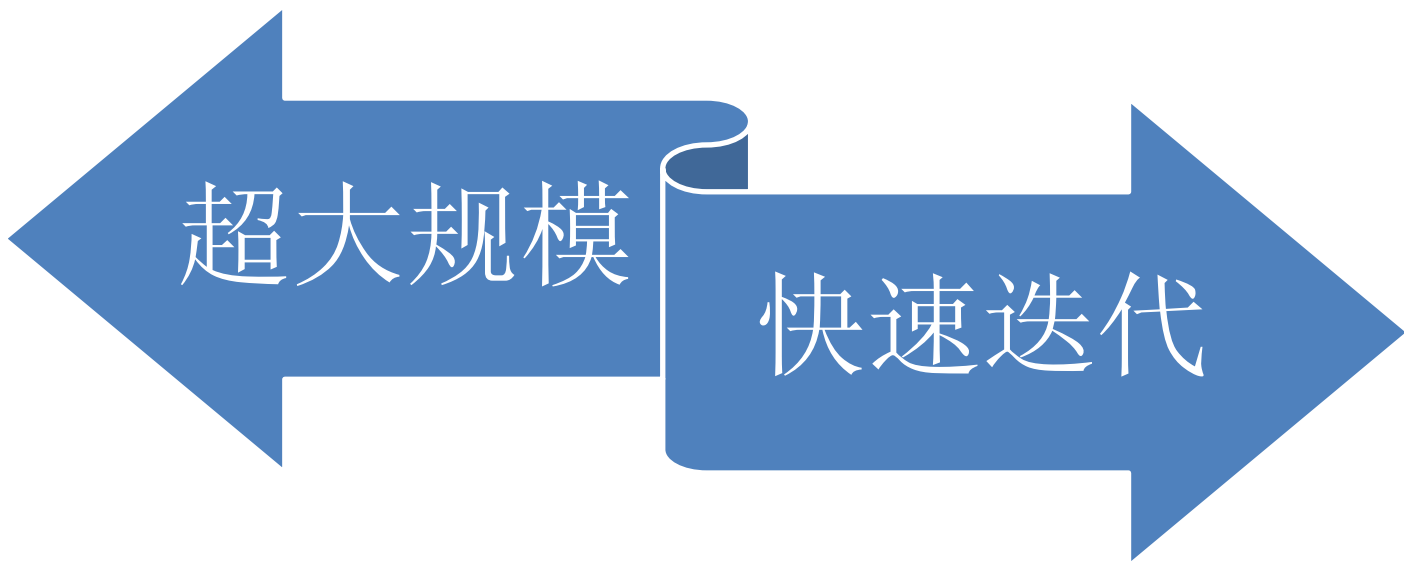


大数据的系统架构支持

@林仕鼎

2013/4/26, BDGS'13

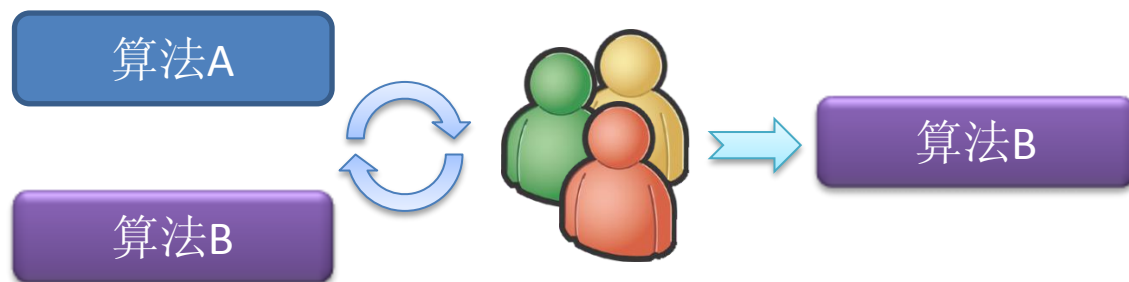
互联网服务的典型技术特点



百度的数据规模

数据总量	• 100~1000PB
数据处理量	• 10~100PB/天
网页	• 千亿~万亿
索引	• 百亿~千亿
更新量	• 十亿~百亿/天
请求	• 十亿~百亿/天
日志	• 100TB~1PB/天

