Database & Big Data

Big Data

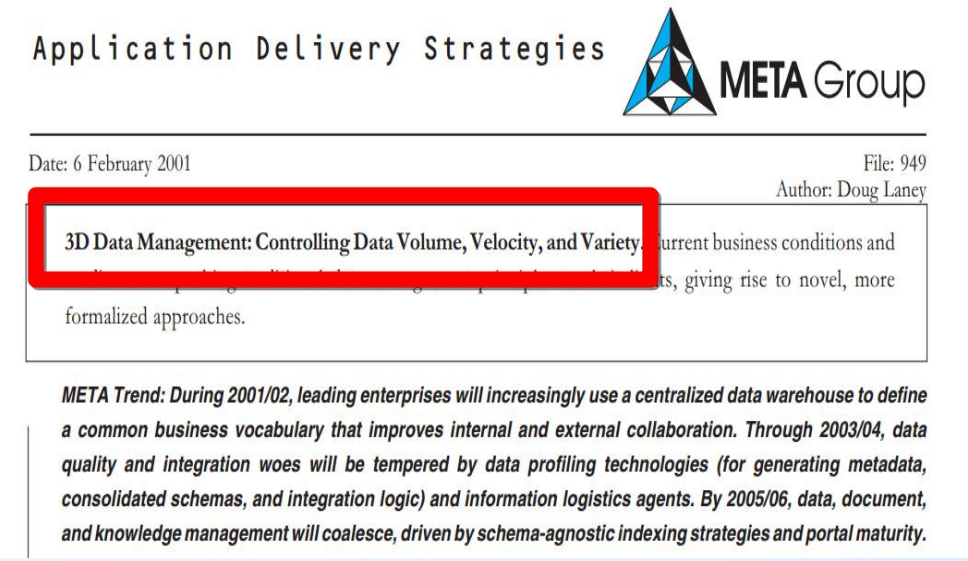
VS

Database

大数据定义

- 维基百科给出的定义
  - 大数据是指利用常用软件工具捕获、管理和处理数据所耗时间超过可容忍时间的数据集。
- 麦肯锡定义
  - 大小超过了常规数据库工具获取、存储、管理和分析能力的数据集。

（一）大数据概念溯源



Application Delivery Strategies

META Group

Date: 6 February 2001

File: 949

Author: Doug Laney

3D Data Management: Controlling Data Volume, Velocity, and Variety

urrent business conditions and

ts, giving rise to novel, more formalized approaches.

META Trend: During 2001/02, leading enterprises will increasingly use a centralized data warehouse to define a common business vocabulary that improves internal and external collaboration. Through 2003/04, data quality and integration woes will be tempered by data profiling technologies (for generating metadata, consolidated schemas, and integration logic) and information logistics agents. By 2005/06, data, document, and knowledge management will coalesce, driven by schema-agnostic indexing strategies and portal maturity.

http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

→ Gartner

2001年，Meta Group（Gartner的前身）指出，随着网络及其应用（电子商务等）发展，数据将呈现出爆炸式增长的趋势，并提出3D Data Management的技术预测，即 Data Volume、Data Velocity 和 Data Variety

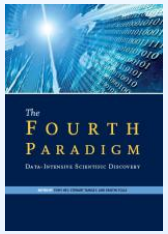
2012年，Gartner的IT技术发展趋势战略报告指出：大数据正在逼近“Tipping Point”（爆发点），40%以上的企业开始大数据方面的投资。

*

（一）大数据概念溯源

2012年4月，欧洲信息学与数学协会会刊ERCIM News出版专刊“Big Data”讨论了数据管理、数据密集型研究等问题

2011年2月11日：Science刊登了一个名为“Dealing with Data”的专辑，联合Science: Signaling、Science: Translational Medicine和Science Careers推出相关专题，讨论数据对科学研究的重要性



2009年10月微软为纪念Jim Gray, 出版了“第四范式—数据密集的科学发现”，认为科学研究范式的发展 “理论科学 →实验科学→计算机仿真→数据科学”

2009年7月O’Reilly Media出版了名为“Beautiful Data”，介绍大数据相关的技术

2008年9月4日《自然》（Nature）刊登了一个名为“Big Data”的专辑，从互联网技术、网络经济学、生物医药等多个方面探讨了大数据的挑战与机遇

（二）大数据概述

21世纪是数据信息大发展的时代，移动互联、社交网络、电子商务等极大拓展了互联网的边界和应用范围，各种数据正在迅速膨胀并变大。

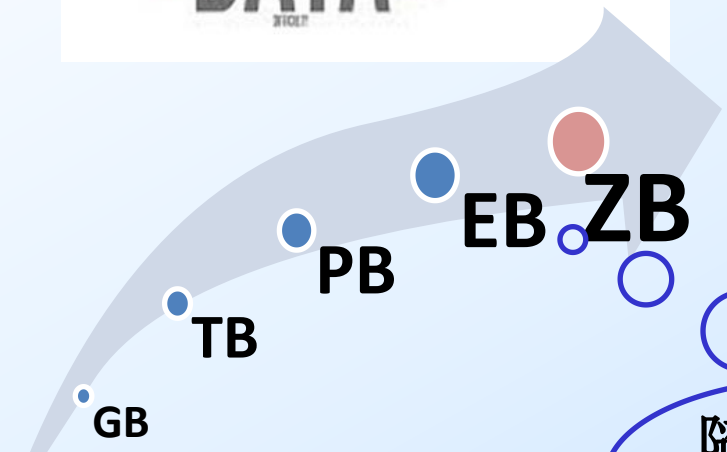
互联网（社交、搜索、电商）、移动互联网（微博）、物联网（传感器，智慧地球）、车联网、GPS、医学影像、安全监控、金融（银行、股市、保险）、电信（通话、短信）都在疯狂产生着数据。

