02

第2章 Spark的设计与运行原理

授课教师：[李阳源]
15138617705@163.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



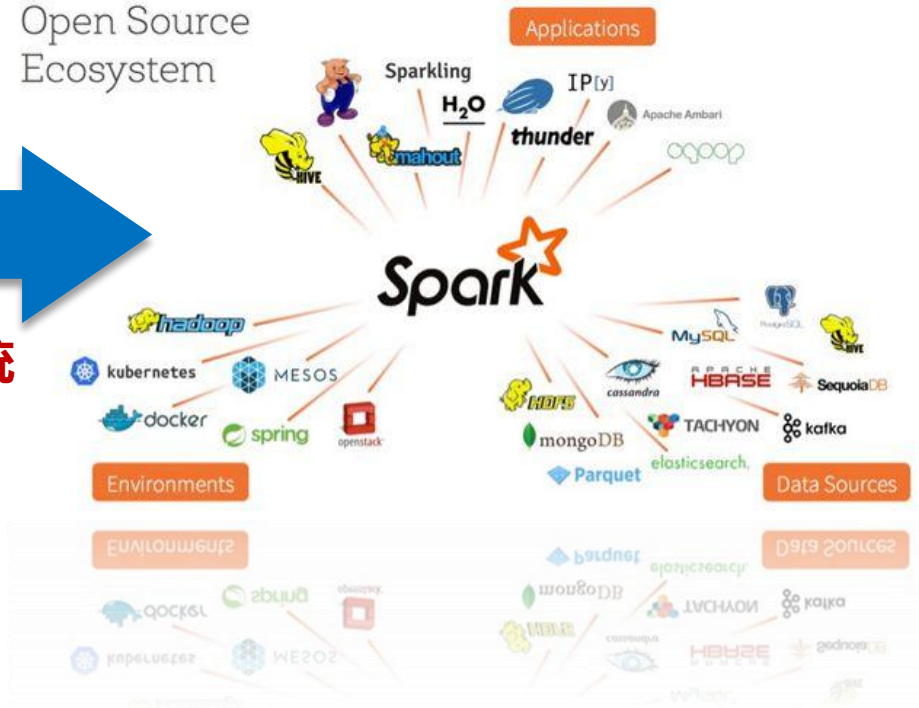
2.2 Spark生态系统

2.2 Spark 生态系统



?

