大数据公开课全国高校巡讲计划

大数据概念、关键技术及其产业化应用案例

林子雨 博士/助理教授
厦门大学计算机科学系
ziyulin@xmu.edu.cn
http://www.cs.xmu.edu.cn/linziyu

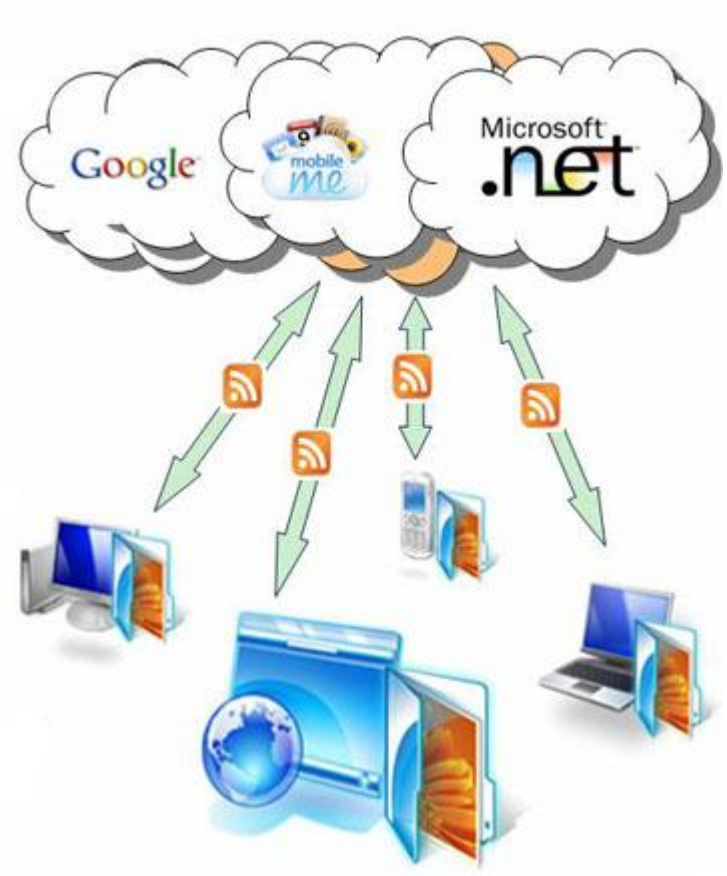
目录 Contents

一 大数据概念

第三次信息化浪潮

2010年前后，以云计算、大数据、物联网的普及为标志
迎来第三次信息化浪潮

云计算



云计算示意图

云计算概念

•通过整合、管理、调配分布在网络各处的计算资源，通过互联网以统一界面，同时向大量的用户提供服务

云计算特点

▪超大规模计算、虚拟化、高可靠性和安全性、通用性、动态扩展性、按需服务、降低成本

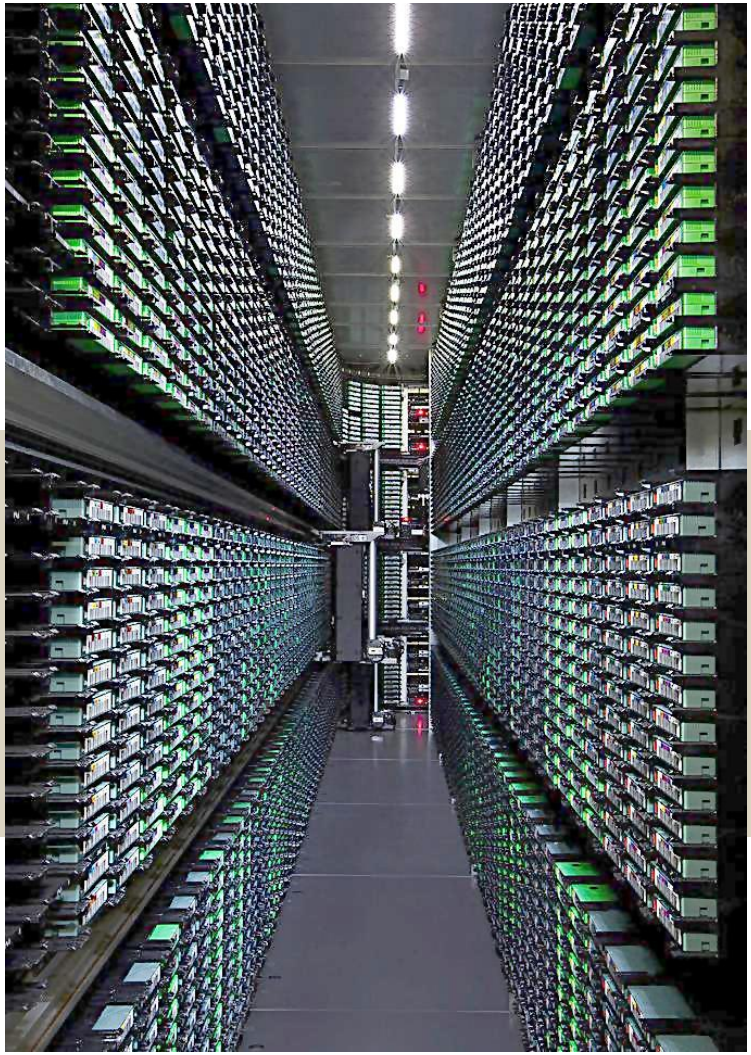
云计算

企业不需要自建IT基础设施，可以租用云端资源

无需自建

企业用户

租用云端资源



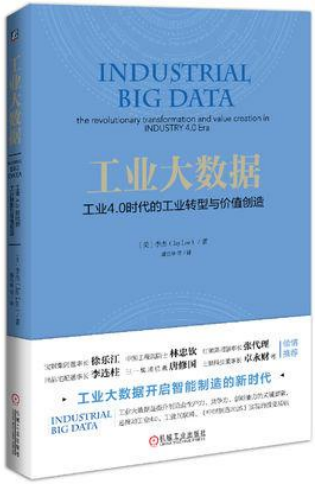
- 数据中心是云计算的温床
- 云计算推动数据中心向虚拟化和云架构的转型，不断提高IT基础架构的灵活性，以降低IT、能源和空间成本，从而让客户能够快速地提高业务敏捷性
- **“西数东输”**
- **安溪-中国国际信息技术（福建）产业园**

“那些正在兴建最大规模数据中心的公司一般都在云计算方面拥有巨大野心”

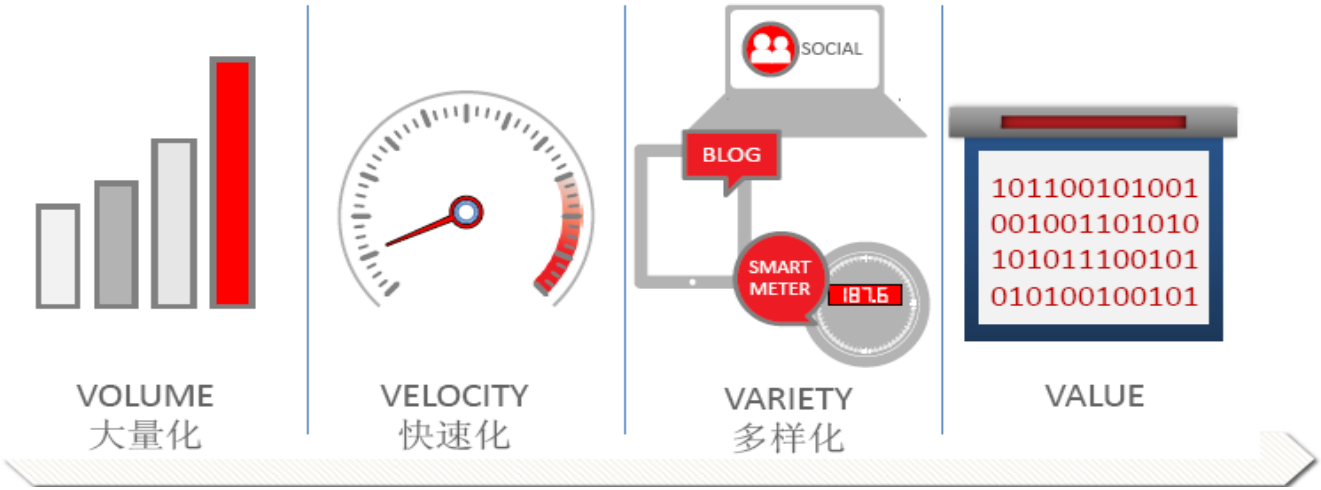
-- 《数据中心知识》杂志主编 Rich Miller

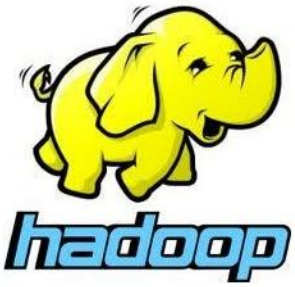
一 大数据概念


BIG DATA


INDUSTRIAL
BIG DATA
工业大数据
工业4.0时代的工业转型与价值创造


BIG
DATA
A REVOLUTION
大数据时代
生活、工作与思维的大变革


SOCIAL
BLOG
SMART
METER
101100101001
001001101010
101011100101
010100100101
VOLUME
大量化
VELOCITY
快速化
VARIETY
多样化
VALUE


hadoop

物联网

1998年

- MIT的 Kevin Ashton第一次提出：把RFID技术与传感器技术应用于日常物品中形成一个“物联网”

