《Spark编程基础》

教材官网： http://dblab.xmu.edu.cn/post/spark/

温馨提示：编辑幻灯片母版，可以修改每页PPT的厦大校徽和底部文字

第1章 大数据技术概述

（PPT版本号：2018年春季学期）

扫一扫访问教材官网

林子雨

厦门大学计算机科学系

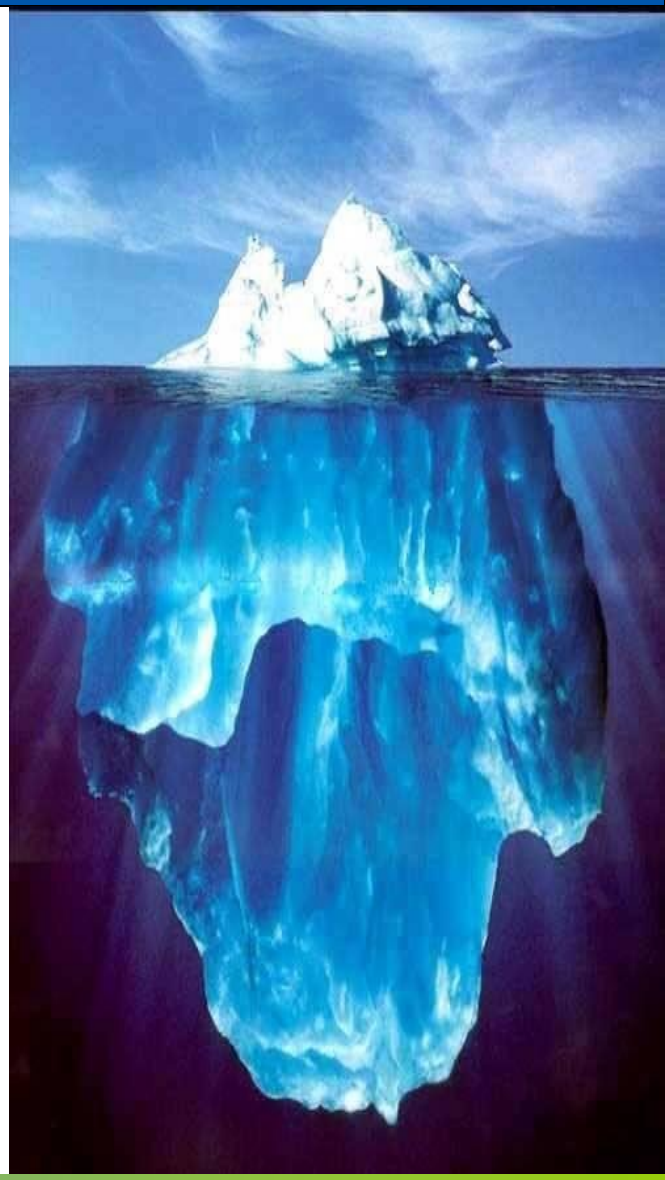
E-mail: ziyulin@xmu.edu.cn

主页：**http://www.cs.xmu.edu.cn/linziyu**

课程配套授课视频

课程在线视频地址：http://dblab.xmu.edu.cn/post/10482/

提纲

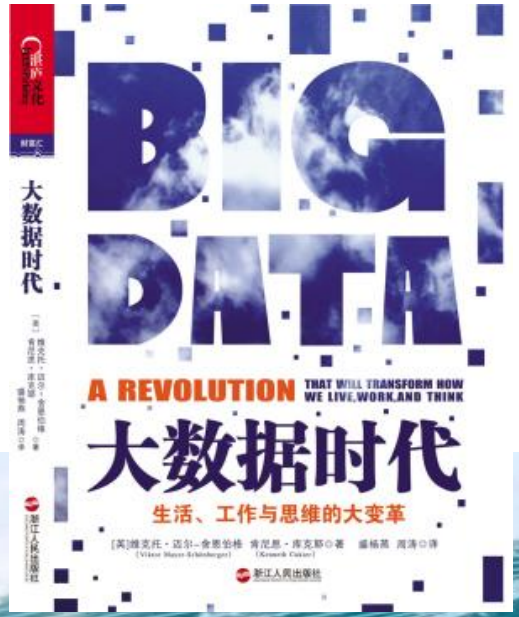


高校大数据课程
公 共 服 务 平 台

百度搜索厦门大学数据库实验室网站访问平台

1.1大数据时代


BIG
DATA
SOFTWARE
ANALYTICS
TECHNOLOGIES
INFORMATION
STORAGE
COMPLEX
DATABASES
INTERNET
SIZE
NEEDED
PETABYTES
MANAGEMENT
SOCIAL
LARGE
EVERY
RESEARCH
SHARED
TIME
INCLUDE
SYSTEMS
NETWORKS
SEARCH
RECORDS
DISK
HUNDREDS
GARTNER
MASSIVELY
SAN
PARALLEL
EXAMPLES
ABILITY


大数据时代
BIG
DATA
A REVOLUTION
THAT WILL TRANSFORM HOW
WE LIVE,WORK,AND THINK
大数据时代
生活、工作与思维的大变革
浙江人民出版社

1.1.1第三次信息化浪潮

- 根据IBM前首席执行官郭士纳的观点，IT领域每隔十五年就会迎来一次重大变革

表1 三次信息化浪潮

信息化浪潮	发生时间	标志	解决问题	代表企业
第一次浪潮	1980年前后	个人计算机	信息处理	Intel、AMD、IBM、苹果、微软、联想、戴尔、惠普等
第二次浪潮	1995年前后	互联网	信息传输	雅虎、谷歌、阿里巴巴、百度、腾讯等
第三次浪潮	2010年前后	物联网、云计算和大数据	信息爆炸	将涌现出一批新的市场标杆企业

