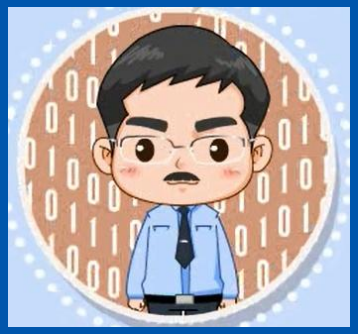


大数据公开课全国高校巡讲计划
厦门华夏学院专场

2019年11月5日 厦门

大数据概念、关键技术和应用

（PPT版本号：2019年11月5日版本）

扫一扫访问专场主页

林子雨

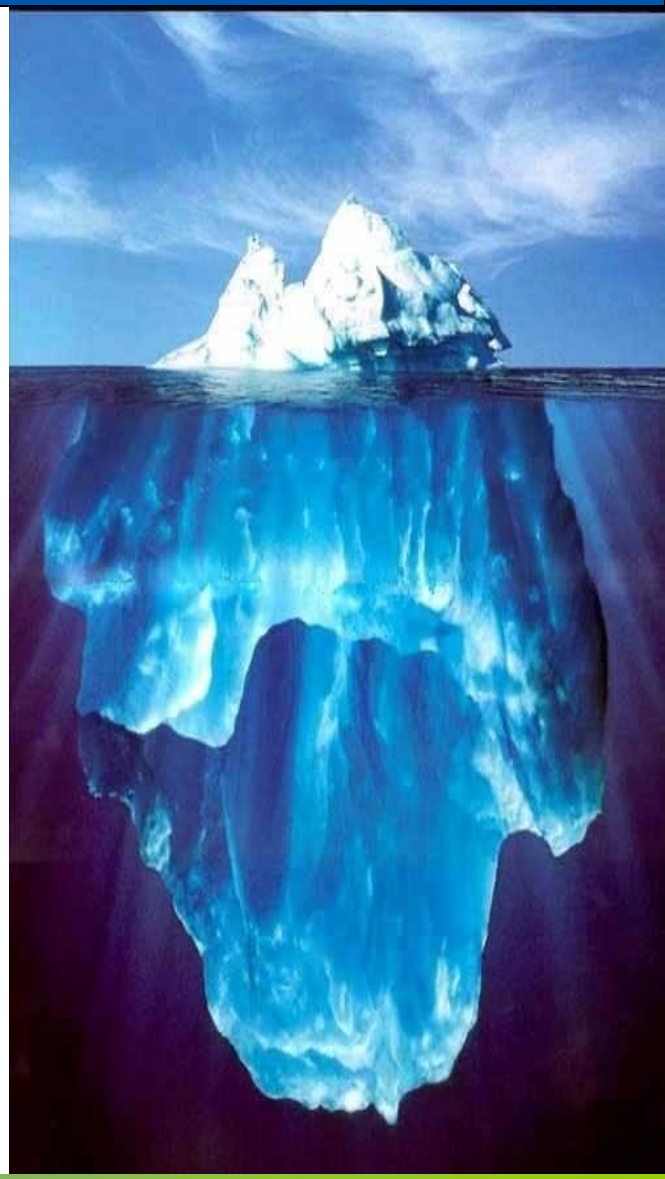
厦门大学计算机科学系

E-mail: ziyulin@xmu.edu.cn

主页：http://cs.xmu.edu.cn/linziyu

提纲

1.大数据的发展
2.大数据的概念
2.大数据关键技术
3.大数据的应用



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

主讲教师

数字教师LOGO

主讲教师：林子雨

中国高校首个“**数字教师**”提出者和建设者
2009年7月从事教师职业以来
累计**免费**网络发布超过500万字高价值教学和科研资料
网络浏览量超过500万次

1.大数据的发展

- 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 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.大数据的发展

表 大数据发展的三个阶段

阶段	时间	内容
第一阶段：萌芽期	上世纪90年代至本世纪初	随着数据挖掘理论和数据库技术的逐步成熟，一批商业智能工具和知识管理技术开始被应用，如数据仓库、专家系统、知识管理系统等。
第二阶段：成熟期	本世纪前十年	Web2.0应用迅猛发展，非结构化数据大量产生，传统处理方法难以应对，带动了大数据技术的快速突破，大数据解决方案逐渐走向成熟，形成了并行计算与分布式系统两大核心技术，谷歌的GFS和MapReduce等大数据技术受到追捧，Hadoop平台开始大行其道
第三阶段：大规模应用期	2010年以后	大数据应用渗透各行各业，数据驱动决策，信息社会智能化程度大幅提高

1.大数据的发展

世界各国的大数据发展战略

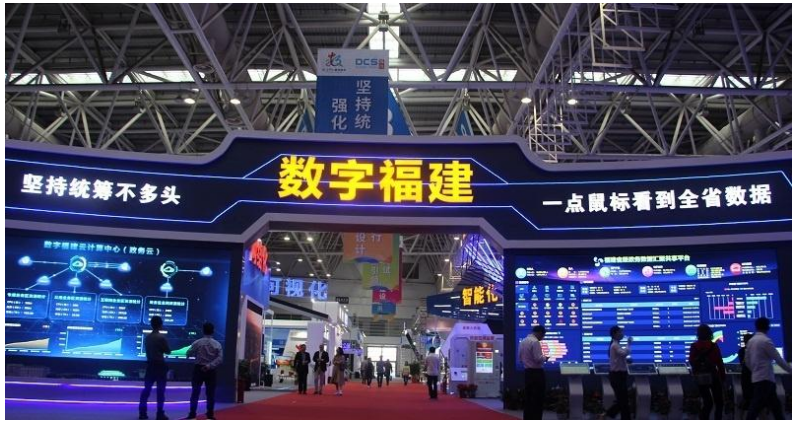
国家	战略
美国	稳步实施“三步走”战略，打造面向未来的大数据创新生态
英国	紧抓大数据产业机遇，应对脱欧后的经济挑战
法国	通过发展创新性解决方案并应用于实践来促进大数据发展
韩国	以大数据等技术为核心应对第四次工业革命
日本	开放公共数据，夯实应用开发
中国	实施国家大数据战略，加快建设数字中国

1.大数据的发展

•2015年8月，国务院印发了《促进大数据发展行动纲要》。党的十八届五中全会将大数据上升为国家战略。在党的十九大报告中，习近平总书记明确指出：“推动互联网、大数据、人工智能和实体经济深度融合”

