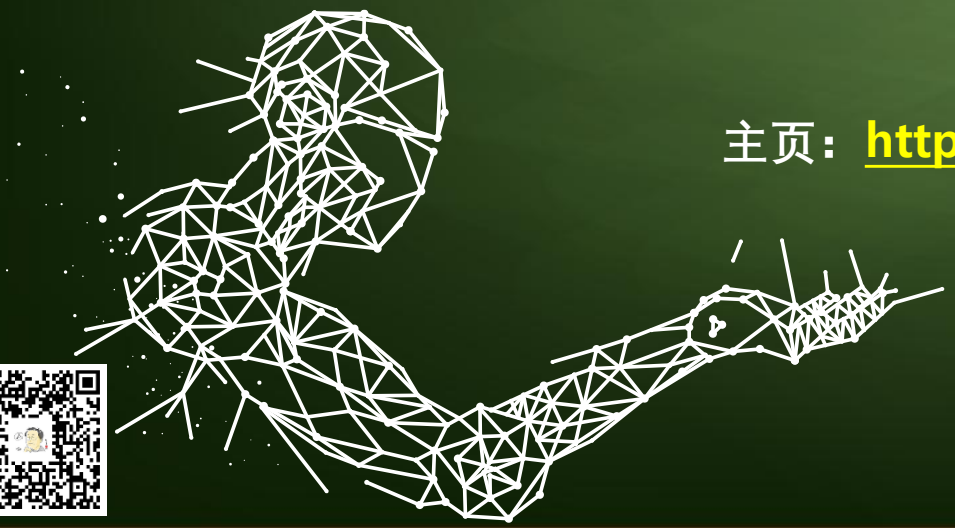
02

第2章 Spark的设计与运行原理

授课教师：[李阳源]
ywlee1212@gmail.com
主页：https://sites.google.com/view/dalguri





第2章 Spark的设计与运行原理



目录

CONTENT

- Spark概述
- Spark生态系统
- Spark运行架构
- Spark的部署方式



2.1 Spark 概述

2.1 Spark概述



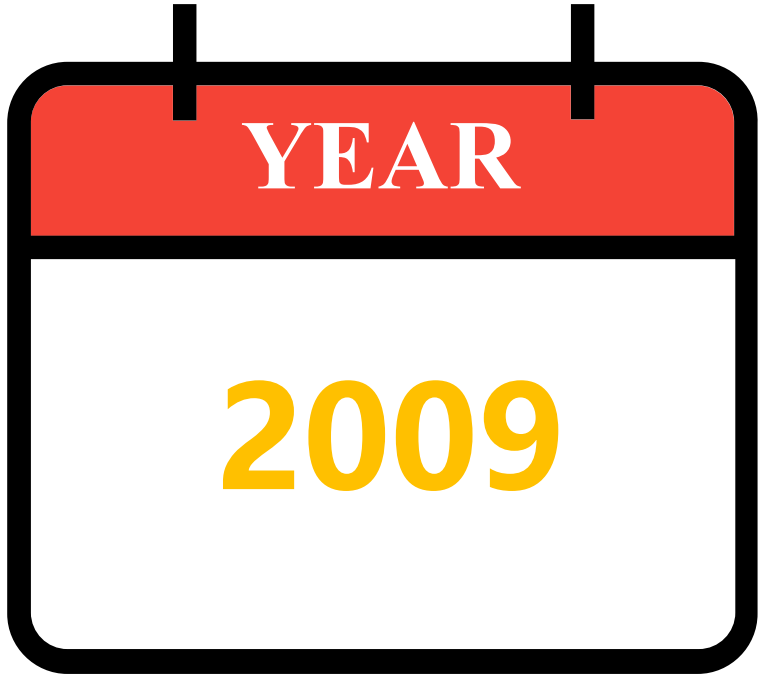
美国加州大学伯克利分校

2.1 Spark概述



AMP实验室

2.1 Spark概述



2.1 Spark概述



2.1 Spark概述





2.1 Spark概述





基于内存的计算框架

构建

构建

大型的

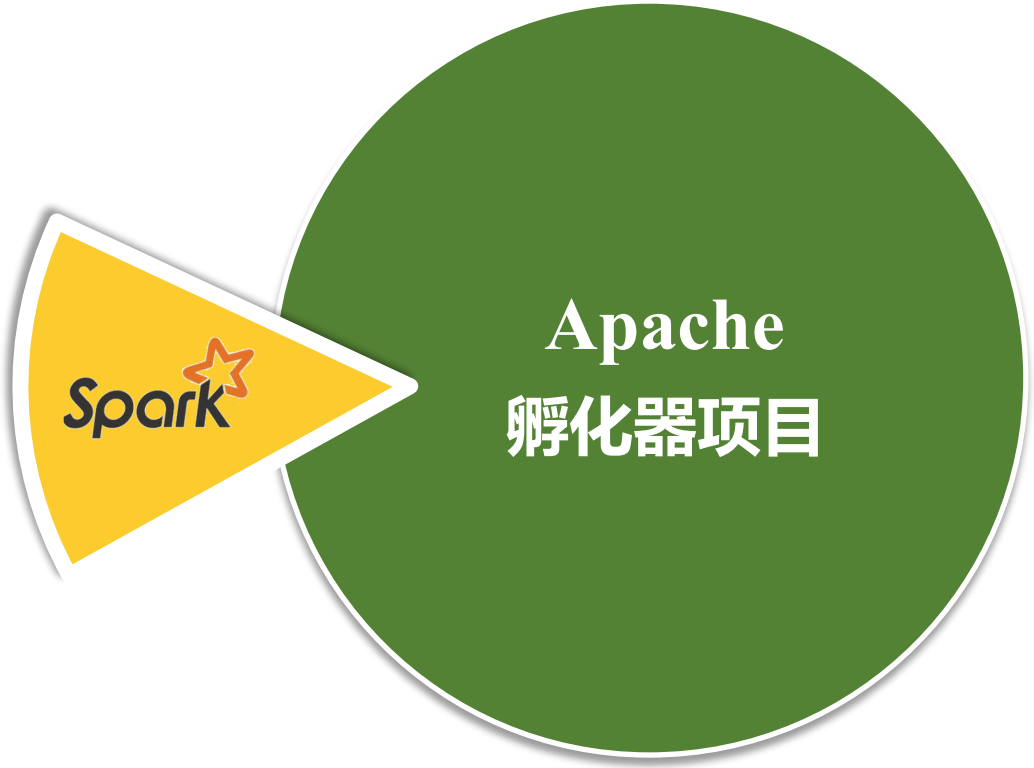
低延迟

数据分析应用程序

2.1 Spark概述



2.1 Spark概述




YEAR
2018

2.1 Spark概述


分布式
计算系统开源项目
hadoop
Spark
APACHE
STORM™
Distributed • Resilient • Real-time

2.1 Spark概述



2.1 Spark概述





基准排序纪录

2.1 Spark概述



Spark

hadoop

206个节点	2000个节点
23分钟	72分钟
100 TB数据	100 TB数据

$\frac{1}{10}$ 计算资源

2.1 Spark概述



Spark

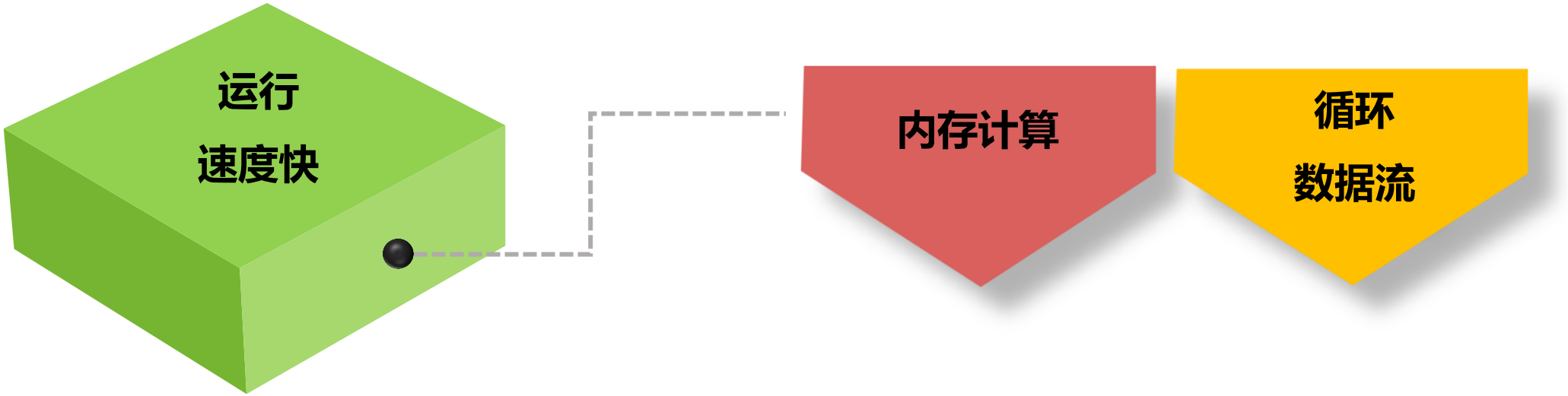
