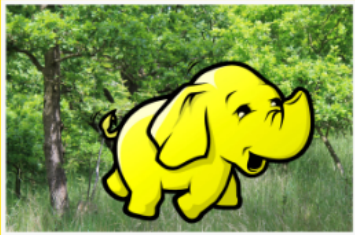


The Hadoop Ecosystem



Hadoop Overview

What is Hadoop?

The open source software platform for distributed storage and processing of large data sets across commodity hardware. It is designed to scale up to large numbers of machines, and to support high throughput, distributed processing.

Why Hadoop?



- Scalable architecture
- Fault tolerance
- High throughput
- Distributed processing
- High availability

Hadoop History



Hadoop was developed by the Apache Software Foundation. It was inspired by Google's MapReduce framework and the distributed file system HDFS. The project was started in 2005 and has since become one of the most popular open-source frameworks for big data processing.

World of Hadoop

Query Engines



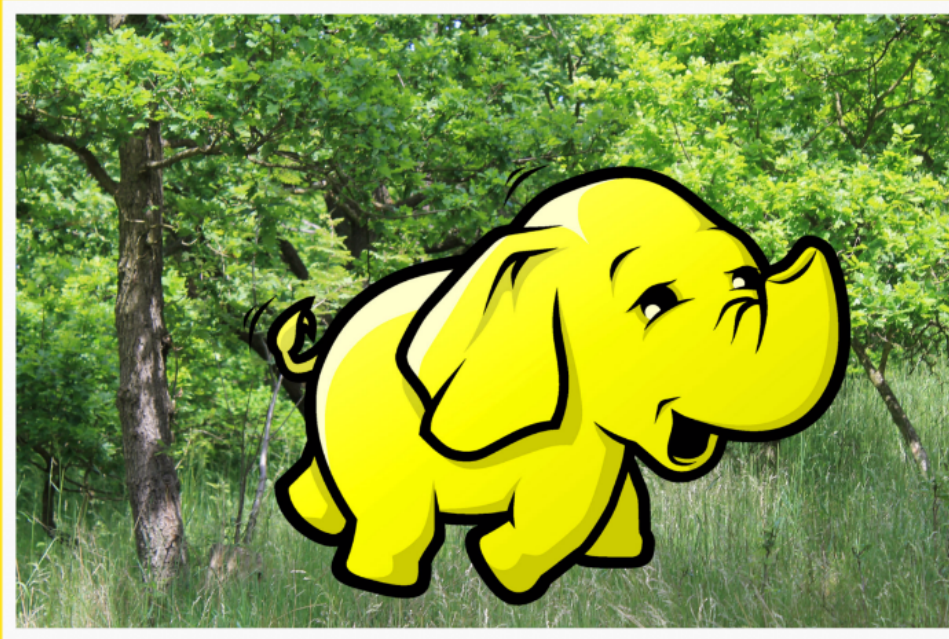
Core Hadoop Ecosystem



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?

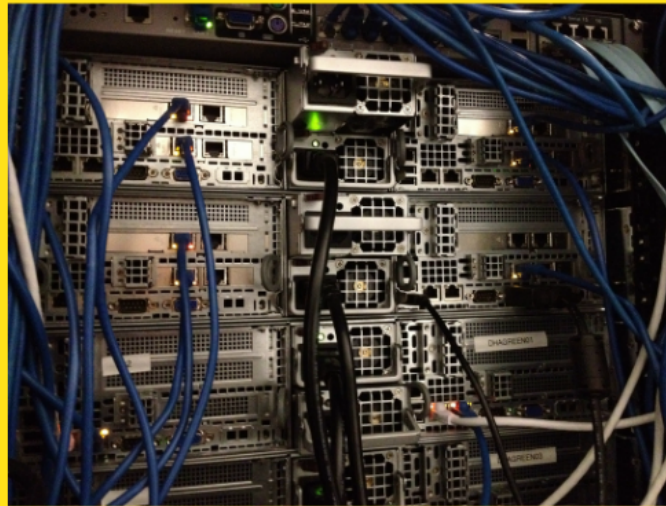
"an open source **software platform** for **distributed storage** and **distributed processing** of **very large data sets** on **computer clusters** built from commodity hardware" - *Hortonworks*

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?