快速迭代是互联网产品的主要创新手段

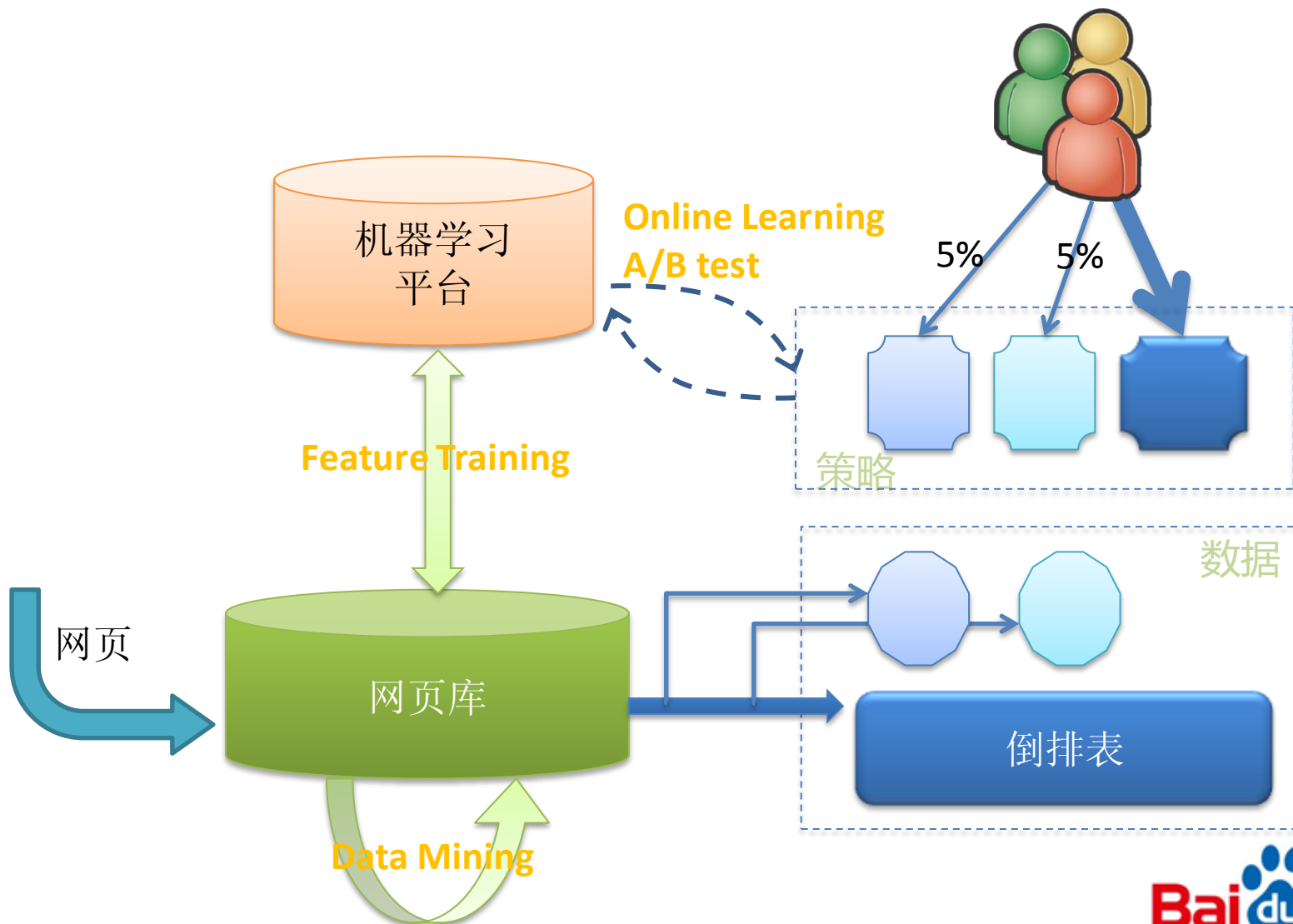


通过反馈来验证算法优劣



离线分析与在线实验相结合

搜索引擎的迭代