- 206个节点
- 23分钟
- 100 TB数据

hadoop

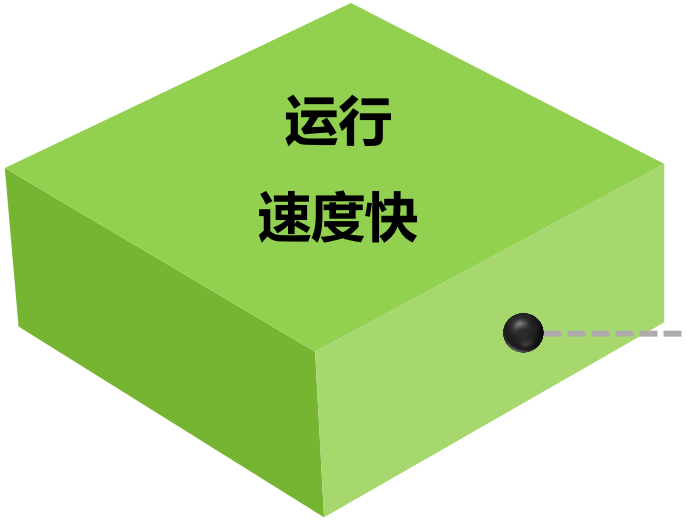
- 2000个节点
- 72分钟
- 100 TB数据

3倍速度

2.1 Spark概述

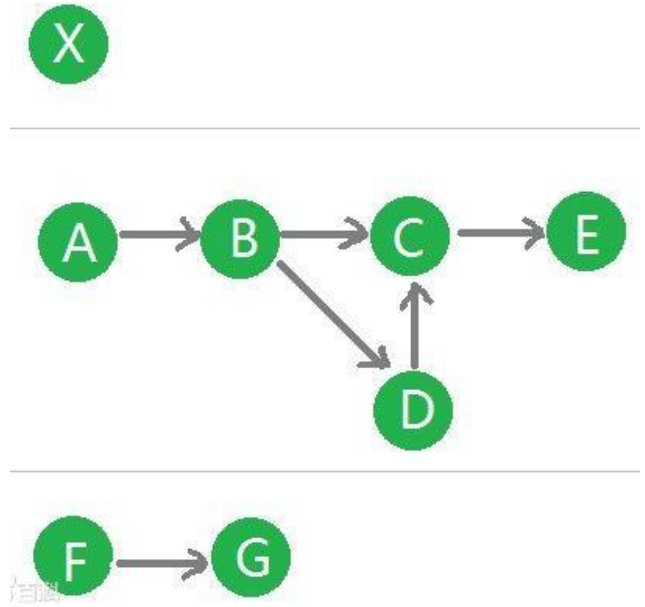


2.1 Spark概述



有向无环图设计机制

DAG有向无环图

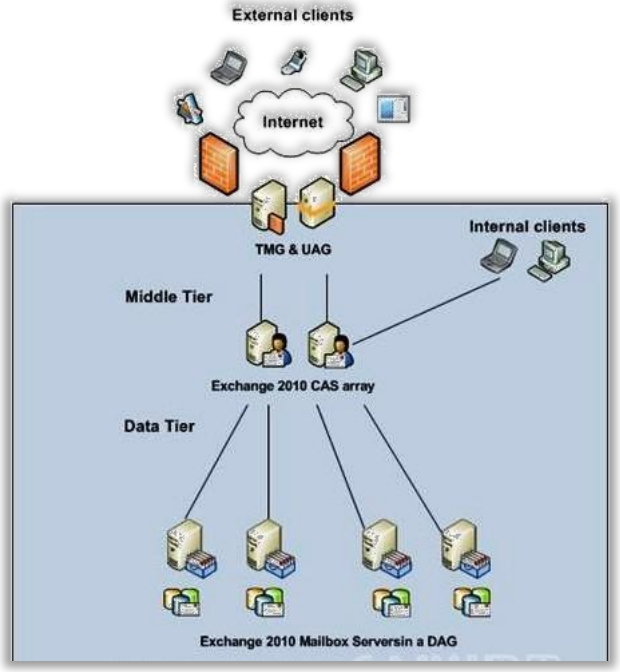


DAG图

2.1 Spark概述



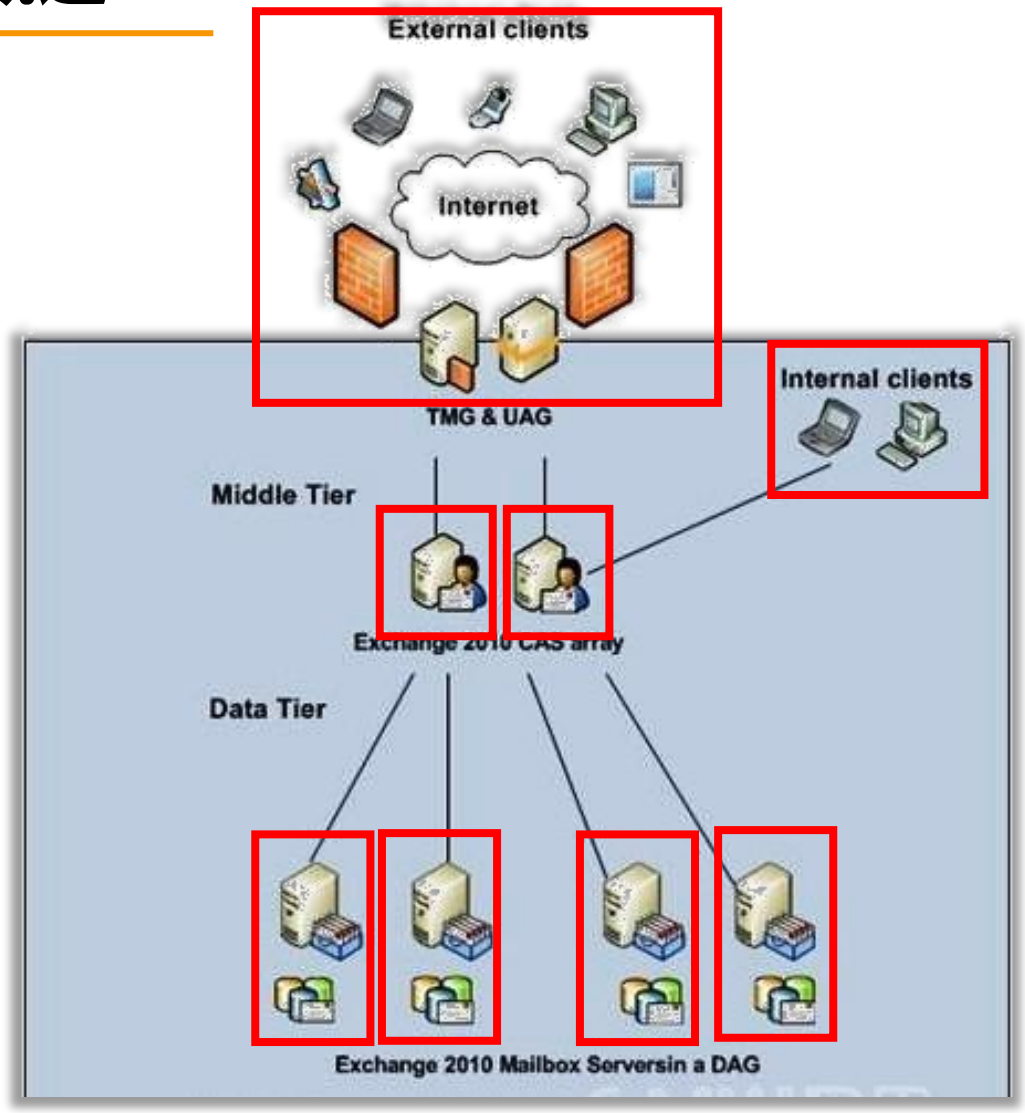
所有针对数据集的操作



有向无环图

2.1 Spark概述

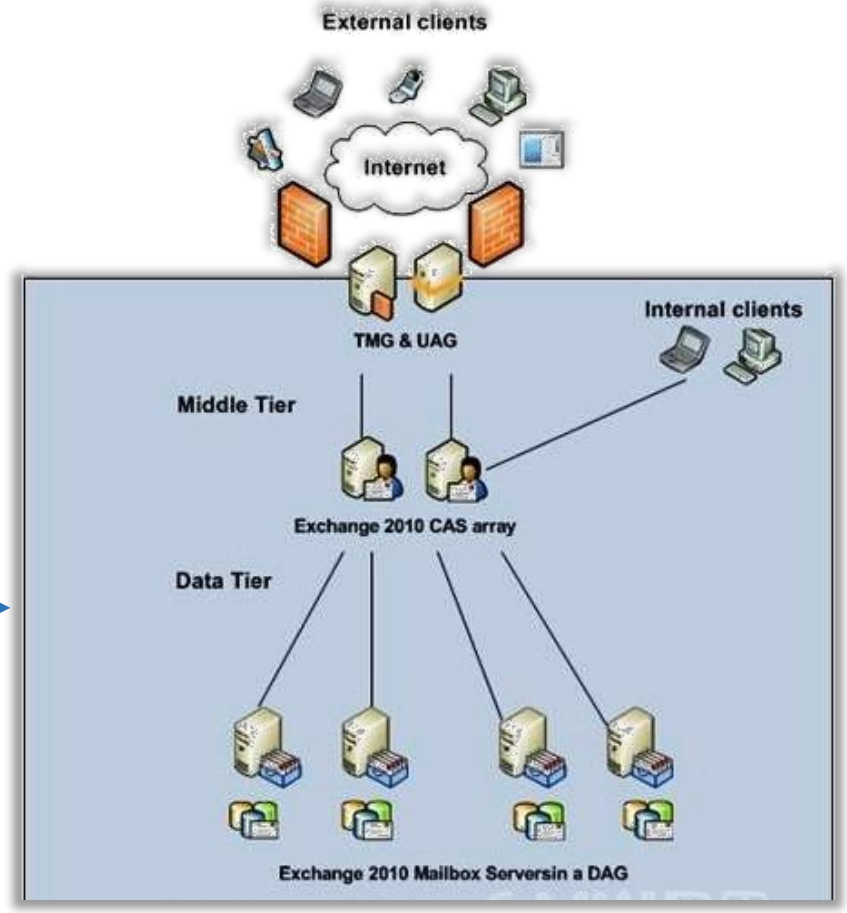
有向无环图



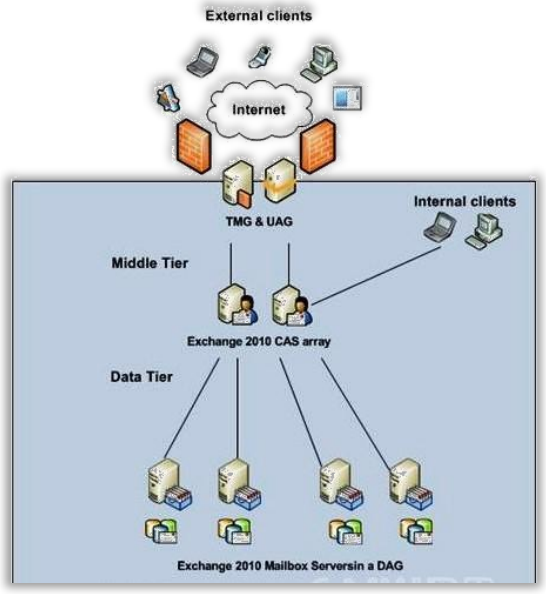
2.1 Spark概述



执行引擎调度 → **基于**



2.1 Spark概述



拆分

阶段

阶段

阶段

拆分

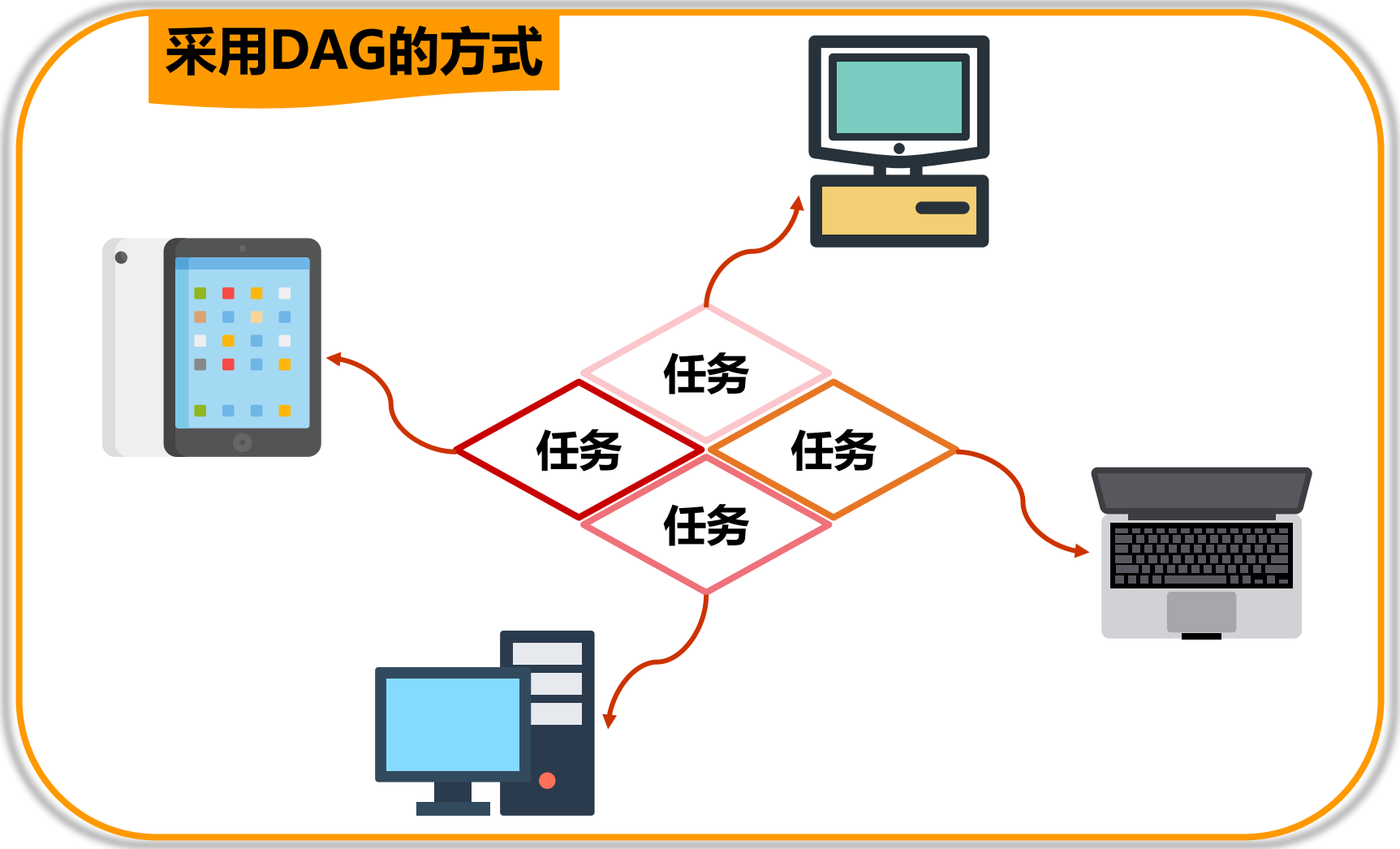
任务

任务

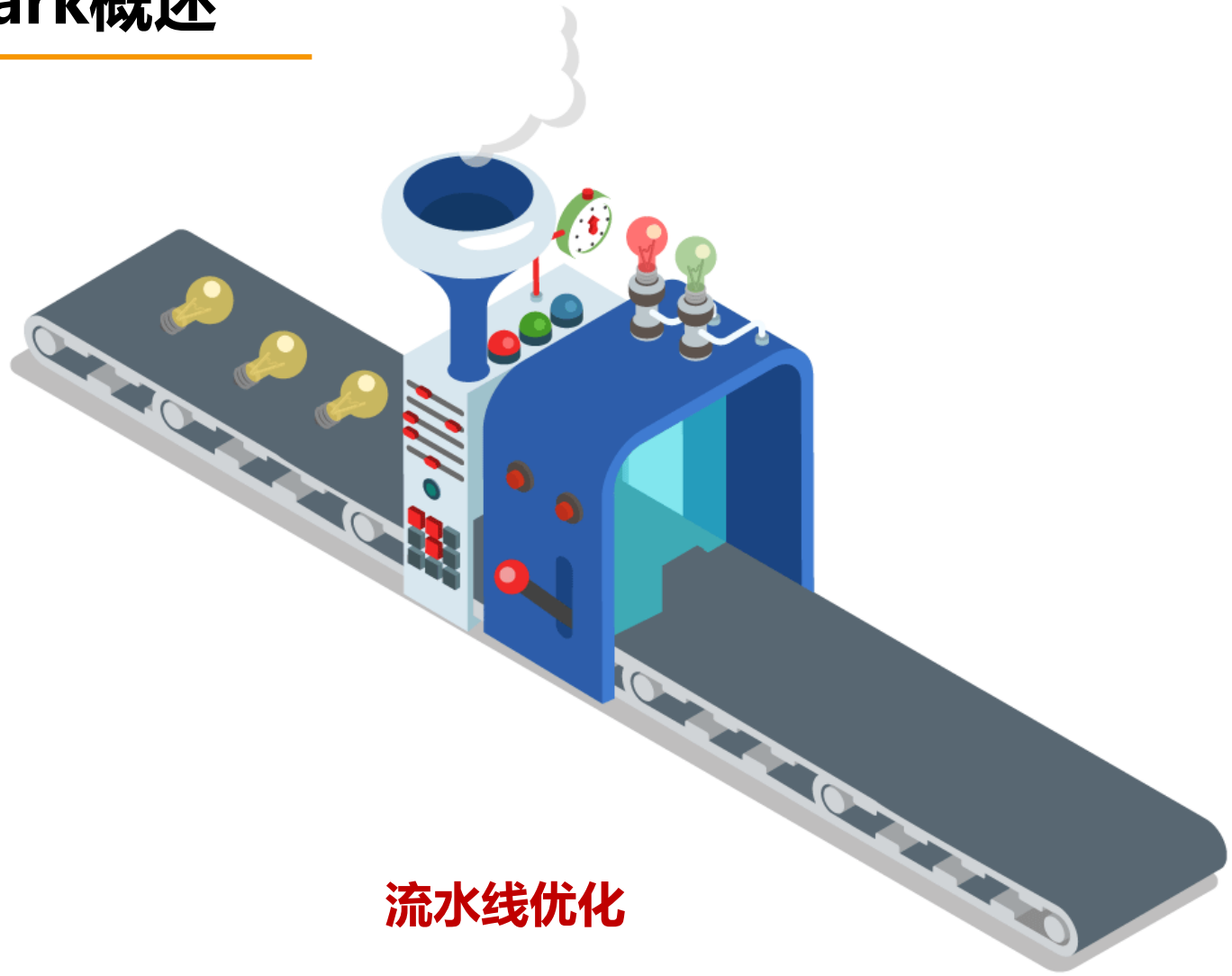
任务

任务

2.1 Spark概述

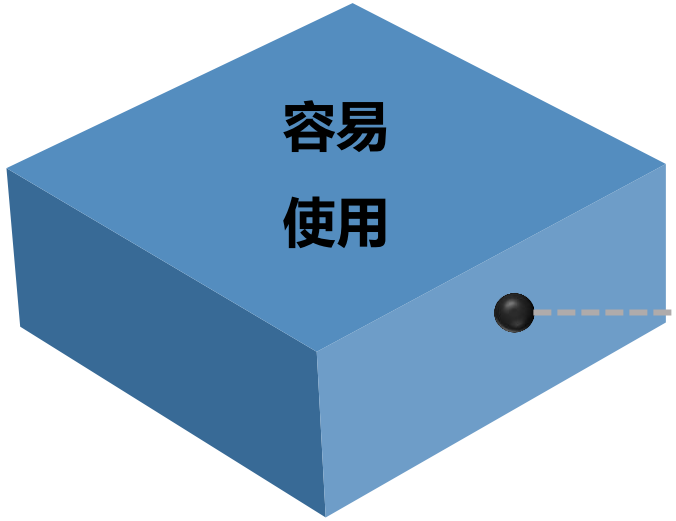


2.1 Spark概述



流水线优化

2.1 Spark概述



Java

R语言

Scala

Python

2.1 Spark概述





2.1 Spark概述







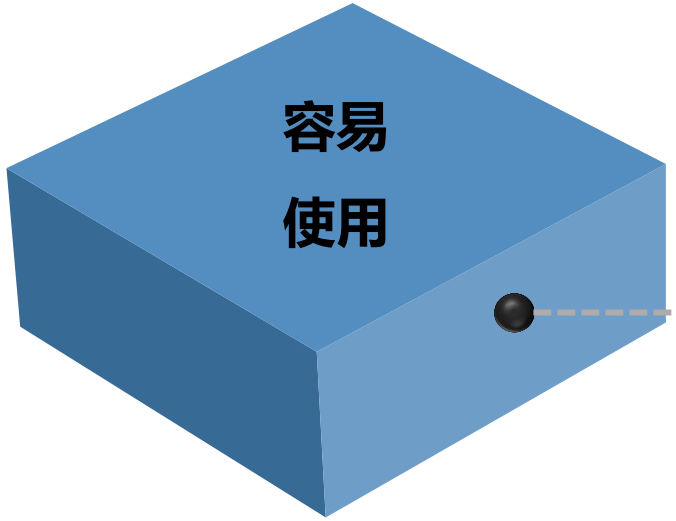