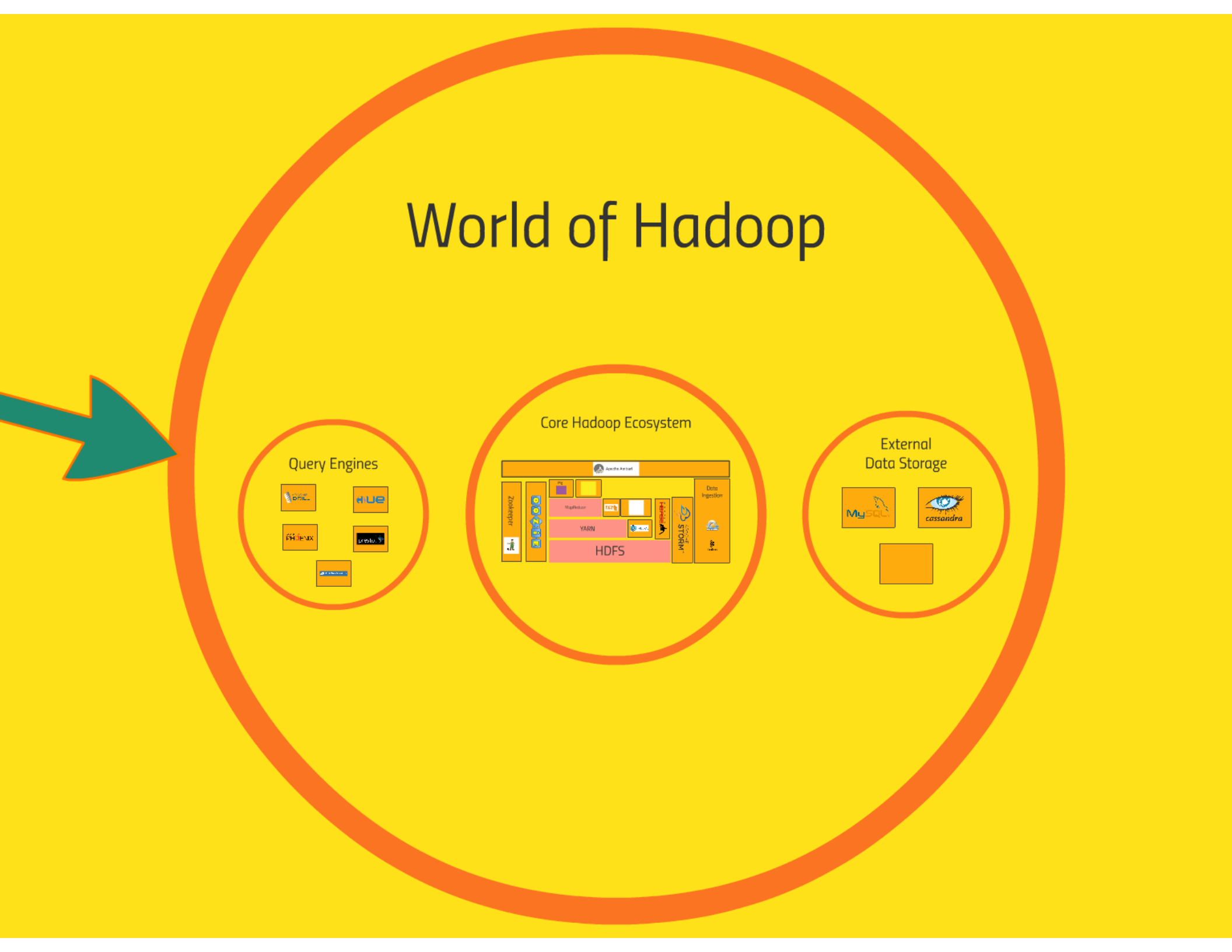


- Data's too darn big - terabytes per day
- Vertical scaling doesn't cut it
 - Disk seek times
 - Hardware failures
 - Processing times
- Horizontal scaling is linear
- Hadoop: It's not just for batch processing anymore