•2017年1月，为加快实施国家大数据战略，推动大数据产业健康快速发展，工业和信息化部印发了《大数据产业发展规划（2016-2020年）》

•2018年4月22日-24日，首届“数字中国”建设峰会在福建省福州市举行



1.大数据的发展

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

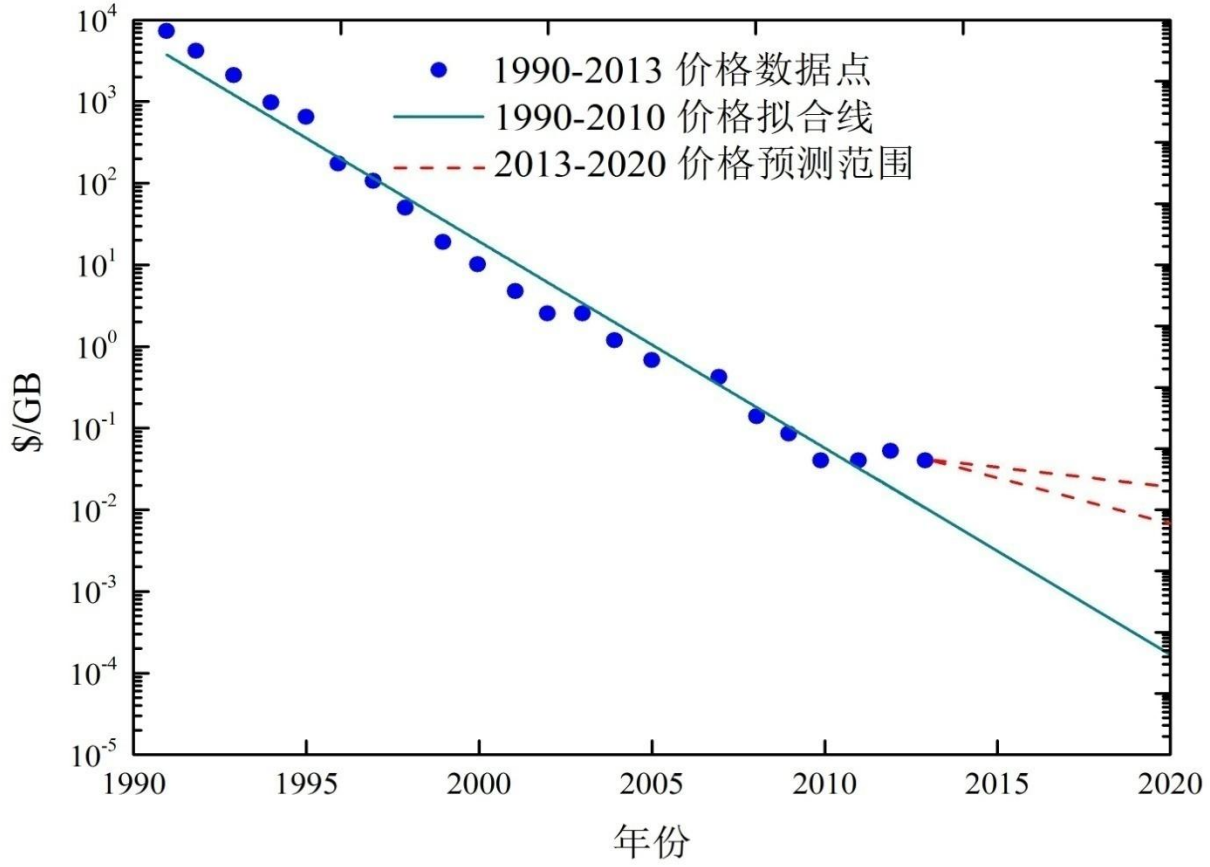


图 存储价格随时间变化情况

1.大数据的发展

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

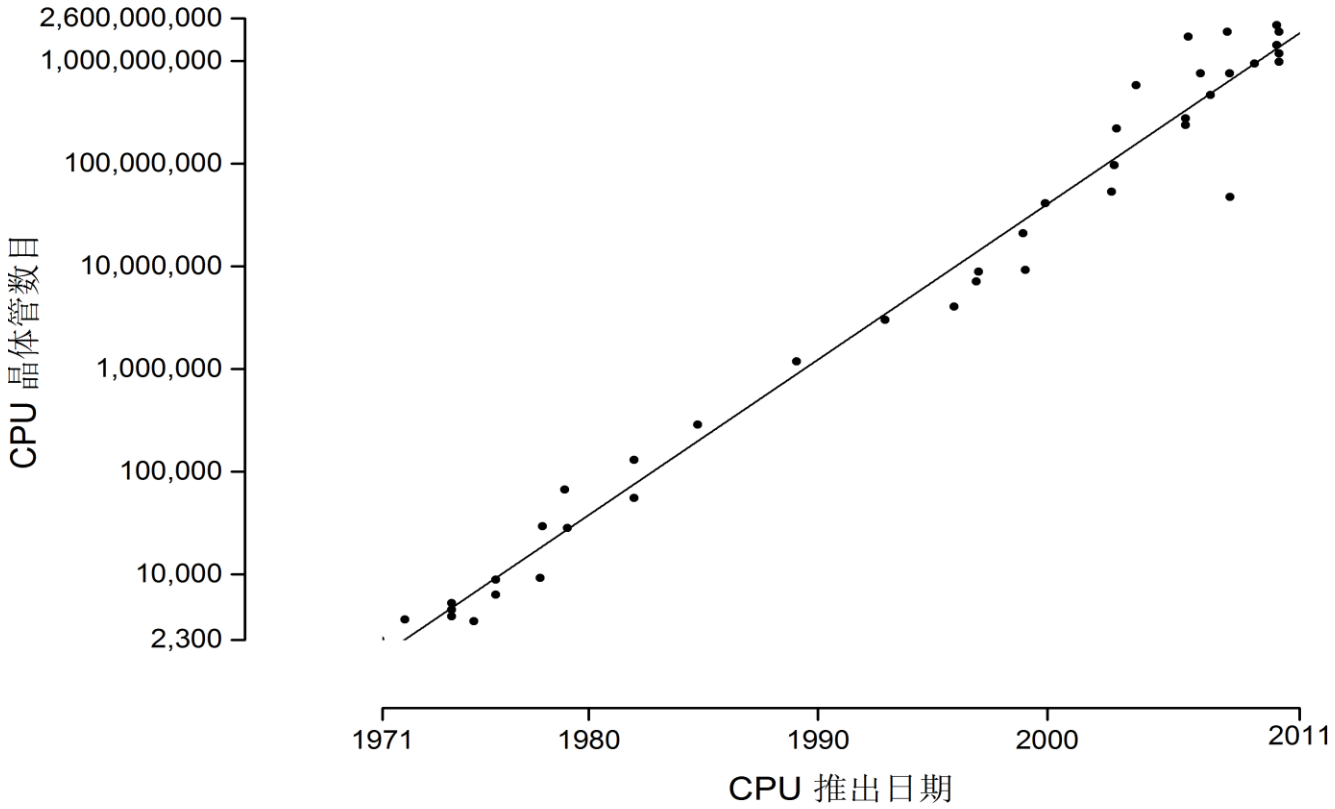


图 CPU晶体管数目随时间变化情况

1.大数据的发展

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

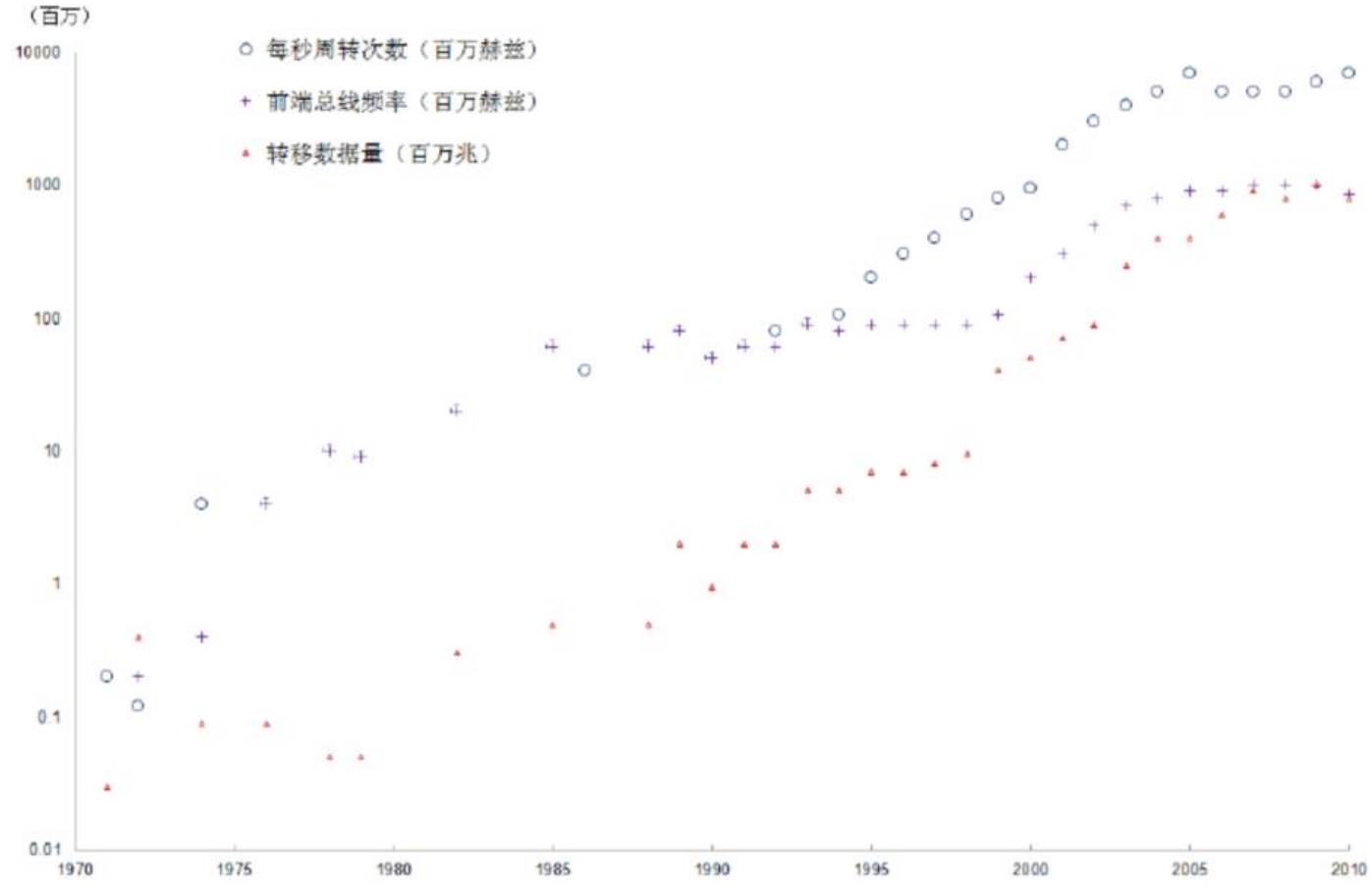


图 网络带宽随时间变化情况

1.大数据的发展

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

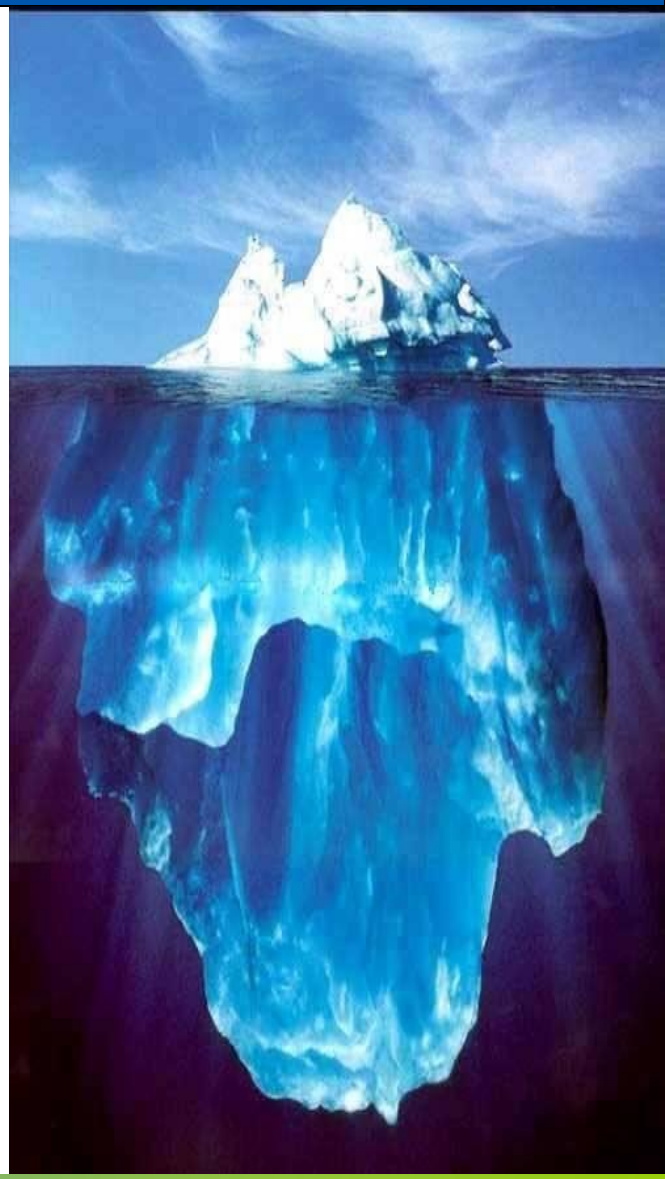
3 感知式系统阶段

•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

提纲

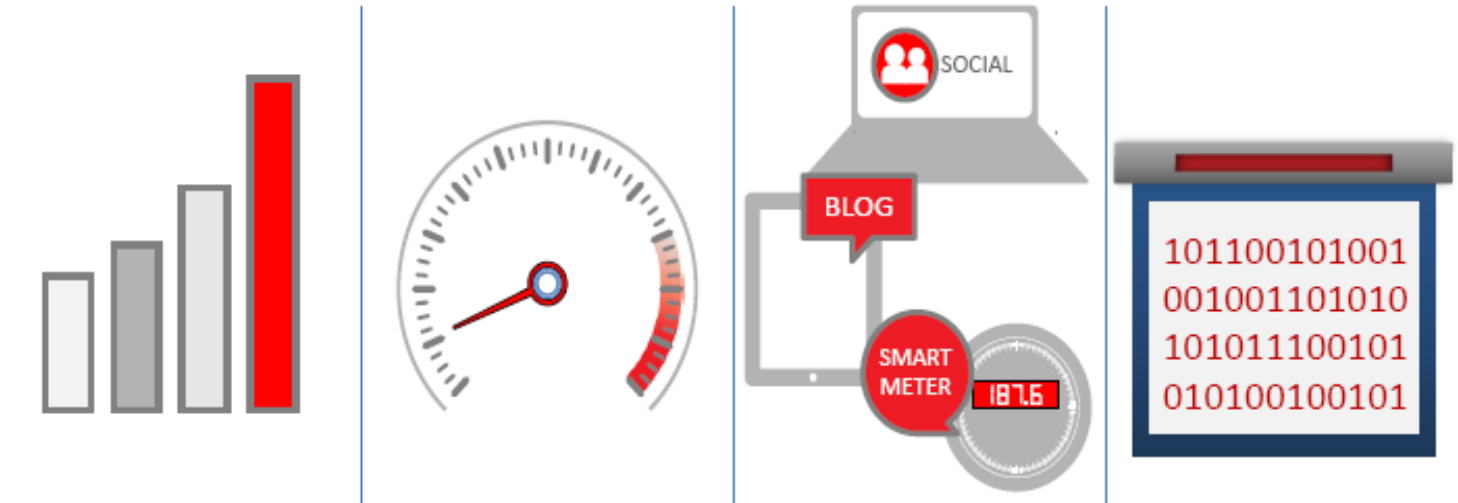
1.大数据的发展
2.大数据的概念
2.大数据关键技术
3.大数据的应用



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

