大数据算法

第一讲 大数据算法概述

哈尔滨工业大学
王宏志
wangzh@hit.edu.cn

本讲内容

1.1 大数据的定义与特点

1.2 大数据算法

1.3 大数据算法设计与分析

什么是大数据？

- 至今没有公认的定义
- 定义1（Kusnetzky, Dan. What is "Big Data?"）

所涉及的数据量规模巨大到无法通过人工，在合理时间内达到截取、管理、处理、并整理成为人类所能解读的信息

- 定义2（维克托·迈尔-舍恩伯格、肯尼斯·库克耶.“大数据时代”）

不用随机分析法（抽样调查）这样的捷径，而采用所有数据的方法

- 定义3（“大数据”（Big data）研究机构Gartner）

“大数据”是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

“大”有多大

1Byte = 8 bit
1 KB = 1,024 Bytes
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

仟 1,000
萬 10^4
億 10^8
兆 10^{12}
京 10^{16}
垓 10^{20}
秭 10^{24}
穰 10^{28}
沟 10^{32}
涧 10^{36}
正 10^{40}
载 10^{44}

俱胝 10^7
阿庾多 10^{14}
那由他 10^{28}
频波罗 10^{56}
矜羯罗 10^{112}
阿伽罗 10^{224}
最胜 10^{448}
摩婆罗 10^{896}
阿婆罗 10^{1792}
多婆罗 10^{3584}
界分 10^{7168}
普摩 10^{14336}
祢摩 10^{28672}
阿婆钤 10^{57344}
弥伽婆 10^{114688}
毗攞伽 10^{229376}
毗伽婆 10^{458752}
僧羯逻摩 10^{917504}
毗萨罗 $10^{1835008}$
毗赡婆 $10^{3670016}$
毗盛伽 $10^{7340032}$
毗素陀 $10^{14680064}$
毗婆诃 $10^{29360128}$
毗薄底 $10^{58720256}$

毗佉担 $10^{117440512}$
称量 $10^{234881024}$
一持 $10^{469762048}$
异路 $10^{939524096}$
颠倒 $10^{1879048192}$
三末耶 $10^{3758096384}$
毗睹罗 $10^{7516192768}$
奚婆罗 $10^{15032385536}$
伺察 $10^{30064771072}$
周广 $10^{60129542144}$
高出 $10^{120259084288}$
最妙 $10^{240518168576}$
泥罗婆 $10^{481036337152}$
诃理婆 $10^{962072674304}$
一动 $10^{1924145348608}$
诃理蒲 $10^{3848290697216}$
诃理三 $10^{7696581394432}$
奚鲁伽 $10^{15393162788864}$
达攞步陀 $10^{30786325577728}$
诃鲁那 $10^{61572651155456}$
摩鲁陀 $10^{123145302310912}$
忏慕陀 $10^{246290604621824}$
鞞攞陀 $10^{492581209243648}$
摩鲁摩 $10^{985162418487296}$

“大”有多大 (续)

调伏 $10^{1970324836974592}$
离憍慢 $10^{3940649673949184}$
不动 $10^{7881299347898368}$
极量 $10^{15762598695796736}$
阿么怛罗 $10^{31525197391593472}$
勃么怛罗 $10^{63050394783186944}$
伽么怛罗 $10^{126100789566373888}$
那么怛罗 $10^{252201579132747776}$
奚么怛罗 $10^{504403158265495552}$
鞞么怛罗 $10^{1008806316530991104}$
鉢罗么怛罗 $10^{2017612633061982208}$
尸婆么怛罗 $10^{4035225266123964416}$
翳罗 $10^{8070450532247928832}$
薜罗 $10^{16140901064495857664}$
谛罗 $10^{32281802128991715328}$
偈罗 $10^{64563604257983430656}$
窣步罗 $10^{129127208515966861312}$
泥罗 $10^{258254417031933722624}$
计罗 $10^{516508834063867445248}$
细罗 $10^{1033017668127734890496}$
睥罗 $10^{2066035336255469780992}$
谜罗 $10^{4132070672510939561984}$
娑攞荼 $10^{8264141345021879123968}$
谜鲁陀 $10^{16528282690043758247936}$
契鲁陀 $10^{33056565380087516495872}$
摩睹罗 $10^{66113130760175032991744}$
娑母罗 $10^{132226261520350065983488}$
阿野娑 $10^{264452523040700131966976}$
迦么罗 $10^{528905046081400263933952}$
摩伽婆 $10^{1057810092162800527867904}$
阿怛罗 $10^{2115620184325601055735808}$
酰鲁耶 $10^{423124036865120211471616}$
薜鲁婆 $10^{846248073730240422943232}$
羯罗波 $10^{1692496147460480844588646 4}$
诃婆婆 $10^{3384992294920961689177292 8}$
毗婆罗 $10^{6769984589841923378354585 6}$
那婆罗 $10^{1353996917968384675670917 12}$
摩攞罗 $10^{2707993835936769351341834 24}$
娑婆罗 $10^{5415987671873538702683668 48}$
迷攞普 $10^{1083197534374707740536733 696}$
者么罗 $10^{2166395068749415481073467 392}$
駄么罗 $10^{4332790137498830962146934 784}$
鉢攞么陀 $10^{8665580274997661924293869 568}$
毗迦摩 $10^{1733116054999532384858773 9136}$
乌波跋多 $10^{3466232109999064769717547 8272}$
演説 $10^{6932464219998129539435095 6544}$
无尽 $10^{1386492843999625907887019 13088}$
出生 $10^{2772985687999251815774038 26176}$