语法琐碎



代码冗长

编写Spark程序较繁琐

2.1 Spark概述



Java

R语言

Scala

Python

可通过Spark Shell进行交互式编程

2.1 Spark概述



2025

1. Python
2. C++
3. Java
4. C
5. C#



2025年语言排榜位居第一

- ✅ 数据挖掘
- ✅ 语法简单，易入门
- ❌ 并发性不好
- ❌ 用于小规模数据分析
- ❌ 不适合企业级大规模数据分析程序





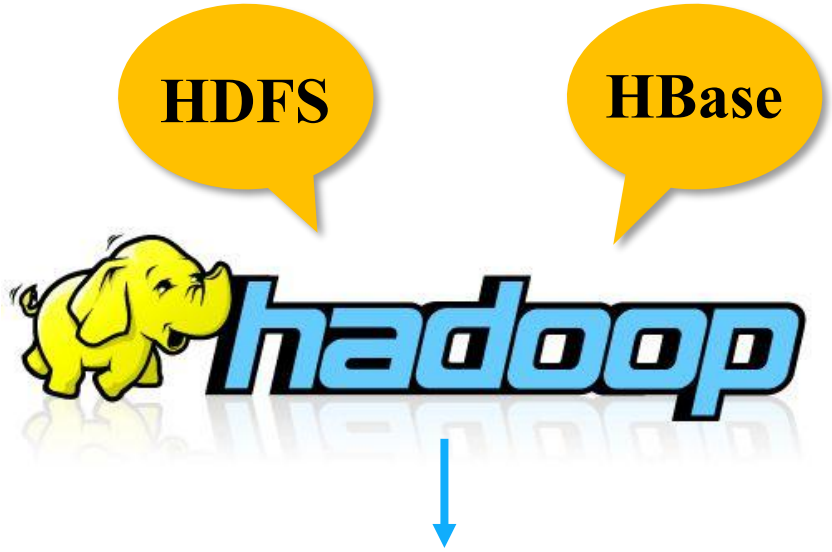
Hadoop平台



2.1 Spark概述

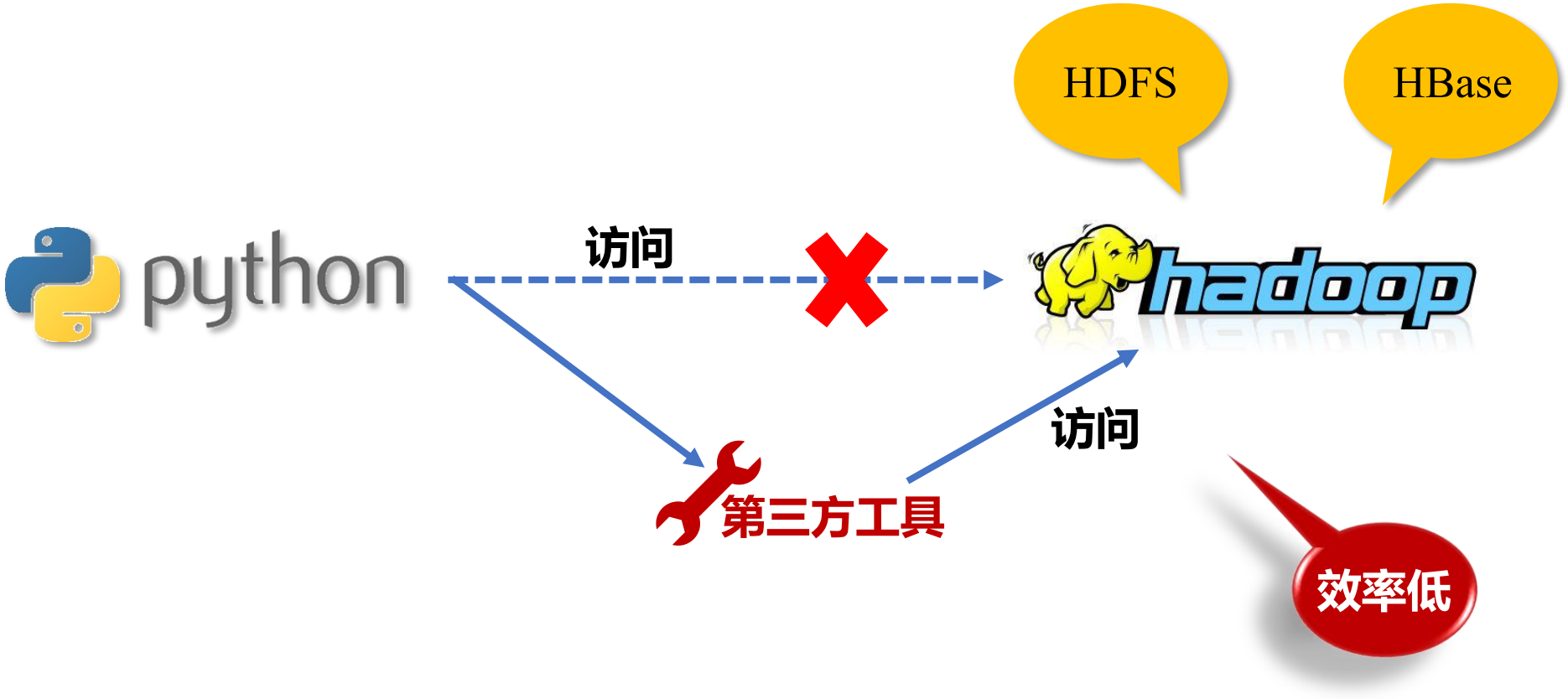


解决计算问题



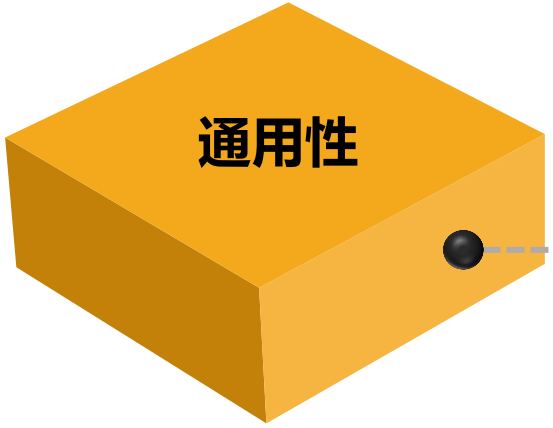
解决存储问题

2.1 Spark概述




Scala
python
开发
编写
Spark
Spark应用程序

2.1 Spark概述



完整的解决方案

技术软件栈

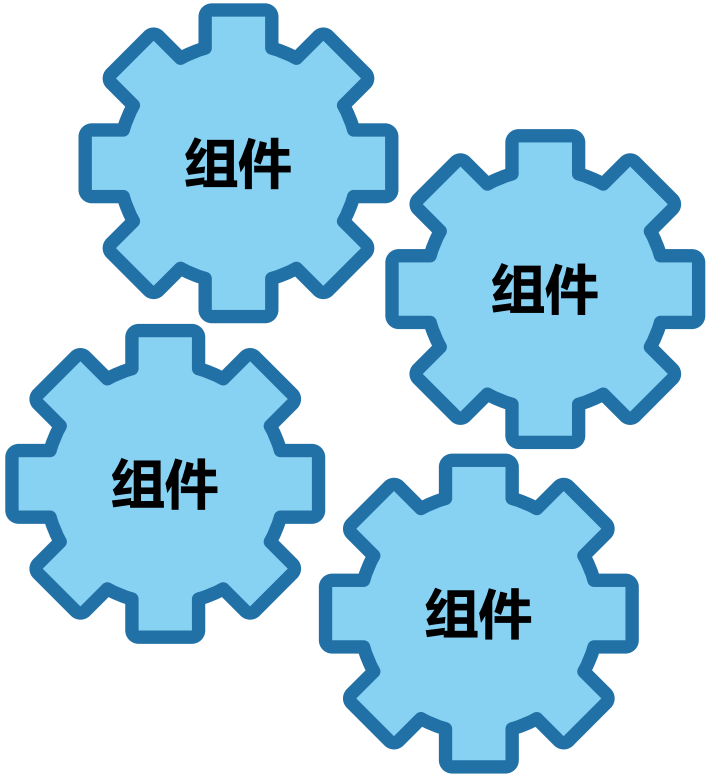




不是一个单一的软件

2.1 Spark概述





2.1 Spark概述



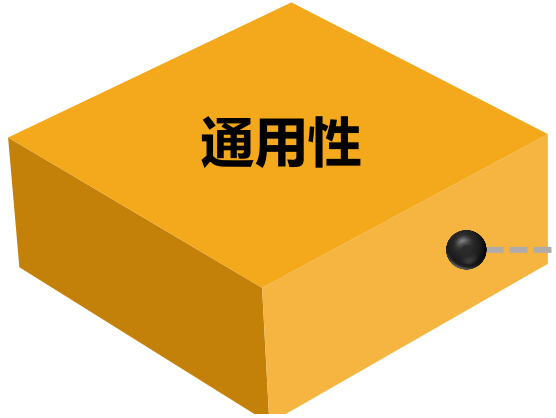
生态系统



生态系统



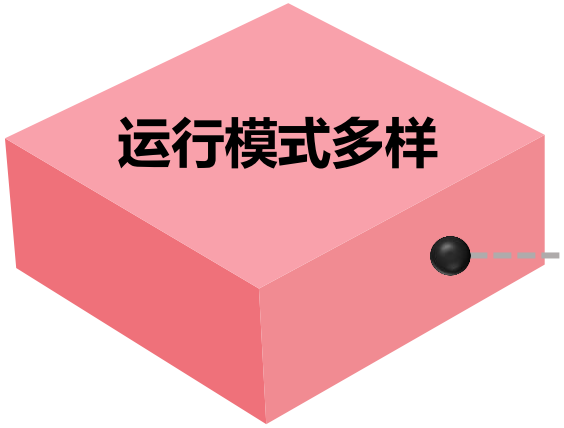
2.1 Spark概述



- SQL查询 → Spark SQL
- 流式计算 → Spark Streaming
- 机器学习 → Spark MLlib
- 图算法组件 → Spark的GraphX