2.大数据的概念



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。

2.大数据的概念

- 云计算、大数据和物联网代表了IT领域最新的技术发展趋势，三者既有区别又有联系

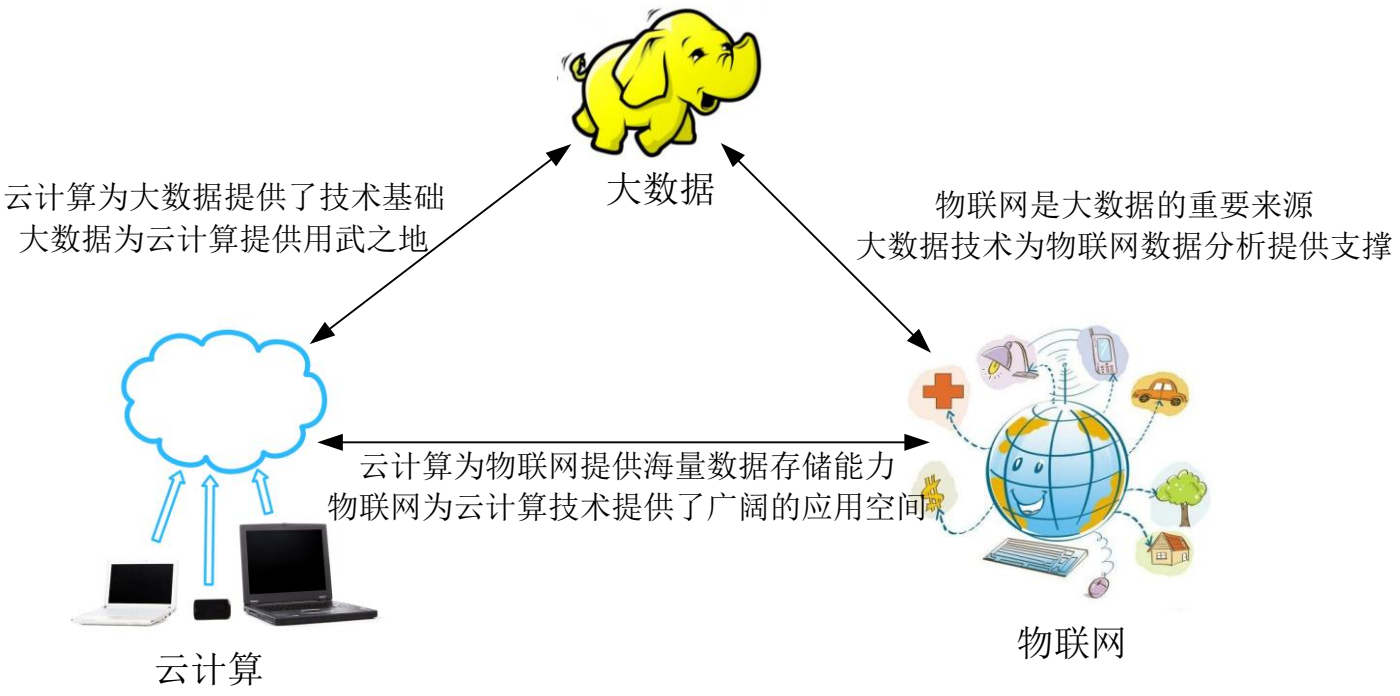
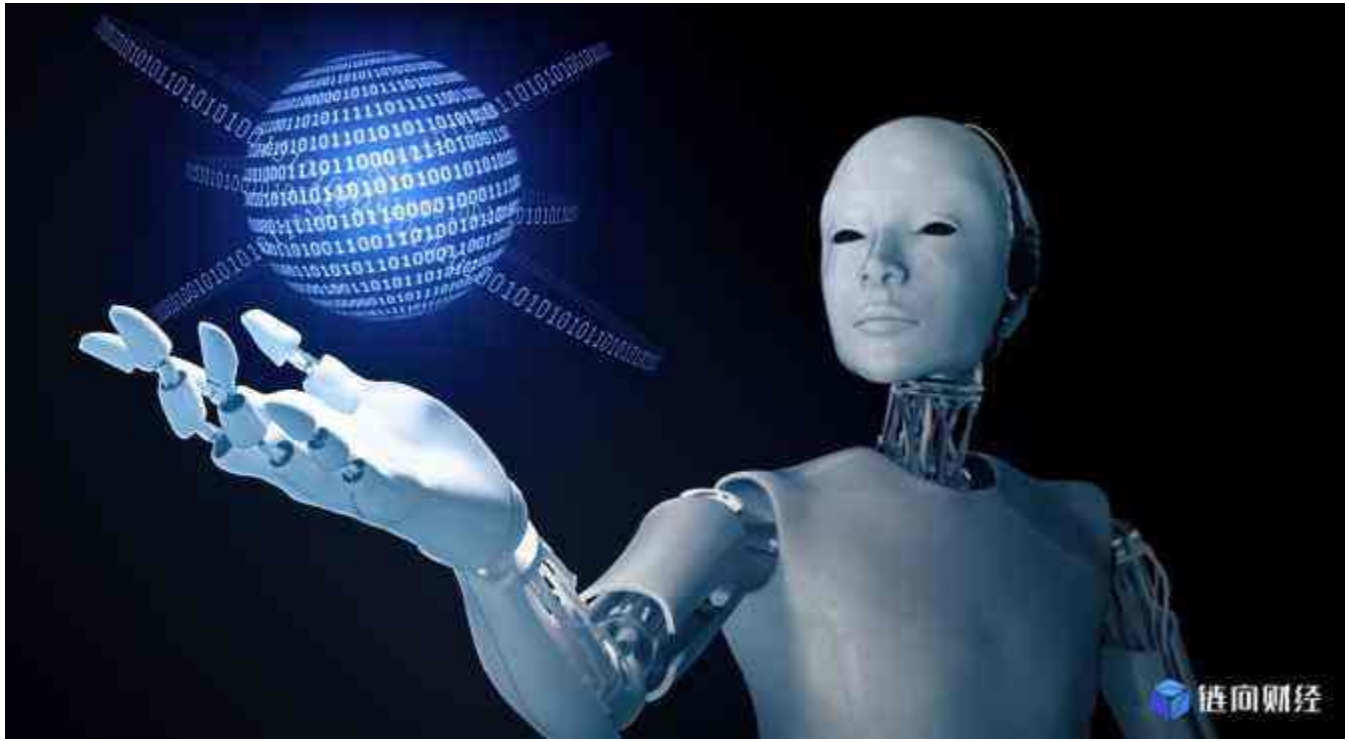


图 大数据、云计算和物联网之间的关系

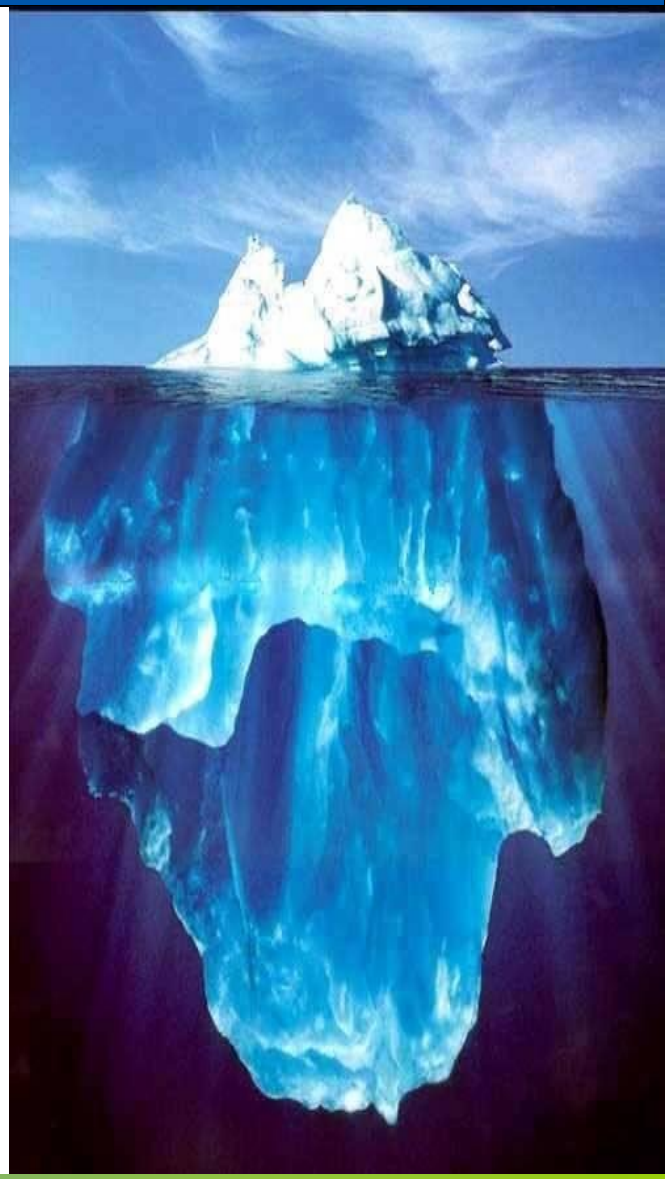
2.大数据的概念

大数据与人工智能的关系



提纲

1.大数据的发展
2.大数据的概念
2.大数据关键技术
3.大数据的应用



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

3.大数据关键技术

表　　大数据技术的不同层面及其功能

技术层面	功能
数据采集与预处理	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；利用日志采集工具（如Flume、Kafka等）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；利用网页爬虫程序到互联网网站中爬取数据
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
数据可视化	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
数据安全和隐私保护	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