为什么要构建一个生态系统



2.2 Spark 生态系统



应用类型	时间跨度
复杂的批量数据处理	数十分钟到数小时
基于历史数据的交互式查询	数十秒到数分钟
基于实时数据流的数据处理	数百毫秒到数秒

2.2 Spark 生态系统



批处理





交互查询



流处理

2.2 Spark 生态系统

无法无缝共享



2.2 Spark 生态系统



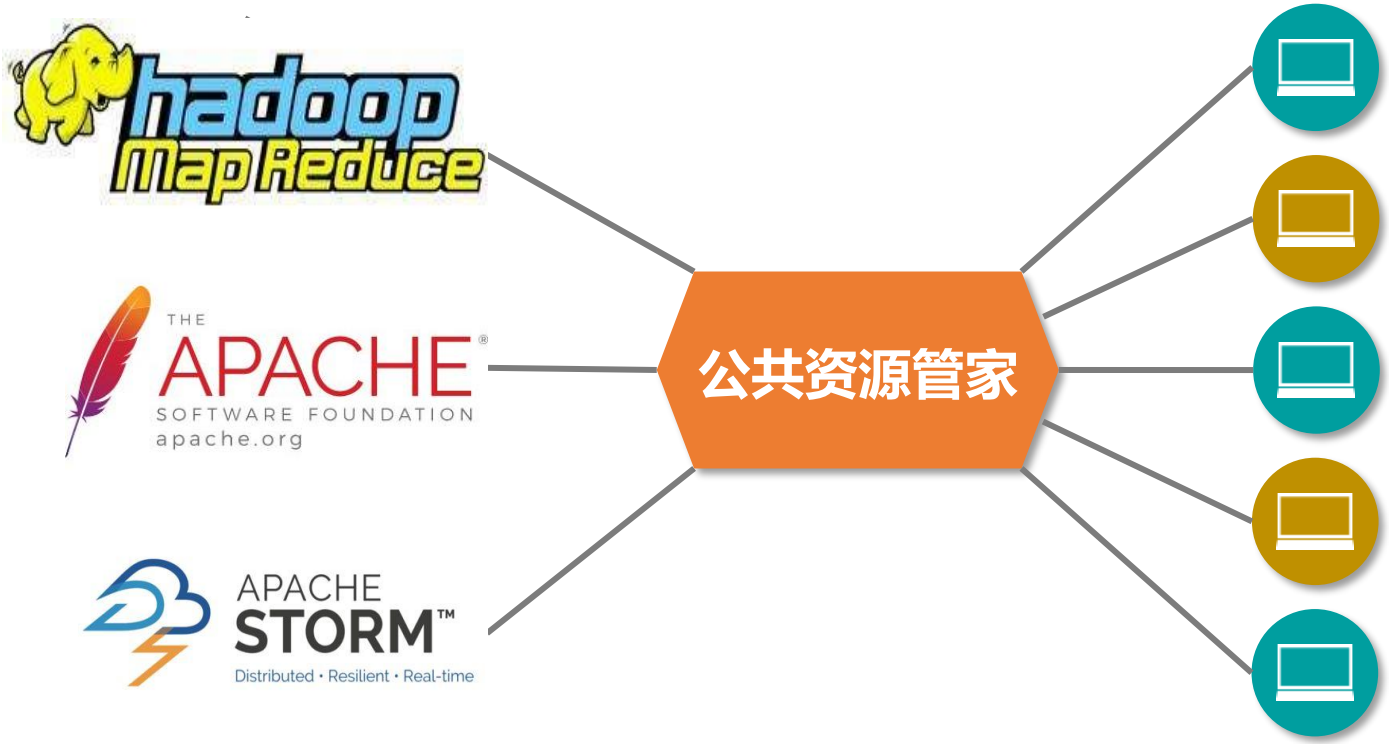
维护成本较高



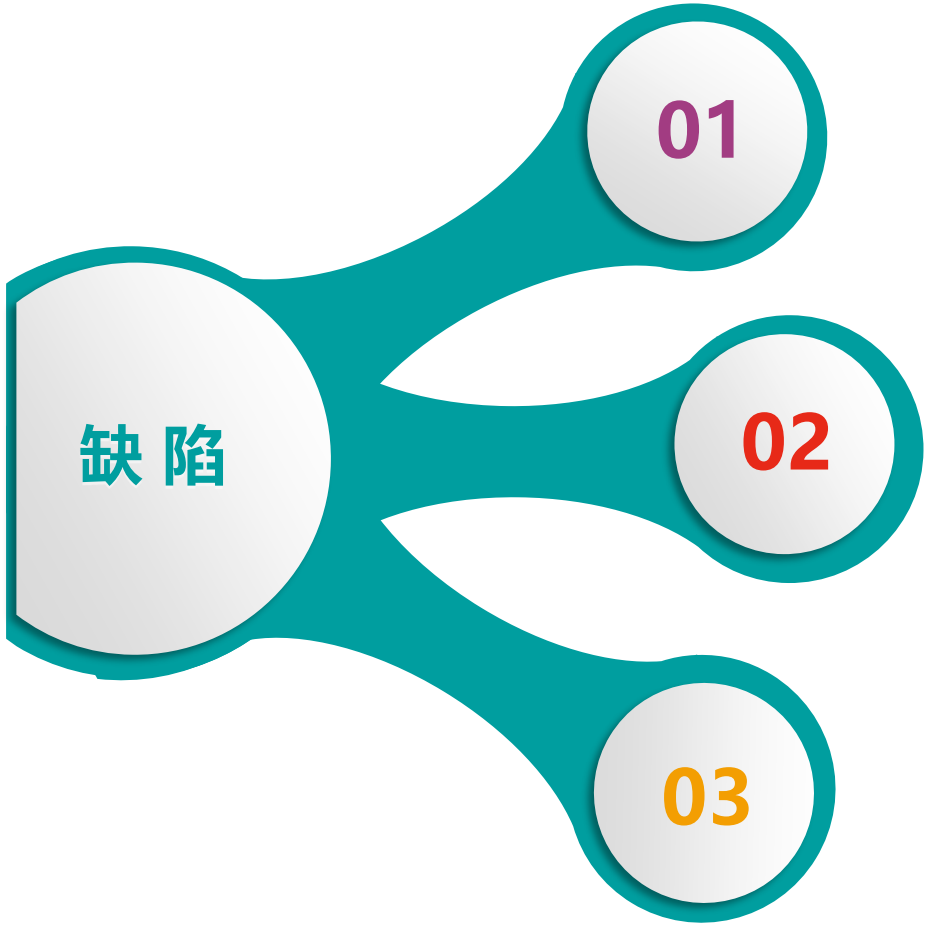
2.2 Spark 生态系统

资源利用不充分

集群



2.2 Spark 生态系统



无法无缝共享

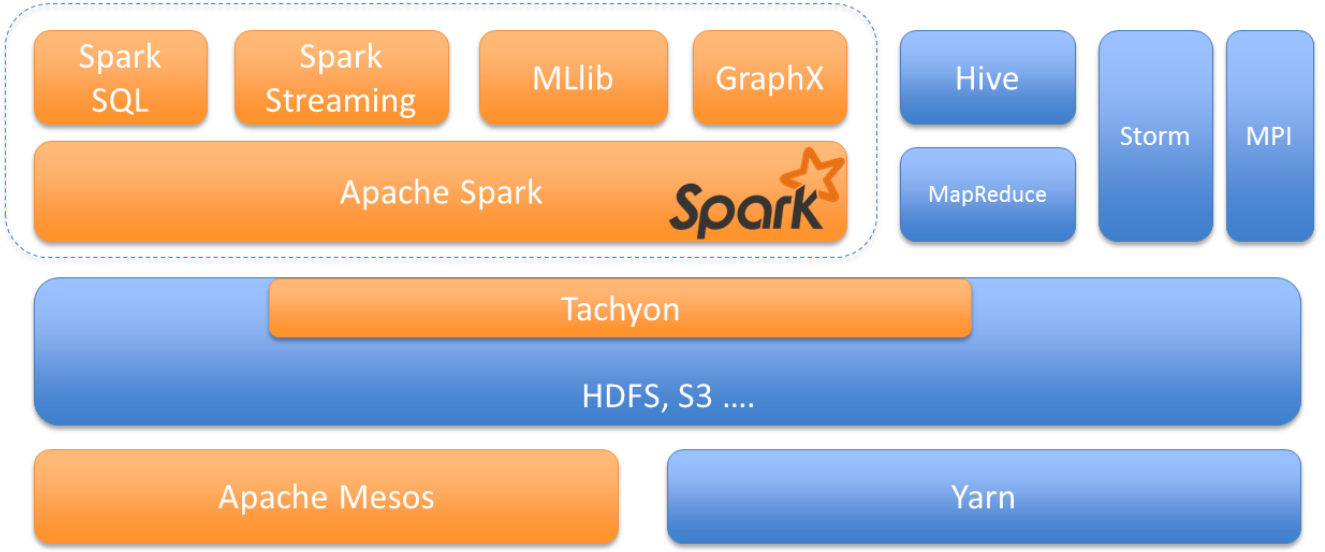