2.1 Spark概述



运行于

独立的
集群模式



2.1 Spark概述



资源管理框架YARN

资源调度管理器

2.1 Spark概述

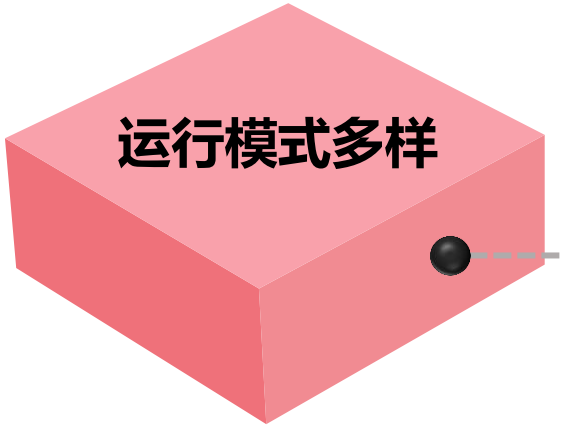


资源管理框架YARN

资源调度管理器

Spark

提供集群资源的调度管理服务



运行于

独立的
集群模式

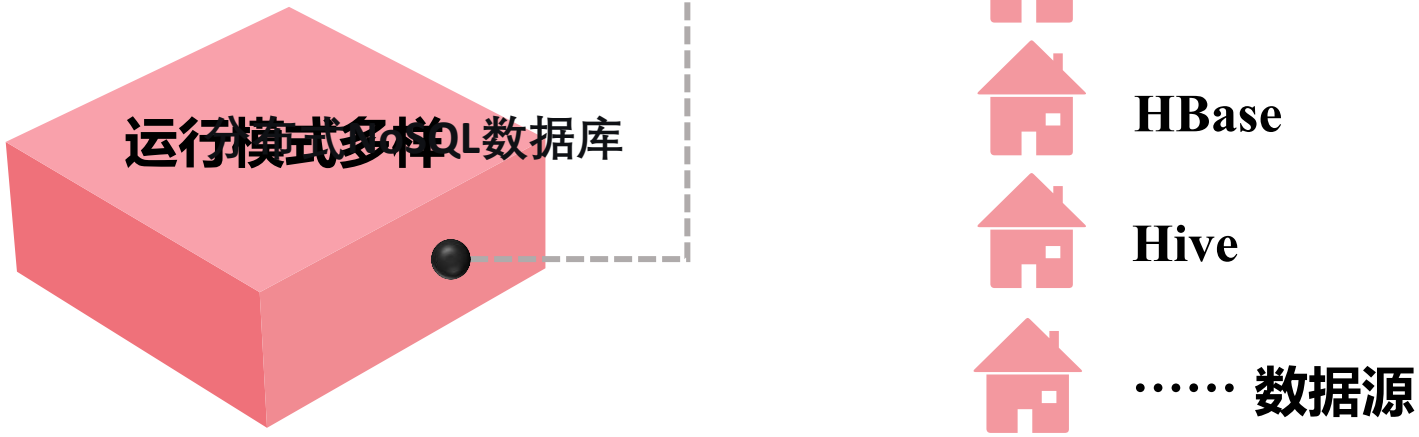
阿里云

腾讯云

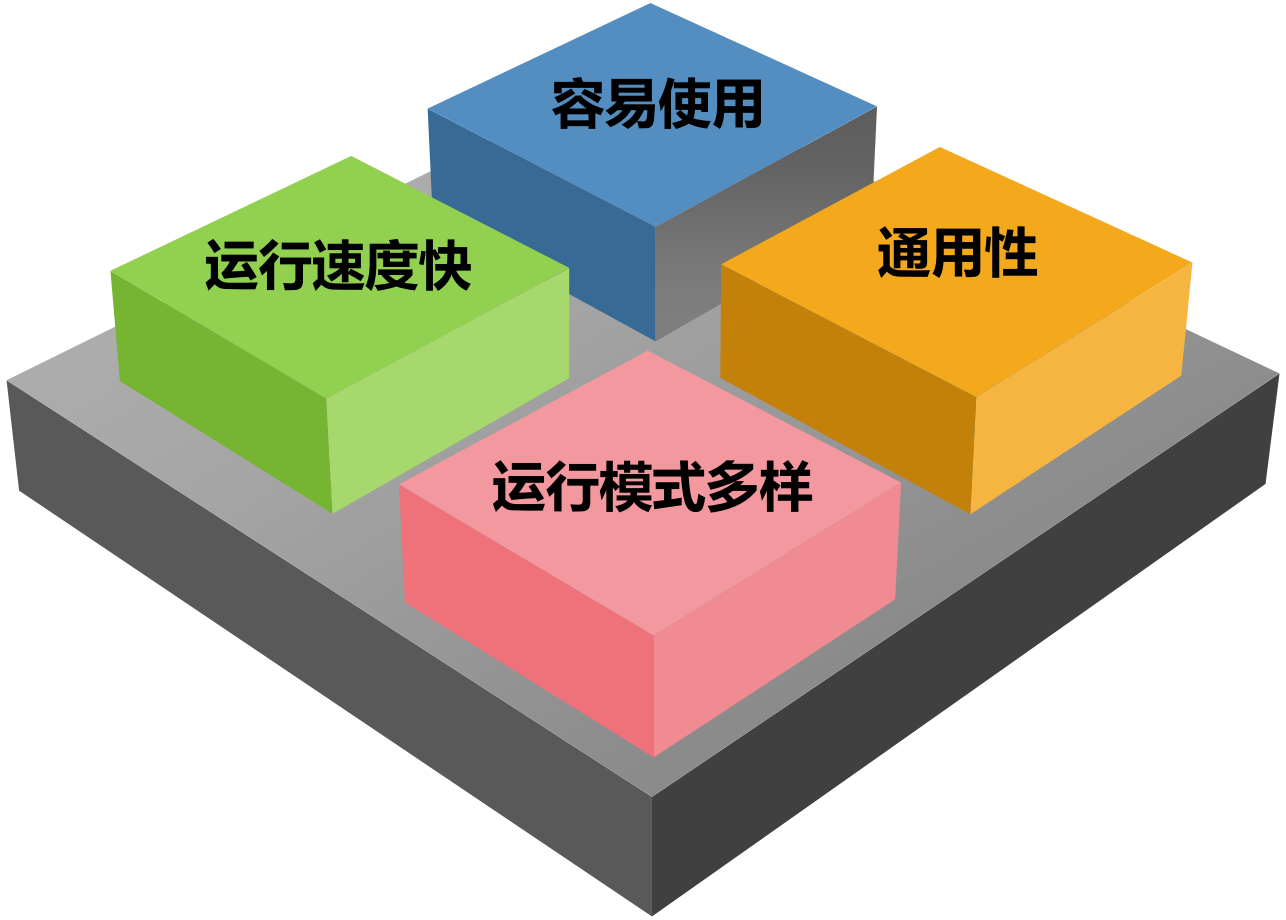
amazon
web services™
EC2

Elastic Compute Cloud (EC2)

2.1 Spark概述



2.1 Spark概述



2.1 Spark概述

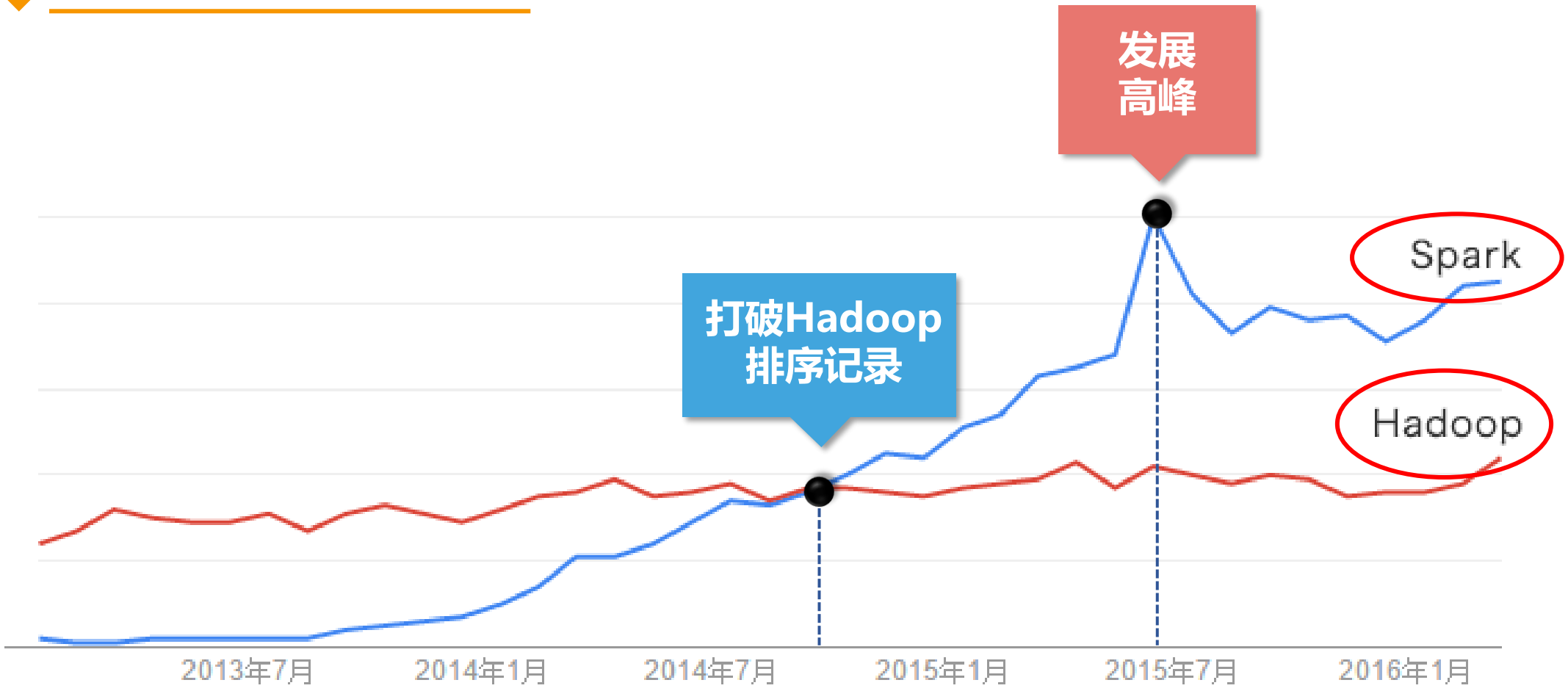


图 谷歌趋势：Spark与Hadoop对比

2.1 Spark概述











2.1 Spark概述



VS



2.1 Spark概述



MapReduce ✗

HDFS ✓

2.1 Spark概述



表达能力有限



Map函数

复杂业务逻辑

Reduce函数



降低了分布式应用开发的复杂性

2.1 Spark概述



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

2.1 Spark概述

磁盘IO开销大

MapReduce

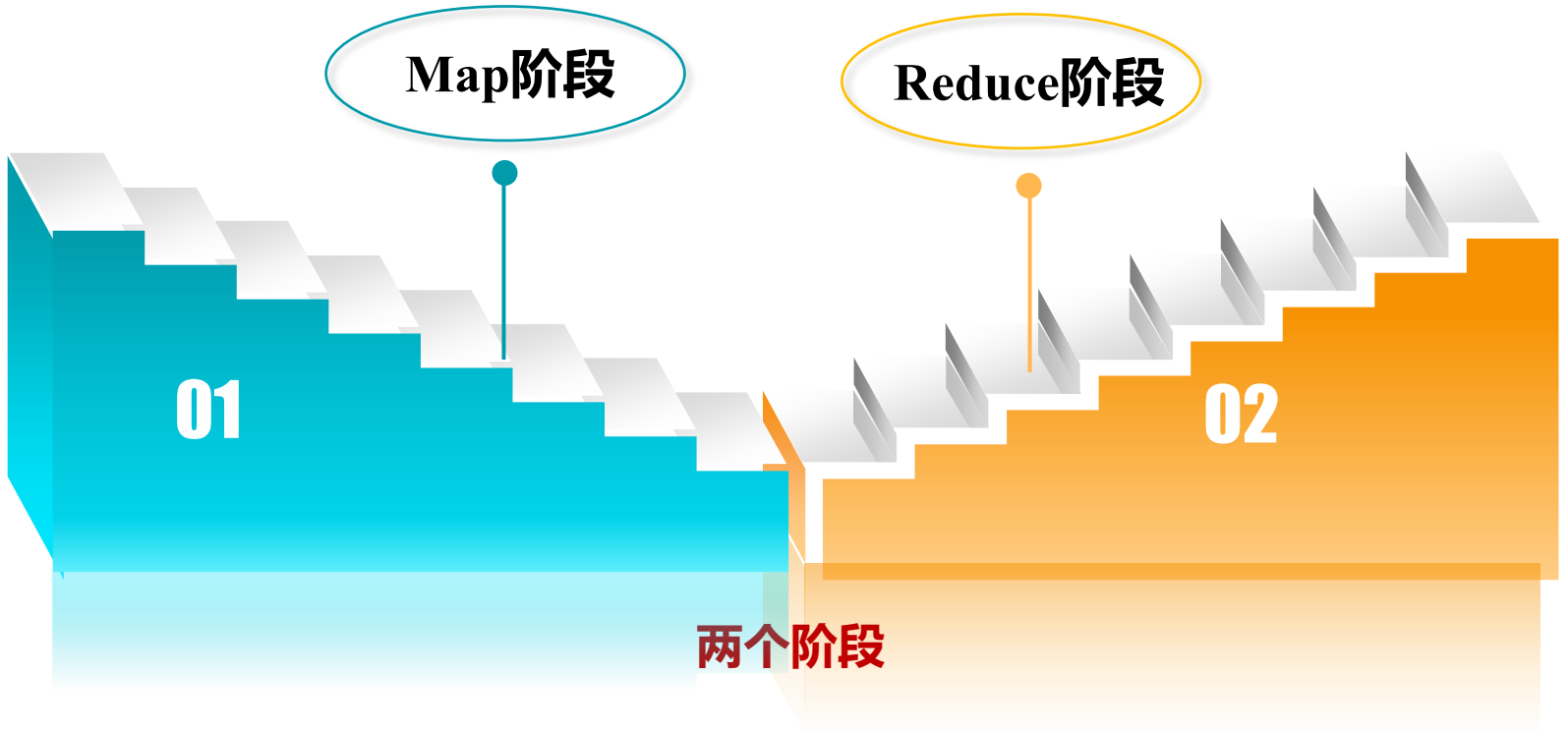




2.1 Spark概述



延迟高



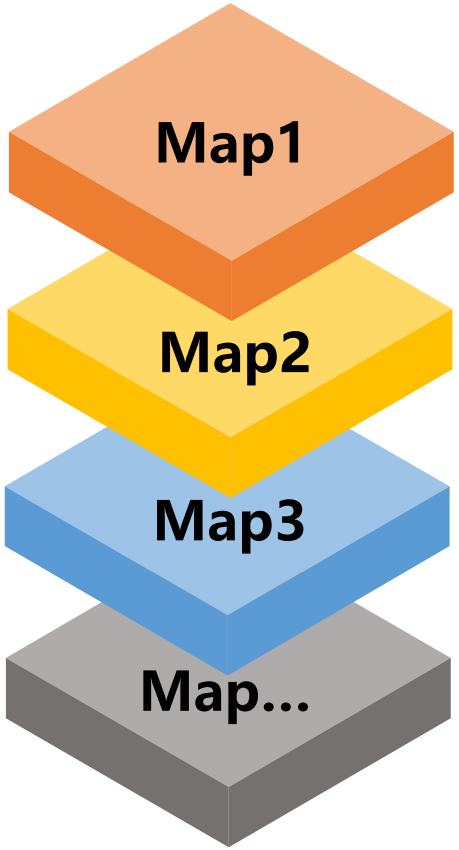
2.1 Spark概述







2.1 Spark概述



并行处理

Reduce1

Reduce2

Reduce3

Reduce...