2005年

- ITU报告：物联网是通过RFID和智能计算等技术实现全世界设备互连的网络

2008年

- IBM提出：把传感器设备安装到各种物体中，并且普遍连接形成网络，即“物联网”，进而在此基础上形成“智慧地球

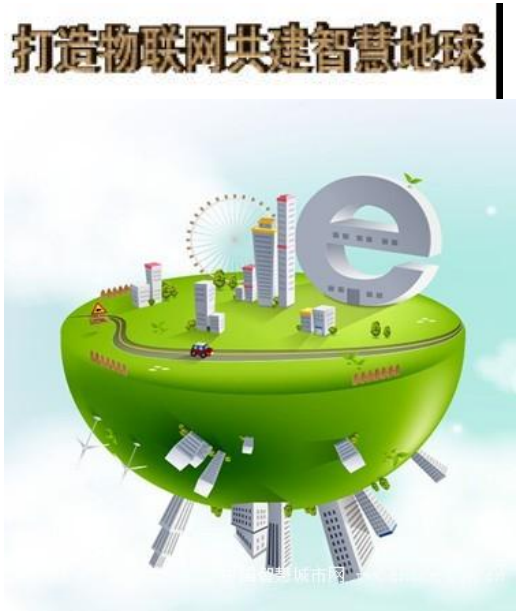
2009年

- 欧洲物联网研究项目工作组制订《物联网战略研究路线图》，介绍传感网/RFID等前端技术和20年发展趋势

物联网形式早已存在，统一意义上的物联网概念提出是架构在互联网发展成熟的基础上

物联网

智慧地球也称为智能地球，就是把感应器嵌入和装备到电网、铁路、桥梁、隧道、公路、建筑、供水系统、大坝、油气管道等各种物体中，并且被普遍连接，形成所谓“物联网”，然后将“物联网”与现有的互联网整合起来，实现人类社会与物理系统的整合。智慧地球的核心是以一种更加智慧的方法，通过利用新一代信息技术改变政府、公司和人们相互交互的方式，以便提高交互的明确性、效率、灵活性和响应速度。



在IBM《智慧地球赢在中国》计划书中，IBM为中国量身打造了六大智慧解决方案：“智慧电力”、“智慧医疗”、“智慧城市”、“智慧交通”、“智慧供应链”和“智慧银行”。

从互联网到物联网

随着网络覆盖的普及，人们提出了一个问题，既然无处不在的网络能够成为人际间沟通的无所不能的工具，为什么我们不能将网络作为物体与物体沟通的工具，人与物体沟通的工具，乃至人与自然沟通的工具？

——中国移动通信集团公司原总经理 王建宙《从互联网到“物联网”》

物联网（IoT：The Internet of Things)物联网就是物物相连的互联网，是互联网的延伸。是利用局部网络或互联网等通信技术把传感器、控制器、机器、人员和物等通过新的方式联在一起，形成人与物、物与物相联，实现信息化、远程管理控制和智能化的网络。