互联网产品的迭代

A/B测试，持续优化



互联网服务

enable

数据智能

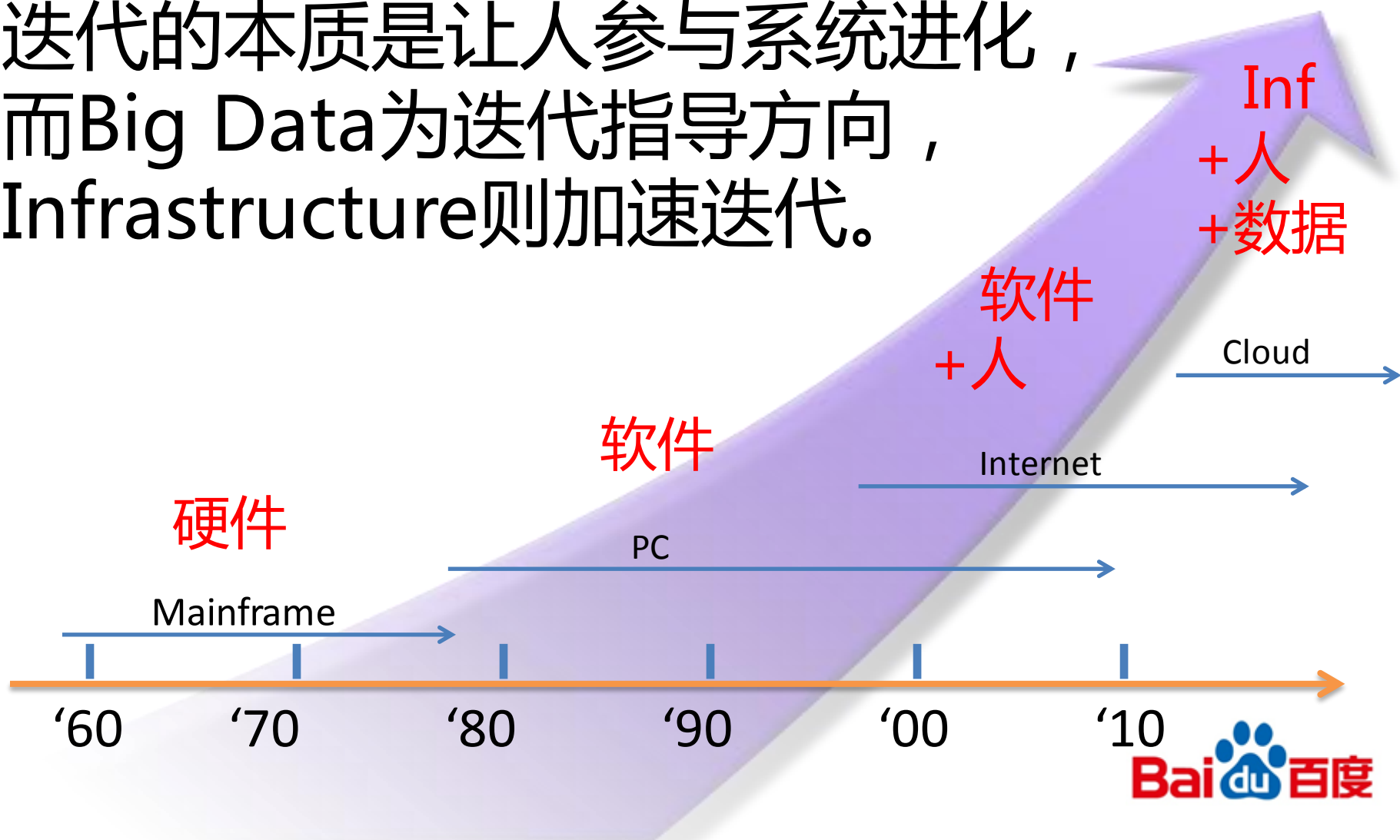
数据

架构
技术



IT产业生产力的变化

