



大数据有多大？

南方科技大学

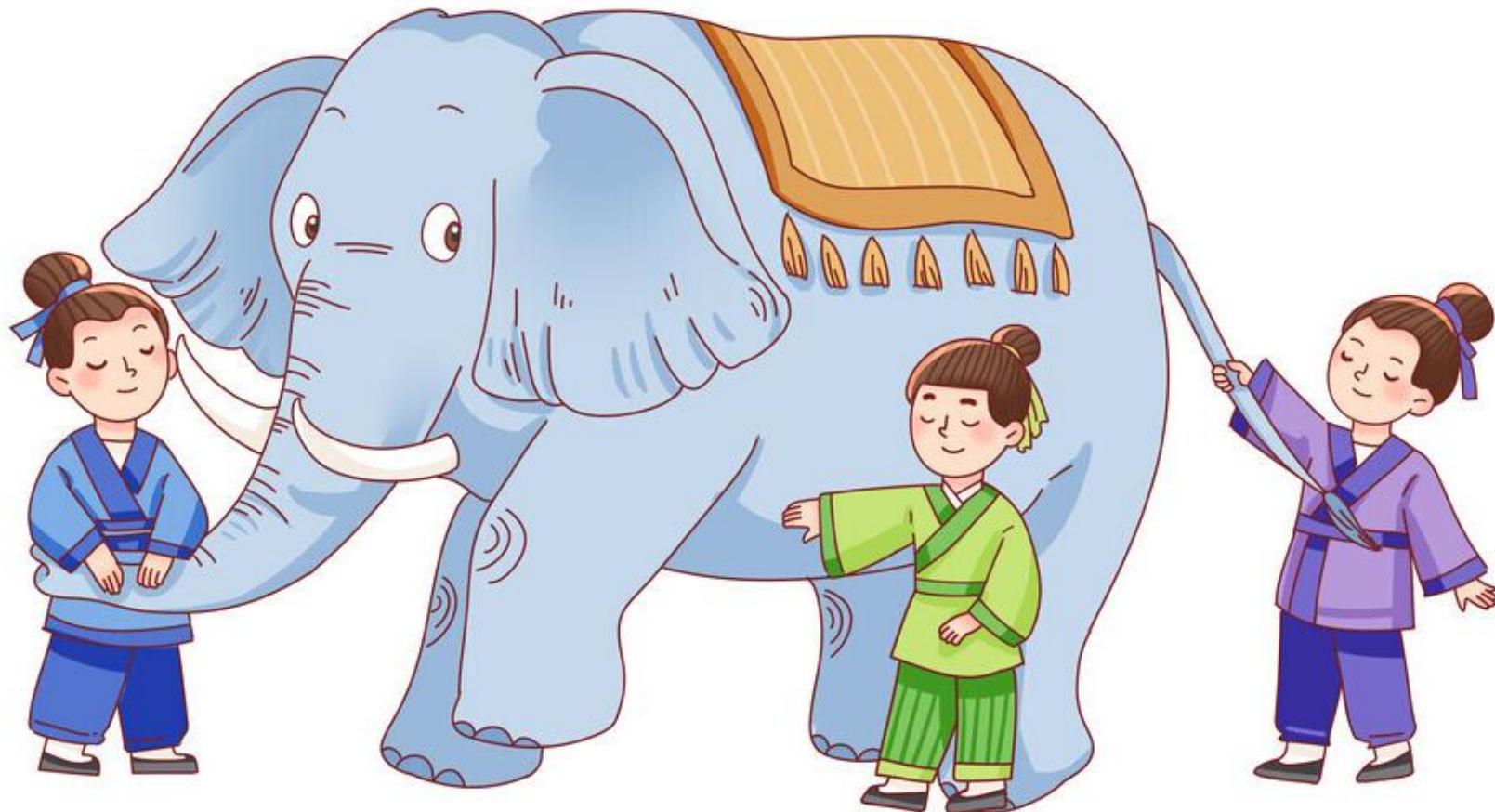
唐 博

tangb3@sustech.edu.cn





盲人摸象





大数据有多大



大数据有多大

唐博

日常生活中常常有这种现象：对于一种习以为常的现象，每个人都谈论它，但没人知道它的由来。大数据（big data）显然就属于这种情况。目前，大数据这个概念已经走入社会的各个角落。一般人都能懵懵懂懂地讨论大数据。在一般人的认知中，大数据就是数量庞大而复杂的数据集合。应用传统的数据处理方法，不能轻易厘清这些数据集合的头绪以及挖掘其中的潜在价值。

但是，这就是大数据的全部吗？大数据的特点到底是什么？

大数据的概念并不是突然蹦出来的，它也经历了一个逐渐演化的过程。大数据的主要特征可以用4个“V”来表示（图2-3）：第一个“V”是容量（volume），这就是一般人最能了解的特征，我们使用的手机容量一般以GB为单位，如64 GB、128 GB等，而大数据处理的数据可以高达十万甚至百万级别GB；第二个“V”是类型（variety），大数据所包括的数据不仅仅是单一的文本文件，同时还包括视频、音频、图片、定位信息，甚至是阿尔法狗下棋所产生的棋谱等其他类型的文件和信息；第三个“V”是速度（velocity），大数

