厦门大学计算机科学系研究生课程

《大数据技术基础》

第2章 大数据关键技术与挑战
（2013年新版）

林子雨

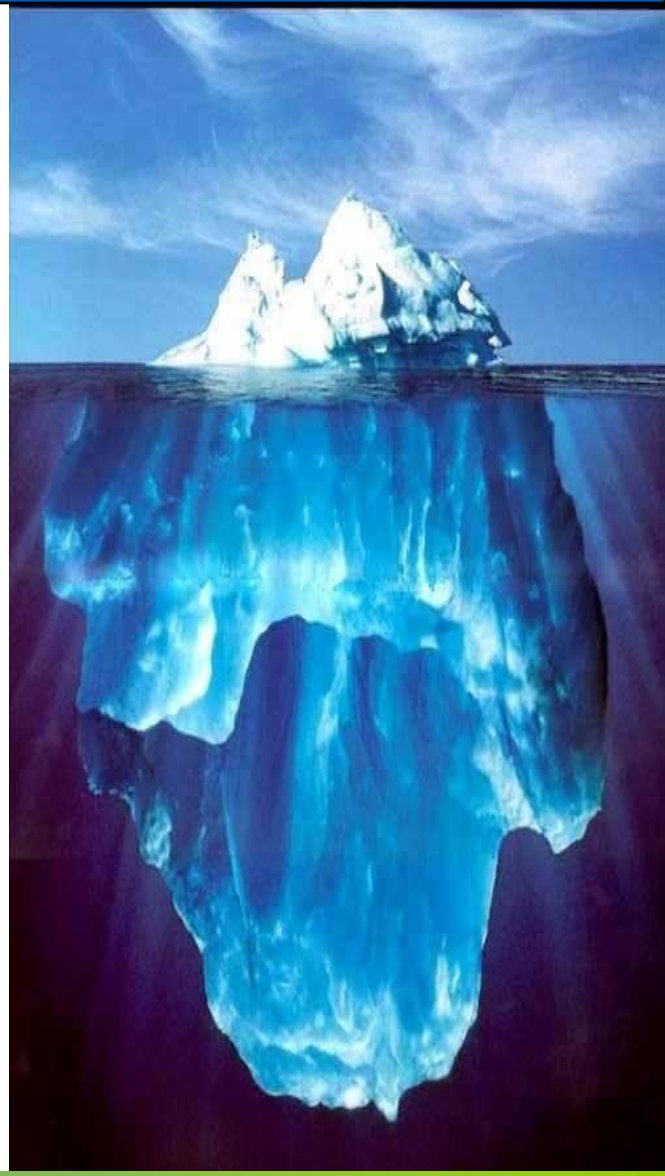
厦门大学计算机科学系

E-mail: ziyulin@xmu.edu.cn

主页：http://www.cs.xmu.edu.cn/linziyu

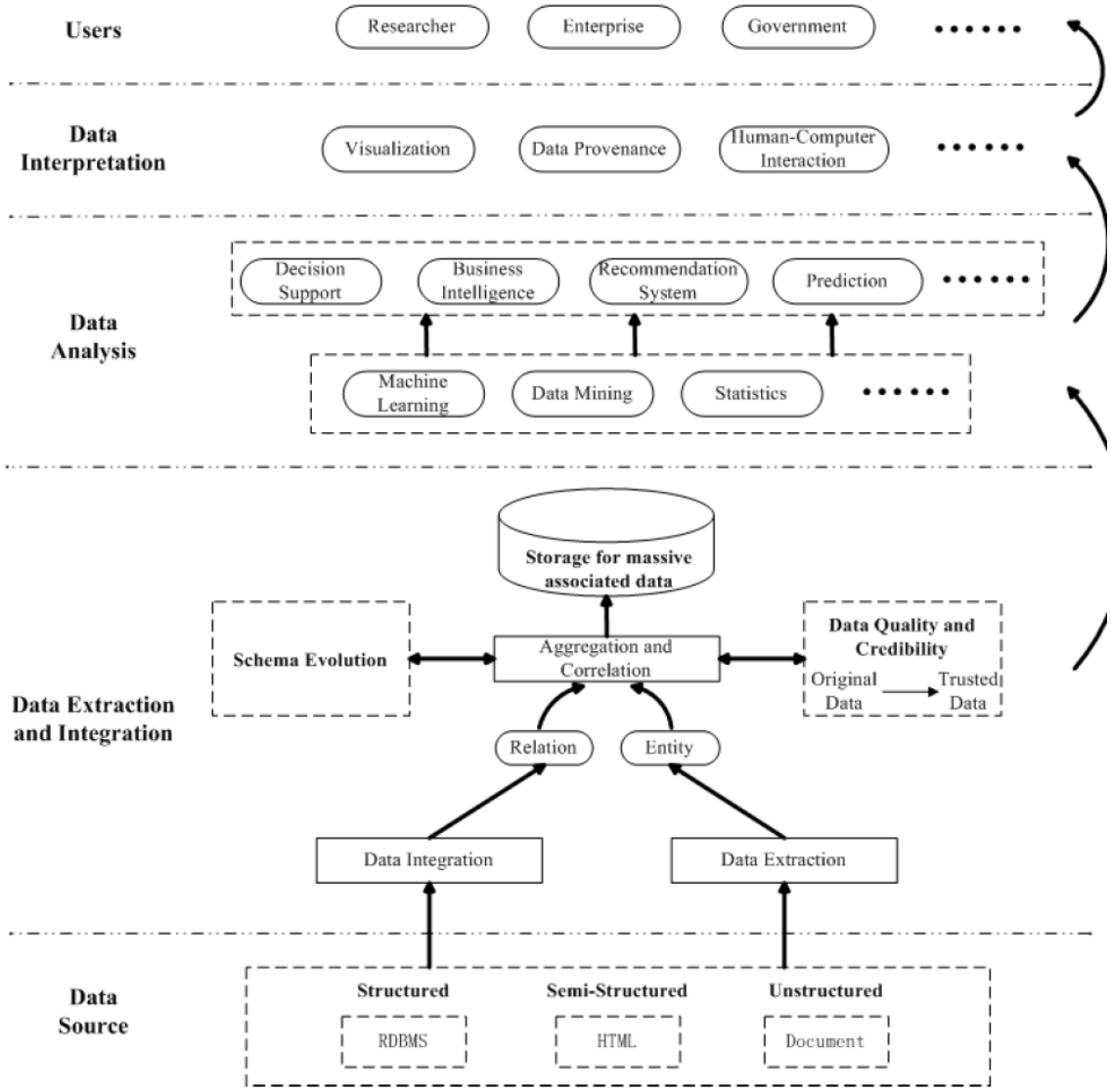
提纲

- 大数据处理的基本流程
- 大数据处理模型
- 大数据关键技术
- 大数据处理工具
- 大数据时代面临的新挑战



本讲义PPT存在配套教材，由林子雨通过大量阅读、收集、整理各种资料后编写而成
下载配套教材请访问《大数据技术基础》2013班级网站：http://dblab.xmu.edu.cn/node/423

大数据处理的基本流程



整个大数据的处理流程可以定义为：在合适工具的辅助下，对广泛异构的数据源进行抽取和集成，结果按照一定的标准进行统一存储，并利用合适的数据分析技术对存储的数据进行分析，从中提取有益的知识并利用恰当的方式将结果展现给终端用户。具体来说，可以分为数据抽取与集成、数据分析以及数据解释。