“大”有多大 (续)

无我 $10^{554597137599850363154807652352}$

阿畔多 $10^{1109194275199700726309615304704}$

青莲华 $10^{221838855039940145261923060940 8}$

鉢头摩 $10^{443677710079880290523846121881 6}$

僧祇 $10^{88735542015976058104769224376 32}$

趣 $10^{17747108403195211620953844875264}$

至 $10^{35494216806390423241907689750528}$

阿僧祇 $10^{70988433612780846483815379501056}$

阿僧祇转 $10^{14197686722556169296763075900211 2}$

无量 $10^{28395373445112338593526151800422 4}$

无量转 $10^{56790746890224677187052303600844 8}$

无边 $10^{11358149378044935437410460720168 96}$

无边转 $10^{22716298756089870874820921440337 92}$

无等 $10^{45432597512179741749641842880675 84}$

无等转 $10^{90865195024359483499283685761351 68}$

不可数 $10^{18173039004871896699856737152270336}$

不可数转 $10^{36346078009743793399713474304540672}$

不可称 $10^{72692156019487586799426948609081344}$

不可称转 $10^{14538431203897517359885389721816 2688}$

不可思 $10^{29076862407795034719770779443632 5376}$

不可思转 $10^{58153724815590069439541558887265 0752}$

不可量 $10^{11630744963118013887908311777453015 04}$

不可量转 $10^{23261489926236027775816623554906030 08}$

不可说 $10^{46522979852472055551633247109812060 16}$

不可说转 $10^{93045959704944111103266494219624120 32}$

不可说不可说 $10^{18609191940988822220653298843924824064}$

不可说不可说转 $10^{37218383881977644441306597687849648128}$

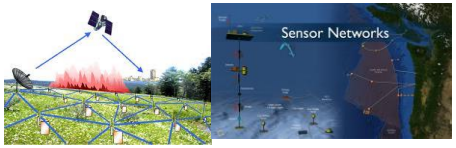
不可说不可说不可说 $10^{74436767763955288882613195375699296256}$

处处皆是大数据

社交网络

科学仪器

移动设备

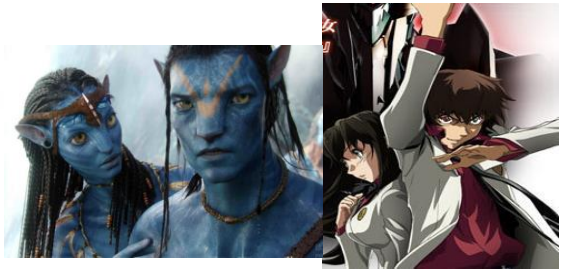


传感网

医疗数据

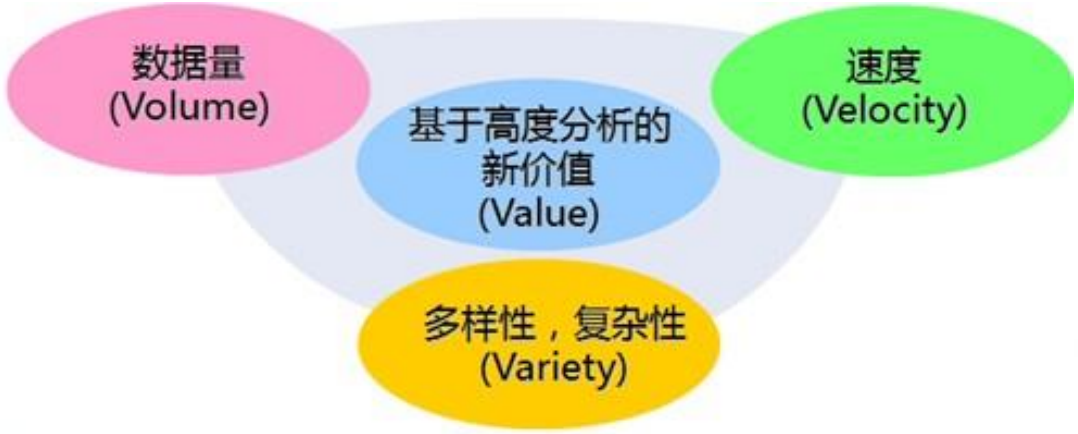
Walmart
Save money. Live better.