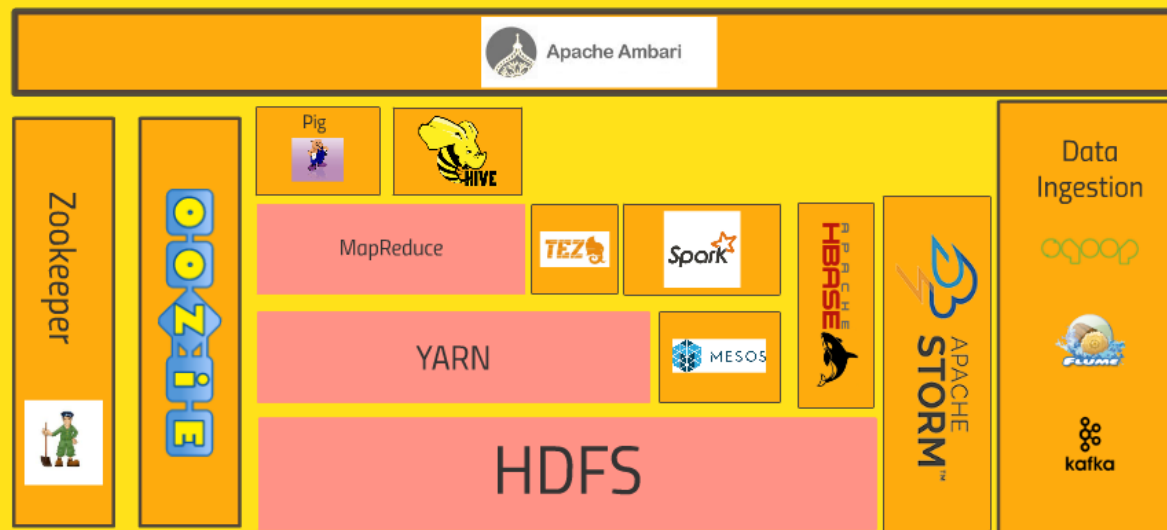
The diagram illustrates the 'World of Hadoop' ecosystem, centered around three main components:

- Core Hadoop Ecosystem:** The central hub, containing HDFS, YARN, MapReduce, and other foundational technologies.
- Query Engines:** Surrounding the core, this layer includes tools like Hive, Pig, Tez, and others for processing data.
- External Data Storage:** The outermost layer, featuring databases like MySQL and Cassandra for storing data outside the Hadoop cluster.

A large blue arrow points from the left towards the 'Query Engines' circle, indicating the flow of data or processing from external sources into the Hadoop ecosystem.