数据抽取与集成

- 大数据的一个重要特点就是多样性，这就意味着数据来源极其广泛，数据类型极为繁杂。这种复杂的数据环境给大数据的处理带来极大的挑战。
- 要想处理大数据，首先必须对所需数据源的数据进行抽取和集成，从中提取出关系和实体，经过关联和聚合之后采用统一定义的结构来存储这些数据。
- 在数据集成和提取时需要对数据进行清洗，保证数据质量及可信性。
- 现有的数据抽取与集成方式可以大致分为以下四种类型：数据整合、数据联邦、数据传播和混合方法等。

数据分析

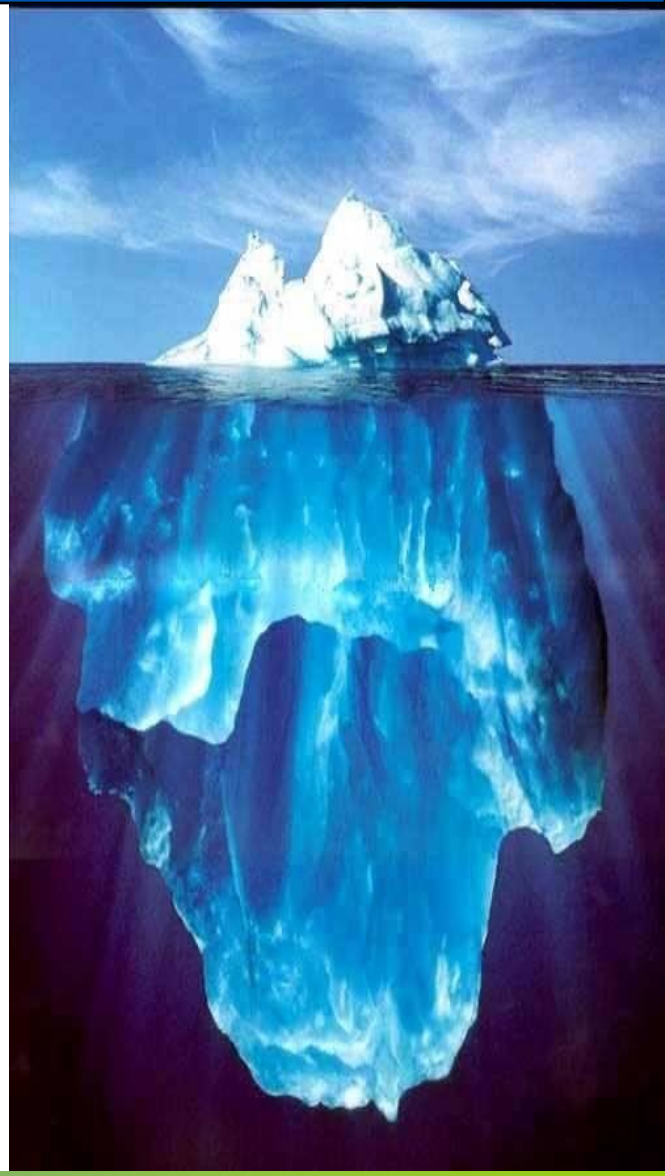
- 传统的分析技术如数据挖掘、机器学习、统计分析等在大数据时代需要做出调整，因为这些技术在大数据时代面临着一些新的挑战，主要有：
  - **数据量大并不一定意味着数据价值的增加，相反这往往意味着数据噪音的增多**
  - **大数据时代的算法需要进行调整**（邦弗朗尼原理）
  - **数据结果好坏的衡量**