物联网时代示意图：万物相联。例如：当司机出现操作失误时汽车会自动报警；公文包会提醒主人忘带了什么东西等等

计算机:自动计算

以独立计算机进行信息处理

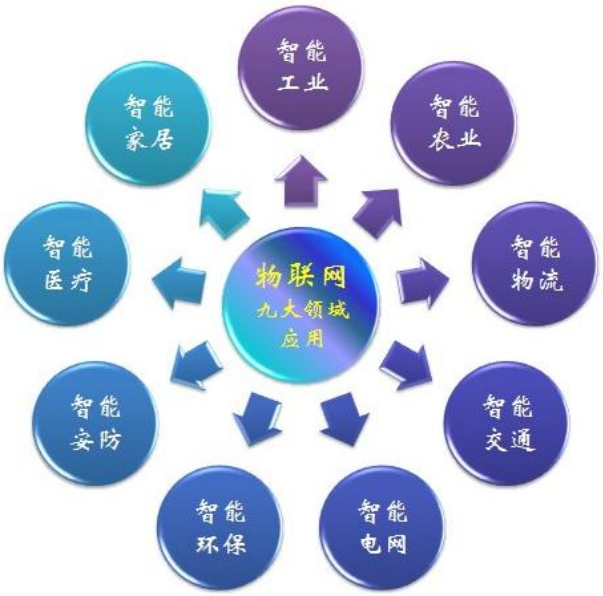
互联网：人与人交互

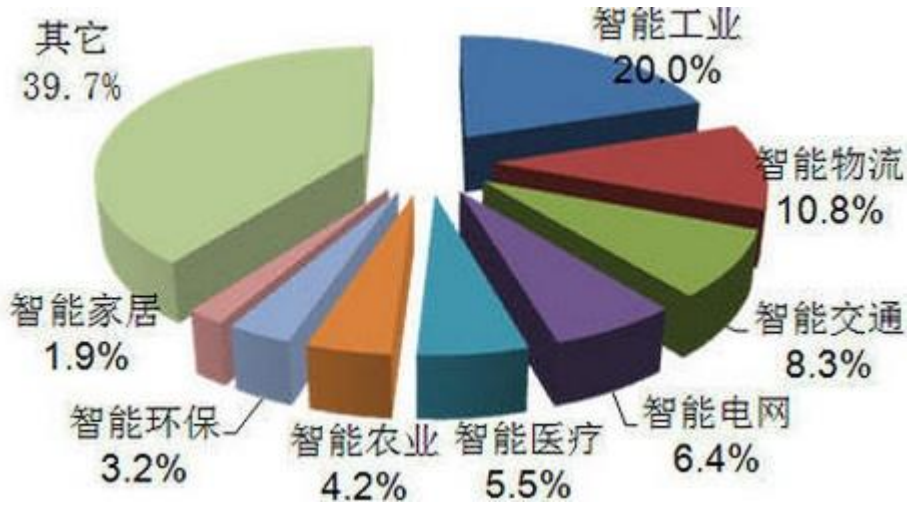
基于计算机网络进行信息交互

物联网：物物&人物交互

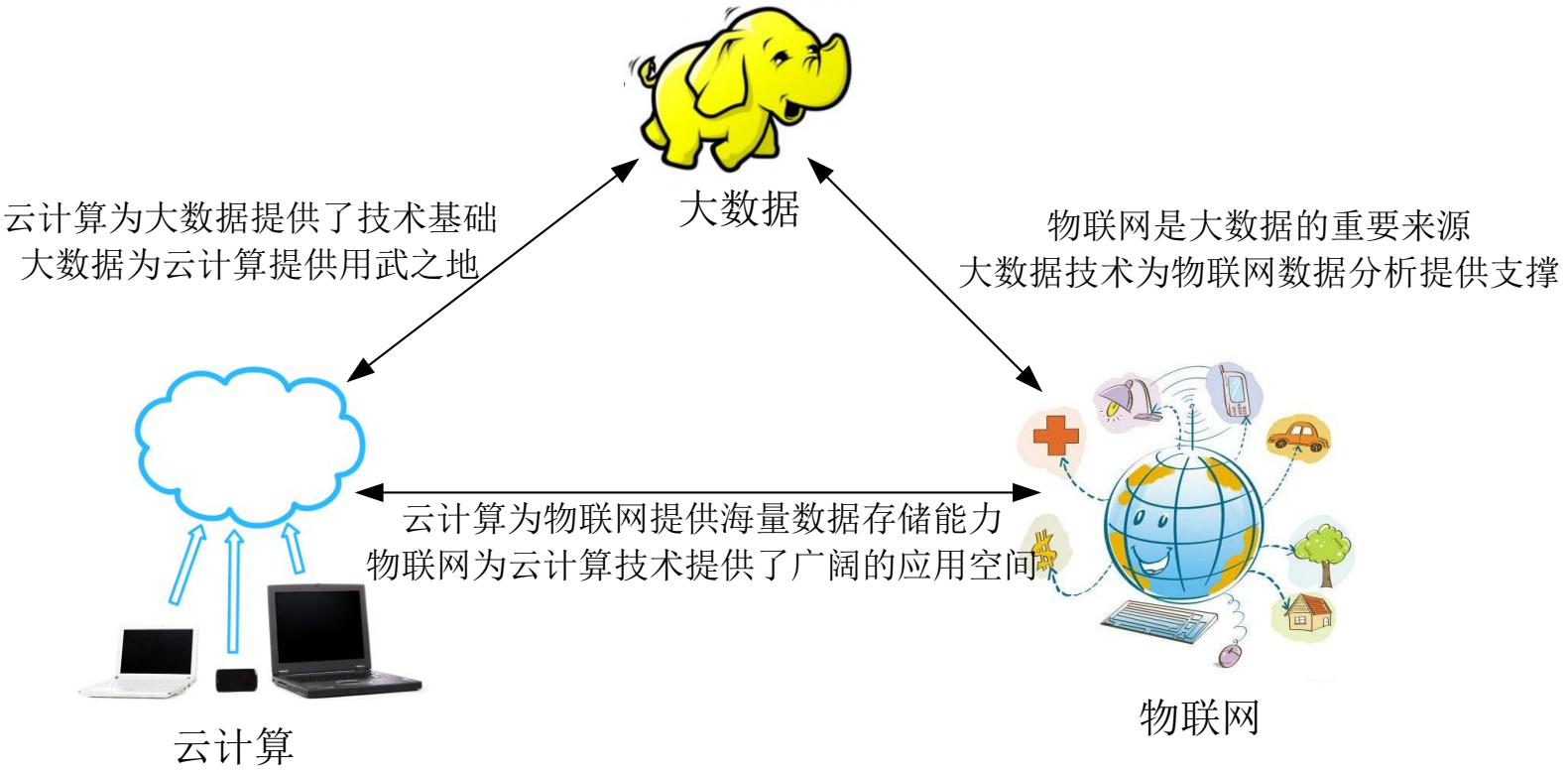
万物的信息感知、交互、处理的网络

物联网

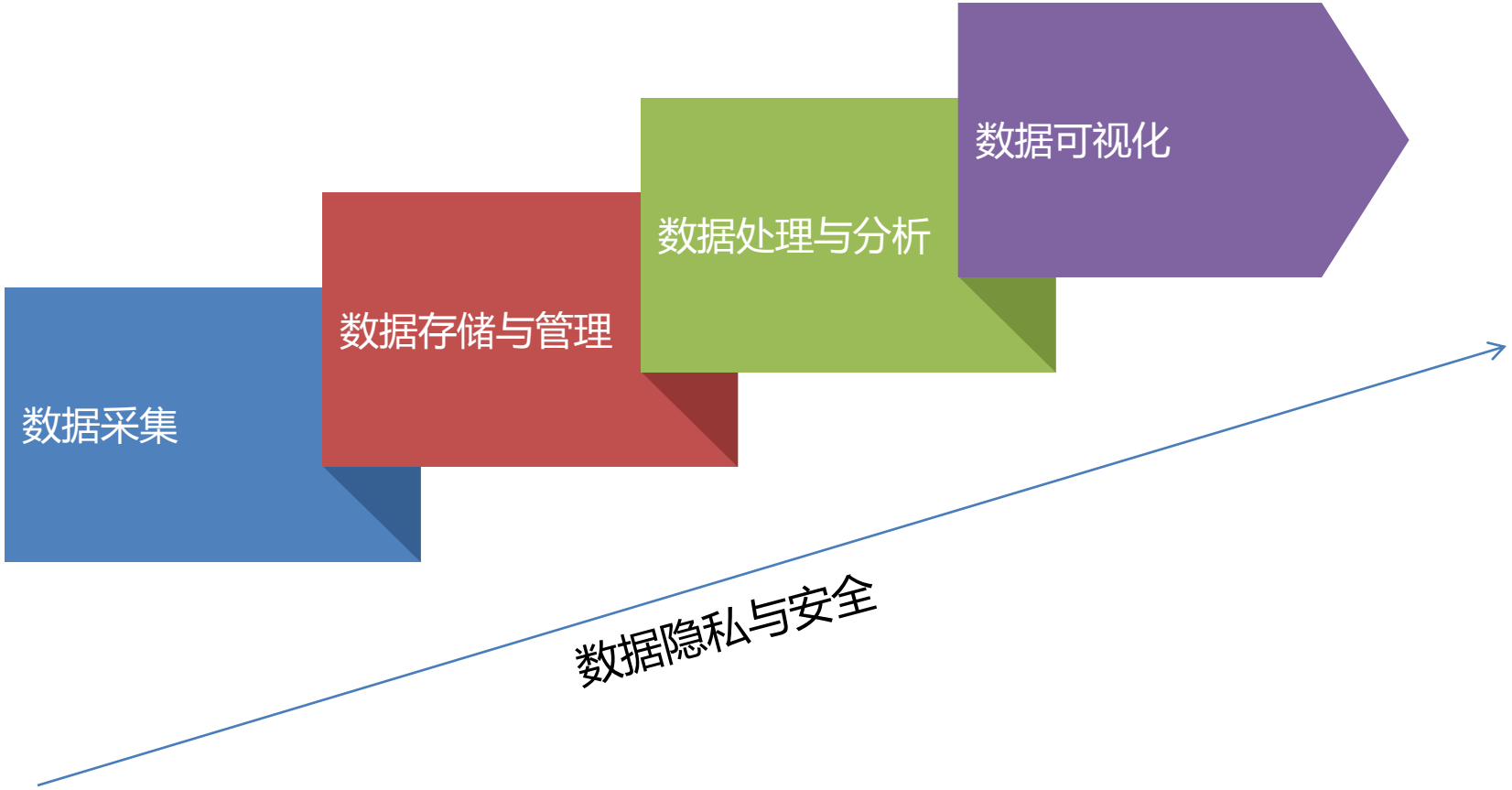

物联网
九大领域
应用
智能工业
智能农业
智能物流
智能交通
智能电网
智能环保
智能安防
智能医疗
智能家居


其它
39.7%
智能工业
20.0%
智能物流
10.8%
智能交通
8.3%
智能电网
6.4%
智能医疗
5.5%
智能农业
4.2%
智能环保
3.2%
智能家居
1.9%