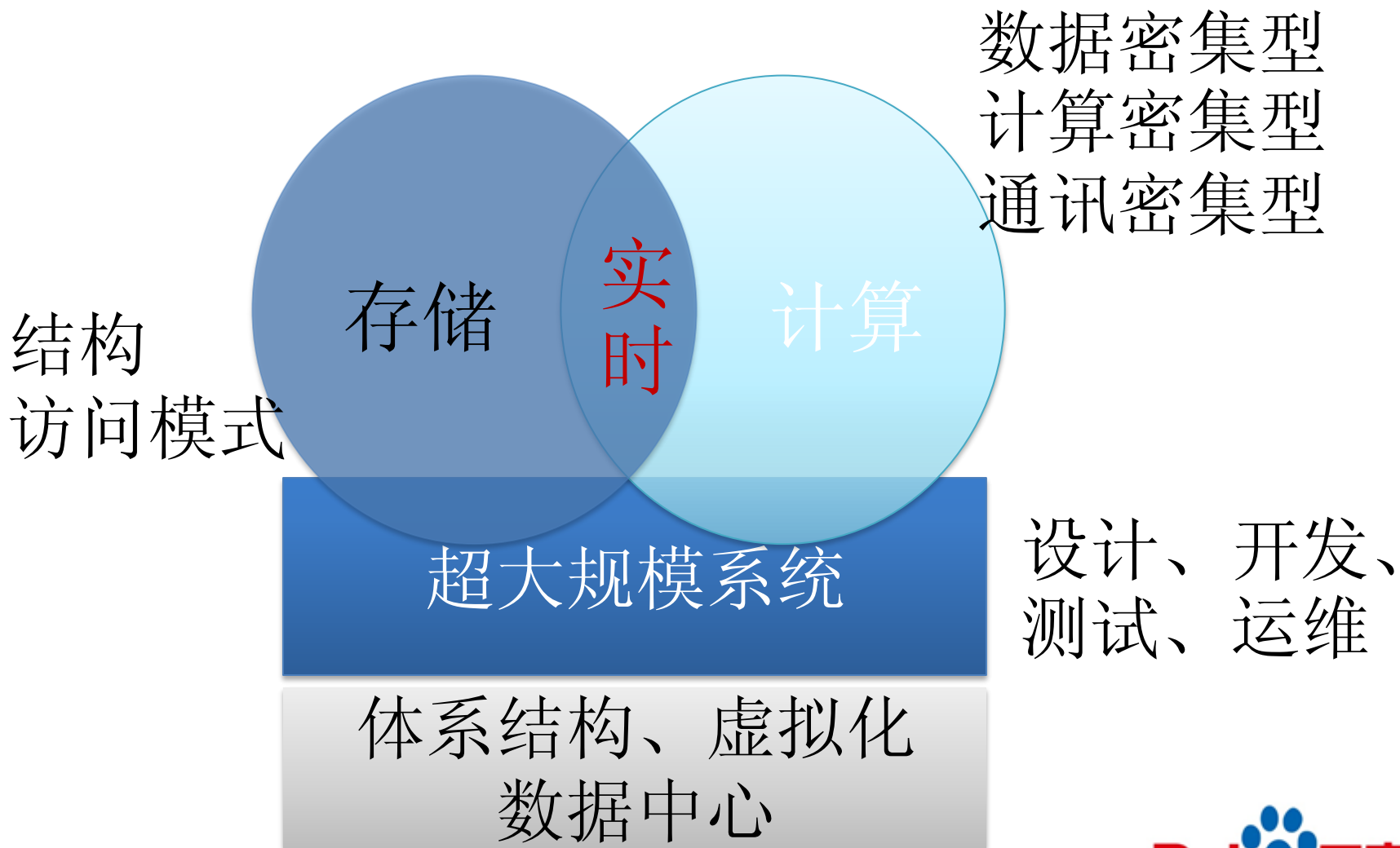
迭代的本质是让人参与系统进化，
而Big Data为迭代指导方向，
Infrastructure则加速迭代。