*

（二）大数据概述

地球上至今总共的数据量：

- 在2006 年，个人用户才刚刚迈进TB时代，全球一共新产生了约180EB的数据；
- 在2011 年，这个数字达到了1.8ZB。
- 市场研究机构预测：到2020 年，整个世界的数据总量将会增长44 倍，达到35.2ZB（1ZB=10 亿TB）！



1TB=1024GB
1PB=1024TB
1EB=1024PB
1ZB=1024EB

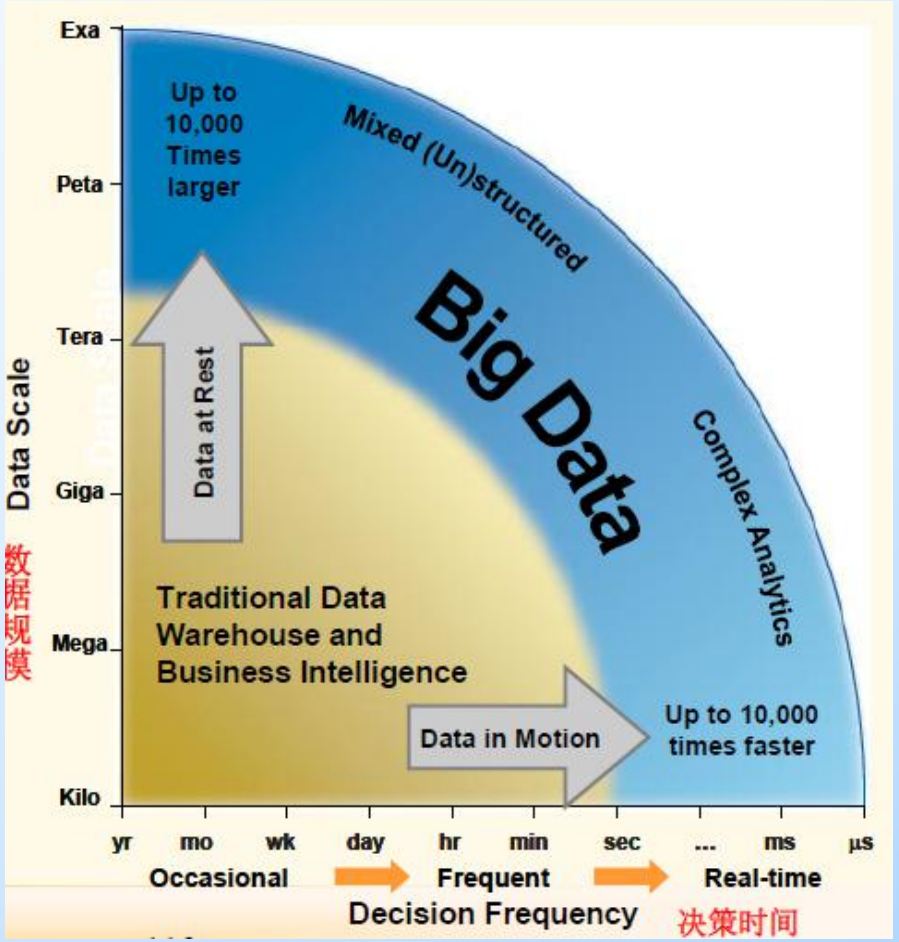
随着信息不断膨胀与爆炸，已经积累到了引发变革的程度。更多的信息增长速度也在指数级加快。

想驾驭这庞大的数据，我们必须了解大数据的特征。

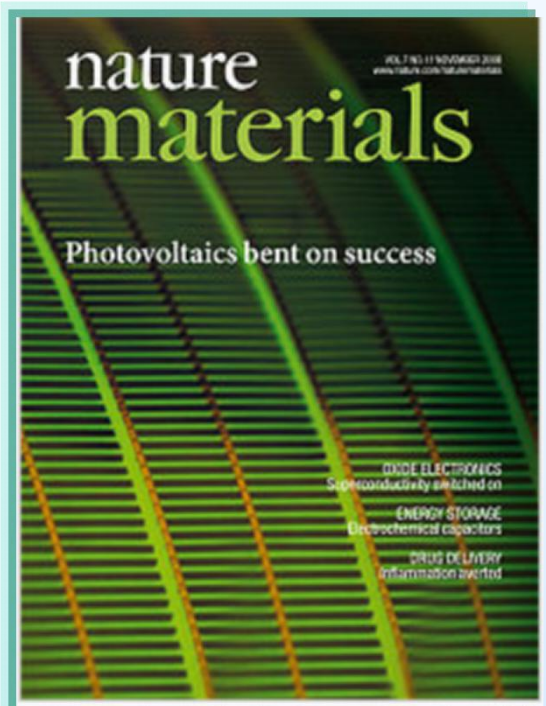
（二）大数据概述

- **大数据无法在容许的时间内，用常规软件工具对其内容进行抓取、管理和处理的数据集合；**
- **大数据规模的标准是持续变化的；**
- **大数据当前泛指单一数据集的大小在几十TB和数PB之间**