Core Hadoop Ecosystem



MapReduce



YARN



HDFS

MapReduce



YARN

HDFS



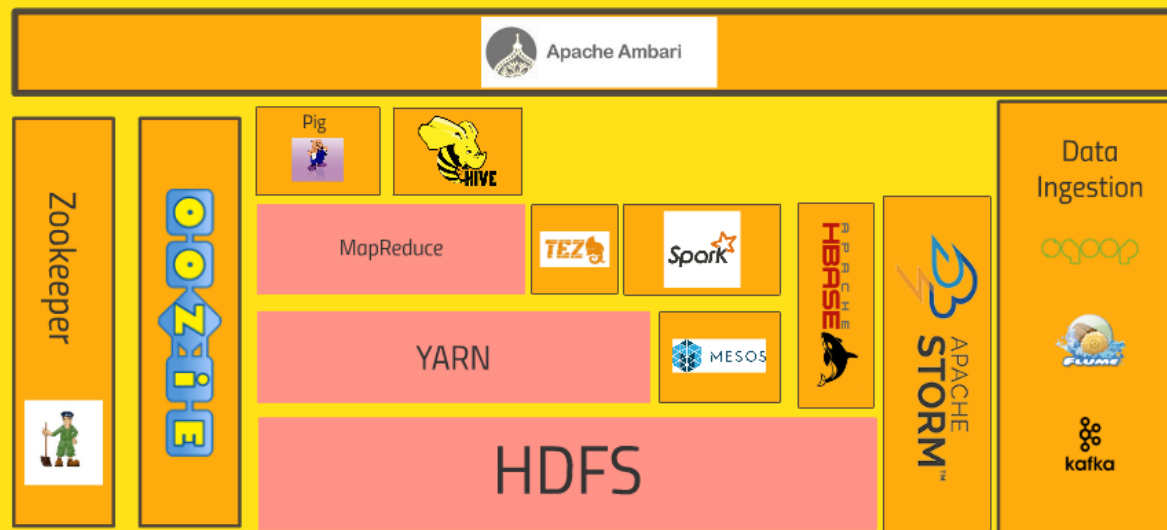
MapReduce

VADN

Pig



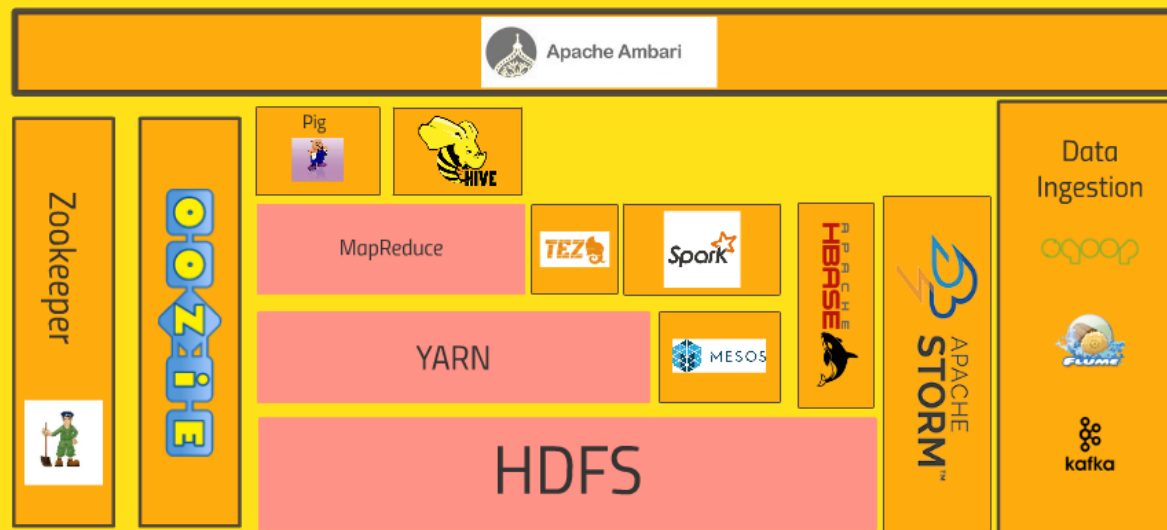
Core Hadoop Ecosystem





HIVE

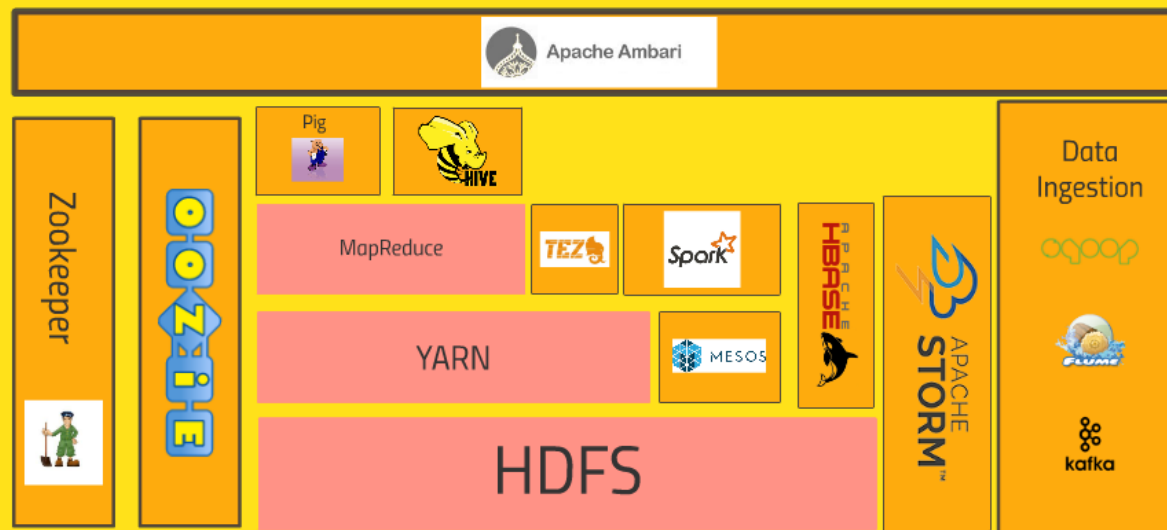
Core Hadoop Ecosystem