使用成本较高

资源利用不充分

2.2 Spark 生态系统

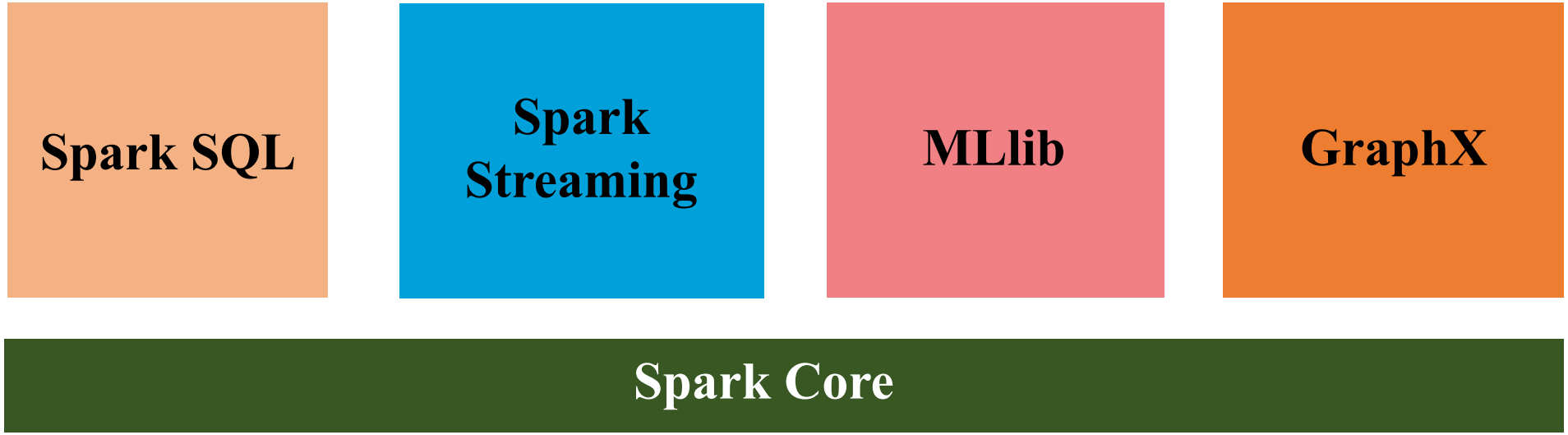


满足不同的应用场景需求



2.2 Spark 生态系统

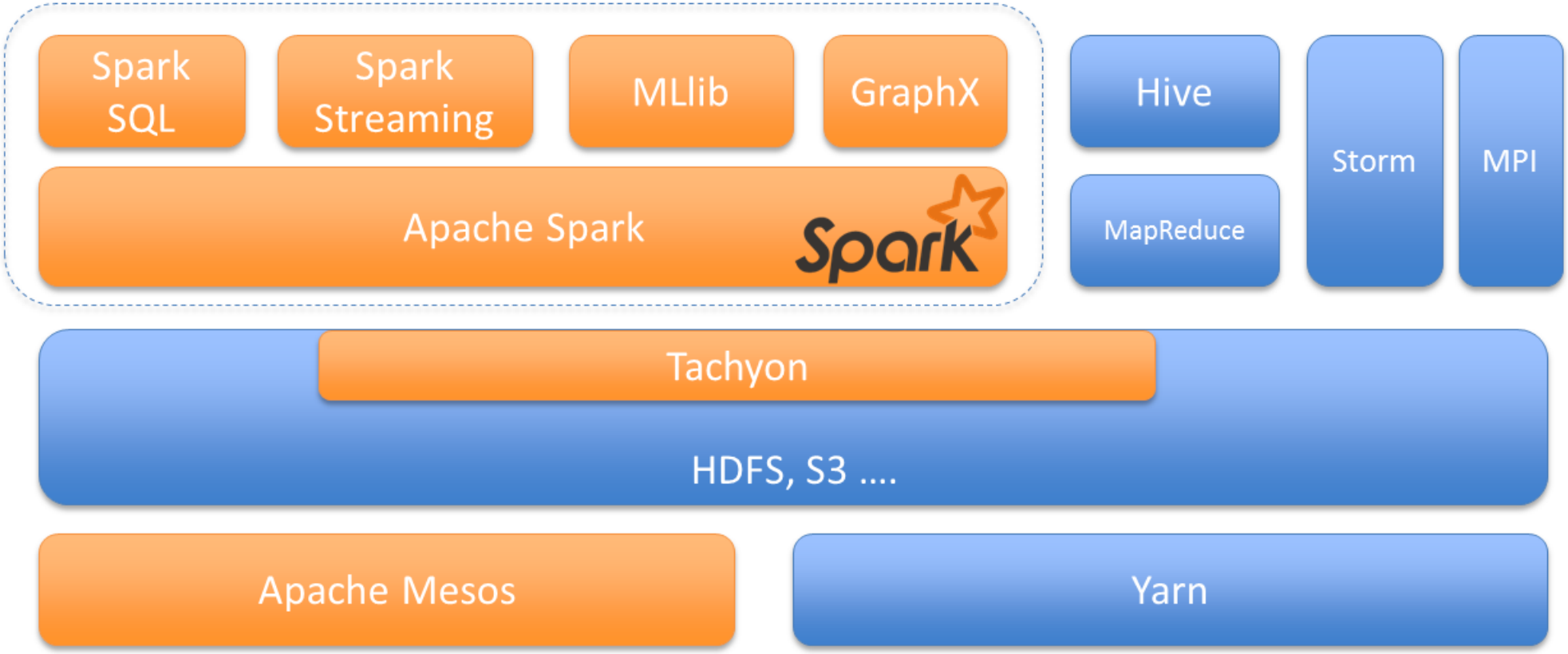




2.2 Spark 生态系统



“一个软件栈满足不同应用场景”



