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

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

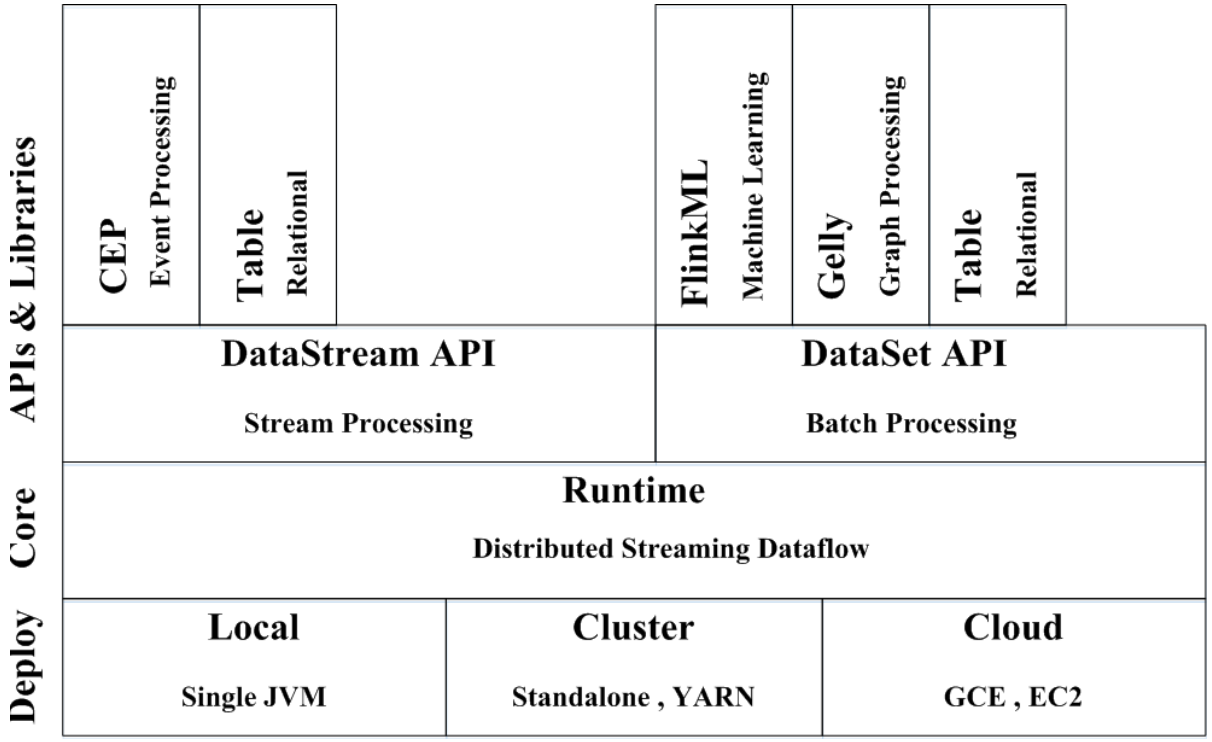
3 Flink

4 Beam



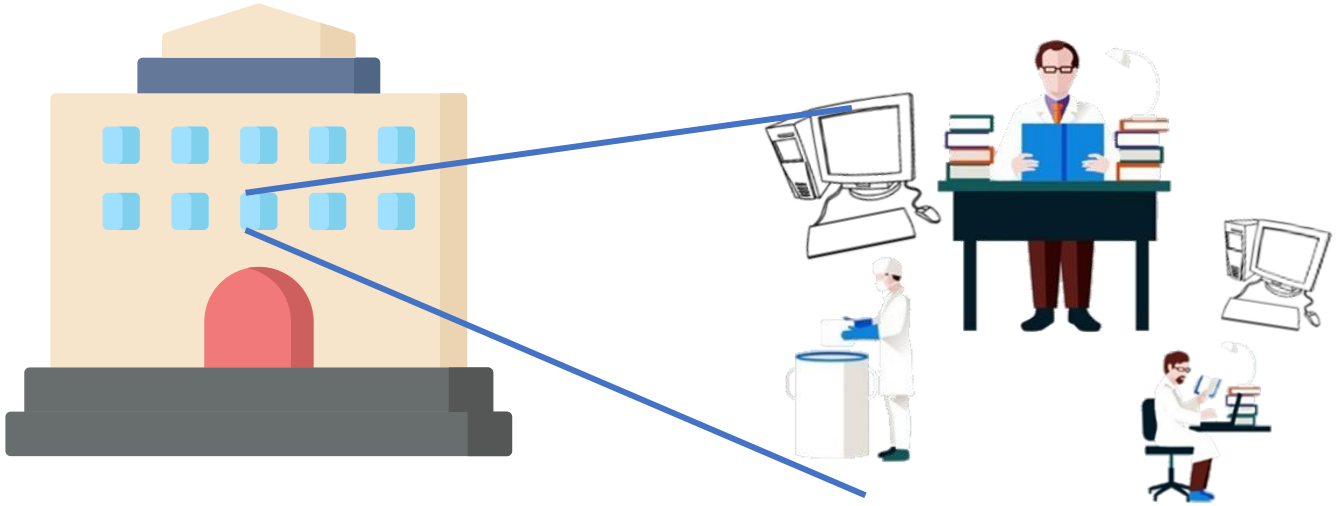
1.6.3/4 Flink和Beam

1.6.3 代表性大数据技术Flink和Beam



Flink架构图