--- 维基百科定义



（二）大数据概述



Nature Materials 2008年11月1日的封面

2008年9月美国《自然》杂志刊登了一个名为“Big Data”的专辑，首次提出**大数据（Big Data）概念**。

大数据，或称巨量资料，是指由**数量巨大、结构复杂、类型众多**的数据所构成的数据集合，必须通过特殊化处理分析才能形成有规律、可预测的信息服务能力。

（二）大数据概述

V

数据体量巨大

Volume

V

数据类型繁多

Variety

V

价值密度低

Value

V

处理速度快

Velocity

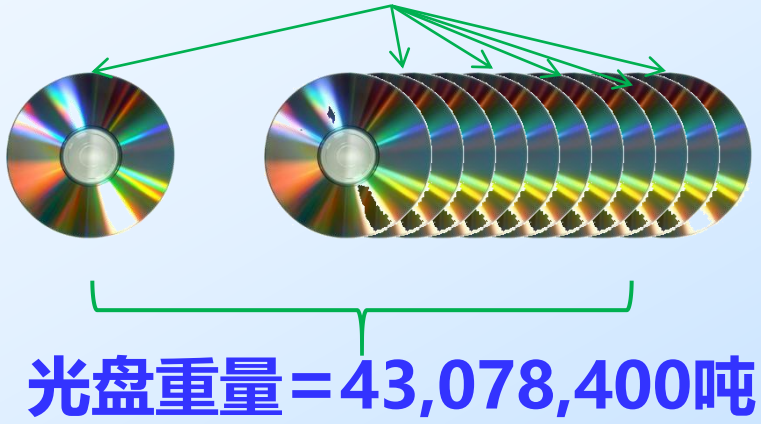
5V

- 1、Volume(大量)
- 2、Velocity(高速)
- 3、Variety(多样)
- 4、Value(低价值密度)
- 5、Veracity(真实性)

- IBM

口数据存储量大、计算量大（Volume）

预计到2020年，中国产生的数据总量将是8.5ZB，全球的数据总量预计将达到40ZB。若以光盘存储，其总重量相当于424艘满载的尼米兹航空母舰。



424艘尼米兹号航母重量

(101,600吨)

□数据来源多、格式多（Variety）

大数据类型分为结构化数据和非结构化数据，结构化数据是指用数据或统一的结构加以表示，如**数字、符号**；非结构化数据是指无法用数字或统一的结构表示，如**文本、图像、声音、网页、地理位置信息**等，这些多类型的数据对数据的处理能力提出了更高要求。

