01

第1章 大数据关键技术

授课教师：[李阳源]

15138617705@163.com

主页：https://sites.google.com/view/dalguri



第1章 大数据技术概述


BIG
DATA
ANALYTICS
COMPLEX
SOCIAL
INTERNET
TECHNOLOGIES
STORAGE
SYSTEMS
SOFTWARE
INFORMATION
DATABASES
NETWORKS
RECORDS
SEARCH
PARALLEL
SIZE
RESEARCH
PETABYTES
INCLUDE
MANAGEMENT
SETS
SHARED
DISK
EXAMPLES
EVERY
ALSO
TOOLS
TARGET
SENSOR
DEFINITION
MOVING
MANAGE
ARCHIVES
PRACTITIONERS
CAPTURE
BUSINESS
NEEDED
TIME
LARGE
ONE
PERFORMANCE
STORE
LOGS
COST
SAN
MASSIVELY
GROW
DIFFICULTY
GARTNER
WORKING
ELAPSED
TOLERABLE
ABILITY
CASE
TERABYTES
DISTRIBUTED
GENOMICS
CURRENT
MAY
WITHIN
USING
PROCESS
RELATIONAL
CONTINUES
LARGER
TENS
CITATION
BIOLOGICAL
PROCESSING
CAPACITY
HUNDREDS
SET
USE

目录

CONTENT

- 大数据时代
- 大数据概念
- 大数据的影响
- 大数据关键技术
- 大数据计算模式
- 代表性大数据技术



1 Hadoop

2 Spark

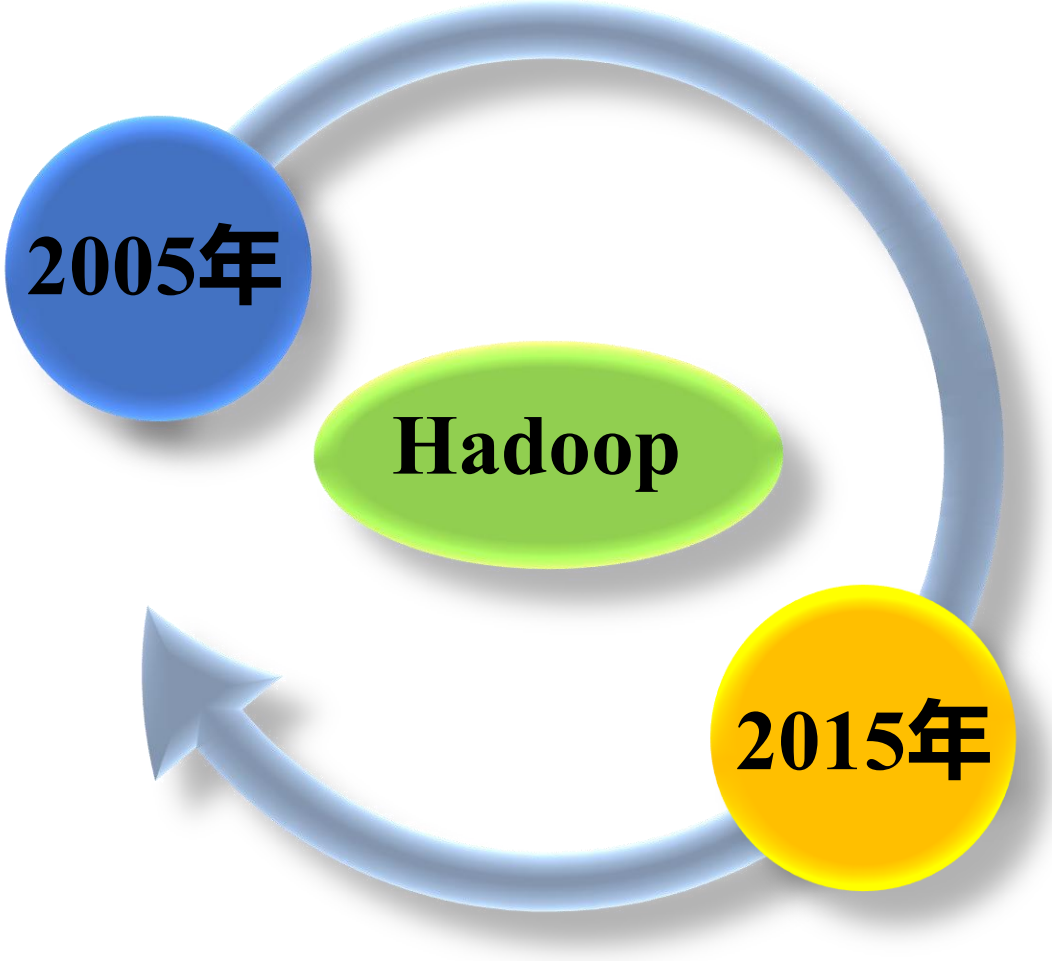
3 Flink

4 Beam



1.6.1 Hadoop

1.6.1 代表性的大数据技术之Hadoop





1.6.1 代表性的大数据技术之Hadoop



1.6.1 代表性的大数据技术之Hadoop



1.6.1 代表性的大数据技术之Hadoop

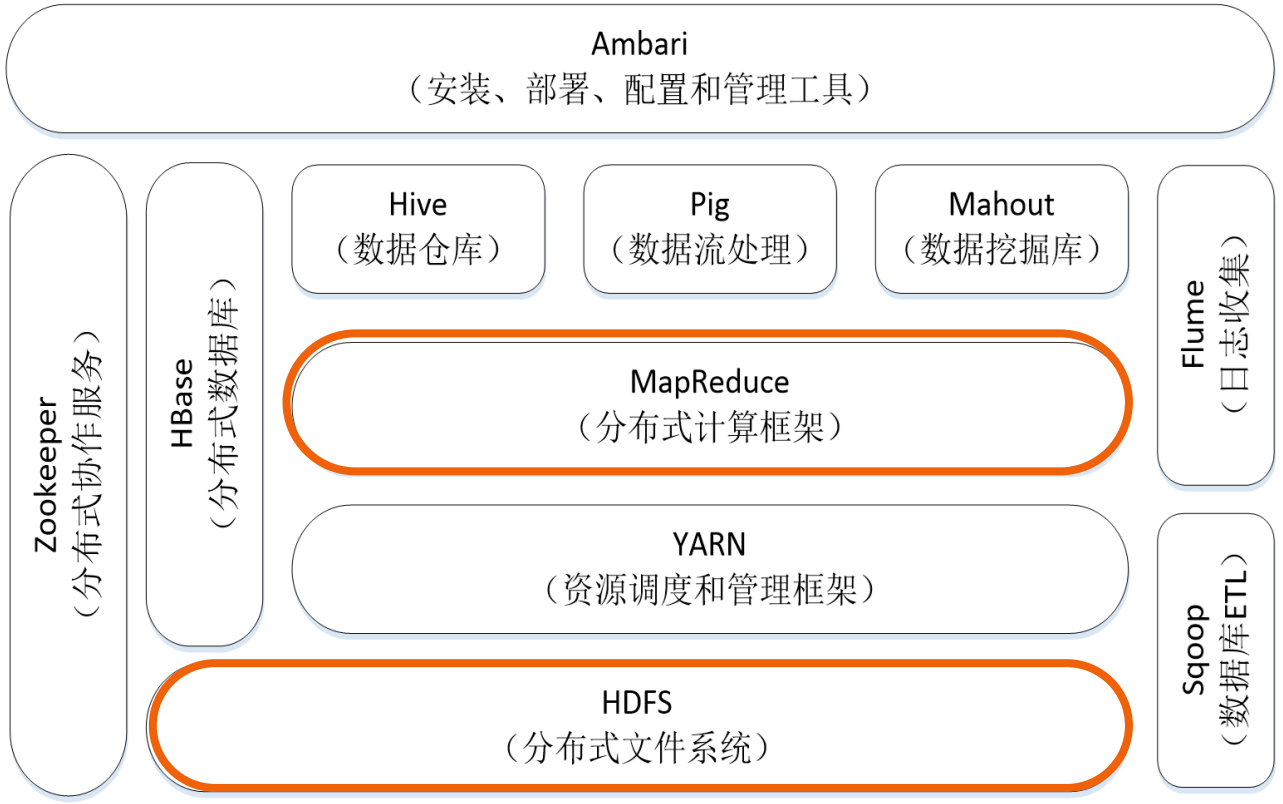


图　Hadoop生态系统

1.6.1 代表性的大数据技术之Hadoop



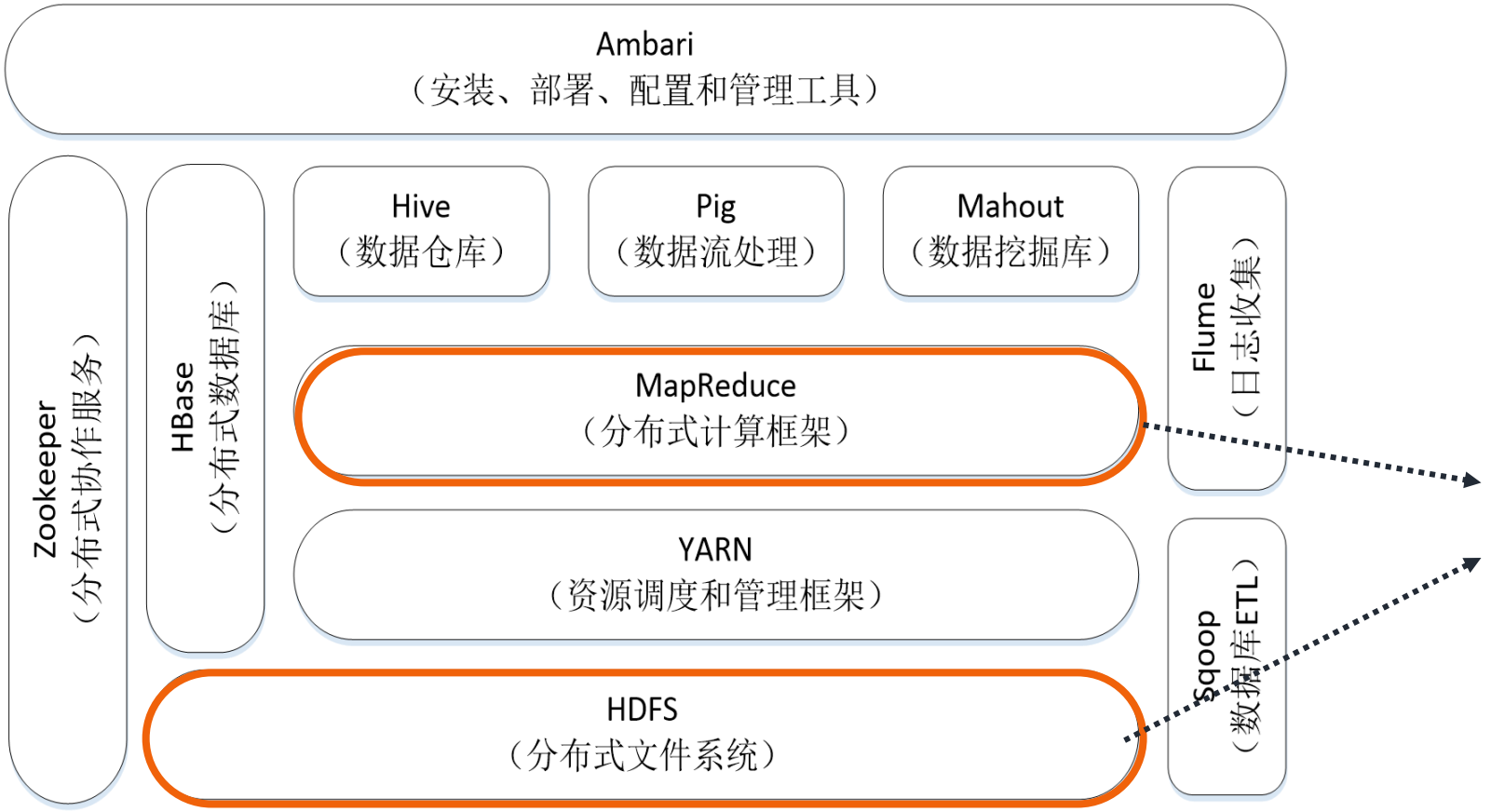


图　Hadoop生态系统

- **两大关键技术**



1.6.1 代表性的大数据技术之Hadoop



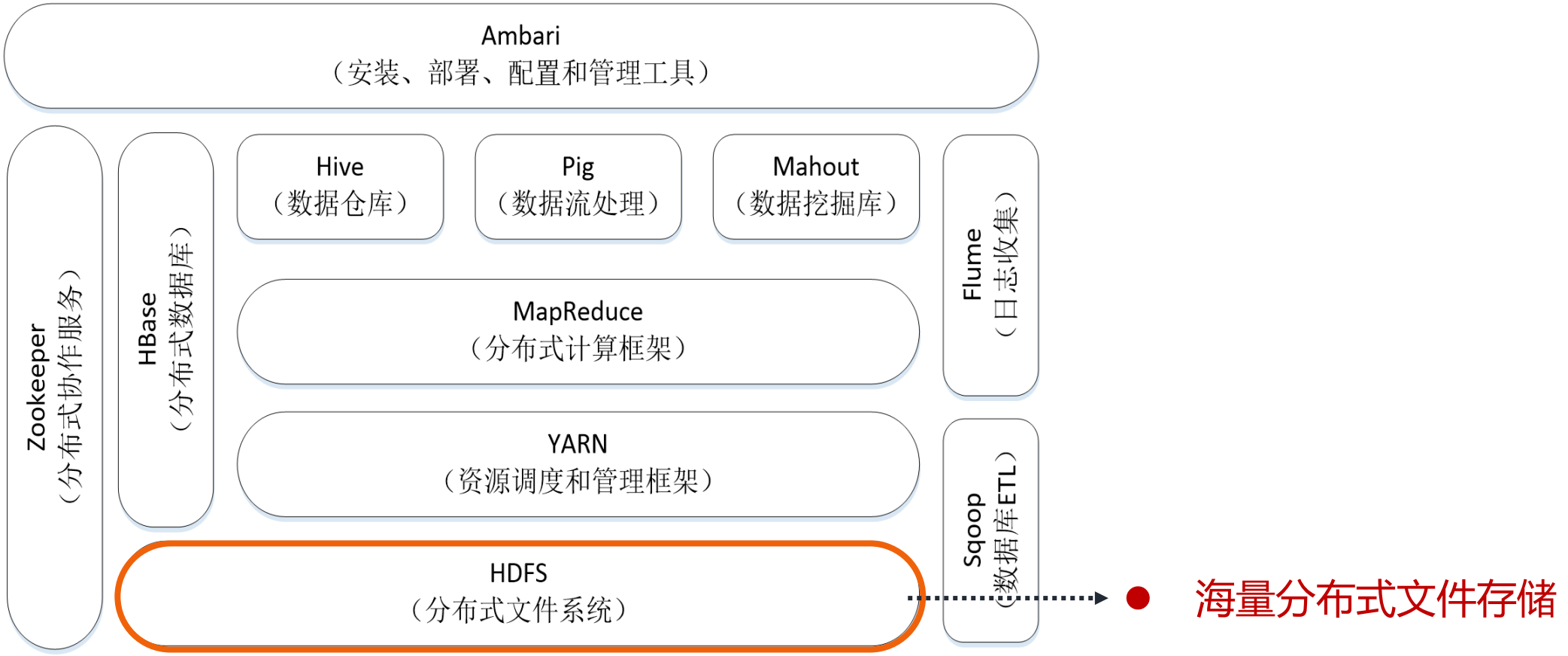
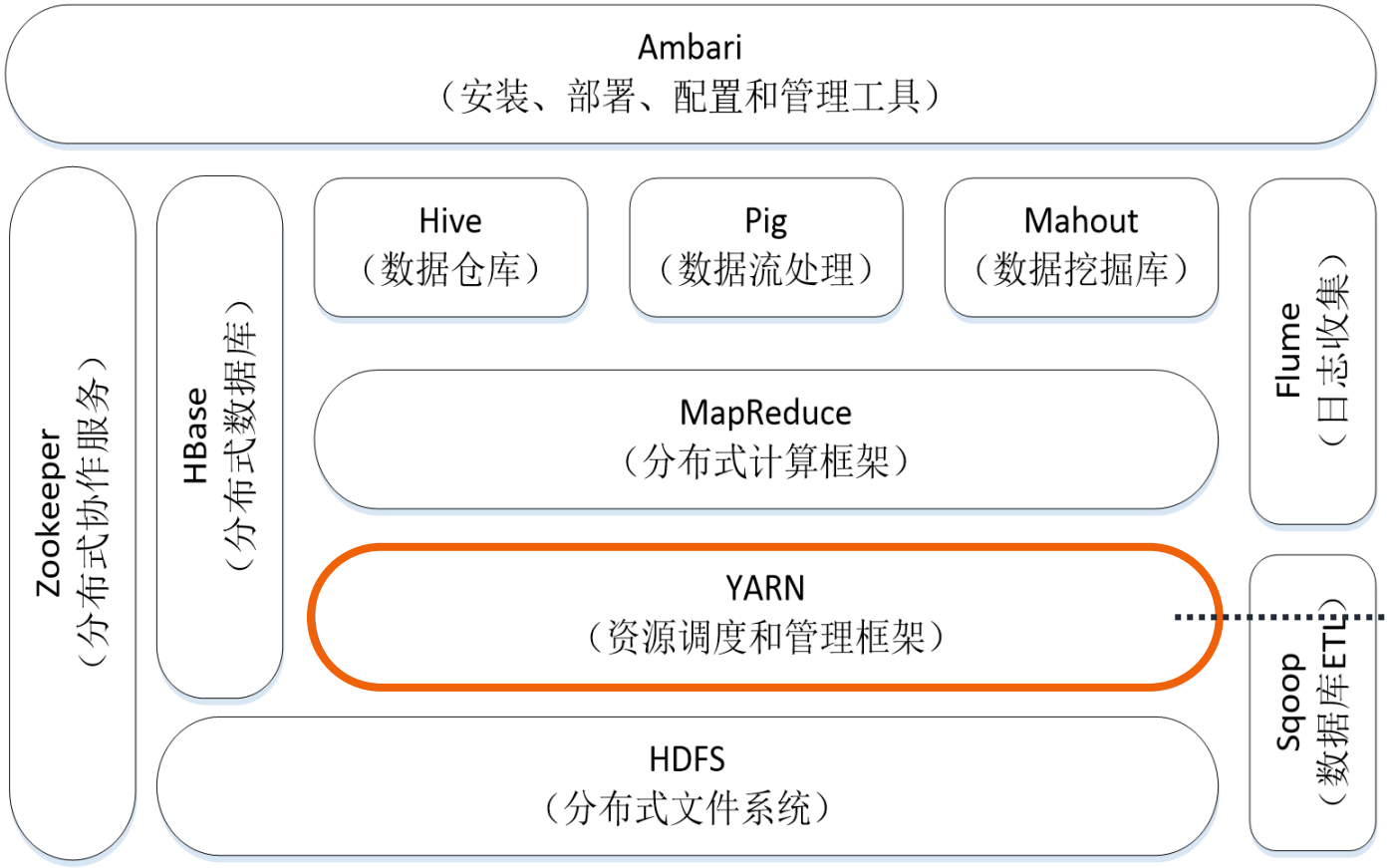