3.大数据关键技术

- （1）数据采集与预处理

3.大数据关键技术

•数据清洗

1.需要清洗的数据的主要类型

•残缺数据
•错误数据
•重复数据

2.数据清洗的内容

•一致性检查
•无效值和缺失值的处理
•估算
•整例删除
•变量删除
•成对删除

3.大数据关键技术

（2）数据存储和管理

- 传统的数据存储和管理技术
  - 文件系统
  - 关系数据库
  - 数据仓库
  - 并行数据库

- 大数据时代的数据存储和管理技术
  - 分布式文件系统
  - NewSQL和NoSQL数据库

3.大数据关键技术

（3）数据处理与分析

- 数据挖掘和机器学习算法
- 大数据处理与分析技术

3.大数据关键技术

•数据挖掘和机器学习是计算机学科中最活跃的研究分支之一。机器学习是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科，专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能，它是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。

•数据挖掘是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘可以视为机器学习与数据库的交叉，它主要利用机器学习界提供的算法来分析海量数据，利用数据库界提供的存储技术来管理海量数据。

3.大数据关键技术

典型的机器学习和数据挖掘算法

- 分类
- 聚类
- 回归分析
- 关联规则

3.大数据关键技术

大数据处理分析技术类型及其代表产品

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

3.大数据关键技术

- 代表性大数据处理分析技术
  - Hadoop
  - Spark
  - Flink
  - Beam

3.大数据关键技术

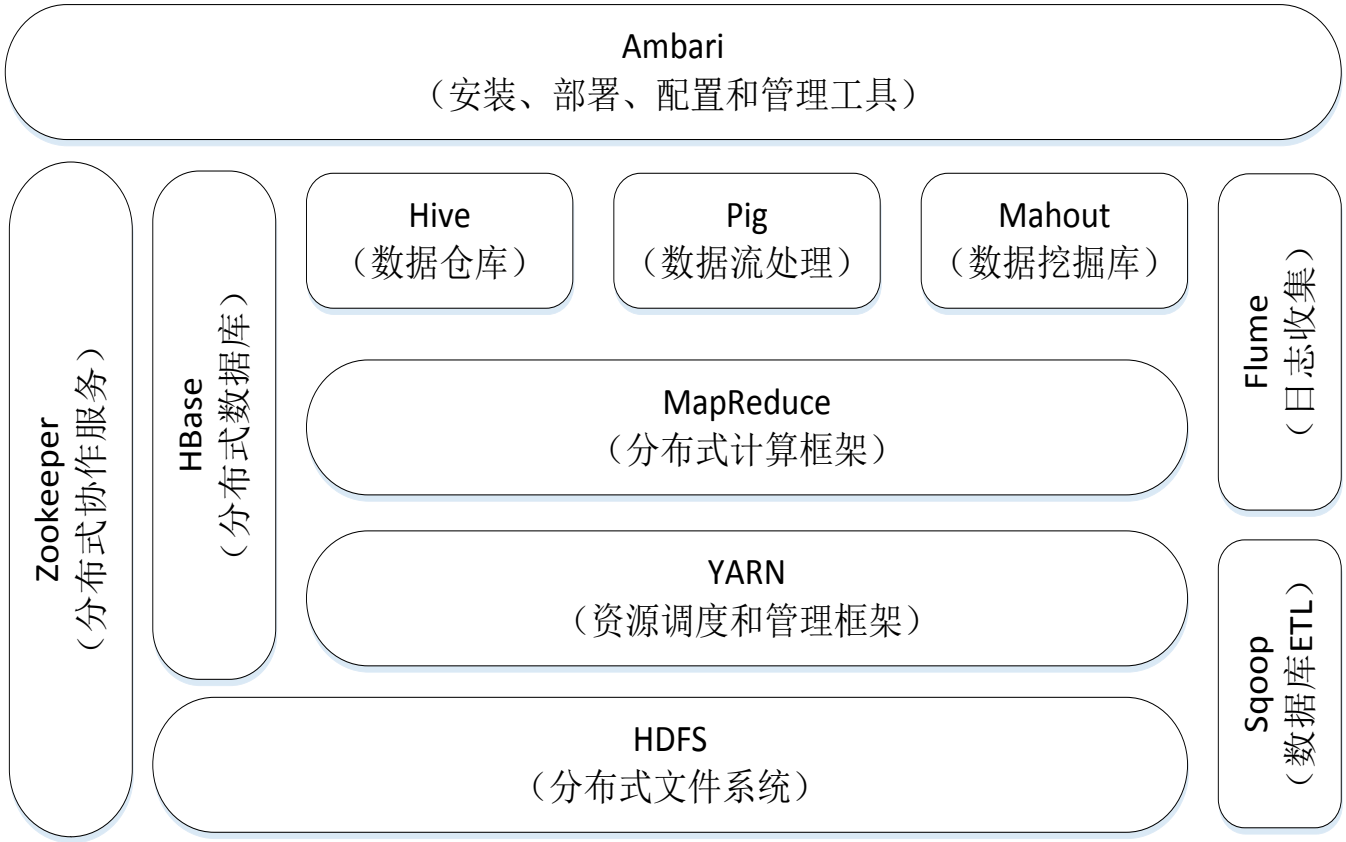
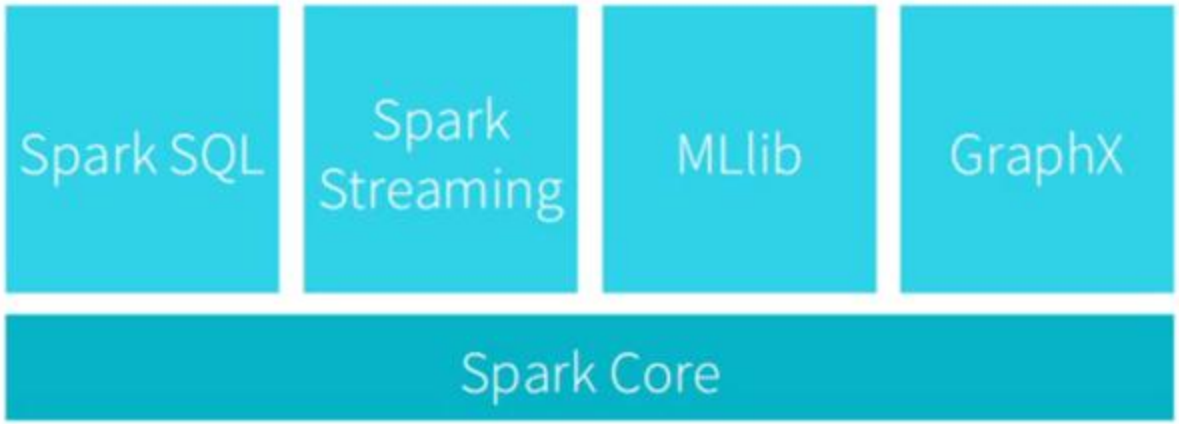