大数据的来源

- 互联网快速发展，社交网络成熟
- 移动终端普及
- 物联网、传感器、监控设备
- 云计算
- ……

主要三类：人和人之间、人和物（机器）之间、物和物之间

数据量骤增，数据来源多样。

主动、被动和自动方式产生数据，数据产生方式的巨大变化

导致大数据的产生。

PC

NoteBook

iPad

iPhone

淘宝

微博

百度

腾讯

国家电网

华大基因

联通

北京公交

RFID

条码

二维码

GPS终端

视频采集设备

移动互联网

物联网

社交网络

电子商务

智慧城市

…

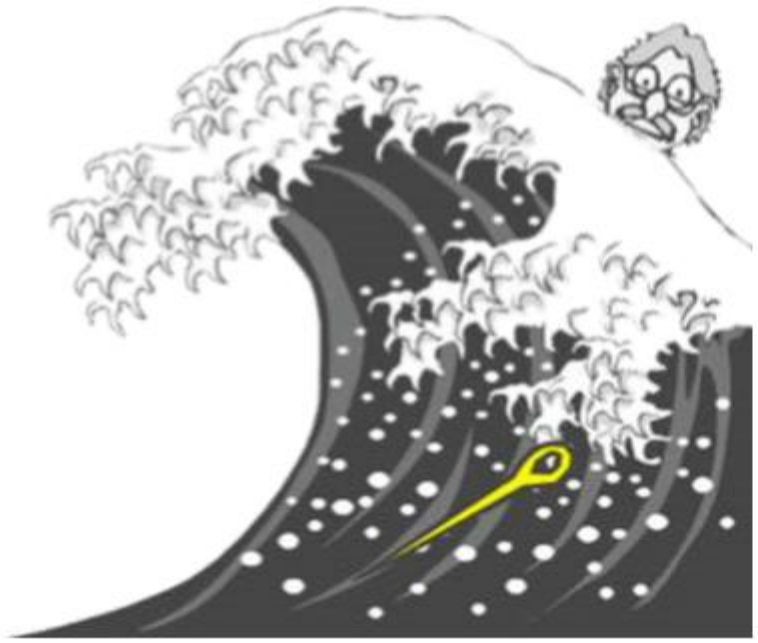
大数据时代已经到来

互联网思维、大数据等信息技术正在改变 ……

- 社会治理
- 企业经营
- 工作方式
- 生活方式
- 行为方式
- 思维方式

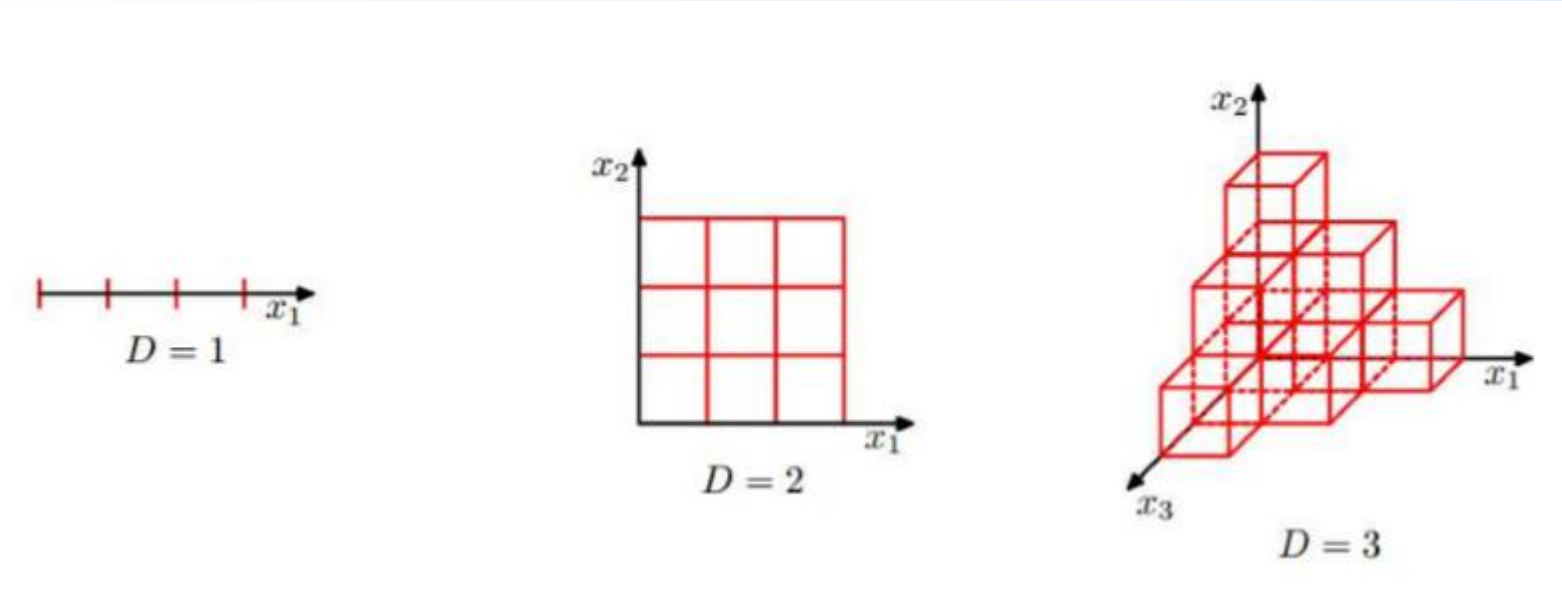
大数据带来的挑战

- 大量

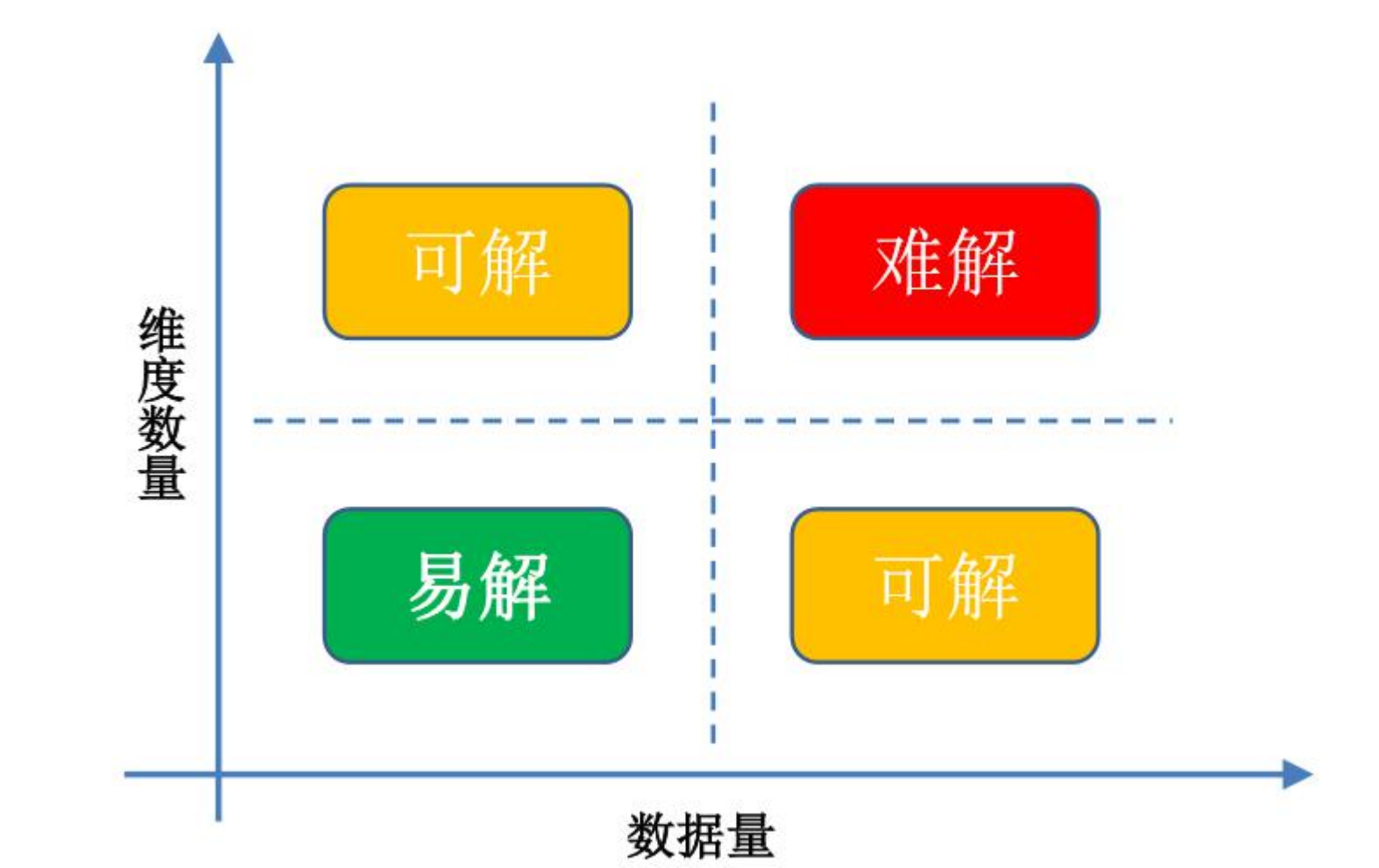


大数据带来的挑战

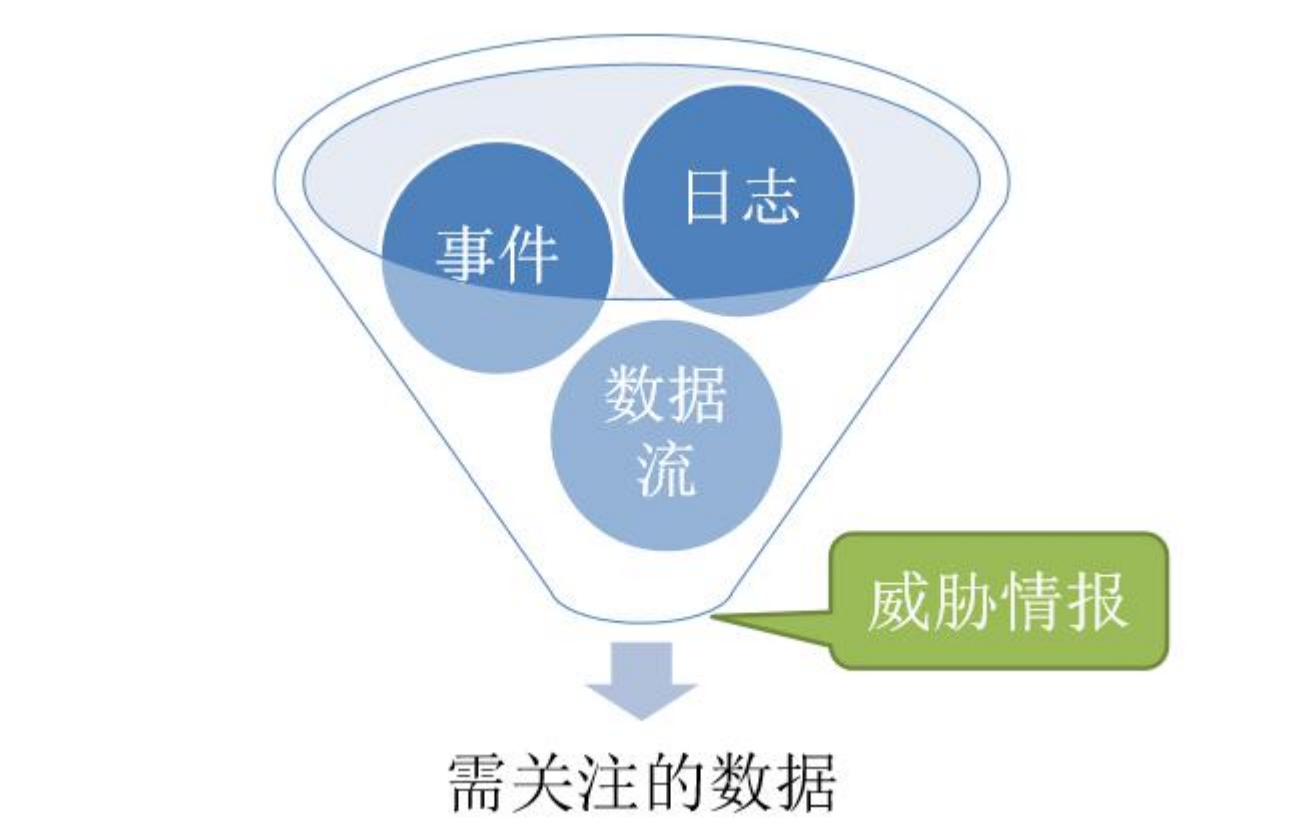
- 高维



大数据带来的挑战



从“大数据”中浓缩有价值的“小数据”



