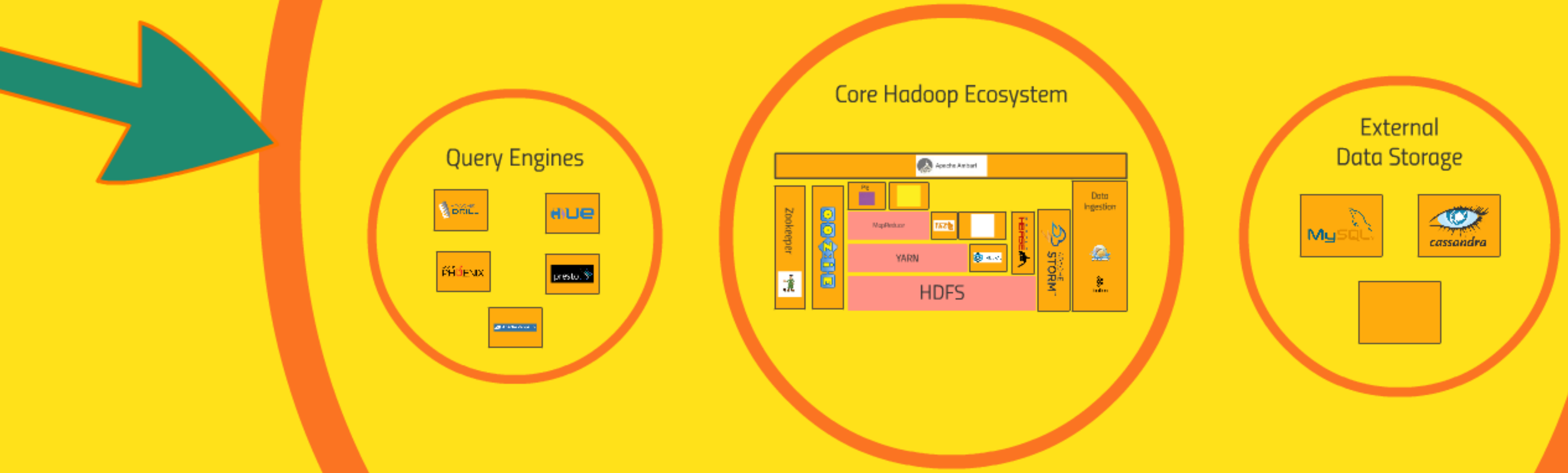
HUE

APACHE
PHOENIX

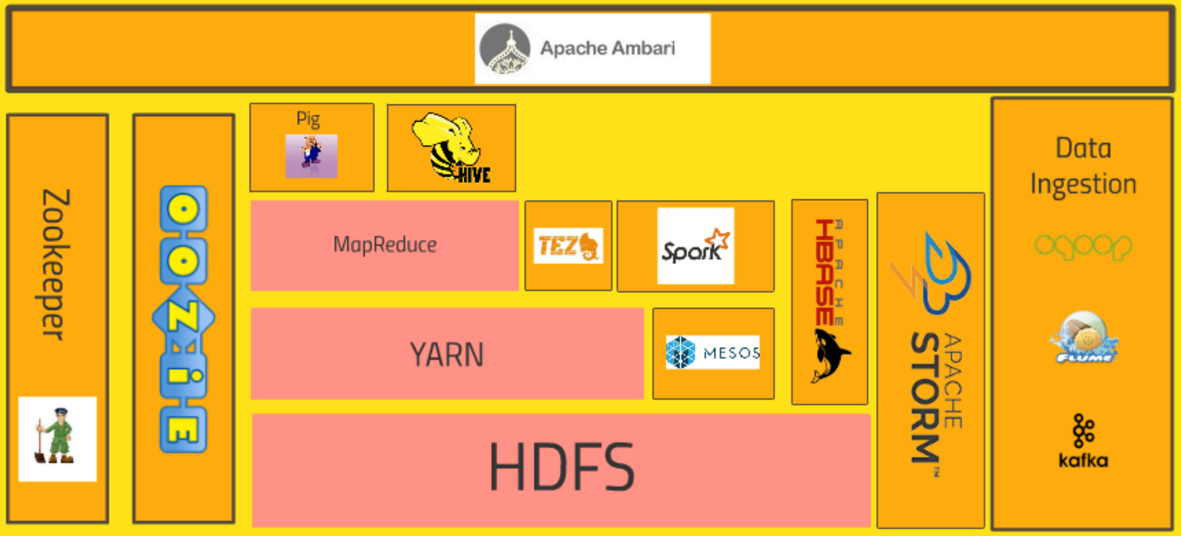
presto

Apache Zeppelin

World of Hadoop



Core Hadoop Ecosystem



External Data Storage

Query Engines

APACHE DRILL

HUE

APACHE PHOENIX

presto

Apache Zeppelin