《十万个高科技为什么》之《大数据有多大？》

量子基石篇

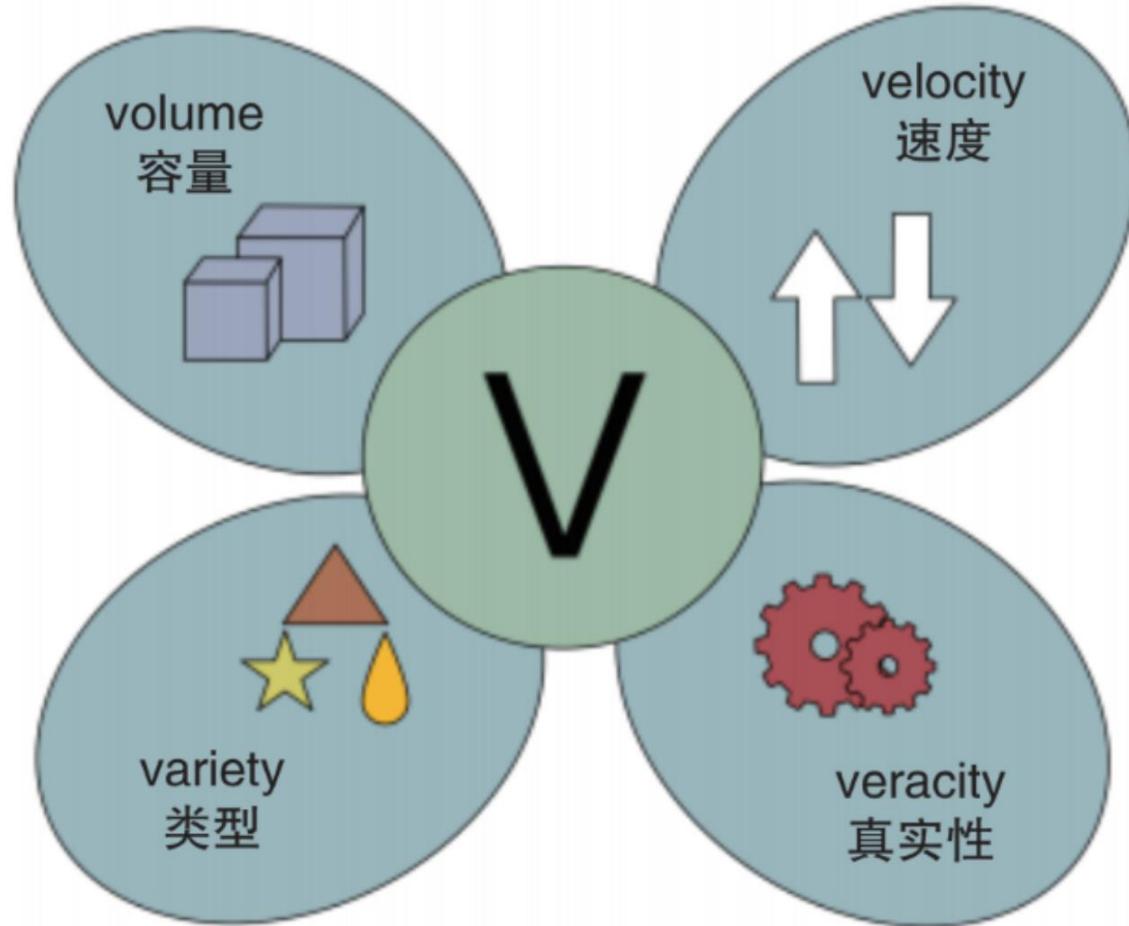
电子与信息篇

材料与化学篇

生命与科技篇

地球与环

大数据的4V特征



大数据“4V”特征



提纲



大数据是什么

- 关于大数据的三两趣事

□ 大数据的研究

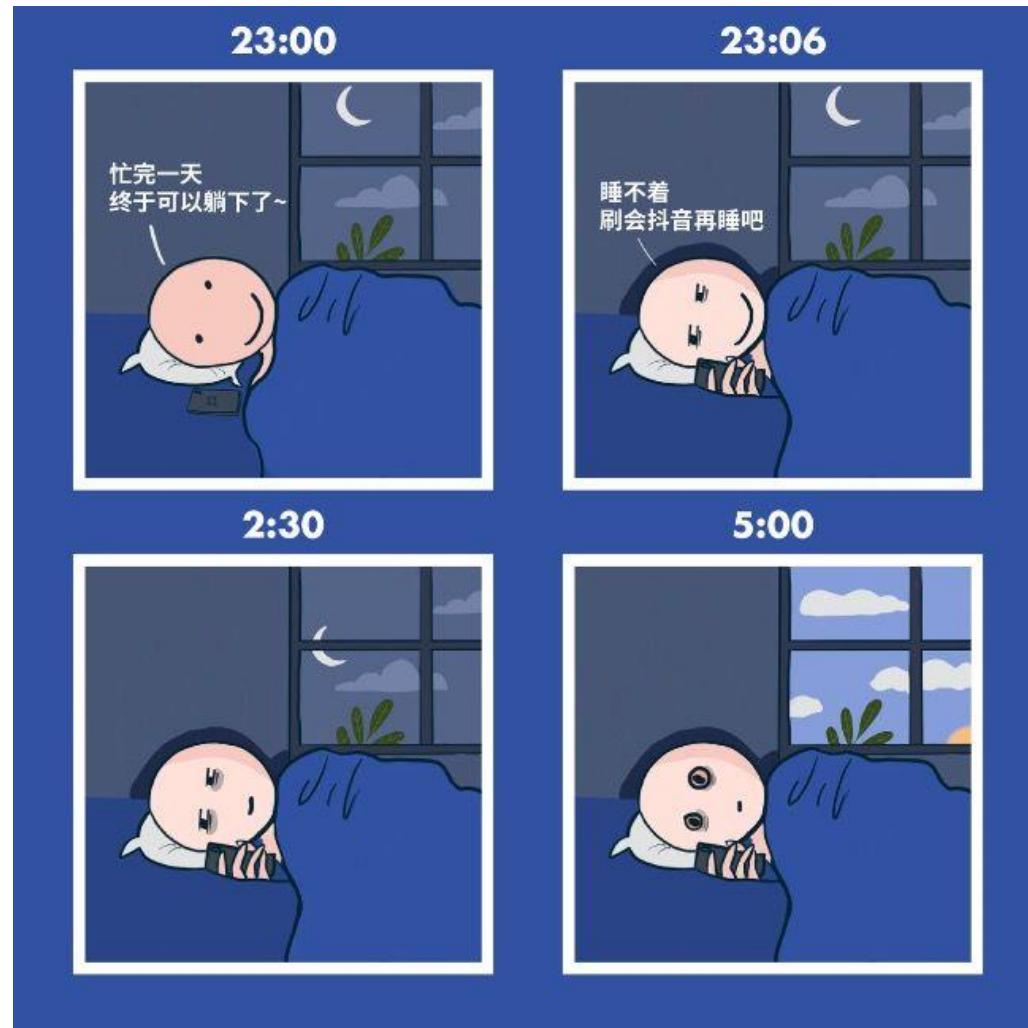
- 从研究初探到未来挑战

□ 大数据的未来

- 让大数据赋能生产生活



什么是大数据（一）



刷抖音是停不下来的！



什么是大数据（一）



Customers who bought this item also bought

Page 1 of 11



Redragon GS520 RGB Desktop Speakers, 2.0 Channel PC Computer Stereo Speaker with 6...
★★★★★ 9,376
 Amazon's Choice in Computer Speakers
 \$31.99
 Get it as soon as Thursday, Nov 14
 FREE Shipping on orders over \$49 shipped by Amazon