图　Hadoop生态系统

1.6.1 代表性的大数据技术之Hadoop





- 提供资源调度和管理服务

图　Hadoop生态系统

1.6.1 代表性的大数据技术之Hadoop

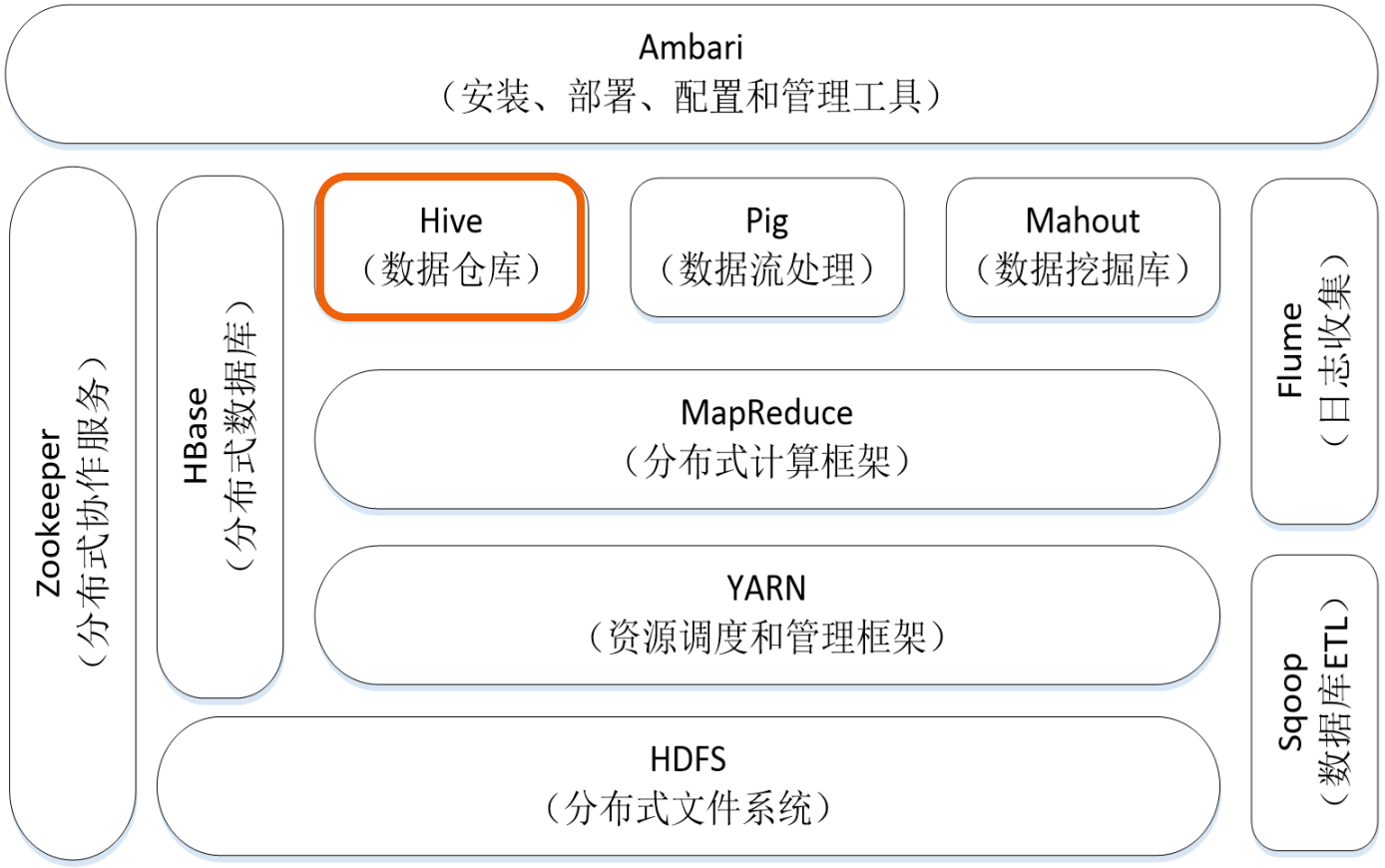


图　Hadoop生态系统

1.6.1 代表性的大数据技术之Hadoop

数据库

- 保存某一时刻的状态数据



库存：10件

1.6.1 代表性的大数据技术之Hadoop



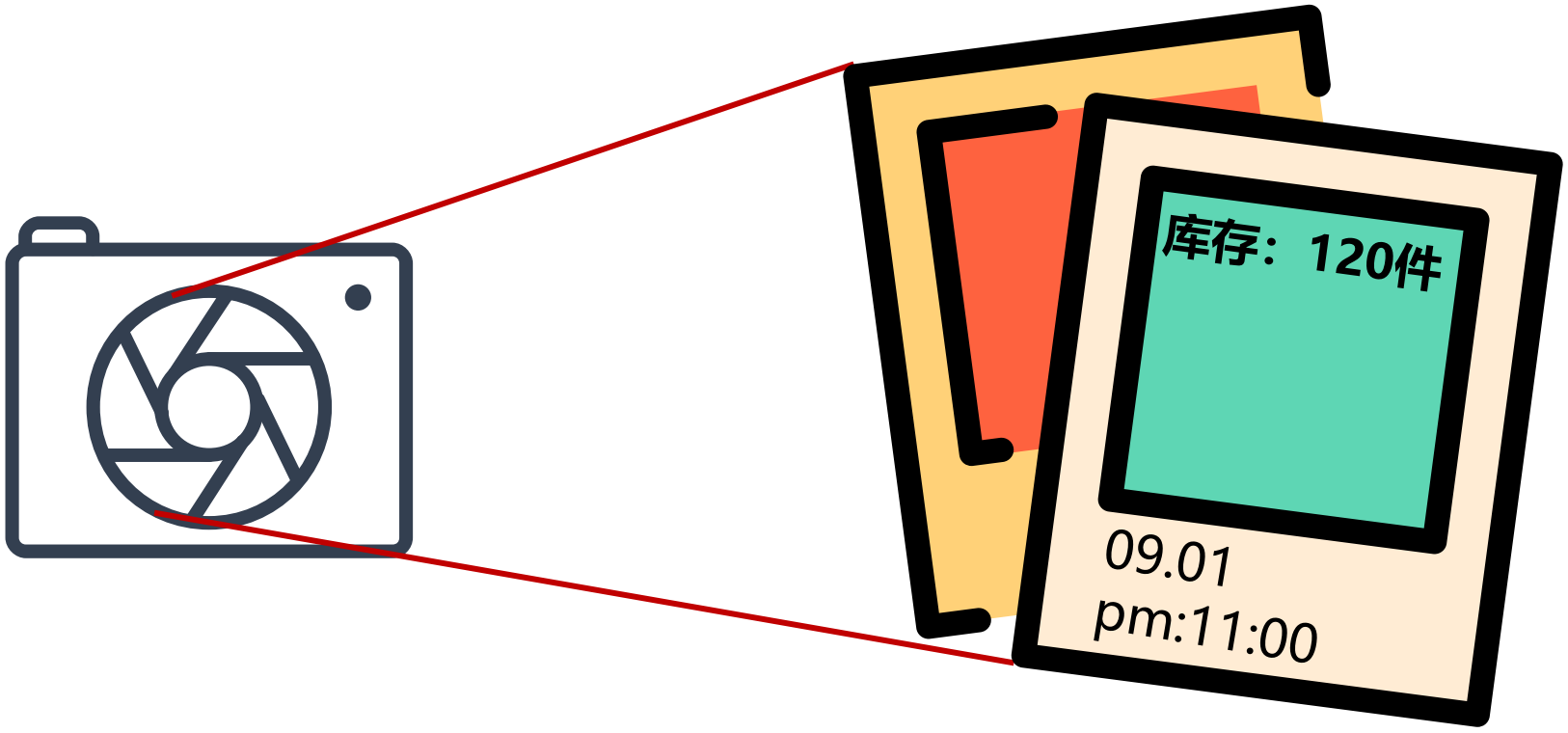
数据库



库存：9件

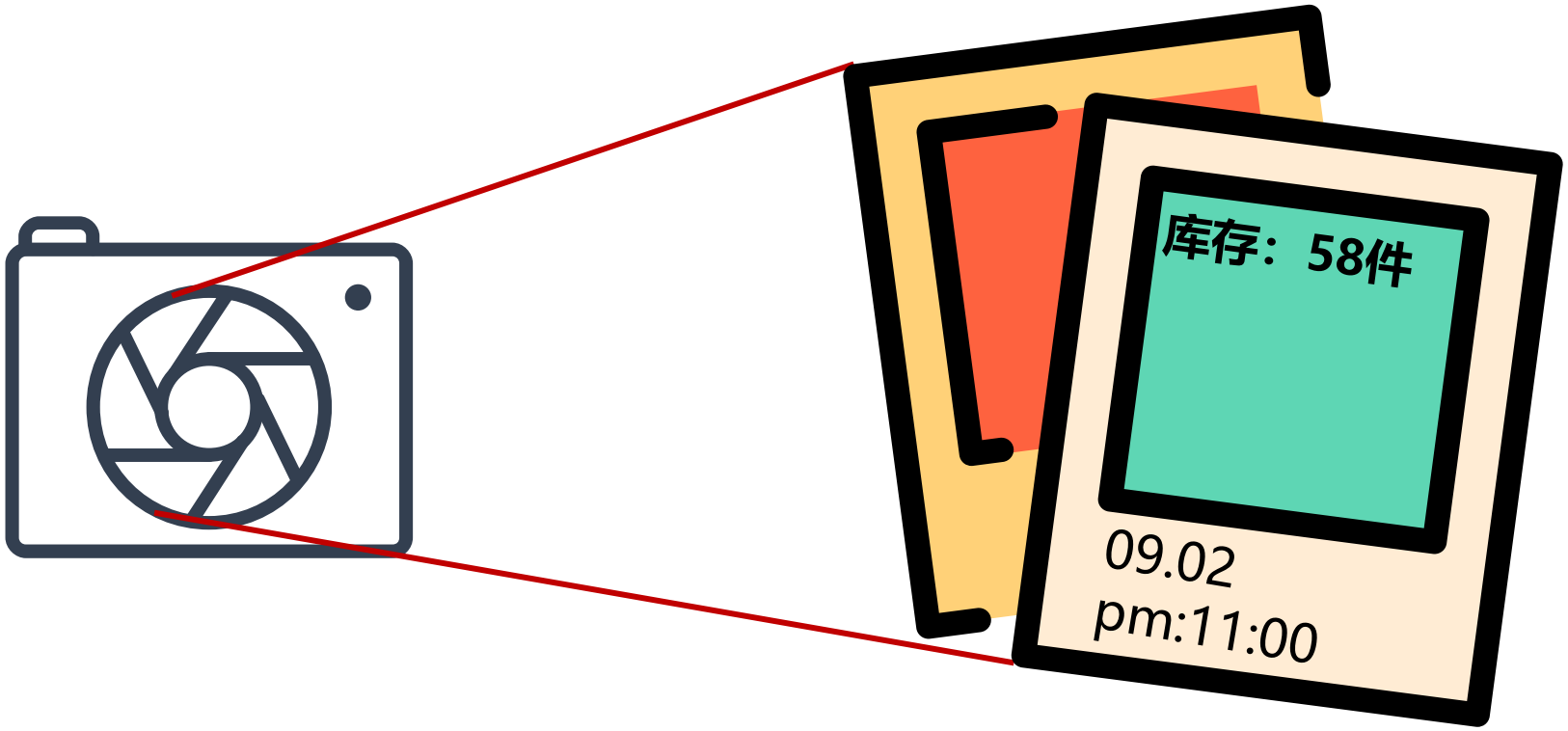
1.6.1 代表性的大数据技术之Hadoop

数据仓库



1.6.1 代表性的大数据技术之Hadoop

数据仓库



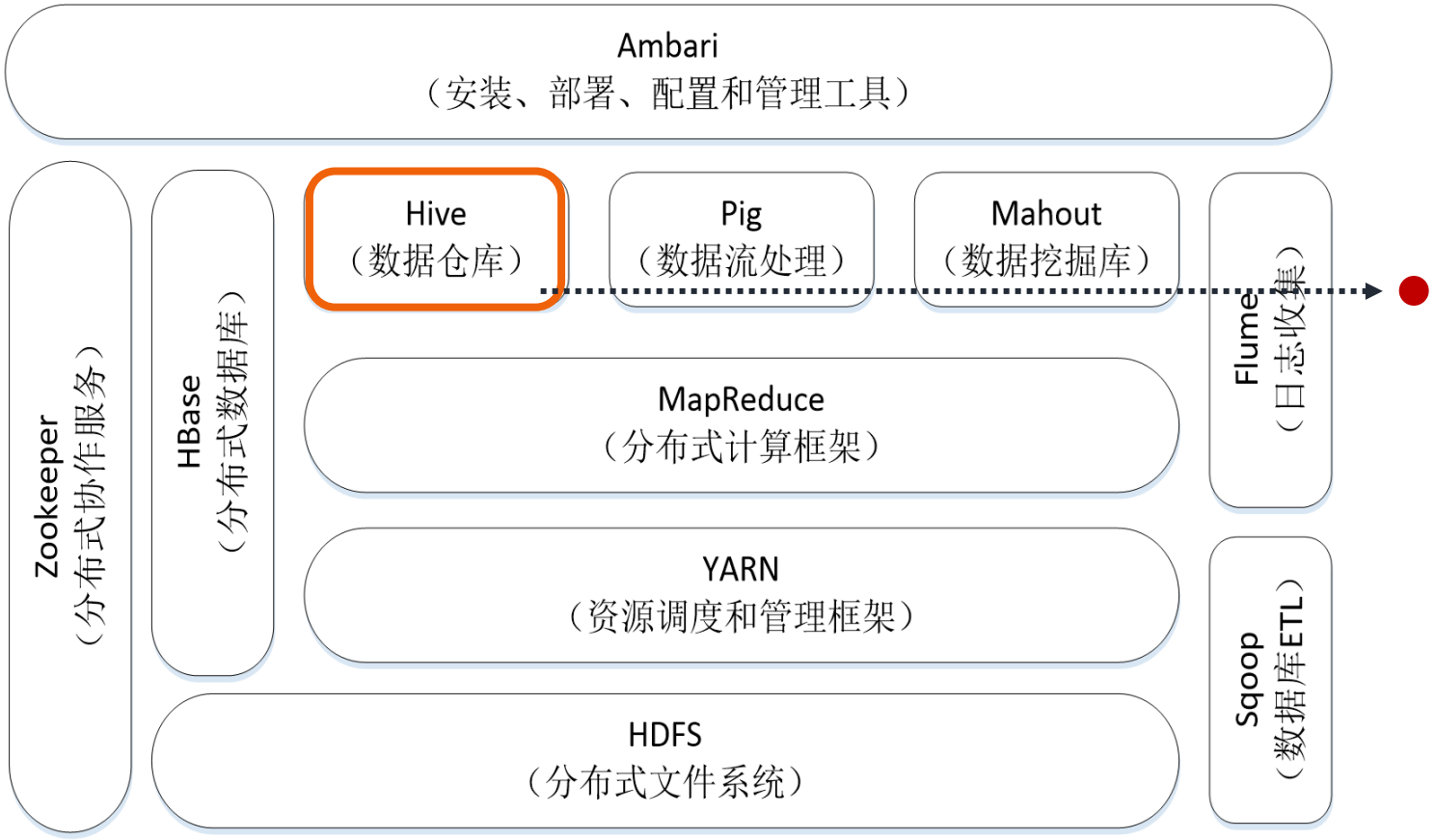
1.6.1 代表性的大数据技术之Hadoop

数据仓库



1.6.1 代表性的大数据技术之Hadoop

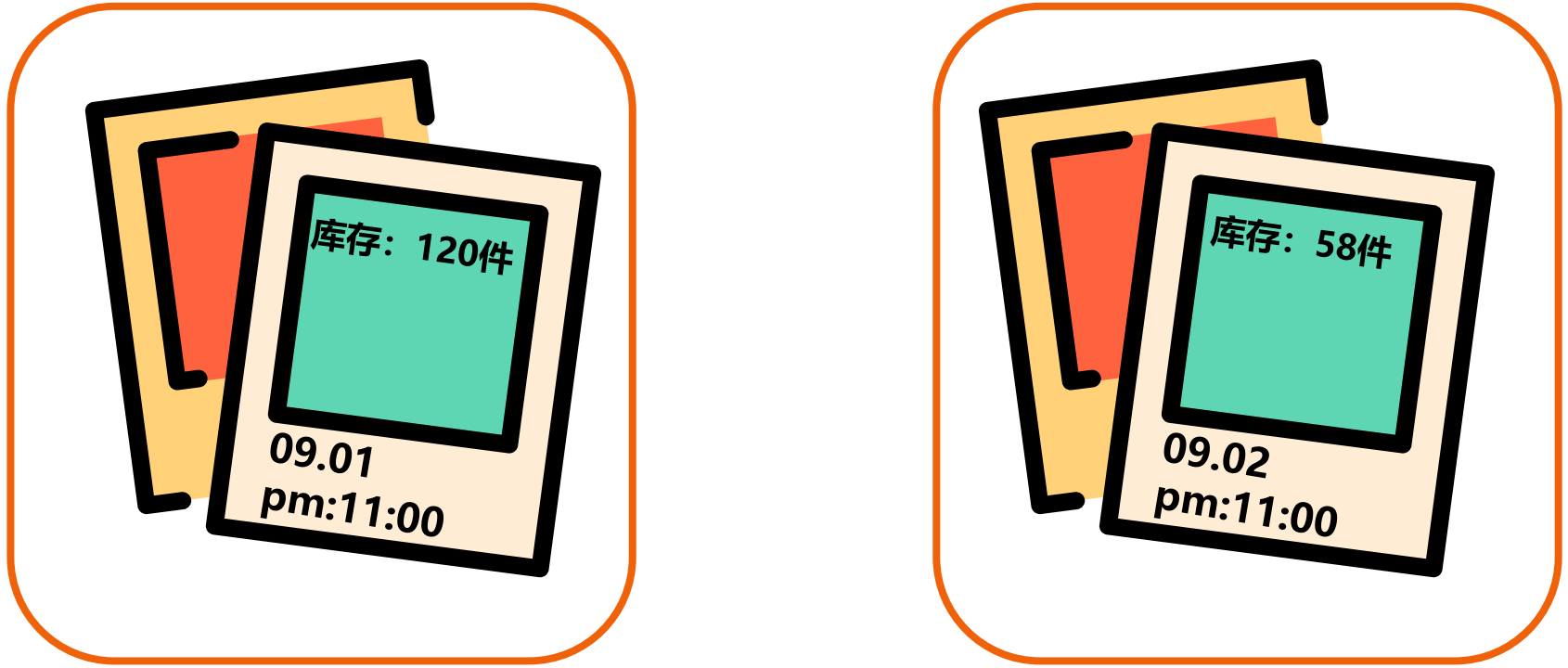




一个时间维度上保存连续数据

图　Hadoop生态系统

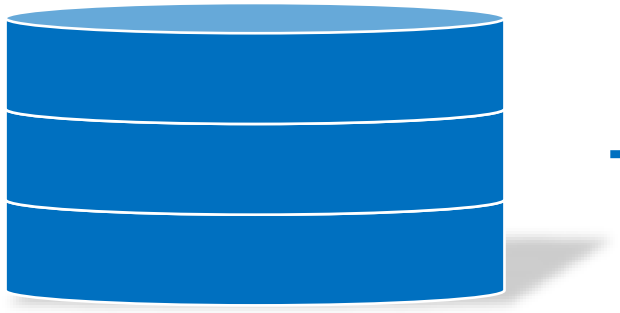
1.6.1 代表性的大数据技术之Hadoop



数据库只能保存某一时刻的状态

1.6.1 代表性的大数据技术之Hadoop

反映时间维度的信息的数据

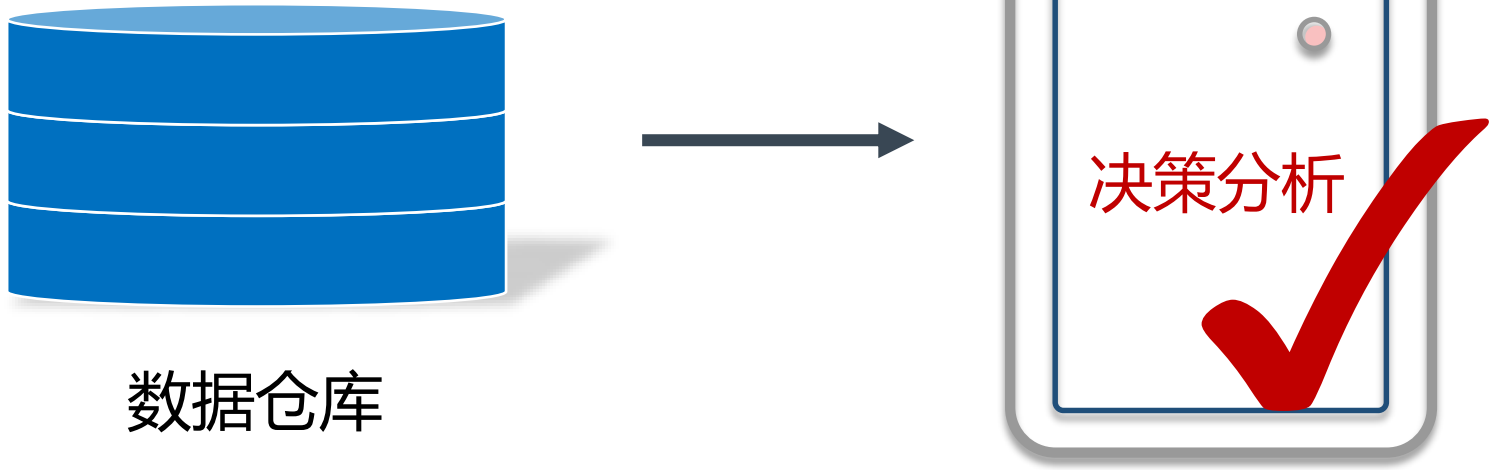


数据仓库

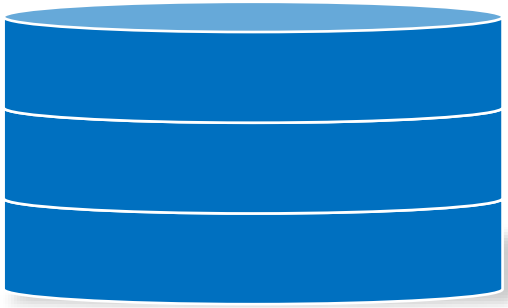
可以做什么呢?