数据解释

- 数据分析是大数据处理的核心，但是用户往往更关心结果的展示。如果分析的结果正确但是没有采用适当的解释方法，则所得到的结果很可能让用户难以理解，极端情况下甚至会误导用户。
- 大数据时代的数据分析结果往往也是海量的，同时结果之间的关联关系极其复杂，采用传统的解释方法基本不可行
- 可以考虑从下面两个方面提升数据解释能力：
  - **引入可视化技术**
  - **让用户能够在一定程度上了解和参与具体的分析过程**

提纲

- 大数据处理的基本流程
- 大数据处理模型
- 大数据关键技术
- 大数据处理工具
- 大数据时代面临的新挑战



大数据之“快”从何说起

- **时间就是金钱**

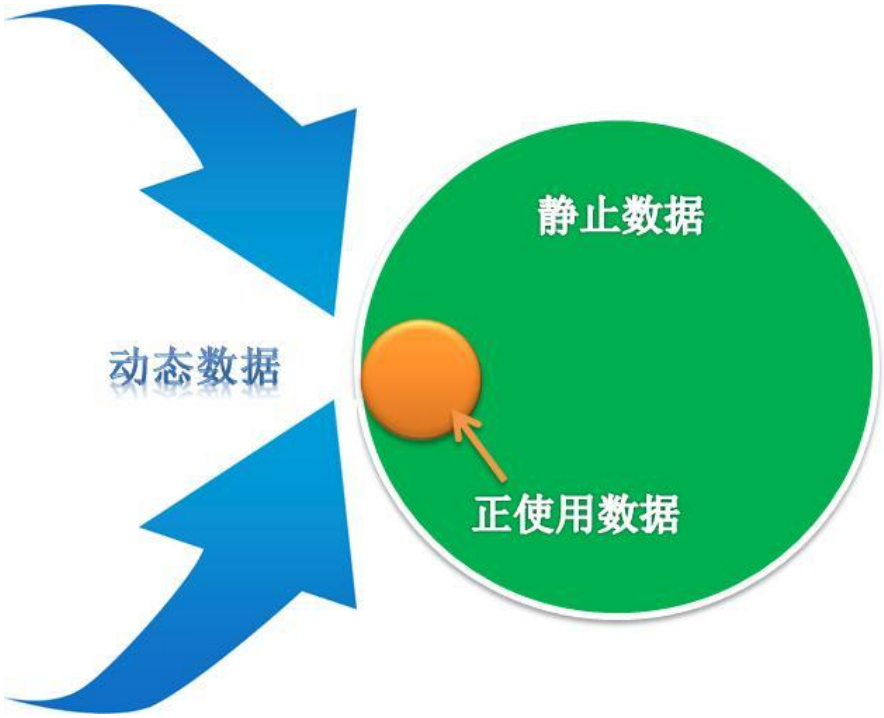
  时间在分母上，越小，单位价值就越大。

- **像其它商品一样，数据的价值会折旧**

  过去一天的数据，比过去一个月的数据可能都更有价值。

- **数据跟新闻和金融行情一样，具有时效性**

大数据的三种状态



大数据的三种状态如上图所示，按照数据的三状态定义，水库里一平如镜（非活跃）的水是“静止数据（data at rest）”，水处理系统中上下翻动的水是“正使用数据（data inuse）”，汹涌而来的新水流就是“动态数据（data in motion）”。

大数据的“快”说的是两个层面

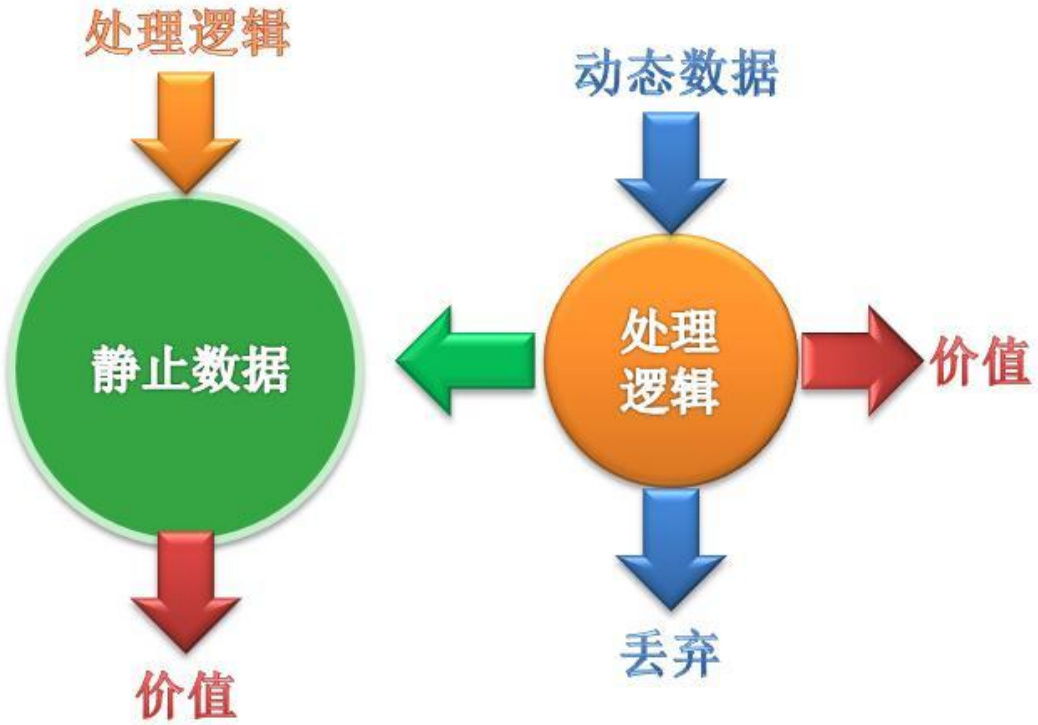
- **“动态数据”来得快**

动态数据有不同的产生模式。有的是burst模式，极端的例子如欧洲核子研究中心（CERN）的大型强子对撞机(Large Hadron Collider，简称LHC)，此机不撞则已，一撞惊人，工作状态下每秒产生PB级的数据。也有的动态数据是涓涓细流的模式，典型的如 clickstream，日志，RFID数据，GPS位置信息，Twitter的firehose流数据等。