1.6.3 代表性大数据技术Flink和Beam



柏林理工大学　　开发团队

1.6.3 代表性大数据技术Flink和Beam



Flink

2008



1.6.3 代表性大数据技术Flink和Beam



Spark

2009



1.6.3 代表性大数据技术Flink和Beam



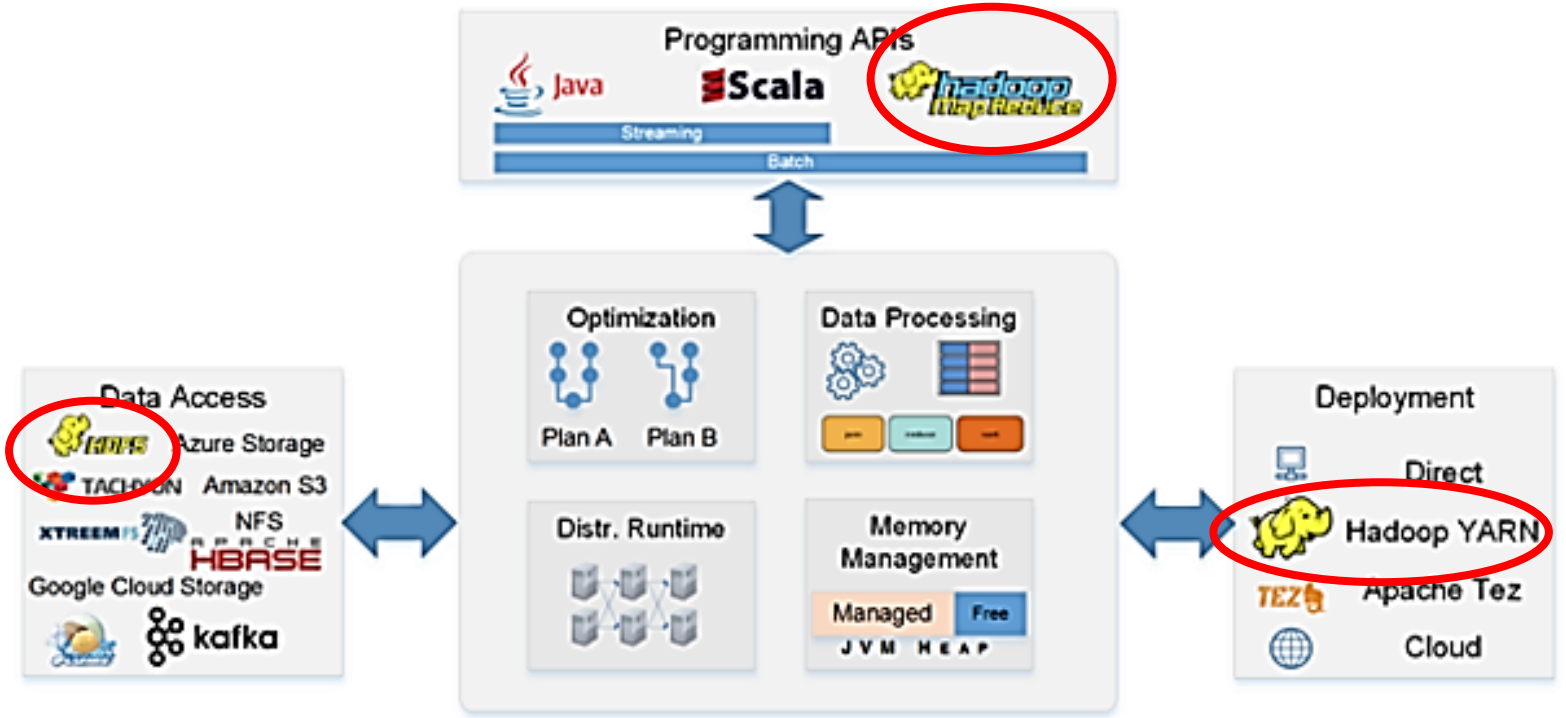
Spark

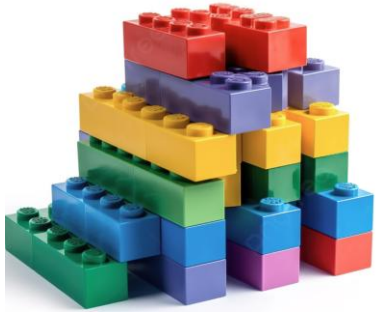


加州大学伯克利分校

1.6.3 代表性大数据技术Flink和Beam