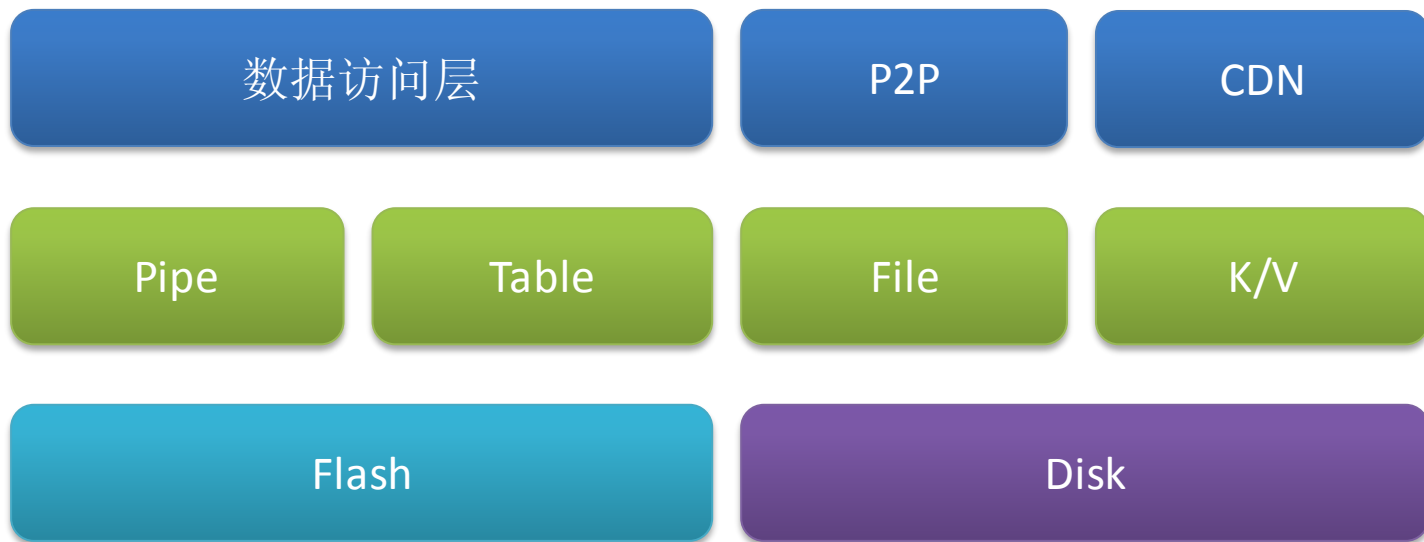
云计算技术体系



主要技术领域

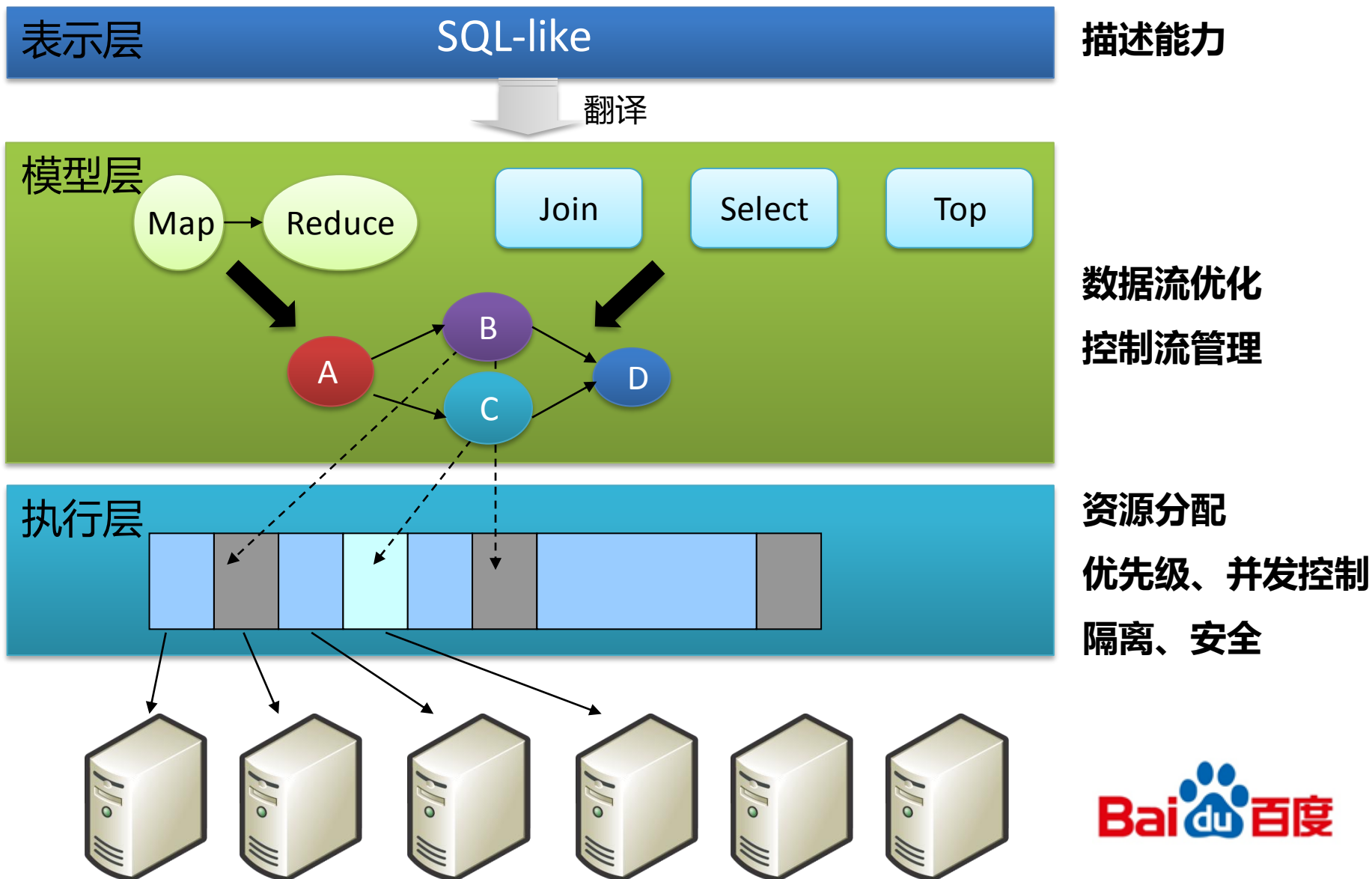


分布式存储



- ✓ 统一存储体系
 - 平衡大容量、高并发、低延迟
 - 不同访问模式通过组合满足
- ✓ 统一访问与传输

分布式计算



实时存储与计算

图查询
平台

kNN查询
平台

机器学习
算法平台

PubSub
引擎