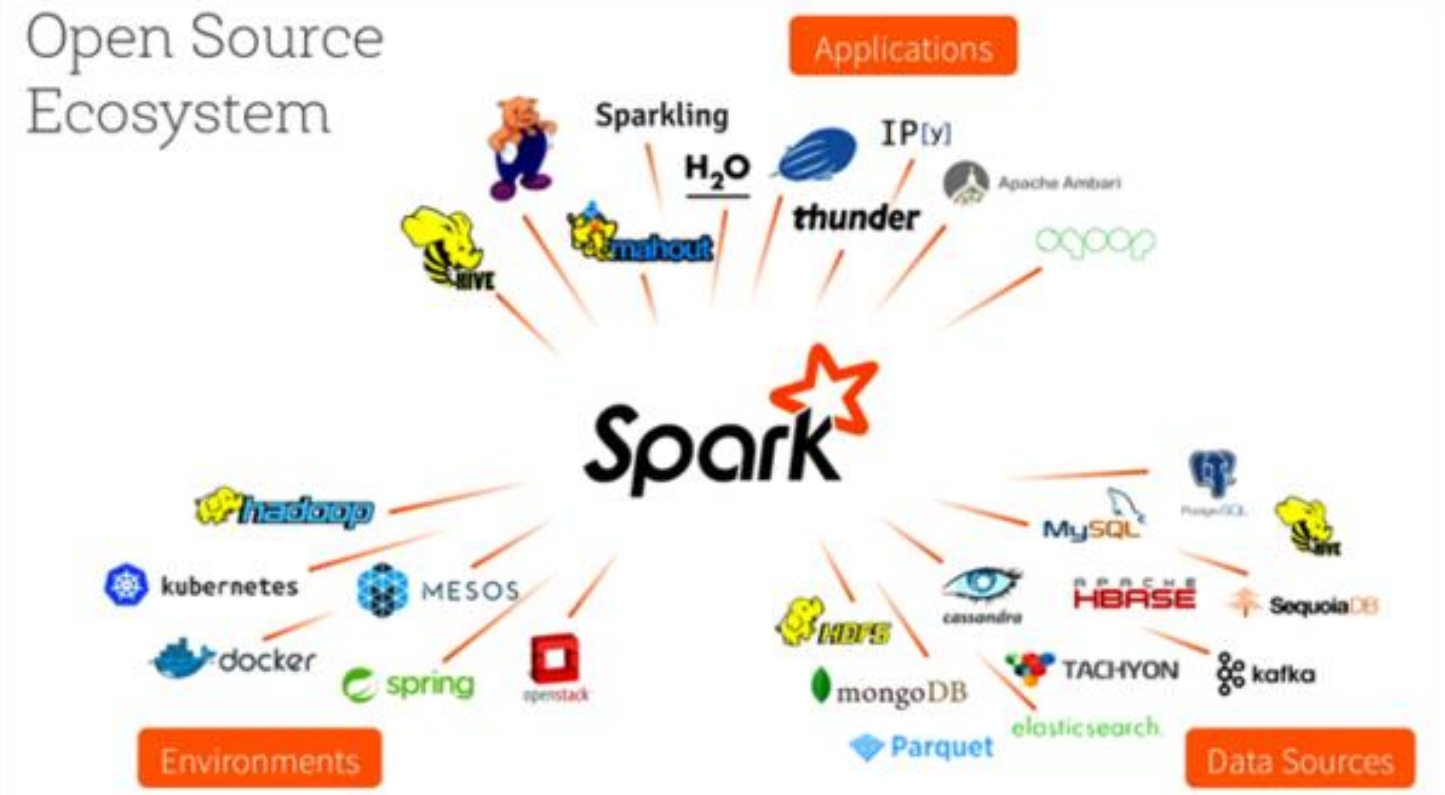
图 Hadoop生态系统

3.大数据关键技术



Spark架构图

3.大数据关键技术



Spark生态系统

3.大数据关键技术

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

3.大数据关键技术

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：

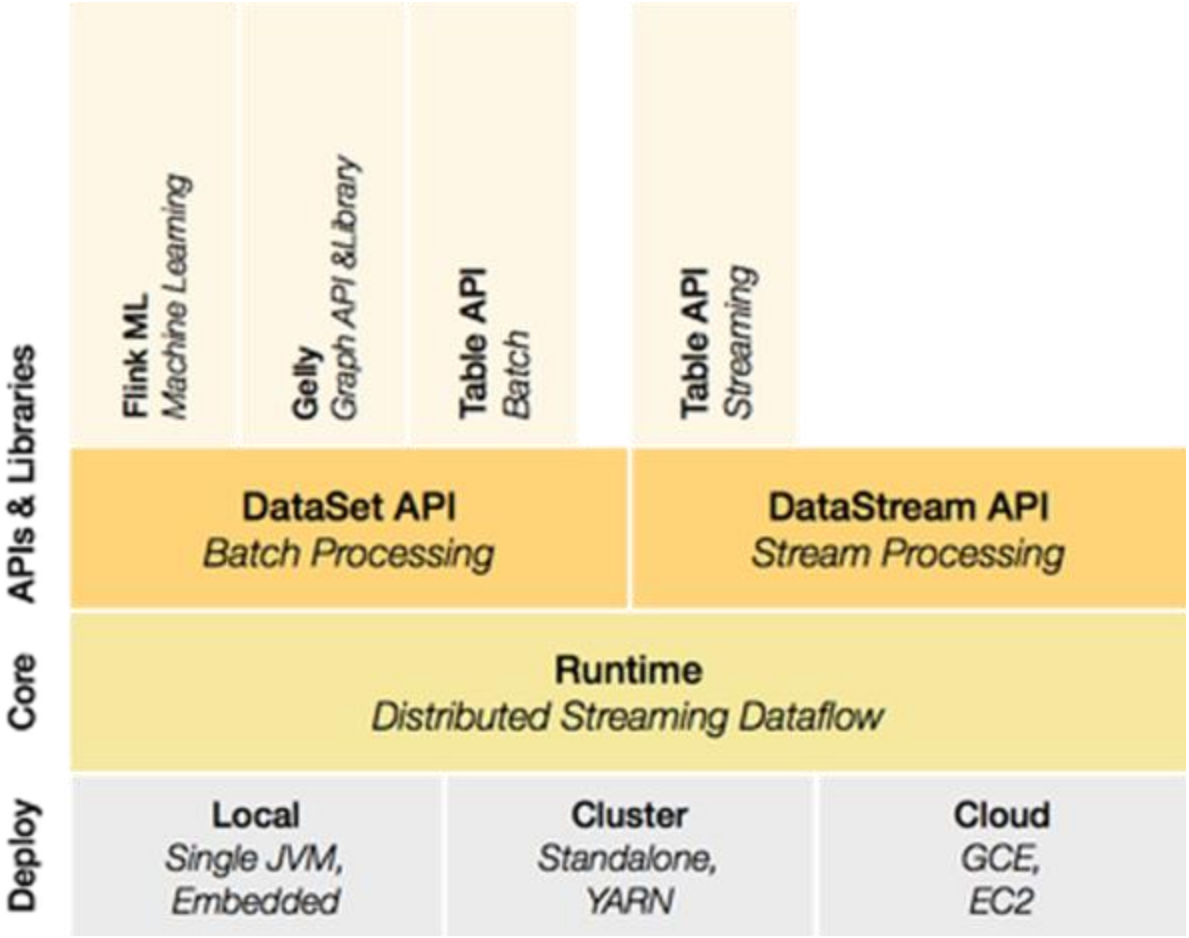
•Spark的计算模式也属于MapReduce，但不局限于Map和Reduce操作，还提供了多种数据集操作类型，编程模型比Hadoop MapReduce更灵活
•Spark提供了内存计算，可将中间结果放到内存中，对于迭代运算效率更高
Spark基于DAG的任务调度执行机制，要优于Hadoop MapReduce的迭代执行机制

3.大数据关键技术

问题：Spark会取代Hadoop吗？

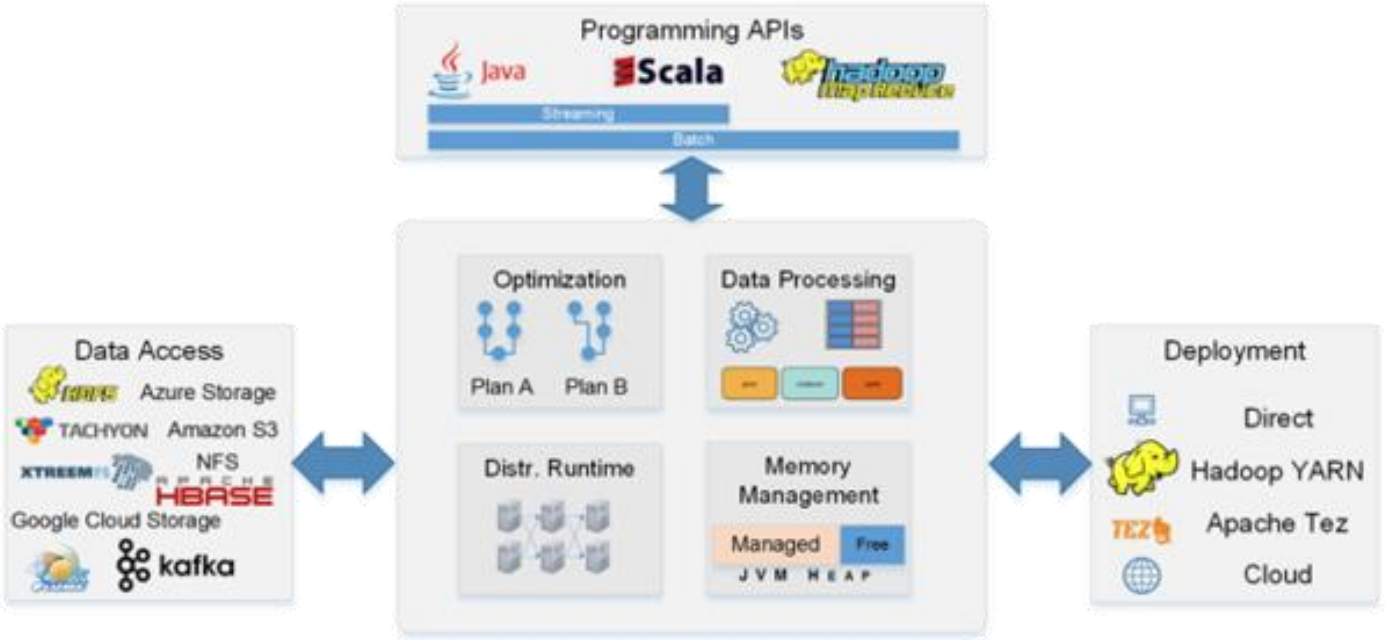
- Hadoop包括两大核心：HDFS和MapReduce
- Spark作为计算框架，与MapReduce是对等的
- 谈到“取代”，Spark应该是取代MapReduce，而不是整个Hadoop
- Spark和Hadoop生态系统共存共荣，Spark借助于Hadoop的HDFS、HBase等来完成数据的存储，然后，由Spark完成数据的计算