常见技术

- hadoop
- Spark
- Storm
- impala

hadoop家族（一）

- Hadoop Common
- HDFS
- MapReduce
- Hive
- Pig
- HBase
- ZooKeeper
- Avro
- Sqoop
- Mahout
- Cassandra
- Chukwa
- Ambari
- HCatalog
- Chukwa

hadoop家族（二）

- Hadoop Common
  - Hadoop体系最底层的一个模块，为Hadoop各子项目提供各种工具，如：配置文件和日志操作等。

hadoop家族（三）

- HDFS
  - 是Hadoop应用程序中主要的分布式储存系统， HDFS集群包含了一个NameNode（主节点），这个节点负责管理所有文件系统的元数据及存储了真实数据的DataNode（数据节点，可以有很多）。HDFS针对海量数据所设计，所以相比传统文件系统在大批量小文件上的优化，HDFS优化的则是对小批量大型文件的访问和存储。

hadoop家族（四）

- MapReduce
  - 是一个软件框架，用以轻松编写处理海量（TB级）数据的并行应用程序，以可靠和容错的方式连接大型集群中上万个节点（商用硬件）。

hadoop家族（五）

- Hive
  - Apache Hive是Hadoop的一个数据仓库系统，促进了数据的综述（将结构化的数据文件映射为一张数据库表）、即席查询以及存储在Hadoop兼容系统中的大型数据集分析。Hive提供完整的SQL查询功能——HiveQL语言，同时当使用这个语言表达一个逻辑变得低效和繁琐时，HiveQL还允许传统的Map/Reduce程序员使用自己定制的Mapper和Reducer。hive类似CloudBase，基于hadoop分布式计算平台上的提供data warehouse的sql功能的一套软件。使得存储在hadoop里面的海量数据 的汇总，即席查询简单化。

hadoop家族（六）

- Pig
  - Apache Pig是一个用于大型数据集分析的平台，它包含了一个用于数据分析应用的高级语言以及评估这些应用的基础设施。Pig应用的闪光特性在于它们的结构经得起大量的并行，也就是说让它们支撑起非常大的数据集。Pig的基础设施层包含了产生Map-Reduce任务的编译器。Pig的语言层当前包含了一个原生语言——Pig Latin，开发的初衷是易于编程和保证可扩展性。
  - Pig是SQL-like语言，是在MapReduce上构建的一种高级查询语言，把一些运算编译进MapReduce模型的Map和Reduce中，并且用户可以定义自己的功能。Yahoo网格运算部门开发的又一个克隆Google的项目Sawzall。
  - pig入门简单操作及语法包括支持数据类型、函数、关键字、操作符等

hadoop家族（七）

- HBase
  - Apache HBase是Hadoop数据库，一个分布式、可扩展的大数据存储。它提供了大数据集上随机和实时的读/写访问，并针对了商用服务器集群上的大型表格做出优化——上百亿行，上千万列。其核心是Google Bigtable论文的开源实现，分布式列式存储。就像Bigtable利用GFS（Google File System）提供的分布式数据存储一样，它是Apache Hadoop在HDFS基础上提供的一个类Bigatable。

Hadoop 架构

- 底层是Hadoop Distributed File System(HDFS),它存储Hadoop集群中所有存储节点上的文件。HDFS（对于本文）的上一层是MapReduce 引擎，该引擎由JobTrackers和TaskTrackers组成。通过对Hadoop分布式计算平台最核心的分布式文件系统HDFS、MapReduce处理过程，以及数据仓库工具Hive和分布式数据库Hbase
- Hadoop的框架最核心的设计就是HDFS，MapReduce和YARN为海量的数据提供了存储和计算、common支持其他Hadoop模块的通用工具HDFS主要是Hadoop的存储，用于海量的数据的存储；MapReduce主要用于分布式计算YARN是Hadoop中的资源管理系统

架构

Spark

- 创始组织：加州大学伯克利分校 AMP 实验室 (Algorithms, Machines, and People Lab) 开发
- Spark 是一种与 Hadoop 相似的开源集群计算环境，但是两者之间还存在一些不同之处，这些有用的不同之处使 Spark 在某些工作负载方面表现得更加优越，换句话说，Spark 启用了内存分布数据集，除了能够提供交互式查询外，它还可以优化迭代工作负载。
- Spark 是在 Scala 语言中实现的，它将 Scala 用作其应用程序框架。与 Hadoop 不同，Spark 和 Scala 能够紧密集成，其中的 Scala 可以像操作本地集合对象一样轻松地操作分布式数据集。
- 尽管创建 Spark 是为了支持分布式数据集上的迭代作业，但是实际上它是对 Hadoop 的补充，可以在 Hadoo 文件系统中并行运行。通过名为 Mesos 的第三方集群框架可以支持此行为。Spark 由加州大学伯克利分校 AMP 实验室 (Algorithms, Machines, and People Lab) 开发，可用来构建大型的、低延迟的数据分析应用程序。

Storm

- 创始人：Twitter
- Twitter将Storm正式开源了，这是一个分布式的、容错的实时计算系统，它被托管在GitHub上，遵循 Eclipse Public License 1.0。Storm是由BackType开发的实时处理系统，BackType现在已在Twitter麾下。GitHub上的最新版本是Storm 0.5.2，基本是用Clojure写的。

