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

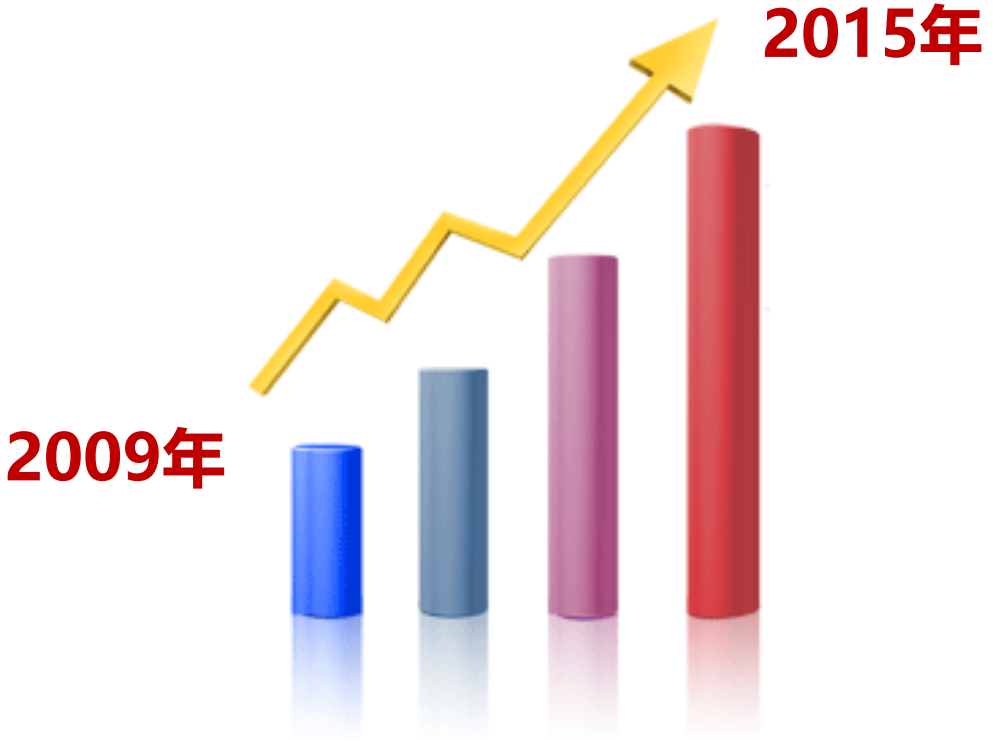
4 Beam



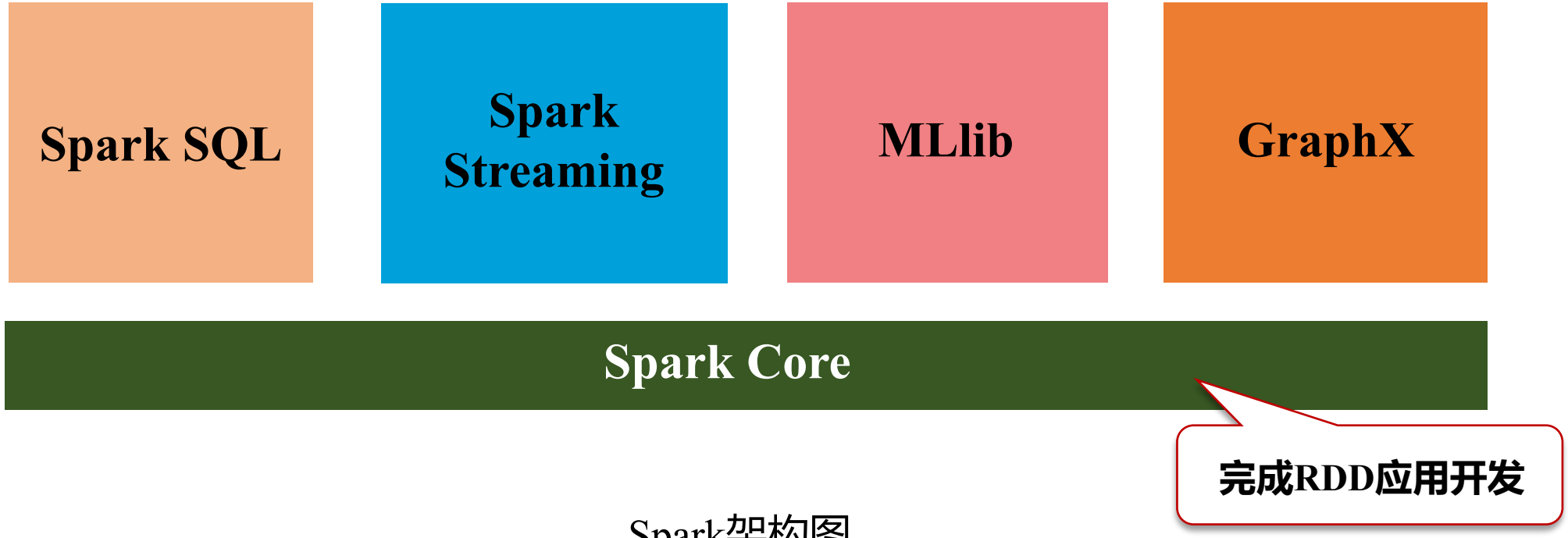
1.6.2 Spark

1.6.2 代表性大数据技术之Spark



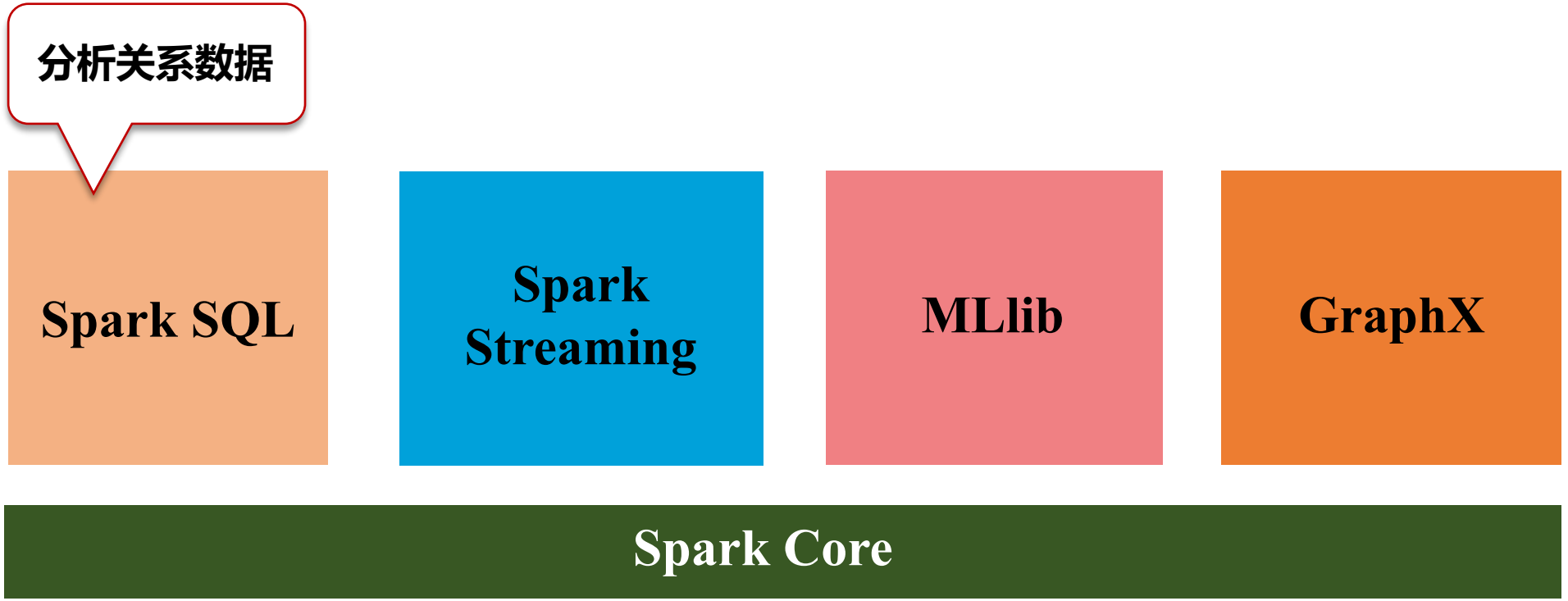


1.6.2 代表性大数据技术之Spark



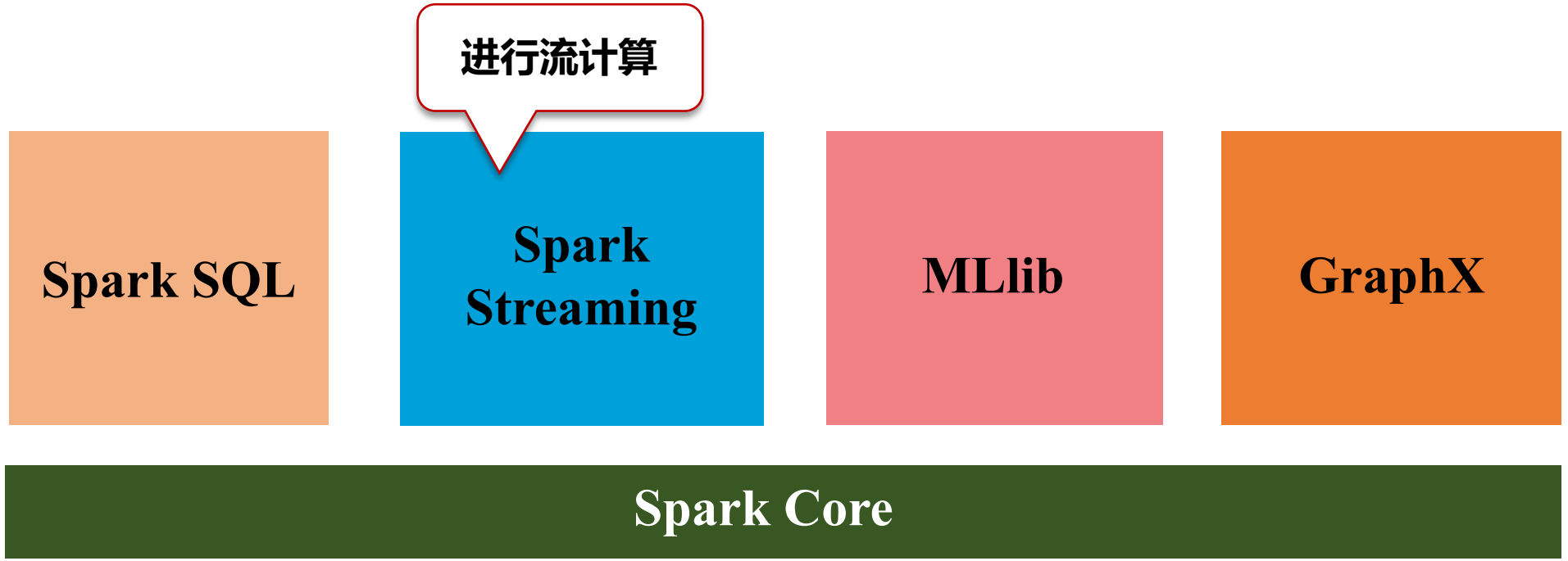
Spark架构图

1.6.2 代表性大数据技术之Spark



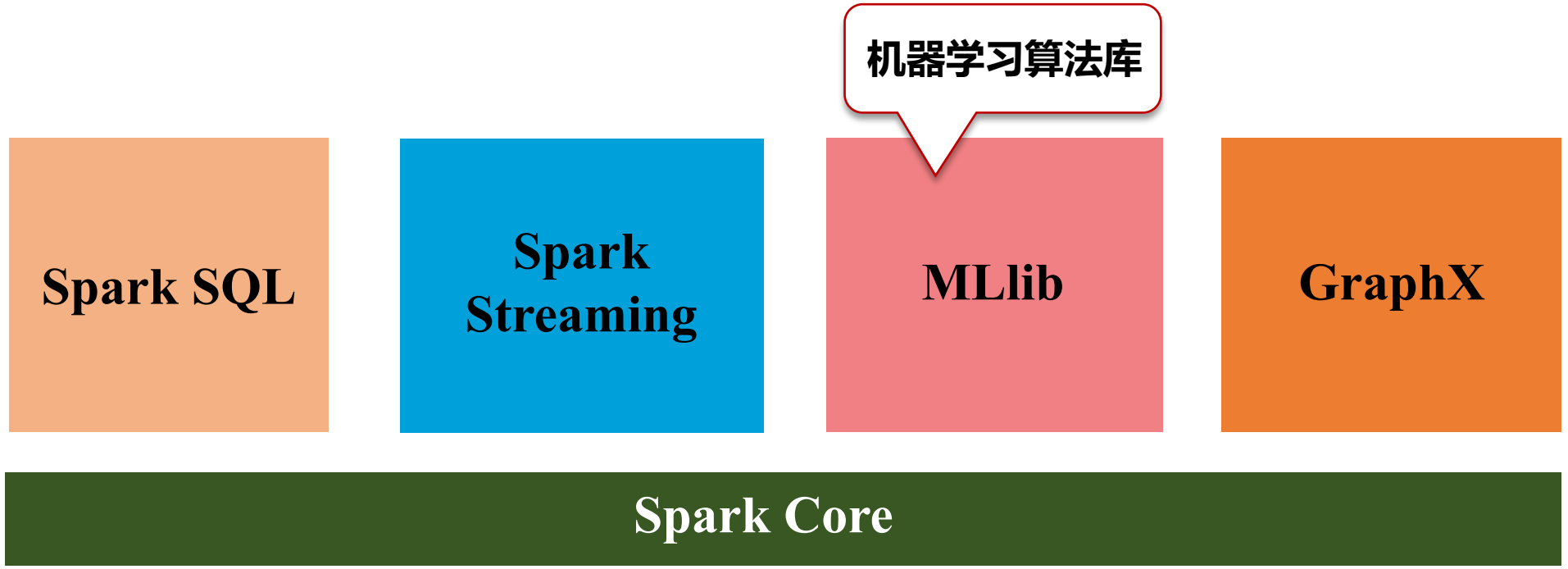
Spark架构图

1.6.2 代表性大数据技术之Spark



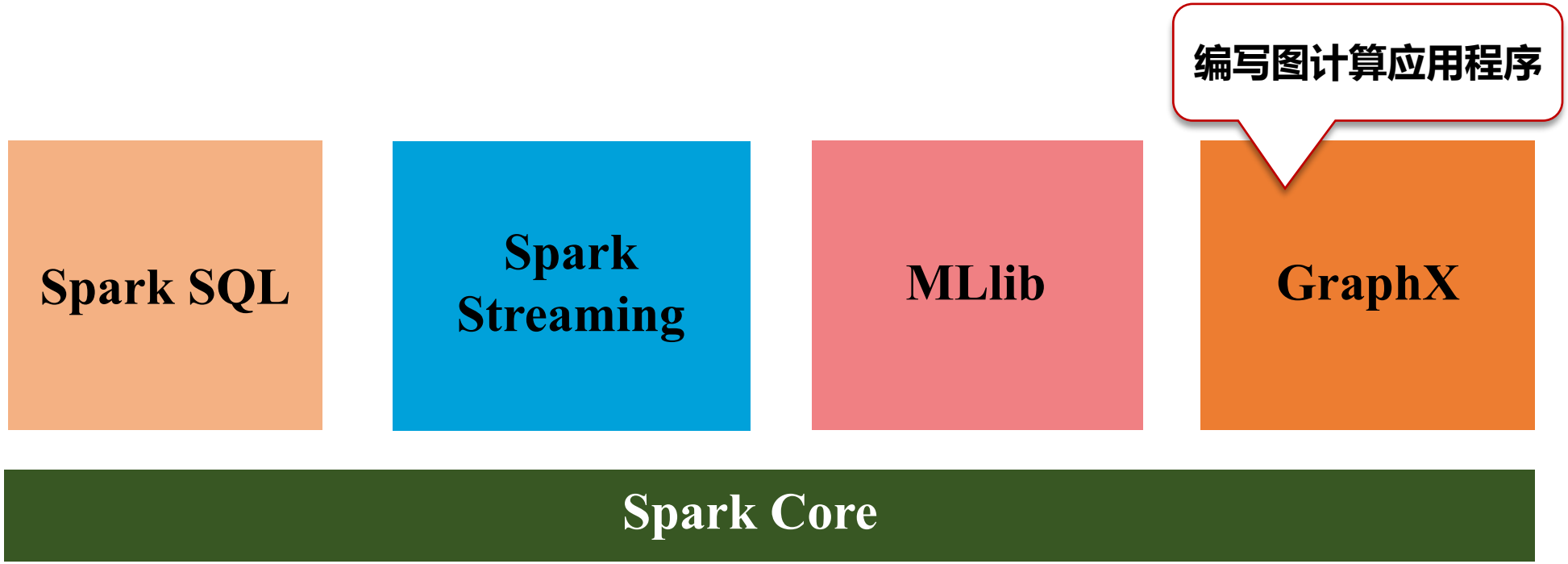
Spark架构图

1.6.2 代表性大数据技术之Spark