- **“正使用数据”处理得快**

水处理系统可以从水库调出水来进行处理（“静止数据”转变为“正使用数据”），也可以直接对涌进来的新水流处理（“动态数据”转变为“正使用数据”）。这对应着两种大相迥异的处理范式：批处理和流处理。

批处理与流处理



左半部是批处理：以“静止数据”为出发点，数据是任尔东西南北风、我自岿然不动，处理逻辑进来，算完后价值出去。**右半部**则是流数据处理范式。这次不动的是逻辑，“动态数据”进来，计算完后价值留下，原始数据加入“静止数据”，或索性丢弃。

批处理与流处理的组合

两种范式常常组合使用，而且形成了一些定式：

- **流处理作为批处理的前端**：比如大型强子对撞机，每秒PB级的数据先经过流处理范式进行过滤，只有那些科学家感兴趣的撞击数据保留下来进入存储系统，留待批处理范式处理。这样，欧洲核子研究中心每年的新增存储存储量可以减到25PB。
- **流处理与批处理肩并肩**：流处理负责动态数据和实时智能，批处理负责静止数据和历史智能，实时智能和历史智能合并成为全时智能。

如何实现“快”的数据处理

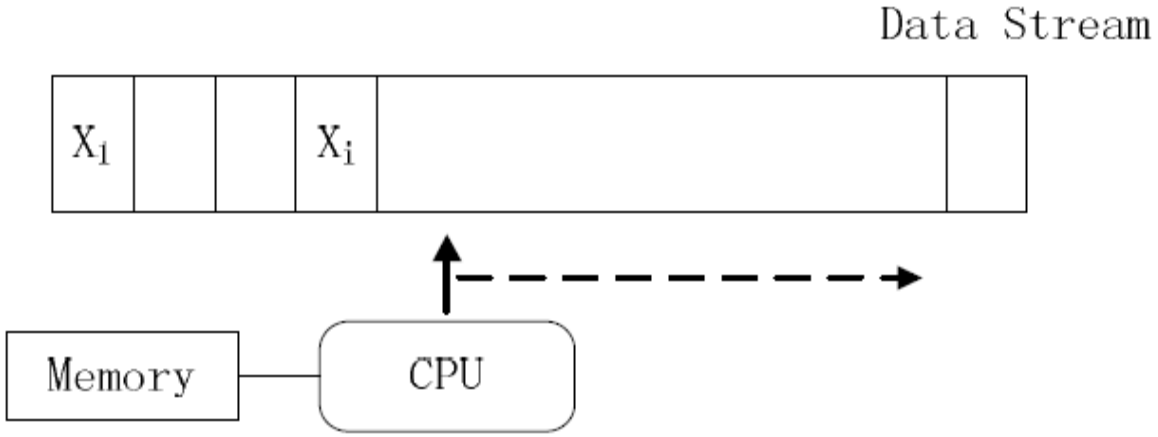
首先，“快”是个相对的概念，可以是实时，也可以秒级、分钟级、小时级、天级甚至更长的延迟。其次，考虑目前的架构是不是有潜力改造到足够“快”。

一些通用的技术思路来实现“快”：

- 如果数据流入量太大，在前端就地采用流处理进行即时处理、过滤掉非重要数据
- 把数据预处理成适于快速分析的格式
- 增量计算--也即先顾眼前的新数据，再去更新老数据
- 很多批处理系统慢的根源是磁盘和I/O，把原始数据和中间数据放在内存里，一定能极大地提升速度
- 降低对精确性的要求