Impala

- 创始人：Cloudera公司
  - 新型查询系统，它提供SQL语义，能查询存储在Hadoop的HDFS和HBase中的PB级大数据。已有的Hive系统虽然也提供了SQL语义，但由于Hive底层执行使用的是MapReduce引擎，仍然是一个批处理过程，难以满足查询的交互性。相比之下，Impala的最大特点也是最大卖点就是它的快速。

大数据处理

- 大数据必然无法用人脑来推算、估测，或者用单台的计算机进行处理，必须采用分布式计算架构，依托云计算的分布式处理、分布式数据库、云存储和虚拟化技术，因此，大数据的挖掘和处理必须用到云技术。

大数据特点

- 不是随机样本，而是全体数据
- 不是精确性，而是混杂性
- 不是因果关系，而是相关关系

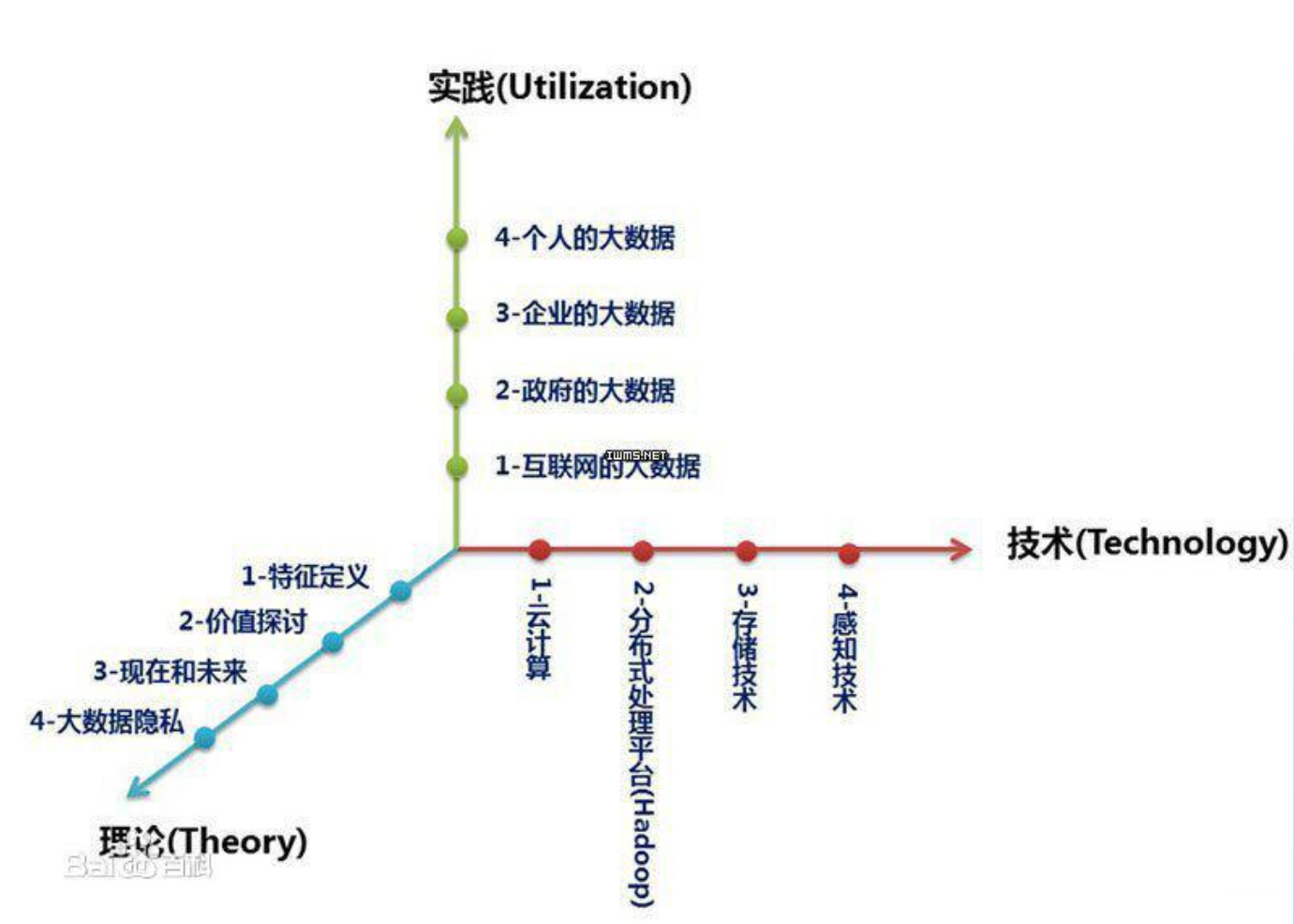
数据量单位

- 1 Byte =8 bit
- 1 KB = 1,024 Bytes = 8192 bit
- 1 MB = 1,024 KB = 1,048,576 Bytes
- 1 GB = 1,024 MB = 1,048,576 KB
- 1 TB = 1,024 GB = 1,048,576 MB
- 1 PB = 1,024 TB = 1,048,576 GB
- 1 EB = 1,024 PB = 1,048,576 TB
- 1 ZB = 1,024 EB = 1,048,576 PB
- 1 YB = 1,024 ZB = 1,048,576 EB
- 1 BB = 1,024 YB = 1,048,576 ZB
- 1 NB = 1,024 BB = 1,048,576 YB
- 1 DB = 1,024 NB = 1,048,576 BB

