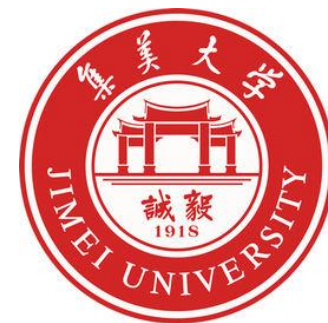




大数据公开课全国高校巡讲计划 集美大学专场

2017年12月6日 集美大学



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>



扫一扫访问专场主页

目录

Contents

一

大数据概念

二

大数据关键技术

三

大数据产业化应用案例

四

大数据时代的个人感悟



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

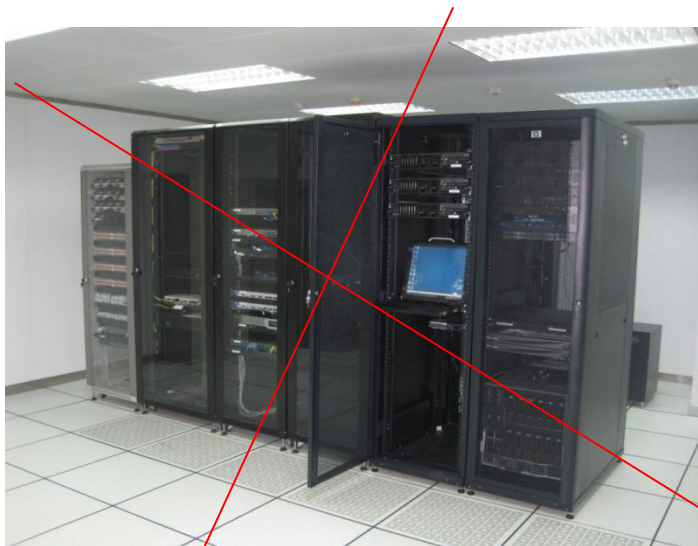
云计算特点

■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

云计算八大优势



企业不需要自建IT基础设施，可以租用云端资源



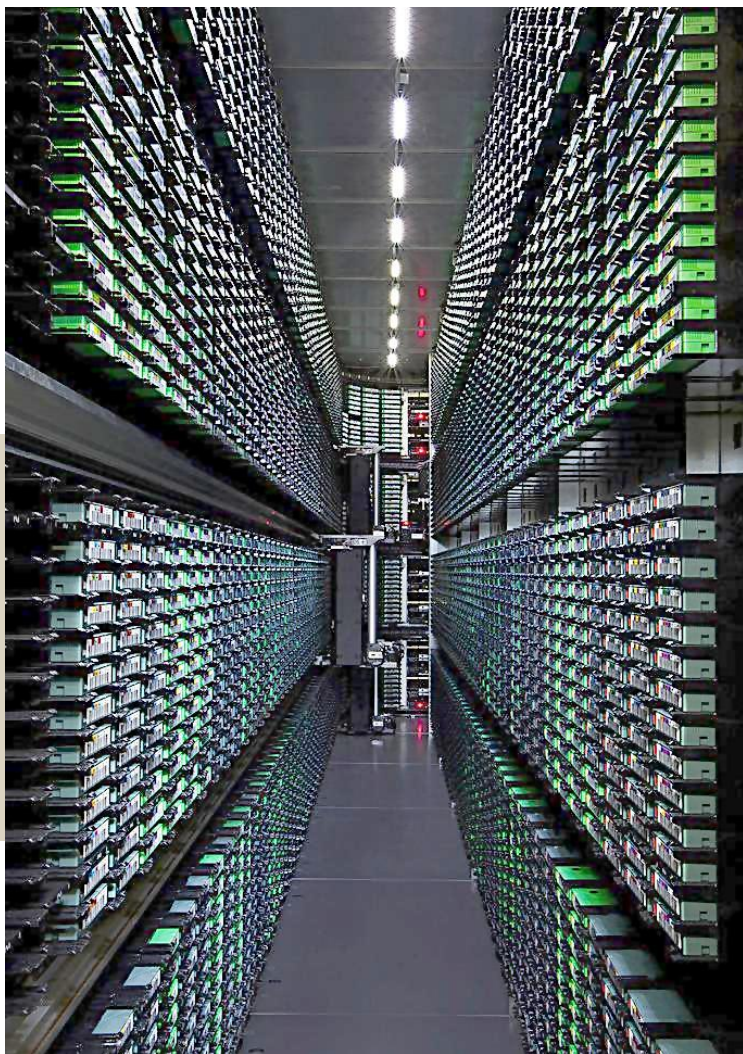
无需自建

企业用户



租用云端资源



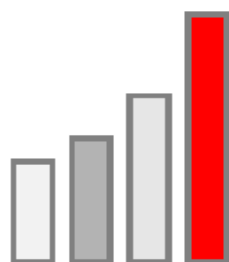


- 数据中心是云计算的温床
- 云计算推动数据中心向虚拟化和云架构的转型，不断提高IT基础架构的灵活性，以降低IT、能源和空间成本，从而让客户能够快速提高业务敏捷性
- “西数东输”
- 安溪-中国国际信息技术（福建）产业园



“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”

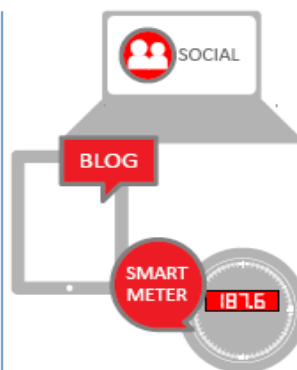
-- 《数据中心知识》杂志主编 Rich Miller



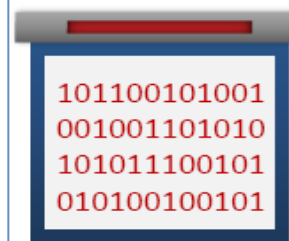
VOLUME
大量化



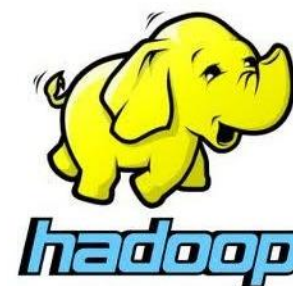
VELOCITY
快速化



VARIETY
多样化



VALUE



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

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。

打造物联网共建智慧地球



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

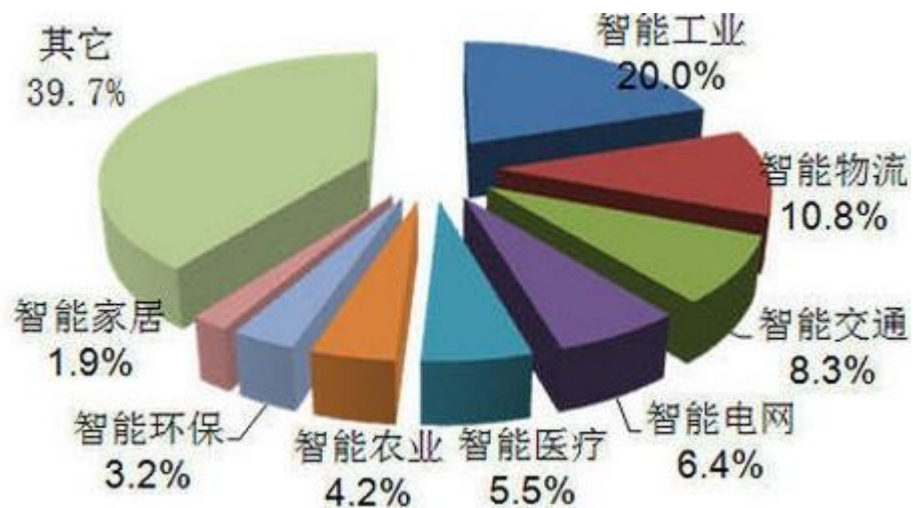
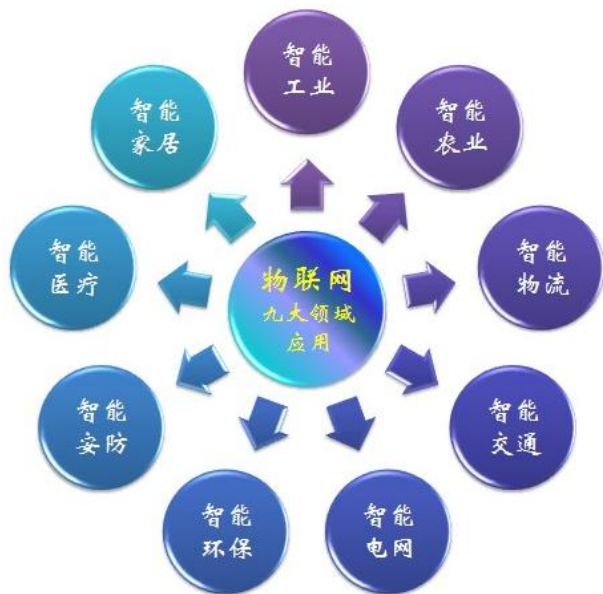