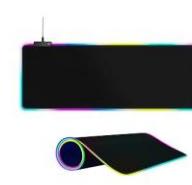
Sceptre Curved 24.5-inch Gaming Monitor up to 240Hz 1080p R1500 1ms DisplayPort x2 HDMI x2...
★★★★★ 6,651
 Amazon's Choice in Computer Monitors
 -7% \$139.05
 List: \$149.97
 Get it as soon as Friday, Nov 15
 FREE Shipping by Amazon



Sceptre 30-inch Curved Gaming Monitor 21:9 2560x1080 Ultra Wide/Slim HDMI DisplayPort up to 200Hz Build-in...
★★★★★ 11,972
 -10% \$179.97
 List: \$199.97
 Get it as soon as Friday, Nov 15
 FREE Shipping by Amazon



Sceptre Curved 24-inch Gaming Monitor 1080p R1500 98% sRGB HDMI x2 VGA Build-in Speakers, VESA Wall Mount Machine Black (C248W-...
★★★★★ 23,093
 1 offer from \$84.89



Large RGB Gaming Mouse Pad -15 Light Modes Touch Control Extended Soft Computer Keyboard...
★★★★★ 4,366
 Amazon's Choice in Mouse Pads
 -50% \$9.99
 List: \$19.99
 Get it as soon as Thursday, Nov 14
 FREE Shipping on orders over \$49 shipped by Amazon



RGB Mousepad Led Mouse Pad, Large Mouse Pad,Led and Big Mouse mat
★★★★★ 6,285
 -50% \$9.99
 List: \$19.99
 Get it as soon as Thursday, Nov 14
 FREE Shipping on orders over \$49 shipped by Amazon



MONTECH XR, ATX Mid-Tower PC Gaming Case, 3 x 120mm ARGB PWM Fans Pre-Installed, Full-View Dual Tempered...
★★★★★ 633
 \$63.90
 Get it Dec 4 - 24
 \$42.05 shipping



单品



自营 商品名称商品名称商品名称商品名称
 ¥499.4 满减 看相似

关键词

手提包

促销活动



厨卫818 火力全开

全场厨卫低至5折，更有万件商品秒杀

品牌



BOSE

自营 BOSE官方旗舰店
 1256927人已关注

资讯



被字字字 [明星]

618打折买三送一，超级好看，而且很好搭，长裙短裤搭配长裙短裤搭配...