Apache Ambari

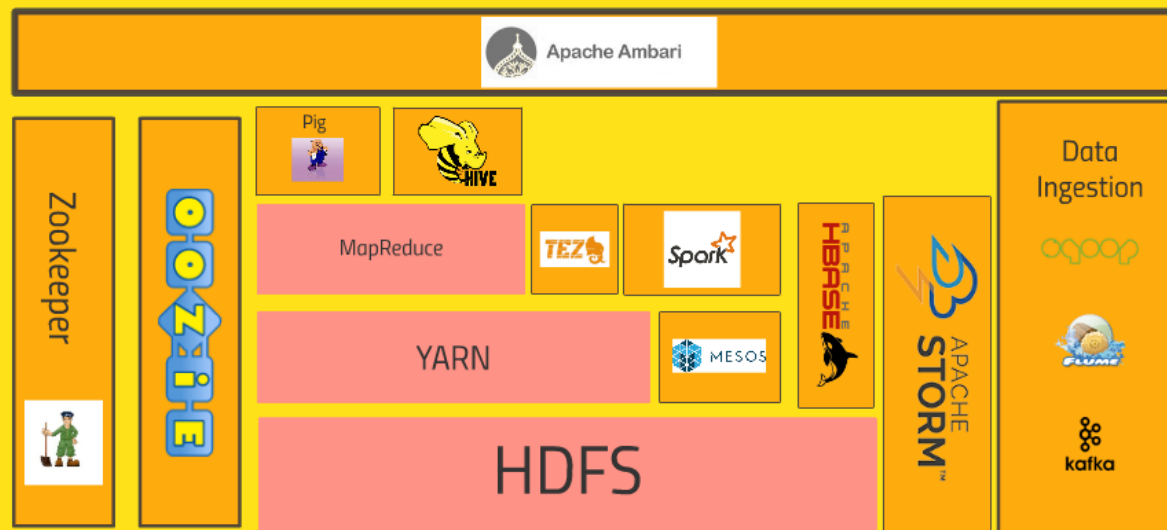
Core Hadoop Ecosystem





MESOS

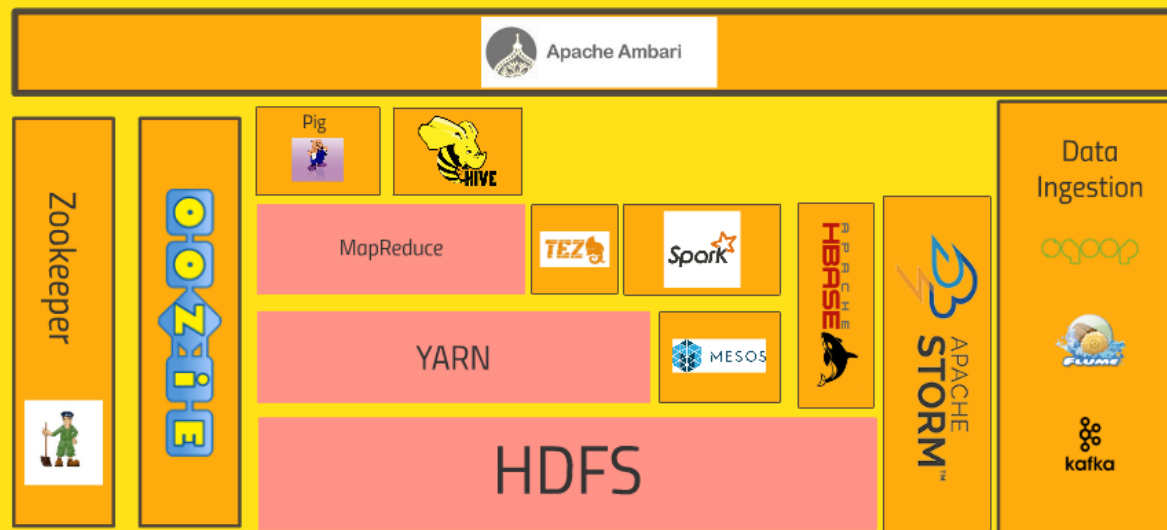
Core Hadoop Ecosystem



Spark

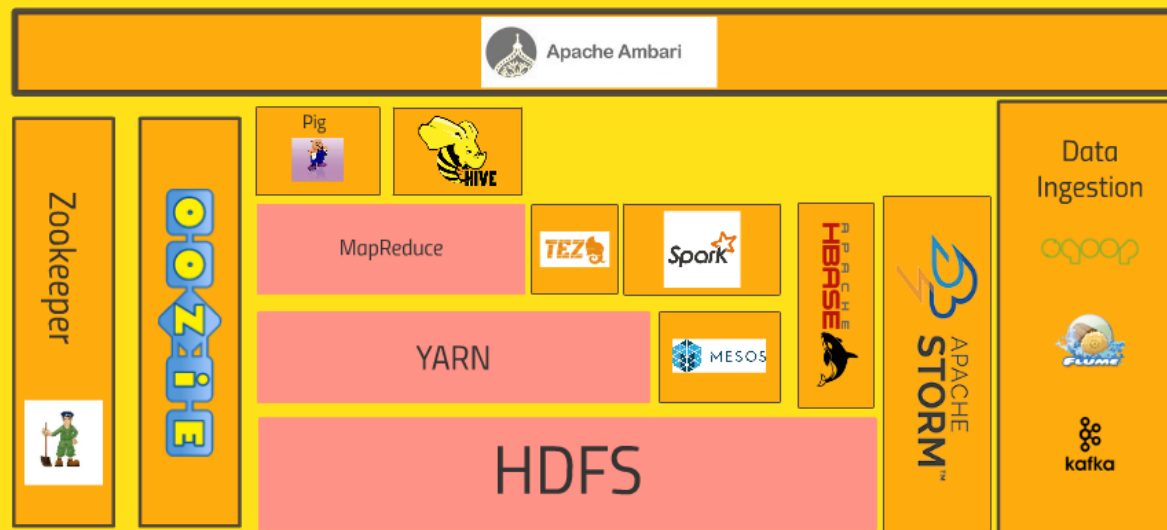
An orange outline of a five-pointed star is positioned behind the letter 'k' in the word 'Spark'. The star is slightly offset to the right and upwards, creating a layered effect. The word 'Spark' is written in a bold, dark grey, sans-serif font. The entire logo is centered on a white background, flanked by two vertical orange bars on the left and right sides.