互联网：人与人交互

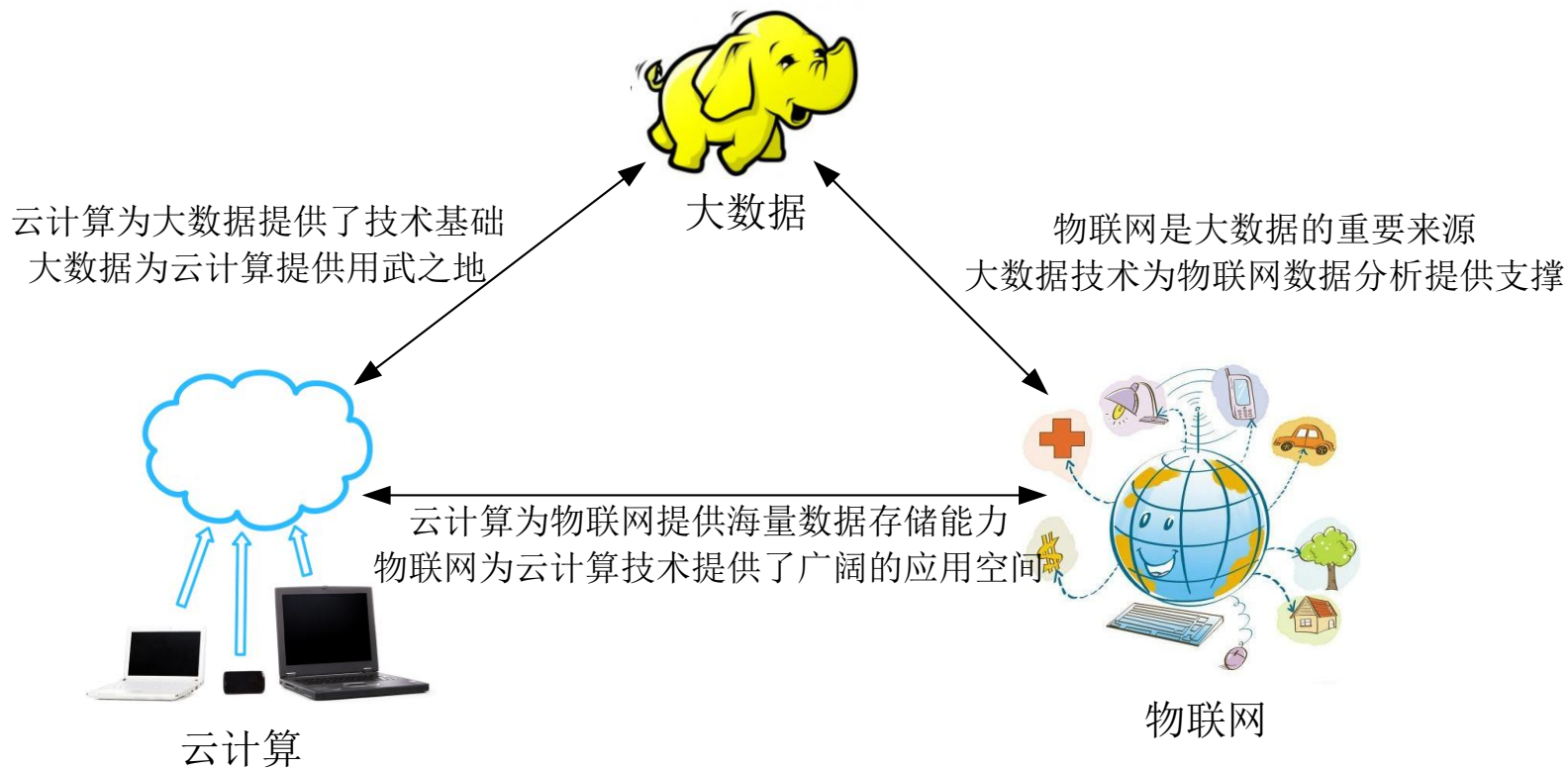


物联网：物物&人物交互





大数据、云计算、物联网密的紧密关系



目录

Contents

一

大数据概念

二

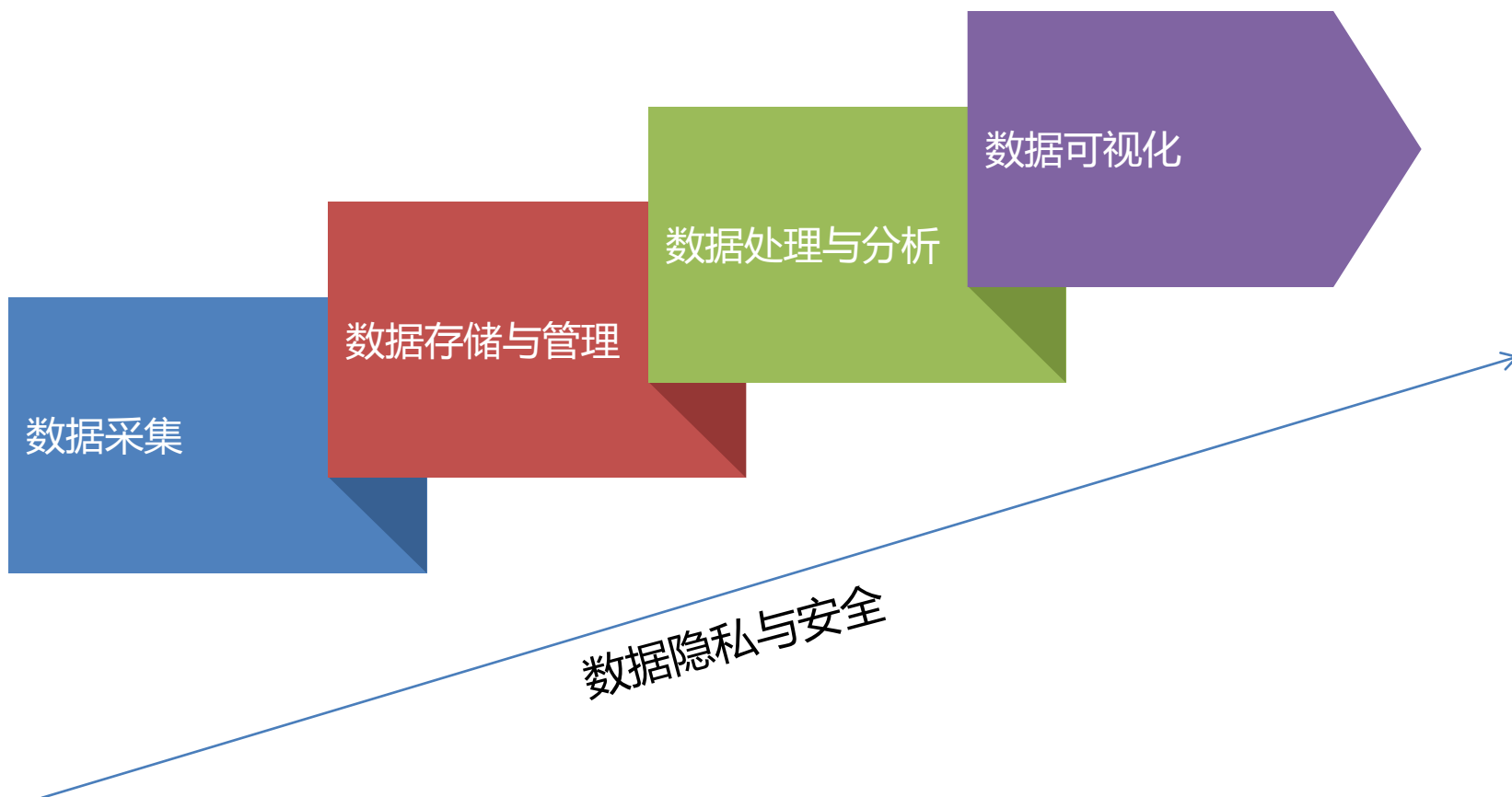
大数据关键技术

三

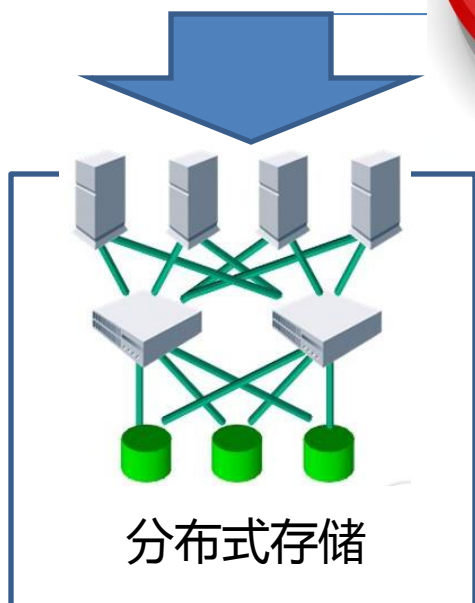
大数据产业化应用案例

四

大数据时代的个人感悟



两大核心技术

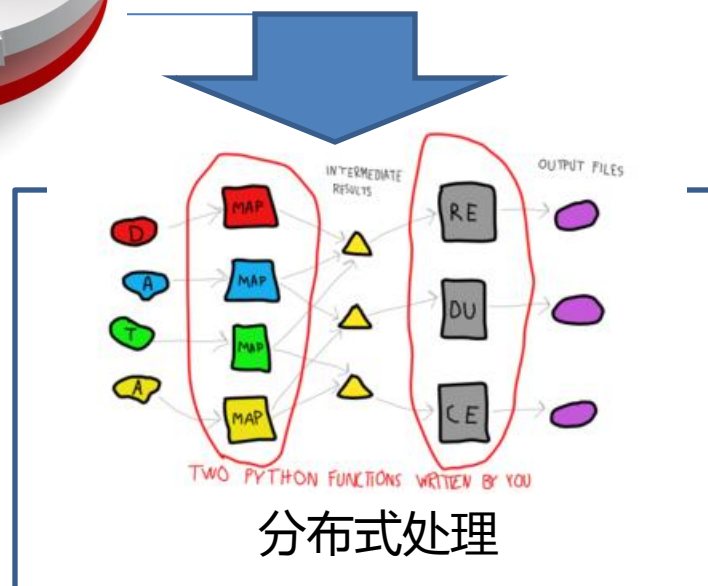


GFS\HDFS

BigTable\HBase

NoSQL (键值、列族、图形、文档数据库)

NewSQL (如: SQL Azure)