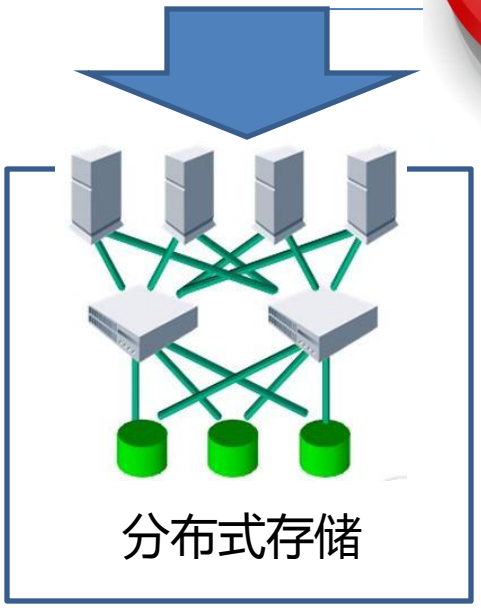
大数据、云计算、物联网密的紧密关系

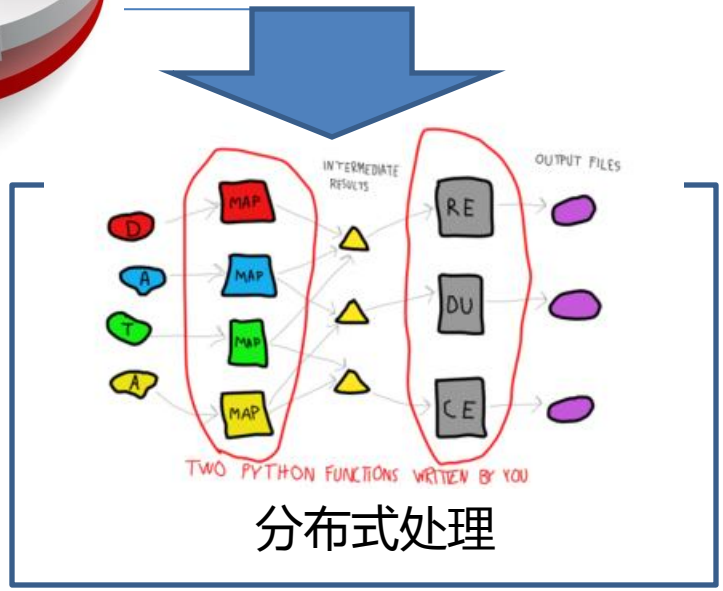


目录 Contents


数据采集
数据存储与管理
数据处理与分析
数据可视化
数据隐私与安全

两大核心技术

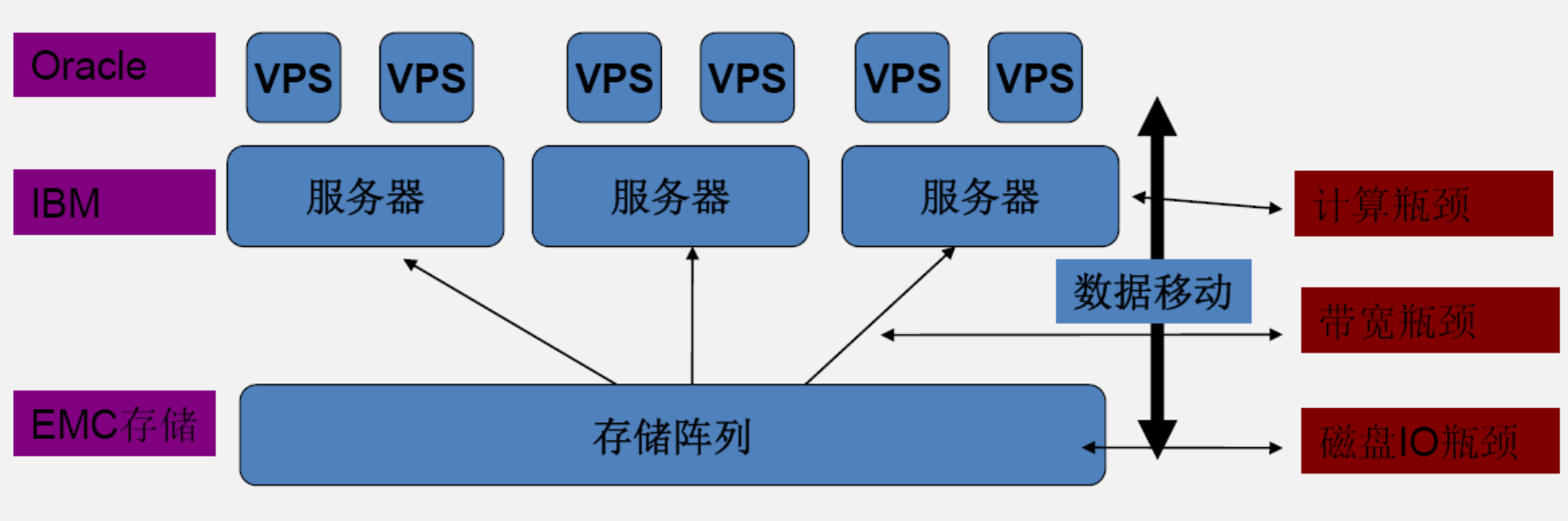




GFS\HDFS
BigTable\HBase
NoSQL（键值、列族、图形、文档数据库）
NewSQL（如：SQL Azure）

MapReduce
Spark
Flink
Beam

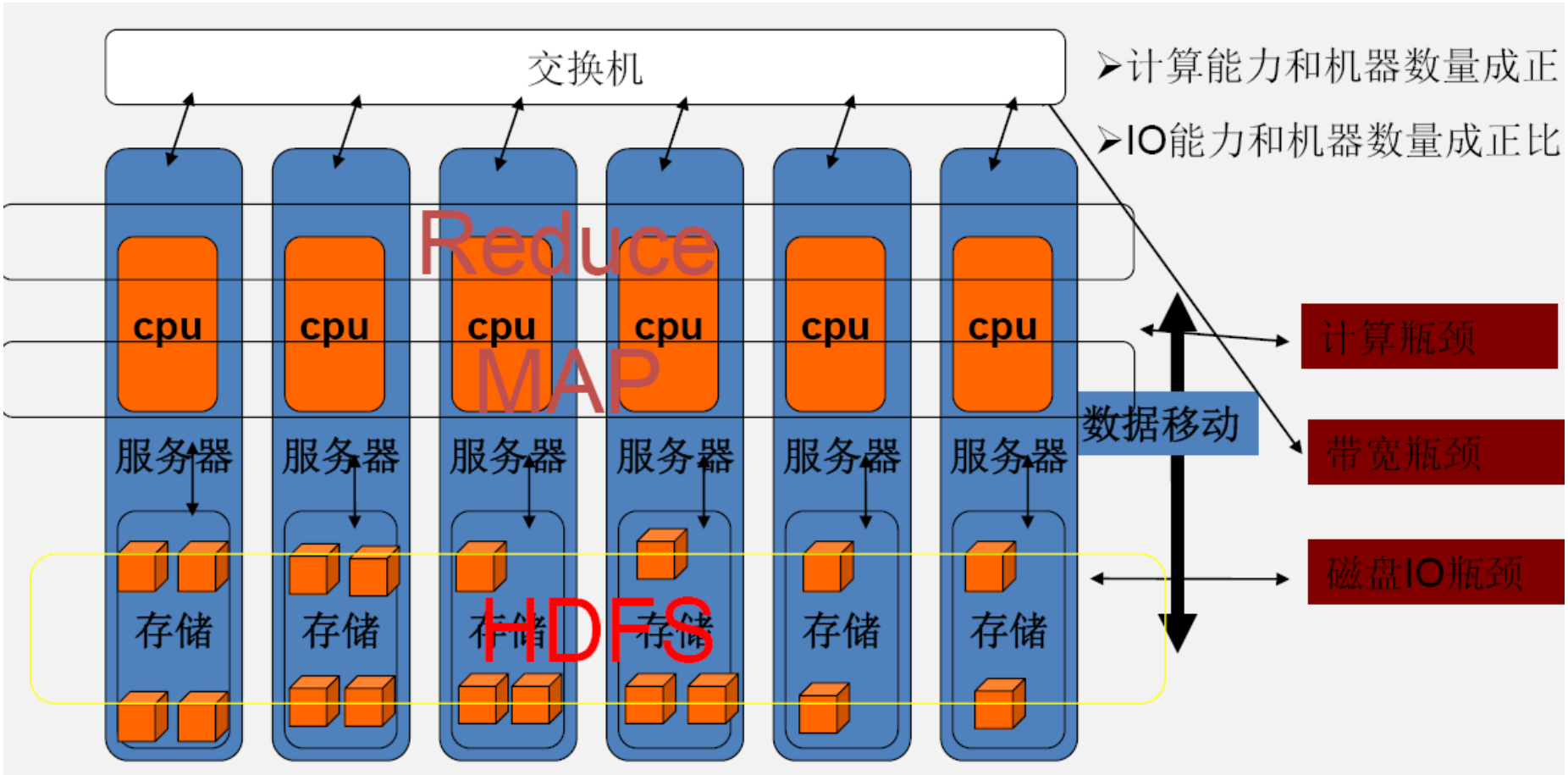
传统架构IOE的瓶颈



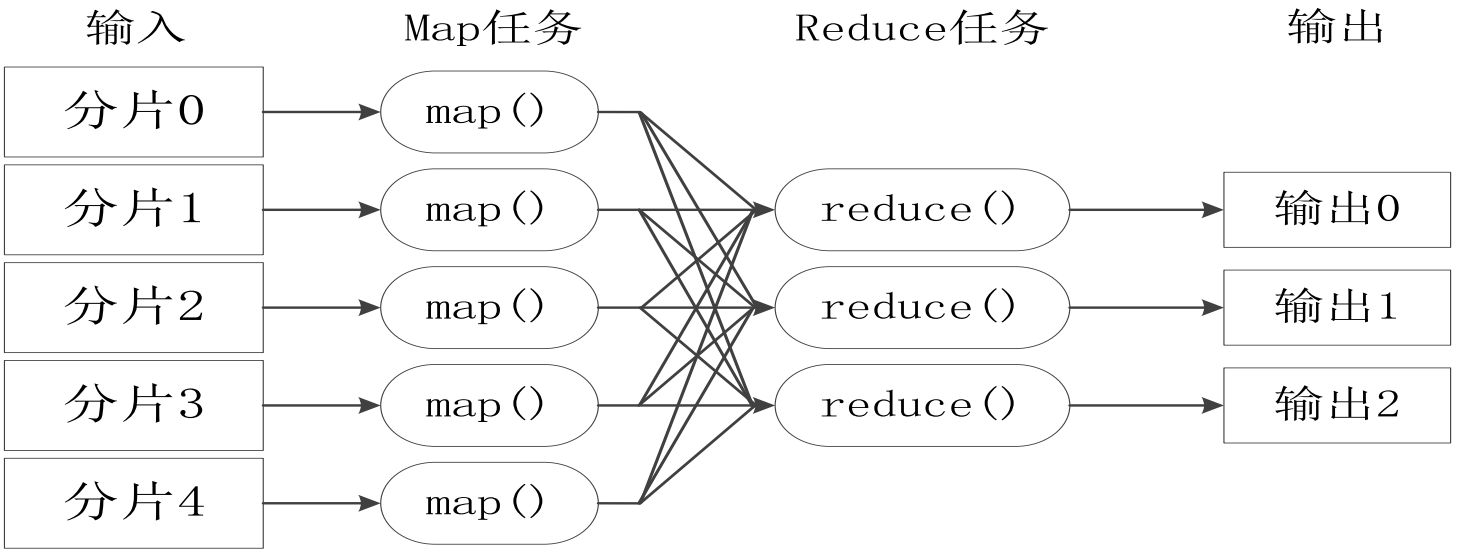
基于共享存储和高性能计算的架构

VPS服务器（虚拟专用服务器）（"Virtual Private Server"，或简称 "VPS"）是利用虚拟服务器软件(如微软的Virtual Server、VMware的ESX server、SWsoft 的Virtuozzo)在一台物理服务器上创建多个相互隔离的小服务器。

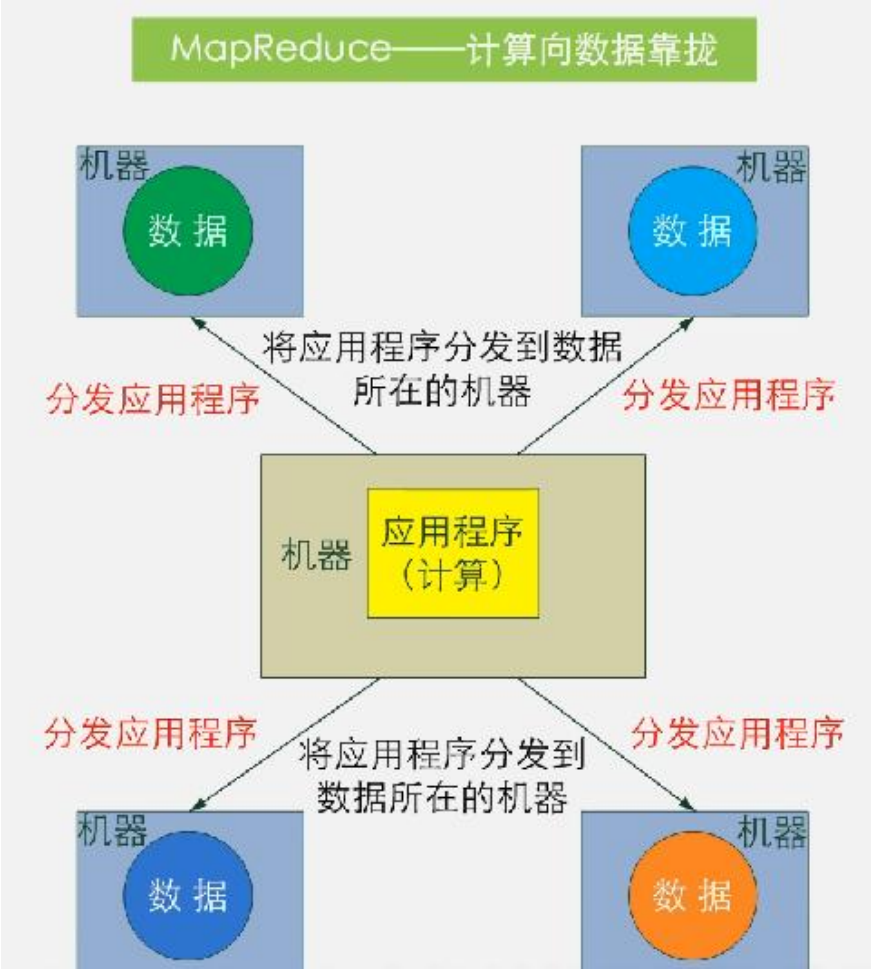
Hadoop架构


交换机
➢计算能力和机器数量成正
➢IO能力和机器数量成正比
Reduce
MAP
cpu
cpu
cpu
cpu
cpu
cpu
服务器
服务器
服务器
服务器
服务器
服务器
存储
存储
存储
存储
HDFS
数据移动
计算瓶颈
带宽瓶颈
磁盘IO瓶颈