Flink生态系统

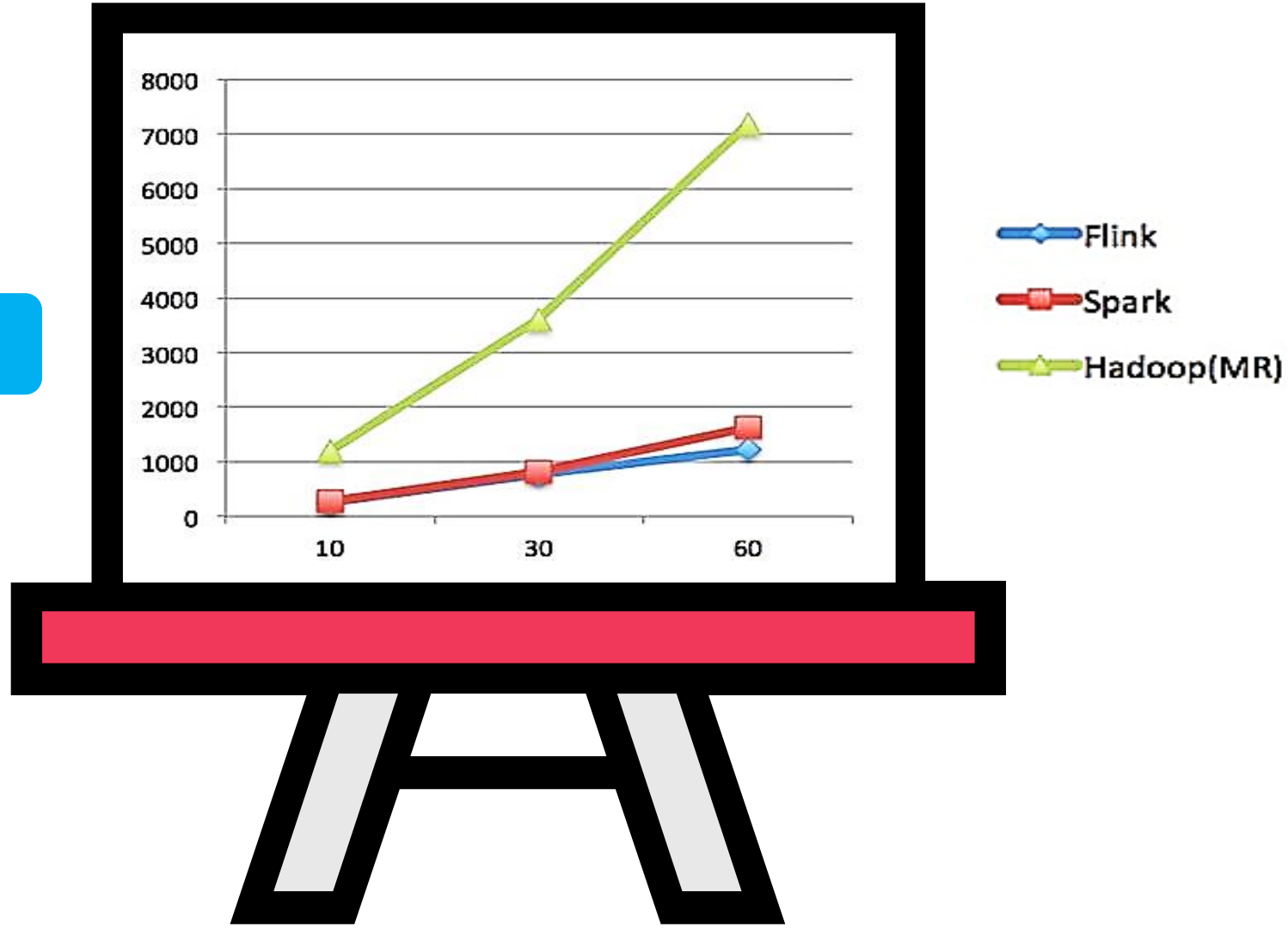
1.6.3 代表性大数据技术Flink和Beam



	Apache Spark	Apache Flink
核心实现	Scala	Java
编程接口	Java,Python,R语言	DataSet API 支持: Java、Scala和Python DataStream API 支持: Java和Scala
计算模型	基于数据片集合的微批处理模型	基于操作符的连续流模型
优缺点	流式处理有延时 只支持秒级计算	流式计算与Storm性能相当 支持毫秒级计算