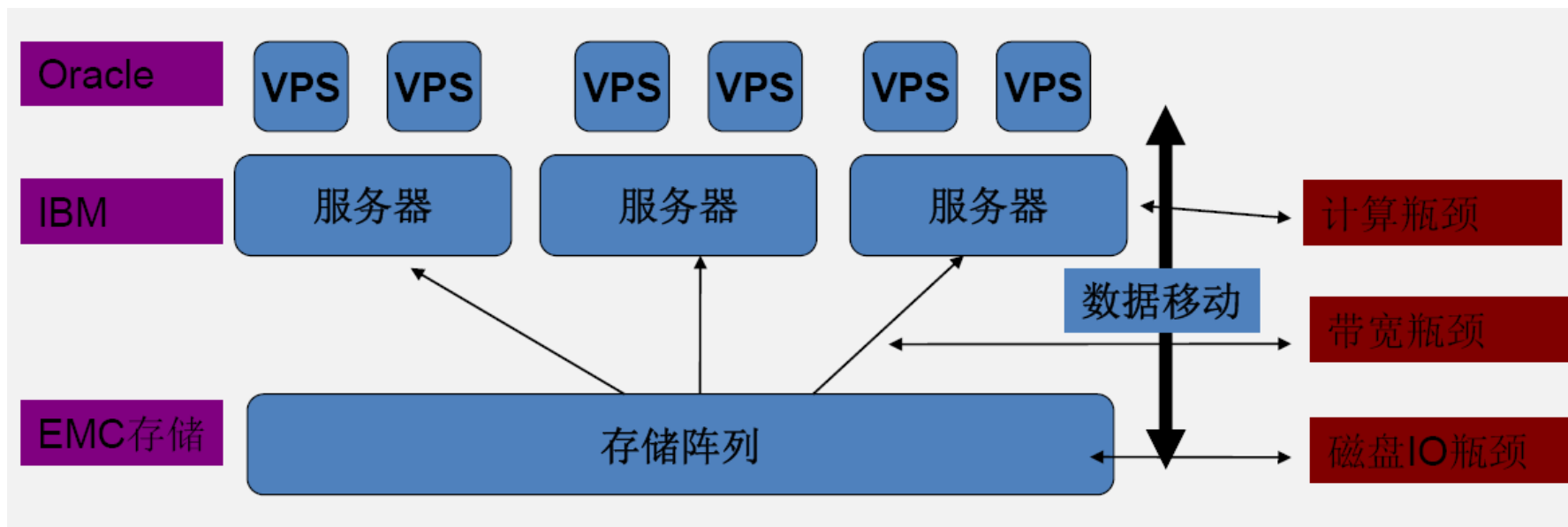
MapReduce

Spark

Flink

Beam

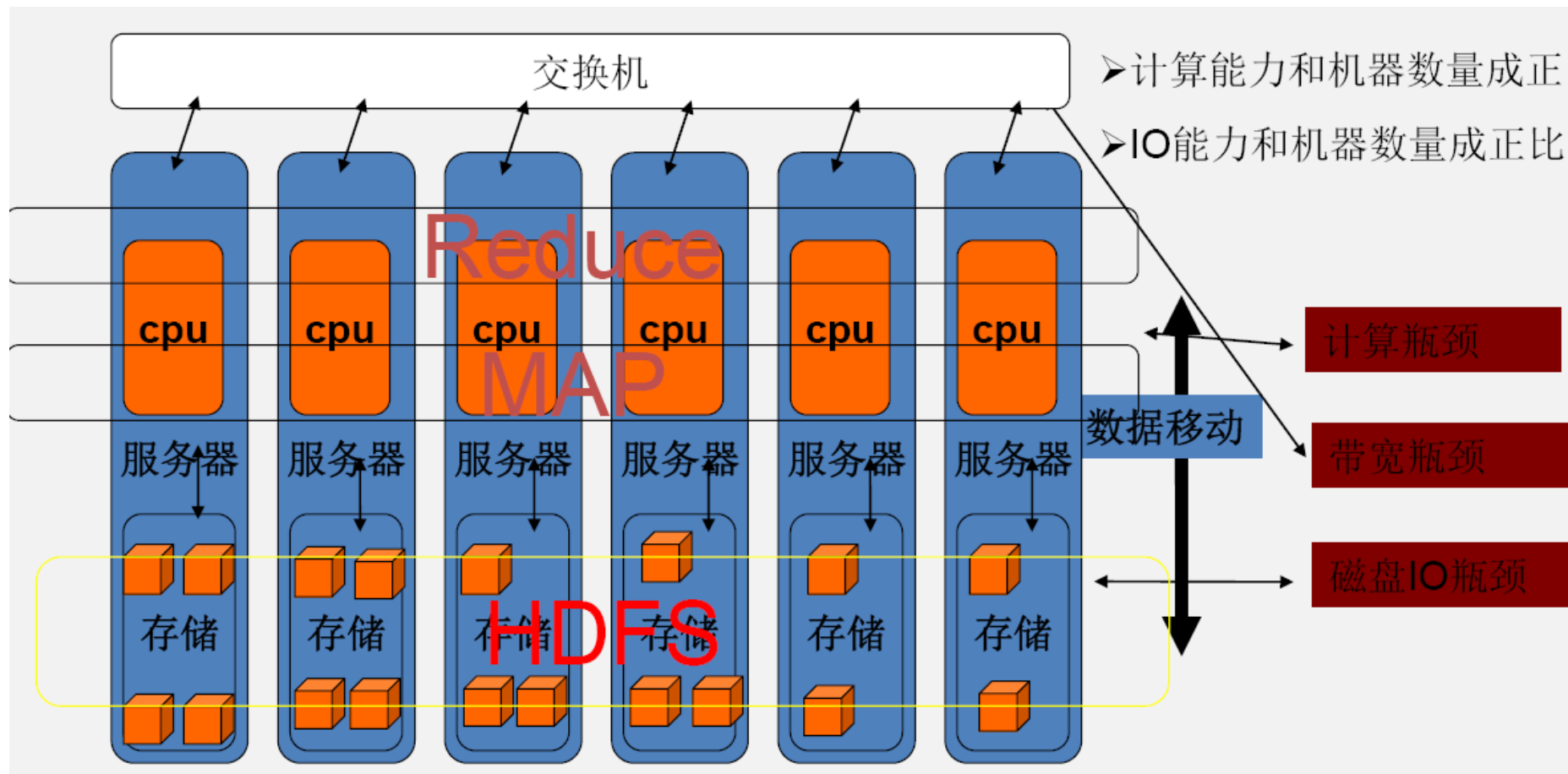
传统架构IOE的瓶颈



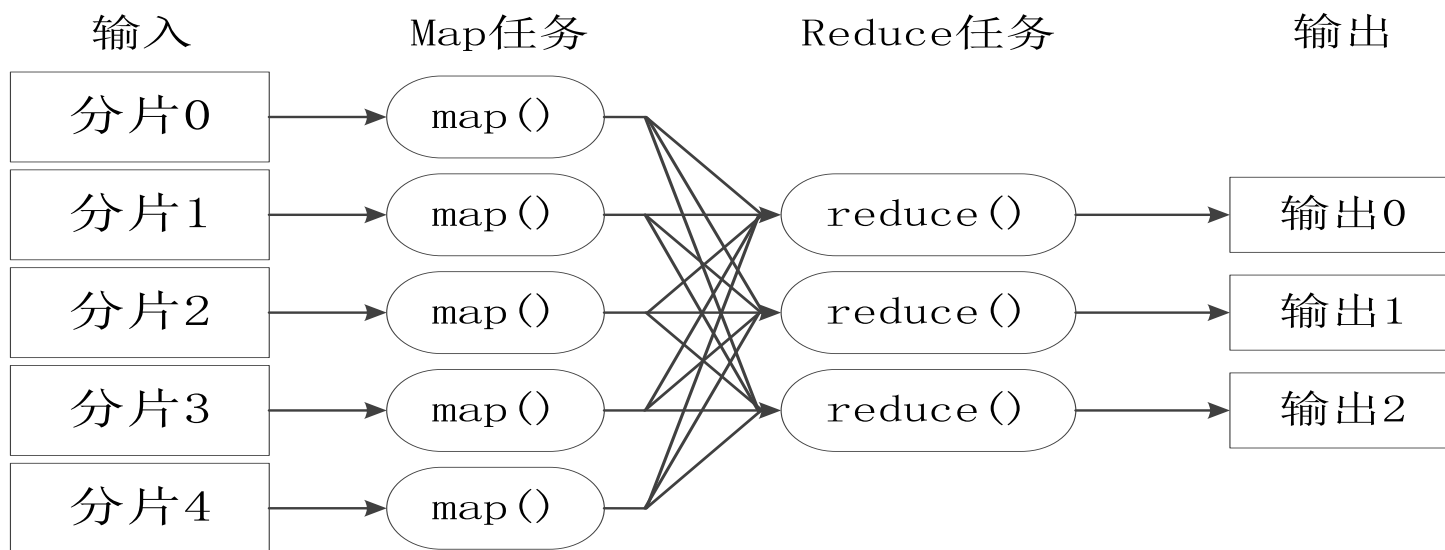
基于共享存储和高性能计算的架构

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。

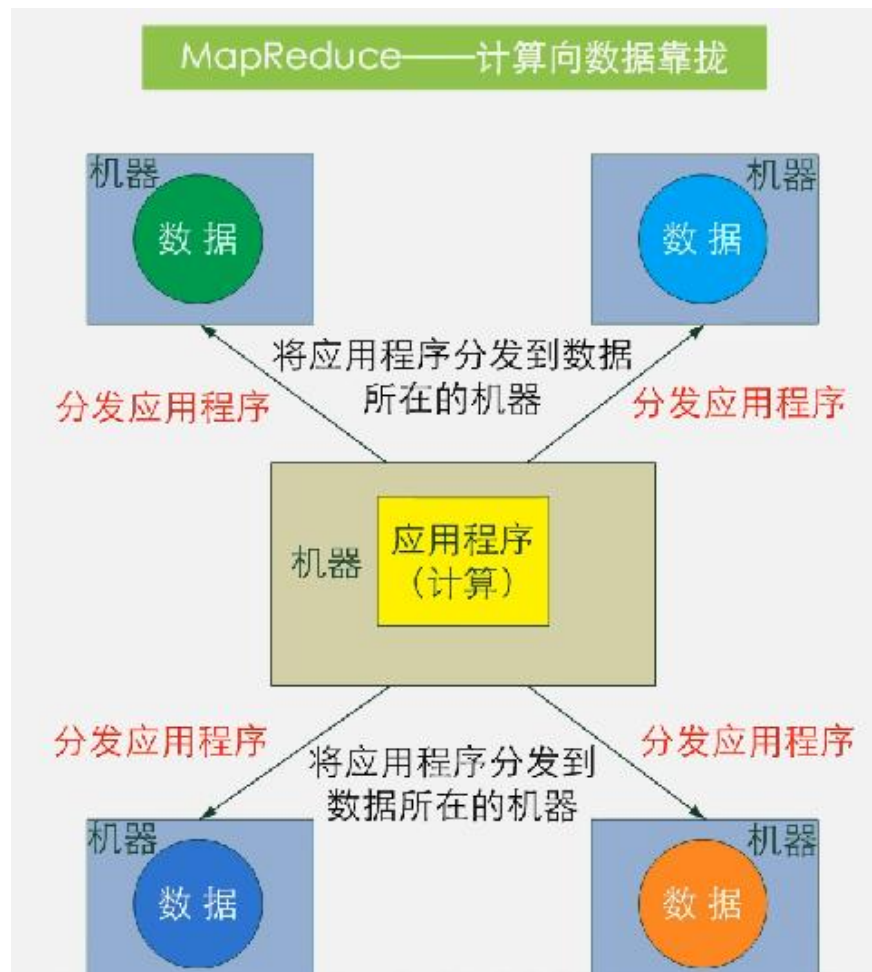
Hadoop架构



MapReduce核心思想之一：分而治之