1.1.2信息科技为大数据时代提供技术支撑

1. 存储设备容量不断增加

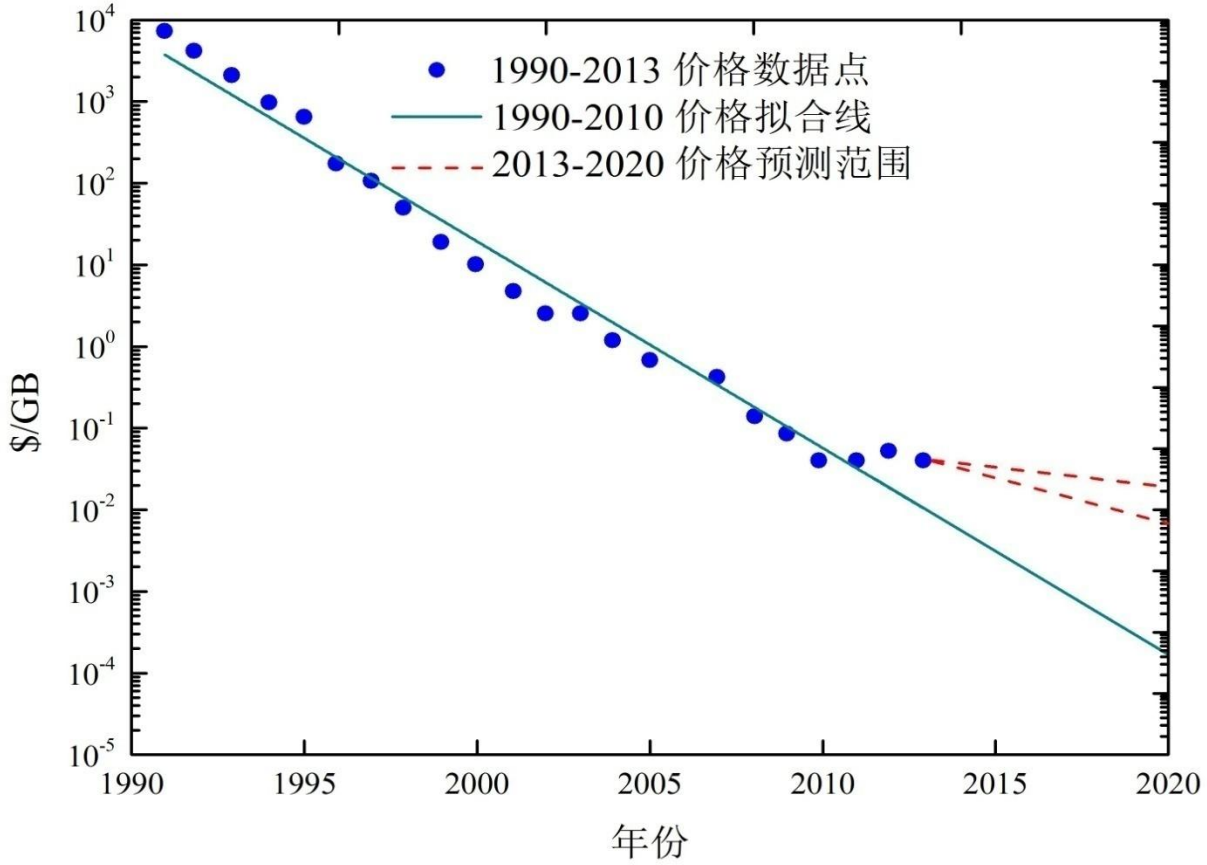


图1-1 存储价格随时间变化情况

1.1.2信息科技为大数据时代提供技术支撑

2. CPU处理能力大幅提升

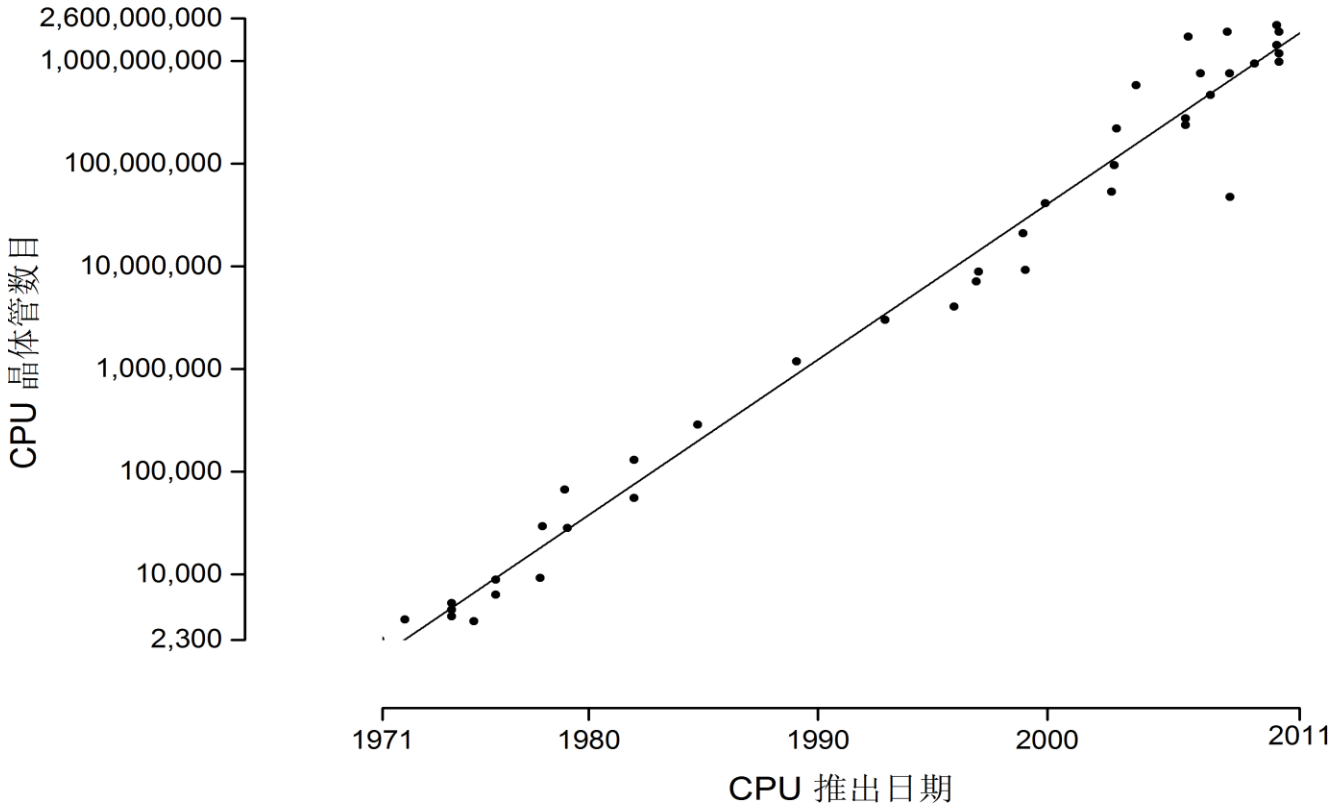


图 CPU晶体管数目随时间变化情况

1.1.2信息科技为大数据时代提供技术支撑

3. 网络带宽不断增加

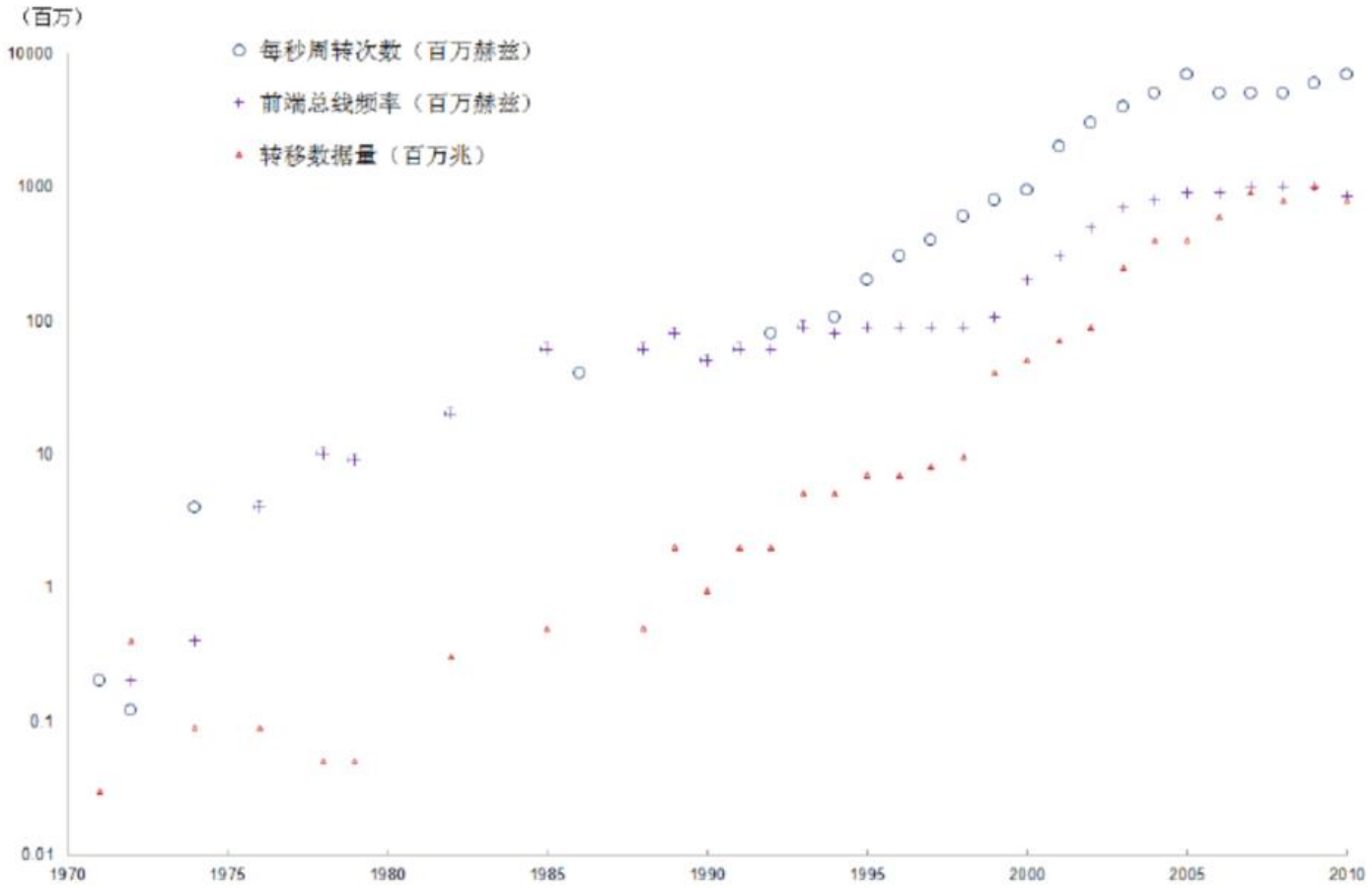


图 网络带宽随时间变化情况

1.1.3数据产生方式的变革促成大数据时代的来临

1 运营式系统阶段

•数据库的出现使得数据管理的复杂度大大降低，数据往往伴随着一定的运营活动而产生并记录在数据库中，数据的产生方式是被动的

2 用户原创内容阶段

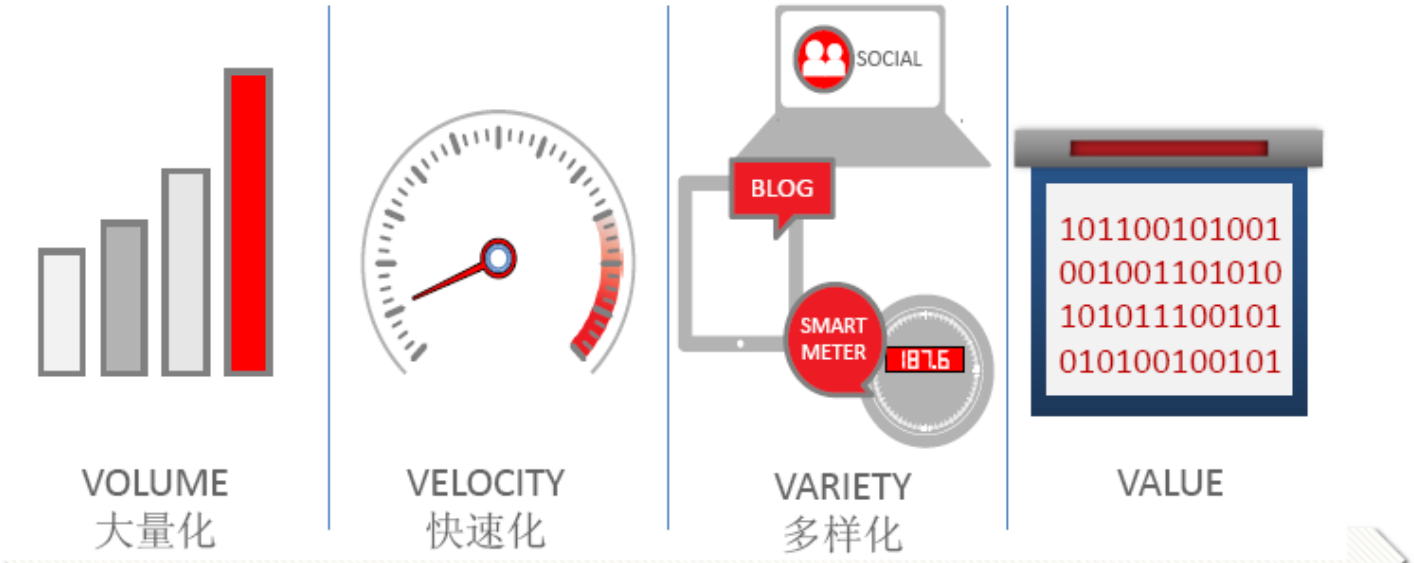
•数据爆发产生于Web 2.0时代，而Web 2.0的最重要标志就是用户原创内容
•智能手机等移动设备加速内容产生
•数据产生方式是主动的