1.6.1 代表性的大数据技术之Hadoop



1.6.1 代表性的大数据技术之Hadoop



数据仓库



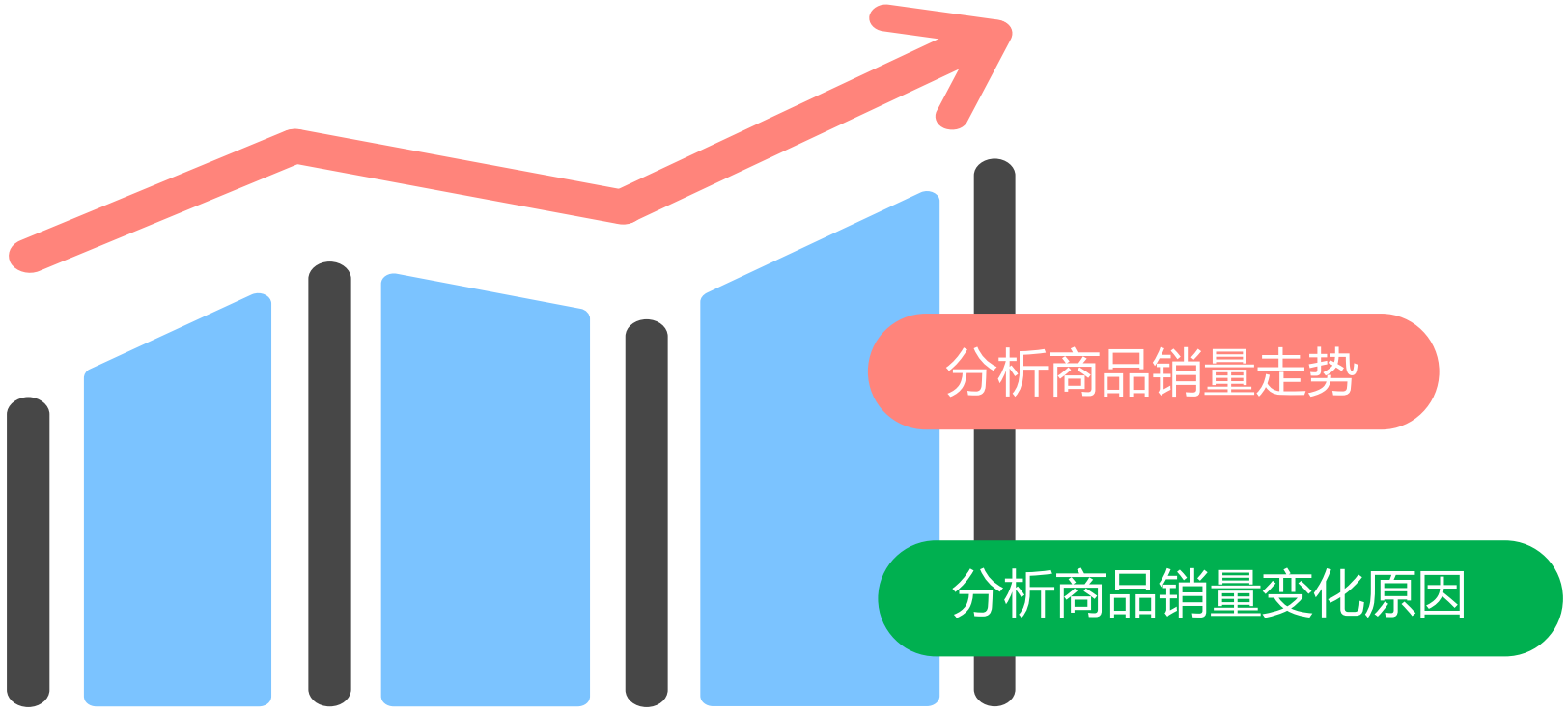


- 进行多维数据分析



1.6.1 代表性的大数据技术之Hadoop

数据仓库

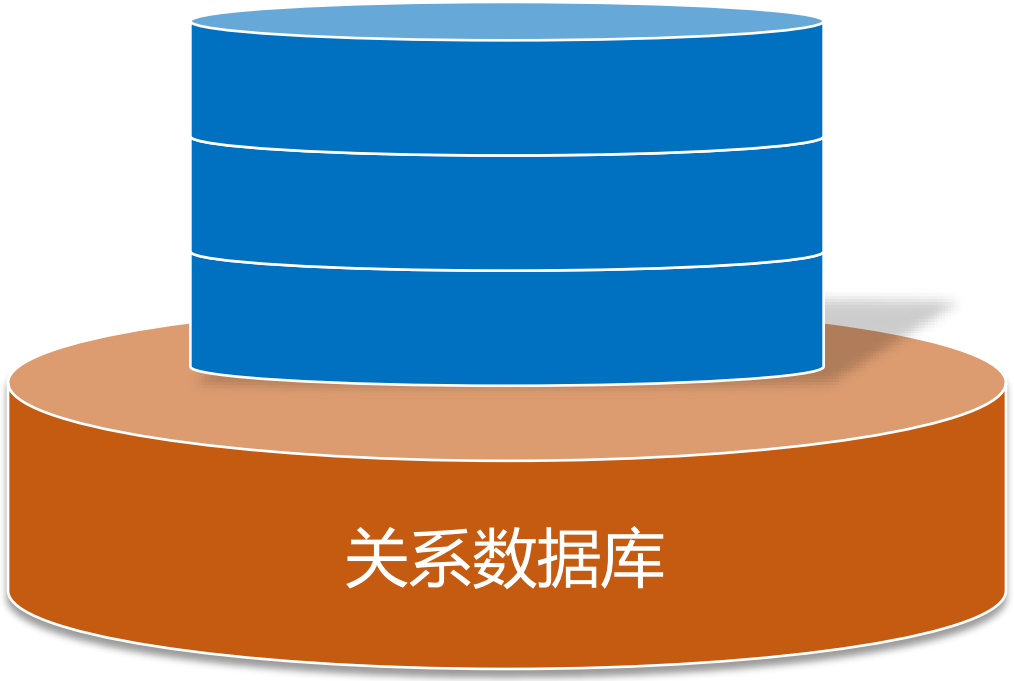


商品销量

1.6.1 代表性的大数据技术之Hadoop



传统数据仓库



1.6.1 代表性的大数据技术之Hadoop





1.6.1 代表性的大数据技术之Hadoop

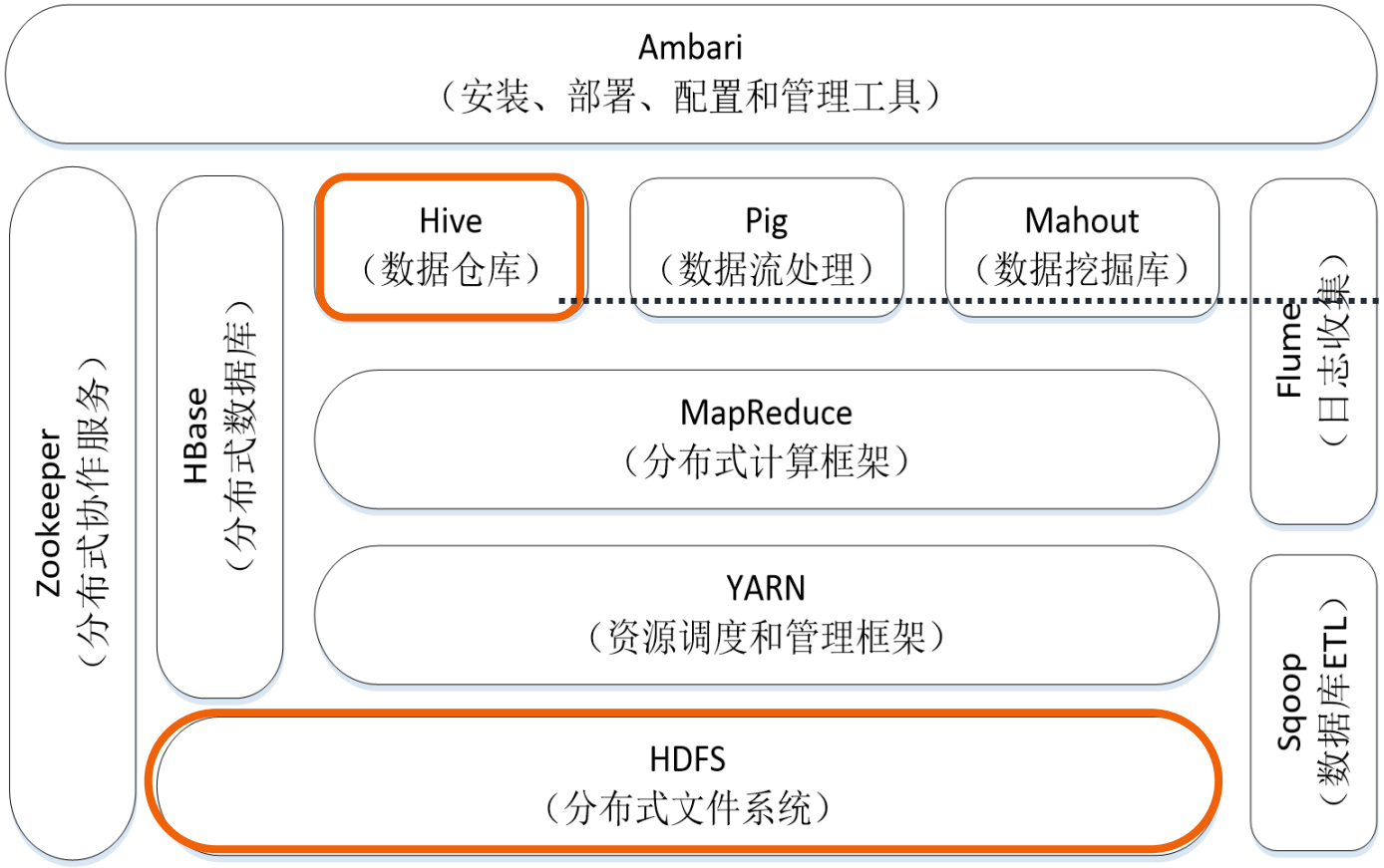


图　Hadoop生态系统

- 查询时写的是SQL语句

1.6.1 代表性的大数据技术之Hadoop



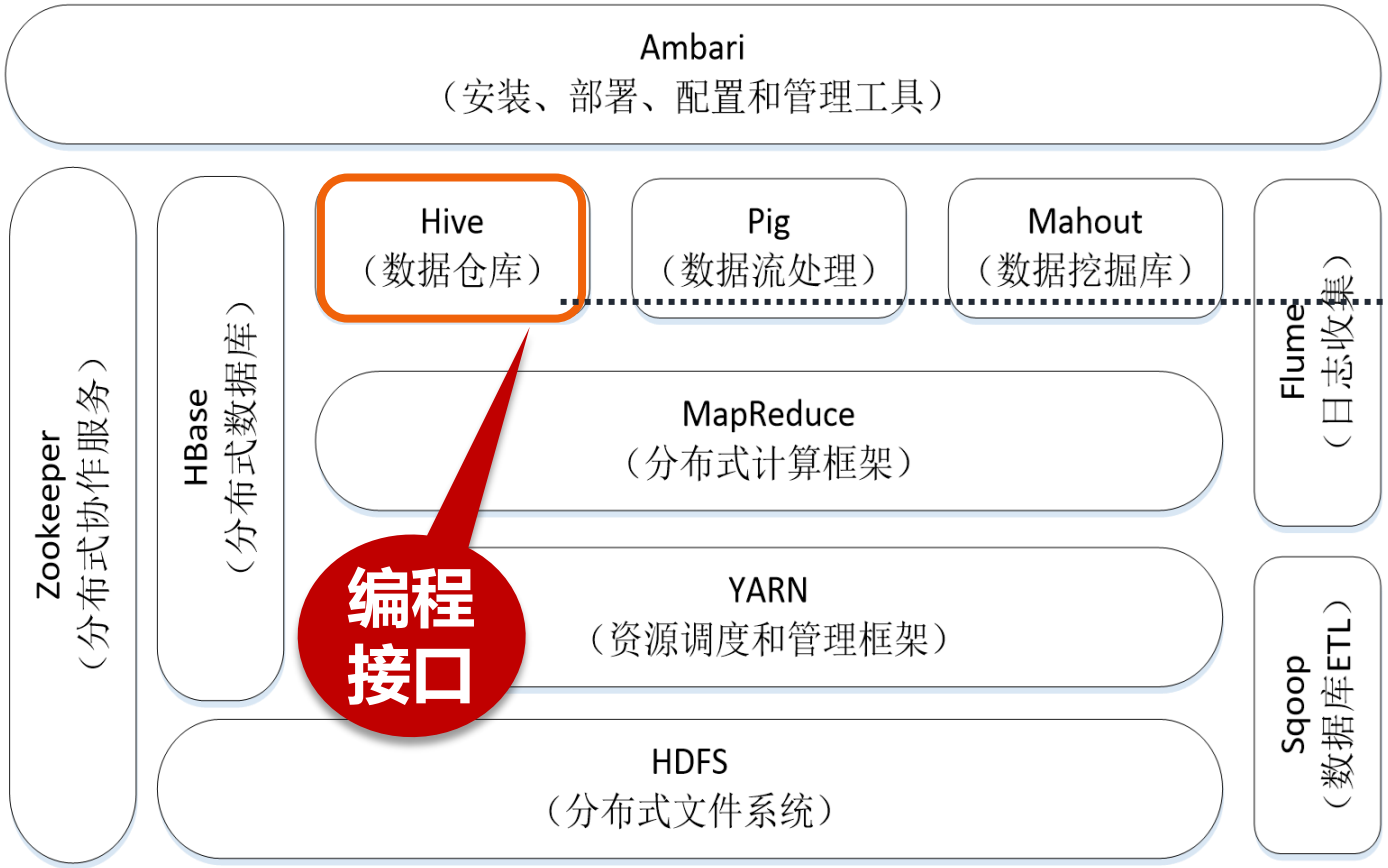