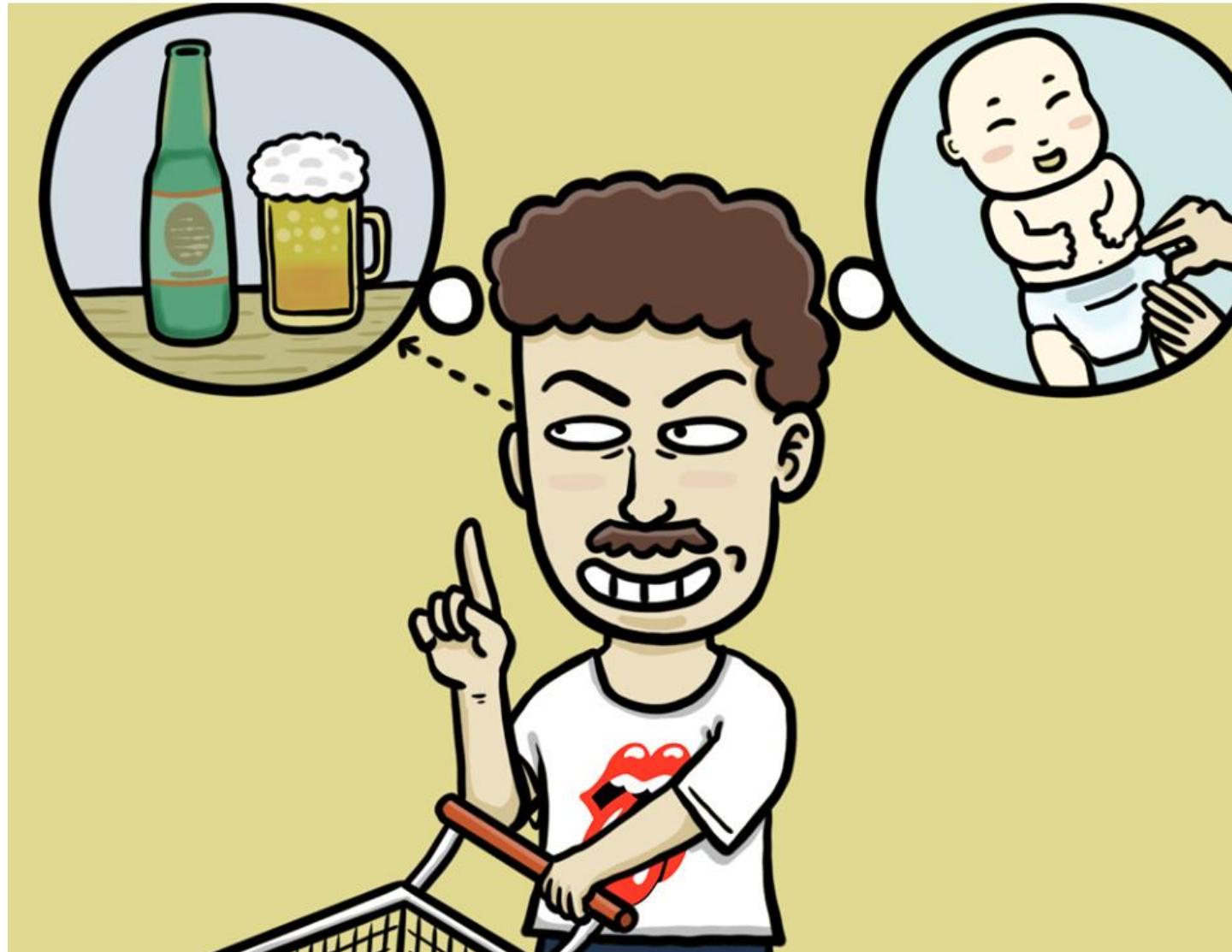
2234

晒单

你还想买什么？猜你喜欢？

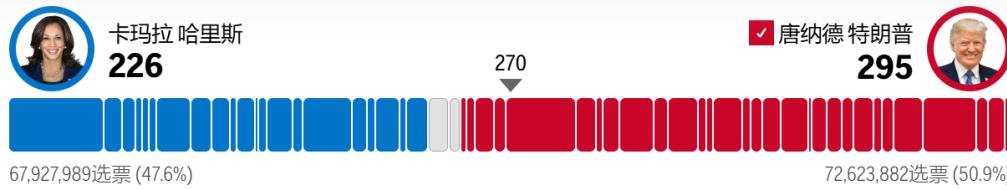


啤酒与纸尿布的故事



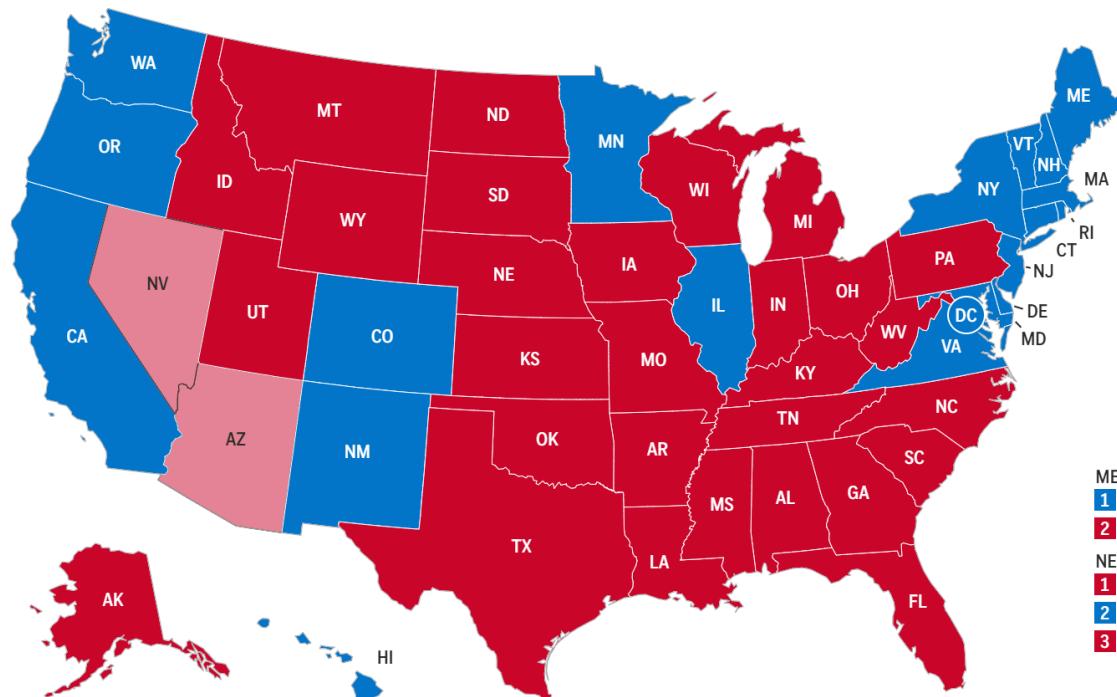


什么是大数据（二）



2024 总统选举

全国地图 ▾



美国大选

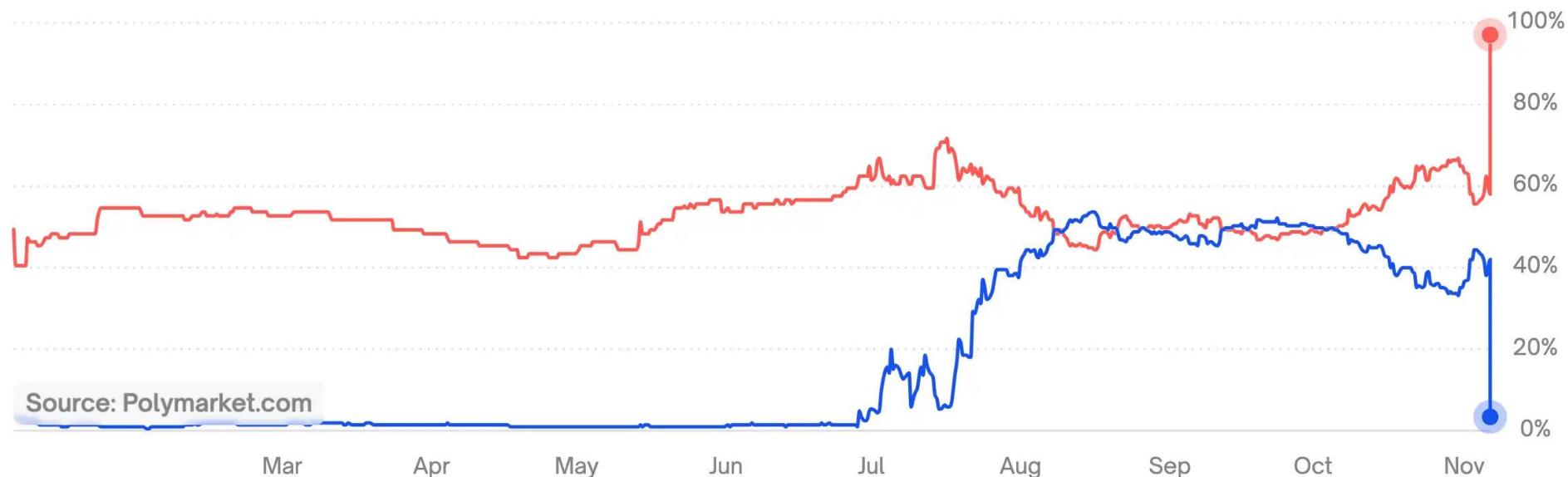


什么是大数据（二）



- Donald Trump 96.9%
- Kamala Harris 3.5%

 Polymarket