2.2 Spark 生态系统



一个软件栈满足不同应用场景

01 SQL即席查询

02 实时流式计算

03 机器学习

04 图计算



2.2 Spark 生态系统



一站式大数据解决方案







YARN

2.2 Spark 生态系统

一站式大数据解决方案



3.2 Spark的生态系统

一站式大数据解决方案



2.2 Spark 生态系统

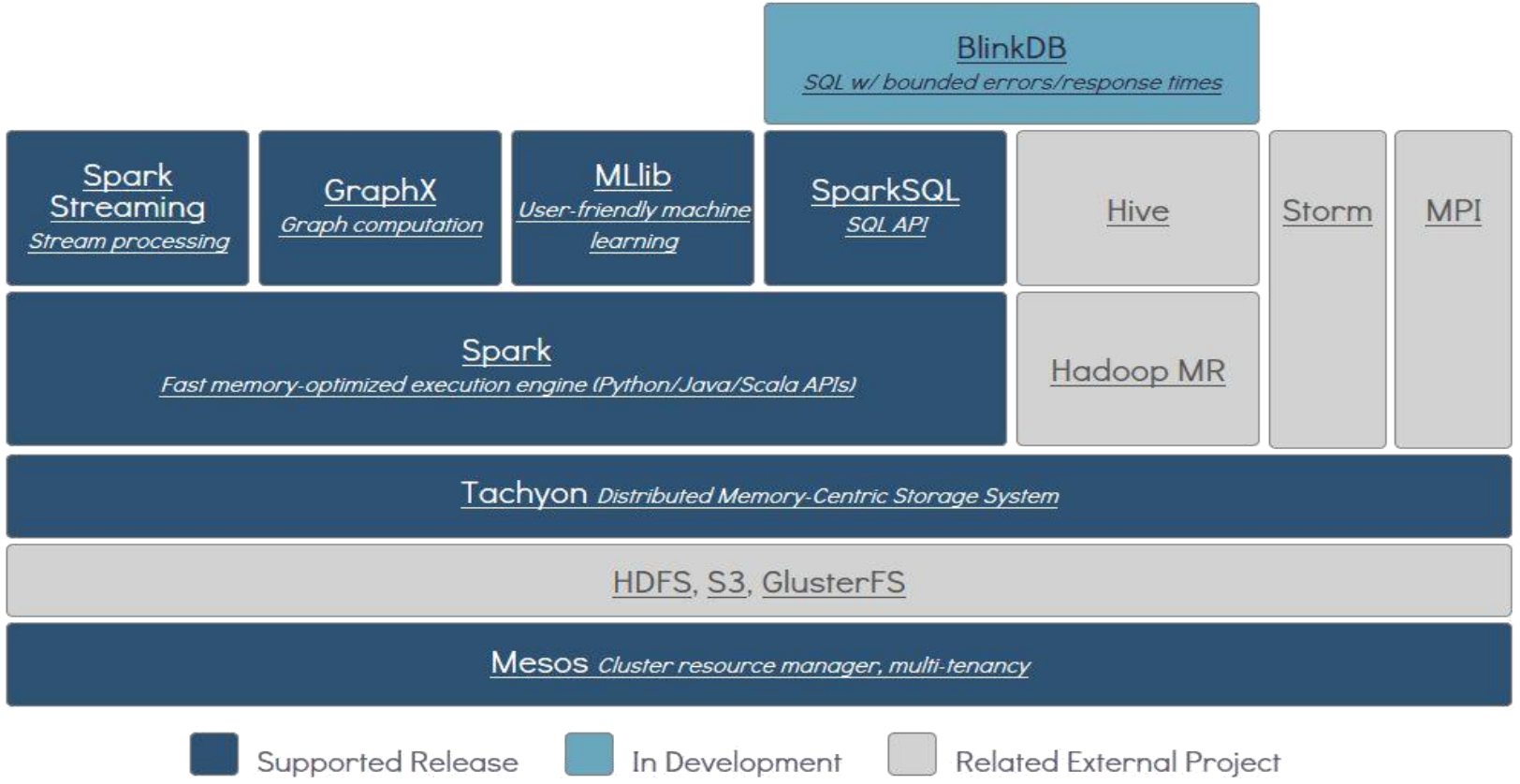




伯克利数据分析软件栈

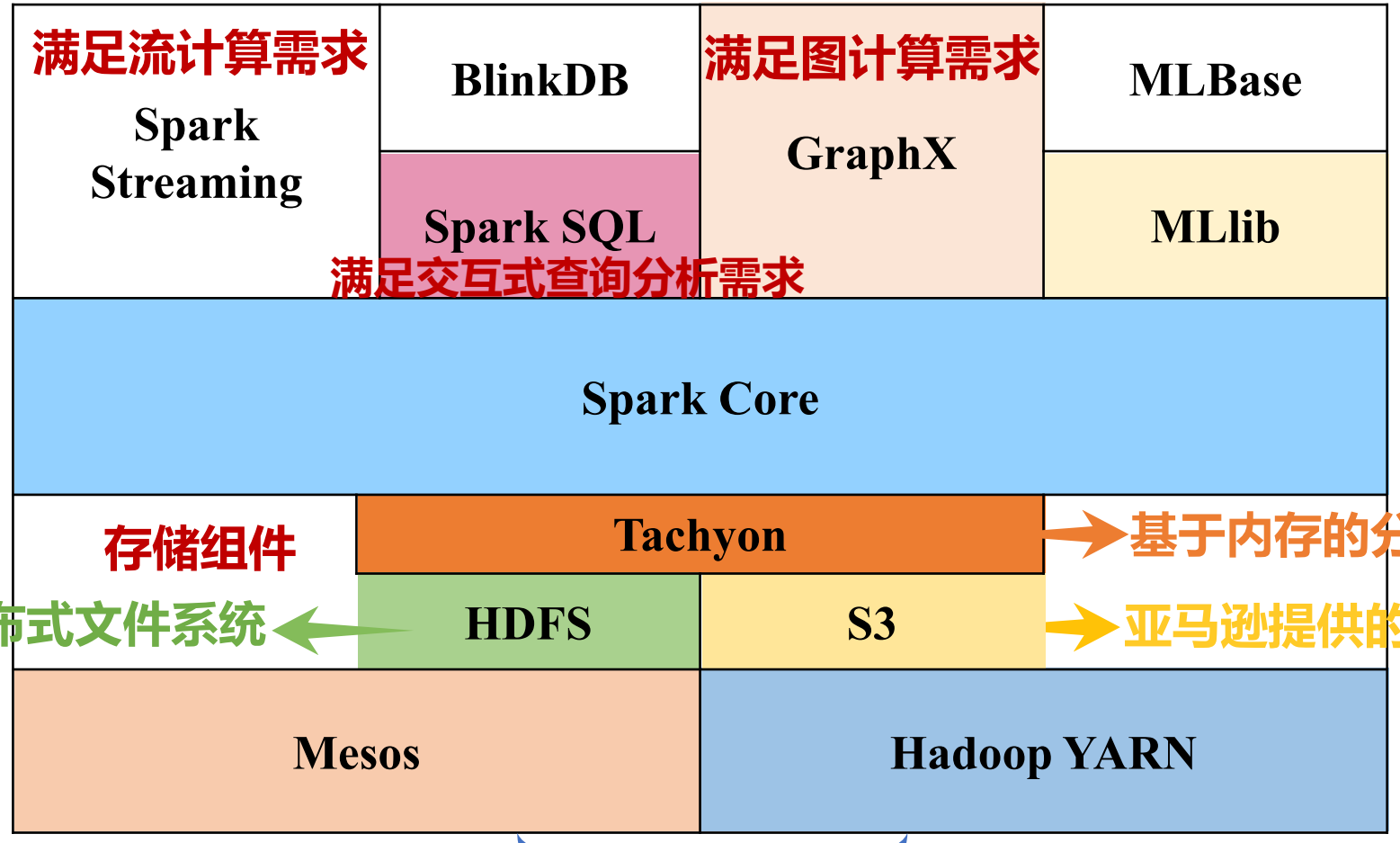
BDAS（Berkeley Data Analytics Stack）

2.2 Spark 生态系统



BDAS（Berkeley Data Analytics Stack）

2.2 Spark 生态系统



2.2 Spark 生态系统



MLlib

机器学习算法库

分类

聚类



用Spark编写

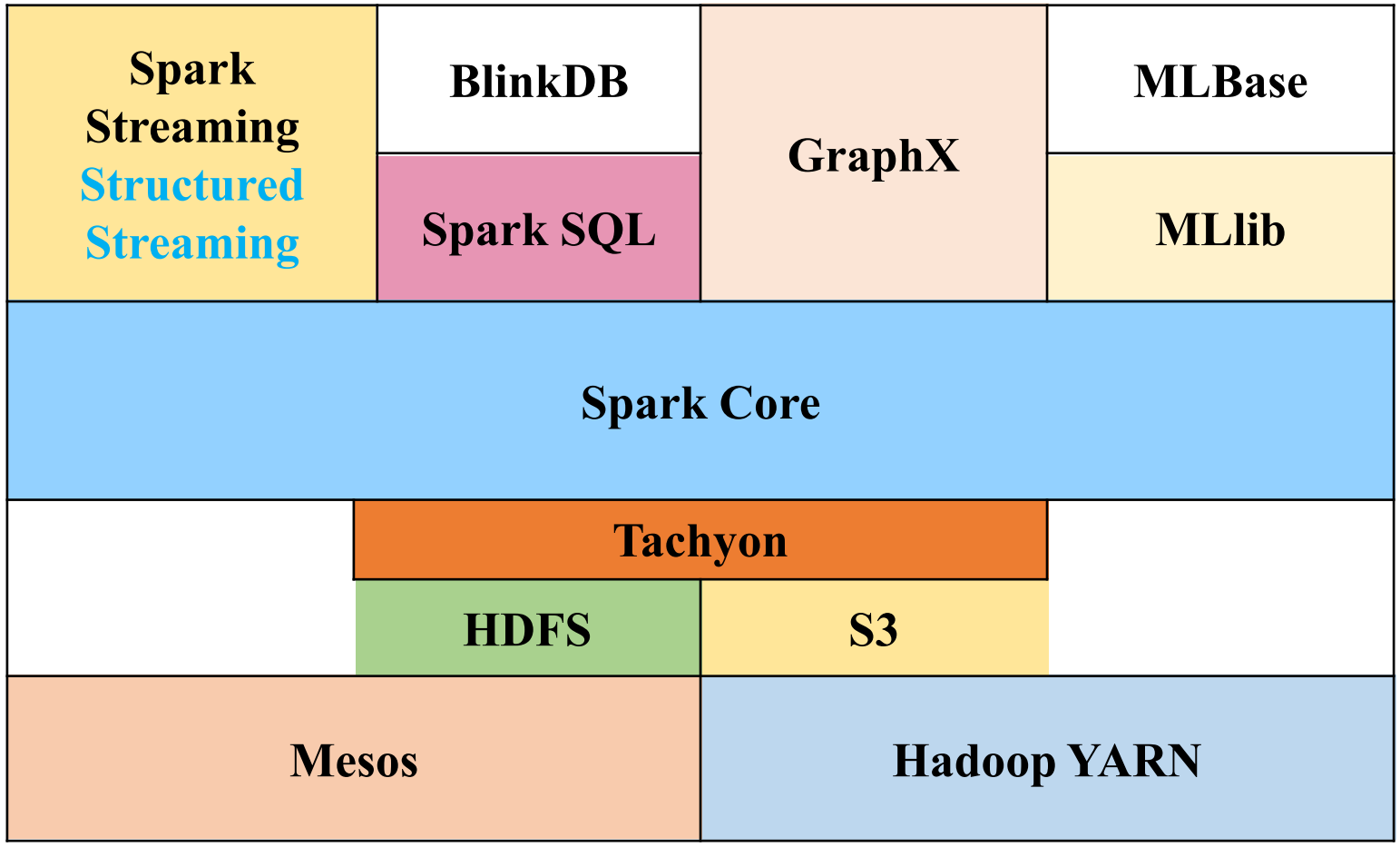
调用接口

传参数

进行训练

输出结果

2.2 Spark 生态系统