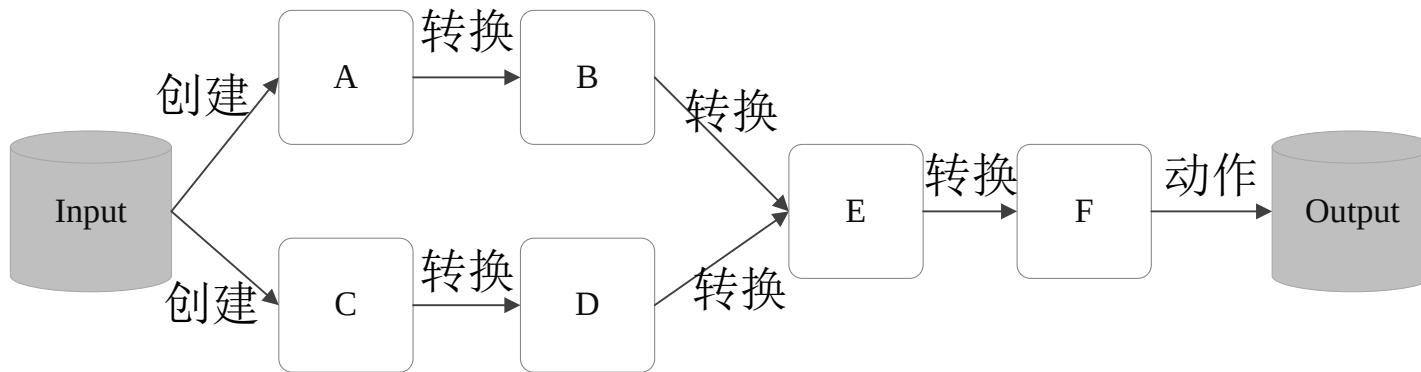


MapReduce核心思想之二：计算向数据靠拢



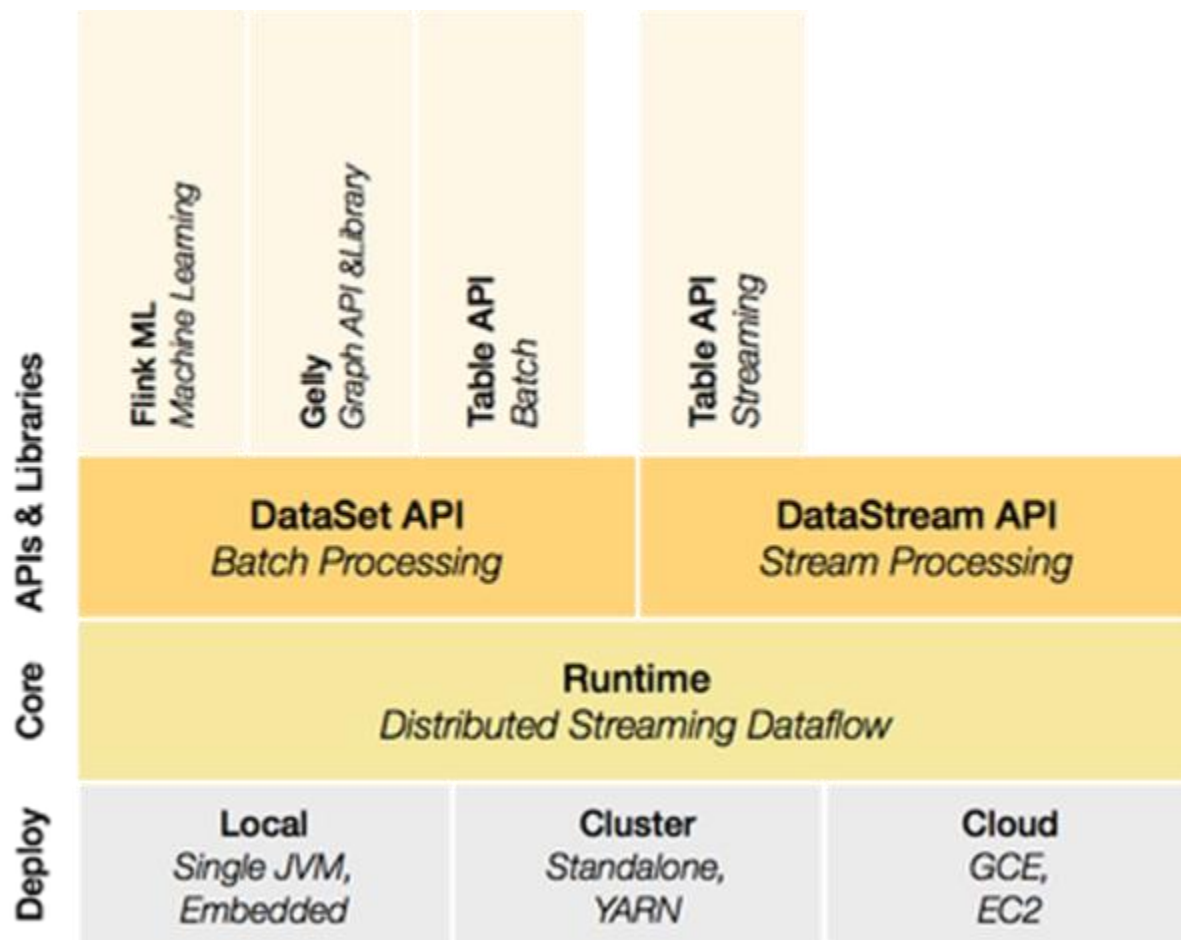


Spark Core, Spark SQL, Spark GraphX, Spark Streaming, Spark MLlib

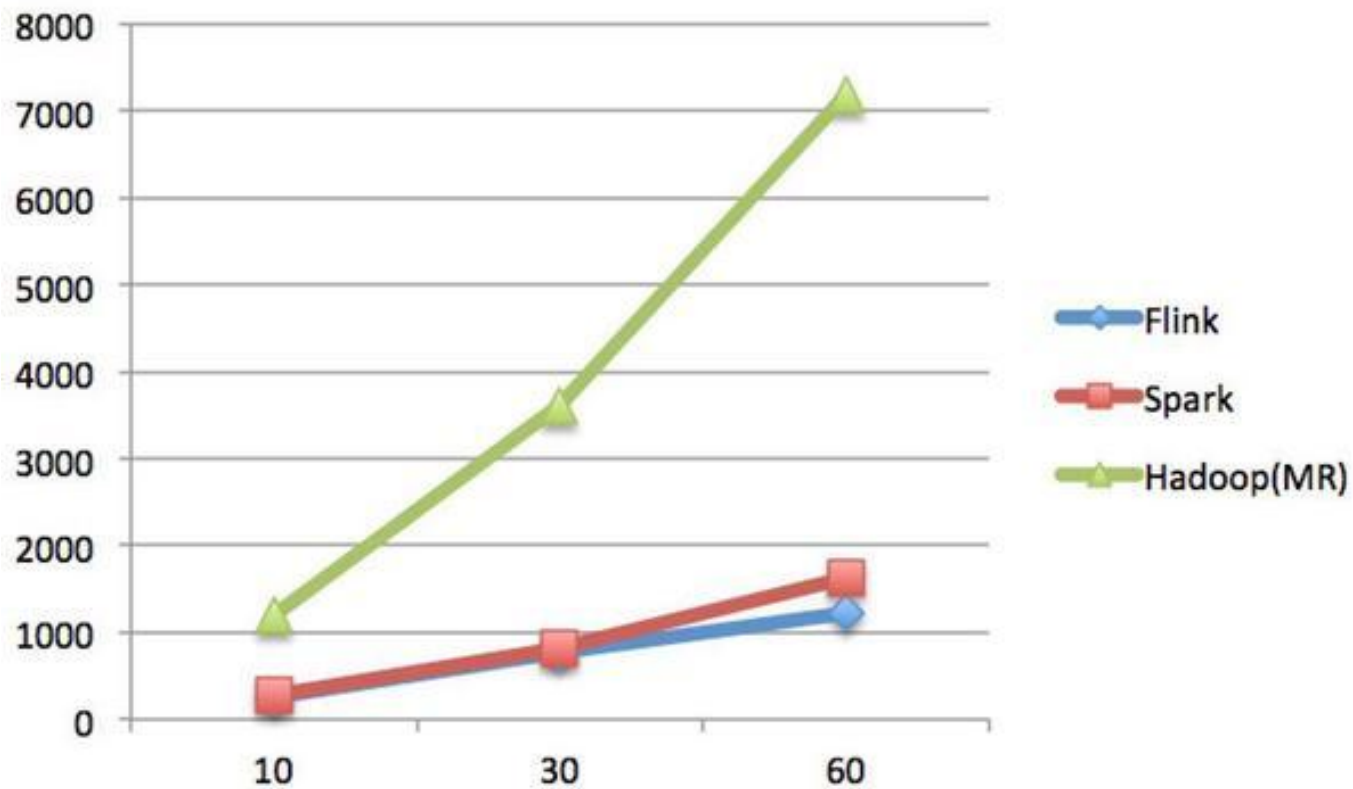


问题：Hadoop是否会被淘汰？





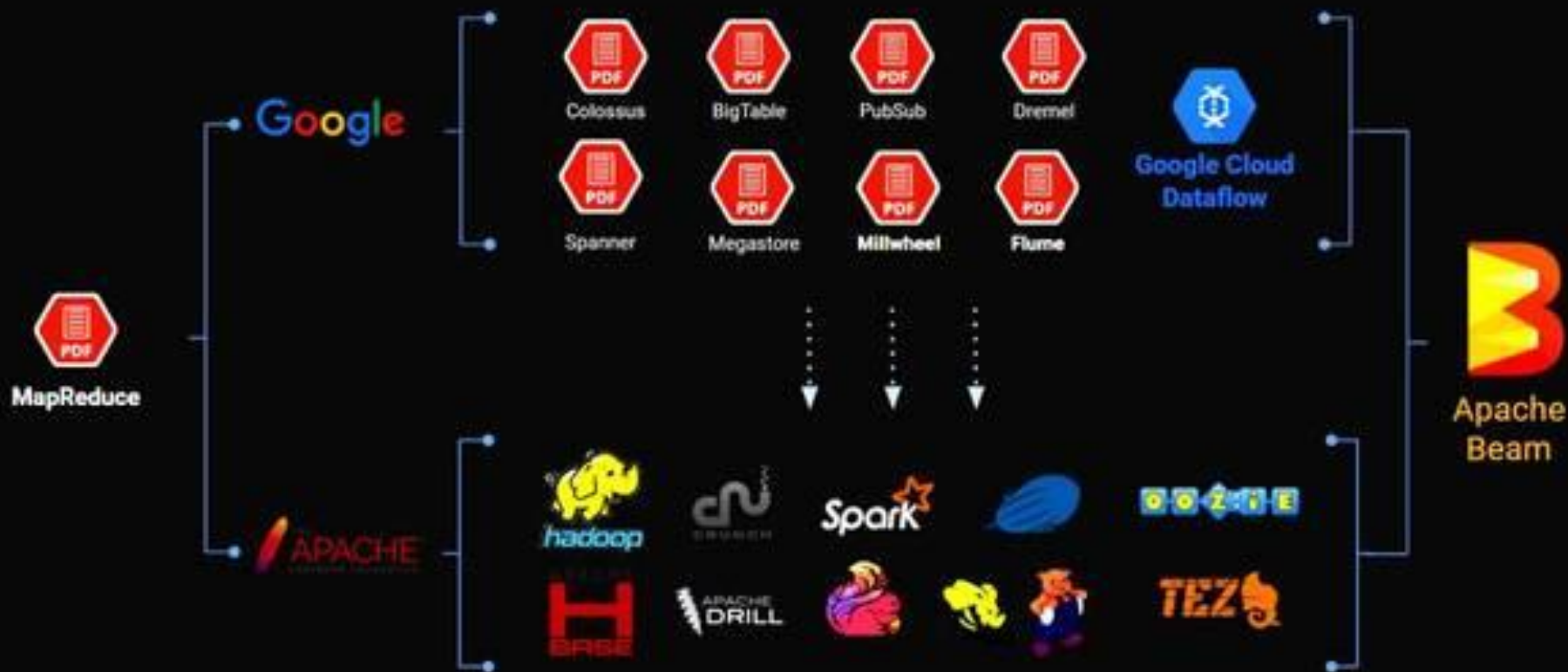
Flink架构图



既生瑜，何生亮！

谷歌，Beam，一统天下？