Trump vs Harris



义乌指数



sina 新浪财经 经济新闻滚动 > 正文

事关2024美国大选，是时候看看“义乌指数”了

2024年11月04日 23:14 媒体滚动

新浪财经APP | A+





什么是大数据（三）



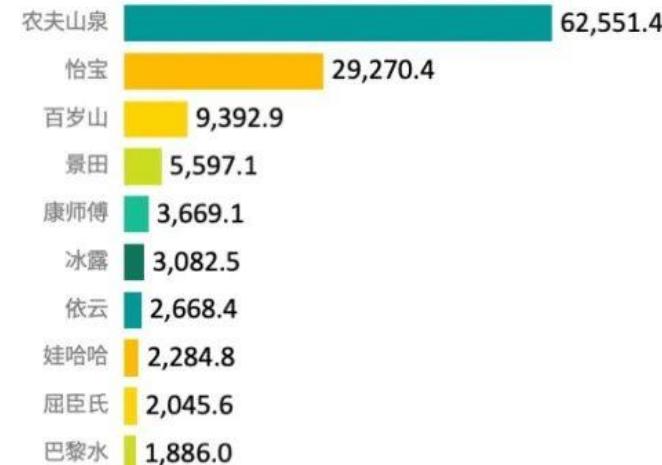
O2OMind x 包装饮用水

消费者更多通过美团（7.5亿元）、饿了么（3.69亿元）等平台购买包装饮用水；
农夫山泉、怡宝、百岁山等包装饮用水O2O到家渠道有出色表现。

2021年各O2O平台包装饮用水营收（亿元）



2021年三大O2O平台包装饮用水TOP10品牌营收(万元)