到底要多大

- 数据库
  - T级别
  - 微软
- 大数据
  - 一般在10TB规模左右，但在实际应用中，很多企业用户把多个数据集放在一起，已经形成了PB级的数据量；

理论、技术、实践


实践(Utilization)
4-个人的大数据
3-企业的大数据
2-政府的大数据
1-互联网的大数据
技术(Technology)
1-云计算
2-分布式处理平台(Hadoop)
3-存储技术
4-感知技术
1-特征定义
2-价值探讨
3-现在和未来
4-大数据隐私
理论(Theory)

大数据分析

- 可视化分析（Analytic Visualizations）
- 数据挖掘算法（Data Mining Algorithms）
- 预测性分析能力（Predictive Analytic Capabilities）
- 语义引擎（Semantic Engines）
- 数据质量和数据管理（Data Quality and Master Data Management）

大数据技术（一）

- 数据采集：ETL工具负责将分布的、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。
- 数据存取：关系数据库、NOSQL、SQL等。
- 基础架构：云存储、分布式文件存储等。

大数据技术（二）

- 数据处理：自然语言处理(NLP，NaturalLanguageProcessing)是研究人与计算机交互的语言问题的一门学科。处理自然语言的关键是要让计算机"理解"自然语言，所以自然语言处理又叫做自然语言理解(NLU，NaturalLanguage Understanding)，也称为计算语言学(Computational Linguistics。一方面它是语言信息处理的一个分支，另一方面它是人工智能(AI，Artificial Intelligence)的核心课题之一。

大数据技术（三）

- 统计分析：假设检验、显著性检验、差异分析、相关分析、T检验、方差分析、卡方分析、偏相关分析、距离分析、回归分析、简单回归分析、多元回归分析、逐步回归、回归预测与残差分析、岭回归、logistic回归分析、曲线估计、因子分析、聚类分析、主成分分析、因子分析、快速聚类法与聚类法、判别分析、对应分析、多元对应分析（最优尺度分析）、bootstrap技术等等。

大数据技术（四）

- 数据挖掘：分类 （Classification）、估计（Estimation）、预测（Prediction）、相关性分组或关联规则（Affinity grouping or association rules）、聚类（Clustering）、描述和可视化、Description and Visualization）、复杂数据类型挖掘(Text, Web ,图形图像，视频，音频等)
- 模型预测：预测模型、机器学习、建模仿真。
- 结果呈现：云计算、标签云、关系图等。

大数据处理（一）

采集

- 大数据的采集是指利用多个数据库来接收发自客户端（Web、App或者传感器形式等）的数据，并且用户可以通过这些数据库来进行简单的查询和处理工作。比如，电商会使用传统的关系型数据库MySQL和Oracle等来存储每一笔事务数据，除此之外，Redis和MongoDB这样的NoSQL数据库也常用于数据的采集。
- 在大数据的采集过程中，其主要特点和挑战是并发数高，因为同时有可能会有成千上万的用户来进行访问和操作，比如火车票售票网站和淘宝，它们并发的访问量在峰值时达到上百万，所以需要在采集端部署大量数据库才能支撑。并且如何在这些数据库之间进行负载均衡和分片的确是需要深入的思考和设计。

大数据处理（二）

- 导入/预处理
  - 虽然采集端本身会有很多数据库，但是如果要对这些海量数据进行有效的分析，还是应该将这些来自前端的数据导入到一个集中的大型分布式数据库，或者分布式存储集群，并且可以在导入基础上做一些简单的清洗和预处理工作。也有一些用户会在导入时使用来自Twitter的Storm来对数据进行流式计算，来满足部分业务的实时计算需求。
  - 导入与预处理过程的特点和挑战主要是导入的数据量大，每秒钟的导入量经常会达到百兆，甚至千兆级别。

大数据处理（三）

- 统计/分析
  - 统计与分析主要利用分布式数据库，或者分布式计算集群来对存储于其内的海量数据进行普通的分析和分类汇总等，以满足大多数常见的分析需求，在这方面，一些实时性需求会用到EMC的GreenPlum、Oracle的Exadata，以及基于MySQL的列式存储Infobright等，而一些批处理，或者基于半结构化数据的需求可以使用Hadoop。
  - 统计与分析这部分的主要特点和挑战是分析涉及的数据量大，其对系统资源，特别是I/O会有极大的占用。

大数据处理（四）

- 挖掘
  - 与前面统计和分析过程不同的是，数据挖掘一般没有什么预先设定好的主题，主要是在现有数据上面进行基于各种算法的计算，从而起到预测（Predict）的效果，从而实现一些高级别数据分析的需求。比较典型算法有用于聚类的Kmeans、用于统计学习的SVM和用于分类的NaiveBayes，主要使用的工具有Hadoop的Mahout等。该过程的特点和挑战主要是用于挖掘的算法很复杂，并且计算涉及的数据量和计算量都很大，常用数据挖掘算法都以单线程为主。
  - 整个大数据处理的普遍流程至少应该满足这四个方面的步骤，才能算得上是一个比较完整的大数据处理。