实时检索
平台

OLAP
引擎

向量计算引擎

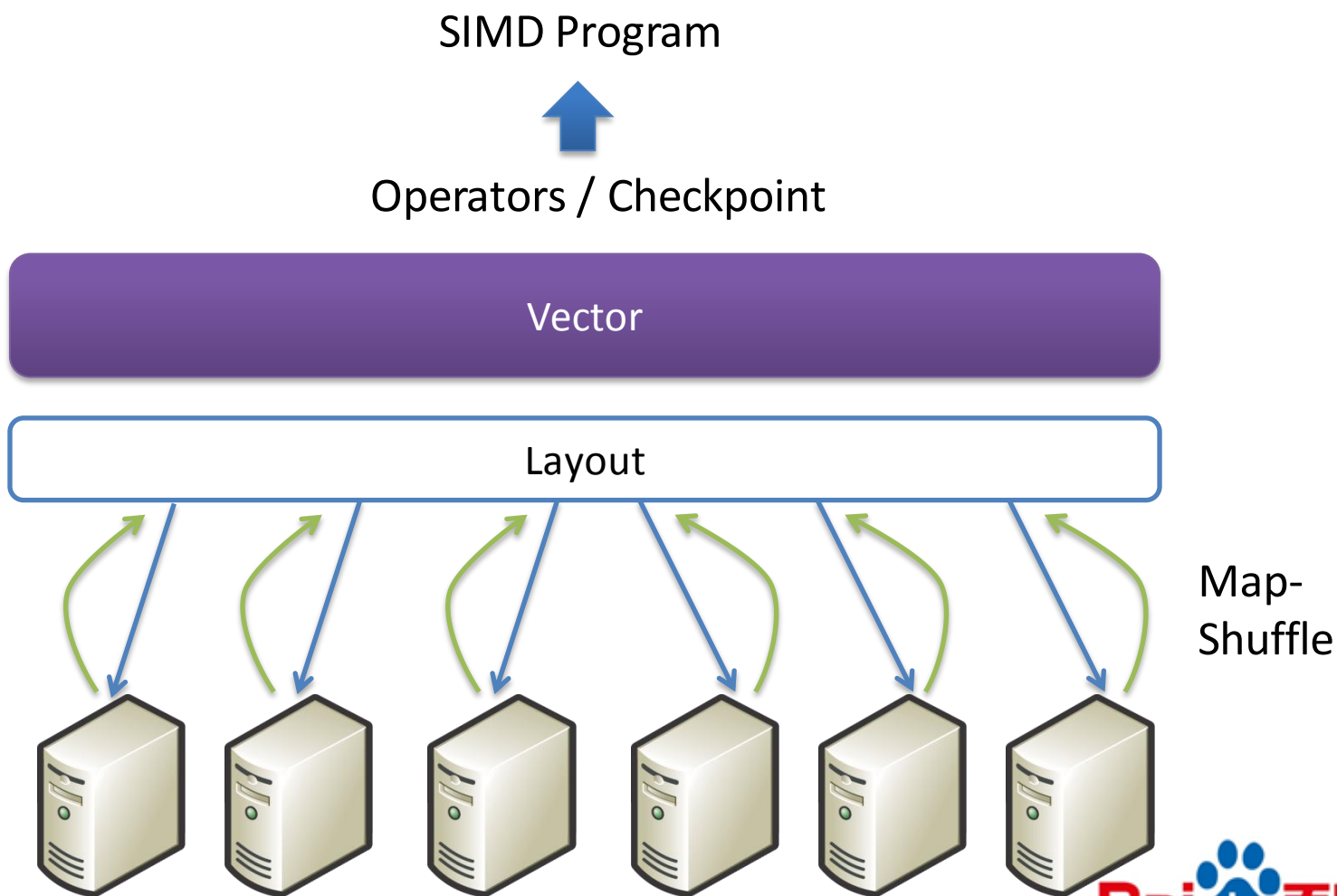
复杂事件处理引擎

流式数据处理引擎

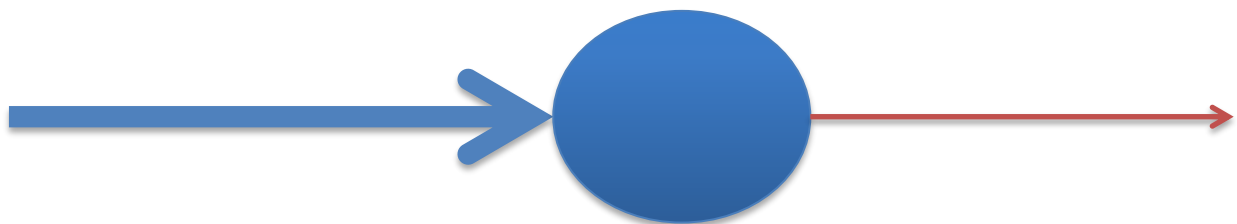
分布式数据结构

超大规模数据仓库

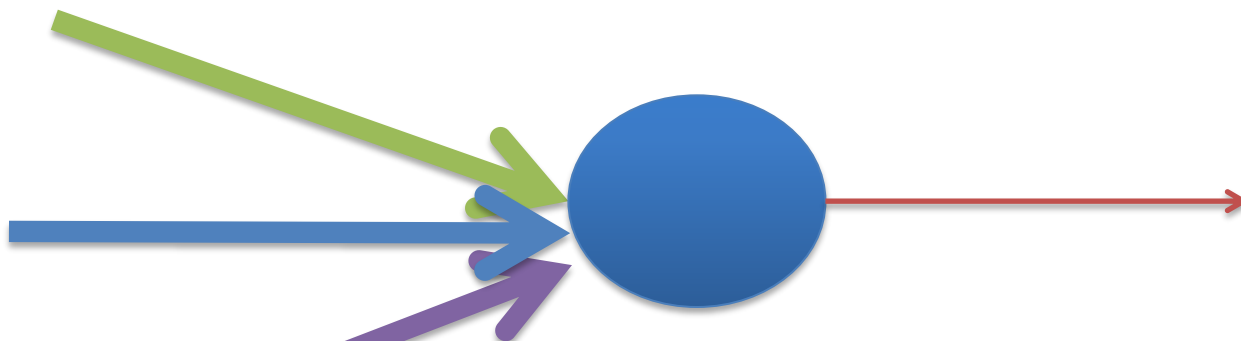
向量计算引擎



复杂事件处理