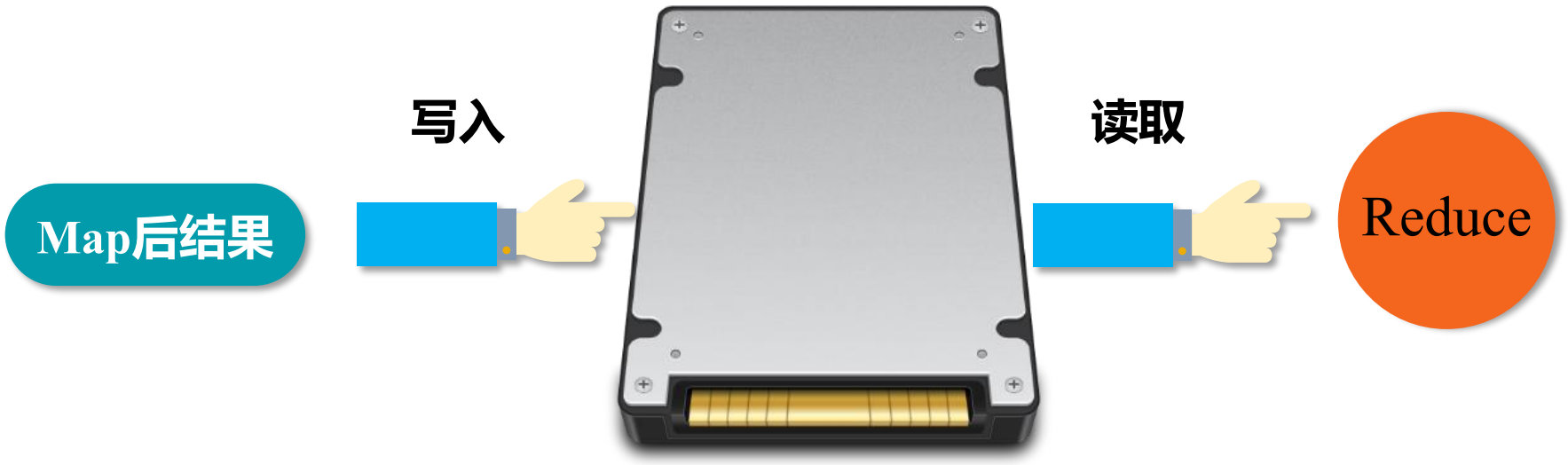
2.1 Spark概述


Map1
Map2
Map3
Map...
输出
中间结果
处理
完成
全部Map
等待
Reduce1
Reduce2
Reduce3
Reduce...