3.大数据关键技术



Flink架构图

3.大数据关键技术



Flink生态系统

3.大数据关键技术

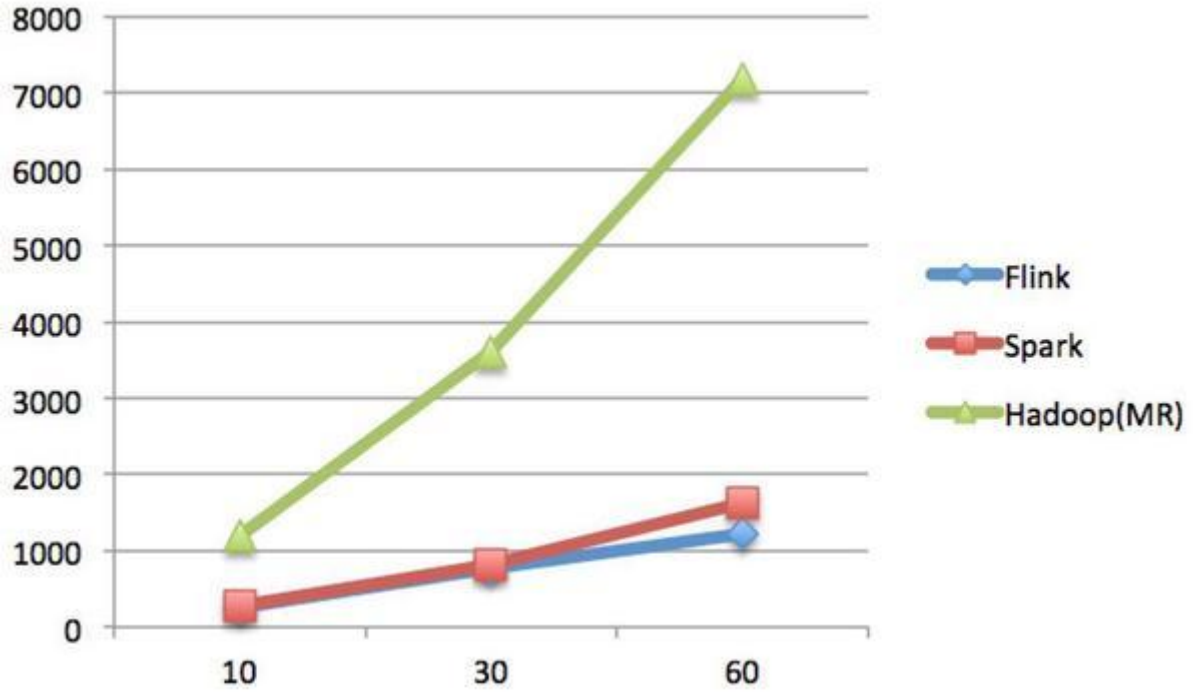
Flink与Spark的比较

	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

2.大数据关键技术

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。

2.大数据关键技术

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。

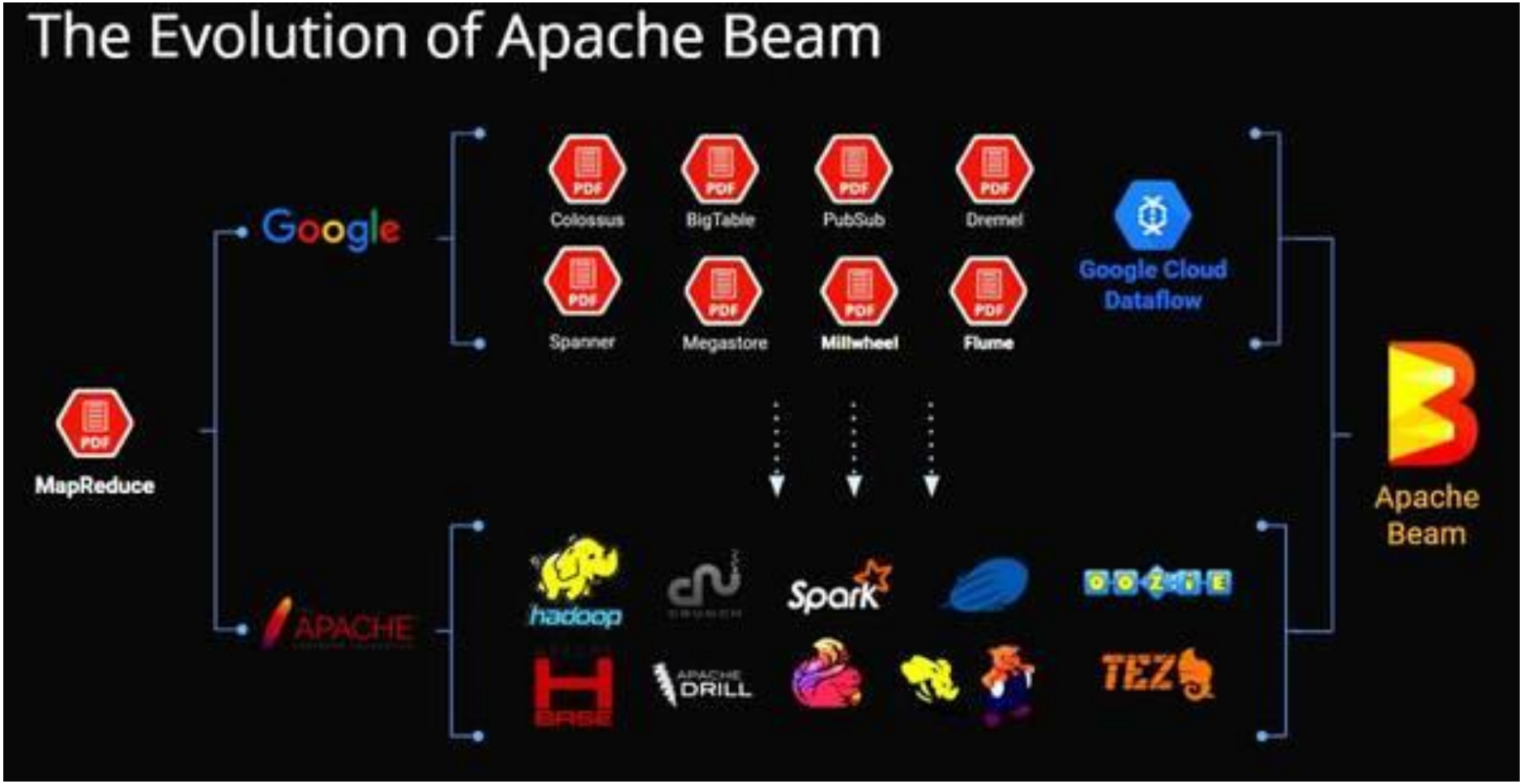
SQL支持

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！

2.大数据关键技术

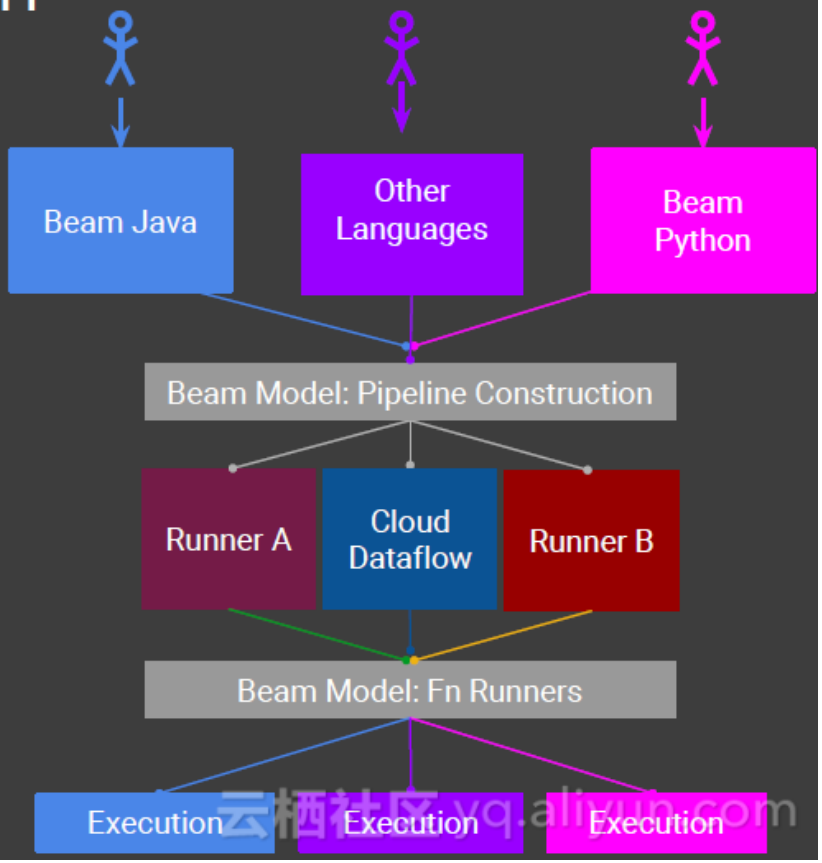
谷歌，Beam，一统天下？



2.大数据关键技术

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



3.大数据关键技术

（4）数据可视化

- 数据可视化是指将大型数据集中的数据以图形图像形式表示，并利用数据分析和开发工具发现其中未知信息的处理过程
- 数据可视化技术的基本思想是将数据库中每一个数据项作为单个图元素表示，大量的数据集构成数据图像，同时将数据的各个属性值以多维数据的形式表示，可以从不同的维度观察数据，从而对数据进行更深入的观察和分析