图　Hadoop生态系统

- 将SQL语句自动转换对HDFS的查询分析　得到结果



1.6.1 代表性的大数据技术之Hadoop



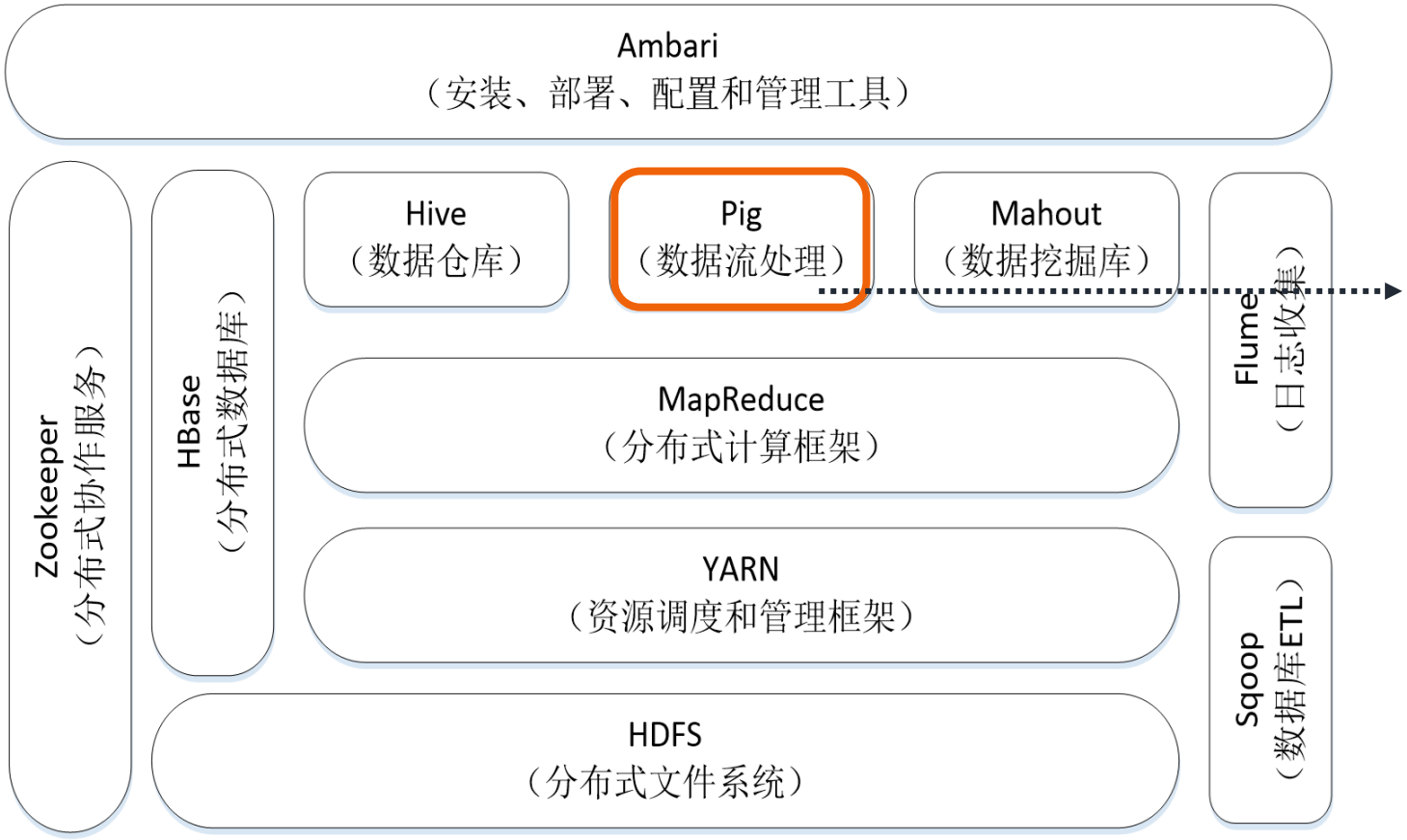


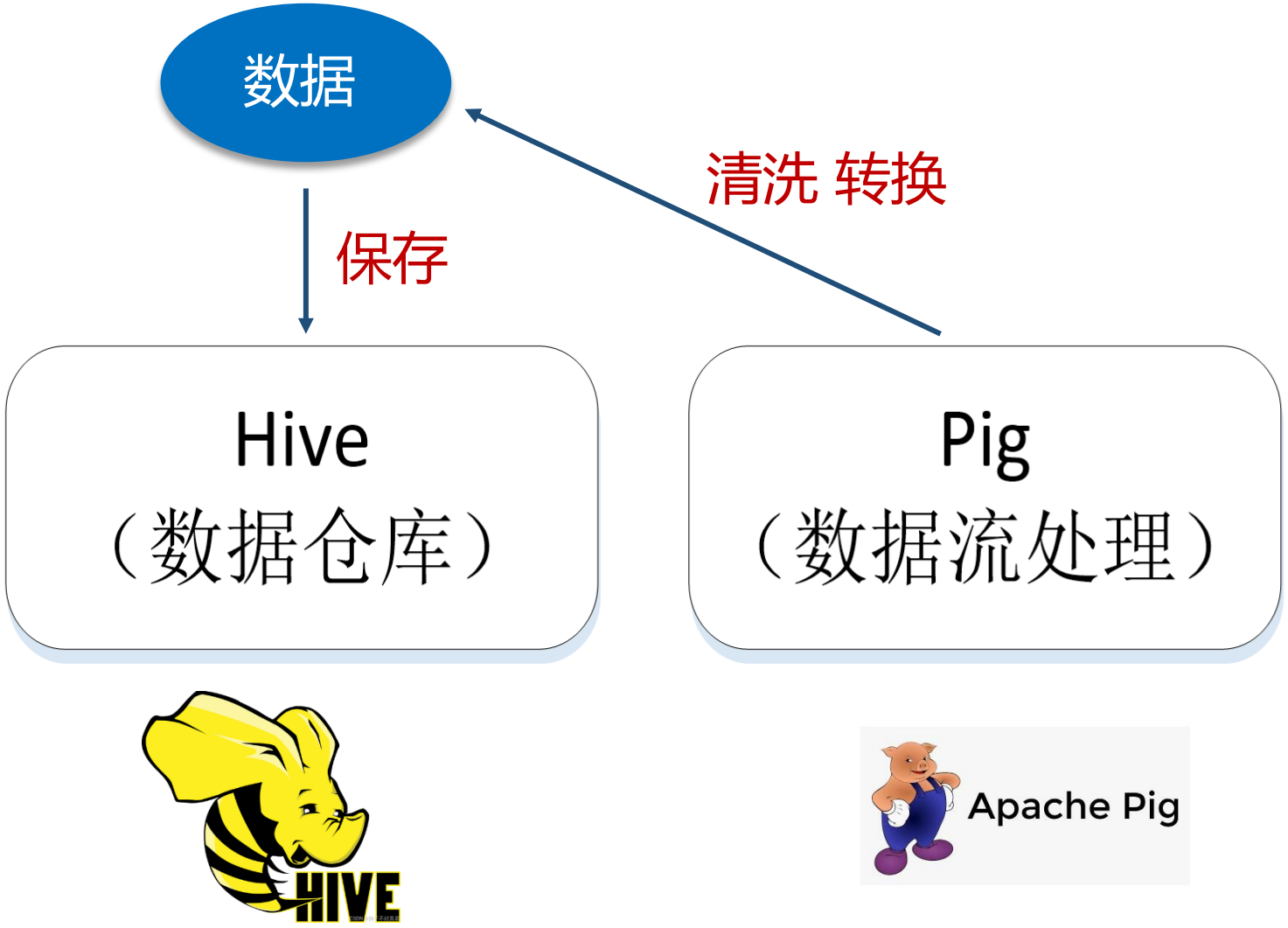
图　Hadoop生态系统

- Pig Latin语言
- 数据清洗 转换 加载



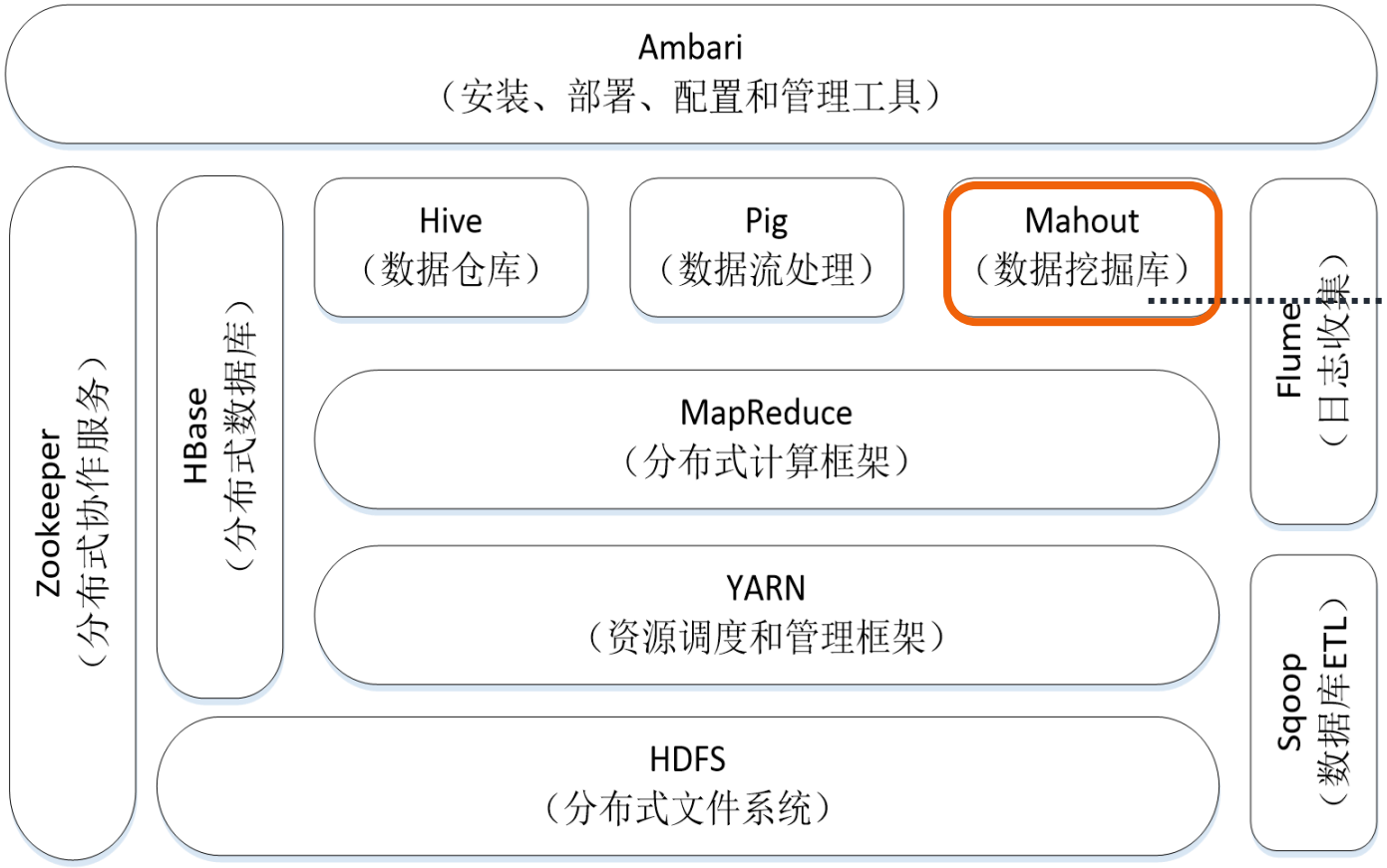
1.6.1 代表性的大数据技术之Hadoop





1.6.1 代表性的大数据技术之Hadoop





- 实现常用数据挖掘算法
  分类 聚类 回归等



图　Hadoop生态系统

1.6.1 代表性的大数据技术之Hadoop