1.6.3 代表性大数据技术Flink和Beam

性能对比



1.6.3 代表性大数据技术Flink和Beam



既生瑜，何生亮！

Spark

- ✓ 社区好
- ✓ 企业支持早

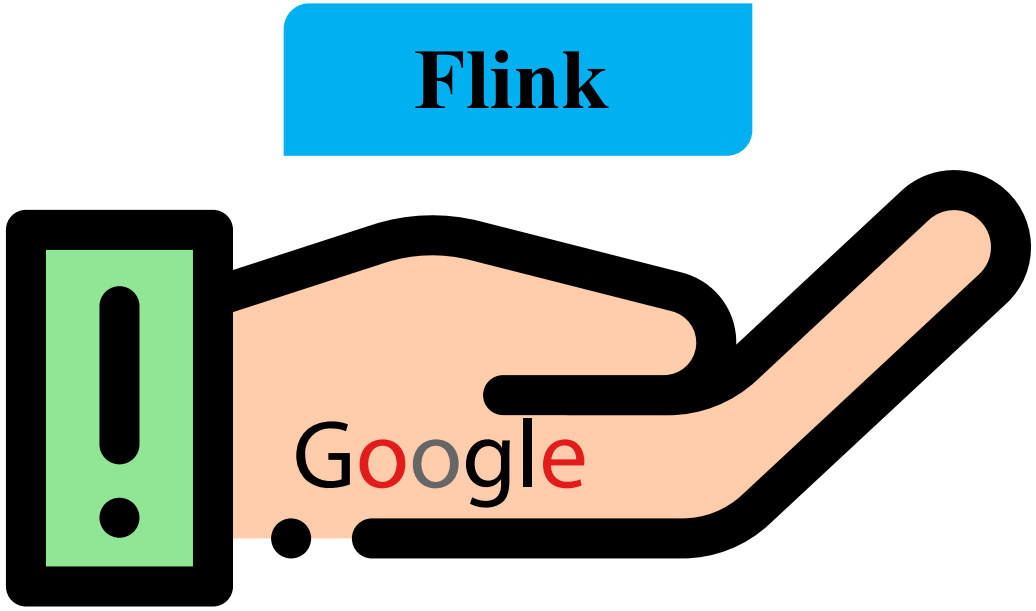
Flink

- ✓ 不愠不火