2.大数据关键技术

在大数据时代，可视化技术可以支持实现多种不同的目标：

（a）观测、跟踪数据

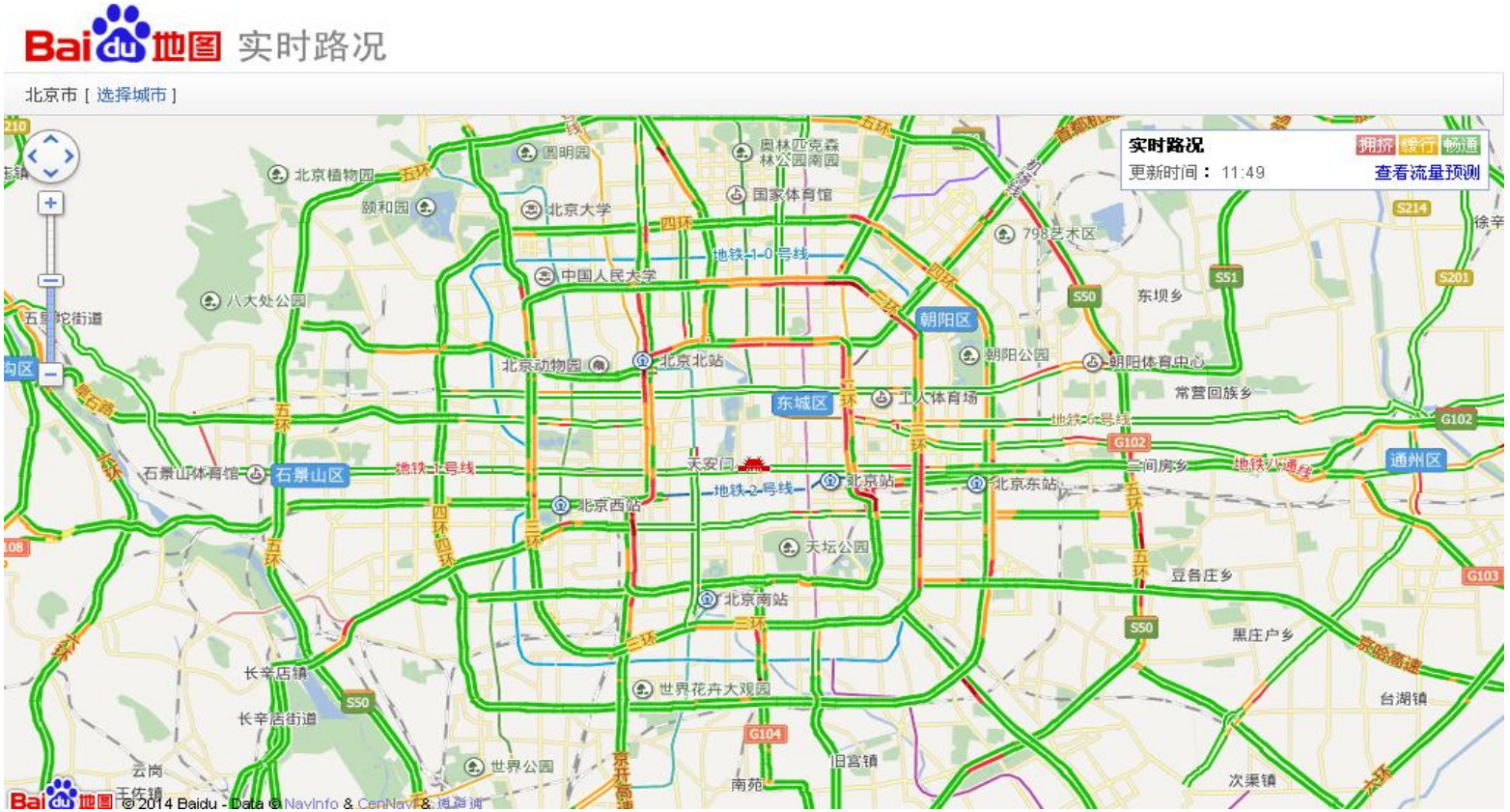


图 百度地图显示的北京市实时交通路况信息

2.大数据关键技术

（b）分析数据

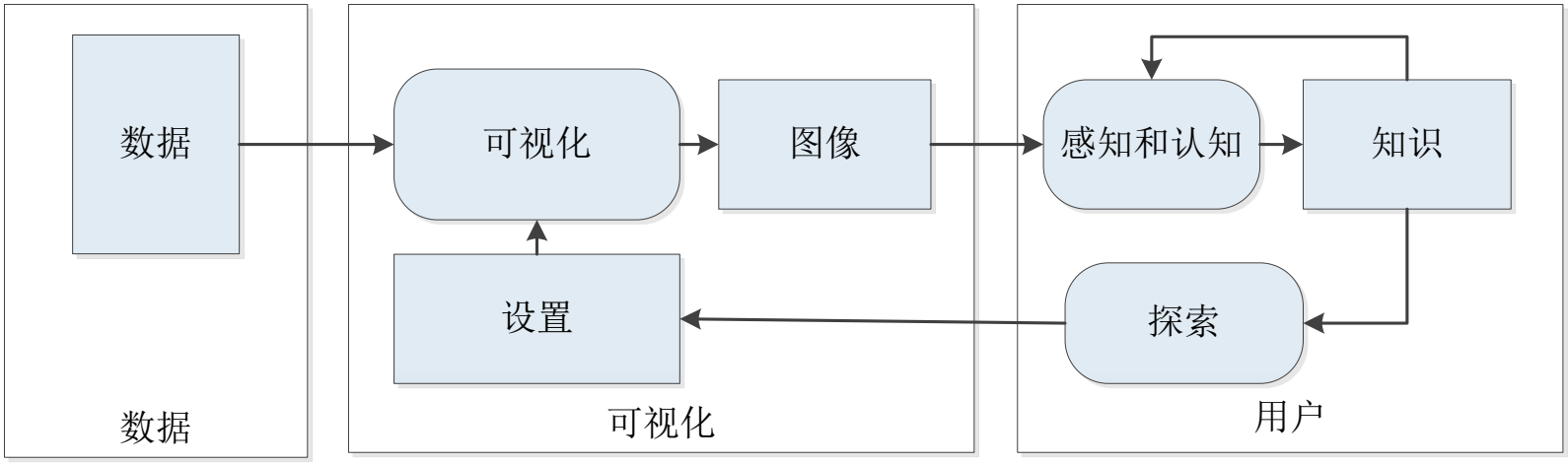


图 用户参与的可视化分析过程

2.大数据关键技术

（c）辅助理解数据

图 微软“人立方”展示的人物关系图

2.大数据关键技术

（d）增强数据吸引力

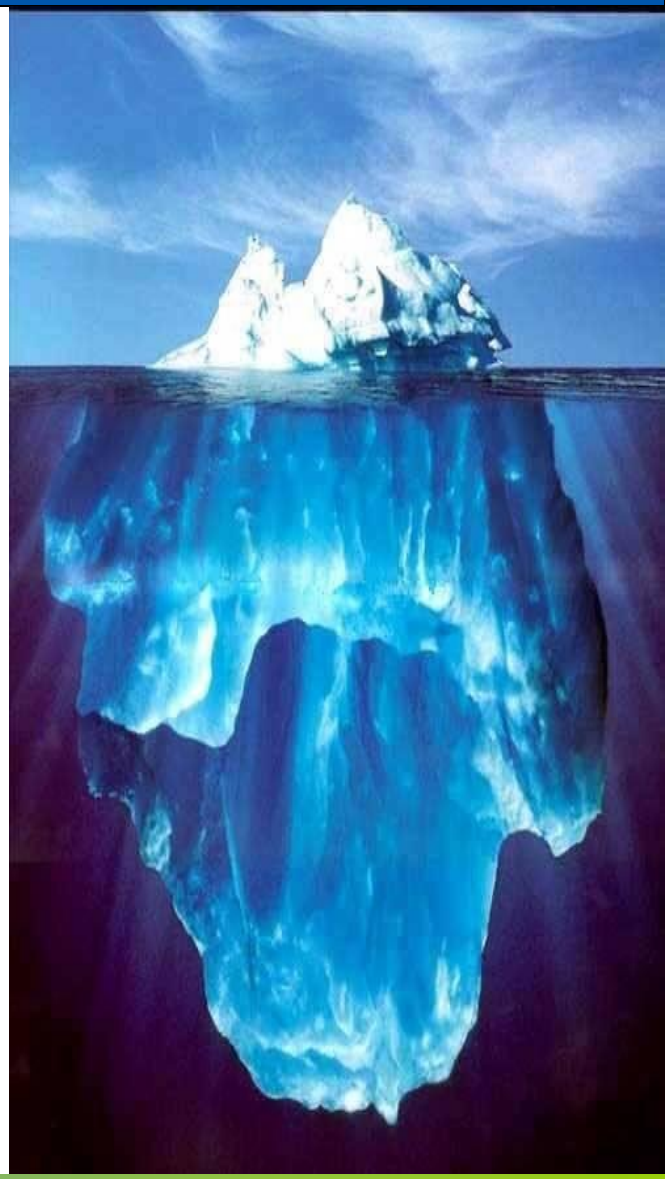
图 一个可视化的图表新闻实例

3.大数据关键技术

（5）数据安全和隐私保护

提纲

1.大数据的发展
2.大数据的概念
2.大数据关键技术
3.大数据的应用

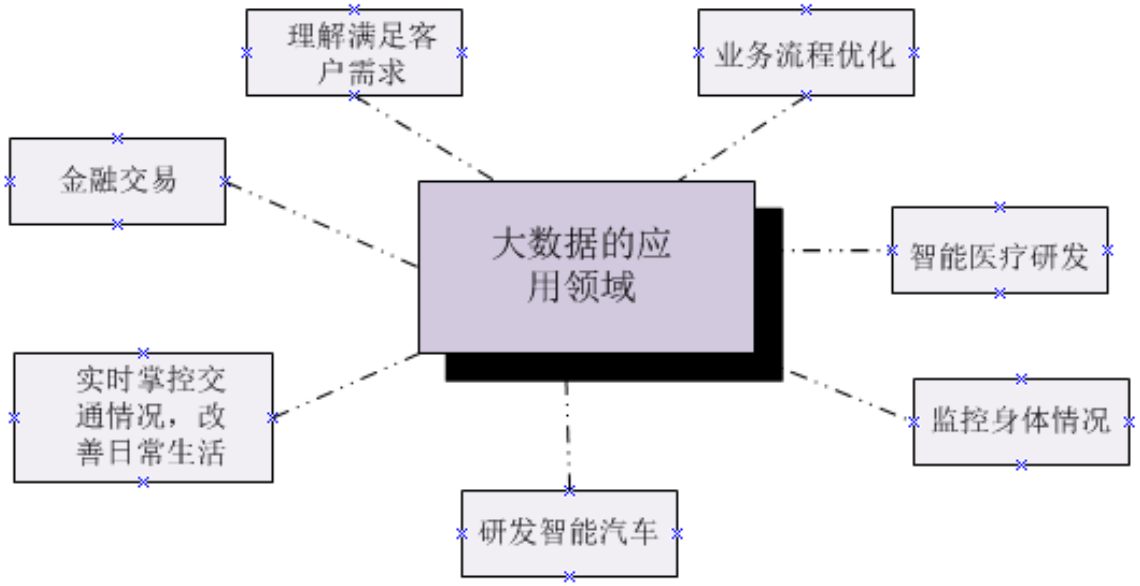


高校大数据课程
公 共 服 务 平 台

百度搜索厦门大学数据库实验室网站访问平台

4.大数据的应用

- 大数据无处不在，包括金融、汽车、零售、餐饮、电信、能源、政务、医疗、体育、娱乐等在内的社会各行各业都已经融入了大数据的印迹



4.大数据的应用

大数据在互联网领域的应用——推荐系统

- 互联网的飞速发展使我们进入了信息过载的时代，搜索引擎可以帮助我们查找内容，但只能解决明确的需求
- 为了让用户从海量信息中高效地获得自己所需的信息，推荐系统应运而生。推荐系统是大数据在互联网领域的典型应用，它可以通过分析用户的历史记录来了解用户的喜好，从而主动为用户推荐其感兴趣的信息，满足用户的个性化推荐需求
- 推荐系统是自动联系用户和物品的一种工具，和搜索引擎相比，推荐系统通过研究用户的兴趣偏好，进行个性化计算。推荐系统可发现用户的兴趣点，帮助用户从海量信息中去发掘自己潜在的需求

4.大数据的应用

- 流行病预测

从谷歌流感趋势看大数据的应用价值

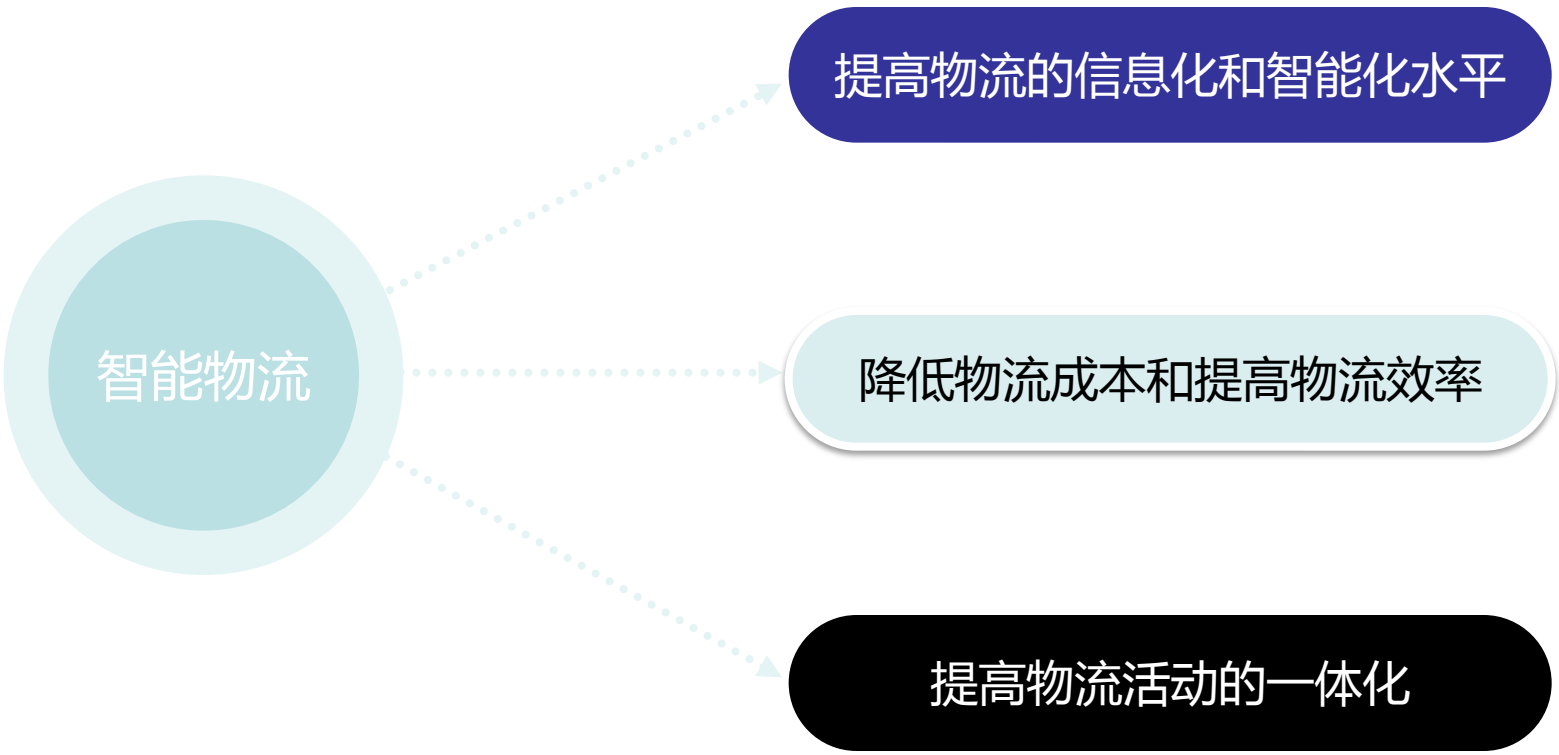
“谷歌流感趋势”，通过跟踪搜索词相关数据来判断全美地区的流感情况

4.大数据的应用

- **生物信息学**
- 生物大数据使得我们可以利用先进的数据科学知识，更加深入地了解生物学过程、作物表型、疾病致病基因等
- 用大数据分析技术，可以从个人健康档案中有效预测个人健康趋势，并为其提供疾病预防建议，达到“治未病”的目的