The Evolution of 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

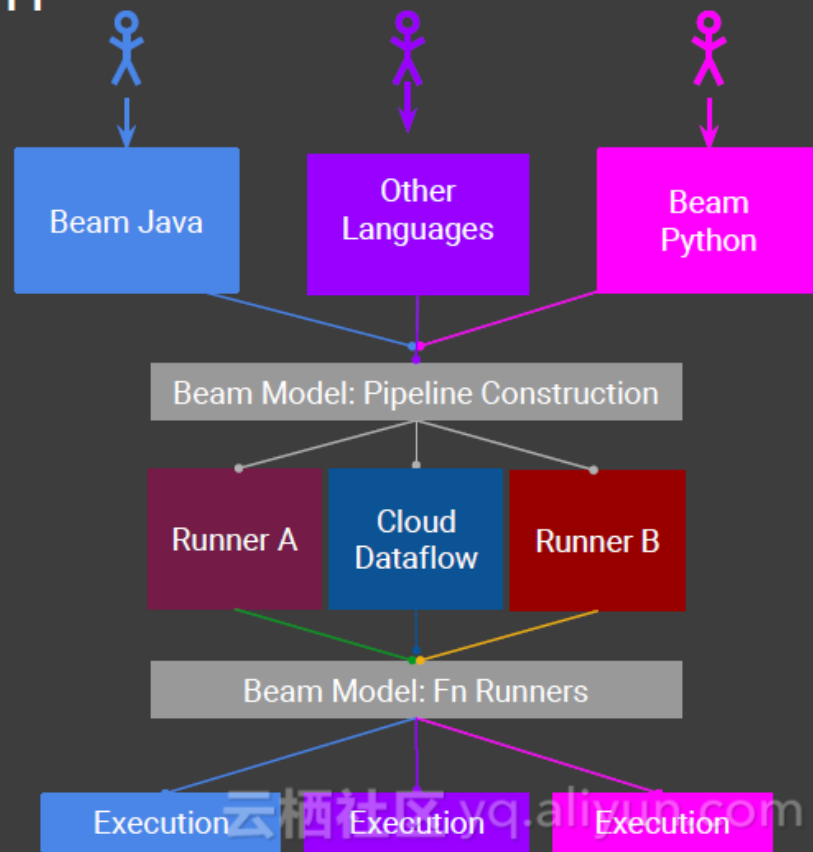


表 大数据计算模式及其代表产品

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

一

大数据概念

二

大数据关键技术

三

大数据产业化应用案例

四

大数据时代的个人感悟

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）



中国智能物流骨干网