Core Hadoop Ecosystem





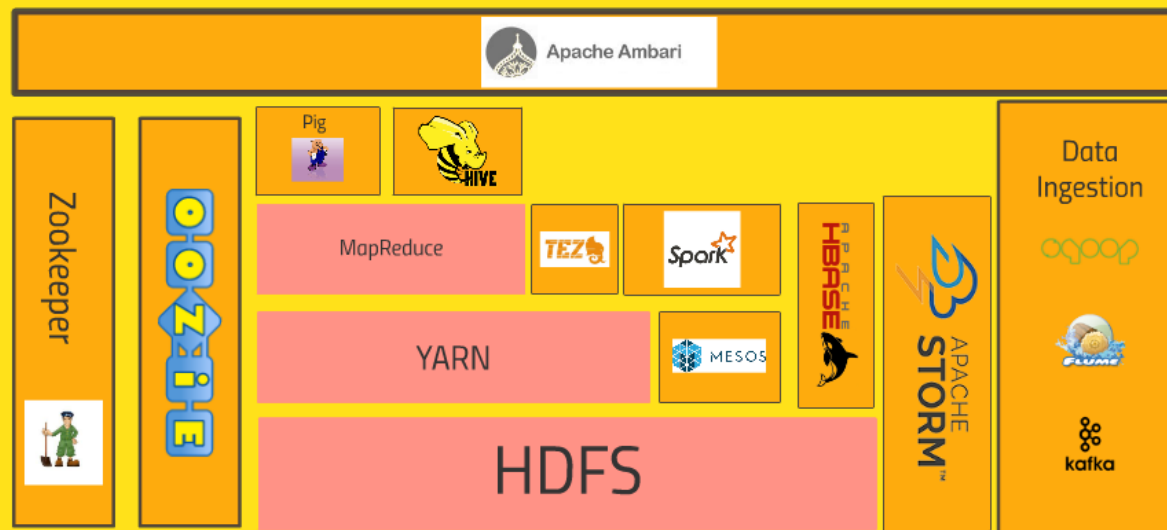
Core Hadoop Ecosystem



A P A C H E
H B A S E



Core Hadoop Ecosystem



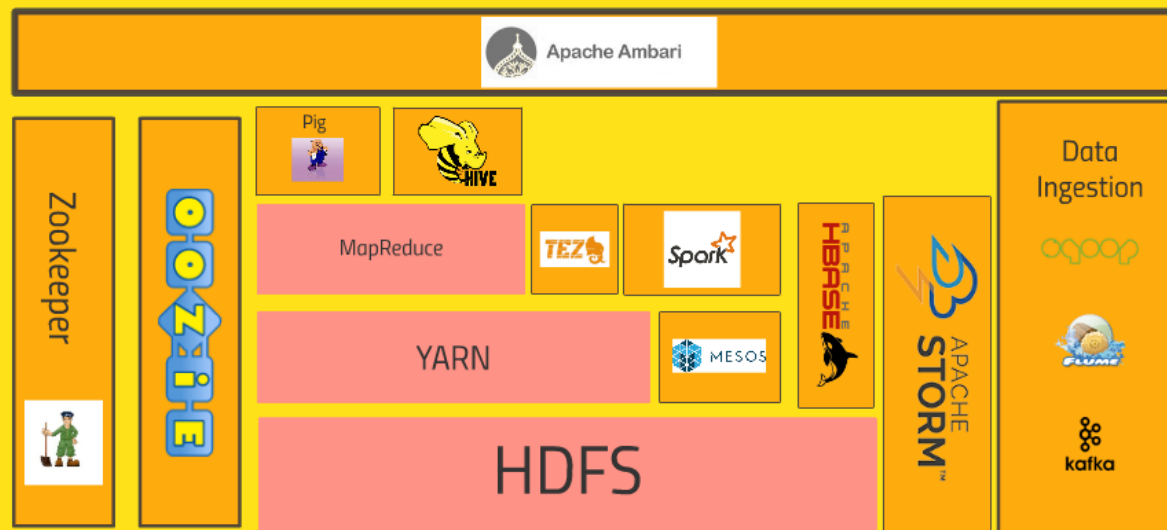


APACHE
STORMTM

P A C H E



Core Hadoop Ecosystem



Zookeeper



OOZIE

Pig



MapReduce

YAR

Pig



M

oozie

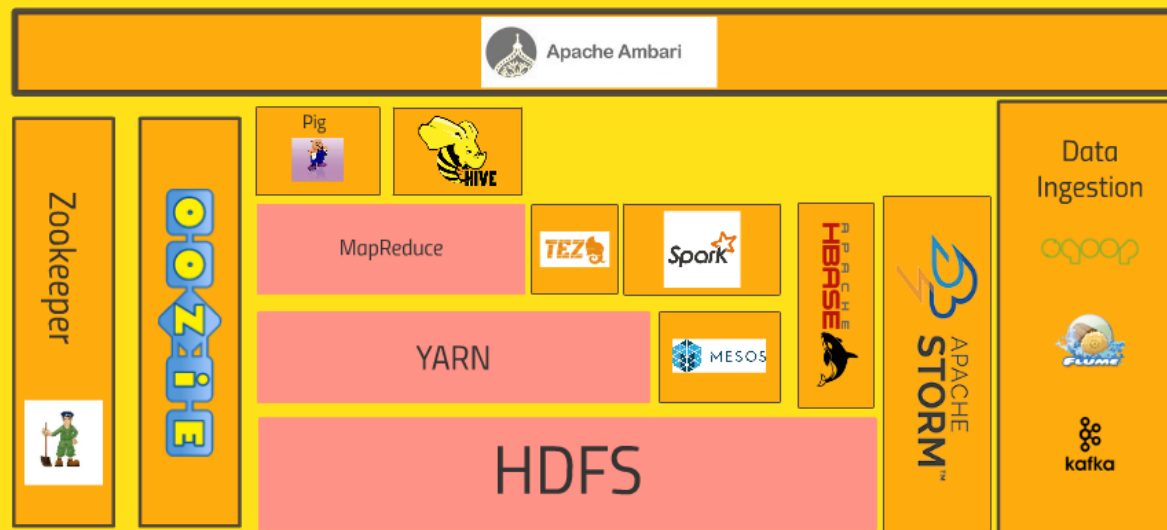
Zookeeper



Data Ingestion



Core Hadoop Ecosystem



External Data Storage







cassandra



mongoDB®

External Data Storage



Query Engines





APACHE
DRILL

hue

APACHE
PHOENIX

The logo for Apache Phoenix. It features the word "APACHE" in a small, black, sans-serif font above the word "PHOENIX". The "PHOENIX" is in a large, bold, black, sans-serif font. The letter "O" is replaced by a stylized phoenix bird, which is depicted in shades of orange and red, with its wings spread and tail feathers visible.