1.6.3 代表性大数据技术Flink和Beam





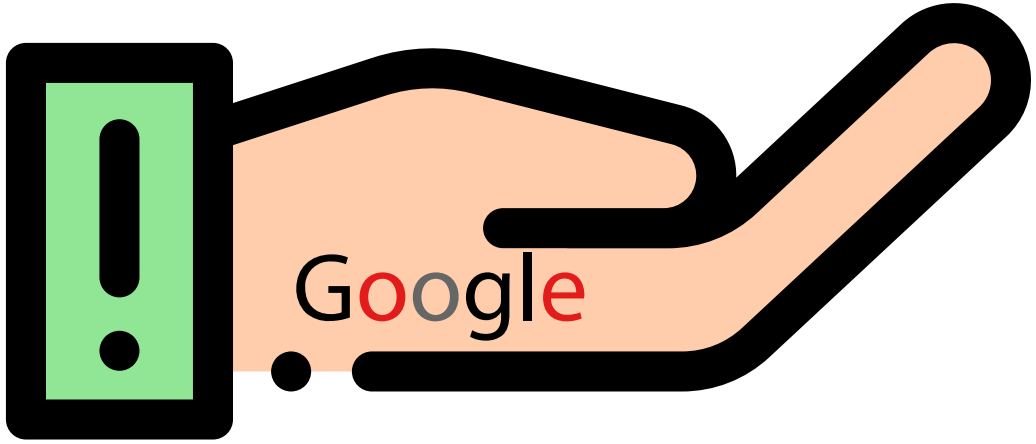




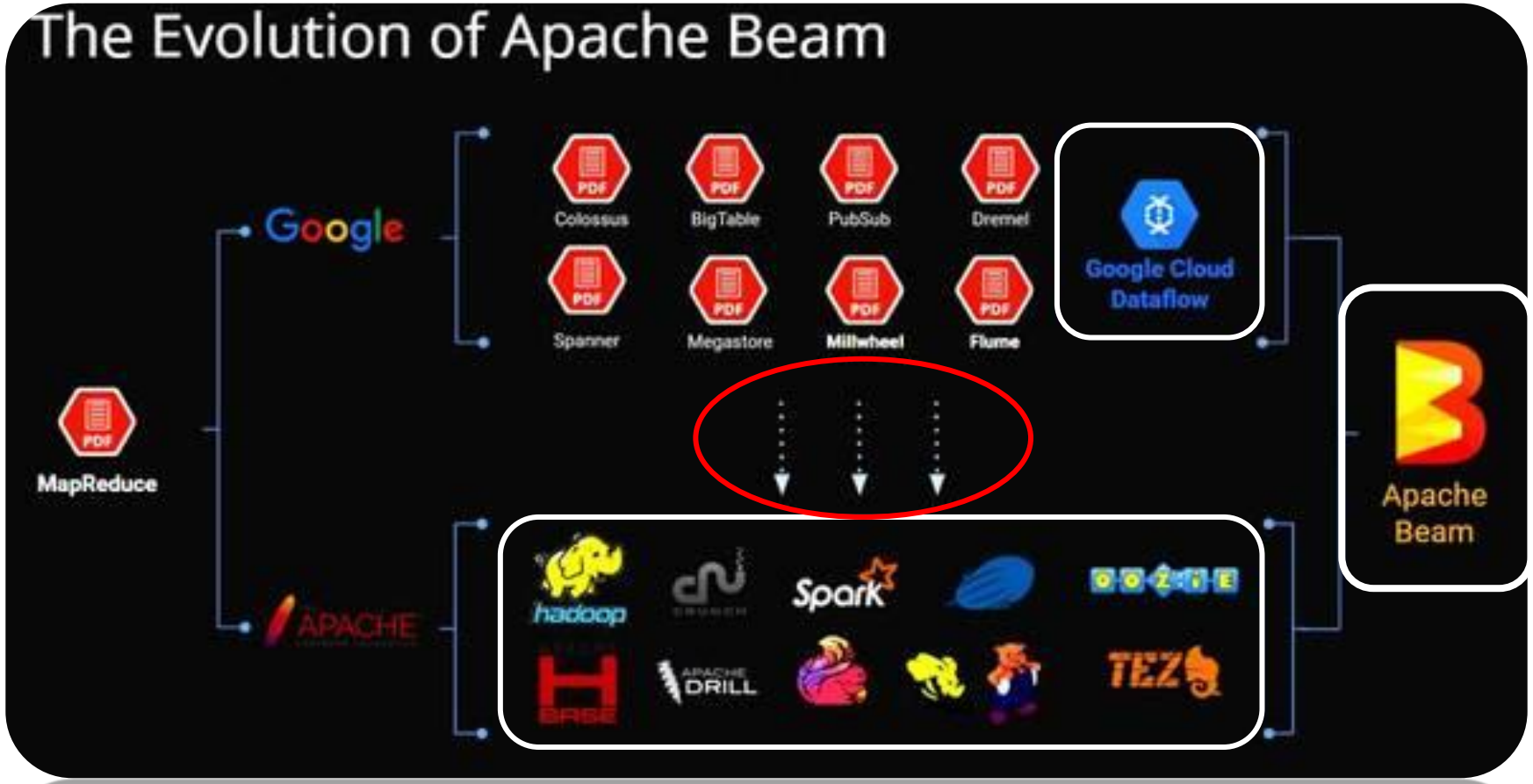
1.6.3 代表性大数据技术Flink和Beam



一统天下?