MapReduce核心思想之一：分而治之

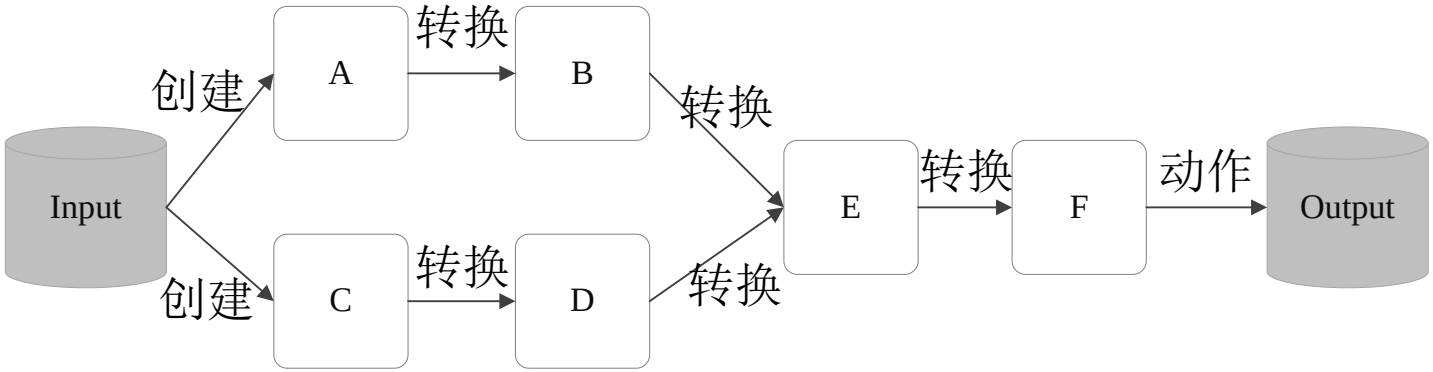


MapReduce核心思想之二：计算向数据靠拢

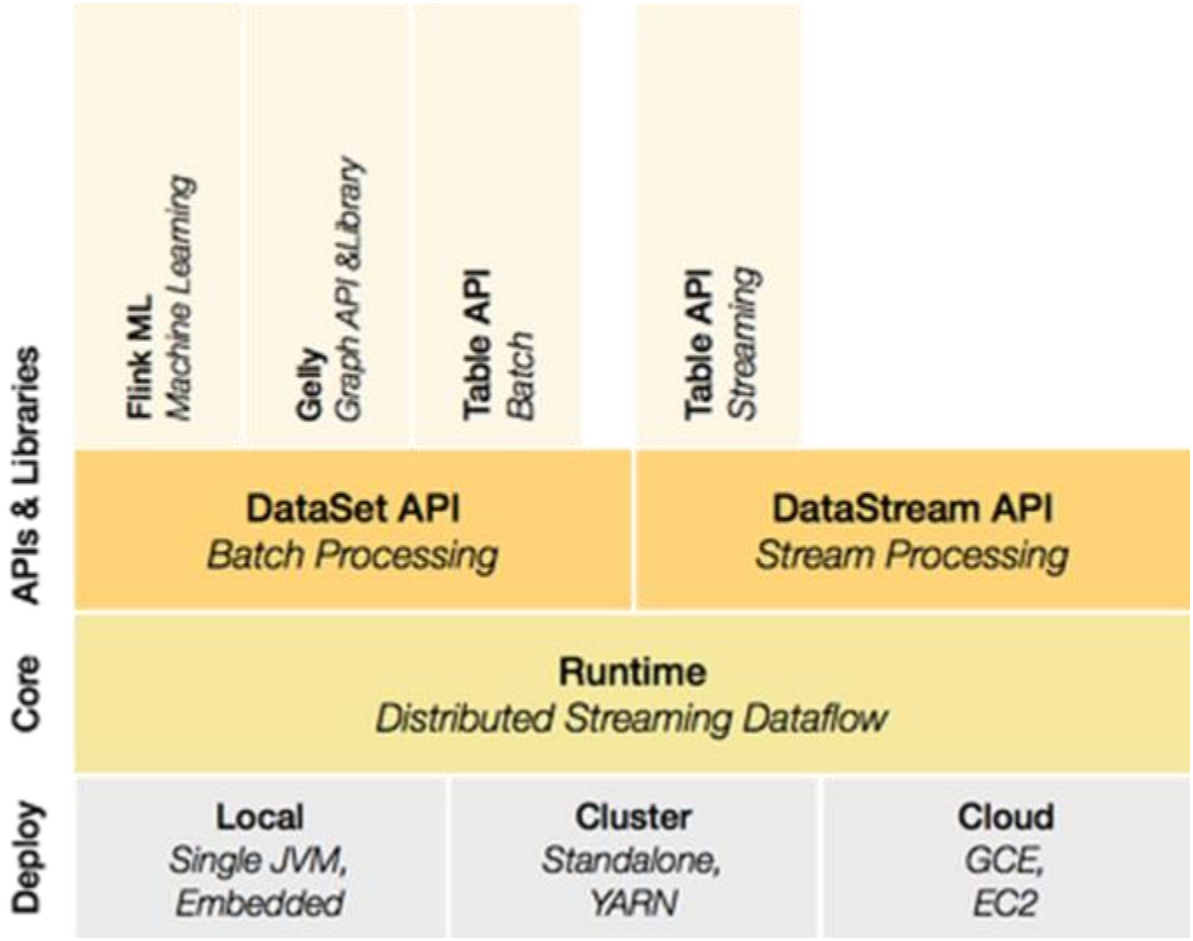


Spark
Lightning-Fast Cluster Computing

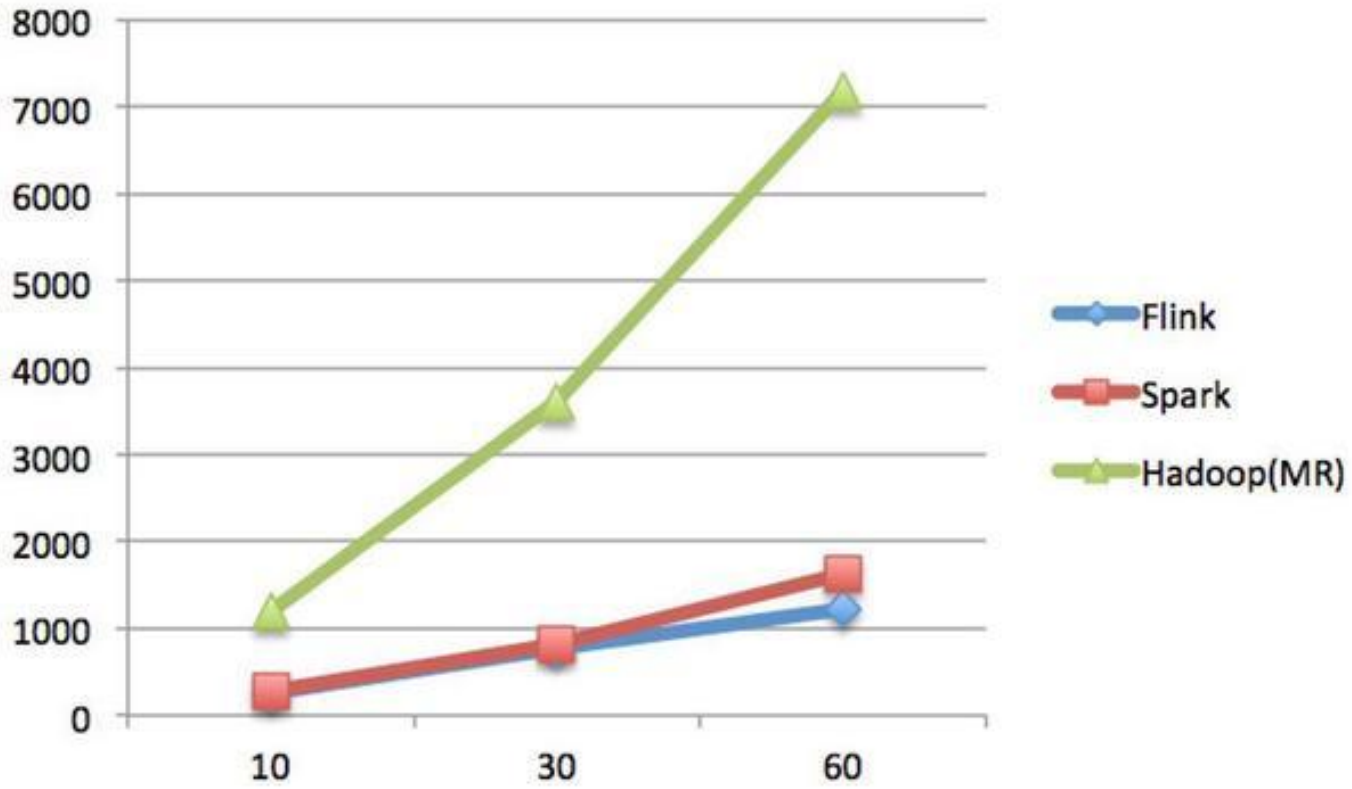
Spark Core, Spark SQL, Spark GraphX, Spark Streaming, Spark MLlib