Spark架构图

1.6.2 代表性大数据技术之Spark

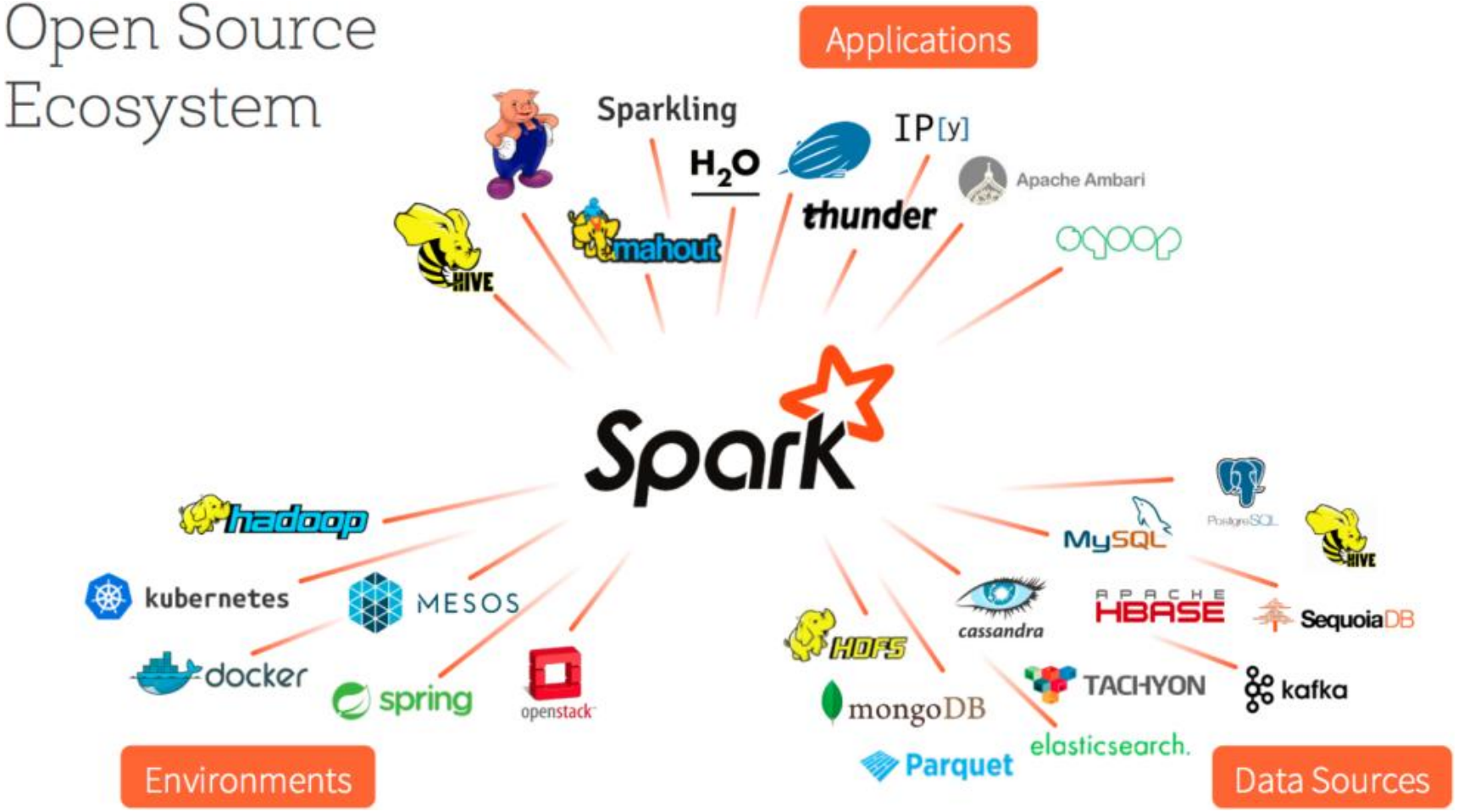


Spark架构图

1.6.2 代表性大数据技术之Spark



1.6.2 代表性大数据技术之Spark

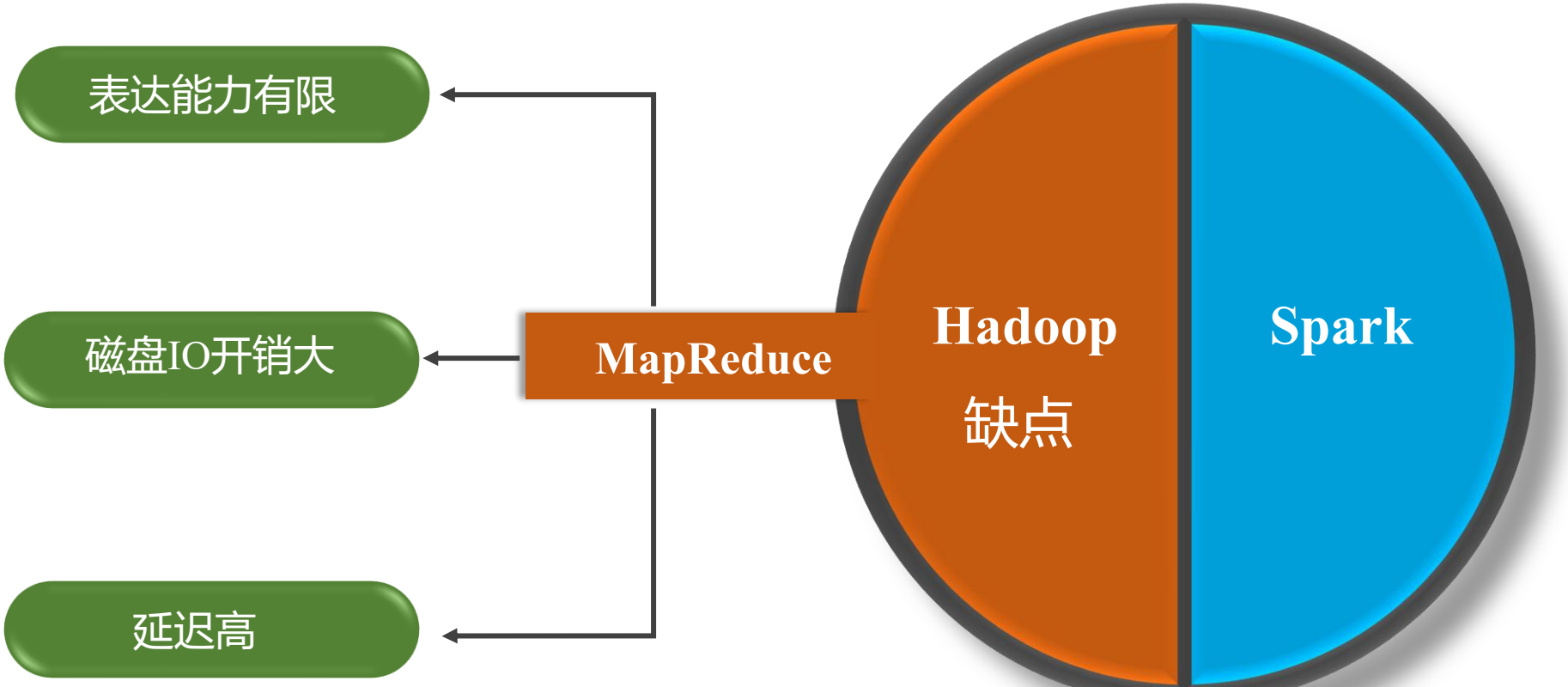


Spark生态系统

1.6.2 代表性大数据技术之Spark

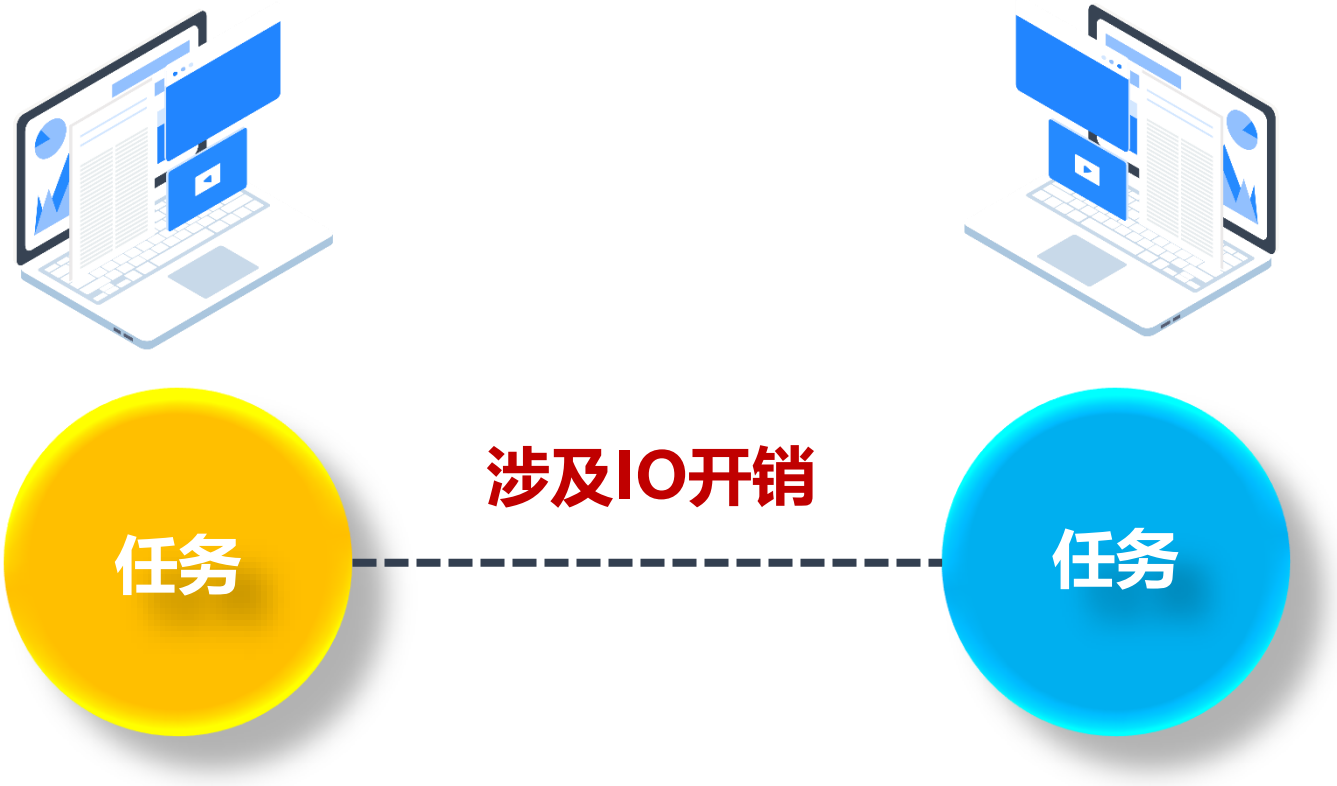


1.6.2 代表性大数据技术之Spark



1.6.2 代表性大数据技术之Spark

延迟高

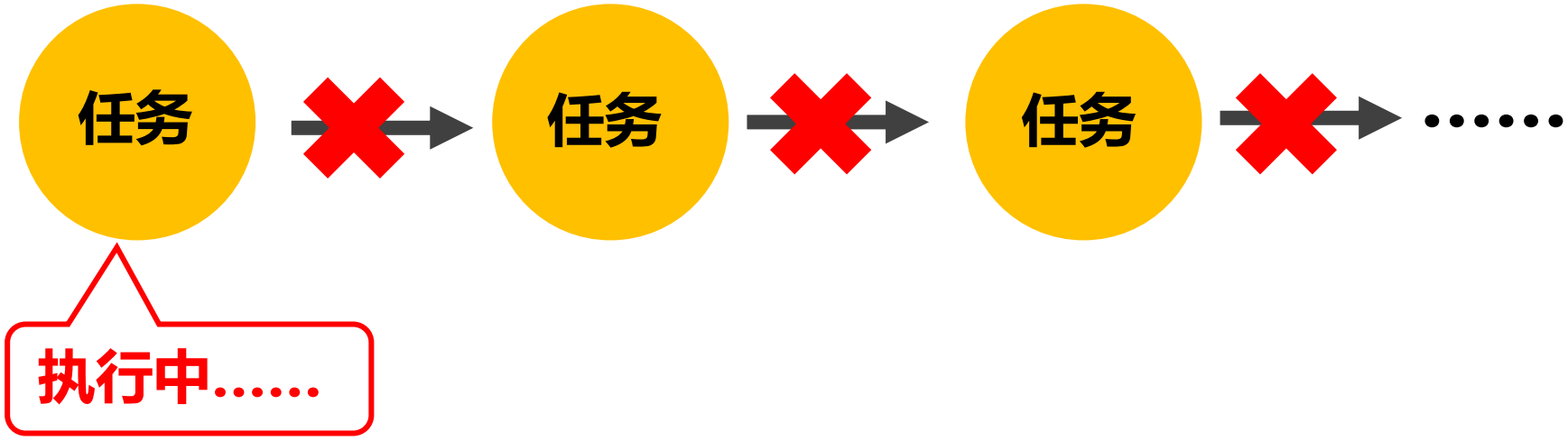


1.6.2 代表性大数据技术之Spark

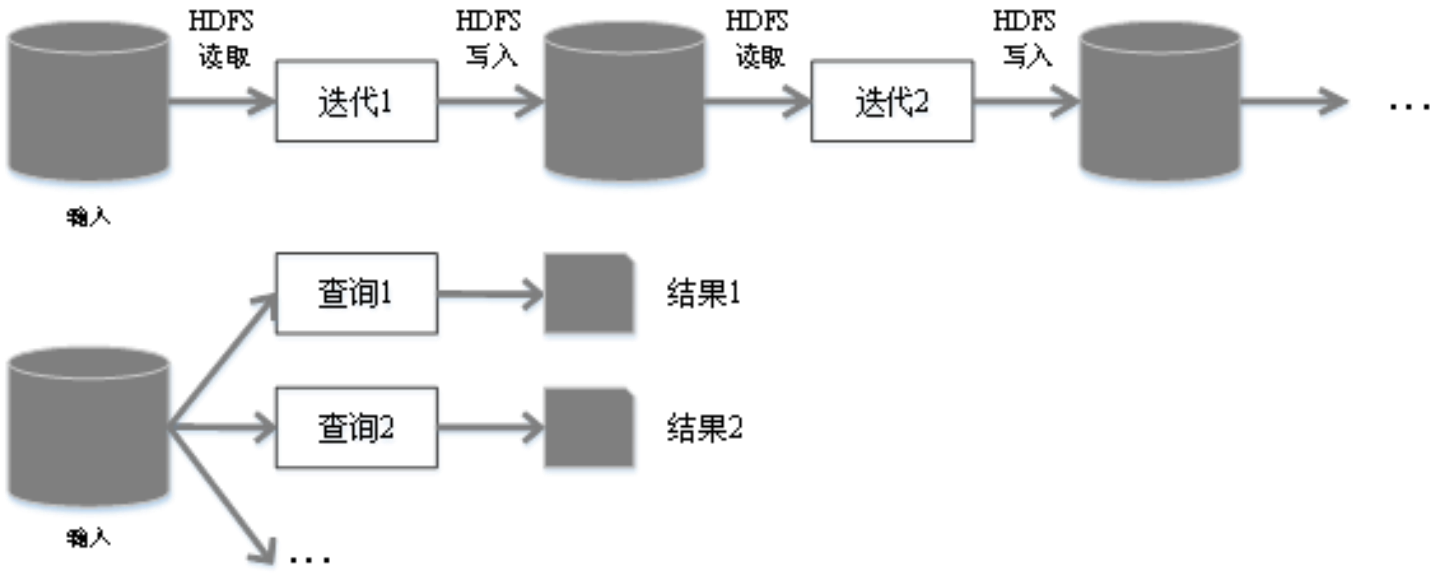


1.6.2 代表性大数据技术之Spark

延迟高