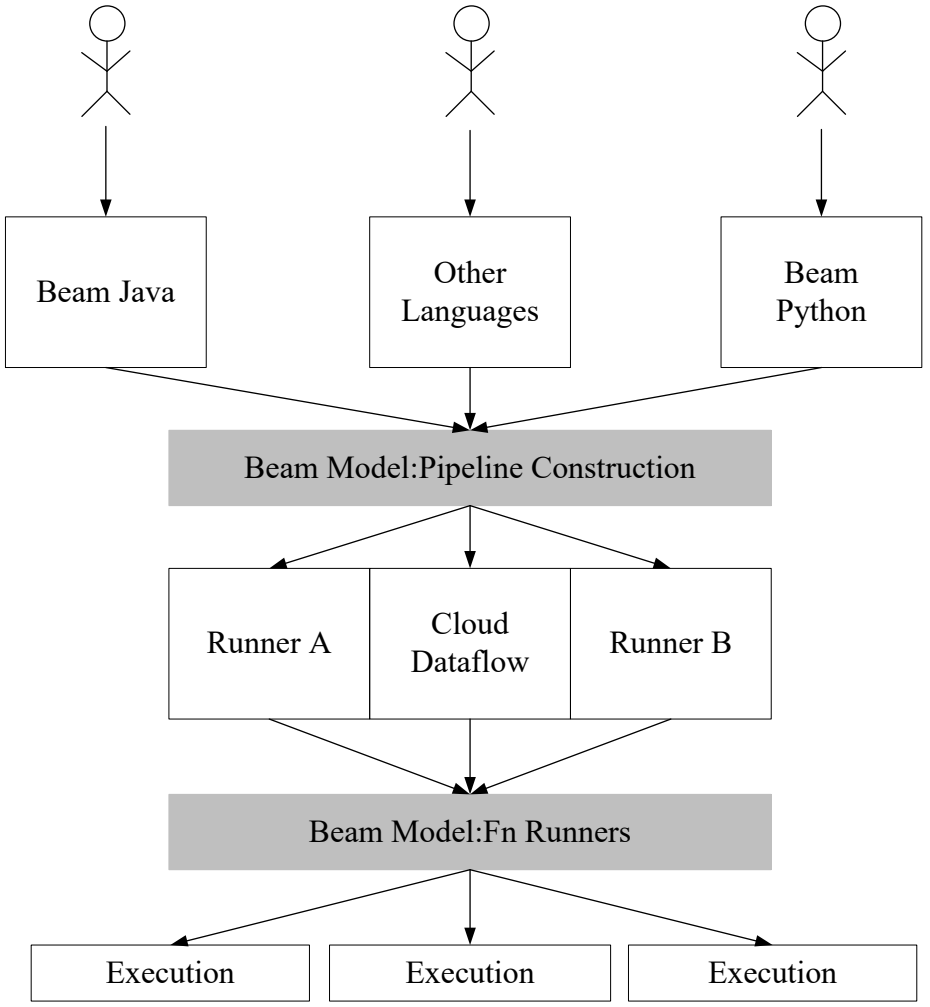


1.6.3 代表性大数据技术Flink和Beam

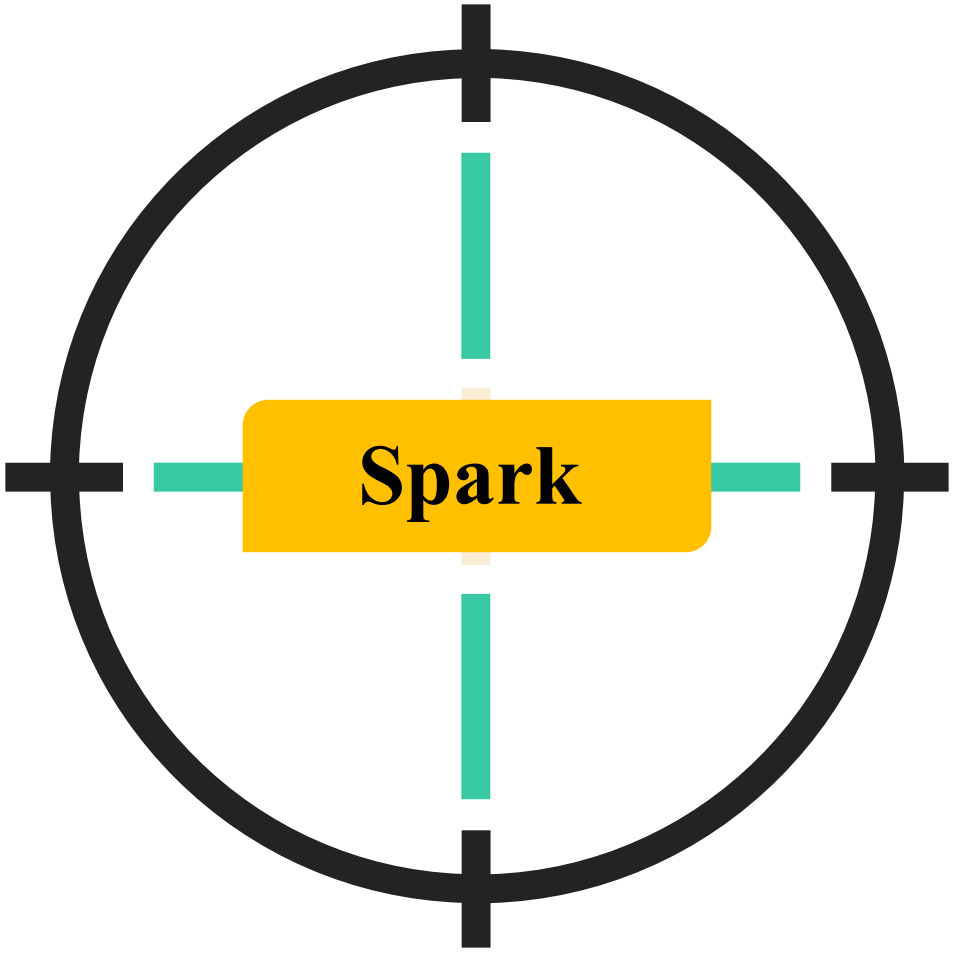


编程接口

1.6.3 代表性大数据技术Flink和Beam



1.6.3 代表性大数据技术Flink和Beam



1.6.3 Flink和Beam理解测验（选择题）



第1题

Apache Flink最初是在哪个大学开发的?

A. 加州大学伯克利分校
B. 柏林理工大学
C. 斯坦福大学
D. 麻省理工学院

1.6.3 Flink和Beam理解测验（选择题）



第2题

关于Apache Flink和Apache Spark的诞生时间，以下说法正确的是：

A. Flink诞生于2009年，Spark诞生于2008年
B. 两者都诞生于2008年
C. Flink诞生于2008年，Spark诞生于2009年
D. Flink诞生于2010年，Spark诞生于2009年

1.6.3 Flink和Beam理解测验（选择题）



第3题

在Flink架构中， DataStream API主要用于处理什么类型的数据？

A. 批处理数据
B. 流处理数据
C. 图数据
D. 时序数据

1.6.3 Flink和Beam理解测验（选择题）



第4题

Apache Spark的核心实现语言是什么？

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

1.6.3 Flink和Beam理解测验（选择题）



第5题

根据课程内容，Flink相比Spark在流处理方面的主要优势是:

A. 支持更多编程语言
B. 有更好的社区支持
C. 支持毫秒级计算，延迟更低
D. 部署更简单

1.6.3 Flink和Beam理解测验（选择题）



第6题

Apache Beam的核心设计理念是什么？

A. 只支持批处理
B. 只支持流处理
C. 一次编写，到处运行
D. 只能在Google Cloud上运行

1.6.3 Flink和Beam理解测验（选择题）



第7题

在Flink生态系统中，以下哪个不是其支持的编程API？

A. Java
B. Scala
C. Python
D. Go

1.6.3 Flink和Beam理解测验（选择题）



第8题

根据性能对比图表，当节点数量从10增加到60时，哪个系统的性能提升最为显著？

A. Flink
B. Spark
C. Hadoop MapReduce
D. 三者提升幅度相同

1.6.3 Flink和Beam理解测验（选择题）



第9题

Apache Beam架构中的"Pipeline Construction"层主要负责什么？

A. 执行具体的数据处理任务
B. 提供编程接口
C. 将用户代码转换为执行计划
D. 管理集群资源

1.6.3 Flink和Beam理解测验（选择题）



第10题

课程中提到Google对Flink的支持表明了什么？

A. Google要收购Flink
B. Flink将替代所有其他大数据技术
C. 即使是技术巨头也认可Flink在流处理方面的优势
D. Flink只能在Google Cloud上使用

1.6.3 Flink和Beam理解测验（选择题）



第11题

根据课程内容，如果一家公司需要构建一个既支持批处理又支持流处理的大数据平台，并且希望避免vendor lock-in（供应商锁定），最合理的技术选择是什么？

A. 只使用Apache Spark，因为它社区支持好

B. 只使用Apache Flink，因为Google支持它

C. 使用Apache Beam作为统一编程接口，可以在Spark、Flink等多个引擎上运行

D. 同时部署Spark和Flink两套独立系统

1.6.3 Flink和Beam理解测验（选择题）



第12题

从Flink（2008年）和Spark（2009年）的发展历程来看，为什么Spark在市场上的接受度一度超过Flink，但Google后来却选择支持Flink？这说明了什么技术趋势？

A. Google的选择是随机的，没有特别意义

B. 早期市场成熟度和社区生态比技术优势更重要，但随着实时流处理需求增长，技术架构的先进性变得更关键

C. Flink比Spark晚一年，所以技术一定更好

D. 只是因为Flink是欧洲开发的，Google想支持国际化

1.6.3 Flink和Beam理解测验（选择题）



第13题

观察第11页的性能对比图，Flink和Spark在不同节点数量下的性能曲线几乎重叠，但Hadoop MapReduce的增长曲线明显更陡。这个现象最合理的技术解释是什么？

A. Flink和Spark都使用了相同的代码库
B. Flink和Spark都采用内存计算模型，具有相似的扩展性特征；而MapReduce的磁盘I/O成为瓶颈，导致扩展性较差
C. 图表数据有误，应该Flink性能最好
D. 节点数量对性能没有影响

1.6.3 Flink和Beam理解测验（选择题）



第14题

根据第9页的Flink生态系统图，如果要将现有的Hadoop生态系统迁移到Flink，同时保持对HDFS和Kafka的支持，需要重点关注哪些层面的兼容性？

A. 只需要关注编程API的兼容性
B. 需要同时考虑Data Access层（数据源连接）和Deployment层（部署环境）的兼容性
C. 只需要更换核心运行时
D. 完全不需要考虑兼容性，直接替换即可

1.6.3 Flink和Beam理解测验（选择题）



第15题

综合考虑课程内容，以下哪个场景最适合使用"Beam + Flink Runner"的组合，而不是直接使用Flink或Spark？

A. 一个只需要简单批处理的离线数据分析项目
B. 一个需要在多个云平台（Google Cloud、AWS等）上部署，且对延迟要求极高的实时推荐系统
C. 一个只在本地开发环境运行的原型项目
D. 一个完全不需要流处理的数据仓库项目



Thank you