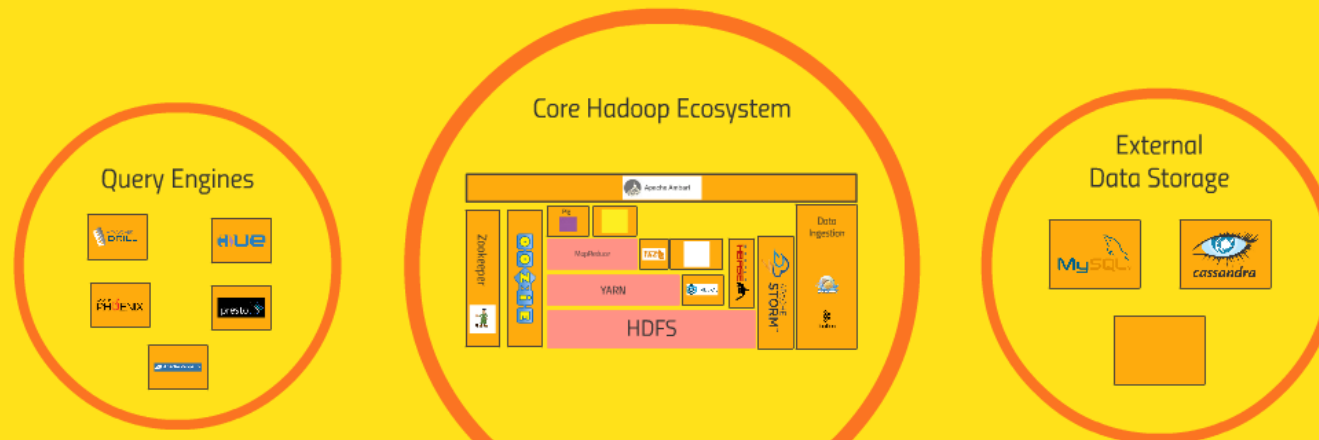
presto



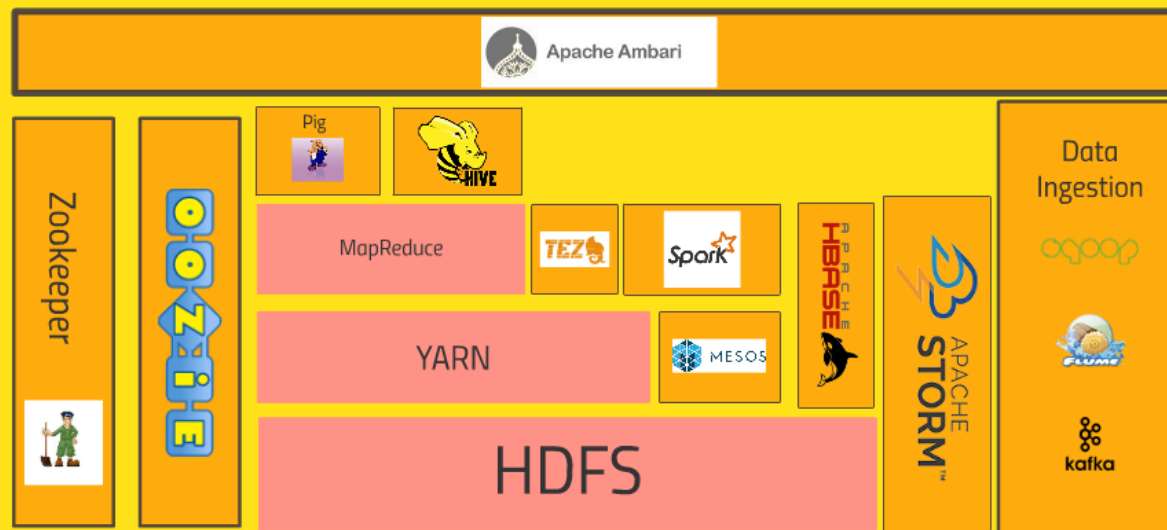


Apache Zeppelin

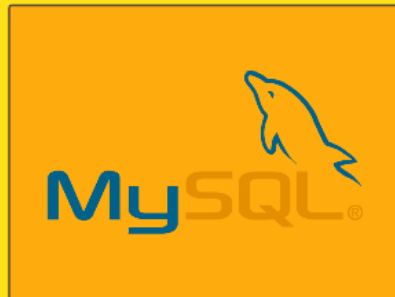
World of Hadoop



Core Hadoop Ecosystem



External Data Storage



Query Engines