2.2 Spark 生态系统



Spark 2.0



Structured Streaming



2.2 Spark 生态系统



流计算

特定应用场景需求

Structured Streaming



DataFrame数据抽象

Spark Streaming



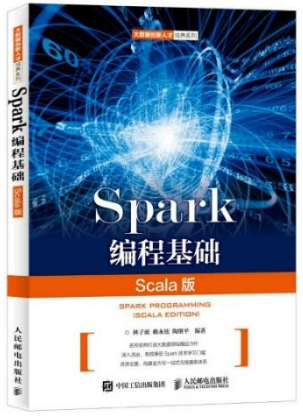
RDD数据抽象

2.2 Spark 生态系统

Spark 2.0 试用版

Structured Streaming

Spark 2.1



Spark 2.2 正式版



2.2 Spark 生态系统



2018年2月28日



2.3

Structured Streaming

实时流计算

2.2 Spark 生态系统



支持毫秒级的流计算

Spark生态系统当中组件非常多

可以很好地去取代一些其他的组件



3.2 Spark的生态系统

应用场景	时间跨度	其他框架	Spark生态系统中的组件
复杂的批量数据处理	小时级	MapReduce Hive	Spark Core
基于历史数据的交互式查询	分钟级 秒级	Impala Dremel Drill	Spark SQL
基于实时数据流的数据处理	毫秒 秒级	Storm S4	Spark Streaming Structured Streaming
基于历史数据的数据挖掘	-	Mahout	MLlib
图结构数据的处理	-	Pregel Hama	GraphX



Mahout

以前是基于MapReduce写的算法库

2.2 Spark 生态系统







日落西山

崛起



2.2 Spark 生态系统



Mahout团队

完全倒向

开始倒戈

Spark



不再支持MapReduce后续更新



Mahout

现在是基于Spark的机器学习算法库

2.2 Spark 生态系统



应用场景	时间跨度	其他框架	Spark生态系统中的组件
复杂的批量数据处理	小时级	MapReduce Hive	Spark Core
基于历史数据的交互式查询	分钟级 秒级	Impala Dremel Drill	Spark SQL
基于实时数据流的数据处理	毫秒 秒级	Storm S4	Spark Streaming Structured Streaming
基于历史数据的数据挖掘	-	Mahout	MLlib
图结构数据的处理	-	Pregel Hama	GraphX