农夫山泉怎么就成了大自然的印钞机？

什么是大数据（三）

⑤ 深層地下水
• 取自地下河床170米
• 年徑流量達15億立方米
• 源自北天山冰川雪融水

⑩ 硫泉水
• 森林覆蓋率：92%
• 地下水資源總量：5.78億立方米
• 漠河地處大興安嶺山脈北麓

② 硫泉水及自然湧出泉水
• 總面積：1,964平方公里
• 森林覆蓋率：88%
• 亞洲東部保存最為完好的典型森林生態系統

⑨ 山泉水
• 森林覆蓋率：93%
• 河北霧靈山是京津地區重要的水源地

⑦ 山泉水
• 森林覆蓋率：94.3%
• 年累計降水量：2.47億立方米
• 太白山是青藏高原以東第一高峰

深層庫水
• 水域面積：745平方公里
• 儲水量：290.5億立方米
• 國家南水北調中綫工程水源地

⑥ 山泉水
• 森林覆蓋率：87%
• 年平均降水量：1,922毫米
• 峨眉山是中國公認的優質水源地

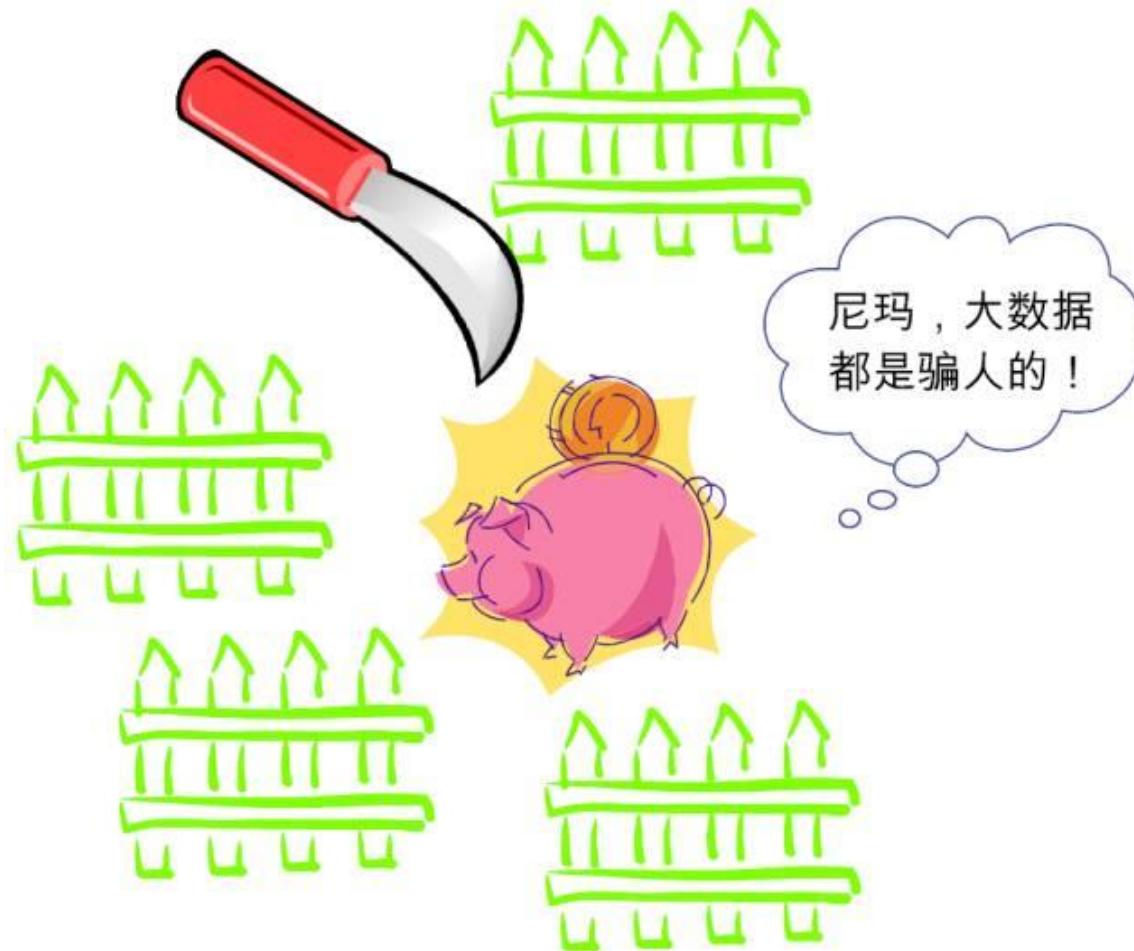
⑧ 山泉水
• 武陵山是中國亞熱帶森林系統核心區
• 長江流域重要的水源涵養地

① 深層湖水
• 水域面積：573平方公里
• 儲水量：178.4億立方米
• 被譽為「天下第一秀水」

④ 深層湖水
• 水域面積：370平方公里
• 儲水量：139億立方米
• 華南地區第一大湖

利用大数据提前布局十大优质水源：广布局-就近买

什么是大数据（三）



自打出生以来，就在猪圈这个世外桃源里美满地生活着。
“猪”生如此，夫复何求？
直到它从小猪长成肥猪.....在春节前的一个下午



分享提纲



✓ 大数据是什么

- 关于大数据的三两趣事

✓ 大数据的研究

- 从研究初探到技术热点

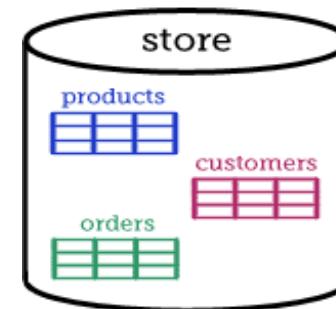
□ 大数据的未来

- 让大数据赋能生活生产

BMW汽车2014-2024的销
量如何？



① 查询



② 检索

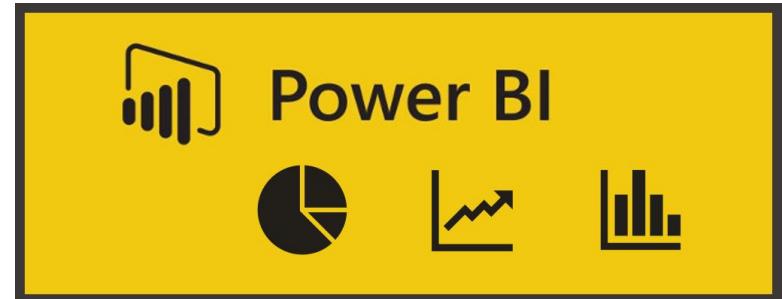
③ 分析



现有商业智能软件

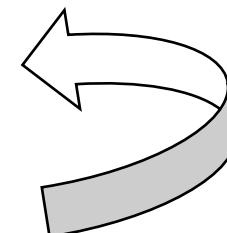


Better Decisions Every Day™

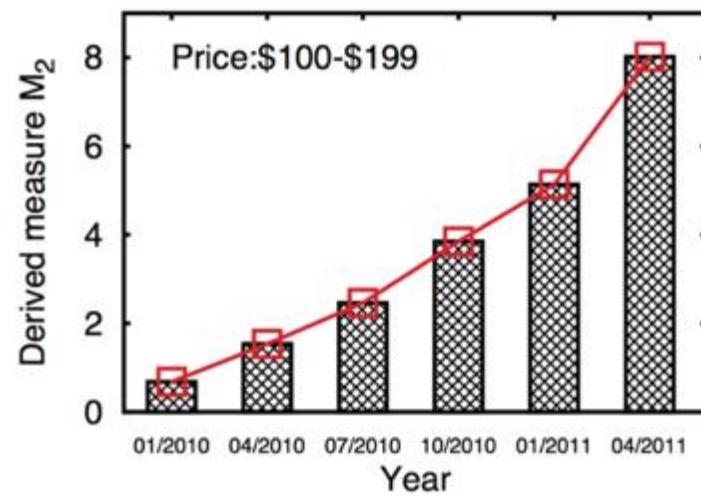
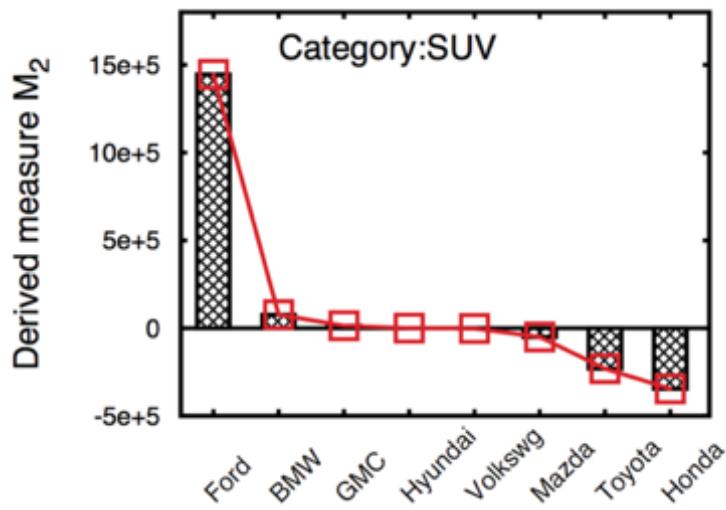


现有大数据分析方法

- ❖ 用户知道自己要分析什么
 - 比如宝马汽车的销量
- ❖ 用户知道底层数据张啥样子
 - 比如schema，数据分布等
- ❖ Hit-and-trial 方法
 - 提交查询
 - 查找结果
 - 分析有趣程度
- ❖ 分析结果质量差异
 - 数据分析师和销售经理



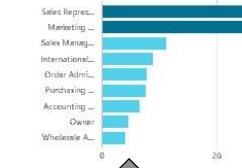
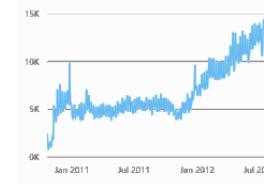
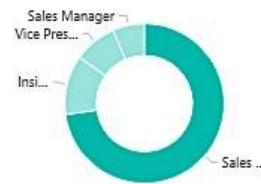
洞察力定义