商业数据



计算机艺术

大数据的特点



大数据的应用

- 预测
- 推荐
- 商业情报分析
- 科学研究

应用中的大数据算法

- 预测
- 推荐
- 商业情报分析
- 科学研究

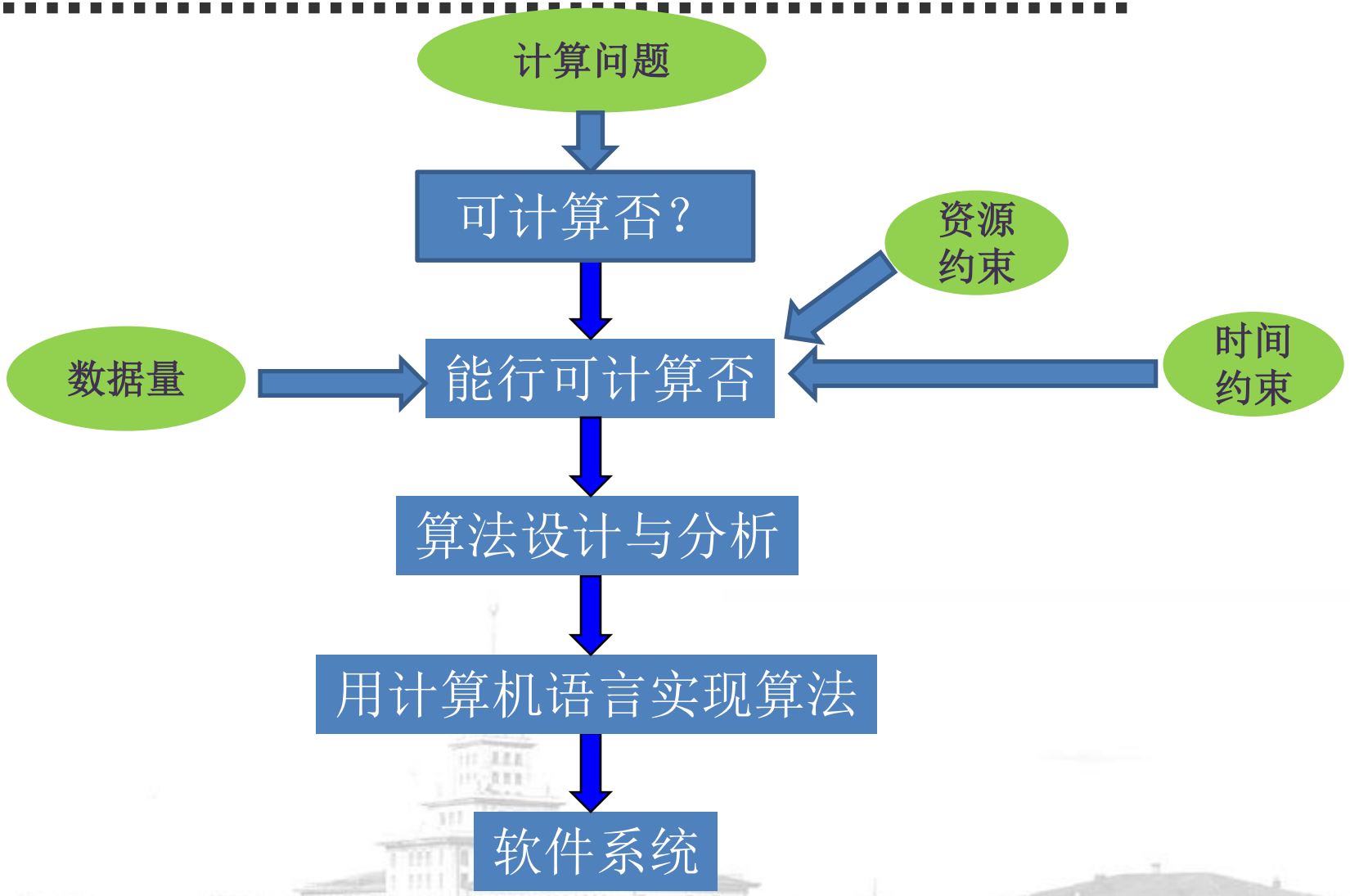
本讲内容

1.1 大数据的定义与特点

1.2 大数据算法

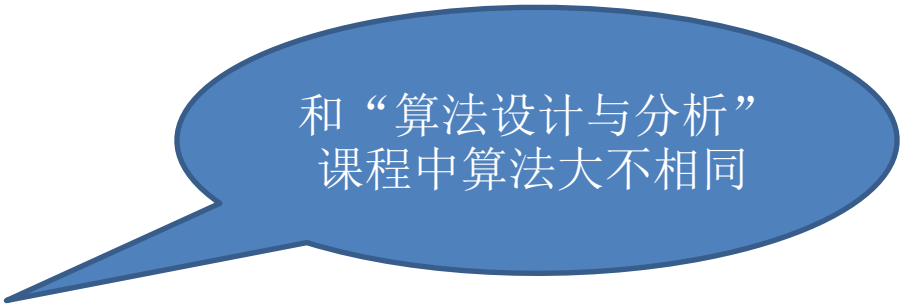
1.3 大数据算法设计与分析

大数据上问题求解计算问题的过程



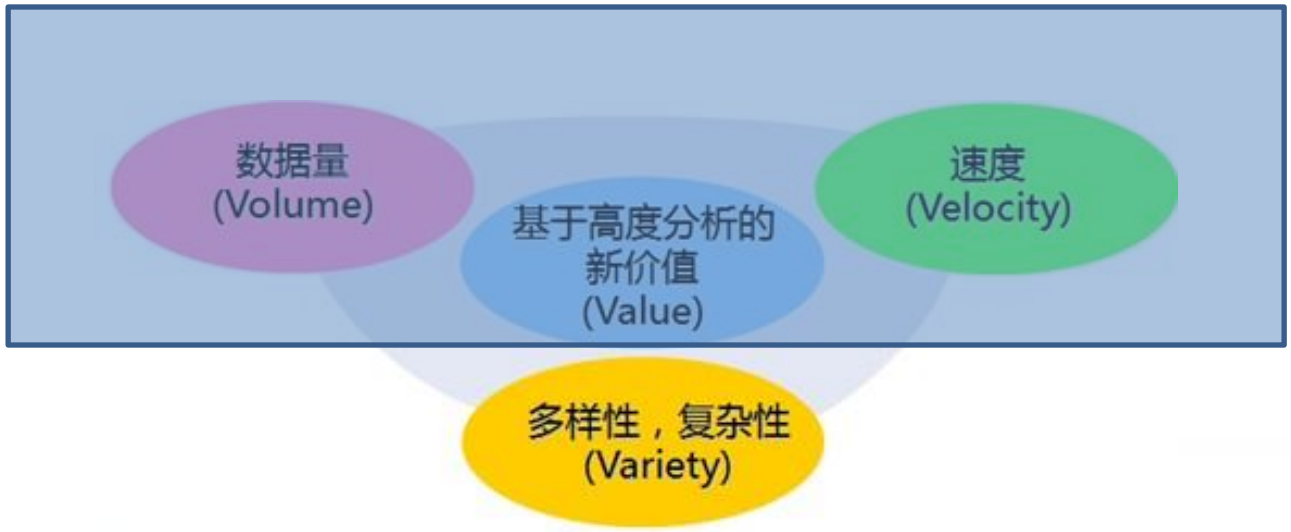
大数据算法

- **大数据算法的定义**
  - 在给定的**资源约束**下，以大数据为输入，在给定**时间约束**内可以生成满足给定**约束**结果的算法。



- **大数据算法可以不是：**
  - 精确算法
  - 内存算法
  - 串行算法
  - 仅在电子计算机上运行的算法
- **大数据算法不仅是：**
  - 云计算
  - MapReduce
  - 大数据分析和挖掘的算法
  - 数据库中的算法

大数据的特点与大数据算法



大数据算法的难度

- **访问全部数据时间过长**
  - 读取部分数据

时间亚线性算法

- **数据难于放入内存计算**
  - 将数据存储到磁盘上
  - 仅基于少量数据进行计算

外存算法

空间亚线性算法

- **单个计算机难以保存全部数据，计算需要整体数据**
  - 并行处理

并行算法

- **计算机计算能力不足或知识不足**
  - 人来帮忙

众包算法

本讲内容

1.1 大数据的定义与特点
1.2 大数据算法
1.3 大数据算法设计与分析

大数据的算法设计技术

- 精确算法设计方法
- 并行算法
- 近似算法
- 随机算法
- 在线算法/数据流算法
- 外存算法
- 面向新型体系结构的算法
- 现代优化算法

大数据的算法分析

- 时间空间复杂性
- IO复杂性
- 结果质量（近似比、competitive ratio）
- 通讯复杂性

这门课的内容

- 亚线性算法
- 外存算法
- 并行算法
- 众包算法