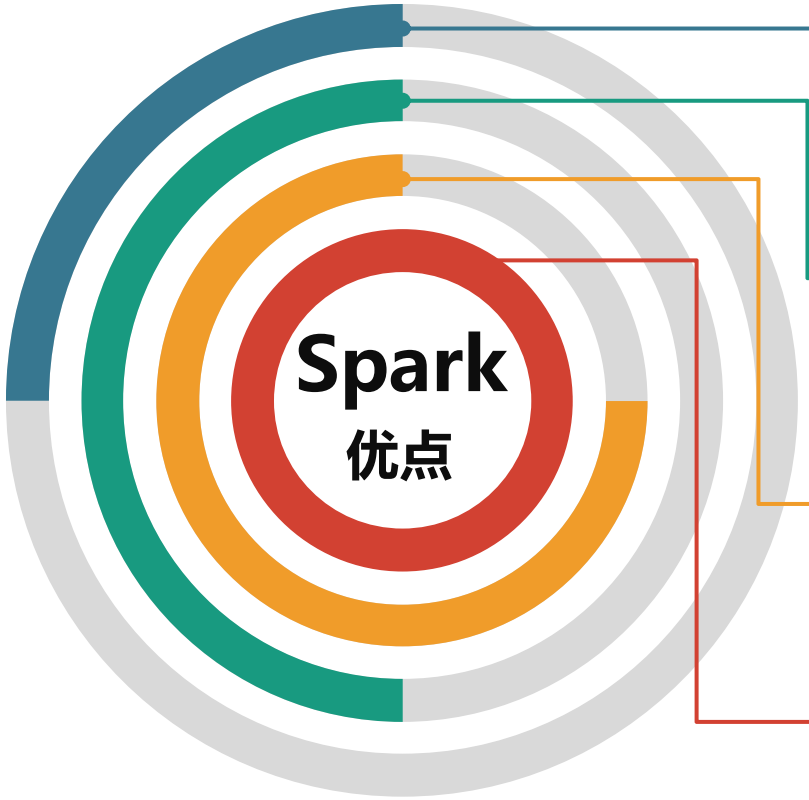


1.6.2 代表性大数据技术之Spark



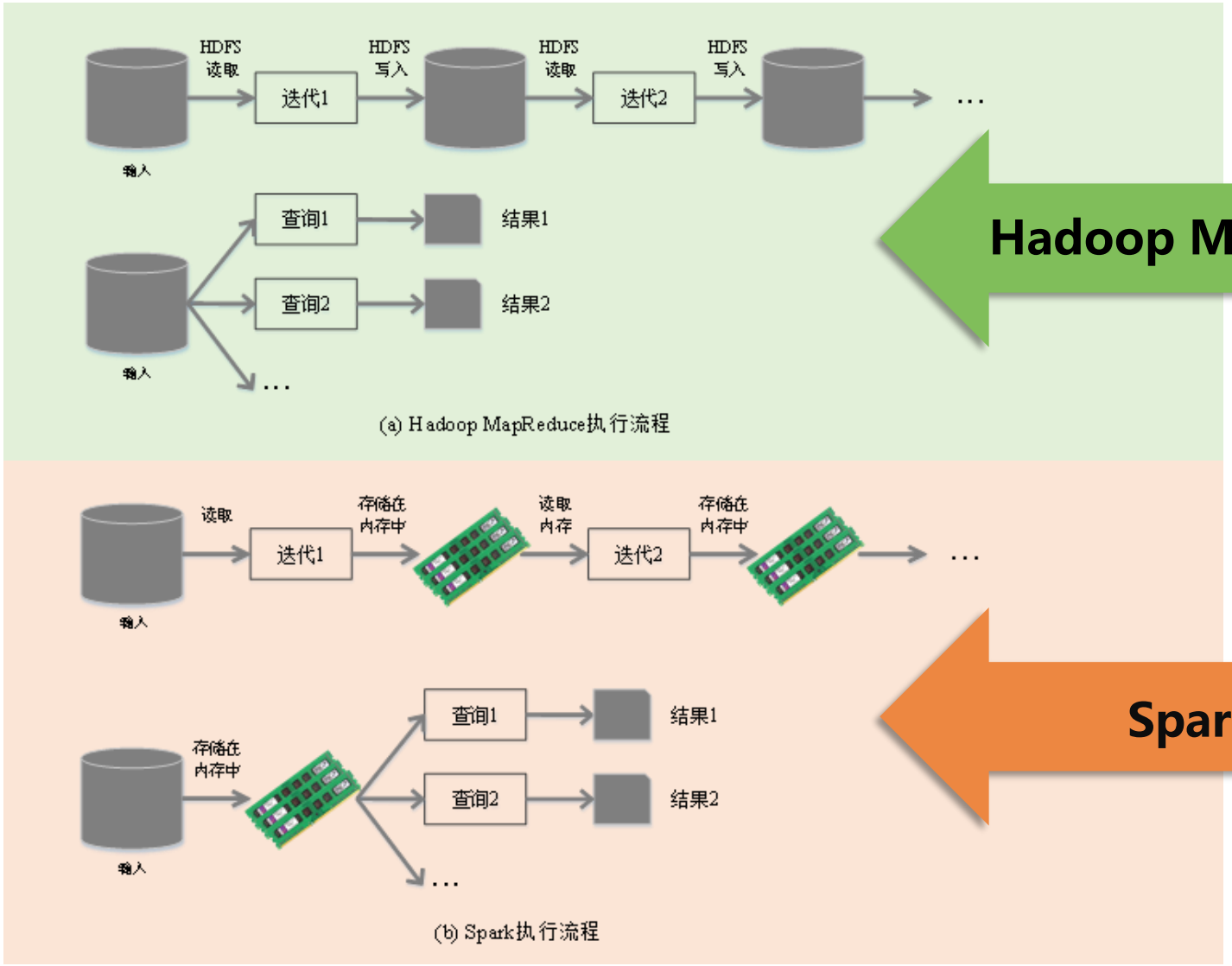
Hadoop MapReduce 执行流程

1.6.2 代表性大数据技术之Spark



- 多种数据集操作类型
- 编程模型更灵活
- 提供了内存计算
- 基于DAG的任务调度执行机制

1.6.2 代表性大数据技术之Spark



Hadoop MapReduce 执行流程

Spark 执行流程

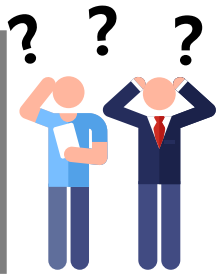
1.6.2 代表性大数据技术之Spark

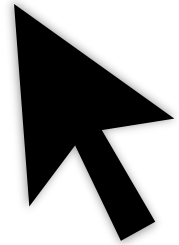


图 Hadoop与Spark执行逻辑斯谛回归的时间对比

1.6.2 代表性大数据技术之Spark

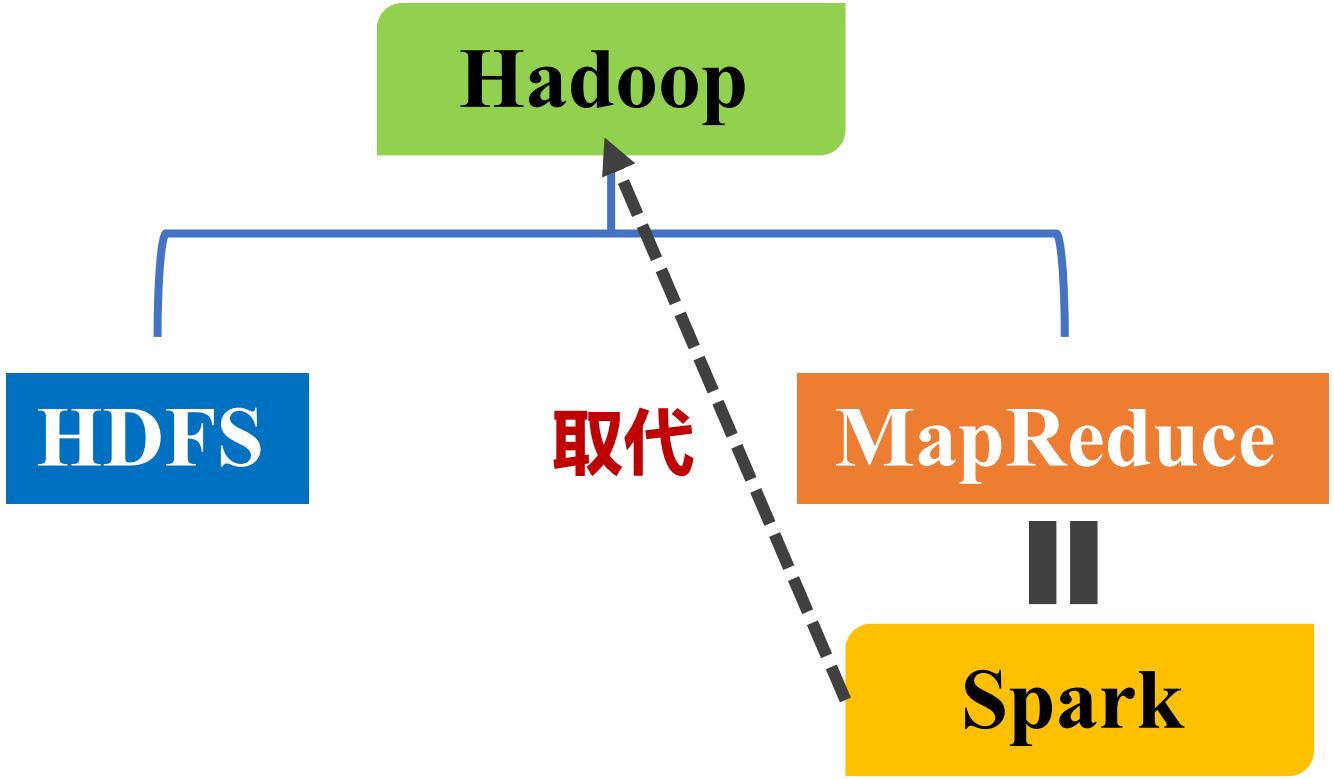
Q1 Spark会取代Hadoop吗?