3 感知式系统阶段

•感知式系统的广泛使用
•人类社会数据量第三次大的飞跃最终导致了大数据的产生

图 数据产生方式的变革

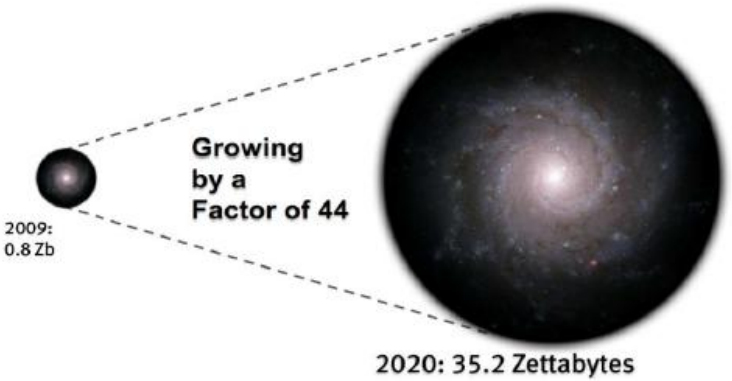
1.2大数据概念

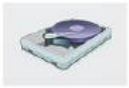


大数据不仅仅是数据的“大量化”，而是包含“快速化”、“多样化”和“价值化”等多重属性。

1.2.1数据量大

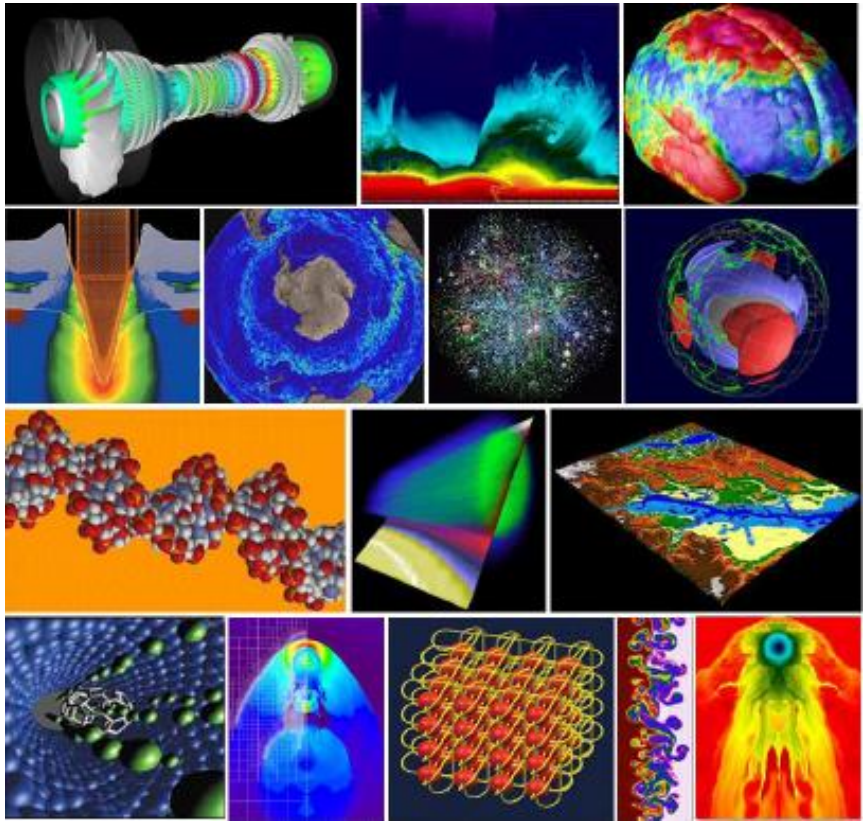
- 根据IDC作出的估测，数据一直都在以每年50%的速度增长，也就是说每两年就增长一倍（大数据摩尔定律）
- 人类在最近两年产生的数据量相当于之前产生的全部数据量
- 预计到2020年，全球将总共拥有35ZB的数据量，相较于2010年，数据量将增长近30倍



TERABYTE	10 的 12 次方	一块 1TB 硬盘		200,000 照片或 mp3 歌曲
PETABYTE	10 的 15 次方	两个数据中心机柜		16 个 Blackblaze pod 存储单元
EXABYTE	10 的 18 次方	2,000 个机柜		占据一个街区的 4 层数据中心
ZETTABYTE	10 的 21 次方	1000 个数据中心		纽约曼哈顿的 1/5 区域
YOTTABYTE	10 的 24 次方	一百万个数据中心		特拉华州和罗德岛州

1.2.2 数据类型繁多

- 大数据是由结构化和非结构化数据组成的
  - 10%的结构化数据，存储在数据库中
  - 90%的非结构化数据，它们与人类信息密切相关



- 科学研究
  - –基因组
  - –LHC 加速器
  - –地球与空间探测
- 企业应用
  - –Email、文档、文件
  - –应用日志
  - –交易记录
- Web 1.0数据
  - –文本
  - –图像
  - –视频
- **Web 2.0**数据
  - –查询日志**/**点击流
  - **–Twitter/ Blog / SNS**
  - **–Wiki**

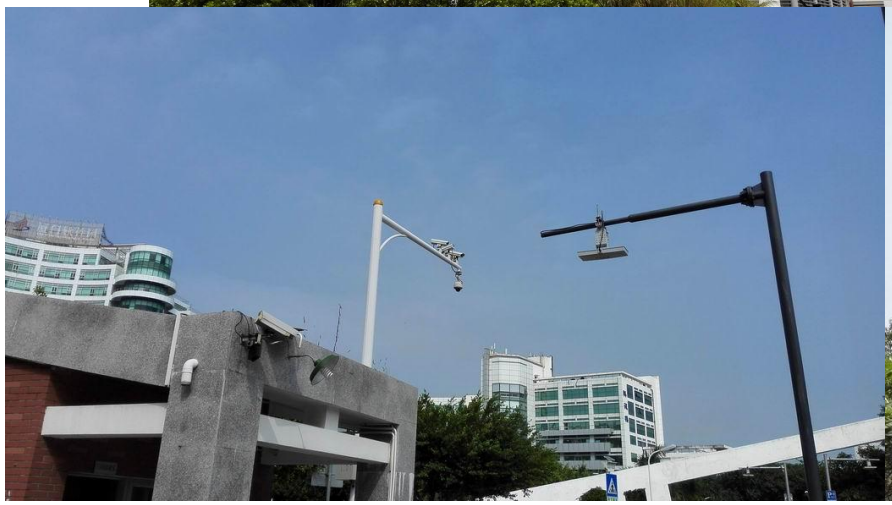
1.2.3处理速度快

- 从数据的生成到消耗，时间窗口非常小，可用于生成决策的时间非常少
- 1秒定律：这一点也是和传统的数据挖掘技术有着本质的不同

1.2.4价值密度低

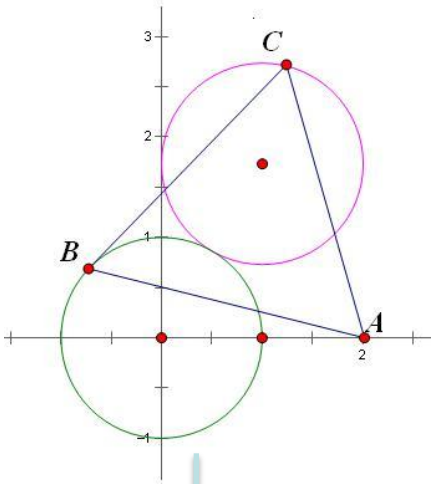
价值密度低，商业价值高

以视频为例，连续不间断监控过程中，可能有用的数据仅仅有一两秒，但是具有很高的商业价值



1.3大数据的影响

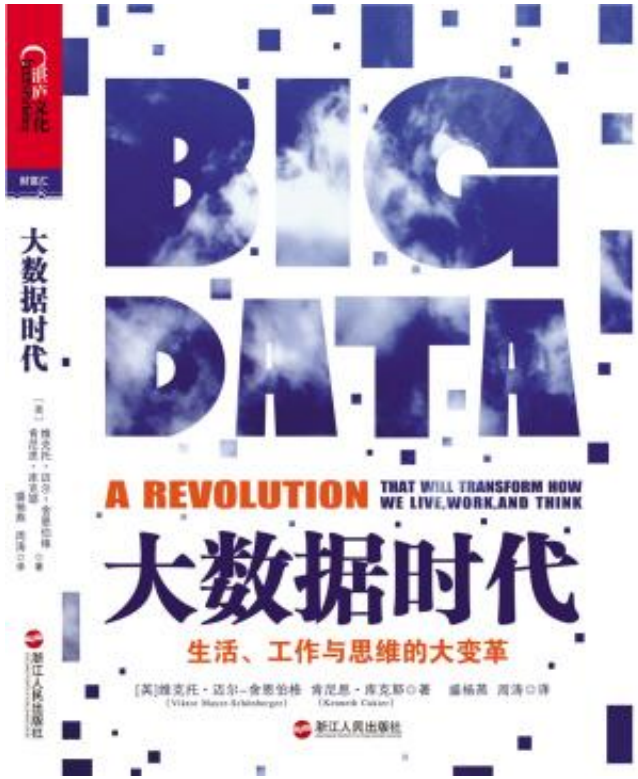
图灵奖获得者、著名数据库专家Jim Gray 博士观察并总结人类自古以来，在科学研究上，先后历经了实验、理论、计算和数据四种范式