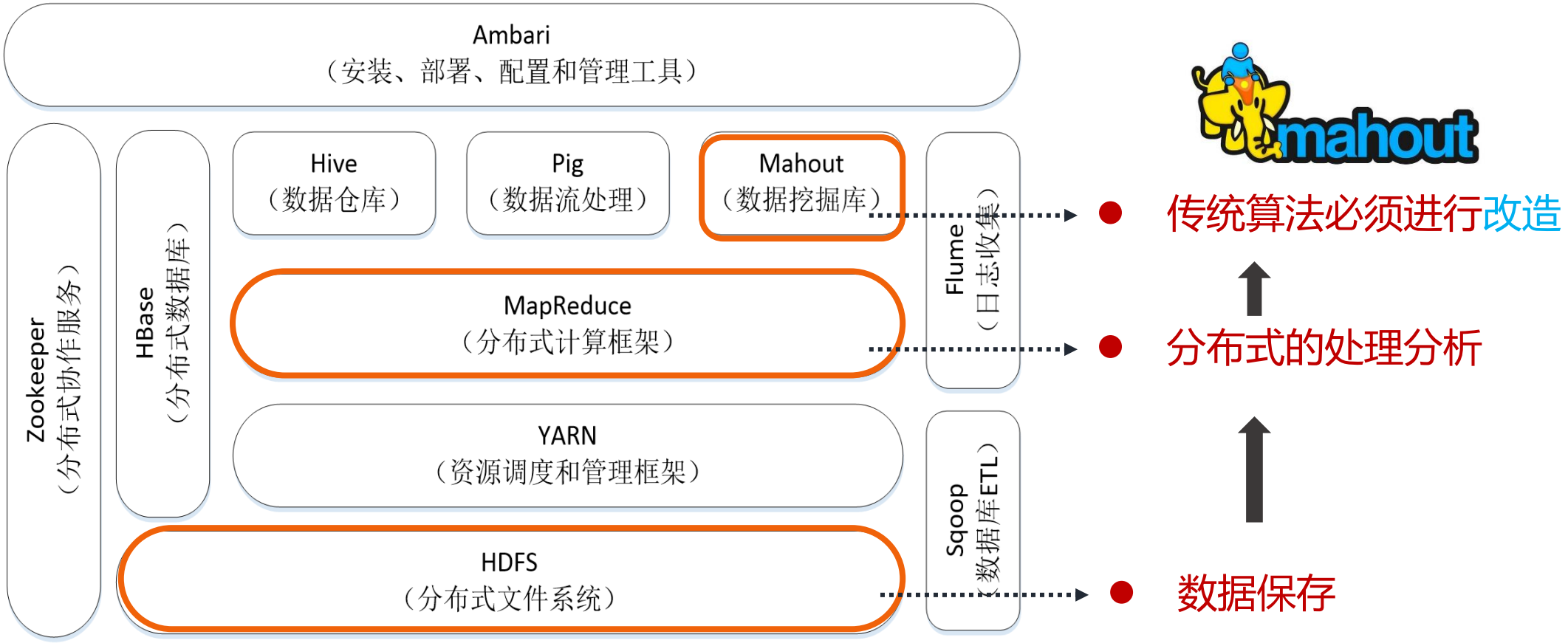
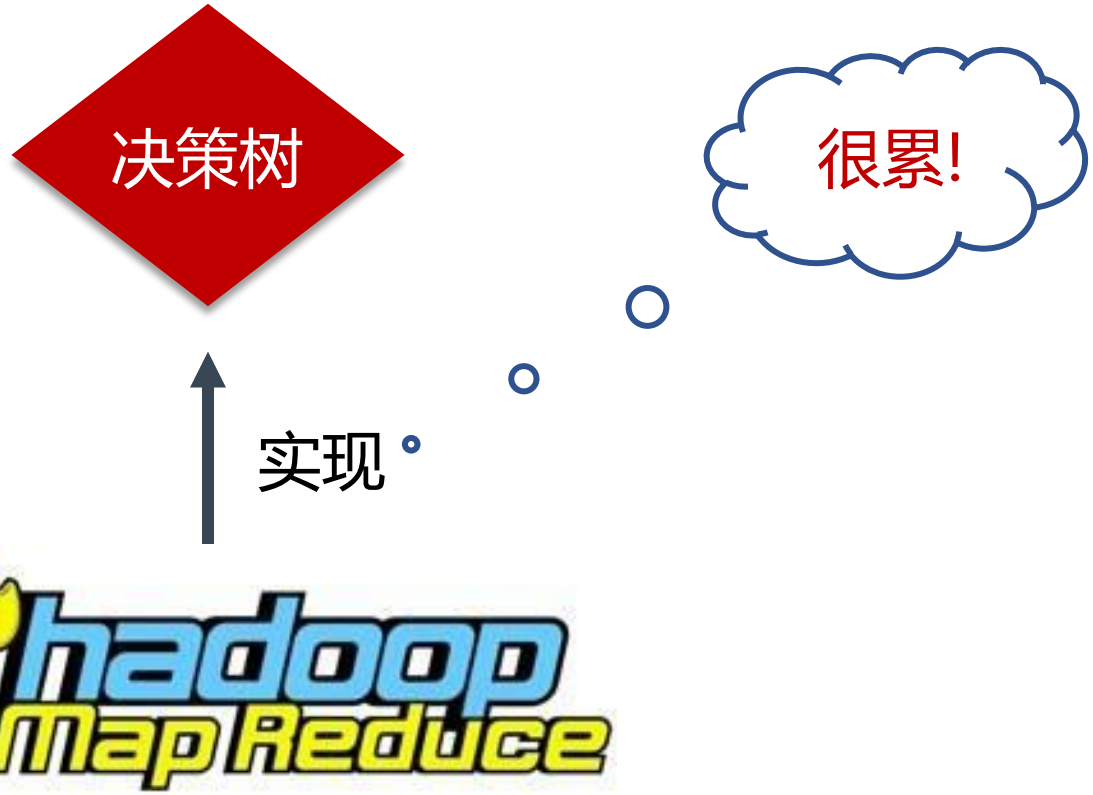


图　Hadoop生态系统

1.6.1 代表性的大数据技术之Hadoop





1.6.1 代表性的大数据技术之Hadoop



Mahout团队

常用的数据挖掘算法

制作

用MapReduce实现的算法库

1.6.1 代表性的大数据技术之Hadoop



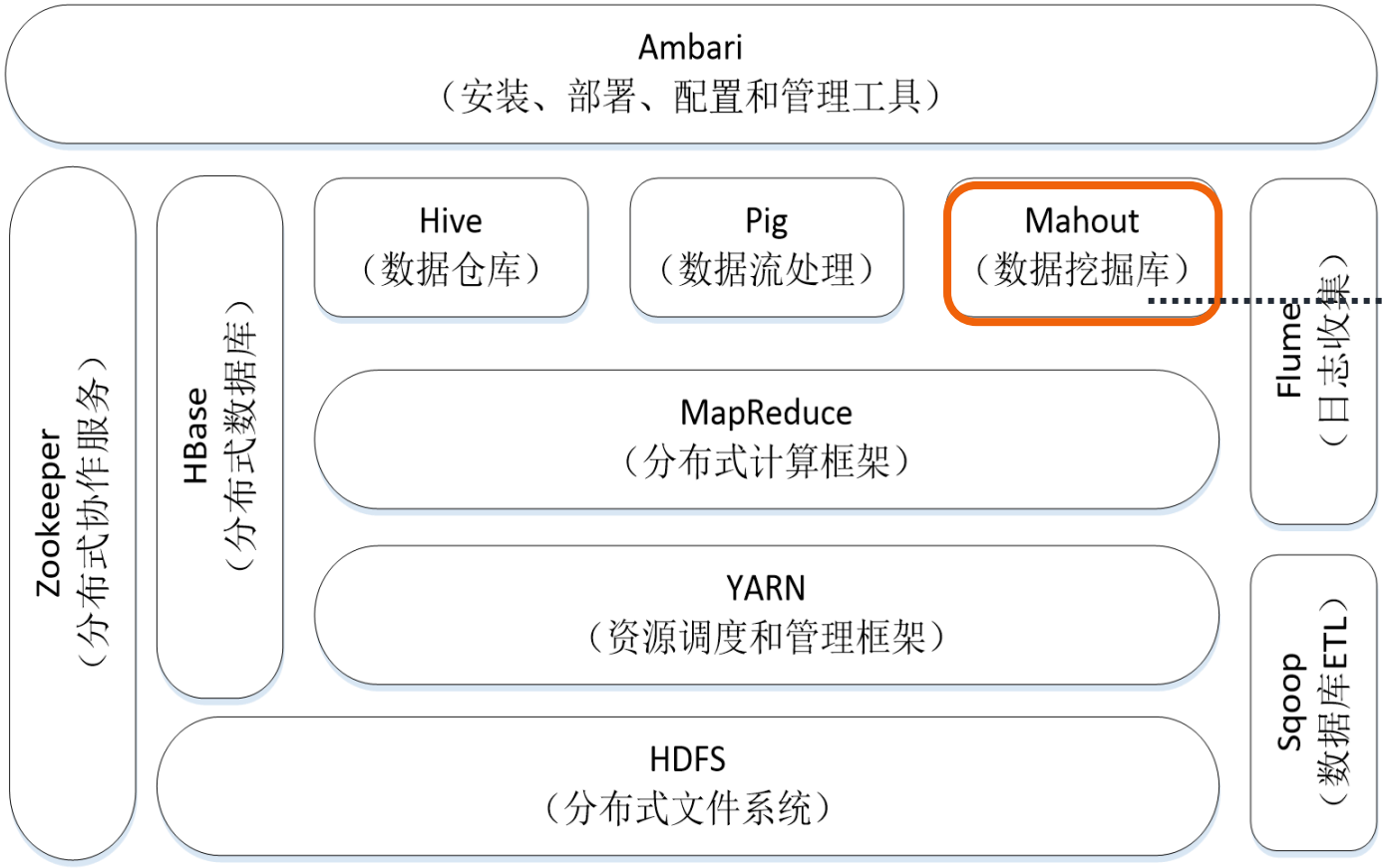


图　Hadoop生态系统

- 调接口 传参数 减少工作量
- 针对海量数据进行数据挖掘分析



1.6.1 代表性的大数据技术之Hadoop



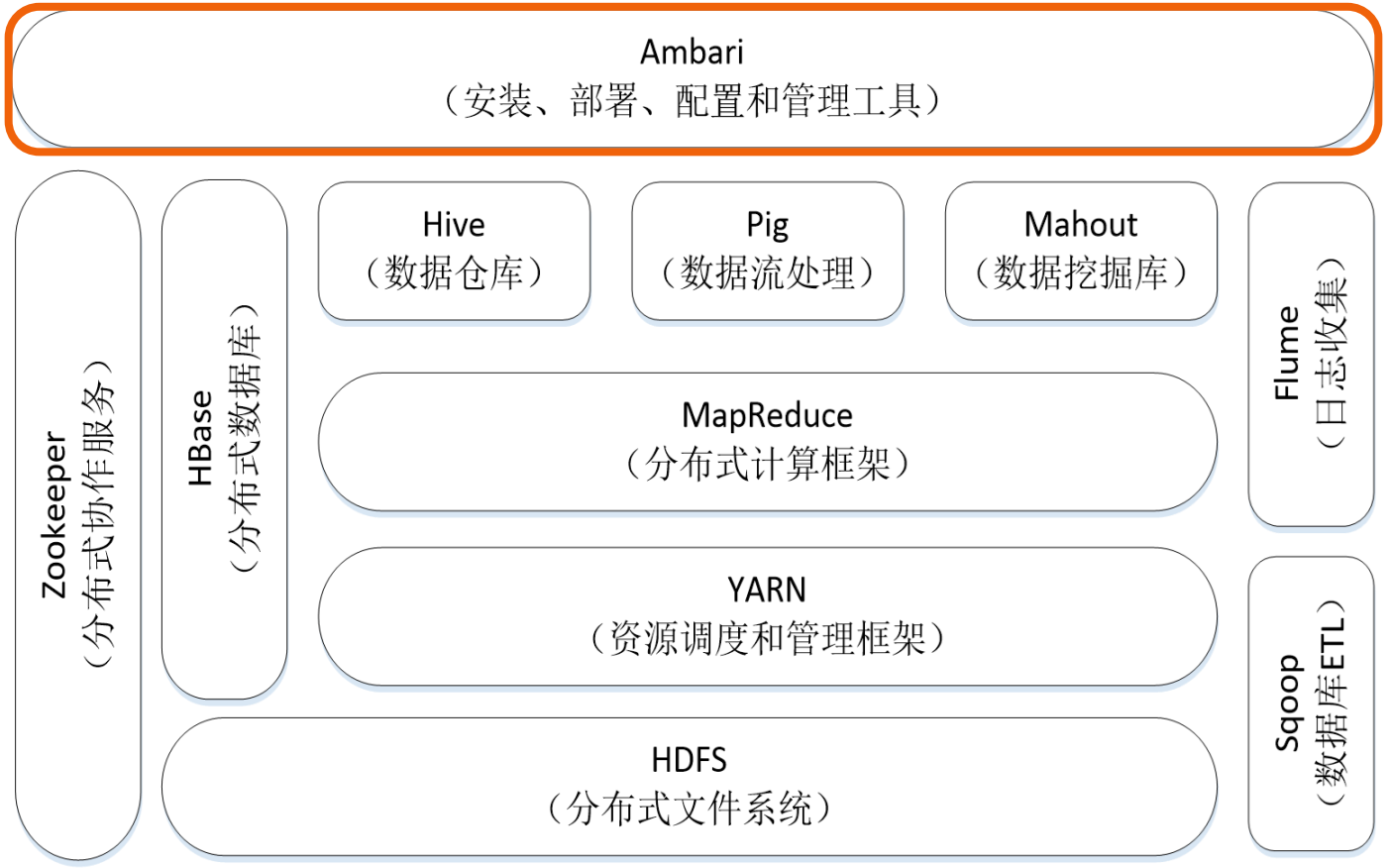
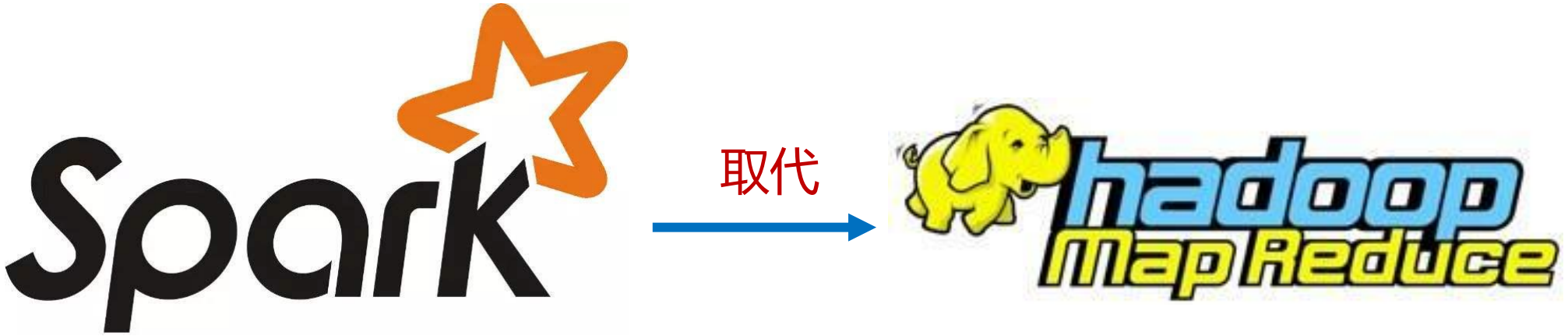