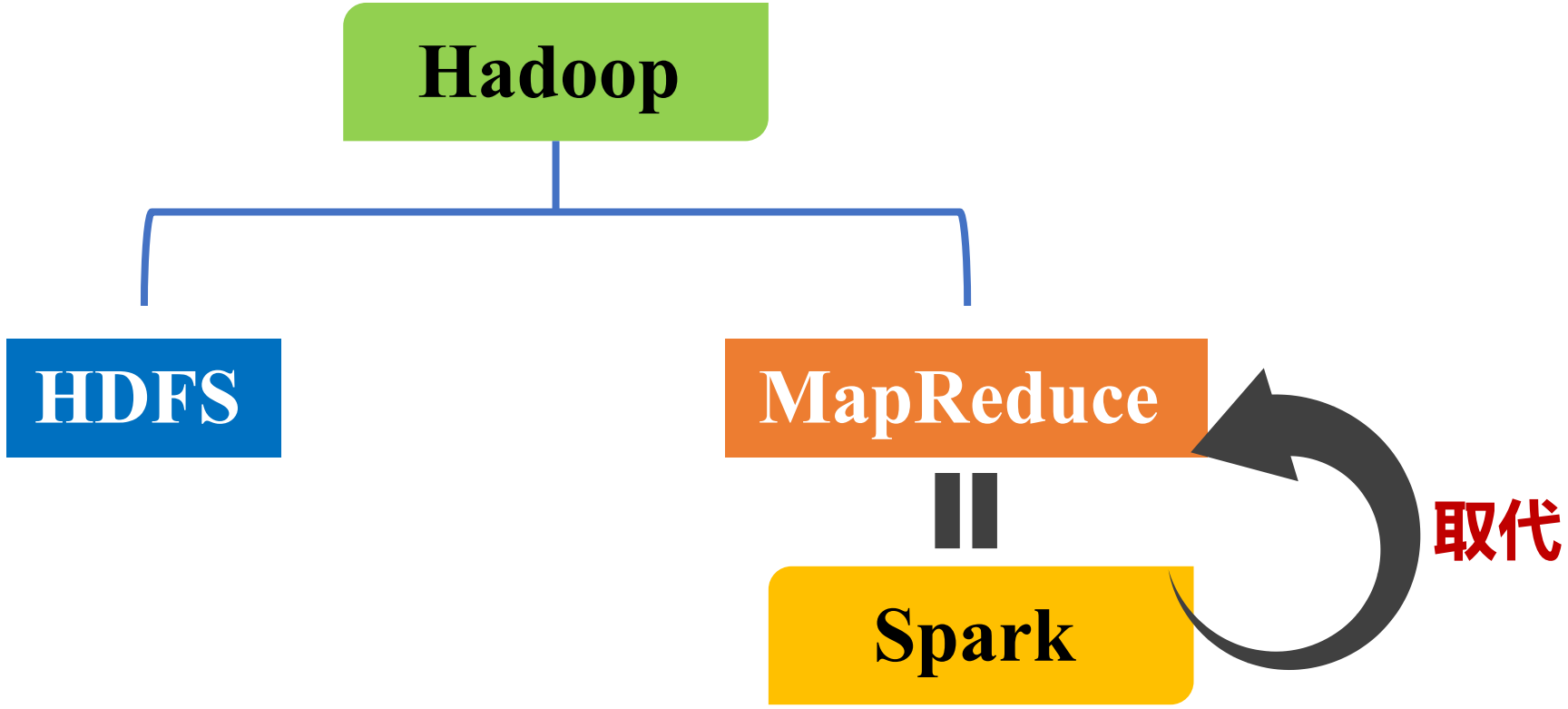




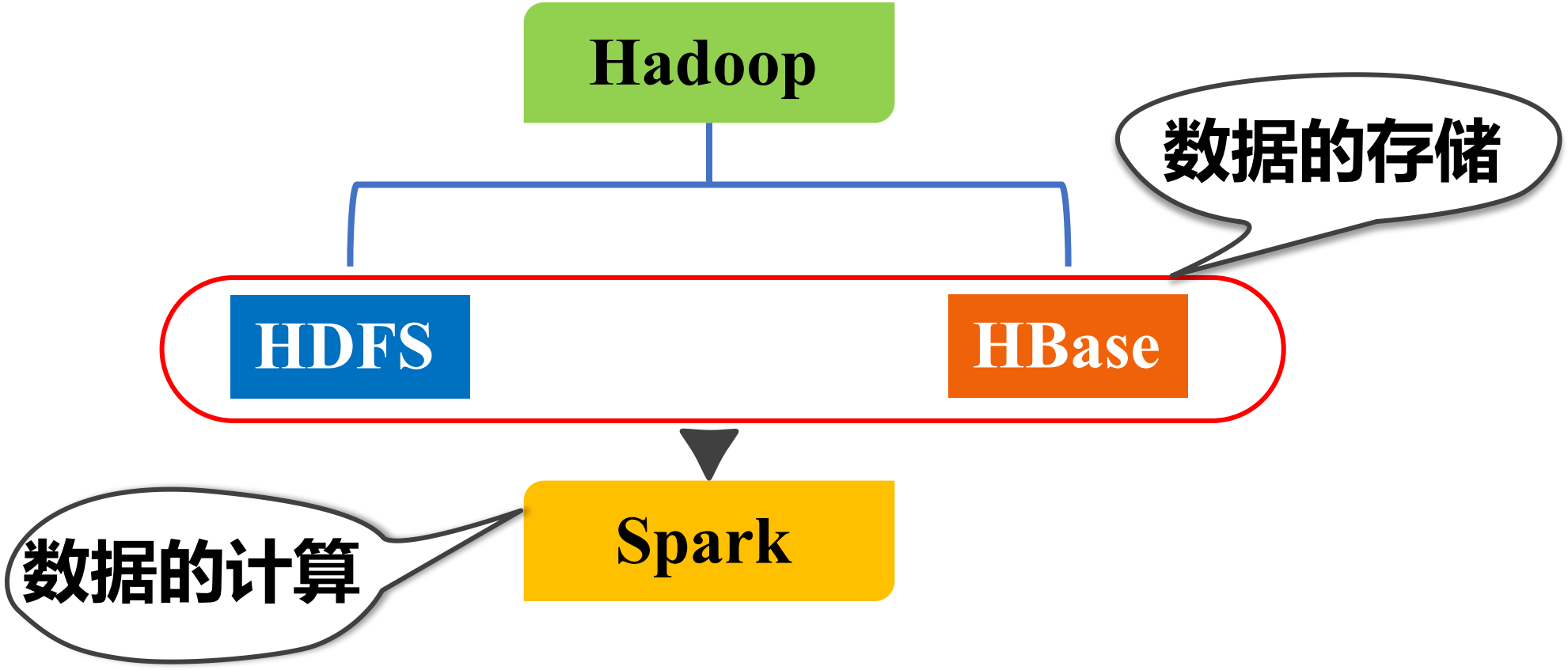
1.6.2 代表性大数据技术之Spark



1.6.2 代表性大数据技术之Spark

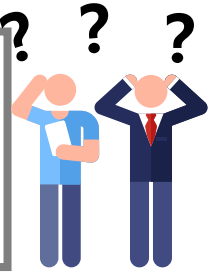


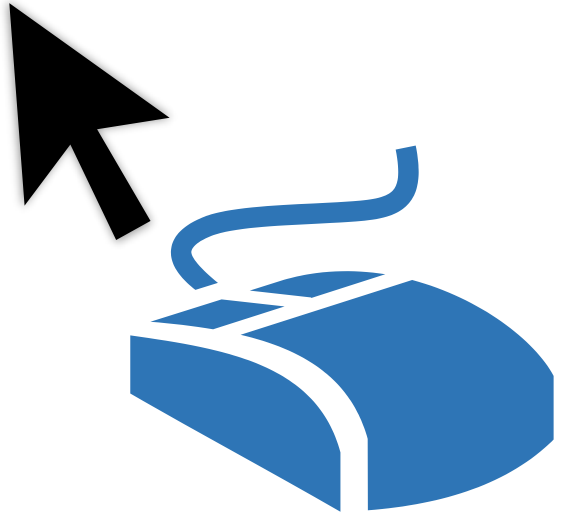
1.6.2 代表性大数据技术之Spark



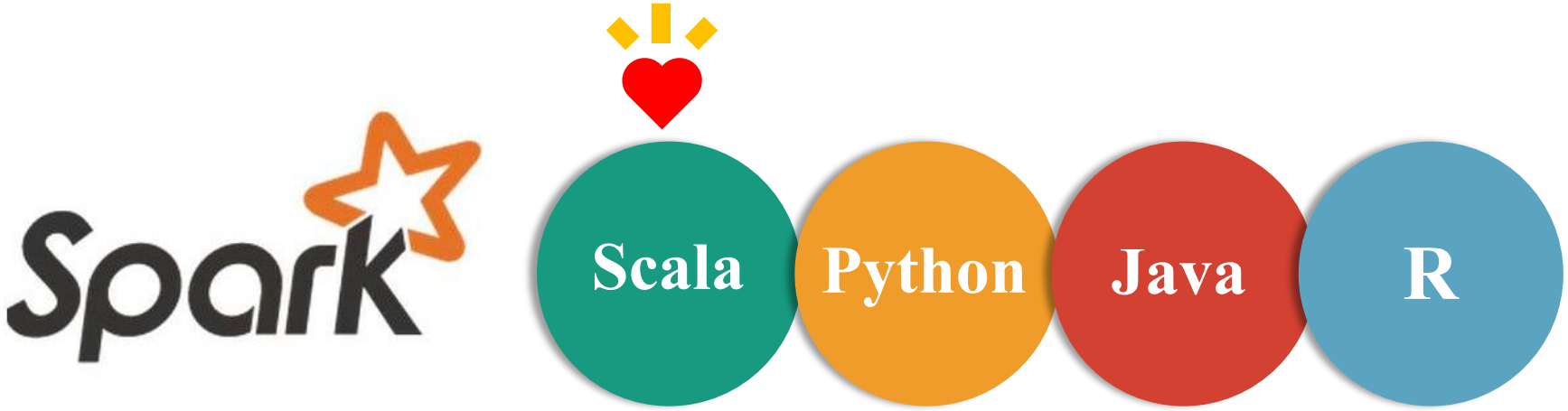
1.6.2 代表性大数据技术之Spark

Q2 开发Spark程序应该使用什么编程语言?