实验　　理论　　计算　　数据

1.3大数据的影响

- 在思维方式方面，大数据完全颠覆了传统的思维方式：
  - 全样而非抽样
  - 效率而非精确
  - 相关而非因果



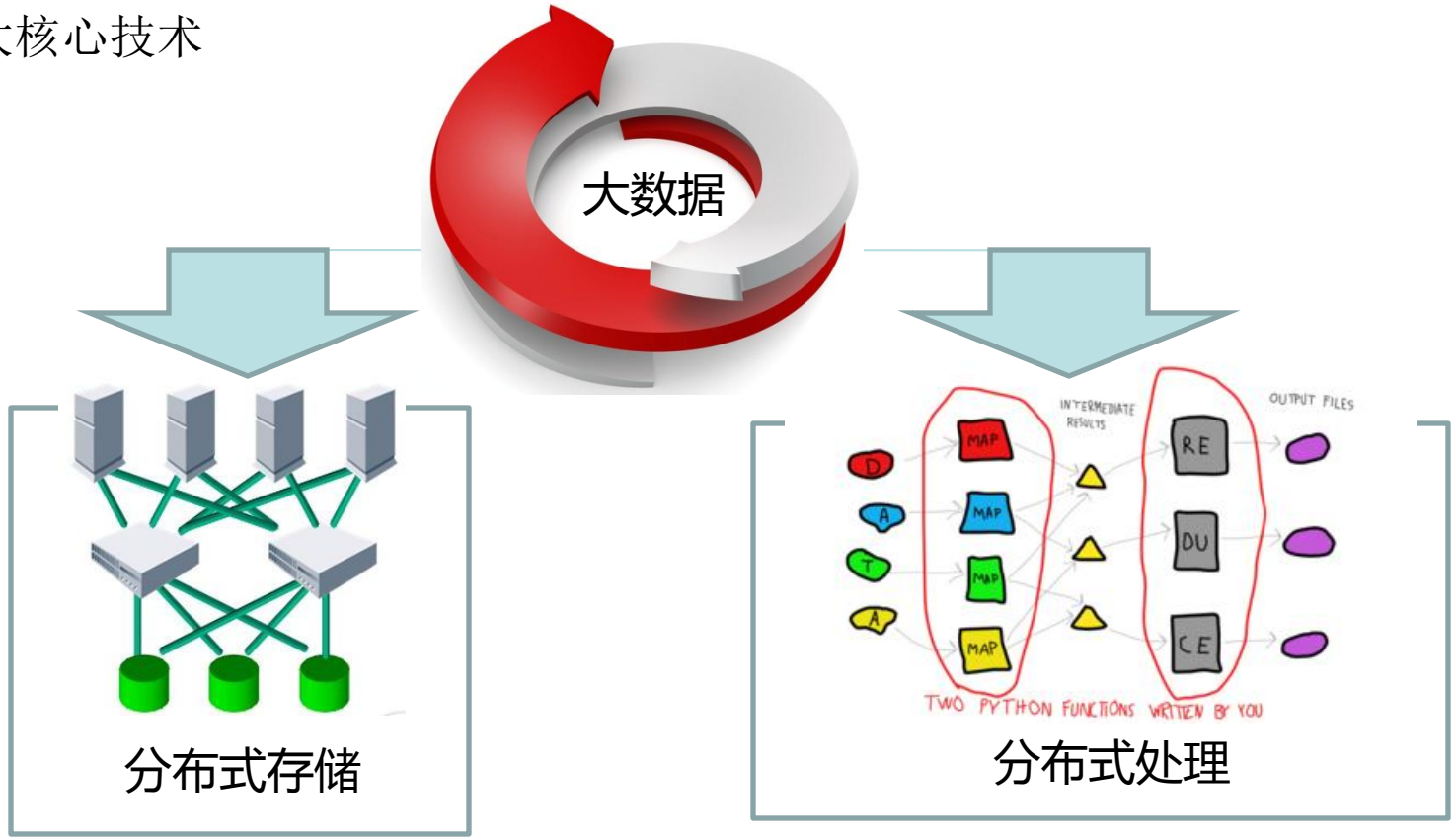
1.4大数据关键技术

表1-5 大数据技术的不同层面及其功能

技术层面	功能
数据采集	利用ETL工具将分布的、异构数据源中的数据如关系数据、平面数据文件等，抽取到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集市中，成为联机分析处理、数据挖掘的基础；或者也可以把实时采集的数据作为流计算系统的输入，进行实时处理分析
数据存储和管理	利用分布式文件系统、数据仓库、关系数据库、NoSQL数据库、云数据库等，实现对结构化、半结构化和非结构化海量数据的存储和管理
数据处理与分析	利用分布式并行编程模型和计算框架，结合机器学习和数据挖掘算法，实现对海量数据的处理和分析；对分析结果进行可视化呈现，帮助人们更好地理解数据、分析数据
数据隐私和安全	在从大数据中挖掘潜在的巨大商业价值和学术价值的同时，构建隐私数据保护体系和数据安全体系，有效保护个人隐私和数据安全

1.4大数据关键技术

两大核心技术



GFS\HDFS
BigTable\HBase
NoSQL（键值、列族、图形、文档数据库）
NewSQL（如：SQL Azure）

MapReduce

1.5 大数据计算模式

表1-3 大数据计算模式及其代表产品

大数据计算模式	解决问题	代表产品
批处理计算	针对大规模数据的批量处理	MapReduce、Spark等
流计算	针对流数据的实时计算	Storm、S4、Flume、Streams、Puma、DStream、Super Mario、银河流数据处理平台等
图计算	针对大规模图结构数据的处理	Pregel、GraphX、Giraph、PowerGraph、Hama、GoldenOrb等
查询分析计算	大规模数据的存储管理和查询分析	Dremel、Hive、Cassandra、Impala等

1.6 代表性大数据技术

1.6.1 Hadoop
1.6.2 Spark
1.6.3 Flink
1.6.4 Beam

1.6.1 Hadoop

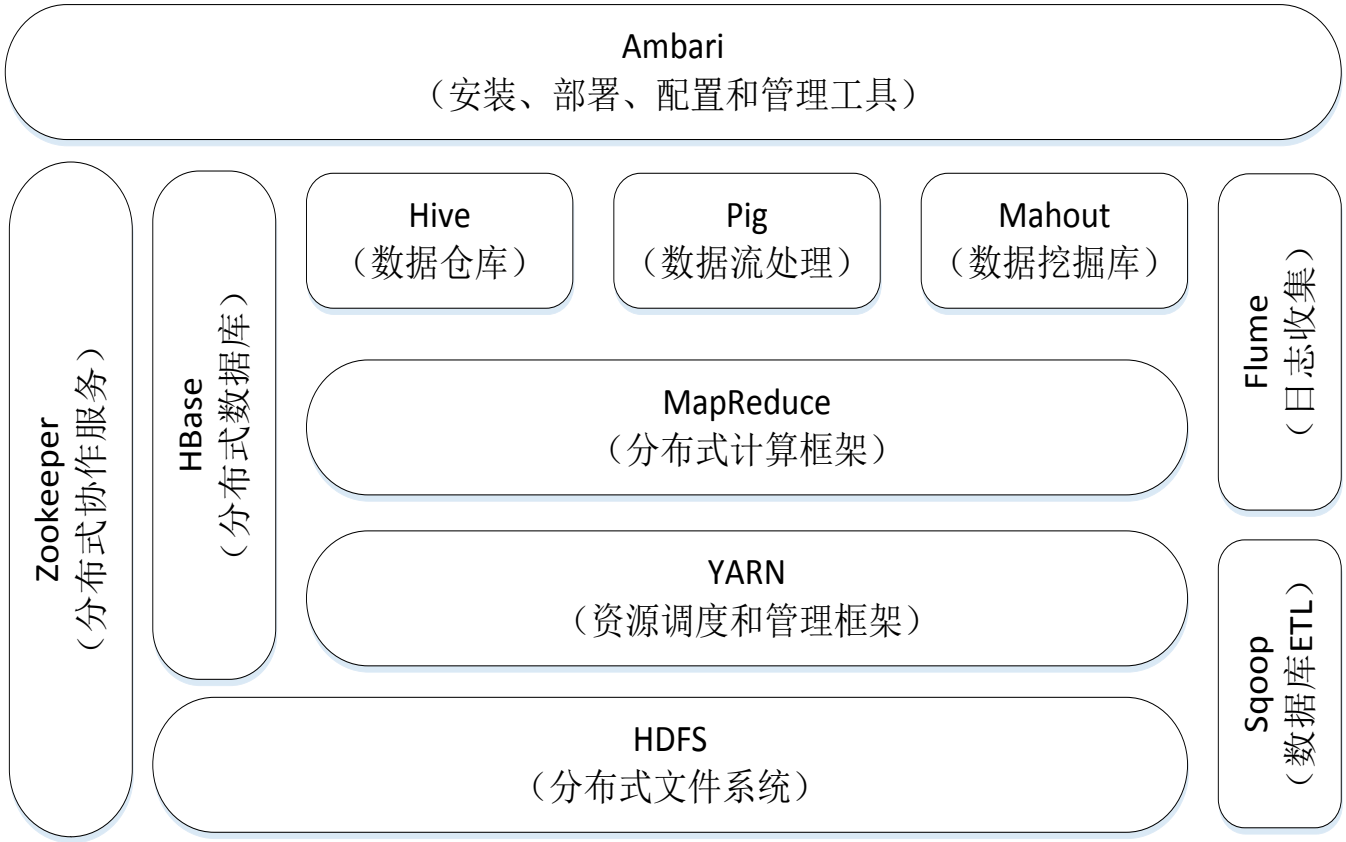


图 Hadoop生态系统

1.6.1 Hadoop——MapReduce

•MapReduce将复杂的、运行于大规模集群上的并行计算过程高度地抽象到了两个函数：Map和Reduce

•编程容易，不需要掌握分布式并行编程细节，也可以很容易把自己的程序运行在分布式系统上，完成海量数据的计算

•MapReduce采用“分而治之”策略，一个存储在分布式文件系统中的大规模数据集，会被切分成许多独立的分片（split），这些分片可以被多个Map任务并行处理

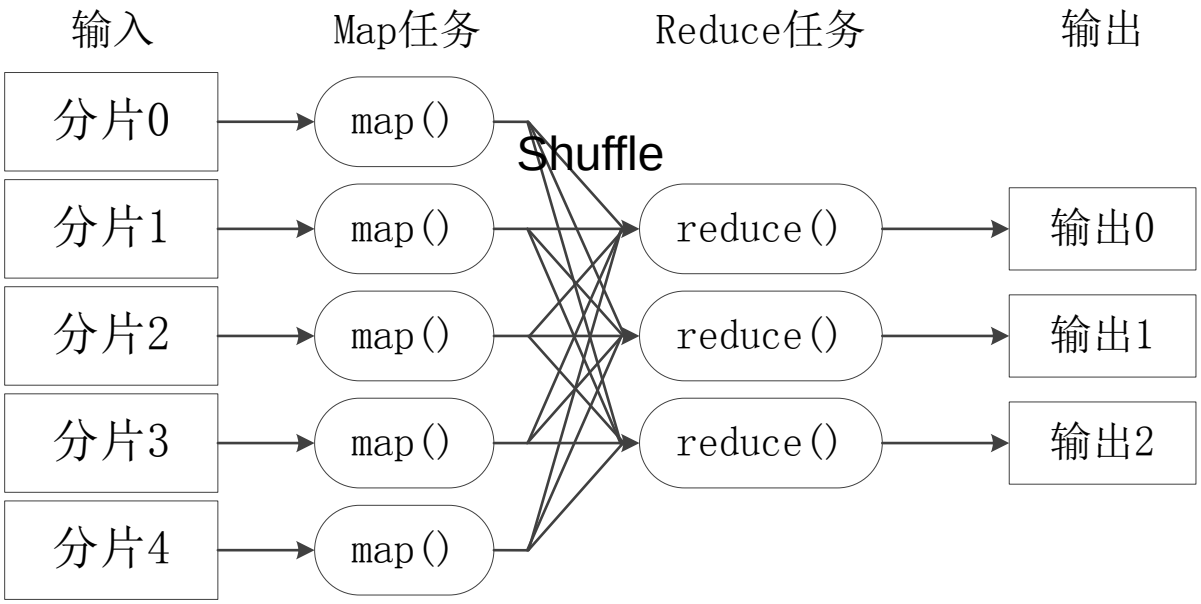


图 MapReduce工作流程

1.6.1 Hadoop——YARN

YARN的目标就是实现“一个集群多个框架”，为什么？

- 一个企业当中同时存在各种不同的业务应用场景，需要采用不同的计算框架
  - MapReduce实现离线批处理
  - 使用Impala实现实时交互式查询分析
  - 使用Storm实现流式数据实时分析
  - 使用Spark实现迭代计算