4.大数据的应用

大数据在物流领域的应用——智能物流



4.大数据的应用

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）”

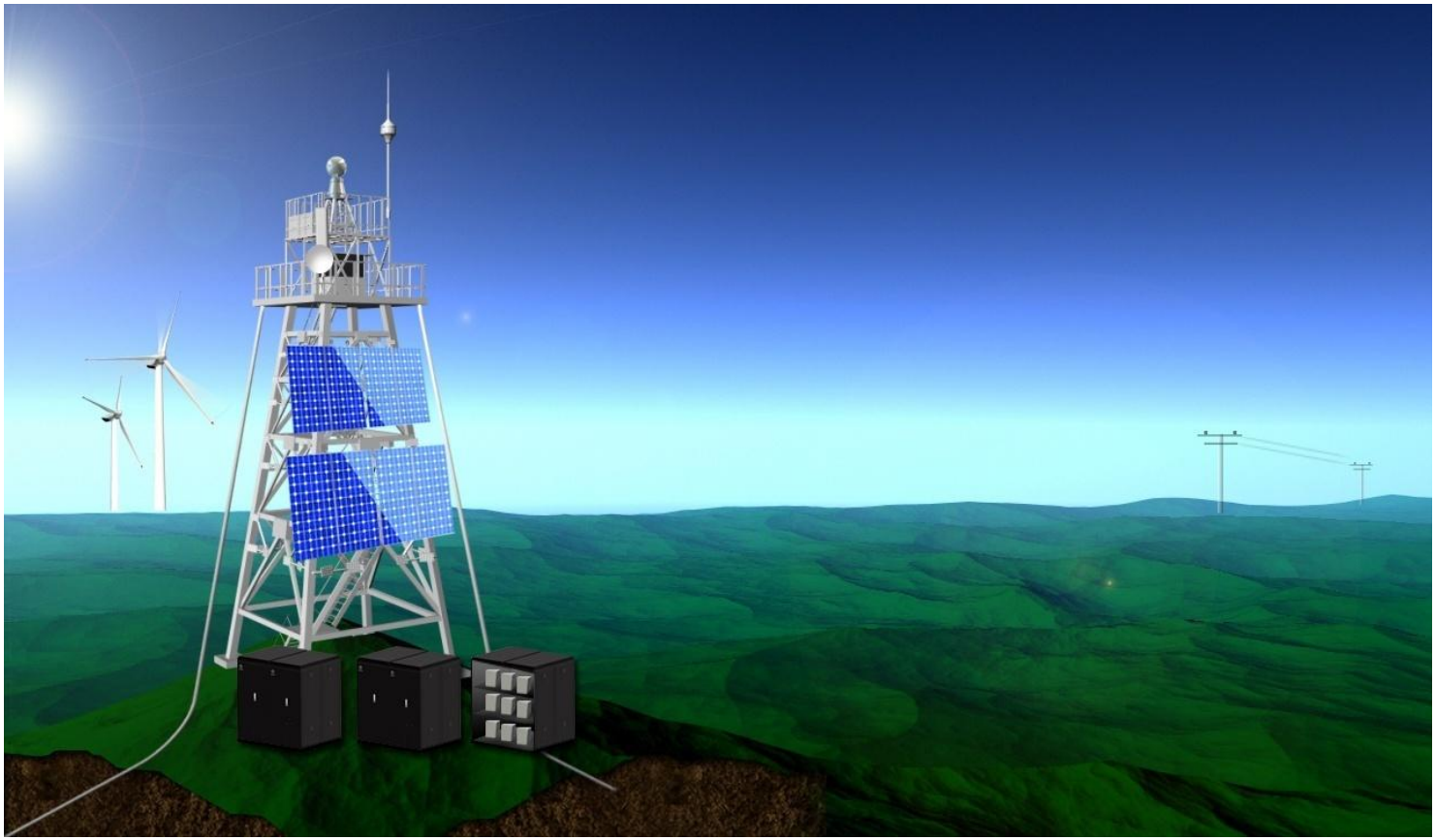
4.大数据的应用

- 智能交通

智能交通将先进的信息技术、数据通信传输技术、电子传感技术、控制技术以及计算机技术等，有效集成并运用于整个地面交通管理，同时可以利用城市实时交通信息、社交网络和天气数据来优化最新的交通情况。

4.大数据的应用

环保监测——森林监视



4.大数据的应用

- **安防领域**
- 中国的很多城市都在开展平安城市建设，在城市的各个角落密布成千上万个摄像头，7×24小时不间断采集各个位置的视频监控数据，数据量之大，超乎想象
- 除了视频监控数据，安防领域还包含大量其他类型的数据，包括结构化、半结构化和非结构化数据

4.大数据的应用

大数据在金融领域的应用

- 高频交易
- 市场情绪分析
- 信贷风险分析
- 大数据征信

4.大数据的应用

•高频交易（High-Frequency Trading，HFT）是指从那些人们无法利用的极为短暂的市场变化中寻求获利的计算机化交易，比如，某种证券买入价和卖出价差价的微小变化，或者某只股票在不同交易所之间的微小价差

•为了从高频交易中获得更高的利润，一些金融机构开始引入大数据技术来决定交易

4.大数据的应用

市场情绪分析是交易者在日常交易工作中不可或缺的一环，根据市场情绪分析、技术分析和基本面分析，可以帮助交易者做出更好的决策。大数据技术在市场情绪分析中大有用武之地。

4.大数据的应用

大数据分析技术已经能够为企业信贷风险分析助一臂之力。通过收集和分析大量中小微企业用户日常交易行为的数据，判断其业务范畴、经营状况、信用状况、用户定位、资金需求和行业发展趋势，解决由于其财务制度的不健全而无法真正了解其真实经营状况的难题，让金融机构放贷有信心、管理有保障

4.大数据的应用

大数据征信就是利用信息技术优势，将不同信贷机构、消费场景、支离破碎的海量数据整合起来，经过数据清洗、模型分析、校验等一系列流程后，加工融合成真正有用的信息。

4.大数据的应用

大数据在汽车领域的应用

- 为了实现无人驾驶的功能，谷歌无人驾驶汽车上配备了大量传感器，包括雷达、车道保持系统、激光测距系统、红外摄像头、立体视觉、GPS导航系统、车轮角度编码器等，这些传感器每秒产生1GB数据，每年产生的数据量将达到约2PB
- 大数据分析技术将帮助无人驾驶系统做出更加智能的驾驶动作决策，比人类驾车更加安全、舒适、节能、环保

图 谷歌无人驾驶汽车

4.大数据的应用

- 发现关联购买行为

啤酒与尿布的故事

4.大数据的应用

•客户群体细分

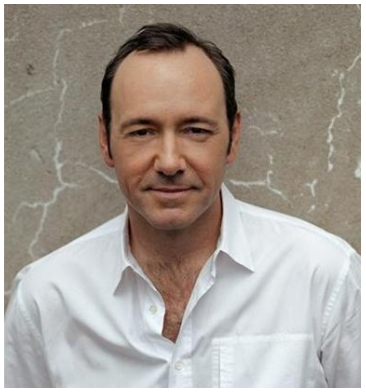
美国Target超市比孩子父亲还早发现他女儿已经怀孕

4.大数据的应用

- **大数据在能源领域的应用**

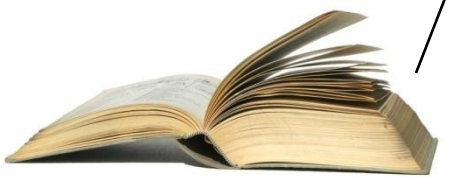
智能电网的发展，离不开大数据技术的发展和应用，大数据技术是组成整个智能电网的技术基石

4.大数据的应用



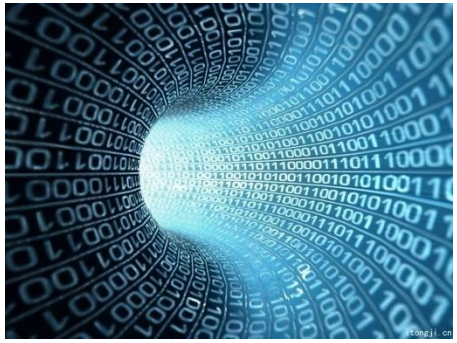
Kevin Spacey

David Fincher



英国同名小说《纸牌屋》

大数据指导影视投拍

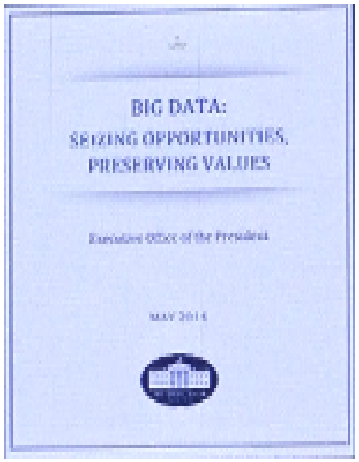


大数据分析

风靡全球的美剧《纸牌屋》

4.大数据的应用

- 大数据与国家安全



美国政府2014年5月发布的大数据报告：大数据可以极大增强国家安全保证能力

美国前国防部长**拉姆斯菲尔德**多次强调：

一枚导弹没有一条情报

能更有效地应对恐怖活动

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。

Department of Computer Science, Xiamen University, 2019