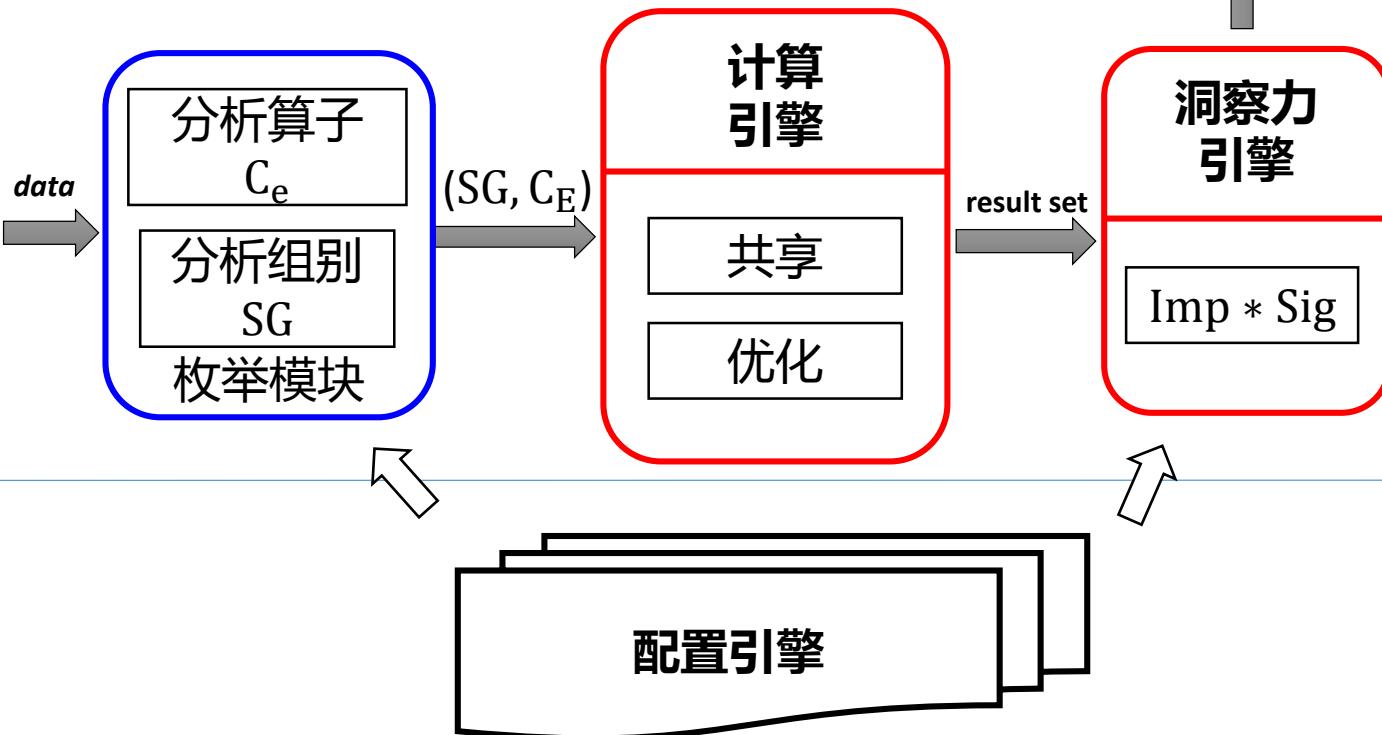
离群点

趋势分析

Datasets in OLAP System



用户
界面



洞察力
挖掘

系统
配置层



行业落地



[微软 Everyday AI 发布会：四大工具把 AI 带到你身边 >](#)

微软研究院 AI 头条...

Excel引入了最新的Insights功能，它能够帮助我们智能分析表格中的数据并给出相应的图文解释。比如当你提供了一个销售表的原始数据后，Insights功能可以自动帮你做出销售量走势、市场份额、产品周期等等图文分析。

A	B	C	D	E	F
Team	Sub Class	Group	Line Item	Fiscal Month	Spend
1 Operations	People	Infrastructure	WW IT Distributions	July 2013	4644
2 Marketing	People	Infrastructure	WW IT Distributions	July 2013	5913
3 Marketing	Shared Operations	Allocated Azure	WW IT Distributions	July 2013	65
4 Marketing	People	Infrastructure	WW IT Distributions	July 2013	61380
5 Engineering	Shared Operations	Allocated Azure	Storage	July 2013	21
6 Engineering	Shared Operations	Allocated Azure	Compute	July 2013	5211
7 Engineering	Shared Operations	Allocated Azure	WW IT Distributions	July 2013	4645
8 Sales	People	Infrastructure	WW IT Distributions	July 2013	663
9 Support	People	Infrastructure	WW IT Distributions	July 2013	15266
10 Hardware	People	Infrastructure	WW IT Distributions	July 2013	5973
11 IT	People	Infrastructure	WW IT Distributions	July 2013	98
12 IT	Shared Operations	Allocated Azure	Storage	July 2013	13940
13 Vendor	People	Infrastructure	WW IT Distributions	July 2013	7634
14 Contractor	People	Infrastructure	WW IT Distributions	July 2013	4647
15 Satellite Engineers	People	Infrastructure	WW IT Distributions	July 2013	1659
16 General Manager	People	Infrastructure	Storage	July 2013	10
17 General Manager	Shared Operations	Allocated Azure	WW IT Distributions	July 2013	5644
18 Satellite Support	People	Infrastructure	WW IT Distributions	July 2013	663
19 Satellite Sales	People	Infrastructure	WW IT Distributions	July 2013	663
20 Satellite Marketing	People	Infrastructure	WW IT Distributions	July 2013	663
21 Satellite Operations	Shared Operations	Allocated Azure	Compute	July 2013	4648
22 Hardware	Shared Operations	Allocated Azure	Compute	July 2013	32
23 IT	Shared Operations	Allocated Azure	WW IT Distributions	July 2013	381
24 General Manager	Shared Operations	Allocated Azure	Compute	July 2013	620
25 Satellite Operator	Shared Operations	Allocated Azure	Compute	July 2013	207
26 Operations	People	Other People	Travel & Entertainment	July 2013	1265
27 Marketing	People	Other People	Travel & Entertainment	July 2013	35600
28 Engineering	People	Other People	Travel & Entertainment	July 2013	61962
29 Sales	People	Other People	Travel & Entertainment	July 2013	1247
30 Support	People	Other People	Travel & Entertainment	July 2013	3045
31 Hardware	People	Other People	Travel & Entertainment	July 2013	18516
32 IT	People	Other People	Travel & Entertainment	July 2013	3528
33 Monitor	People	Other People	Travel & Entertainment	July 2013	14487