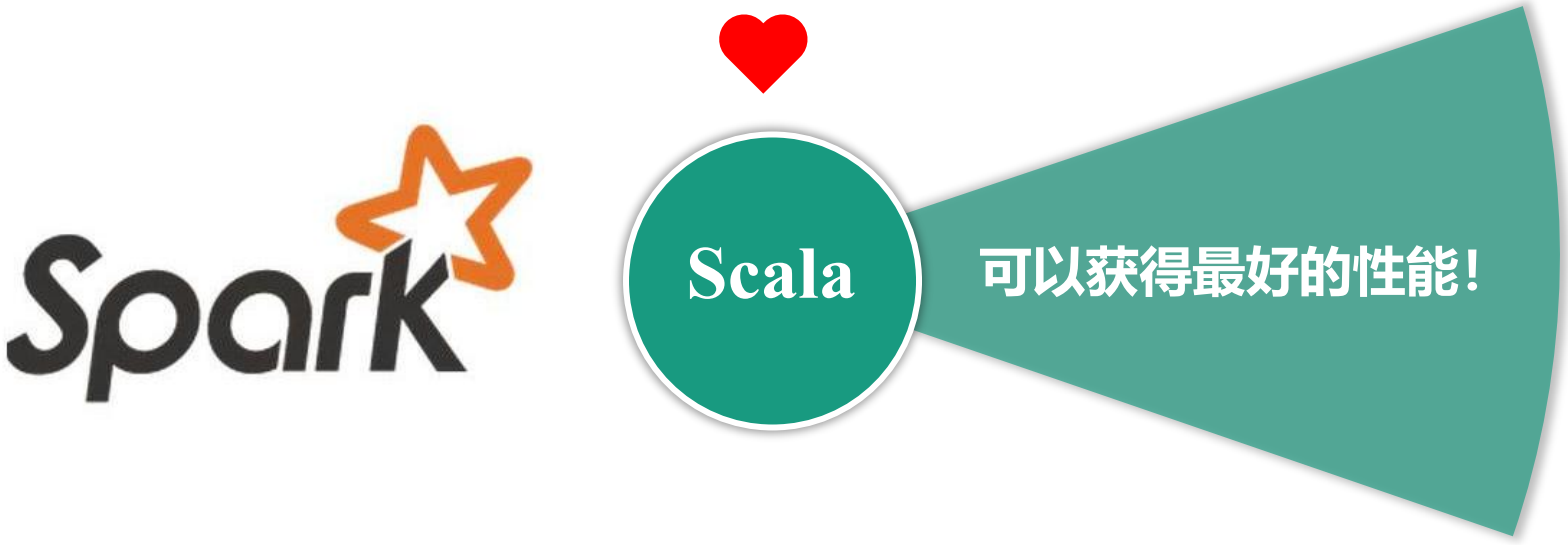




1.6.2 代表性大数据技术之Spark



1.6.2 代表性大数据技术之Spark



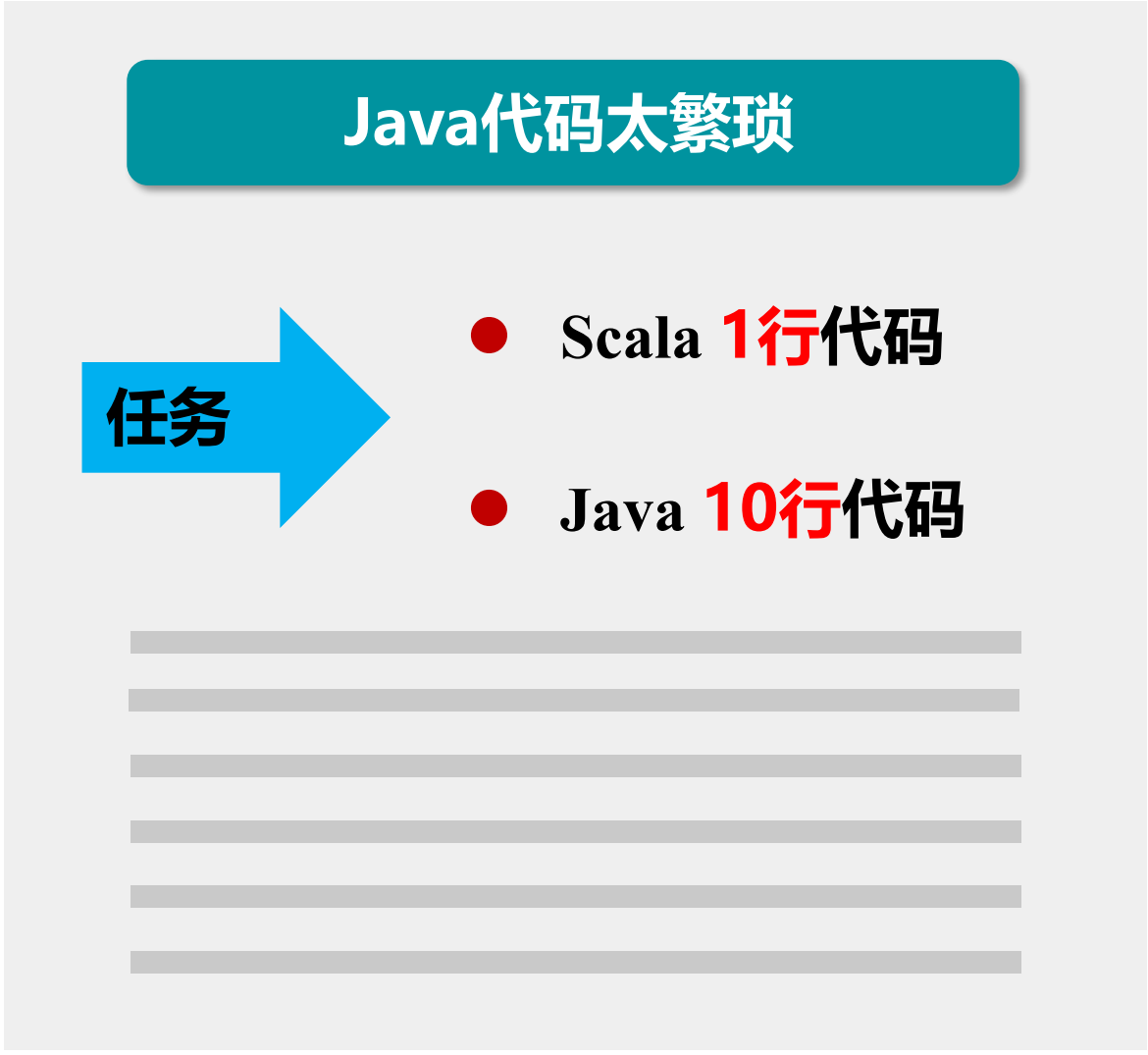
1.6.2 代表性大数据技术之Spark



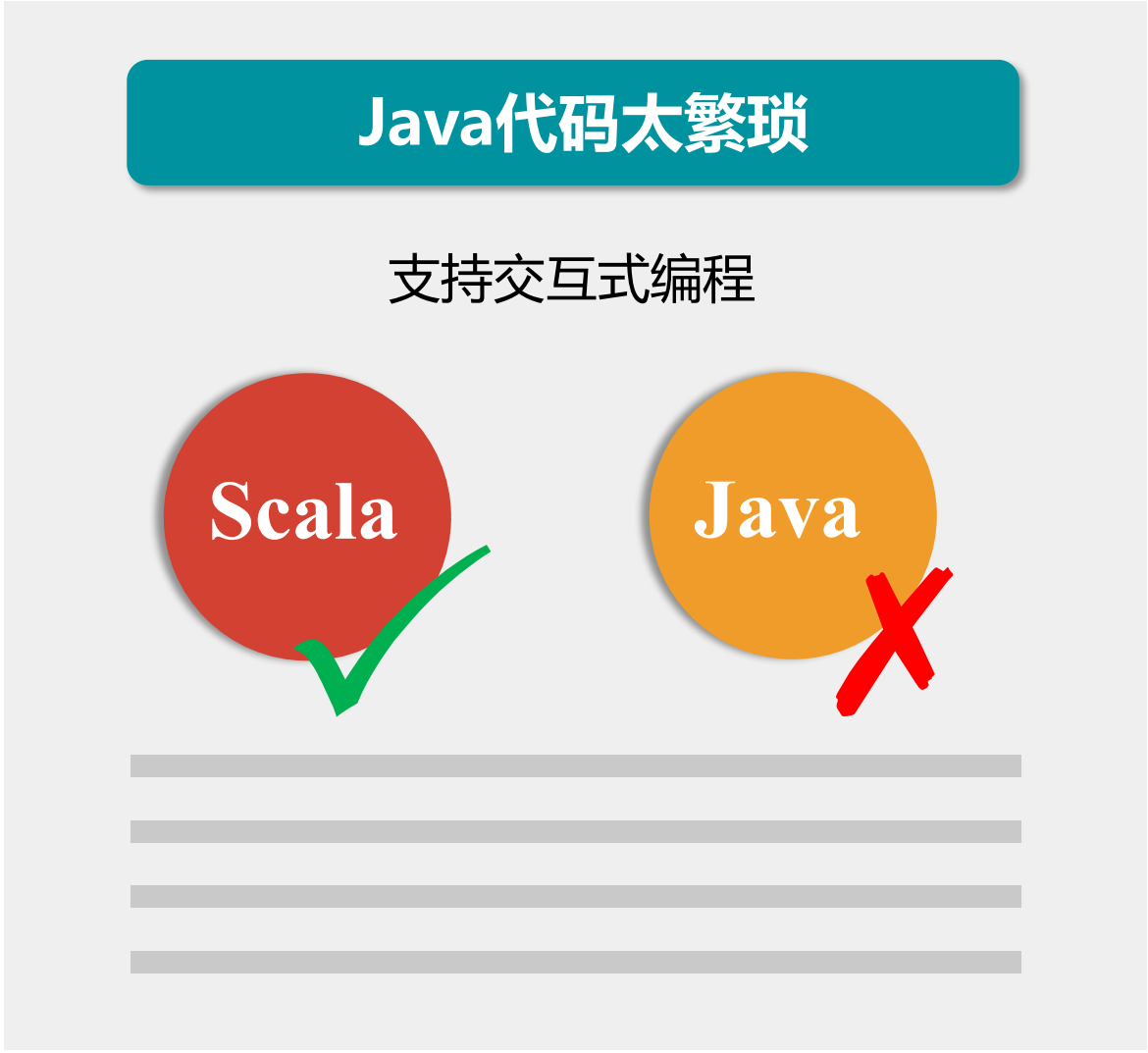
Java代码太繁琐

Python语言并发性能不好

1.6.2 代表性大数据技术之Spark



1.6.2 代表性大数据技术之Spark



1.6.2 代表性大数据技术之Spark



Java代码太繁琐

Python语言并发性能不好

1.6.2 代表性大数据技术之Spark



1.6.2 代表性大数据技术之Spark

Python语言并发性能不好

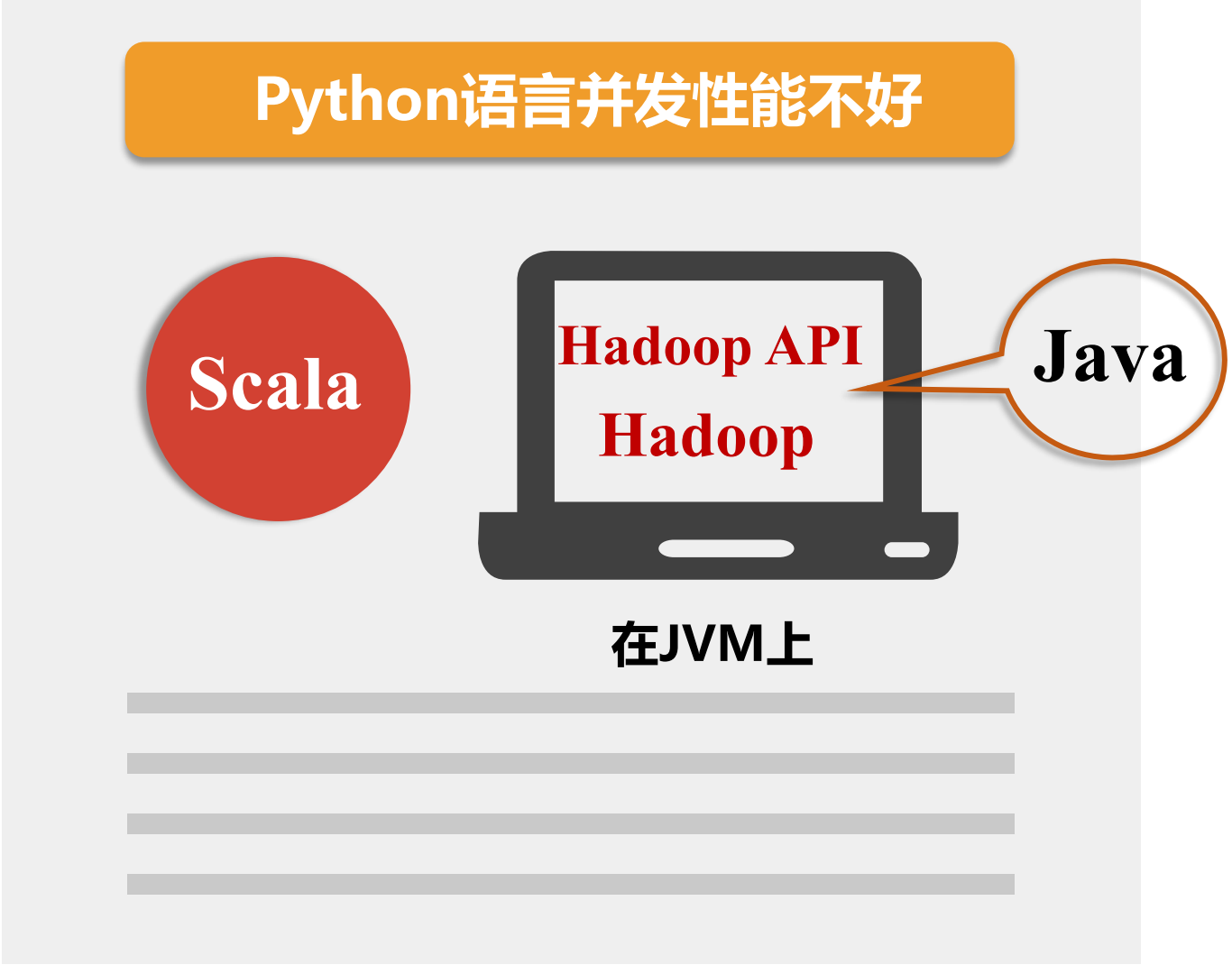


- 静态类型(static typing)
- 便于开发大型大数据项目

1.6.2 代表性大数据技术之Spark



1.6.2 代表性大数据技术之Spark



1.6.2 代表性大数据技术之Spark



1.6.2 代表性大数据技术之Spark

Python语言并发性能不好

Python

企业级应用 ✗

科研的时候 ✓

1.6.2 代表性大数据技术之Spark