问题：Hadoop是否会被淘汰？

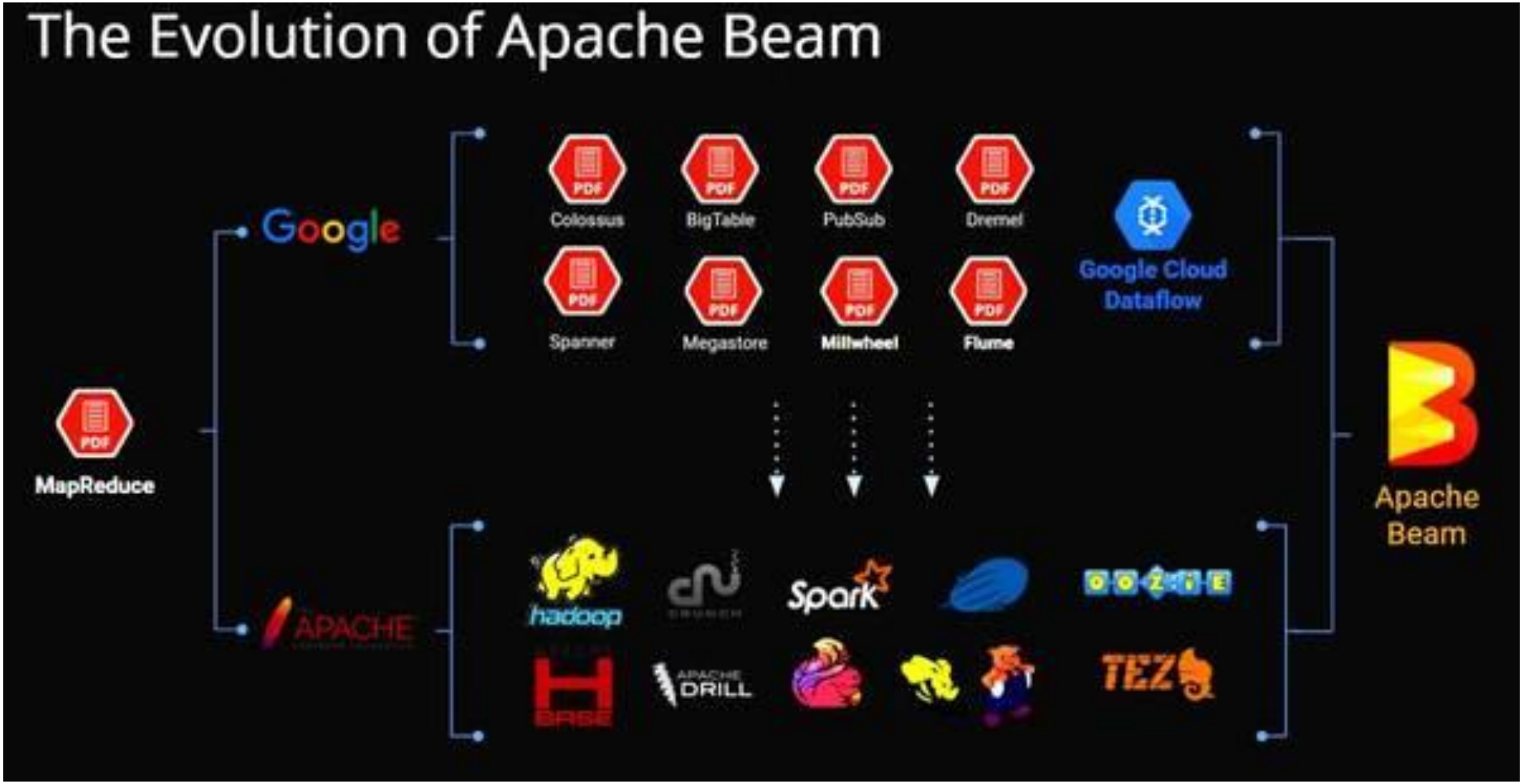

APIs & Libraries
Core
Deploy
Flink ML
Machine Learning
Gelly
Graph API &Library
Table API
Batch
Table API
Streaming
DataSet API
Batch Processing
DataStream API
Stream Processing
Runtime
Distributed Streaming Dataflow
Local
Single JVM,
Embedded
Cluster
Standalone,
YARN
Cloud
GCE,
EC2

Flink架构图


8000
7000
6000
5000
4000
3000
2000
1000
0
10
30
60
Flink
Spark
Hadoop(MR)

既生瑜，何生亮！

谷歌，Beam，一统天下？


The Evolution of Apache Beam
Google
Colossus
BigTable
PubSub
Dremel
Spanner
Megastore
Millwheel
Flume
Google Cloud Dataflow
MapReduce
APACHE
hadoop
Spark
APACHE DRILL
TEZ
Apache 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

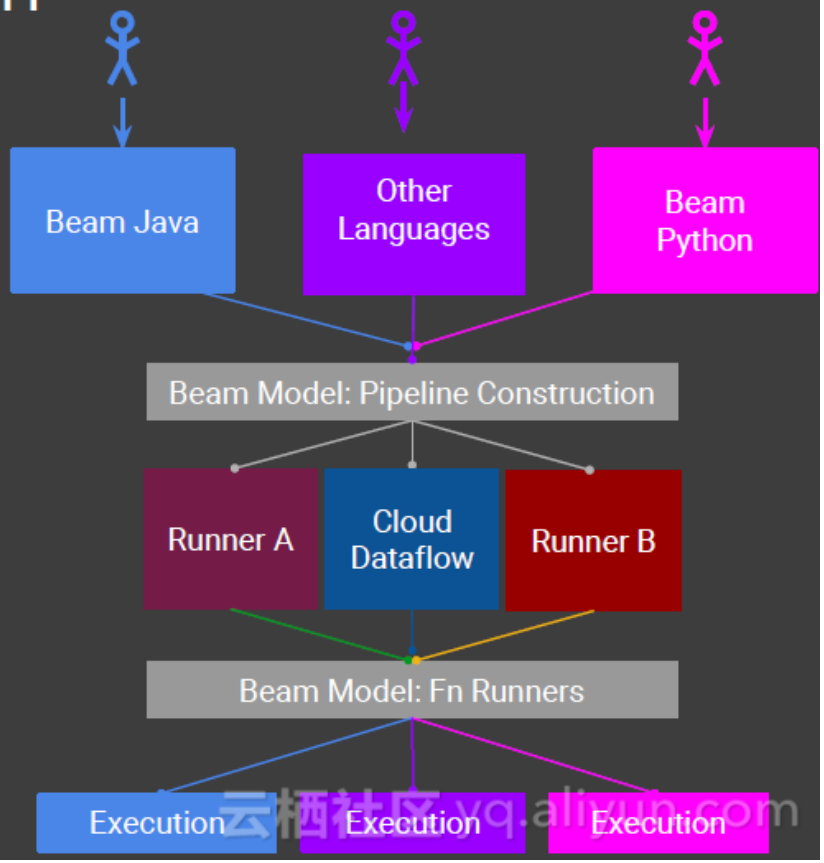


表 大数据计算模式及其代表产品

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

问题：如何在科研工作中使用大数据技术？

- 编程、算法、数学、统计学、机器学习、数据挖掘、分布式框架（Hadoop、Spark）
- 关注传统算法的分布式实现，搭建小集群加快科研数据分析速度
- 灵活运用数据采集、清洗、存储、处理、分析、可视化工具

目录 Contents

产业化应用案例之一：智能物流

智能物流，又称智慧物流，是利用集成智能化技术，使物流系统能模仿人的智能，具有思维、感知、学习、推理判断和自行解决物流中某些问题的能力，从而实现物流资源优化调度和有效配置、物流系统效率提升的现代化物流管理模式。

智能物流案例：阿里巴巴的中国智能物流骨干网（地网）

菜鸟网络到底是什么?

- 中国智能物流骨干网，又名“菜鸟”
- **菜鸟网络计划在5到8年内，打造一个全国性的超级物流网。**
- 这个网络能在24小时内将货物运抵国内任何地区，能支撑日均300亿元(年度约10万亿元)的巨量网络零售额。

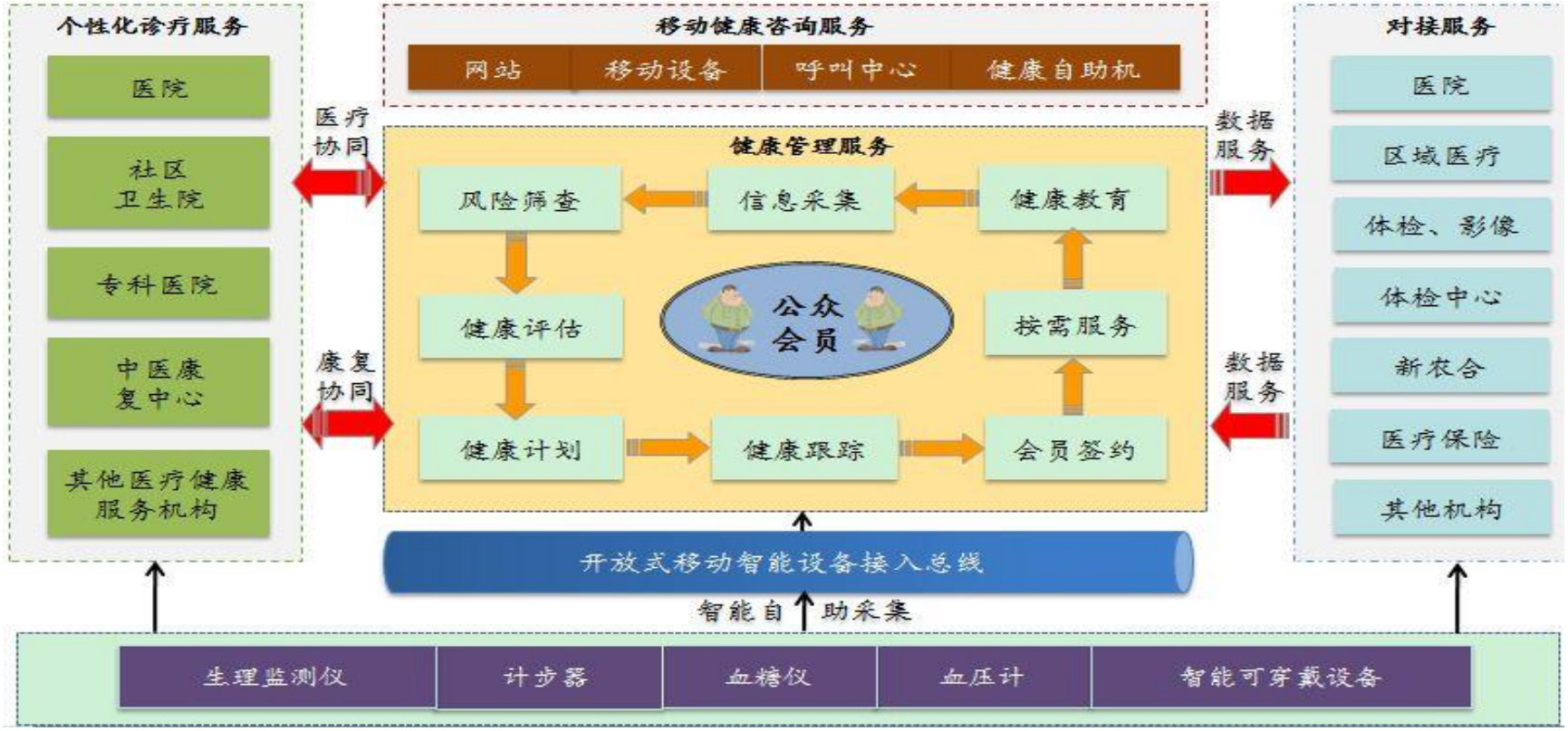
1000亿元投资物流基础设施　强强联手共建智能骨干网络物流信息系统向所有的制造商、网商、快递公司、第三方物流公司完全开放