图　Hadoop生态系统

1.6.1 代表性的大数据技术之Hadoop





Spark2015年开始逐渐地取代MapReduce

1.6.1 代表性的大数据技术之Hadoop



Mahout团队

两年前开始倒戈

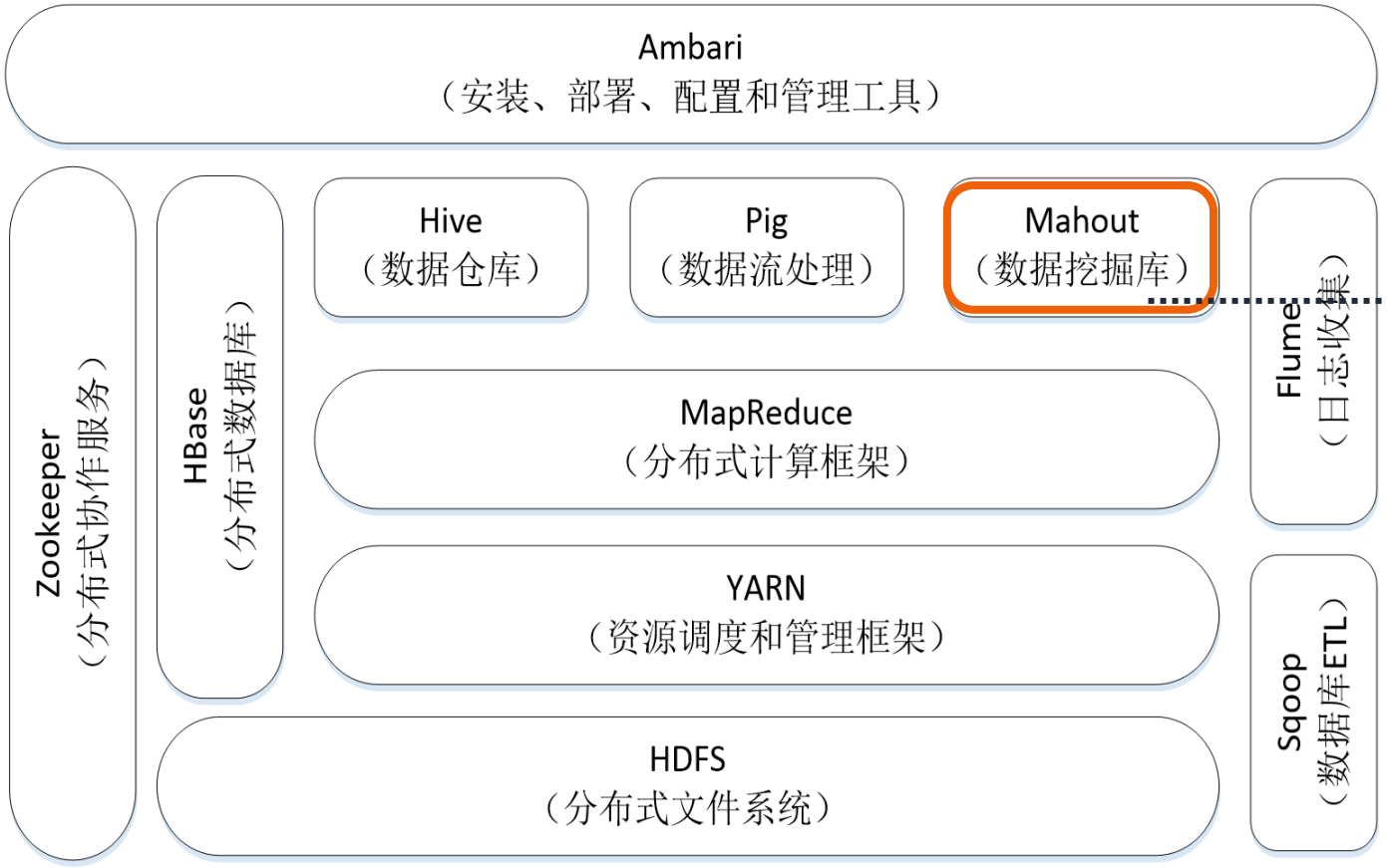
1.6.1 代表性的大数据技术之Hadoop

Mahout



不再支持MapReduce后续更新

1.6.1 代表性的大数据技术之Hadoop

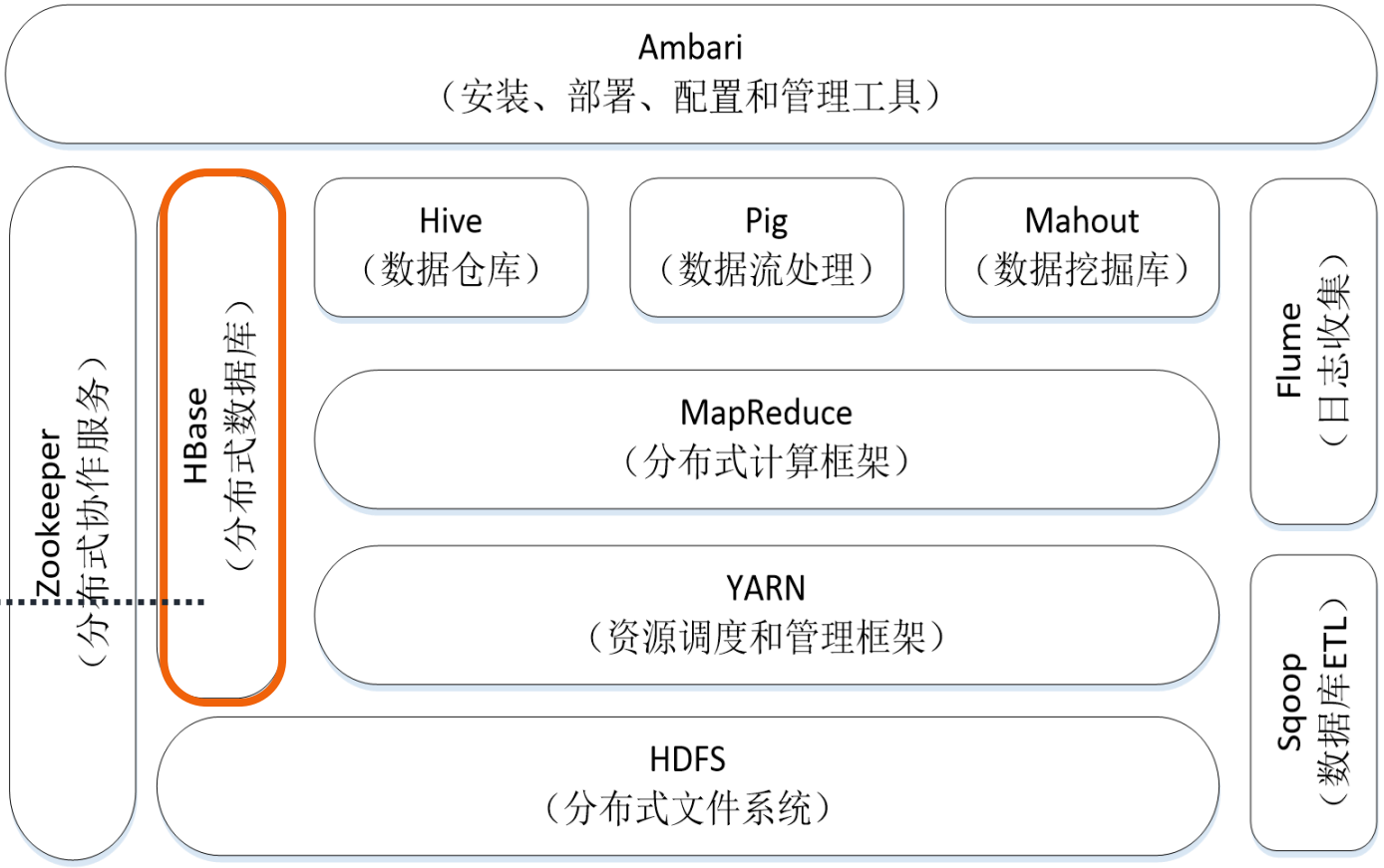


- 全面转向Spark

图　Hadoop生态系统

1.6.1 代表性的大数据技术之Hadoop





一主多从架构

图　Hadoop生态系统

1.6.1 代表性的大数据技术之Hadoop



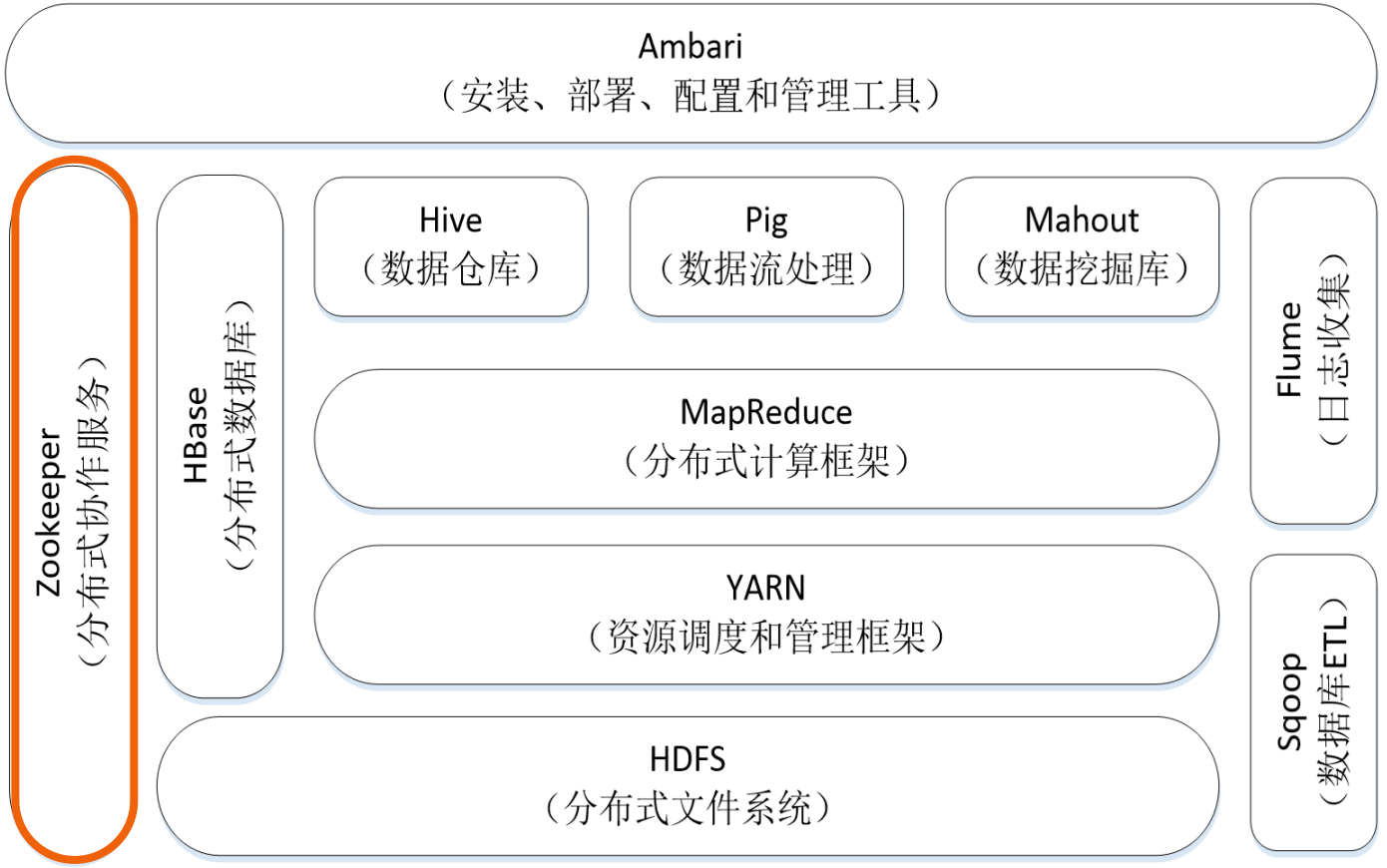


图　Hadoop生态系统

1.6.1 代表性的大数据技术之Hadoop