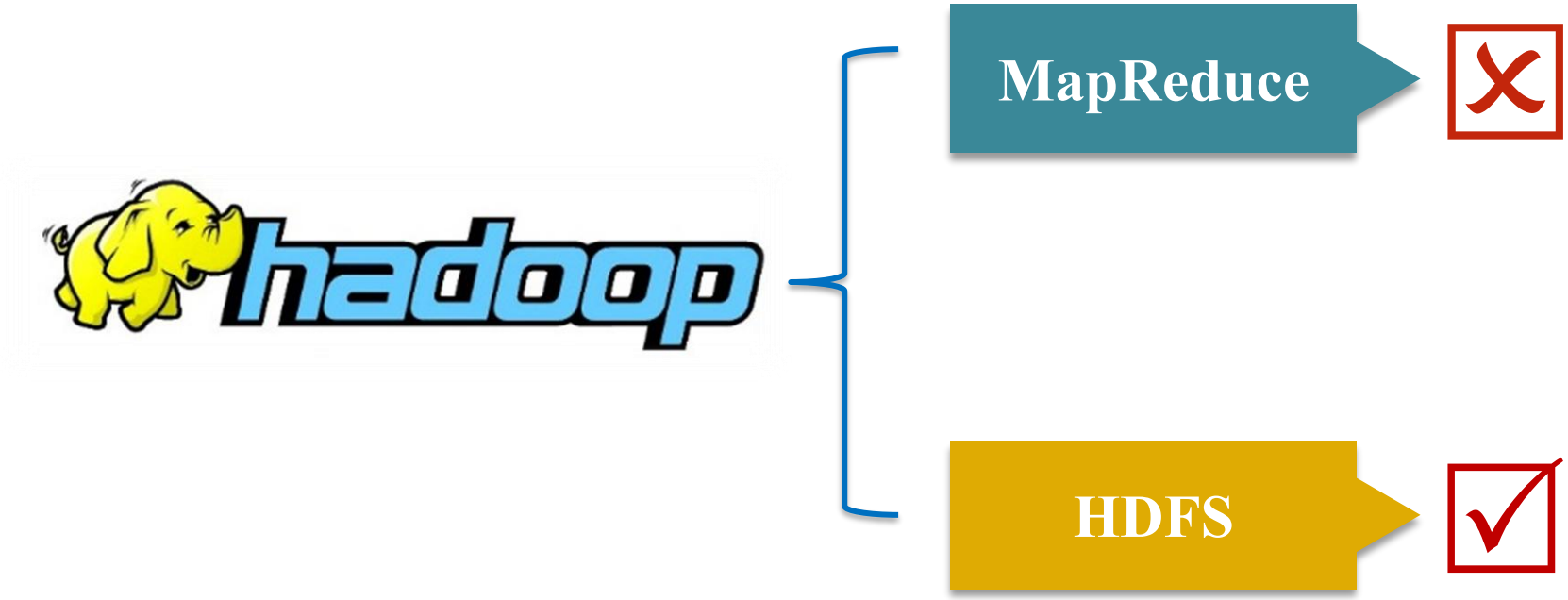
Python语言并发性能不好

Scala

教学首选

1.6.2 代表性大数据技术之Spark





1.6.2 代表性大数据技术之Spark



表达能力有限



复杂业务逻辑

Map函数

Reduce函数



降低了分布式应用开发的复杂性

1.6.2 代表性大数据技术之Spark



表达能力有限



复杂业务逻辑

Map函数

Reduce函数



限制了表达能力

1.6.2 代表性大数据技术之Spark

磁盘IO开销大

MapReduce

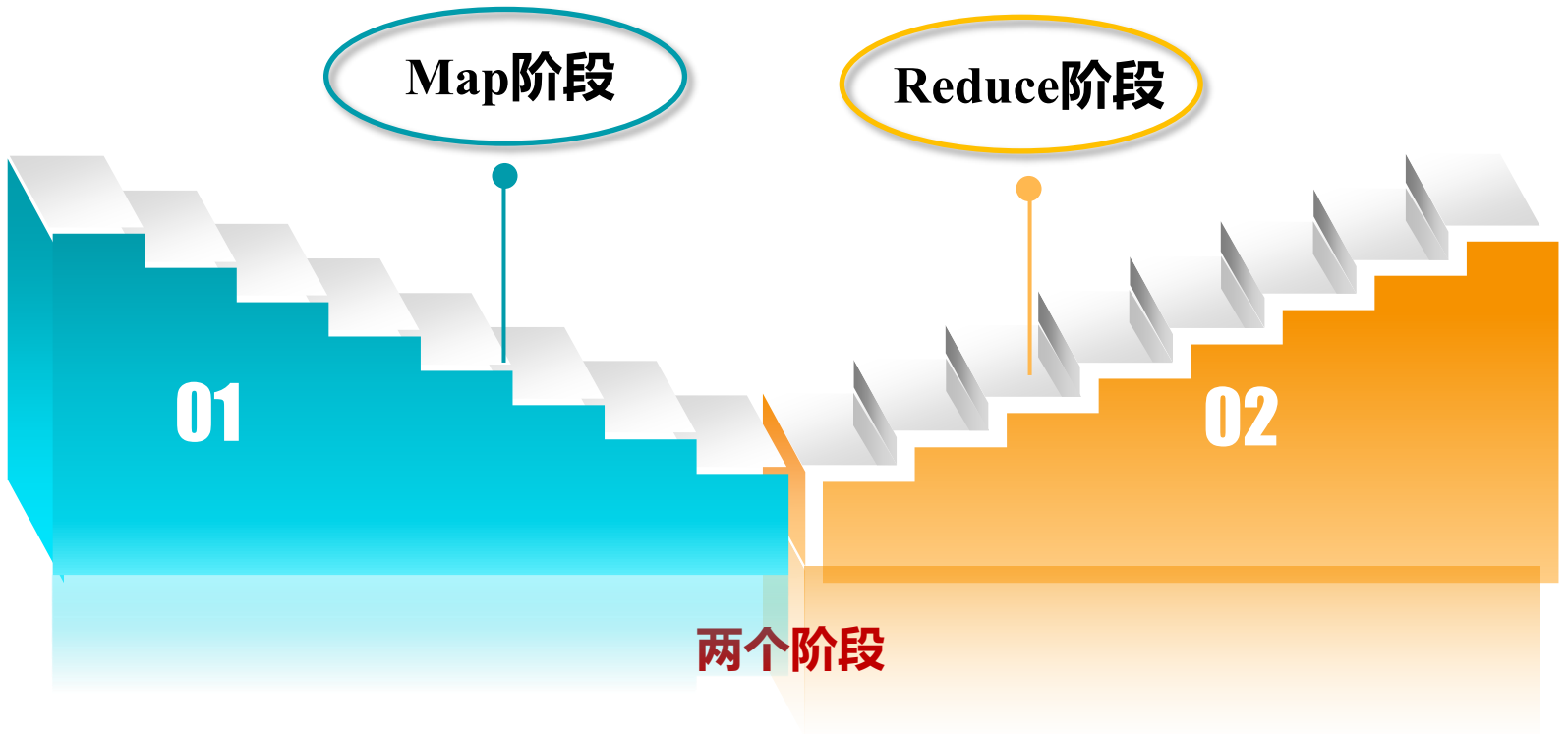




1.6.2 代表性大数据技术之Spark

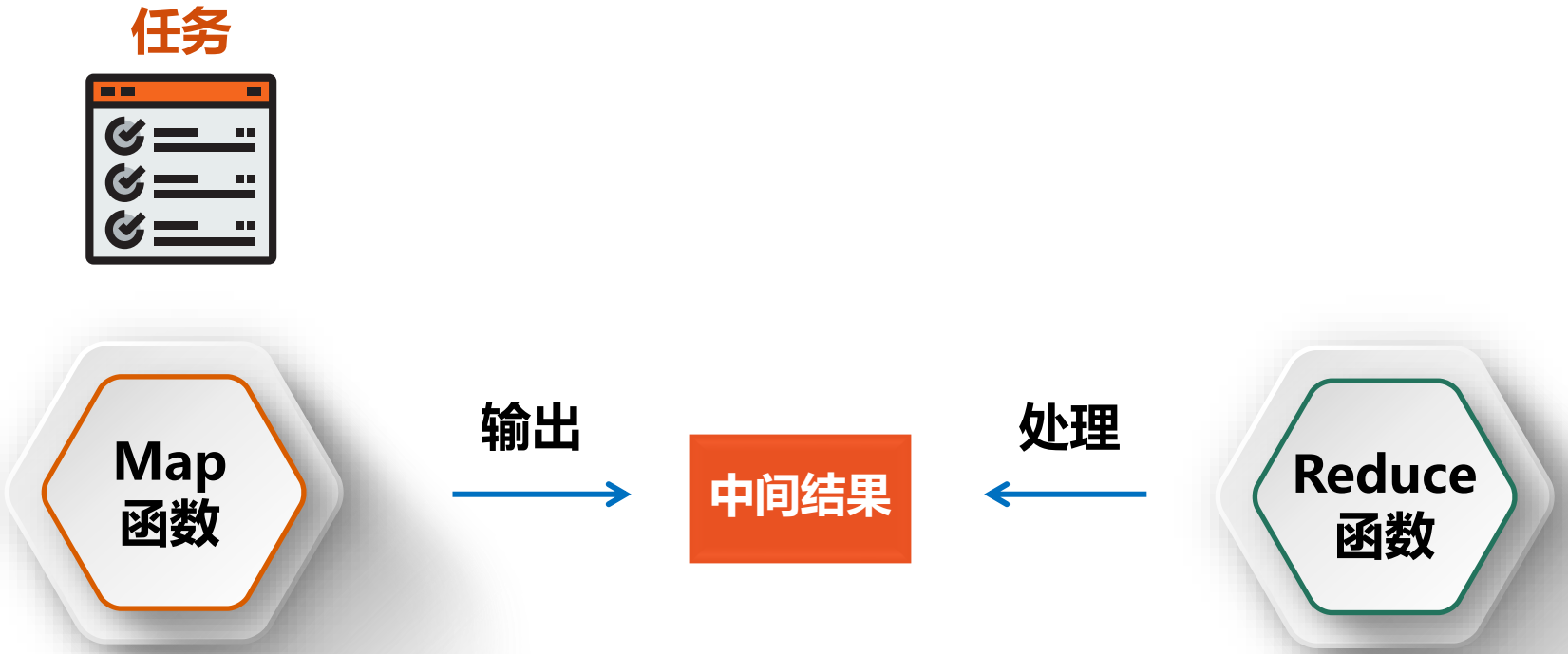


延迟高

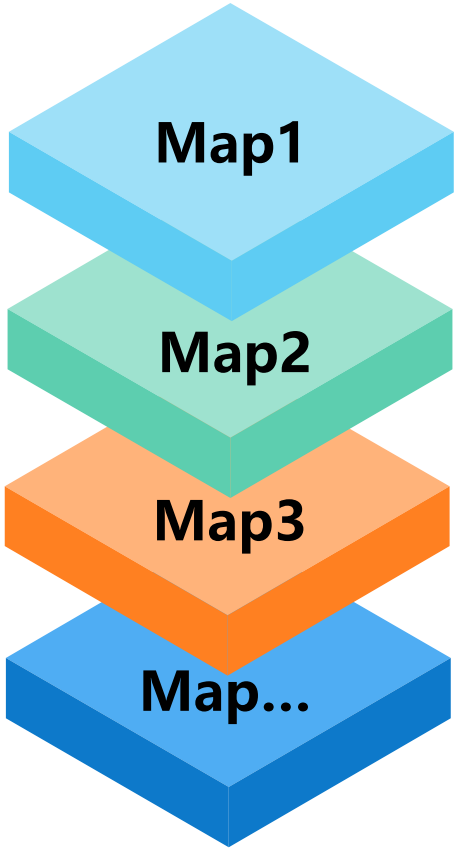


1.6.2 代表性大数据技术之Spark





1.6.2 代表性大数据技术之Spark



并行处理

Reduce1

Reduce2

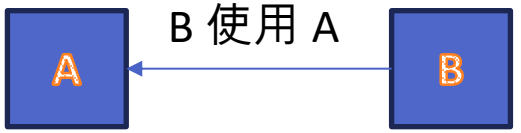
Reduce3

Reduce...

1.6.2 代表性大数据技术之Spark



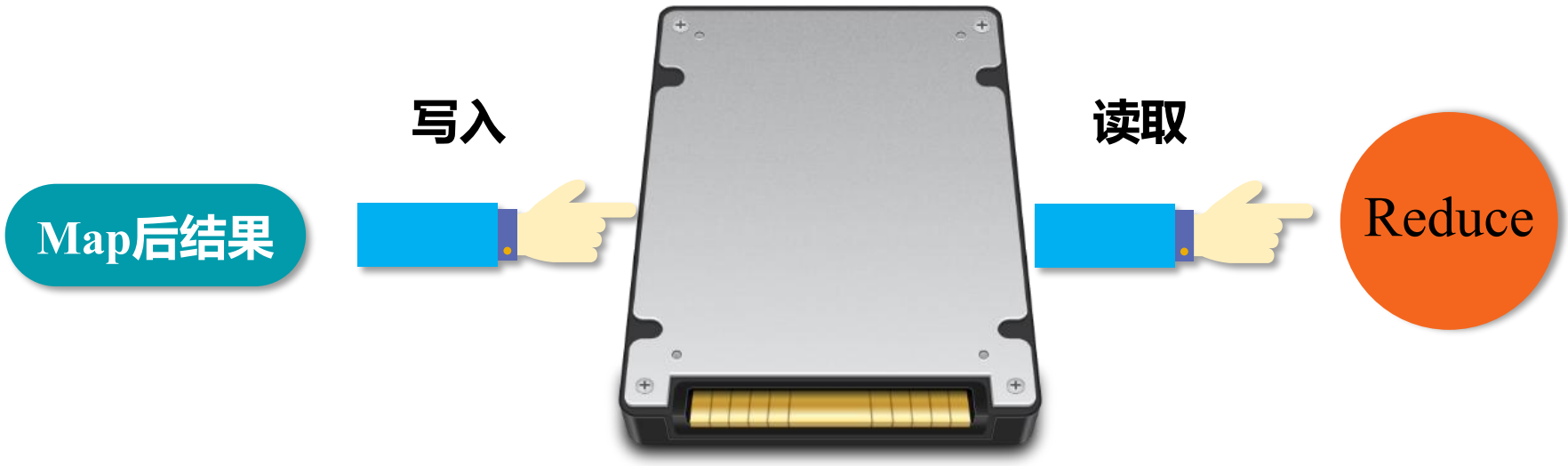
依赖关系 (Dependency)