- 这些产品通常来自不同的开发团队，具有各自的资源调度管理机制
- 为了避免不同类型应用之间互相干扰，企业就需要把内部的服务器拆分成多个集群，分别安装运行不同的计算框架，即“一个框架一个集群”

- 导致问题
  - 集群资源利用率低
  - 数据无法共享
  - 维护代价高

1.6.1 Hadoop——YARN

- YARN的目标就是实现“一个集群多个框架”，即在一个集群上部署一个统一的资源调度管理框架YARN，在YARN之上可以部署其他各种计算框架
- 由YARN为这些计算框架提供统一的资源调度管理服务，并且能够根据各种计算框架的负载需求，调整各自占用的资源，实现集群资源共享和资源弹性收缩
- 可以实现一个集群上的不同应用负载混搭，有效提高了集群的利用率
- 不同计算框架可以共享底层存储，避免了数据集跨集群移动

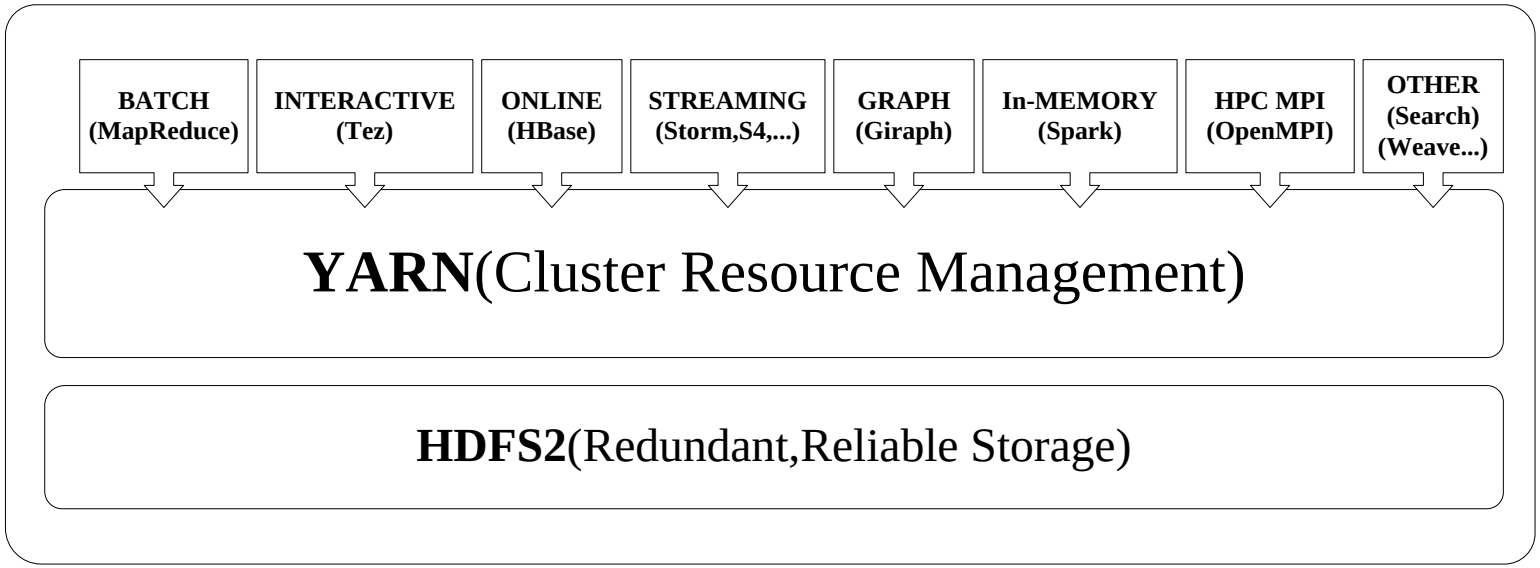
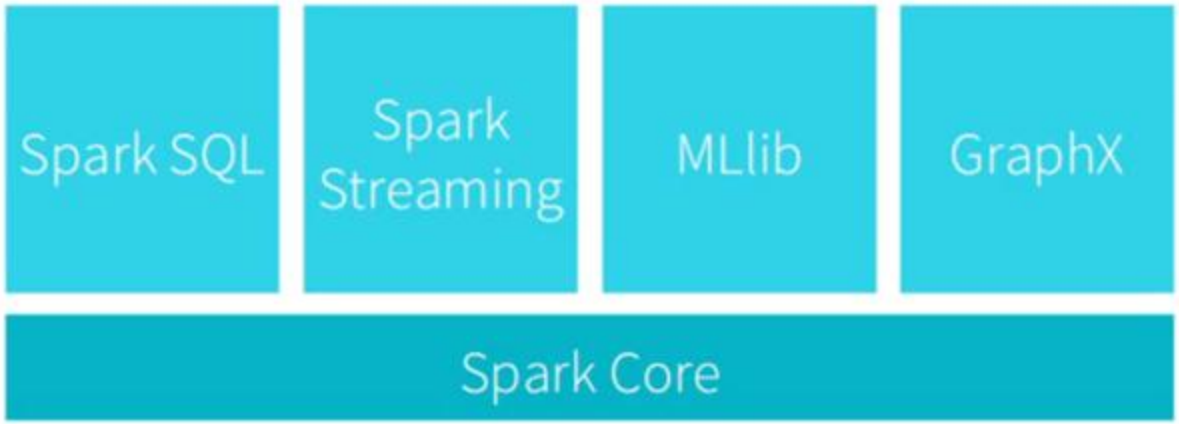


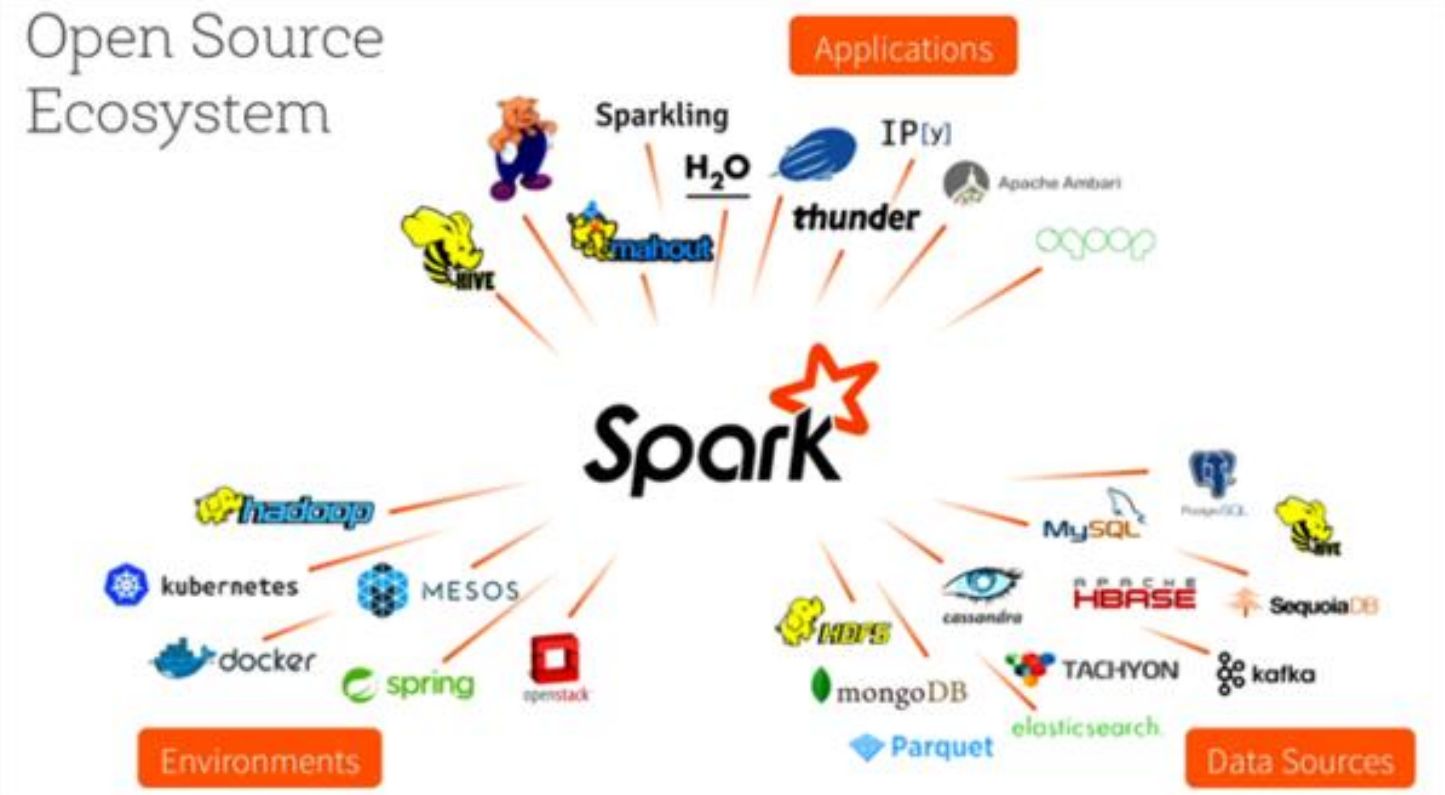
图 在YARN上部署各种计算框架

1.6.2 Spark



Spark架构图

1.6.2 Spark



Spark生态系统

1.6.2 Spark

Hadoop与Spark的对比

Hadoop存在如下一些缺点：

- 表达能力有限
- 磁盘IO开销大
- 延迟高
  - 任务之间的衔接涉及IO开销
  - 在前一个任务执行完成之前，其他任务就无法开始，难以胜任复杂、多阶段的计算任务