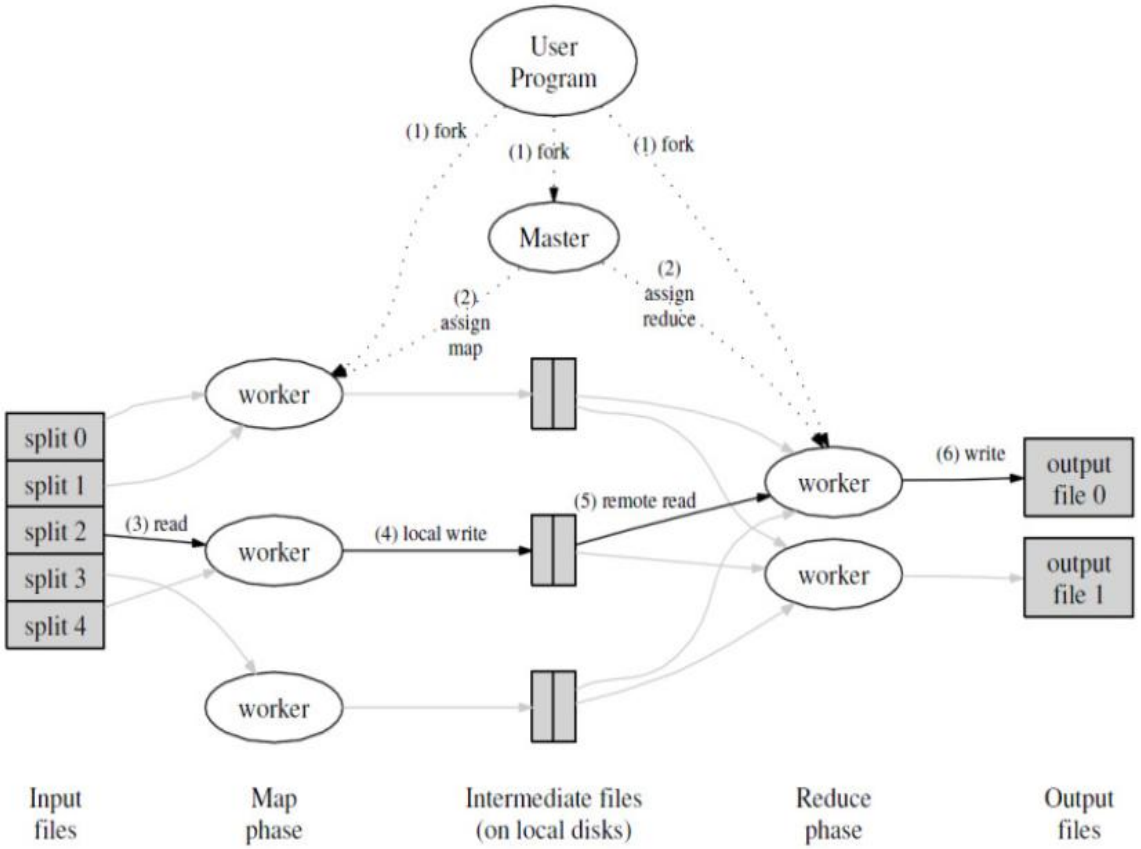
流处理

流处理的处理模式将数据视为流，源源不断的数据组成了数据流。当新的数据到来时就立刻处理并返回所需的结果。



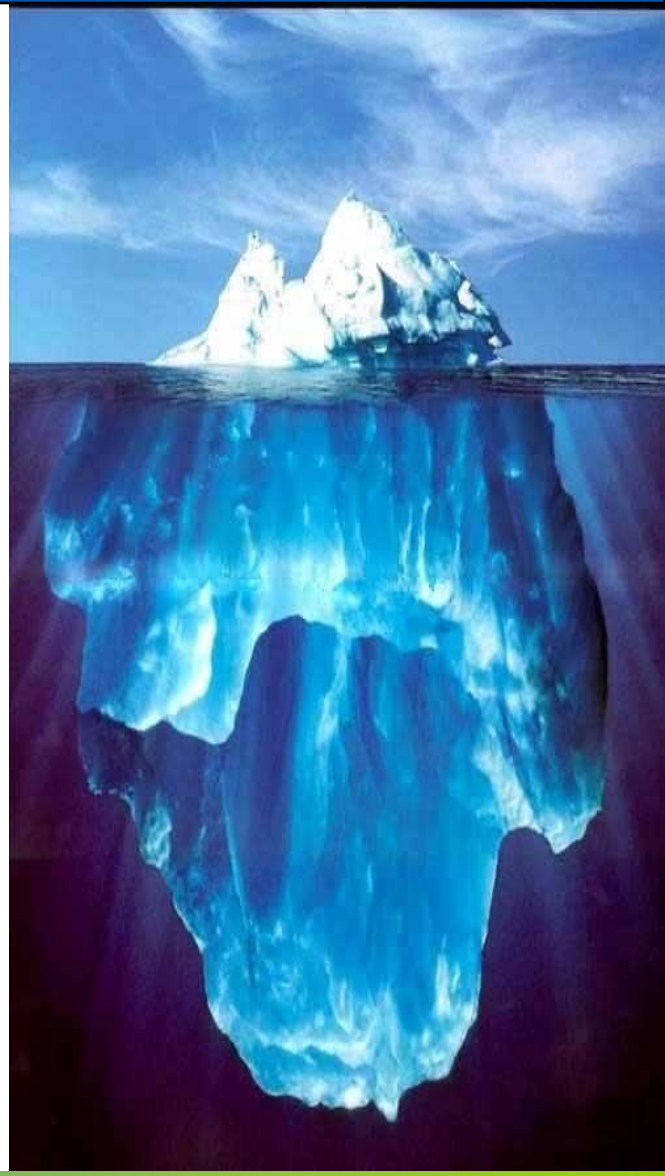
批处理理

Google 公司在2004年提出的MapReduce编程模型是最具代表性的批处理模式。一个完整的MapReduce 过程如图所示：



提纲

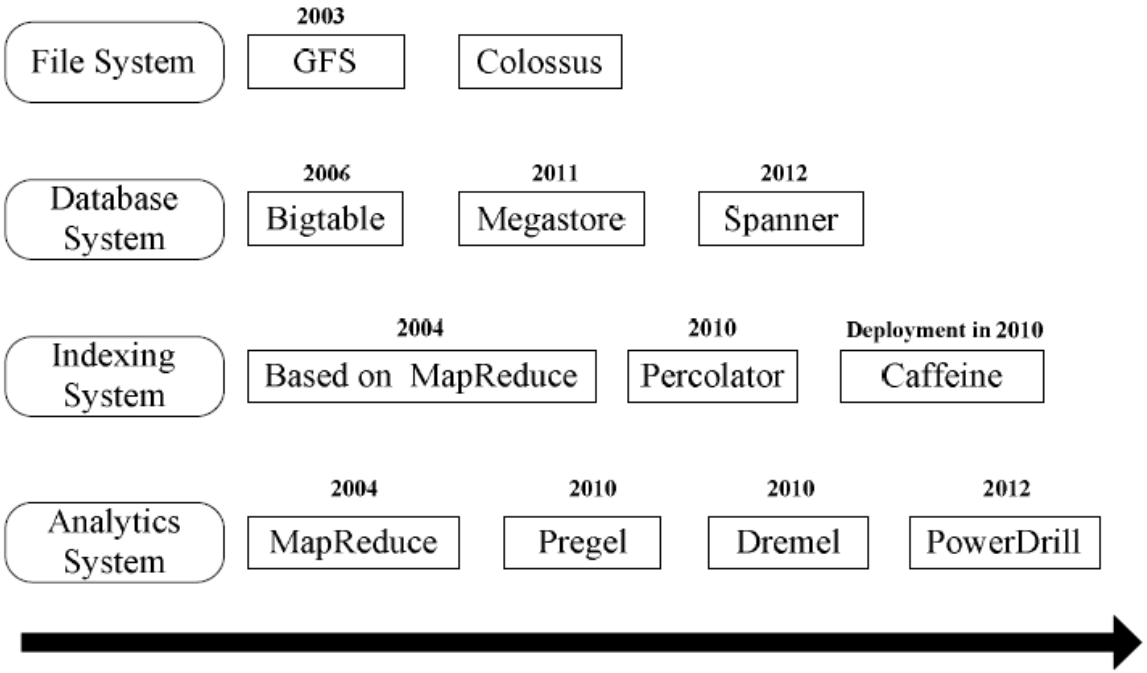
- 大数据处理的基本流程
- 大数据处理模型
- 大数据关键技术
- 大数据处理工具
- 大数据时代面临的新挑战



大数据关键技术

Google 于2006 年首先提出了云计算的概念，并研发了一系列云计算技术和工具。难能可贵的是Google 并未将这些技术完全封闭，而是以论文的形式逐步公开其实现。正是这些公开的论文，使得以GFS、MapReduce、Bigtable 为代表的一系列大数据处理技术被广泛了解并得到应用，同时还催生出以Hadoop为代表的一系列云计算开源工具。

下图展示了Google的技术演化过程：



文件系统

包括Google、微软、Facebook和淘宝在内的众多企业和学者从不同方面对满足大数据存储需求的文件系统进行了详尽的研究。并自行开发出支持其自身业务的文件系统：

- GFS
- Colosuss
- HDFS
- CloudStore
- Haystack
- TFS
- FastDFS

数据库系统

直接采用关系模型的分布式数据库并不能适应大数据时代的数据存储，主要因为：

1. 规模效应所带来的压力
2. 数据类型的多样化
3. 设计理念的冲突
4. 数据库事务特性

面对这些挑战，以Google 为代表的一批技术公司纷纷推出了自己的解决方案：

1. Google的Bigtable
2. Amazon的Dynamo
3. Yahoo的PNUTS

NoSQL技术

Bigtable、Dynamo、PNUTS等的成功促使人们开始对关系数据库进行反思，由此产生了一批未采用关系模型的数据库，这些方案现在被统一的称为NoSQL(Not Only SQL)。NoSQL 并没有一个准确的定义，但一般认为NoSQL 数据库应当具有以下的特征：