“菜鸟”将物流资源重组，欲将运力变得更集中、高效



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

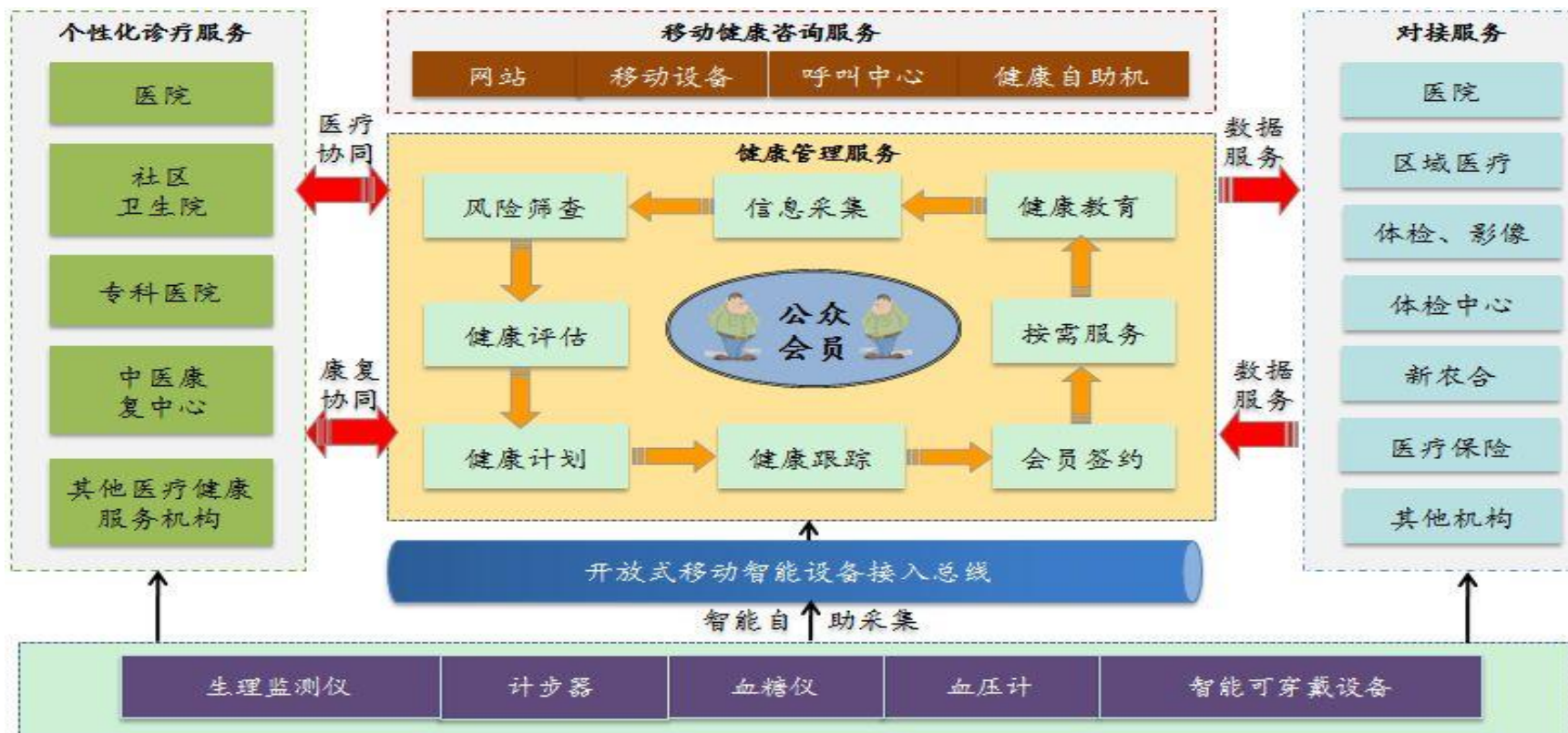
天猫牵头负责与各大物流快递公司对接的数据平台

地网

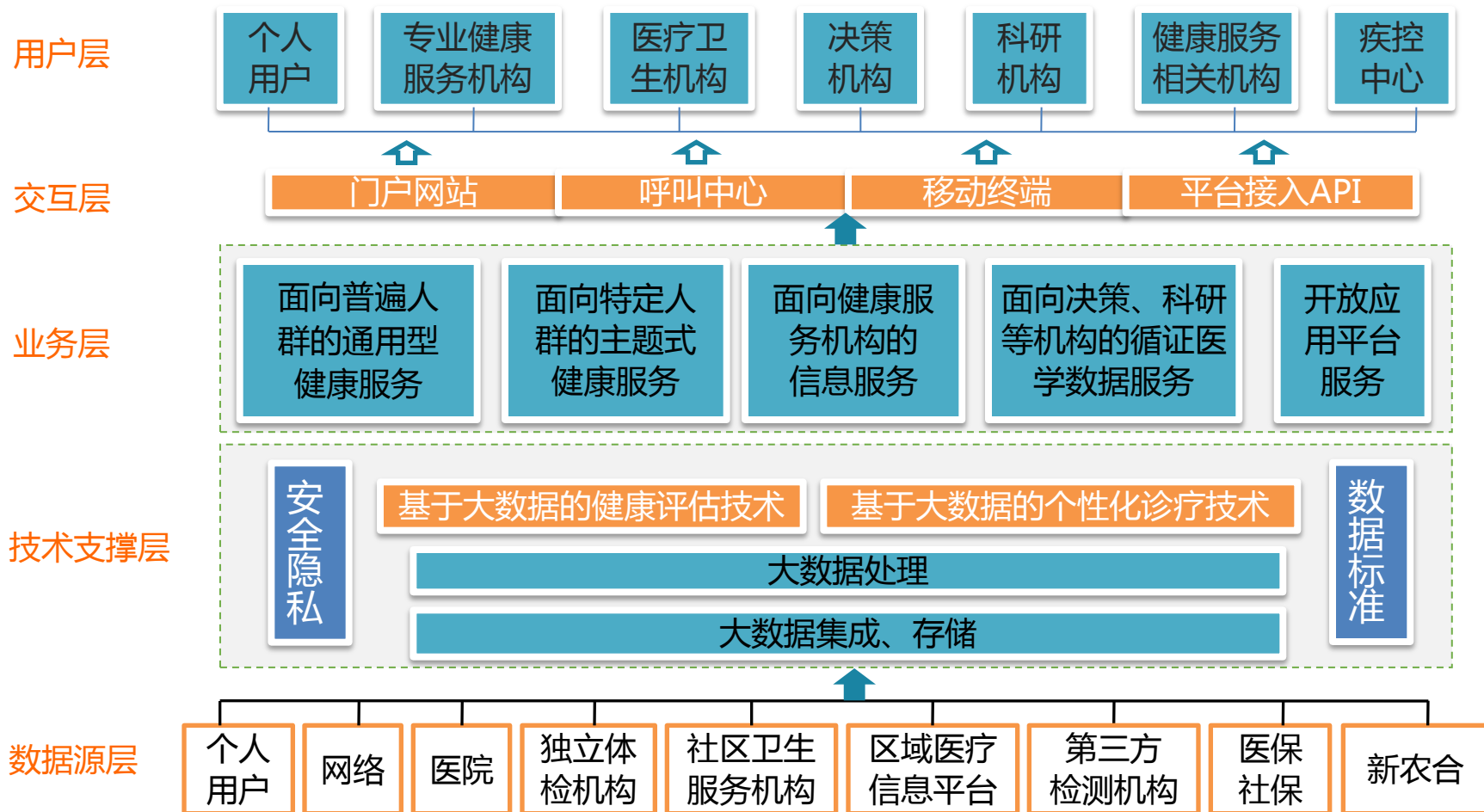
即“菜鸟”，又称“中国智能物流骨干网（CSN）”

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。



基于大数据的综合健康服务平台