2.2 Spark生态系统理解测验（选择题）



문제 1 (개념 이해)

传统大数据处理架构(MapReduce, Storm, Apache)相比Spark生态系统的主要劣势不包括以下哪一项？
A) 无法实现不同组件间的无缝数据共享，需要进行格式转换
B) 各组件需要独立的维护团队，增加人力成本
C) 集群资源无法动态分配，资源利用率低
D) 单个组件的处理性能比Spark组件更低

2.2 Spark生态系统理解测验（选择题）



문제 2 (아키텍처 분석)

Spark生态系统采用分层架构设计的最重要意义是什么？
A) 提高各个组件的独立性，减少相互依赖
B) 通过统一的Spark Core确保组件间无缝集成
C) 降低系统的内存使用量
D) 简化用户的学习成本

2.2 Spark生态系统理解测验（选择题）



문제 3 (응용 시나리오 분석)
一个企业需要处理以下任务：实时监控用户行为(毫秒级响应)、每日批量分析用户画像(小时级处理)、交互式查询历史数据(秒级响应)。在Spark生态系统中，最适合的组件组合是？

A) Spark Streaming + Spark Core + Spark SQL
B) Structured Streaming + Spark Core + Spark SQL
C) Spark Streaming + MLlib + Spark SQL
D) GraphX + Spark Core + MLlib

2.2 Spark生态系统理解测验（选择题）



문제 4 (기술 발전 이해)
Mahout团队从支持MapReduce转向完全倒向Spark，这一技术演进反映了什么趋势？

A) MapReduce技术本身存在根本缺陷
B) 统一计算引擎比分离式架构更符合大数据发展需求
C) Spark在所有方面都优于MapReduce
D) 机器学习算法更适合基于内存的计算模式

2.2 Spark生态系统理解测验（选择题）



문제 5 (설계 철학 이해)
Spark提出"一个软件栈满足不同应用场景"的设计理念主要解决了什么问题？

A) 提高单一任务的处理速度
B) 减少硬件资源的需求
C) 消除不同大数据处理场景间的技术壁垒
D) 降低软件开发的复杂度

2.2 Spark生态系统理解测验（选择题）



문제 6 (버전 진화 분석)
Structured Streaming从Spark 2.0的试用版发展到2.3的正式版，相比传统的Spark Streaming，其主要改进体现在哪个方面？

A) 处理速度更快
B) 基于DataFrame抽象，提供更高级的API
C) 占用内存更少
D) 支持更多的数据源

2.2 Spark生态系统理解测验（选择题）



문제 7 (자원 관리 이해)
在Spark生态系统中，YARN和Mesos作为资源管理器的存在，主要解决了传统分离式架构的哪个问题？

A) 数据格式不兼容
B) 维护成本过高
C) 集群资源利用不充分
D) 处理延迟过高

2.2 Spark生态系统理解测验（选择题）



문제 8 (생태계 구성 분석)
BDAS(Berkeley Data Analytics Stack)中，Tachyon作为基于内存的分布式文件系统，在整个架构中的战略意义是什么？

A) 替代HDFS的所有功能
B) 提供高速数据访问，支撑上层计算组件的性能需求
C) 减少网络传输的数据量
D) 简化数据存储的管理复杂度

2.2 Spark生态系统理解测验（选择题）



문제 9 (응용 시나리오 선택)
一个电商公司需要：分析用户购买行为图谱、实时推荐商品、批量处理订单数据、查询商品销售报表。如果要最大化利用Spark生态系统的优势，应该如何分配这些任务？

A) GraphX + Spark Streaming + Spark Core + Spark SQL
B) MLlib + Structured Streaming + MapReduce + Spark SQL
C) GraphX + MLlib + Spark Core + Spark SQL
D) Spark SQL + Spark Streaming + MLlib + GraphX

2.2 Spark生态系统理解测验（选择题）



문제 10 (종합적 사고)
如果你是一家初创公司的技术架构师，需要设计一个大数据处理平台。考虑到团队规模小、技术人员有限、业务需求多样化，Spark生态系统相比传统分离式架构的最大优势是什么？

A) 技术门槛更低，容易学习
B) 硬件成本更低
C) 处理性能绝对最优
D) 统一技术栈减少维护负担，一套技能解决多种需求



Thank you