阿里物流体系	
天网	地网
天猫牵头负责与各大物流快递公司对接的数据平台	即“菜鸟”，又称“中国智能物流骨干网（CSN）”

产业化应用案例之二：智慧医疗

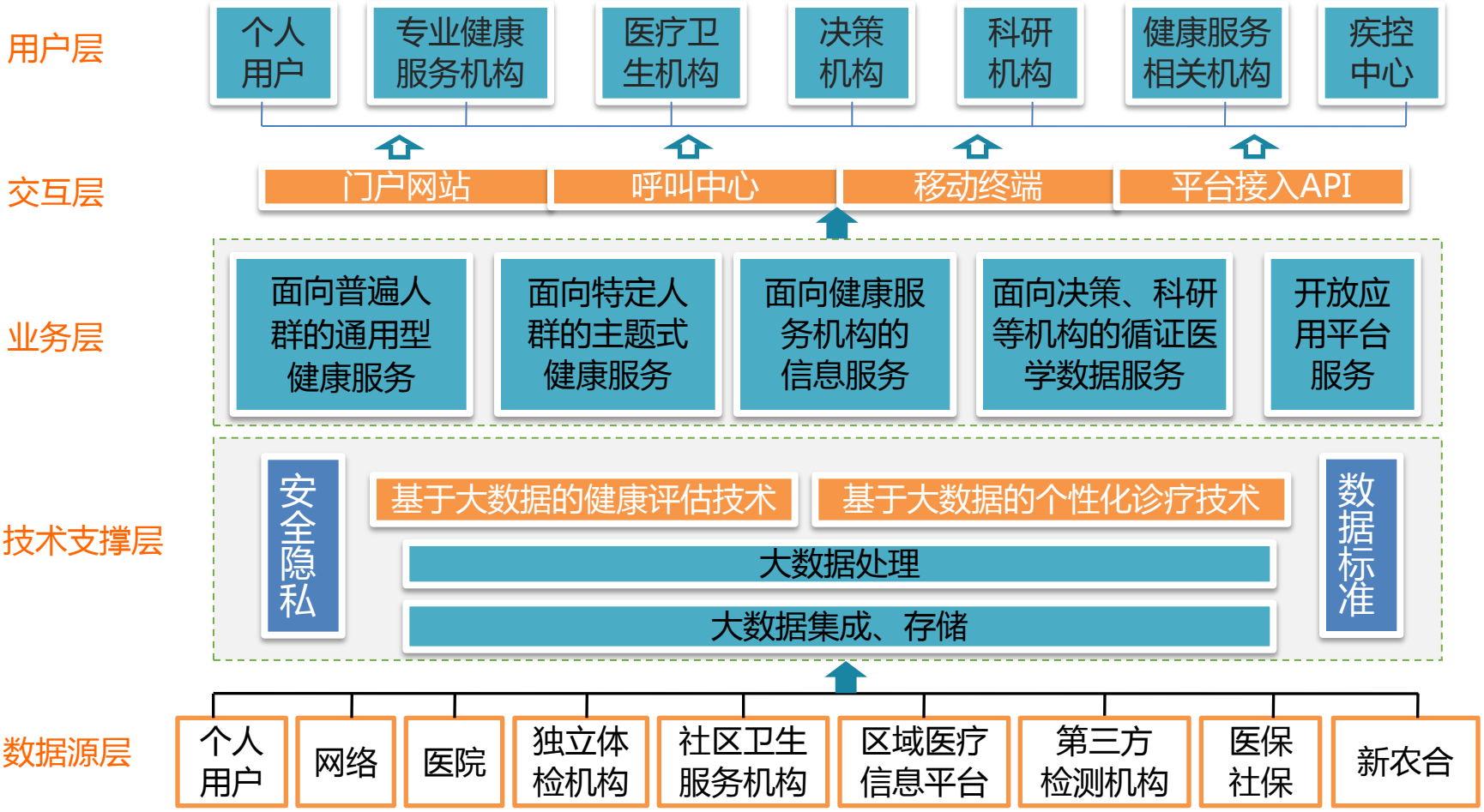
基于大数据的综合健康服务平台

建设目标：构建覆盖全生命周期、内涵丰富、结构合理的以人为本全面连续的综合健康服务体系，利用大数据技术和智能设备技术，提供线上线下相结合的公众健康服务，实现“未病先防、已病早治、既病防变、愈后防复”，满足社会公众多层次、多方位的健康服务需求，提升人民群众的身心健康水平。



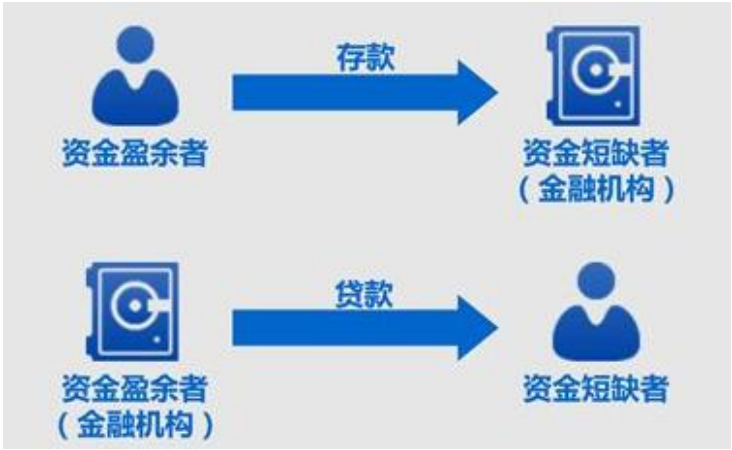
三 产业化应用案例之二：智慧医疗

基于大数据的综合健康服务平台



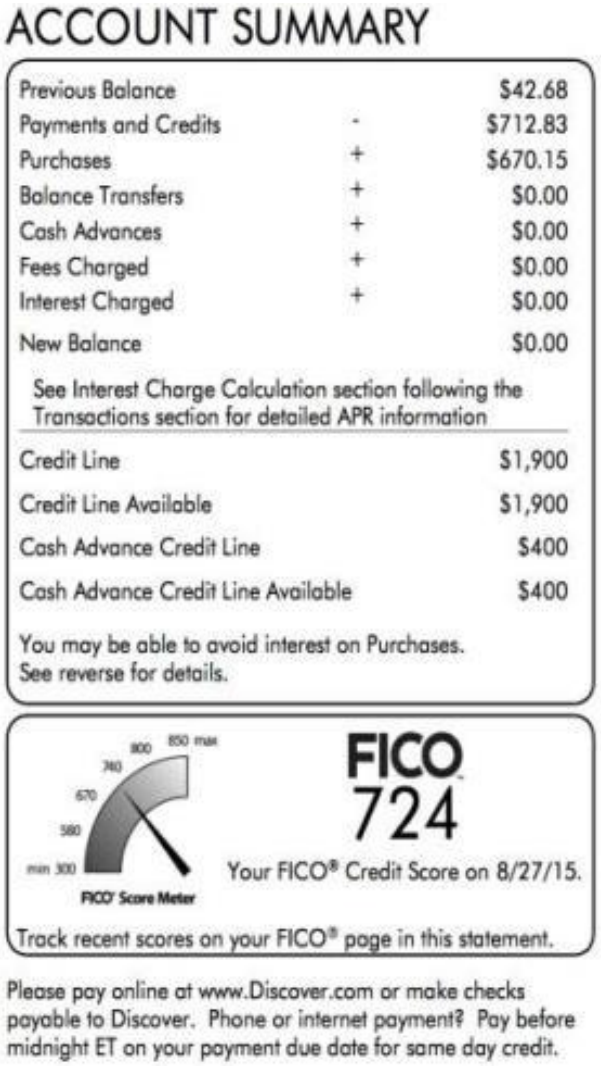
产业化应用案例之三：大数据征信

- 金融的本质
- 信用是一切商业活动与金融的基础
- 金融行业的信用是价值创造的根本
- 数据是信用评估的素材



FICO评分模型

- 信用分达到680分以上，信用卓著；
- 信用分低于620分，增加担保，或者寻找各种理由拒绝贷款；
- 信用分介于620～ 680分之间，进一步的调查核实，采用其它的信用分析工具，作个案处理；
- 信用分低于600分，借款人违约的比例是1/8，信用分介于700～ 800分，违约率为 1/123，信用分高于800分，违约率为 1/1292。
- 银行贷款：LendingClub（A-G 7级FICO 660）



ACCOUNT SUMMARY

Previous Balance		$42.68
Payments and Credits	-	$712.83
Purchases	+	$670.15
Balance Transfers	+	$0.00
Cash Advances	+	$0.00
Fees Charged	+	$0.00
Interest Charged	+	$0.00
New Balance		$0.00

See Interest Charge Calculation section following the Transactions section for detailed APR information

Credit Line	$1,900
Credit Line Available	$1,900
Cash Advance Credit Line	$400
Cash Advance Credit Line Available	$400

You may be able to avoid interest on Purchases. See reverse for details.