箭头从使用者指向被使用者

1.6.2 代表性大数据技术之Spark

衔接的IO开销大



1.6.2 代表性大数据技术之Spark



延迟高

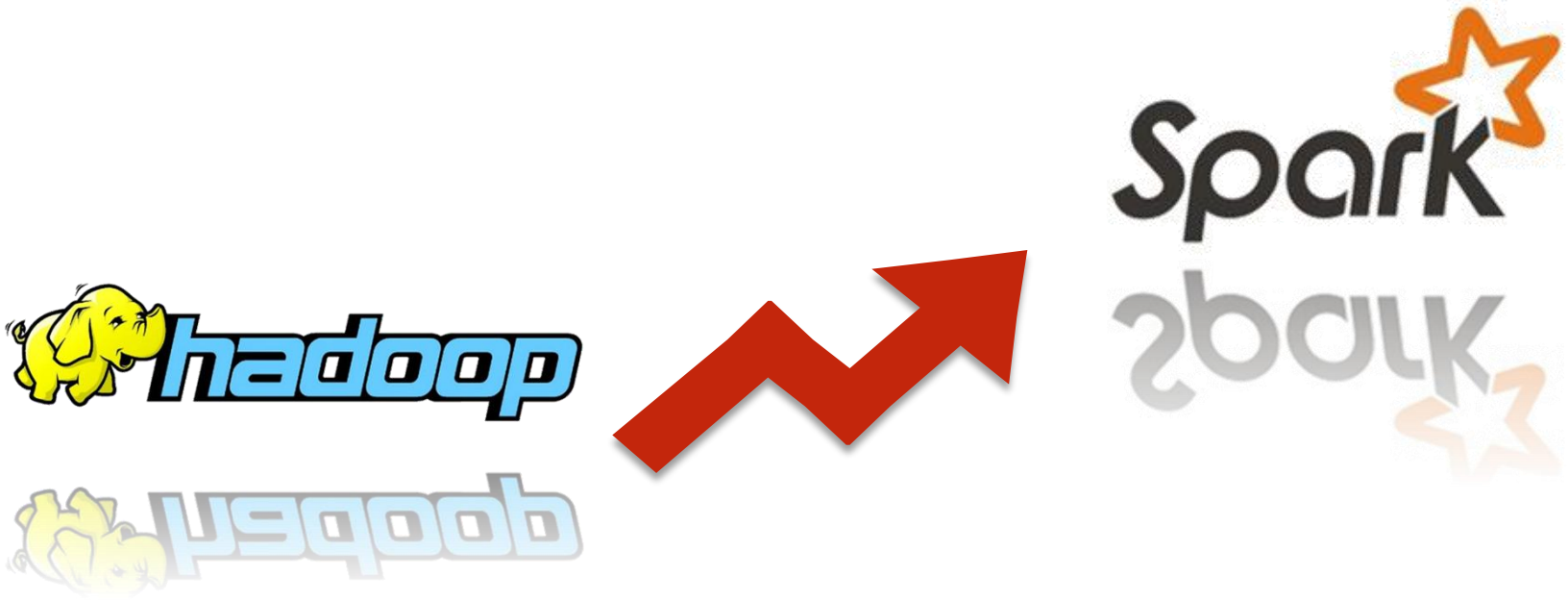


磁盘的IO开销大

衔接的IO开销大

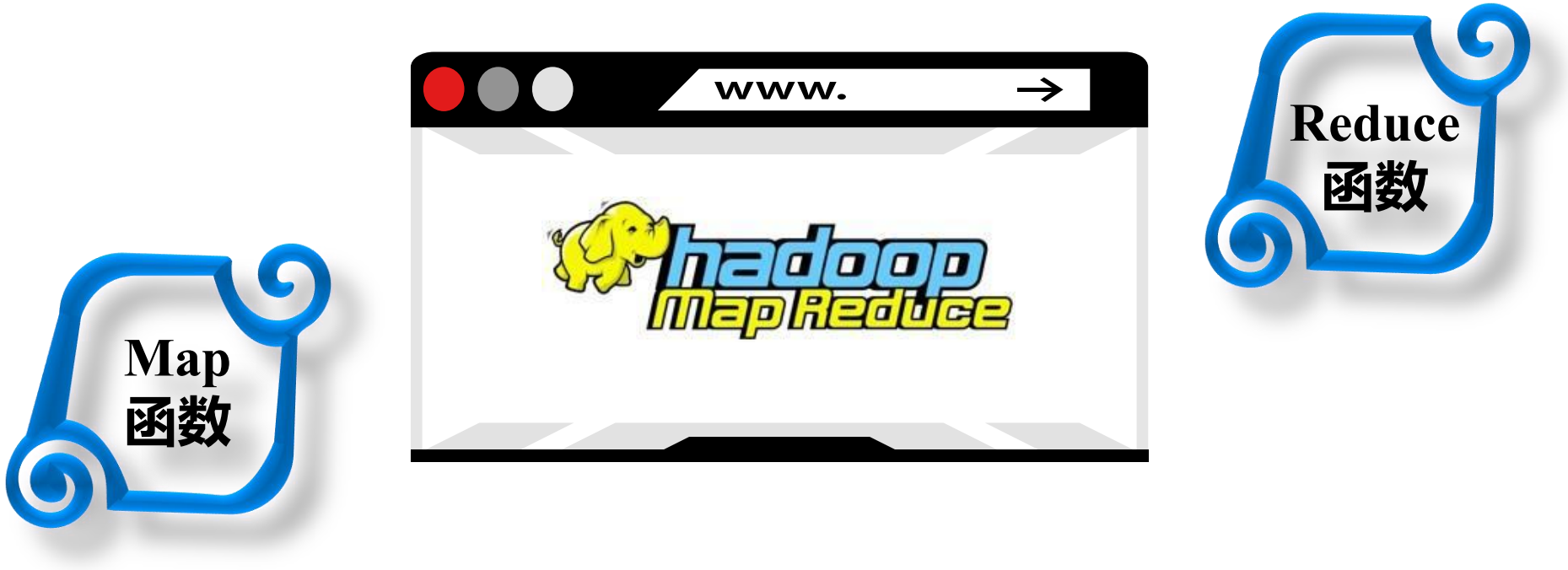
MapReduce的缺陷

1.6.2 代表性大数据技术之Spark



Spark对Hadoop进行缺陷改进与继承

1.6.2 代表性大数据技术之Spark



1.6.2 代表性大数据技术之Spark





1.6.2 代表性大数据技术之Spark



操作类型更多

表达能力更强

1.6.2 代表性大数据技术之Spark



内存计算

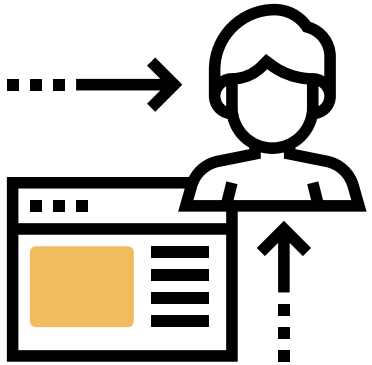




1.6.2 代表性大数据技术之Spark



内存计算



再次访问





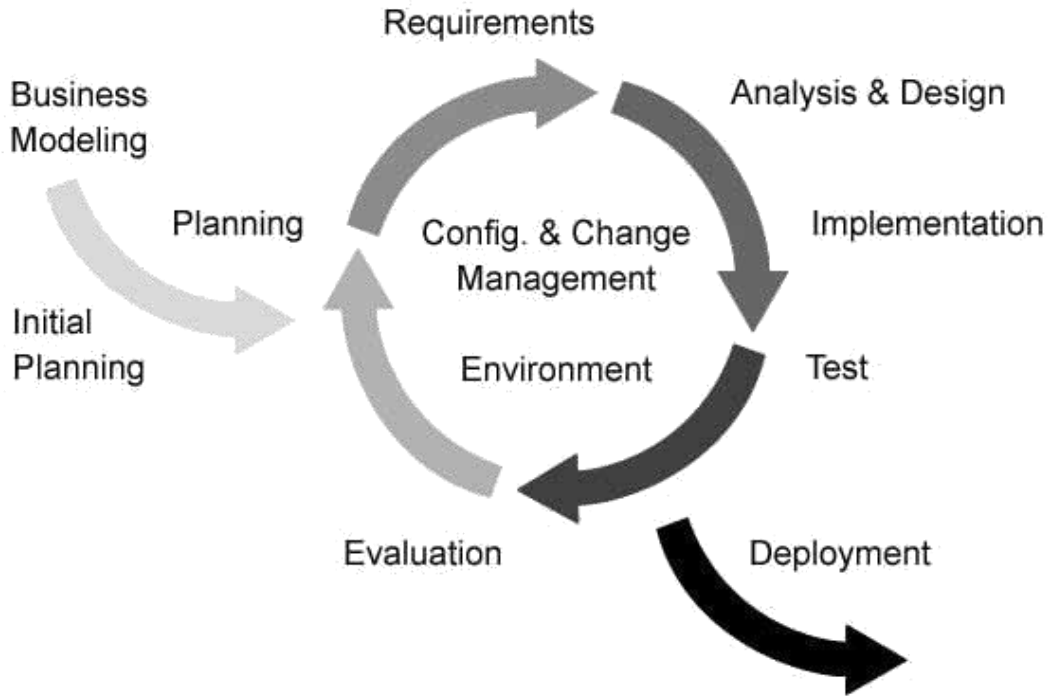
低延迟

1.6.2 代表性大数据技术之Spark



运行效率提升

高效提高迭代运算

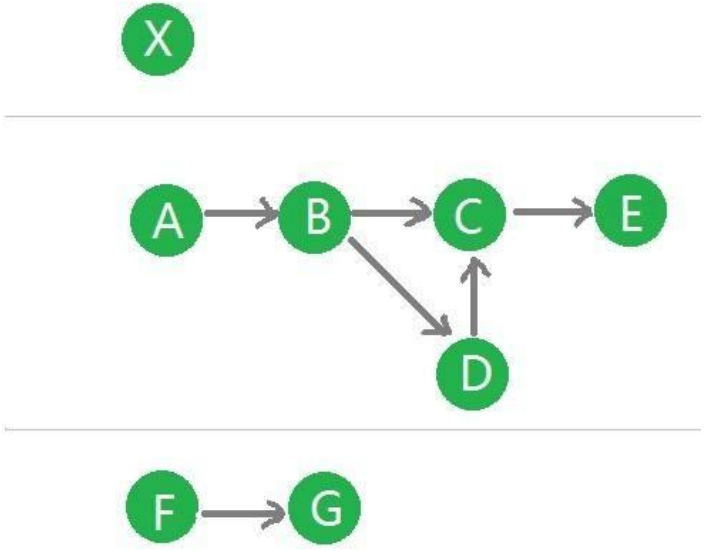


1.6.2 代表性大数据技术之Spark





DAG任务调度执行机制



DAG—有向无环图

1.6.2 代表性大数据技术之Spark

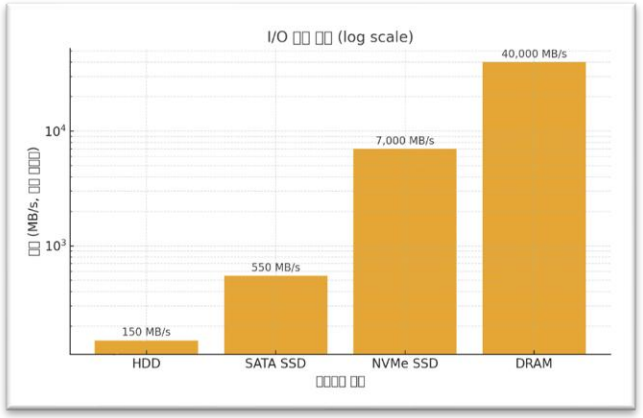


使用MapReduce进行反复迭代

迭代1







结果1

速度差异总结 (大致比例)

对比对象	速度比例	说明
HDD : SATA SSD	1 : 30~50	SSD 比机械硬盘快约 30~50 倍
HDD : NVMe SSD	1 : 100~500	NVMe 比 HDD 快数百倍
NVMe SSD : DRAM	1 : 1000	DRAM 比 NVMe 快约 1000 倍
HDD : DRAM	1 : 100,000	DRAM 比 HDD 快约 10 万倍