1.6.2 Spark

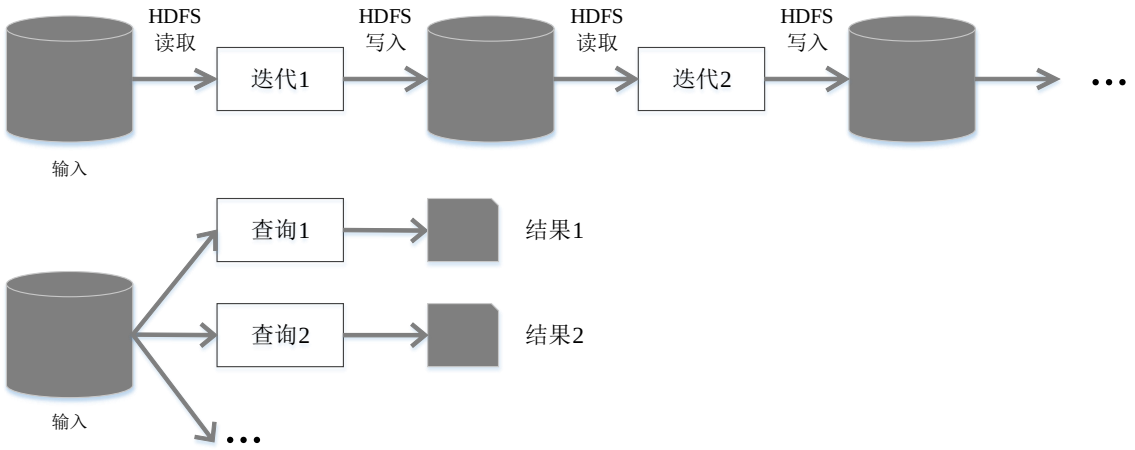
Hadoop与Spark的对比

Spark在借鉴Hadoop MapReduce优点的同时，很好地解决了MapReduce所面临的问题

相比于Hadoop MapReduce，Spark主要具有如下优点：

•Spark的计算模式也属于MapReduce，但不局限于Map和Reduce操作，还提供了多种数据集操作类型，编程模型比Hadoop MapReduce更灵活
•Spark提供了内存计算，可将中间结果放到内存中，对于迭代运算效率更高
Spark基于DAG的任务调度执行机制，要优于Hadoop MapReduce的迭代执行机制

1.6.2 Spark



(a) Hadoop MapReduce执行流程

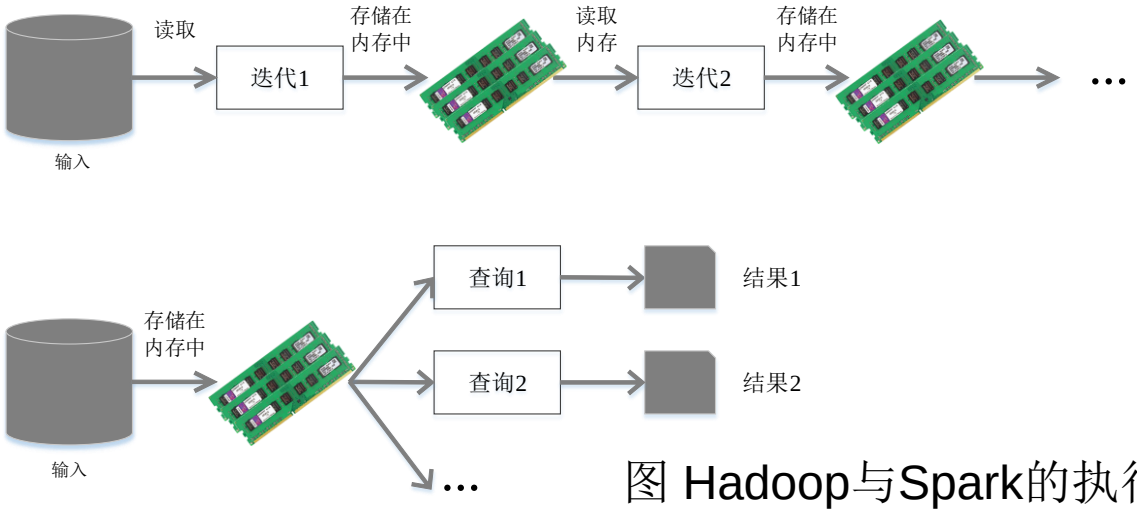


图 Hadoop与Spark的执行流程对比

(b) Spark执行流程

1.6.2 Spark

•使用Hadoop进行迭代计算非常耗资源
•Spark将数据载入内存后，之后的迭代计算都可以直接使用内存中的中间结果作运算，避免了从磁盘中频繁读取数据

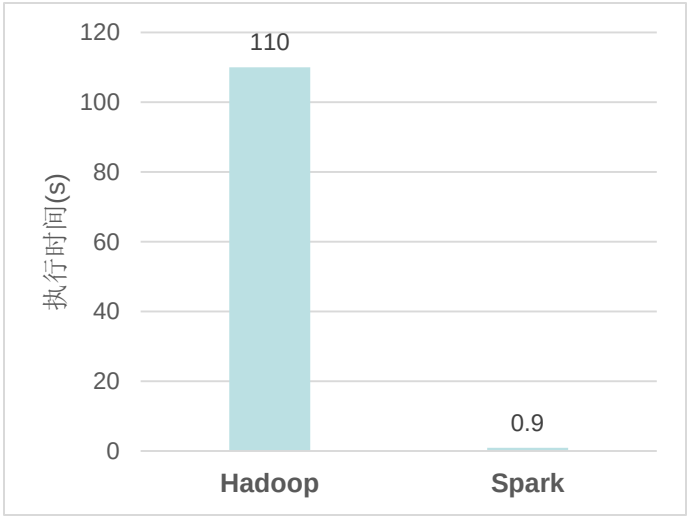


图 Hadoop与Spark执行逻辑回归的时间对比

1.6.2 Spark

问题1：Spark会取代Hadoop吗？

- Hadoop包括两大核心：HDFS和MapReduce
- Spark作为计算框架，与MapReduce是对等的
- 谈到“取代”，Spark应该是取代MapReduce，而不是整个Hadoop
- Spark和Hadoop生态系统共存共荣，Spark借助于Hadoop的HDFS、HBase等来完成数据的存储，然后，由Spark完成数据的计算

1.6.2 Spark

问题2：开发Spark程序应该使用什么编程语言？

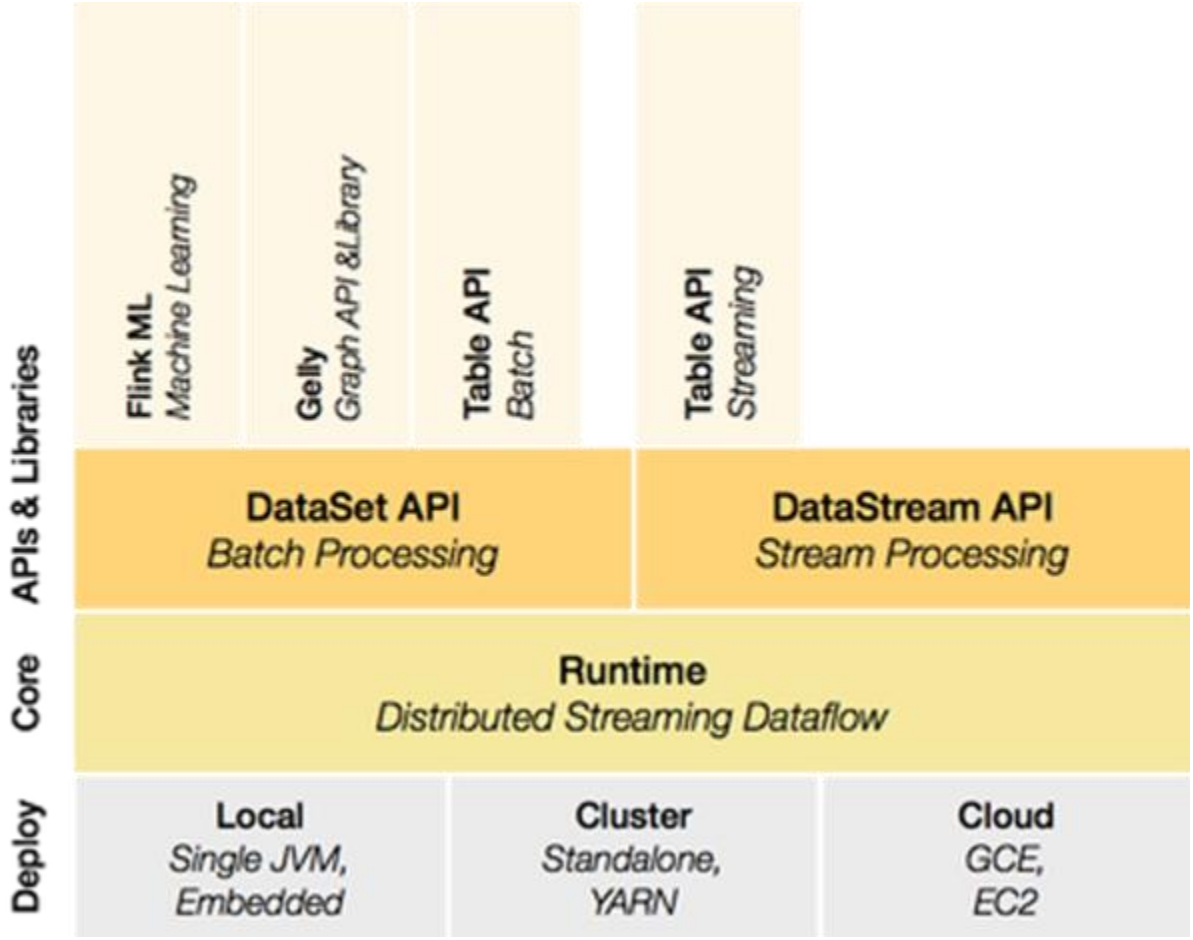
开发Spark应用程序时，可以采用Scala、Python、Java和R等语言，首选语言是Scala，因为Spark这个软件本身就是使用Scala语言开发的，采用Scala语言编写Spark应用程序，可以获得最好的性能。