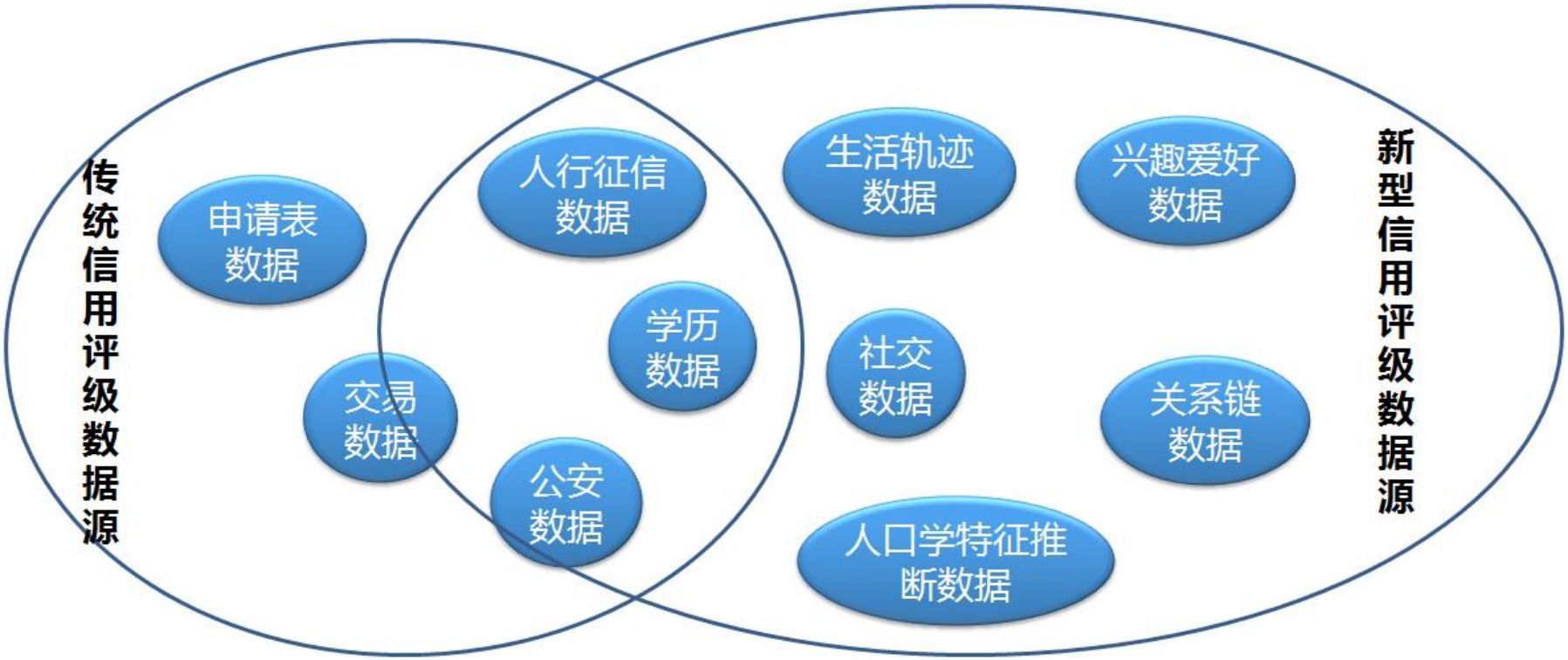
FICO 724

Your FICO® Credit Score on 8/27/15.

Track recent scores on your FICO® page in this statement.

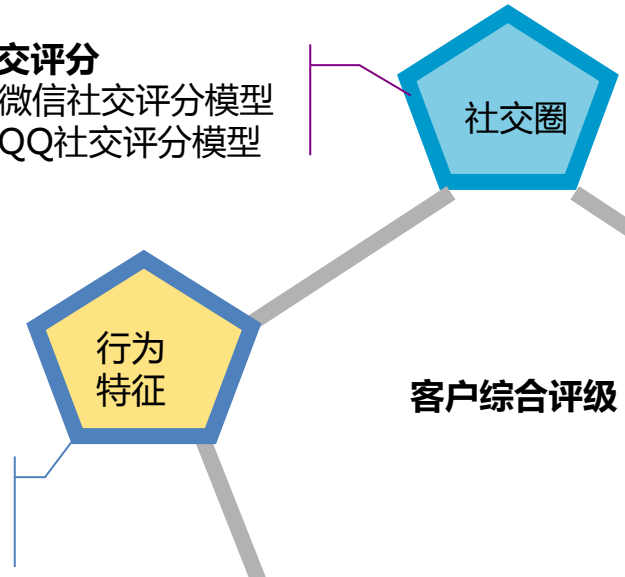
Please pay online at www.Discover.com or make checks payable to Discover. Phone or internet payment? Pay before midnight ET on your payment due date for same day credit.

基于大数据的信用评级



说明：所有数据都是经过客户授权使用；数据不涉及客户隐私，只是基于客户行为进行分析

从5大维度对用户综合评级



社交评分
- 微信社交评分模型
- QQ社交评分模型

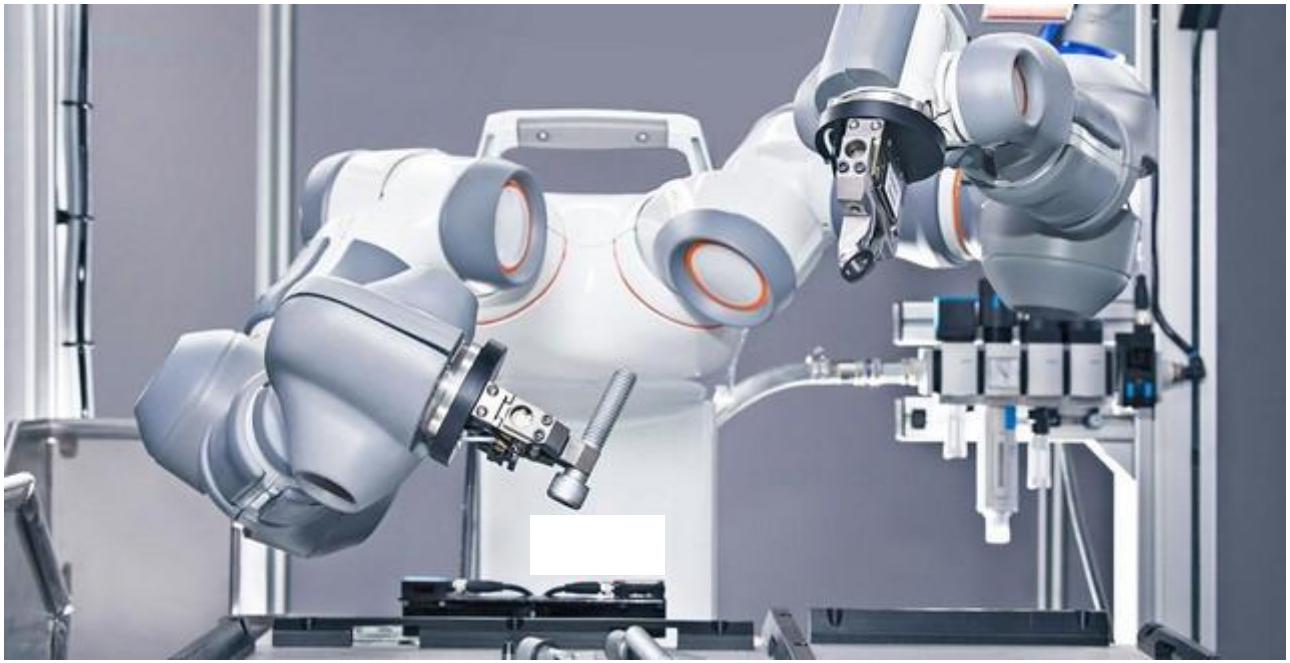
行为评分
- 生活轨迹
- 兴趣爱好
- 小善小恶
- 。。。

交易评分
基于客户在腾讯平台的交易信息进行评分

人行评分
- 违约模型
- 资金饥渴模型

画像评分
- 学历评分模型
- 基本特征
- 社会特征
- 。。。

目录 Contents



THANKS

敬 请 批 评 指 正

附录：报告人介绍

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

扫一扫访问林子雨个人主页

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。

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

附录：《大数据技术原理与应用》教材

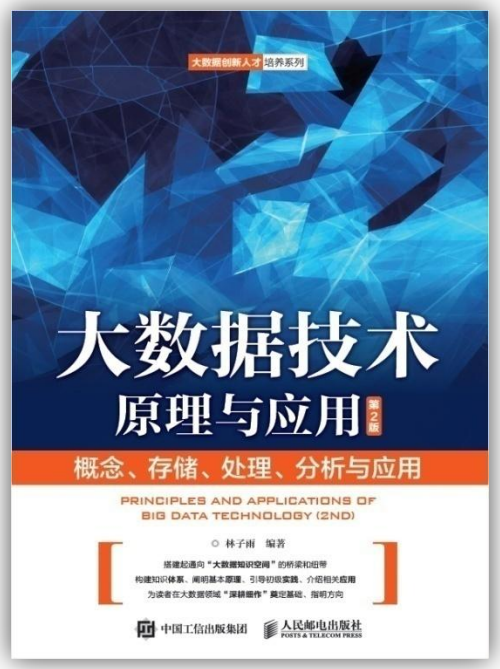
《大数据技术原理与应用——概念、存储、处理、分析与应用（第2版）》，由厦门大学计算机科学系林子雨老师编著，是国内高校第一本系统介绍大数据知识的专业教材。人民邮电出版社 ISBN:978-7-115-44330-4 定价：49.80元

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



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

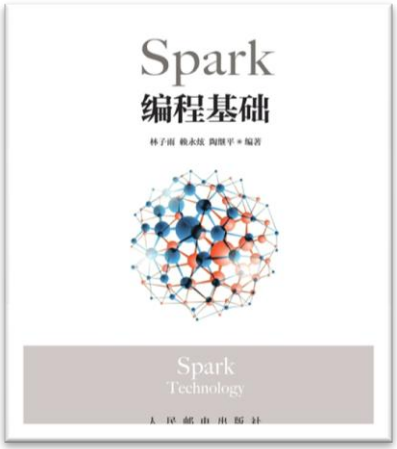
1+1黄金组合

厦门大学林子雨编著

配套实验指导书

清华大学出版社 ISBN:978-7-302-47209-4 定价：59元

- 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



《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：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

服务政府
服务企业

科学分析
科学决策

厦门大学数据库实验室

厦门大学云计算与大数据研究中心 | 厦门大学数据库实验室
地址：福建省厦门大学厦门大学海韵园科研2号楼
电话：(0595)2580033
传真：(0595)2580033
邮编：361005　E-mail: ziyulin@xmu.edu.cn
网址：http://dblab.xmu.edu.cn