1. 模式自由(schema-free)
2. 支持简易备份(easy replication support)
3. 简单的应用程序接口(simple API)
4. 最终一致性(或者说支持BASE特性，不支持ACID)
5. 支持海量数据(Huge amount of data)。

索引和查询技术

不太可能将已有的成熟索引方案直接应用于大数据。

表　一些索引方案直接应用在Facebook上的性能估计

Algorithms	Index Size for Facebook	Index Time for Facebook	Query Time on Facebook(s)
Ullmann[Ullmann 76]	-	-	>1000
VF2[CordellaFSV04]	-	-	>1000
RDF-3X[NeumannW10]	1T	>20 days	>48
BitMat[AtreCZH10]	2.4T	>20days	>269
Subdue[HolderCD94]	-	>67 years	-
SpiderMine[ZhuQLYHY11]	-	>3 years	-
R-Join[ChengYDYW08]	>175T	$>10^{15}$ years	>200
Distance-Join[ZouCO09]	>175T	$>10^{15}$ years	>4000
GraphQL[HeS08]	>13T(r=2)	>600 years	>2000
Zhao[ZhaoH10]	>12T(r=2)	>600 years	>600
GADDI[ZhangLY09]	$>2*10^5$T(L=4)	$>4*10^5$ years	>400

索引和查询技术

NoSQL 数据库针对主键的查询效率一般较高，因此有关的研究集中在NoSQL数据库的多值查询优化上。针对NoSQL数据库上的查询优化研究主要有两种思路：

1.采用MapReduce并行技术优化多值查询：当利用MapReduce并行查询NoSQL数据库时，每个MapTask处理一部分的查询操作，通过实现多个部分之间的并行查询来提高多值查询的效率。此时每个部分的内部仍旧需要进行数据的全扫描。