2.1 Spark概述



衔接的IO开销大





延迟高



磁盘的IO开销大

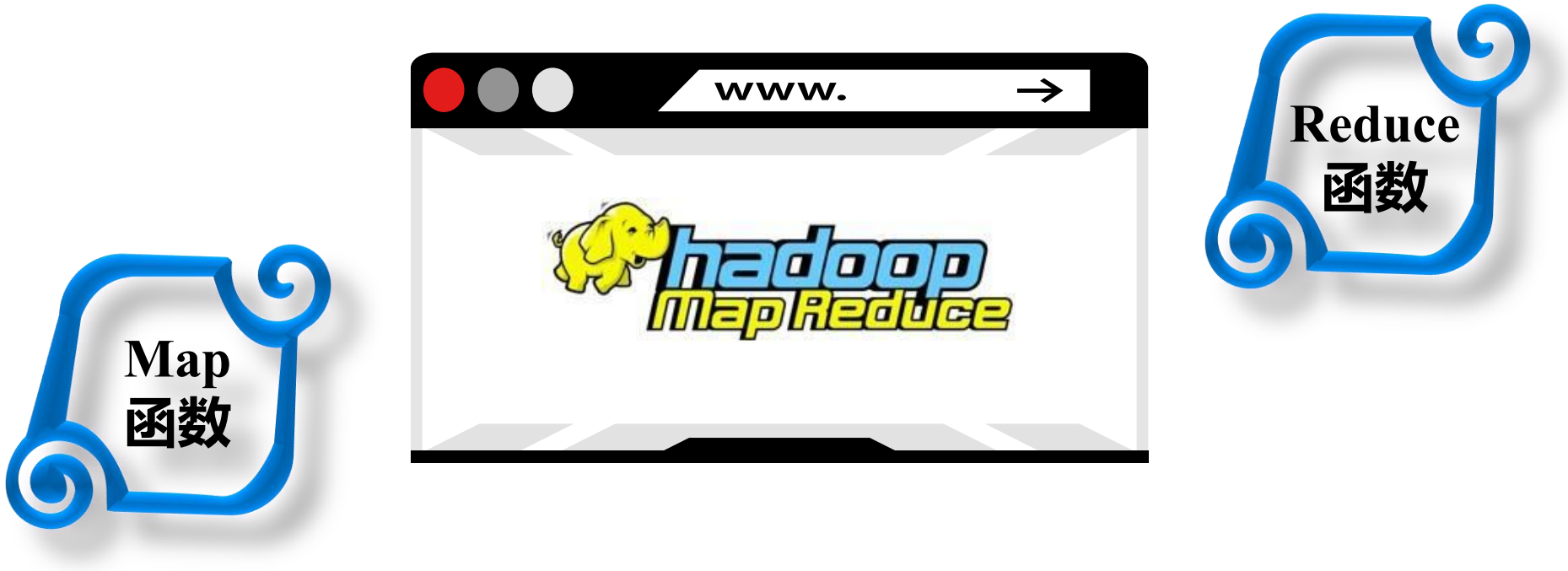
衔接的IO开销大

MapReduce的缺陷

2.1 Spark概述



Spark对Hadoop进行缺陷改进与继承


Map
函数
www.
hadoop
MapReduce
Reduce
函数

2.1 Spark概述









操作类型更多

表达能力更强

2.1 Spark概述



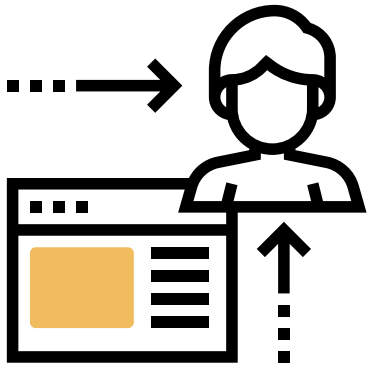
内存计算







内存计算



再次访问





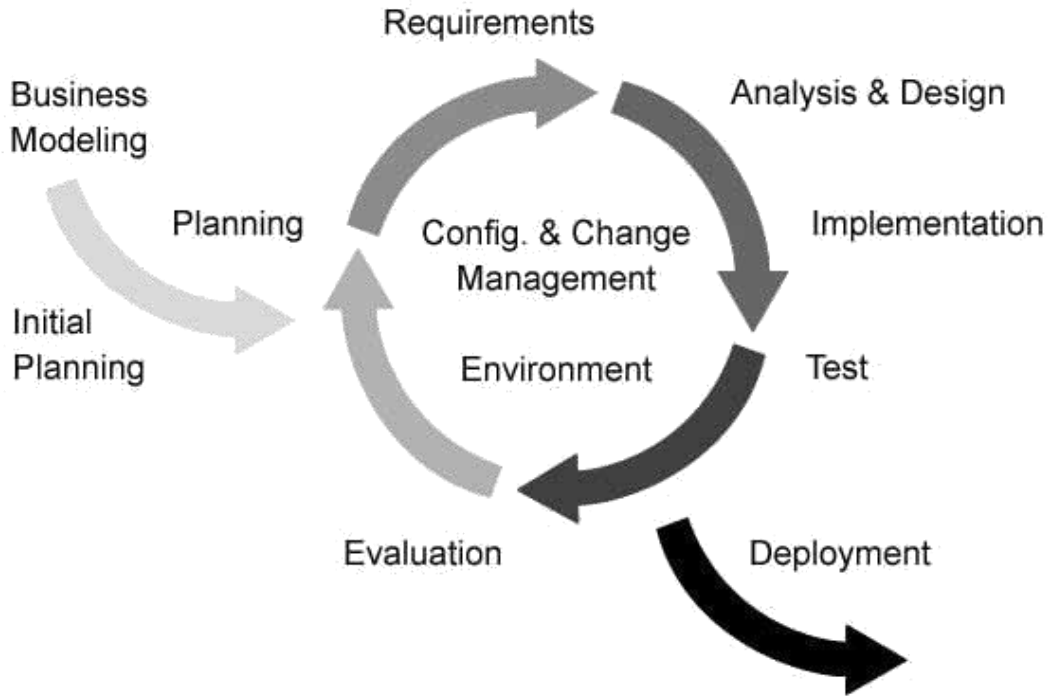
低延迟

2.1 Spark概述



运行效率提升

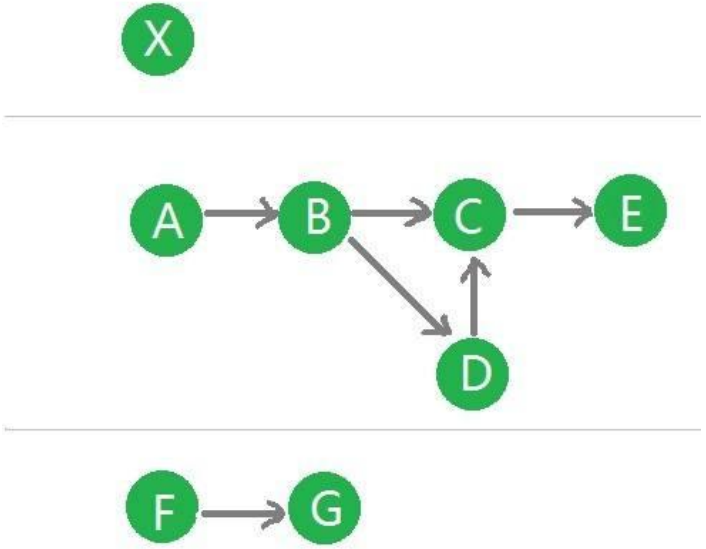
高效提高迭代运算



2.1 Spark概述



DAG任务调度执行机制



DAG—有向无环图



使用MapReduce进行反复迭代

迭代1



结果1





使用MapReduce进行反复迭代

迭代1





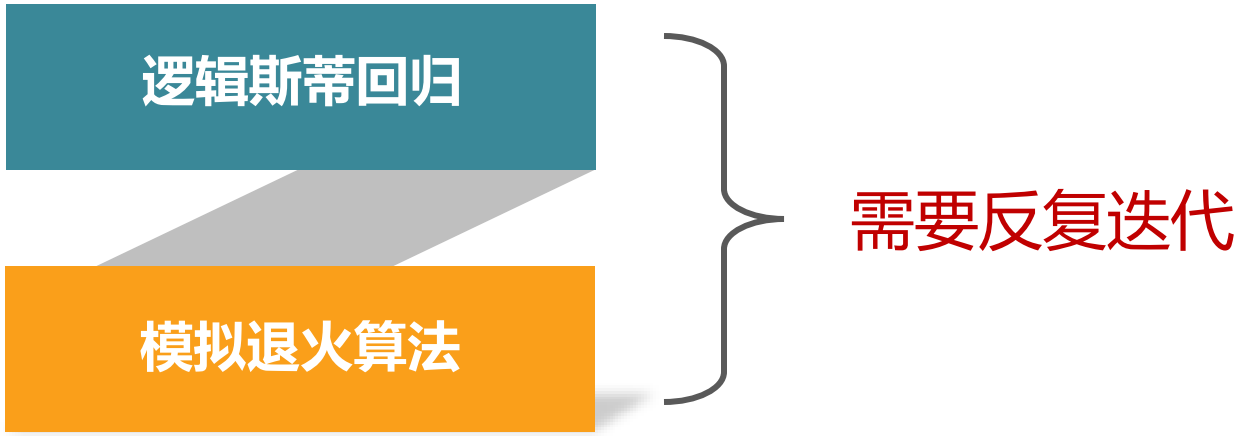
迭代2



结果2

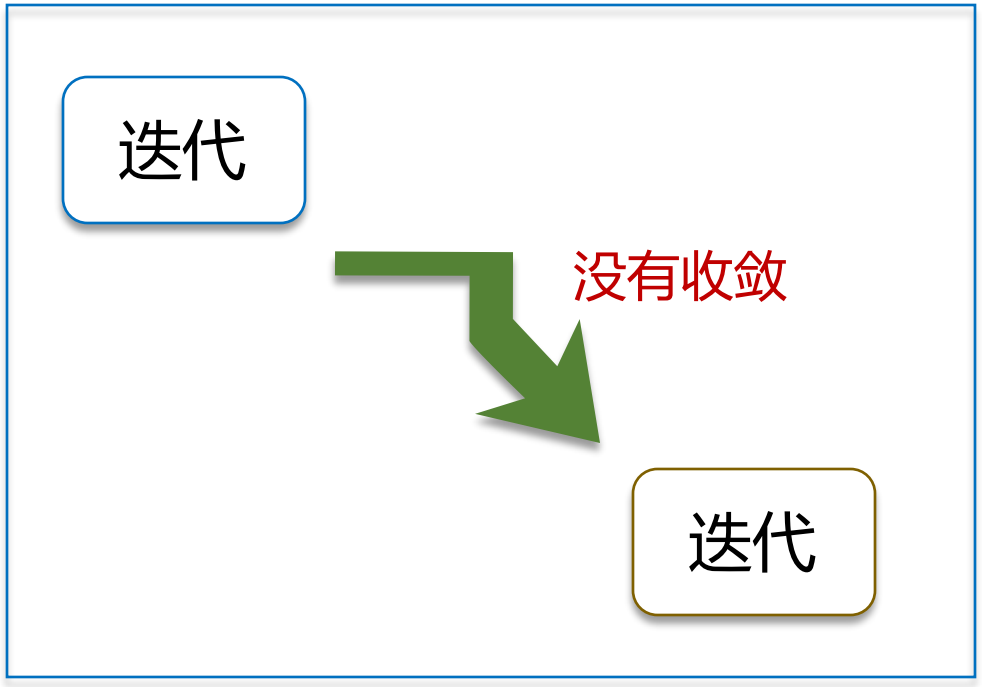


使用MapReduce进行反复迭代



2.1 Spark概述

使用MapReduce进行反复迭代



实现





2.1 Spark概述

使用MapReduce进行反复迭代

迭代1

Map + Reduce

结果1





使用MapReduce进行反复迭代

迭代1





迭代2



结果2

2.1 Spark概述



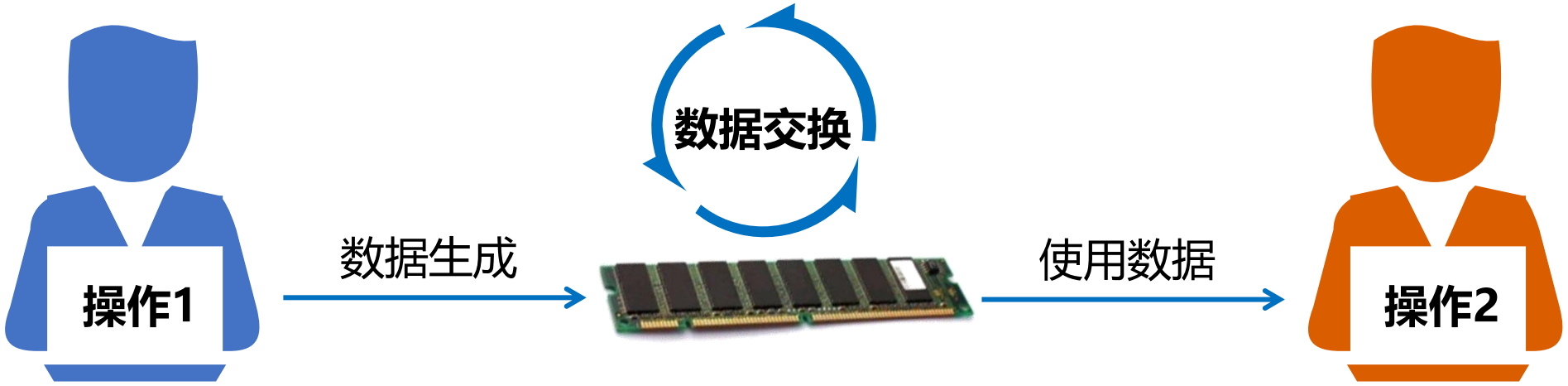
避免数据落地



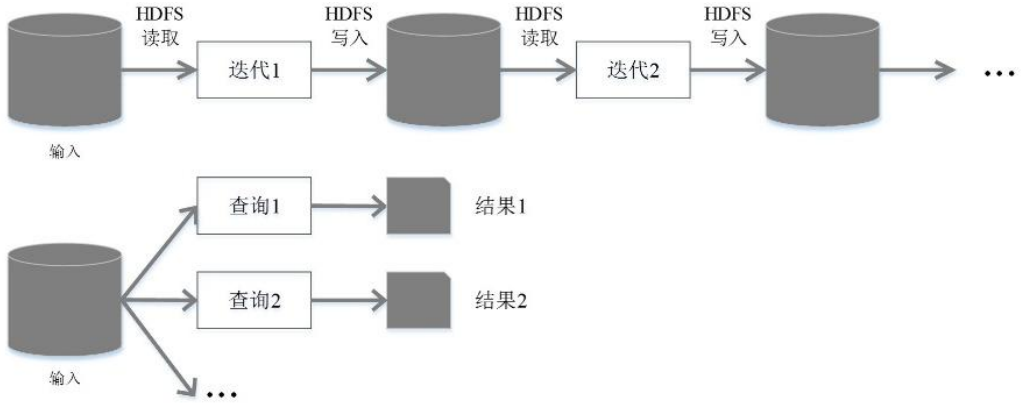