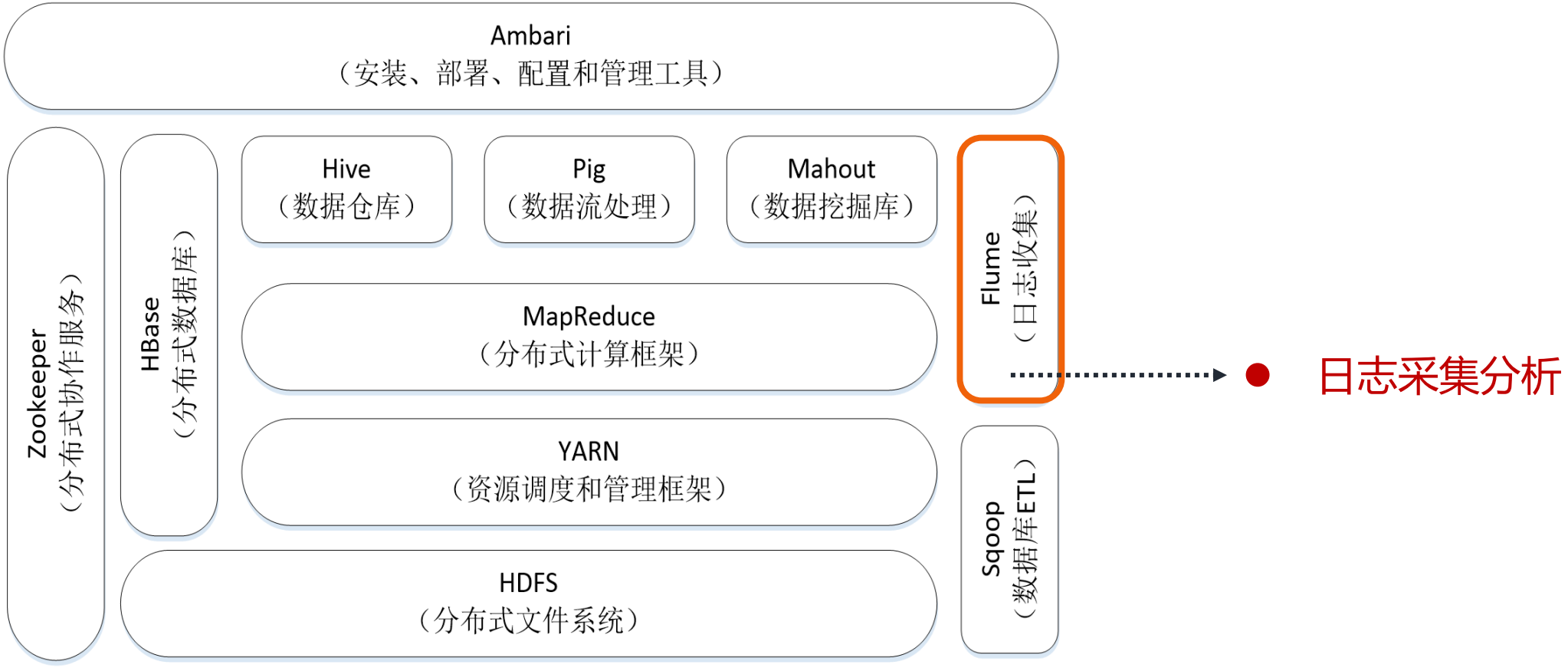
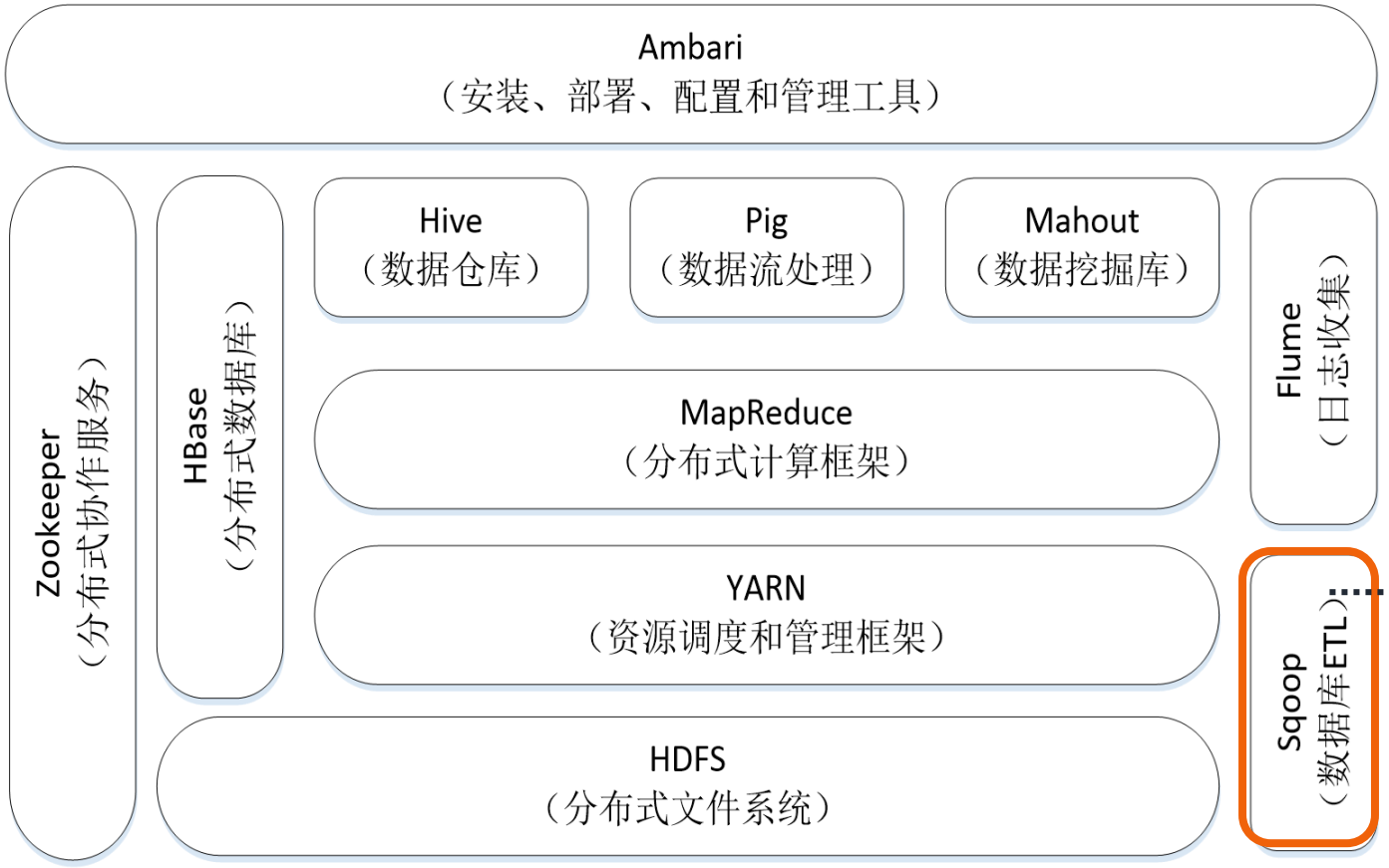


图　Hadoop生态系统

1.6.1 代表性的大数据技术之Hadoop

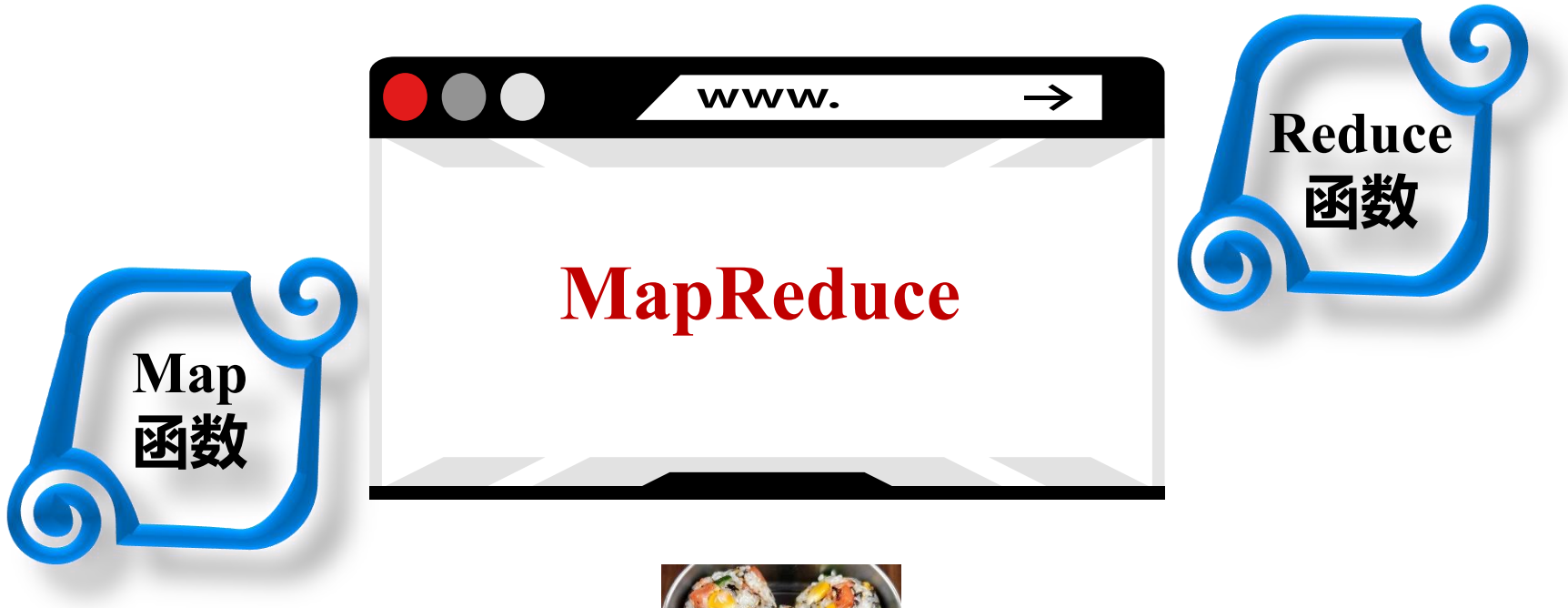


Apache Sqoop

- 完成Hadoop系统组件之间的互通

图　Hadoop生态系统

1.6.1 代表性的大数据技术之Hadoop





1.6.1 代表性的大数据技术之Hadoop



编程容易
屏蔽底层分布式并行编程细节

1.6.1 代表性的大数据技术之Hadoop





采用“分而治之”策略

1.6.1 代表性的大数据技术之Hadoop



分而治之

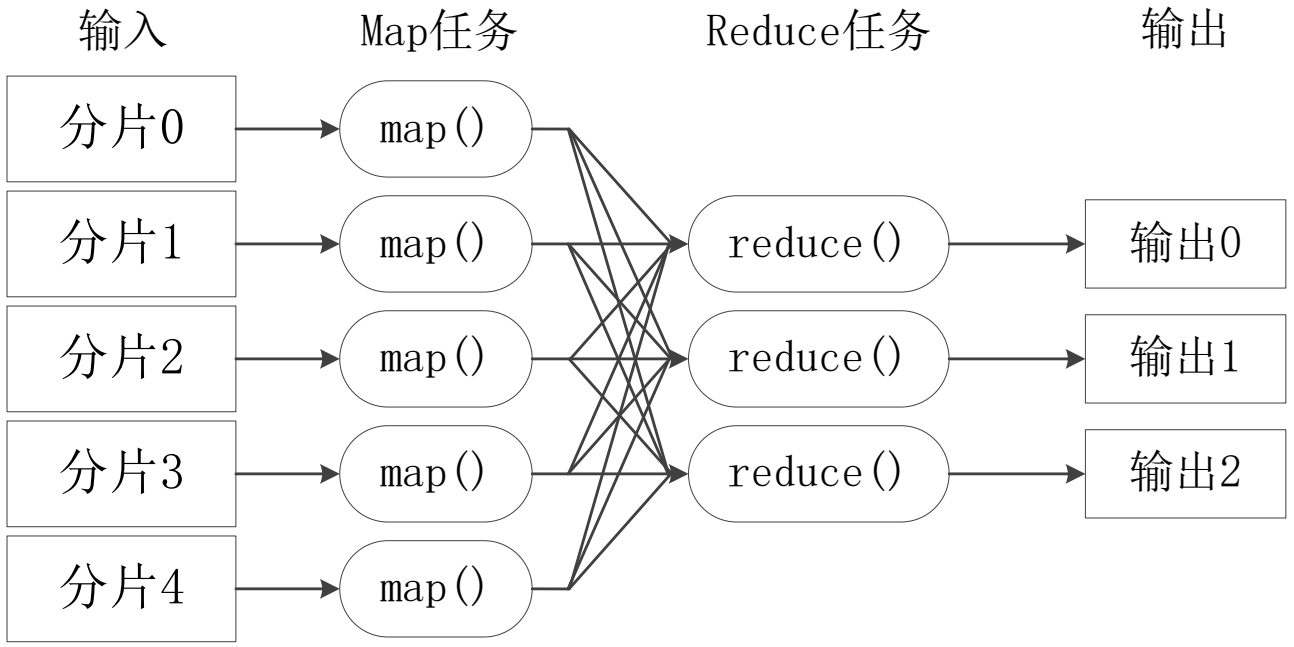
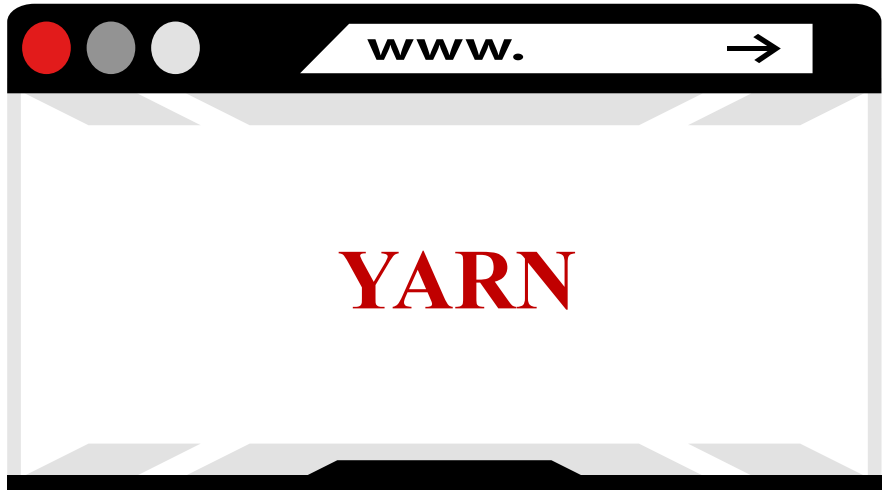
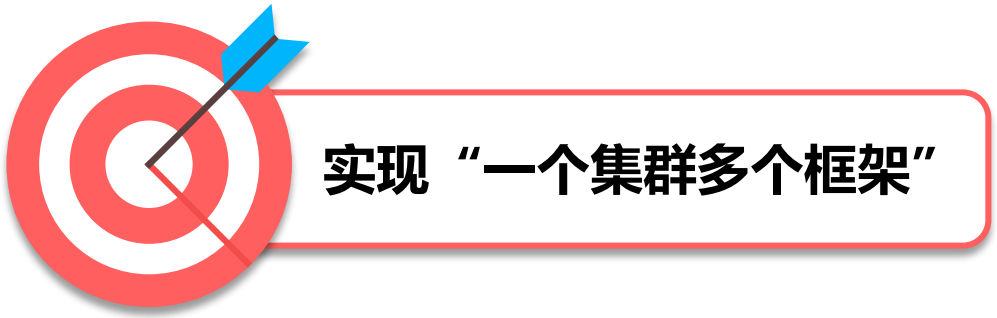