关于采用哪种语言编写Spark应用程序，这里强调两点：

（1）Java代码太繁琐。在大数据应用场景中，不太适合使用Java，因为，完成同样的任务，Scala只需要一行代码，而Java则可能需要10行代码；而且，Scala语言可以支持交互式编程，大大提高了程序开发效率，而Java则不支持交互式执行，必须编译以后运行。

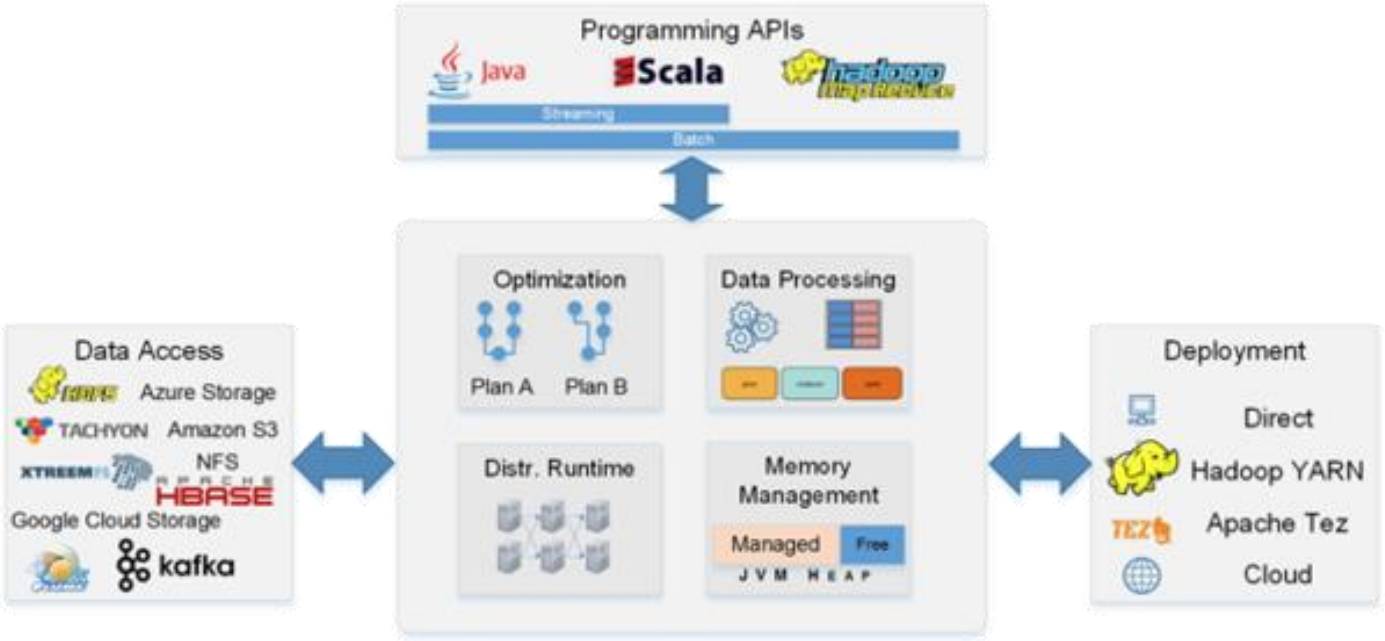
（2）Python语言并发性能不好。在并发性能方面，Scala要明显优于Python，而且，Scala是静态类型，可以在编译阶段就抛出错误，便于开发大型大数据项目，此外，Scala兼容Java，运行在JVM上，可以直接使用Java中的Hadoop API来和Hadoop进行交互，但是，Python与Hadoop之间的交互非常糟糕，通常都需要第三方库（比如hadoopy）。

1.6.3 Flink



Flink架构图

1.6.3 Flink



Flink生态系统

1.6.3 Flink

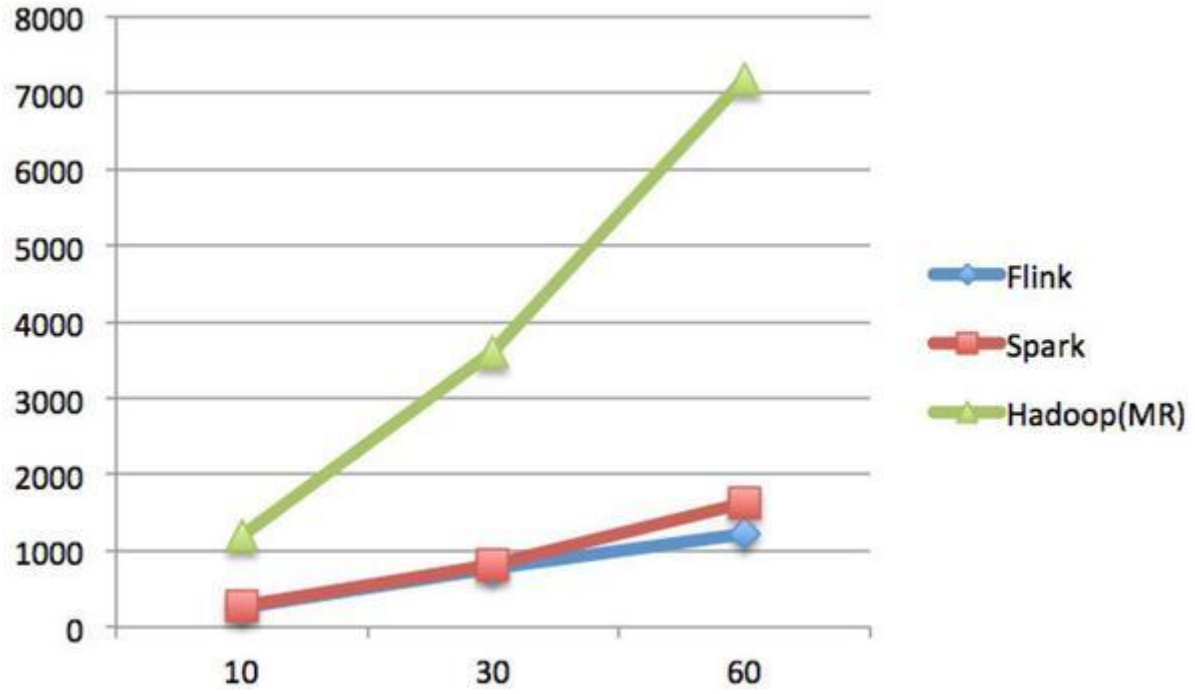
Flink与Spark的比较

	Apache Spark	Apache Flink
核心实现	Scala	Java
编程接口	Java，Python 以及 R 语言	DataSet API 支持 Java、Scala 和 Python。DataStream API 支持 Java and Scala
计算模型	Spark 是基于数据片集合（RDD）进行小批量处理，采用了微批处理模型	Flink 是一行一行处理，基于操作符的连续流模型。
优缺点	在流式处理方面，不可避免增加一些延时，Spark 则只能支持秒级计算。	Flink 的流式计算跟 Storm 性能差不多，支持毫秒级计算

1.6.3 Flink

性能对比

首先它们都可以基于内存计算框架进行实时计算，所以都拥有非常好的计算性能。经过测试，Flink计算性能上略好。



Spark和Flink全部都运行在Hadoop YARN上，性能为Flink > Spark > Hadoop(MR)，迭代次数越多越明显，性能上，Flink优于Spark和Hadoop最主要的原因是Flink支持增量迭代，具有对迭代自动优化的功能。

1.6.3 Flink

流式计算比较

它们都支持流式计算，Flink是一行一行处理，而Spark是基于数据片集合（RDD）进行小批量处理，所以Spark在流式处理方面，不可避免增加一些延时。Flink的流式计算跟Storm性能差不多，支持毫秒级计算，而Spark则只能支持秒级计算。