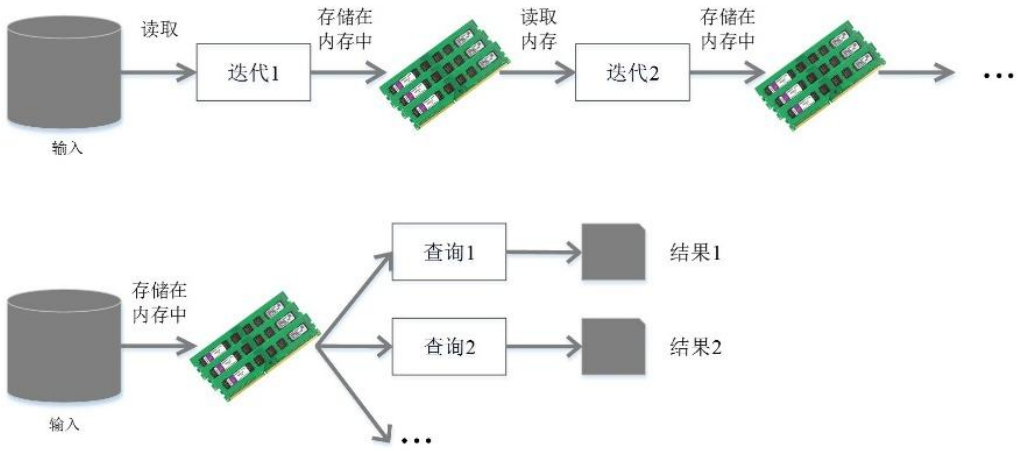
流水线优化



2.1 Spark概述

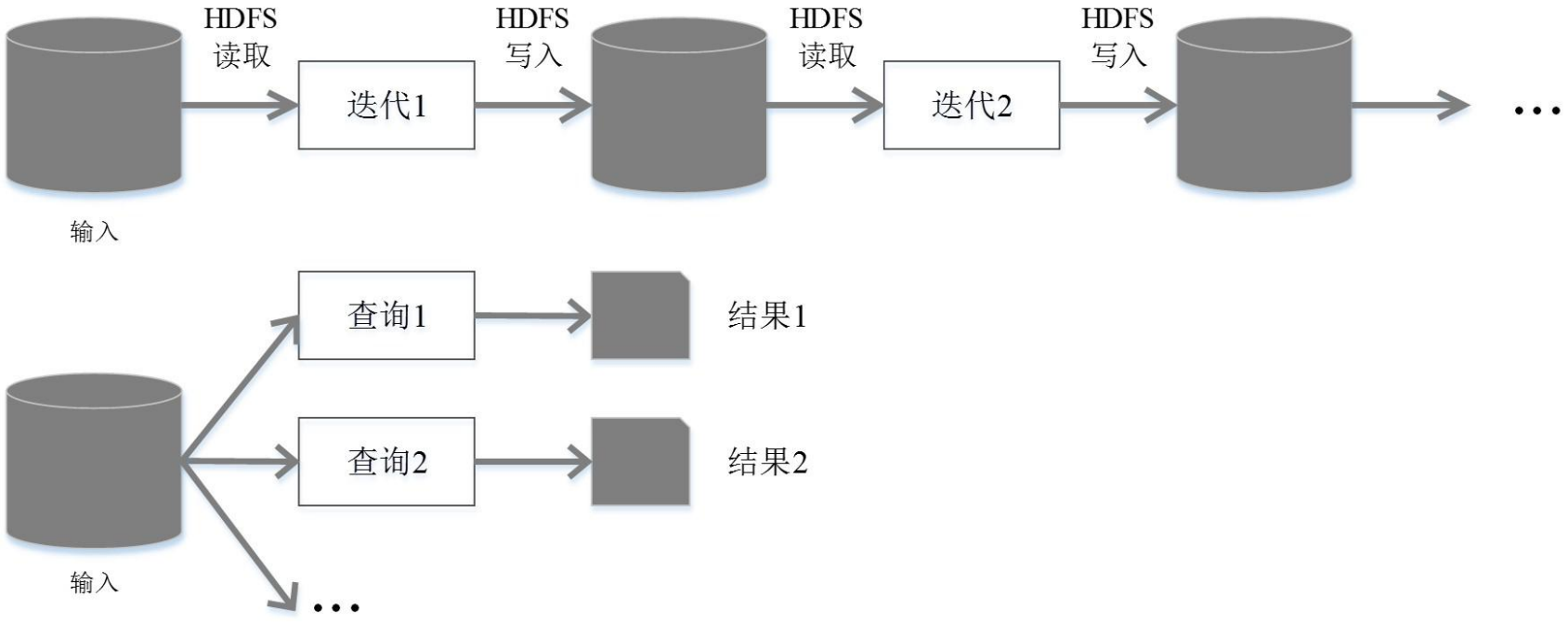


(a) Hadoop MapReduce执行流程



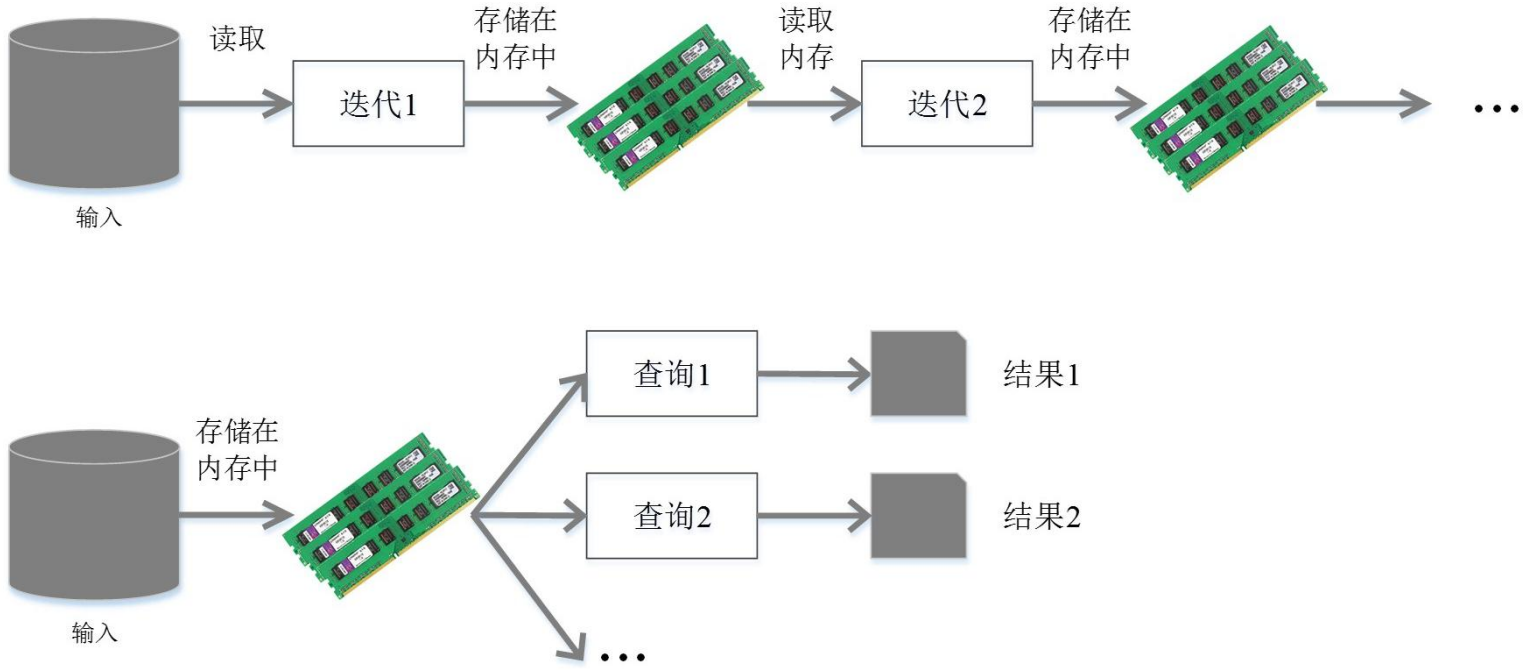
(b) Spark执行流程

2.1 Spark概述



(a) Hadoop MapReduce执行流程

2.1 Spark概述



(b) Spark执行流程

2.1 Spark概述

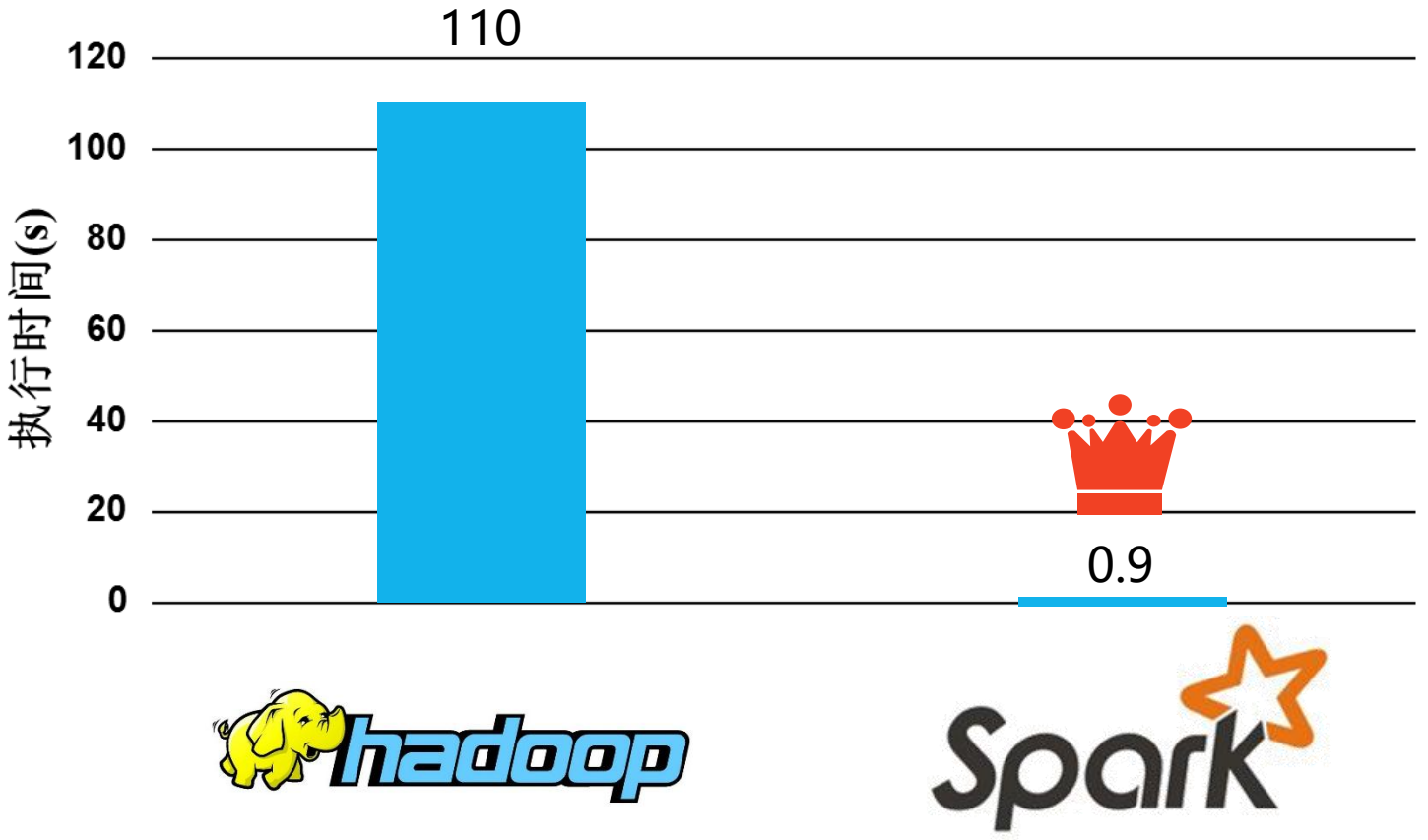


图 Hadoop和Spark执行逻辑斯蒂回归的实验对比



2.1 Spark概述理解测验（选择题）

题目1．Spark项目最初是在哪个大学的实验室诞生的？

A) 斯坦福大学
B) 麻省理工学院
C) 加州大学伯克利分校
D) 卡内基梅隆大学



2.1 Spark概述理解测验（选择题）



题目2．Spark项目开始的年份是？

A)2008年
B) 2009年
C) 2010年
D) 2011年



2.1 Spark概述理解测验（选择题）



题目3．AMP实验室的三个核心要素是什么？

A) Algorithm, Memory, Performance
B) Algorithm, Machine, People
C) Application, Machine, Processing
D) Analysis, Method, Platform



2.1 Spark概述理解测验（选择题）



题目4．Spark相比MapReduce在性能方面取得的突破性成就是？

A) 在2013年打破了Hadoop的排序纪录
B) 在2014年打破了Hadoop的排序纪录
C) 在2015年打破了Hadoop的排序纪录
D) 在2016年打破了Hadoop的排序纪录