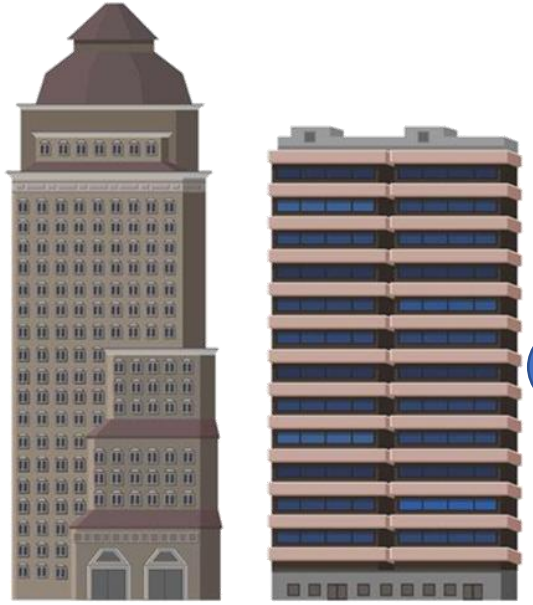
图　MapReduce工作流程

1.6.1 代表性的大数据技术之Hadoop





1.6.1 代表性的大数据技术之Hadoop



企业

业务

应用场景1

应用场景2

……

计算框架1

计算框架2

……

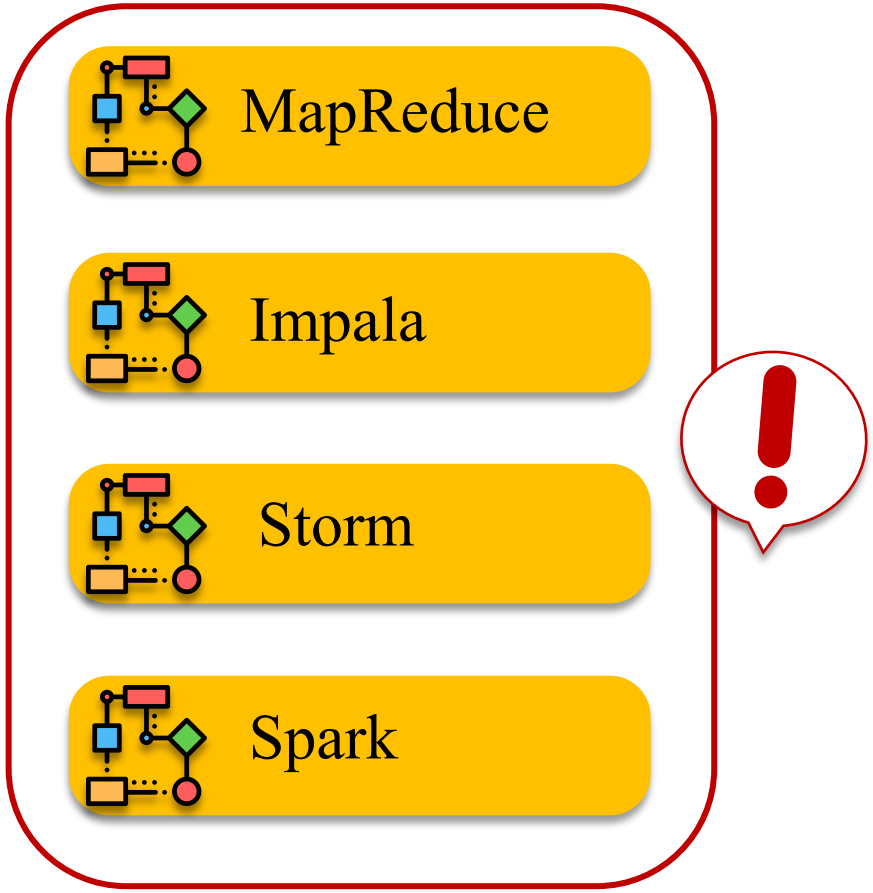
1.6.1 代表性的大数据技术之Hadoop



离线批处理	→	MapReduce
实时交互式查询分析	→	Impala
流式数据实时分析	→	Storm
迭代计算	→	Spark

1.6.1 代表性的大数据技术之Hadoop



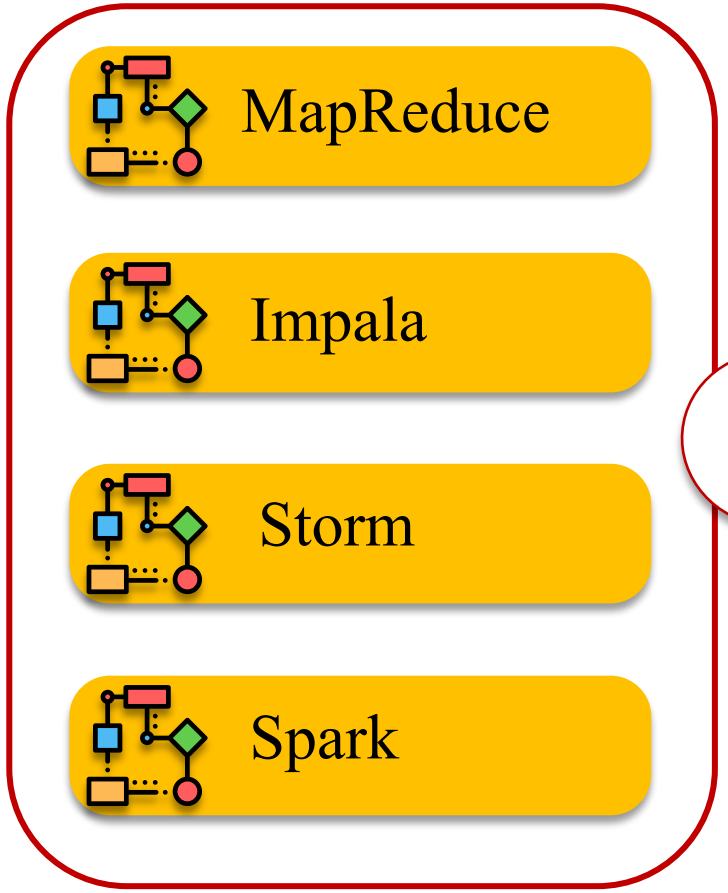




不同的开发团队

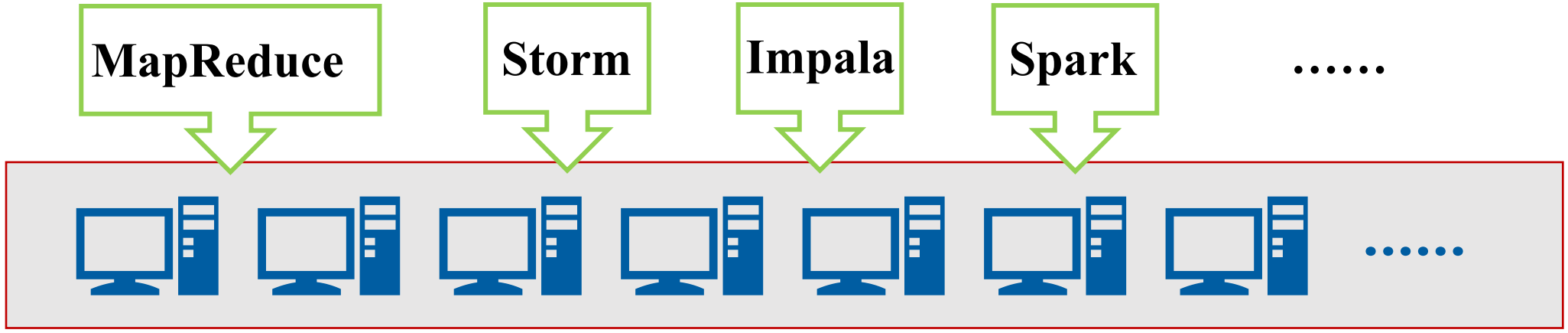
1.6.1 代表性的大数据技术之Hadoop





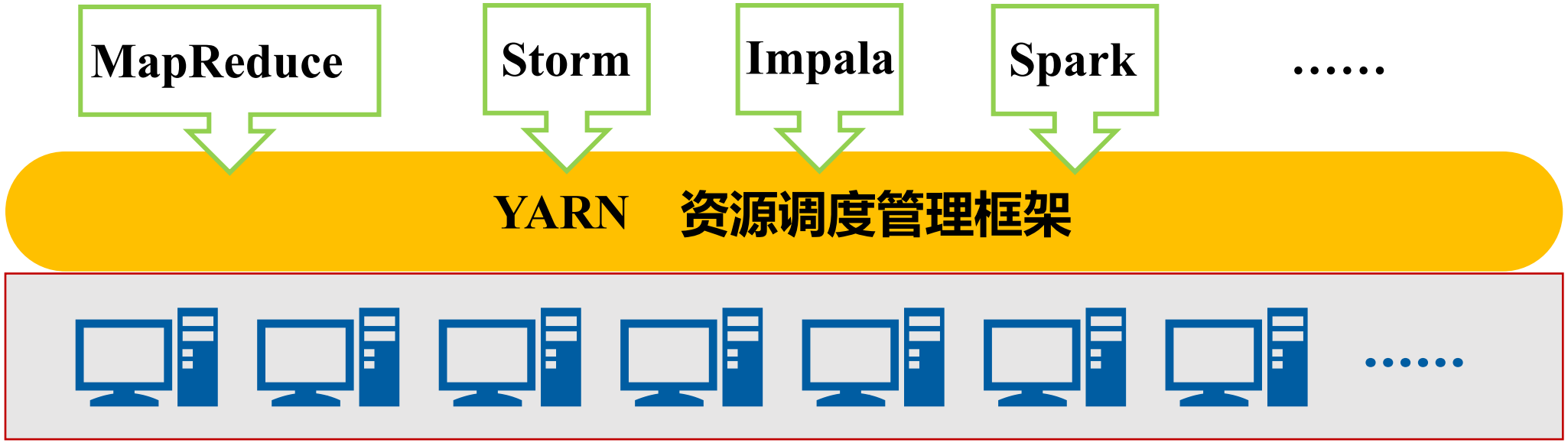
- 开发成本高
- 集群资源利用率低

1.6.1 代表性的大数据技术之Hadoop



1.6.1 代表性的大数据技术之Hadoop





1.6.1 代表性的大数据技术之Hadoop



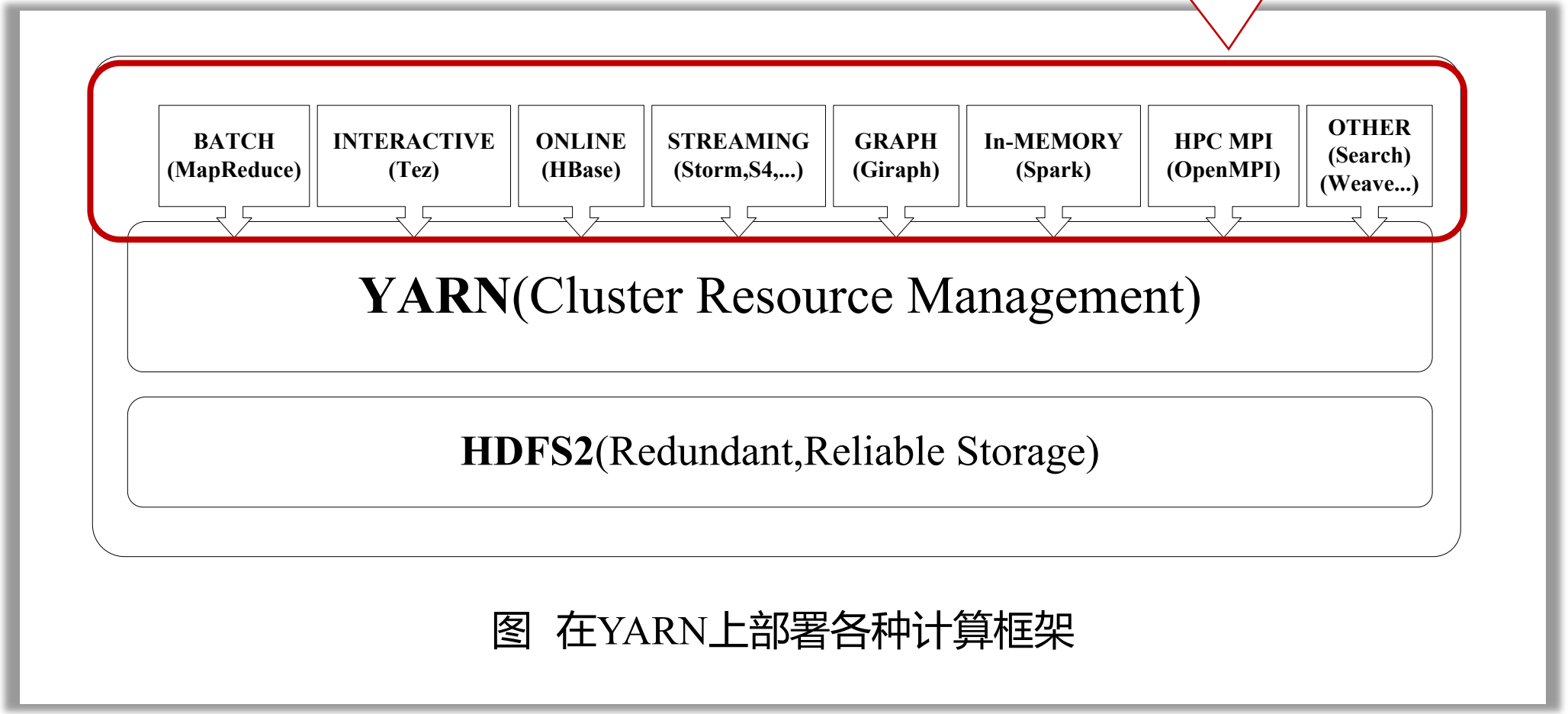


图 在YARN上部署各种计算框架

1.6.1 代表性的大数据技术之Hadoop

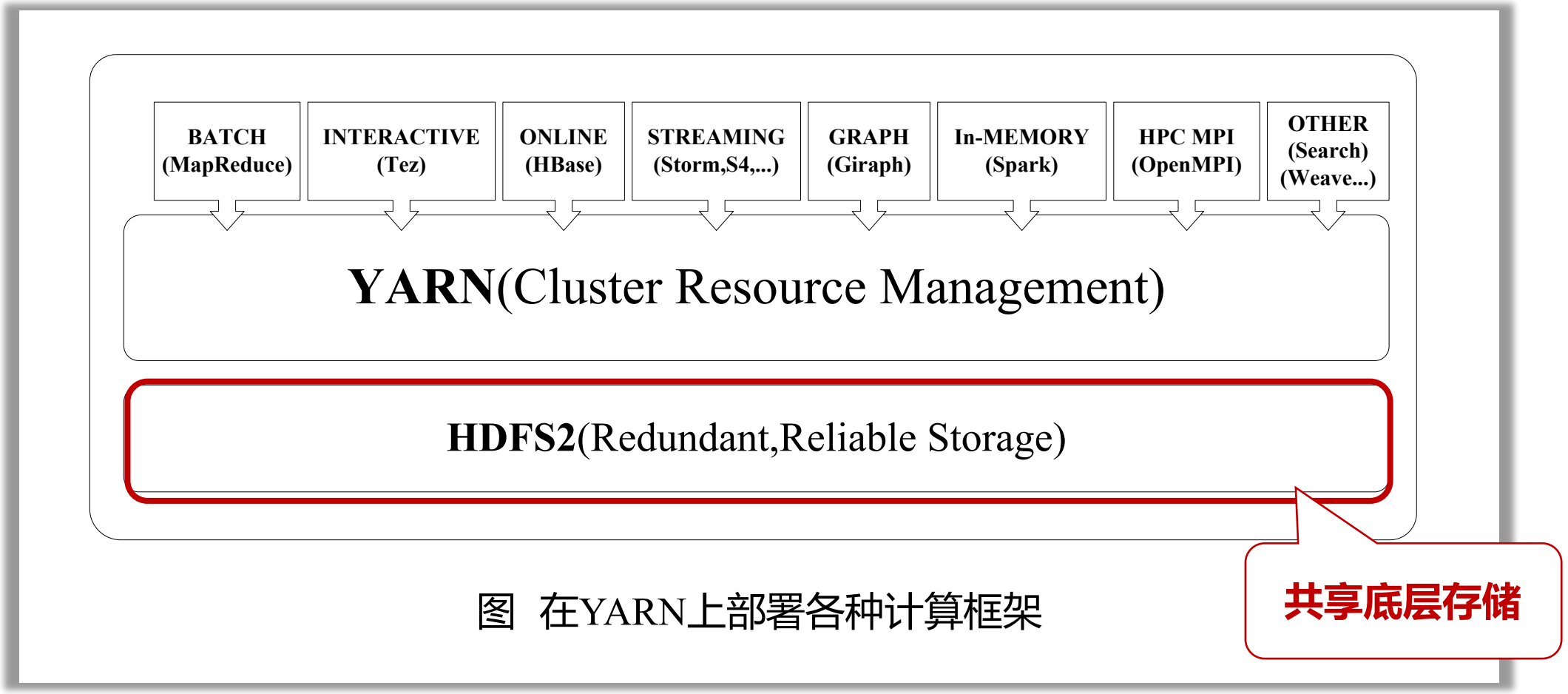


图 在YARN上部署各种计算框架

1.6.1 Hadoop理解测验（选择题）



第1题

以下关于 Hadoop 的说法，**错误**的是？

A. Hadoop 是一个大数据处理的生态系统

B. Hadoop 包含 HDFS、MapReduce、YARN 等核心组件

C. Hadoop 只能运行在单台机器上

D. Hadoop 支持分布式存储与计算

1.6.1 Hadoop理解测验（选择题）



第2题
HDFS 的主要功能是?

A. 图数据分析
B. 分布式任务调度
C. 海量数据的分布式存储
D. 实时数据查询

1.6.1 Hadoop理解测验（选择题）



第3题

下列哪个组件用于 Hadoop 的资源管理与任务调度？

A. Hive

B. Mahout

C. HDFS

D. YARN

1.6.1 Hadoop理解测验（选择题）



第4题
Hive 的主要特点是？

A. 支持图计算
B. 使用 SQL 接口处理结构化数据
C. 替代 HDFS 进行数据存储
D. 专注于机器学习任务

1.6.1 Hadoop理解测验（选择题）



第5题
关于 Pig 的描述，正确的是？

A. Pig 是一种关系型数据库
B. Pig 使用 SQL 编程语言
C. Pig 适合工程师进行数据清洗和转换
D. Pig 用于图形用户界面分析

1.6.1 Hadoop理解测验（选择题）



第6题

Mahout 最初基于哪种计算框架开发？

A. Tez

B. Spark

C. Flink

D. MapReduce

1.6.1 Hadoop理解测验（选择题）



第7题
关于 MapReduce 的“分而治之”策略，下列描述正确的是?

A. 所有任务集中在一个节点处理
B. 拆分大任务为子任务，分发处理后汇总结果
C. 先汇总后处理
D. 使用图计算方式组织数据

1.6.1 Hadoop理解测验（选择题）



第8题

以下哪个不是运行在 YARN 上的框架？

A. Storm

B. Spark

C. MySQL

D. MapReduce

1.6.1 Hadoop理解测验（选择题）



第9题

为什么 Mahout 后来放弃 MapReduce 支持而转向 Spark？

A. Spark 只能用于存储，不支持计算
B. Spark 更适合迭代计算和内存处理，性能更高
C. Spark 无法运行在 YARN 上
D. MapReduce 支持更多算法

1.6.1 Hadoop理解测验（选择题）



第10题

关于 YARN 的功能，说法正确的是？

A. YARN 只能运行 MapReduce 程序
B. YARN 是 Hadoop 的存储引擎
C. YARN 可以统一管理多个计算框架的资源
D. YARN 是用于数据压缩的工具



Thank you