1.6.2 代表性大数据技术之Spark



使用MapReduce进行反复迭代

迭代1





迭代2

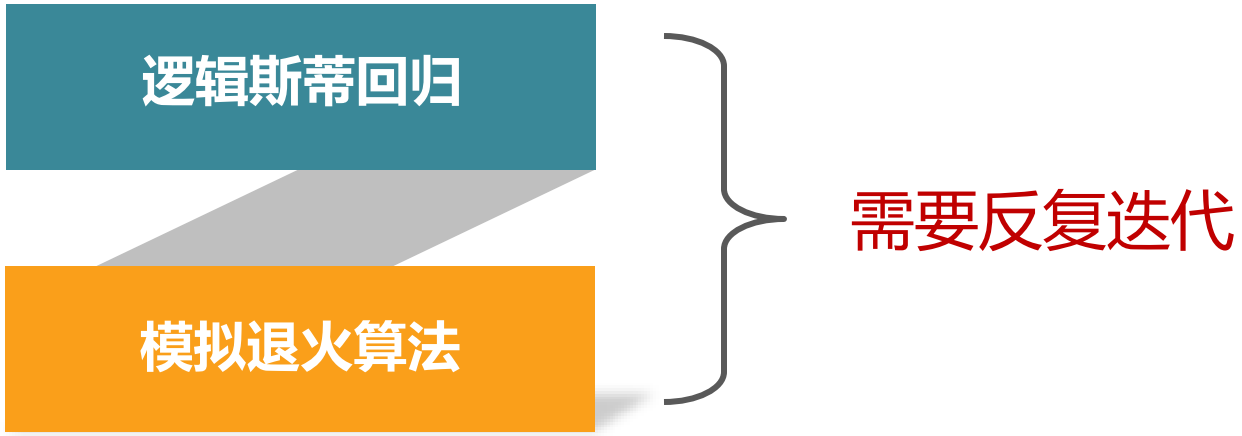


结果2

1.6.2 代表性大数据技术之Spark



使用MapReduce进行反复迭代



1.6.2 代表性大数据技术之Spark



使用MapReduce进行反复迭代



1.6.2 代表性大数据技术之Spark



使用MapReduce进行反复迭代

迭代1

Map ✚ Reduce

结果1



1.6.2 代表性大数据技术之Spark



使用MapReduce进行反复迭代

迭代1





迭代2



结果2

1.6.2 代表性大数据技术之Spark



避免数据落地

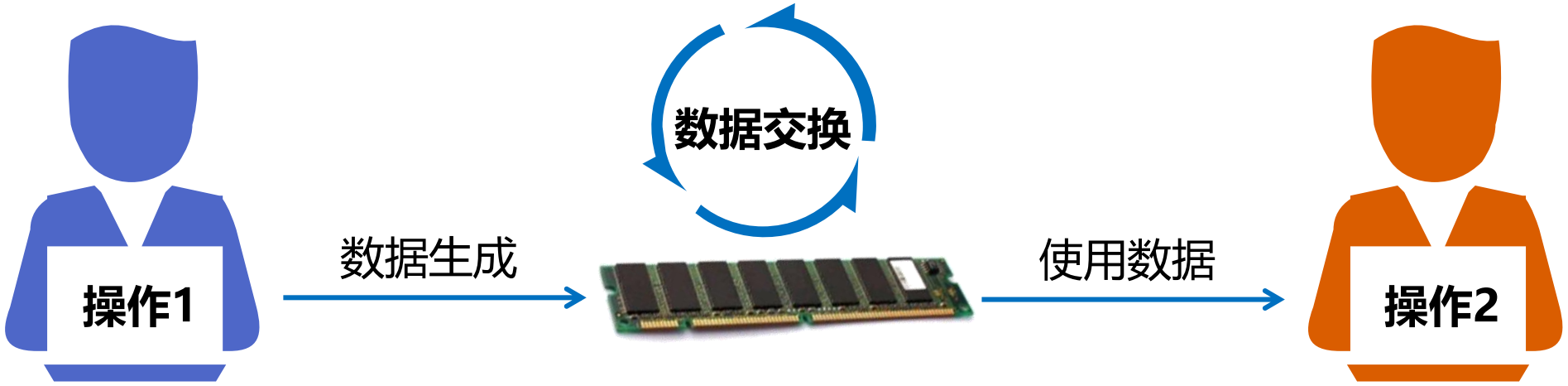




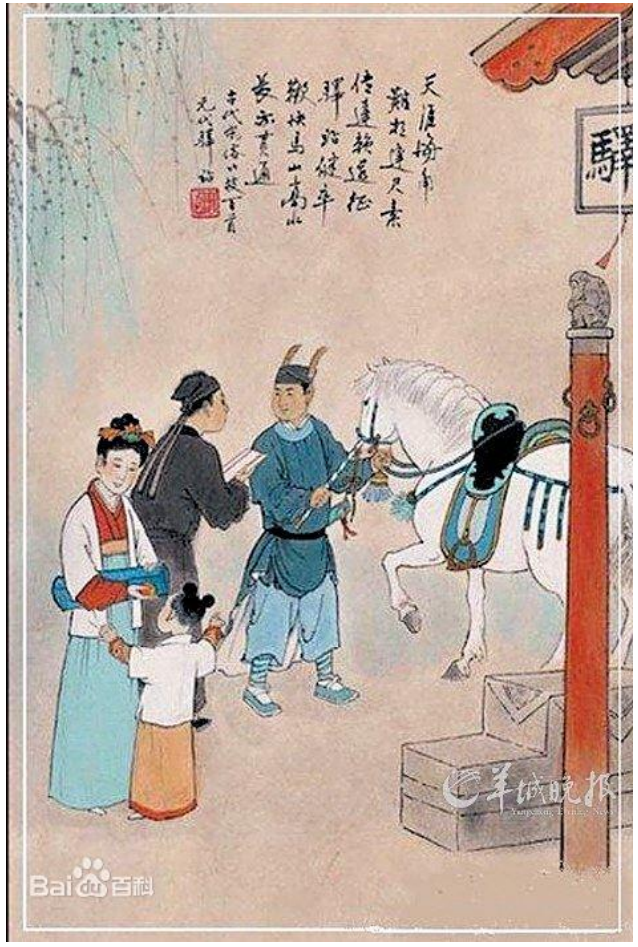
1.6.2 代表性大数据技术之Spark

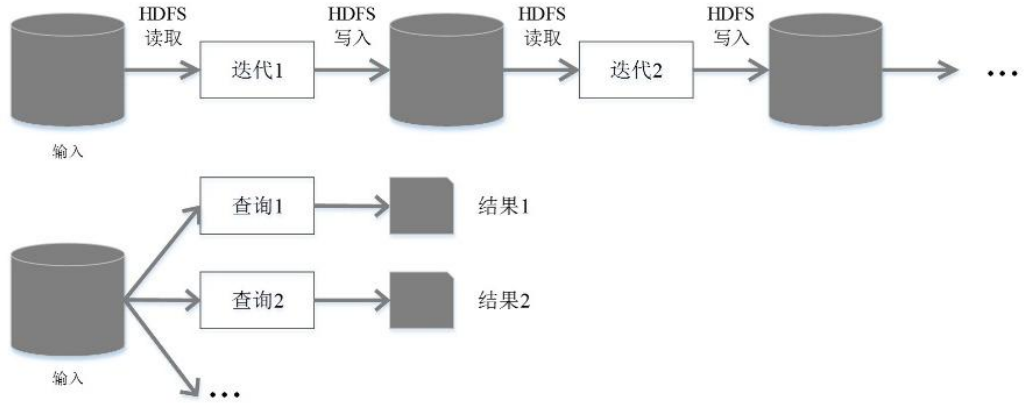


流水线优化

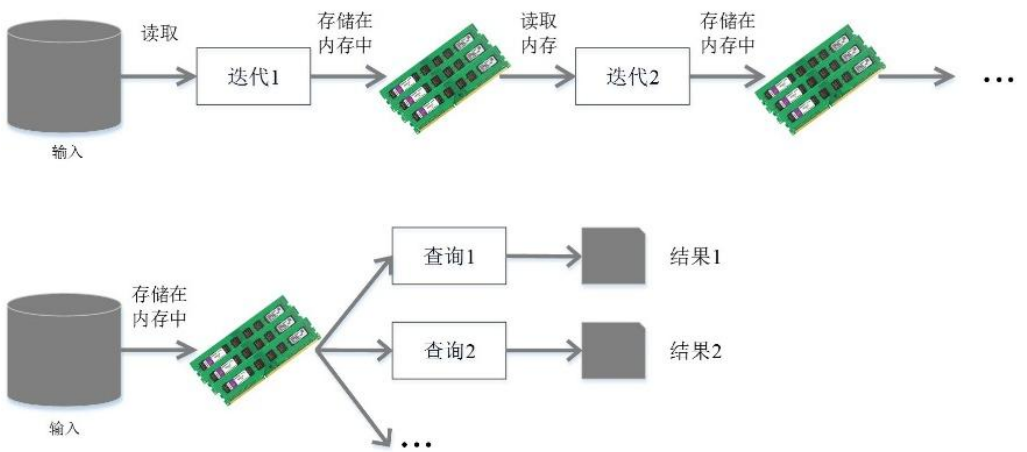


1.6.2 代表性大数据技术之Spark





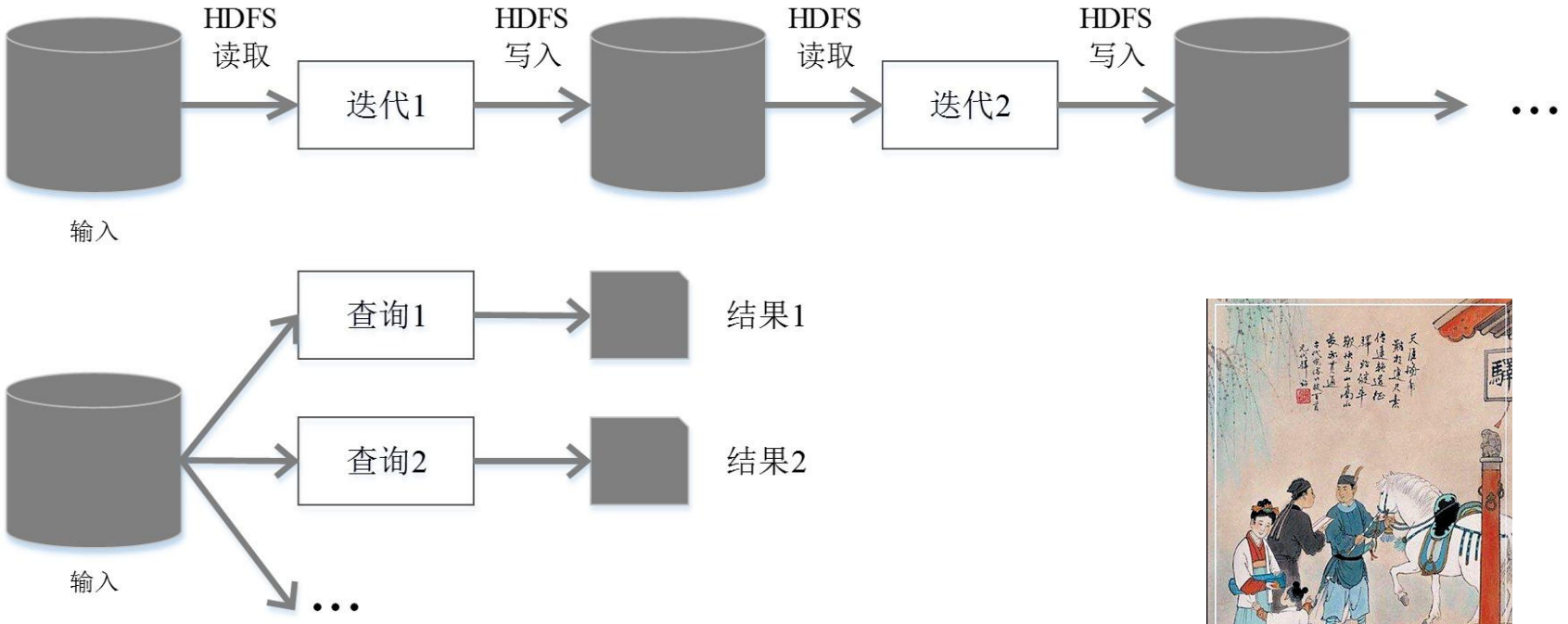
(a) Hadoop MapReduce执行流程



(b) Spark执行流程

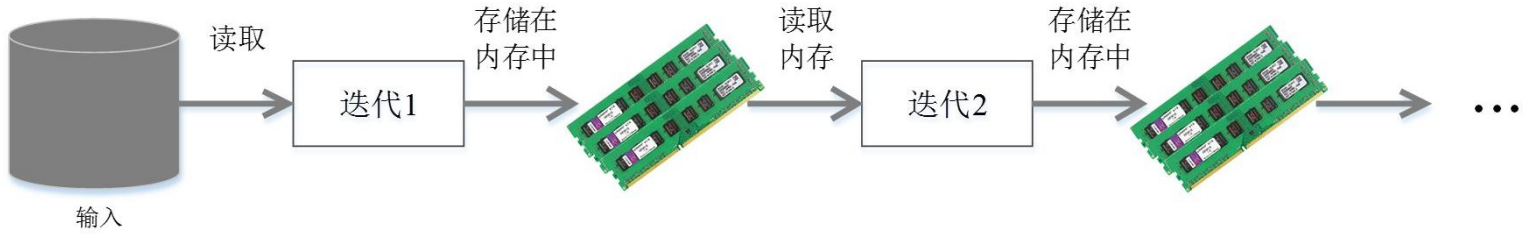


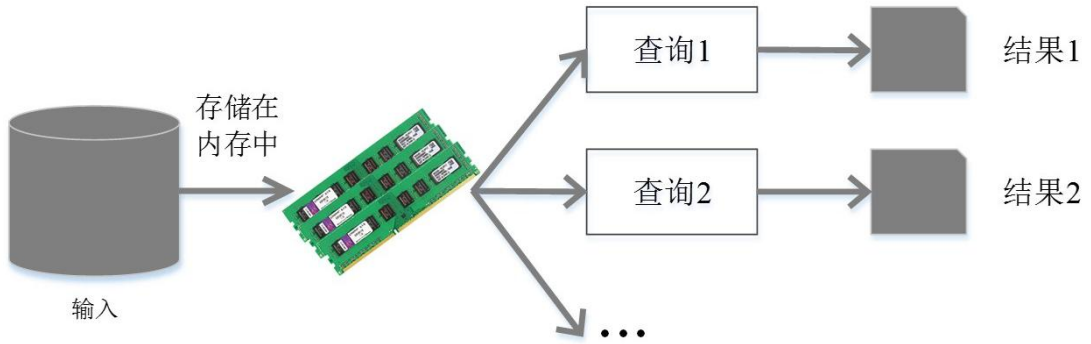
1.6.2 代表性大数据技术之Spark



(a) Hadoop MapReduce执行流程

1.6.2 代表性大数据技术之Spark





(b) Spark执行流程



1.6.2 代表性大数据技术之Spark

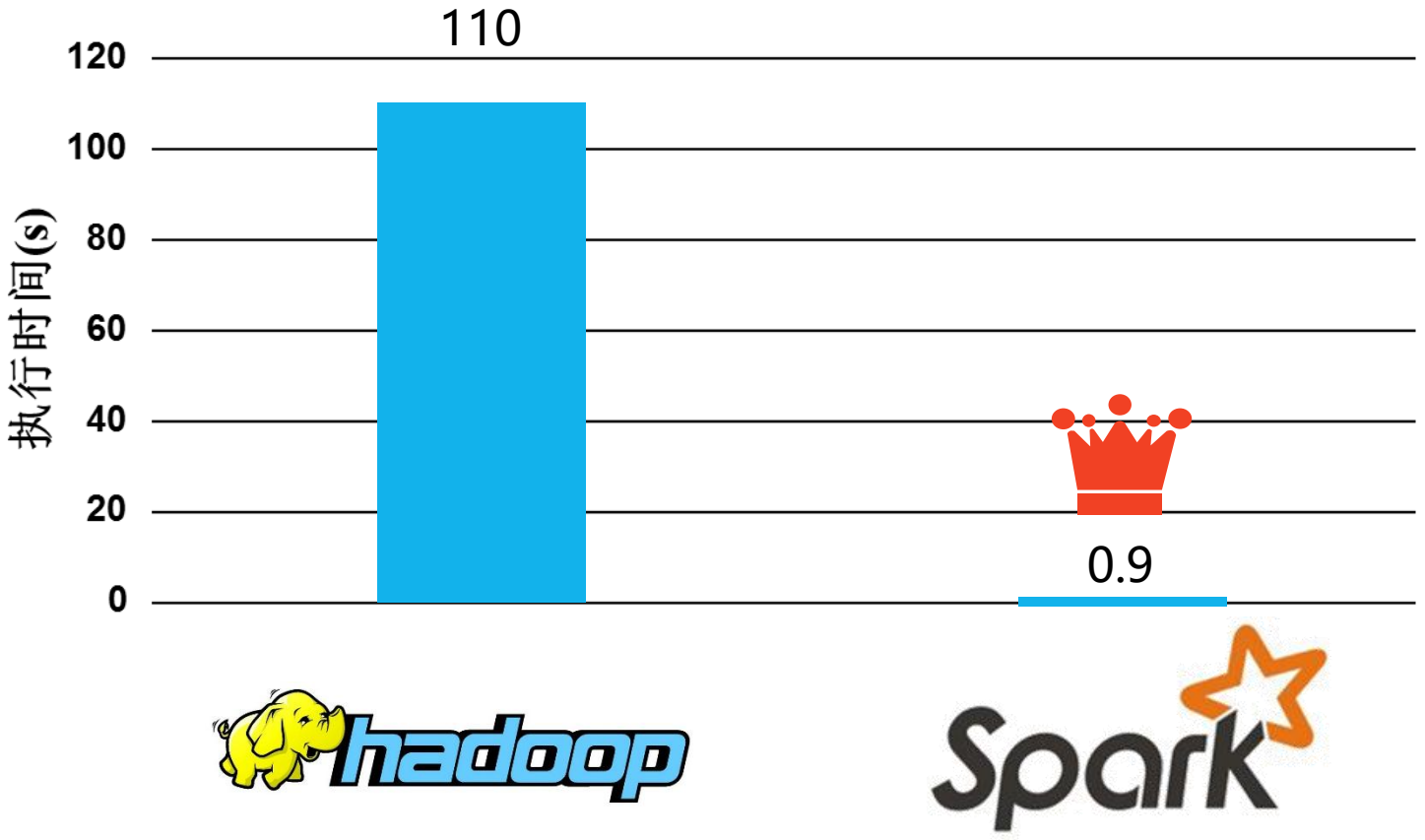


图 Hadoop和Spark执行逻辑斯蒂回归的实验对比



第1题

Apache Spark最初发布于哪一年？

A) 2007年
B) 2009年
C) 2012年
D) 2015年

1.6.2 代表性大数据技术之Spark



第2题
Spark架构中，以下哪个组件是核心基础组件？

A) Spark SQL
B) Spark Streaming
C) Spark Core
D) MLlib

1.6.2 代表性大数据技术之Spark



第3题
根据课程内容，开发Spark程序的首选编程语言是什么？

A) Java
B) Python
C) R
D) Scala

1.6.2 代表性大数据技术之Spark



第4题
Hadoop MapReduce的主要缺点不包括以下哪一项？

A) 表达能力有限
B) 磁盘IO开销大
C) 延迟高
D) 内存使用效率低

1.6.2 代表性大数据技术之Spark



第5题
Spark相比Hadoop
MapReduce在执行逻辑斯蒂回归算法时，性能提升了大约多少倍？

A) 10倍
B) 50倍
C) 100倍
D) 200倍

1.6.2 代表性大数据技术之Spark



第6题
Spark的哪个特性使其特别适合处理需要反复迭代的机器学习算法？

A) 多种编程语言支持
B) 内存计算
C) 图形化界面
D) 自动并行化

1.6.2 代表性大数据技术之Spark



第7题
DAG在Spark中的含义是什么？

A) 数据聚合图
B) 动态分配图
C) 有向无环图
D) 分布式应用图



1.6.2 代表性大数据技术之Spark



第8题
以下哪个组件负责Spark中的机器学习功能？

A) Spark SQL
B) Spark Streaming
C) MLlib
D) GraphX

1.6.2 代表性大数据技术之Spark



第9题
Spark会完全取代Hadoop吗？

A) 是的，Spark会完全取代Hadoop
B) 不会，Spark主要取代MapReduce，仍需要HDFS
C) 只在某些特定场景下取代
D) 两者完全不相关

1.6.2 代表性大数据技术之Spark



第10题
选择Scala作为Spark开发语言的主要原因不包括以下哪一项？

A) 代码简洁，表达能力强
B) 与Java兼容，运行在JVM上
C) 支持交互式编程
D) 拥有最大的开发者社区



Thank you