2.1 Spark概述理解测验（选择题）



题目5．根据课程内容，Spark在逻辑斯蒂回归实验中的性能表现是？

A) 比Hadoop快10倍
B) 比Hadoop快50倍
C) 比Hadoop快100倍以上
D) 与Hadoop性能相当



2.1 Spark概述理解测验（选择题）

题目6．Spark的核心技术DAG指的是什么？

A) Data Analysis Graph
B) Directed Acyclic Graph
C) Dynamic Algorithm Graph
D) Distributed Application Graph



2.1 Spark概述理解测验（选择题）



题目7．Spark支持的编程语言不包括以下哪一种？

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



题目8．MapReduce模型的主要缺陷不包括？

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



题目9．Spark生态系统的四个主要组件是？

A) Spark Core, Spark SQL, Spark Streaming, Spark MLlib
B) Spark SQL, Spark Streaming, Spark MLlib, Spark GraphX
C) Spark Core, Spark SQL, Spark MLlib, Spark GraphX
D) Spark SQL, Spark Streaming, Spark GraphX, Spark Shell



2.1 Spark概述理解测验（选择题）

题目10．关于Spark和Hadoop的关系，以下说法正确的是？

A)Spark完全替代了Hadoop
B) Spark主要解决计算问题，Hadoop主要解决存储问题
C) Spark和Hadoop是竞争关系，不能同时使用
D) Spark只能在Hadoop平台上运行



Thank you