2.采用索引技术优化多值查询：很多的研究工作尝试从添加多维索引的角度来加速NoSQL 数据库的查询速度。

数据分析技术

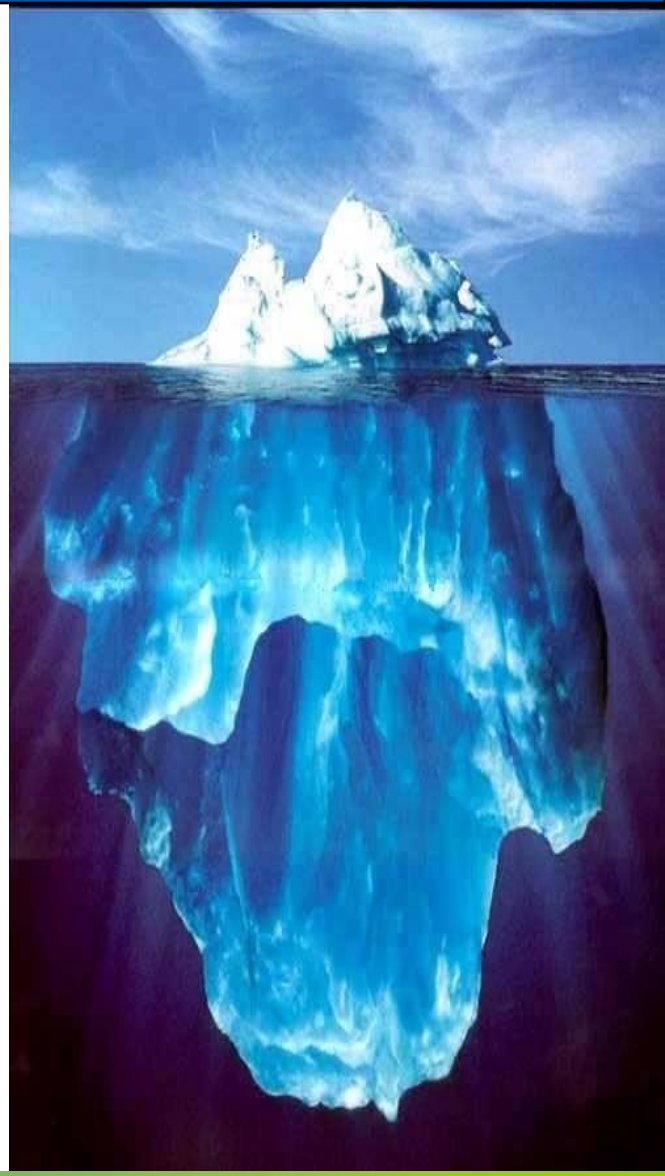
•Mapreduce是谷歌最早采用的计算模型，适合批处理
•谷歌设计了Pregrel用于图计算
• Dremel适用于Web数据级别的交互式数据分析系统
•谷歌的PowerDrill主要用于大数据量的核心数据集分析

实时数据处理是大数据分析的一个核心需求。很多研究工作正是围绕这一需求展开的。前面介绍了大数据处理的两种基本模式，而在实时处理的模式选择中，主要有三种思路：

- 采用流处理模式：Storm
- 采用批处理模式：Percolator\Nectar\DryadInc实现大规模数据的增量计算
- 二者的融合

提纲

- 大数据处理的基本流程
- 大数据处理模型
- 大数据关键技术
- 大数据处理工具
- 大数据时代面临的新挑战



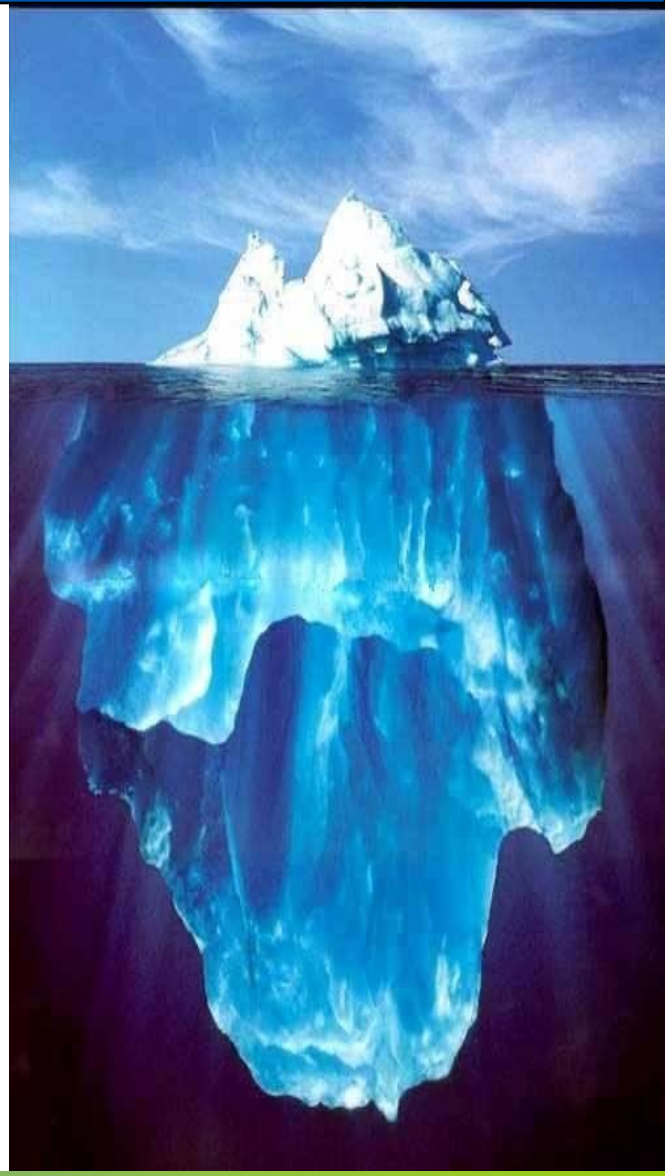
大数据处理工具

Hadoop 是目前最为流行的大数据处理平台。除了Hadoop，还有很多针对大数据的处理工具。这些工具有些是完整的处理平台，有些则是专门针对特定的大数据处理应用。下表归纳总结了现今一些主流的处理平台和工具。（了解专业术语很重要）