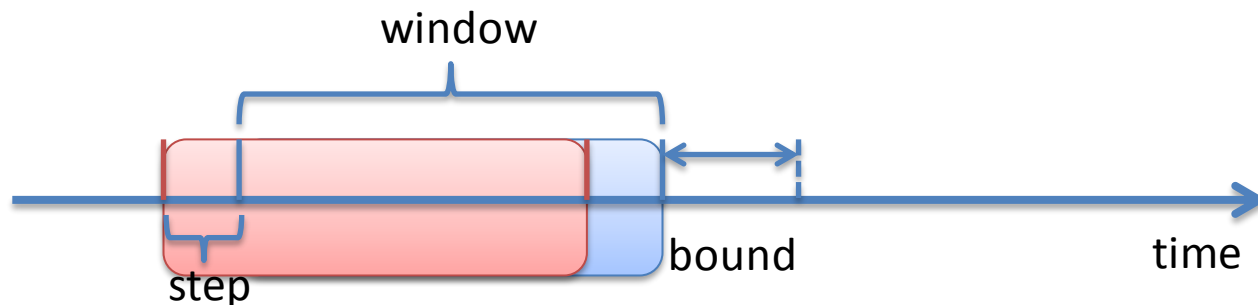


average(price)
trigger(?,b,c)
filter(b)



pattern(a->b->c)
condition(func(a,b,c))

流式计算模型



$M = \text{Stream} \langle \text{window}, \text{step}, \text{bound} \rangle$

目标

海量

- 1000PB

高维、多维

- 10亿维特征训练
- 100维条件查询

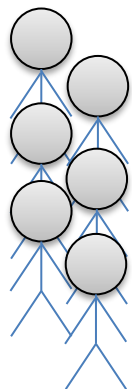
实时

- 流式
- 触发式

更大、更复杂、更快!

数据智能

人计算



大规模人工
辅助标注系统

Web Contents

分布式
存储与计算

Logs

流式处理

Machine
Learning
算法

向量引擎

PubSub

Apps

自动评估

推荐系统

智能交通

商业智能

决策辅助

Thanks!