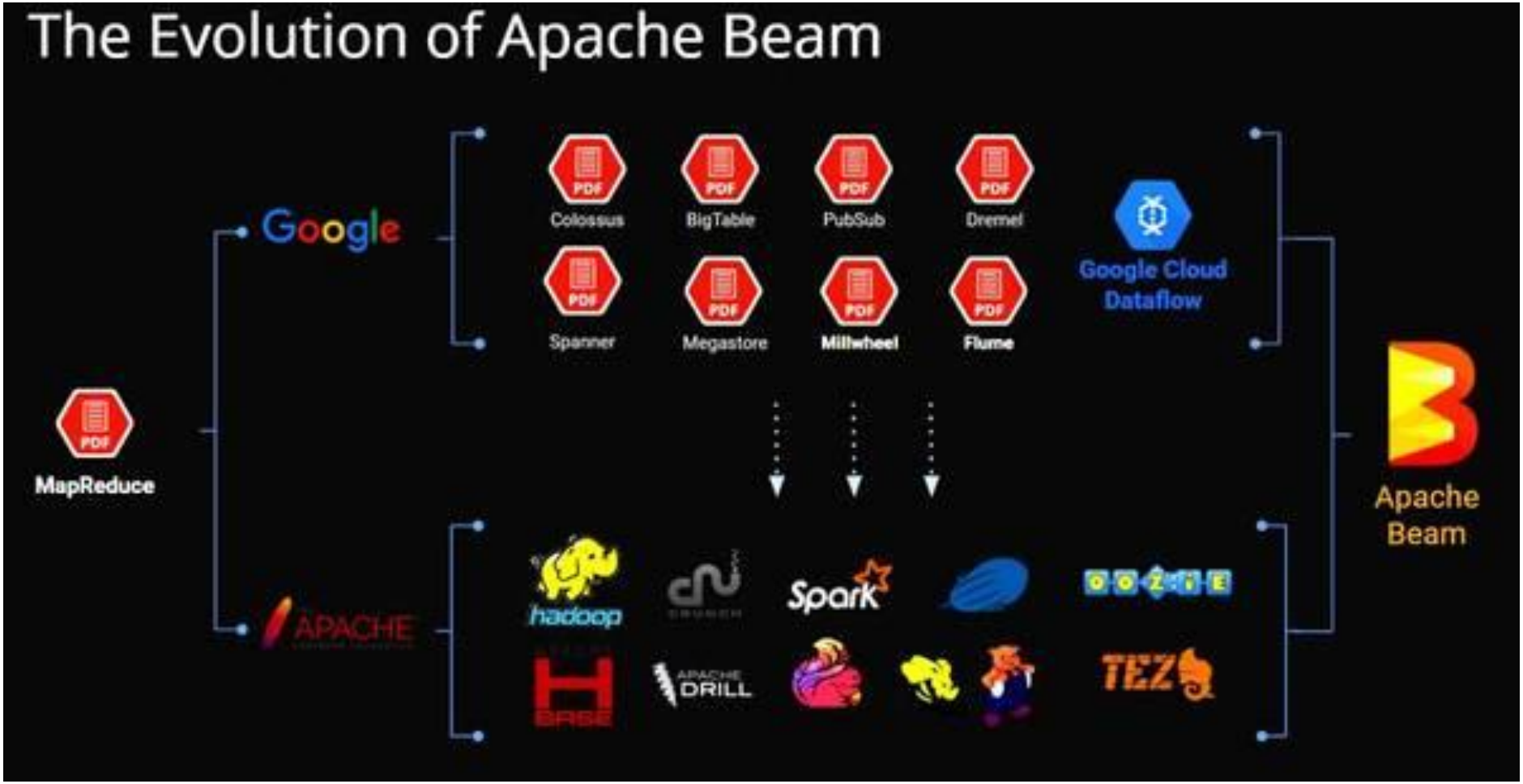
SQL支持

都支持SQL，Spark对SQL的支持比Flink支持的范围要大一些，另外Spark支持对SQL的优化，而Flink支持主要是对API级的优化。

既生瑜，何生亮！

1.6.4 Beam

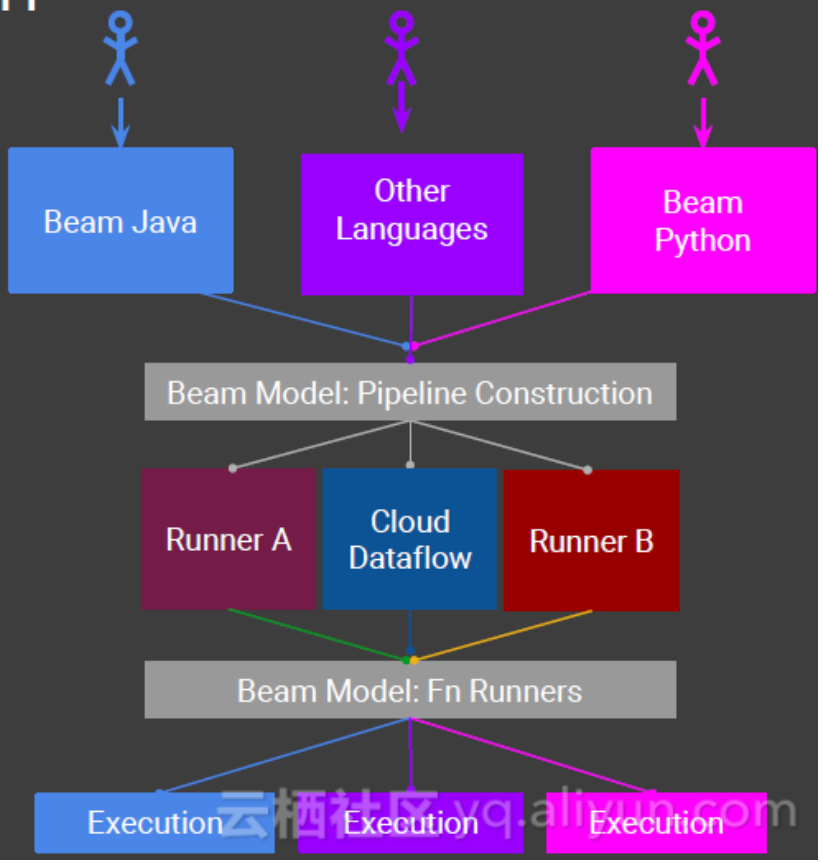
谷歌，Beam，一统天下？



1.6.4 Beam

Apache Beam Technical Vision

1. **End users:** who want to write pipelines or transform libraries in a language that's familiar.
2. **SDK writers:** who want to make Beam concepts available in new languages.
3. **Runner writers:** who have a distributed processing environment and want to support Beam pipelines



附录A：主讲教师林子雨简介

主讲教师：林子雨

单位：厦门大学计算机科学系
E-mail: ziyulin@xmu.edu.cn
个人网页：http://www.cs.xmu.edu.cn/linziyu
数据库实验室网站：http://dblab.xmu.edu.cn

扫一扫访问个人主页

林子雨，男，1978年出生，博士（毕业于北京大学），现为厦门大学计算机科学系助理教授（讲师），曾任厦门大学信息科学与技术学院院长助理、晋江市发展和改革局副局长。中国计算机学会数据库专业委员会委员，中国计算机学会信息系统专业委员会委员。国内高校首个“数字教师”提出者和建设者，厦门大学数据库实验室负责人，厦门大学云计算与大数据研究中心主要建设者和骨干成员，2013年度和2017年度厦门大学教学类奖教金获得者，荣获2017年福建省精品在线开放课程和2017年厦门大学高等教育成果二等奖。主要研究方向为数据库、数据仓库、数据挖掘、大数据、云计算和物联网，并以第一作者身份在《软件学报》《计算机学报》和《计算机研究与发展》等国家重点期刊以及国际学术会议上发表多篇学术论文。作为项目负责人主持的科研项目包括1项国家自然科学青年基金项目(No.61303004)、1项福建省自然科学青年基金项目(No.2013J05099)和1项中央高校基本科研业务费项目(No.2011121049)，主持的教改课题包括1项2016年福建省教改课题和1项2016年教育部产学协作育人项目，同时，作为课题负责人完成了国家发改委城市信息化重大课题、国家物联网重大应用示范工程区域试点泉州市工作方案、2015泉州市互联网经济调研等课题。中国高校首个“数字教师”提出者和建设者，2009年至今，“数字教师”大平台累计向网络免费发布超过500万字高价值的研究和教学资料，累计网络访问量超过500万次。打造了中国高校大数据教学知名品牌，编著出版了中国高校第一本系统介绍大数据知识的专业教材《大数据技术原理与应用》，并成为京东、当当网等网店畅销书籍；建设了国内高校首个大数据课程公共服务平台，为教师教学和学生学习大数据课程提供全方位、一站式服务，年访问量超过100万次。

附录B：大数据学习路线图

大数据学习路线图访问地址：http://dblab.xmu.edu.cn/post/10164/

附录C：《大数据技术原理与应用》教材

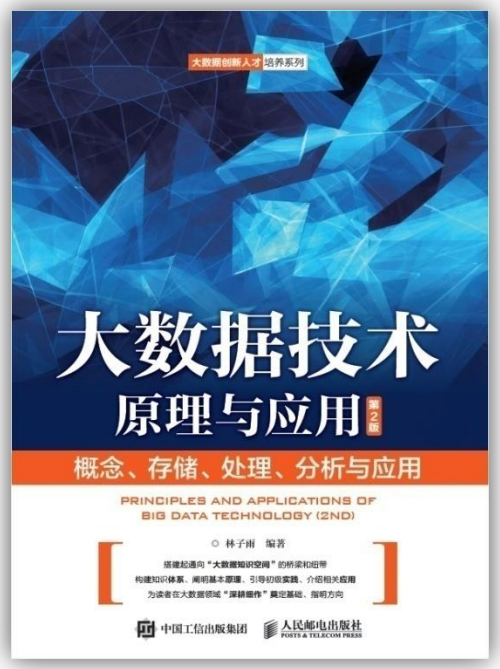
《大数据技术原理与应用——概念、存储、处理、分析与应用（第2版）》，由厦门大学计算机科学系林子雨老师编著，是国内高校第一本系统介绍大数据知识的专业教材。人民邮电出版社 ISBN:978-7-115-44330-4 定价：49.80元

全书共有15章，系统地论述了大数据的基本概念、大数据处理架构Hadoop、分布式文件系统HDFS、分布式数据 库HBase、NoSQL数据库、云数据库、分布式并行编程模型MapReduce、Spark、流计算、图计算、数据可视化以及大数据在互联网、生物医学和物流等各个领域的应用。在Hadoop、HDFS、HBase和MapReduce等重要章节，安排了入门级的实践操作，让读者更好地学习和掌握大数据关键技术。

本书可以作为高等院校计算机专业、信息管理等相关专业的大数据课程教材，也可供相关技术人员参考、学习、培训之用。

欢迎访问《大数据技术原理与应用——概念、存储、处理、分析与应用》教材官方网站：
http://dblab.xmu.edu.cn/post/bigdata

扫一扫访问教材官网



附录D：《大数据基础编程、实验和案例教程》

本书是与《大数据技术原理与应用（第2版）》教材配套的唯一指定实验指导书

大数据教材

1+1黄金组合

厦门大学林子雨编著

配套实验指导书

- 步步引导，循序渐进，详尽的安装指南为顺利搭建大数据实验环境铺平道路
- 深入浅出，去粗取精，丰富的代码实例帮助快速掌握大数据基础编程方法
- 精心设计，巧妙融合，五套大数据实验题目促进理论与编程知识的消化和吸收
- 结合理论，联系实际，大数据课程综合实验案例精彩呈现大数据分析全流程

清华大学出版社 ISBN:978-7-302-47209-4 定价：59元

附录E：《Spark编程基础》

《Spark编程基础》

厦门大学 林子雨，赖永炫，陶继平 编著

披荆斩棘，在大数据丛林中开辟学习捷径
填沟削坎，为快速学习Spark技术铺平道路
深入浅出，有效降低Spark技术学习门槛
资源全面，构建全方位一站式在线服务体系

人民邮电出版社出版发行，ISBN:978-7-115-47598-5
教材官网：http://dblab.xmu.edu.cn/post/spark/

本书以Scala作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Scala语言基础、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、授课视频、上机实验指南等。

附录F：高校大数据课程公共服务平台

高校大数据课程

公 共 服 务 平 台

http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/

扫一扫访问平台主页

扫一扫观看3分钟FLASH动画宣传片

Department of Computer Science, Xiamen University, 2018