Category		Examples
Platform	Local	Hadoop、MapR、Cloudera、Hortonworks、InfoSphere BigInsights、ASTERIX
	Cloud	AWS、Google Compute Engine、Azure
Database	SQL	Greenplum、Aster Data、Vertica
	NoSQL	HBase、Cassandra、MongoDB、Redis
	NewSQL	Spanner、Megastore、F1
Data Warehouse		Hive、HadoopDB、Hadapt
Data Processing	Batch	MapReduce、Dryad
	Stream	Storm、S4、Kafka
Query Language		HiveQL、Pig Latin、DryadLINQ、MRQL、SCOPE
Statistic and Machine Learning		Mahout、Weka、R
Log Processing		Splunk、Loggly

提纲

- 大数据处理的基本流程
- 大数据处理模型
- 大数据关键技术
- 大数据处理工具
- 大数据时代面临的新挑战



大数据集成

数据的广泛存在性使得数据越来越多的散布于不同的数据管理系统中，为了便于进行数据分析需要进行数据的集成。数据集成看起来并不是一个新的问题，但是大数据时代的数据集成却有了新的需求，因此也面临着新的挑战：

- 广泛的异构性
  - 数据类型从以结构化数据为主转向结构化、半结构化、非结构化三者的融合
  - 数据产生方式的多样性带来的数据源变化：移动设备产生的数据时空特性
  - 数据存储方式的变化
- 数据质量
  - 数据量大不一定就代表信息量或者数据价值的增大，相反很多时候意味着信息垃圾的泛滥

大数据分析

随着大数据时代的到来，半结构化和非结构化数据量的迅猛增长，给传统的分析技术带来了巨大的冲击和挑战，主要体现在：

- 数据处理的实时性(Timeliness)
- 动态变化环境中索引的设计
- 先验知识的缺乏

大数据隐私问题

隐私问题由来已久，计算机的出现使得越来越多的数据以数字化的形式存储在电脑中，互联网的发展则使数据更加容易产生和传播，数据隐私问题越来越严重。

- 隐性的数据暴露：面临技术和人力（众包）的双重考验
- 数据公开与隐私保护的矛盾:隐私保护数据挖掘
- 数据动态性：现有隐私保护技术基于静态数据

大数据能耗问题

在能源价格上涨、数据中心存储规模不断扩大的今天，高能耗已逐渐成为制约大数据快速发展的一个主要瓶颈。

从已有的一些研究成果来看，可以考虑以下两个方面来改善大数据能耗问题：

- 采用新型低功耗硬件
- 引入可再生的新能源

大数据处理与硬件的协同

硬件的快速升级换代有力的促进了大数据的发展，但是这也在一定程度上造成了大量不同架构硬件共存的局面。日益复杂的硬件环境给大数据管理带来的主要挑战有：

- 硬件异构性带来的大数据处理难题
  - 木桶效应：Mapreduce处理时间取决于处理时间最长的节点
- 新硬件给大数据处理带来的变革

大数据管理易用性问题

从数据集成到数据分析，直到最后的数据解释，易用性应当贯穿整个大数据的流程。易用性的挑战突出体现在两个方面：

- 首先大数据时代的数据量大，分析更复杂，得到的结果形式更加的多样化。其复杂程度已经远远超出传统的关系数据库。
- 其次大数据已经广泛渗透到人们生活的各个方面，很多行业都开始有了大数据分析的需求。

要想达到易用性，需要关注以下三个基本原则：

- 可视化原则(Visibility)
- 匹配原则(Mapping)
- 反馈原则(Feedback)

性能的测试基准

目前尚未有针对大数据管理的测试基准，构建大数据测试基准面临的主要挑战有：

- 系统复杂度高
- 用户案例的多样性
- 数据规模庞大
- 系统的快速演变
- 重新构建还是复用现有的测试基准

本章小结

本章内容首先介绍了大数据处理的基本流程和大数据处理模型，接着介绍了大数据的关键技术，其中，云计算是大数据的基础平台和支撑技术，本章以Google 的相关技术为主线，详细介绍Google 以及其他众多学者和研究机构在大数据技术方面已有的一些工作，包括文件系统、数据库系统、索引和查询技术、数据分析技术等；接下来，介绍了大数据处理平台和工具，就目前技术发展现状而言，Hadoop 已经成为了大数据处理工具事实上的标准。最后，介绍大数据时代面临的新挑战，包括大数据集成、大数据分析、大数据隐私问题、大数据能耗问题、大数据处理与硬件的协同、大数据管理易用性问题以及性能测试基准。

参考文献

- [1] 孟小峰, 慈祥. 大数据管理：概念、技术与挑战. 计算机学报, 2013年第8期.

主讲教师和助教

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

助教：赖明星

单位：厦门大学计算机科学系数据库实验室2011级硕士研究生（导师：林子雨）
E-mail: mingxinglai@gmail.com
个人主页：http://mingxinglai.com

欢迎访问《大数据技术基础》2013班级网站：http://dblab.xmu.edu.cn/node/423
本讲义PPT存在配套教材《大数据技术基础》，请到上面网站下载。

Department of Computer Science, Xiamen University, Sep, 2013