- 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)**

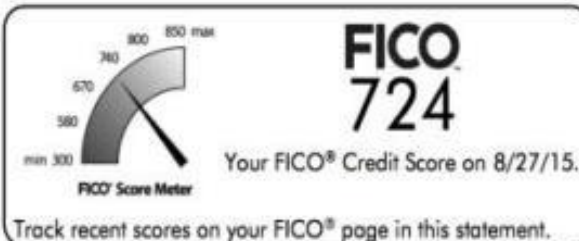
ACCOUNT SUMMARY

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

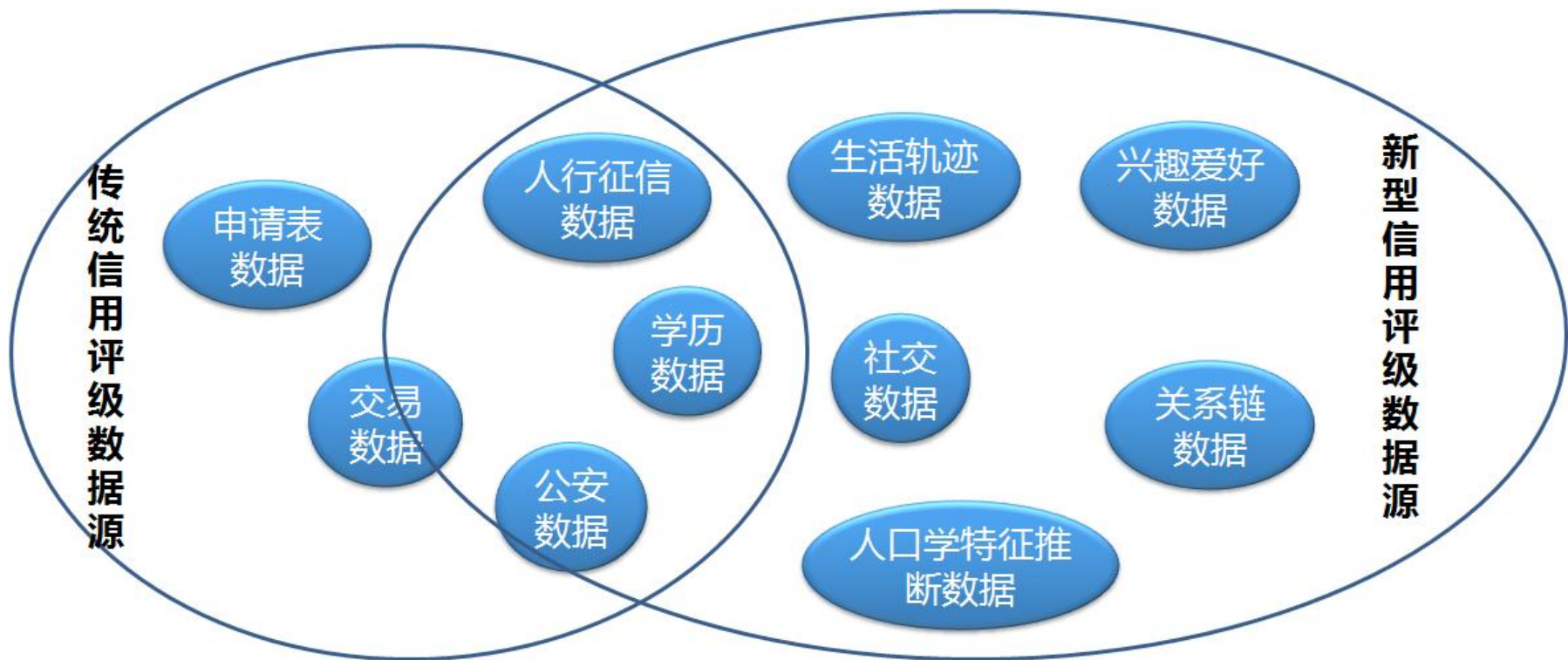
Credit Line	\$1,900
Credit Line Available	\$1,900
Cash Advance Credit Line	\$400
Cash Advance Credit Line Available	\$400

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



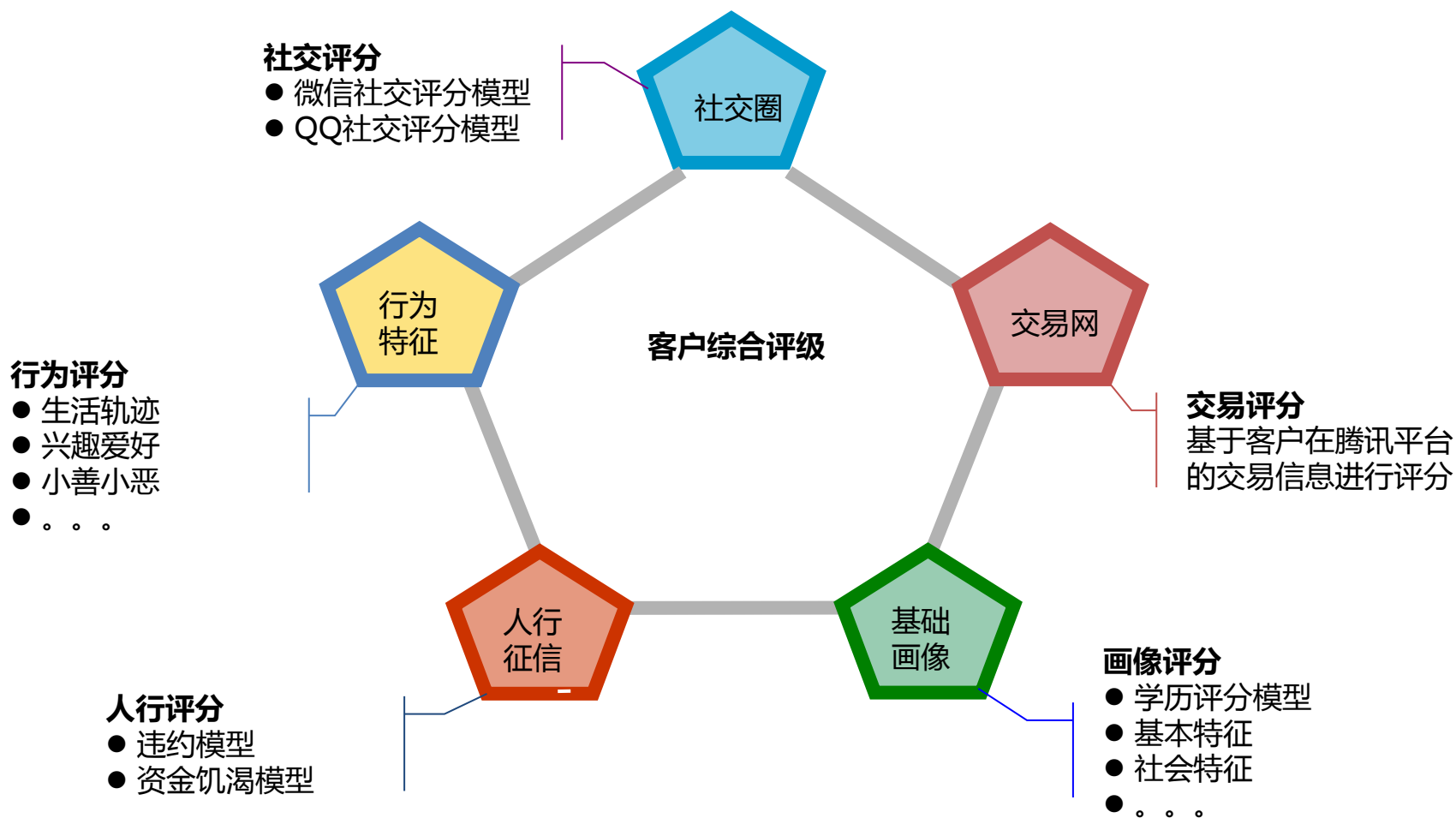
Please pay online at www.Discover.com or make checks payable to Discover. Phone or internet payment? Pay before midnight ET on your payment due date for same day credit.

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



目录

Contents

一

大数据概念

二

大数据关键技术

三

大数据产业化应用案例

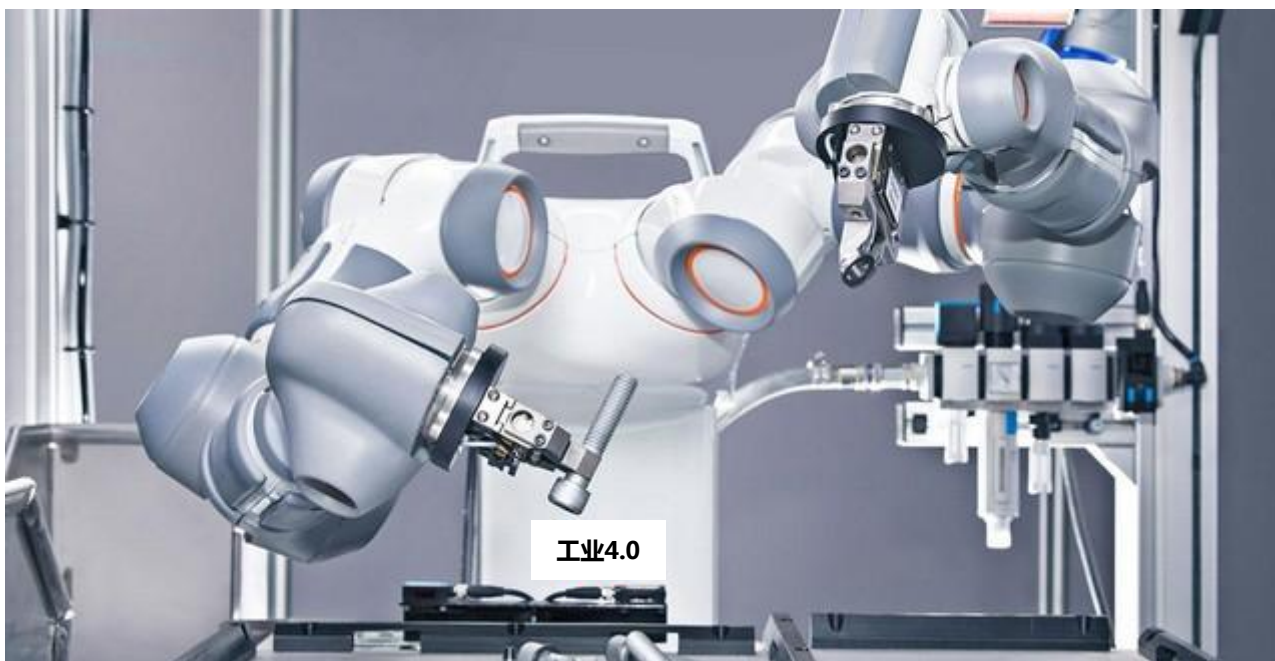
四

大数据时代的个人感悟

【大数据思维：以数据为中心】

大数据时代

这是一个以数据为中心的世界.....



THANKS

敬 请 批 评 指 正



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

简介：林子雨，男，1978年出生，北京大学博士，现为厦门大学计算机科学系助理教授，中国计算机学会数据库专委会委员，中国计算机学会信息系统专委会委员，获得2013年度和2017年度厦门大学教学类奖教金，提出并建设了国内高校首个“数字教师”。曾任厦门大学信息科学与技术学院院长助理，晋江市发展和改革委员会副局长。研究领域包括数据库、数据仓库、大数据、云计算和物联网。主持或参与了包括国家自然科学基金、863计划在内的多项国家重点科研项目，并以第一作者身份在《计算机学报》《软件学报》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，并编著出版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》，建设了国内高校首个大数据课程公共服务平台，平台年访问量超过100万次。建设了国内高校首个大数据课程教师培训交流基地，2016年暑假来自全国60多所高校的100余位教师参加了培训交流。具有丰富的政府和企业信息化培训经验，厦门大学管理学院EDP中心、浙江大学管理学院EDP中心、泉州市科技培训中心、厦门大学国际关系学院特邀培训讲师，曾给中国移动通信集团公司、福州马尾区政府、福建龙岩卷烟厂、福建省物联网科学研究院、石狮市物流协会、厦门市物流协会、浙江省中小企业家、四川泸州企业家、江苏沛县企业家等开展信息化培训，累计培训人数达3000人以上。

社会服务：面向企业和政府部门，提供科研支持、软件开发、企业信息化培训、科技项目申请书与产业规划文档撰写等服务。

项目经历：作为项目负责人主持完成的项目主要包括《城市信息化顶层设计框架研究报告》、《2015年泉州市互联网经济调研报告》、《厦门市云计算产业发展技术路线图》、《石狮市物流园区建设方案研究报告》、《晋江市海洋生物科技园区建设方案建议书》、《国家物联网重大应用示范工程重点项目福建省物联网科学院平台实施方案》、《国家物联网重大应用示范工程区域试点泉州市总体工作方案》、《基于地面远程控制的新一代智能塔吊系统实施方案》、《基于大数据的综合健康服务平台2015科技部支撑计划项目申报书》等。



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>