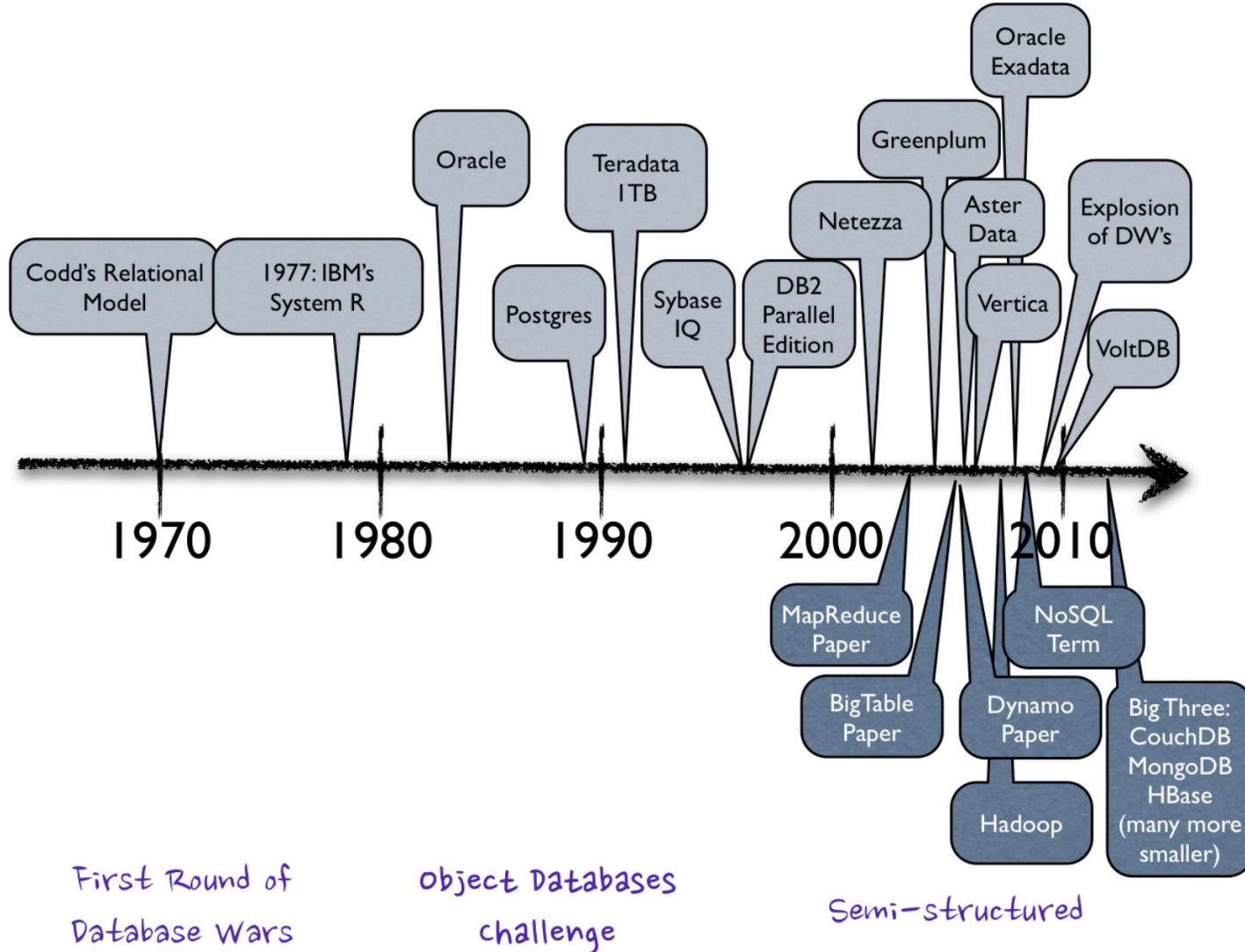


微软办公软件Excel集成洞察力功能

从跟随者到引领者



大数据系统研究历史



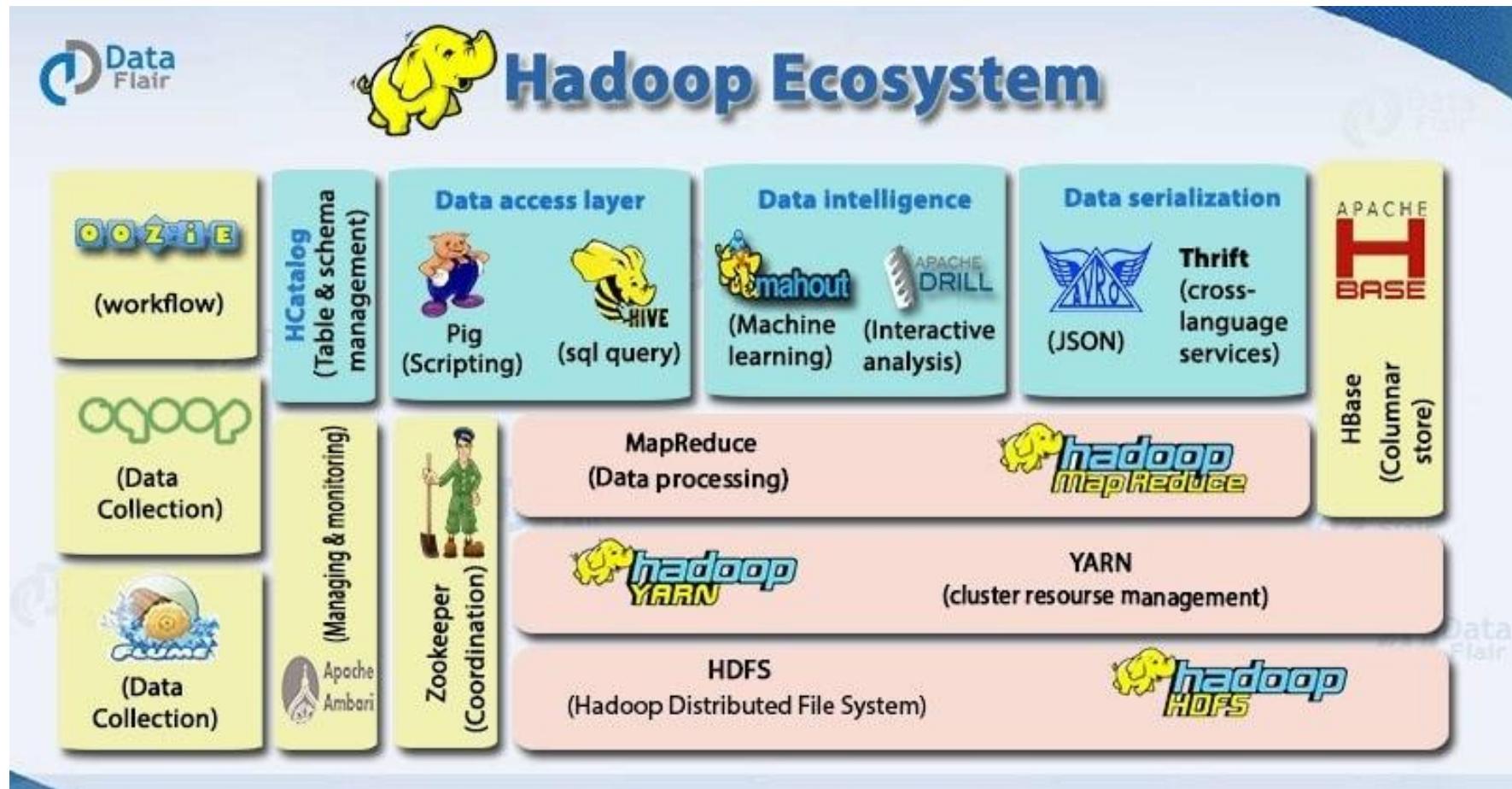


大数据系统产品





第一款以大数据命名的系统





伯克利AMPLab明星产品



APACHE SPARK ECOSYSTEM

Spark SQL

Spark
Streaming
(Streaming)

MLlib
(Machine
learning)

GraphX
(Graph
Computation)

SparkR
(R on spark)

Apache Spark Core API

R

SQL

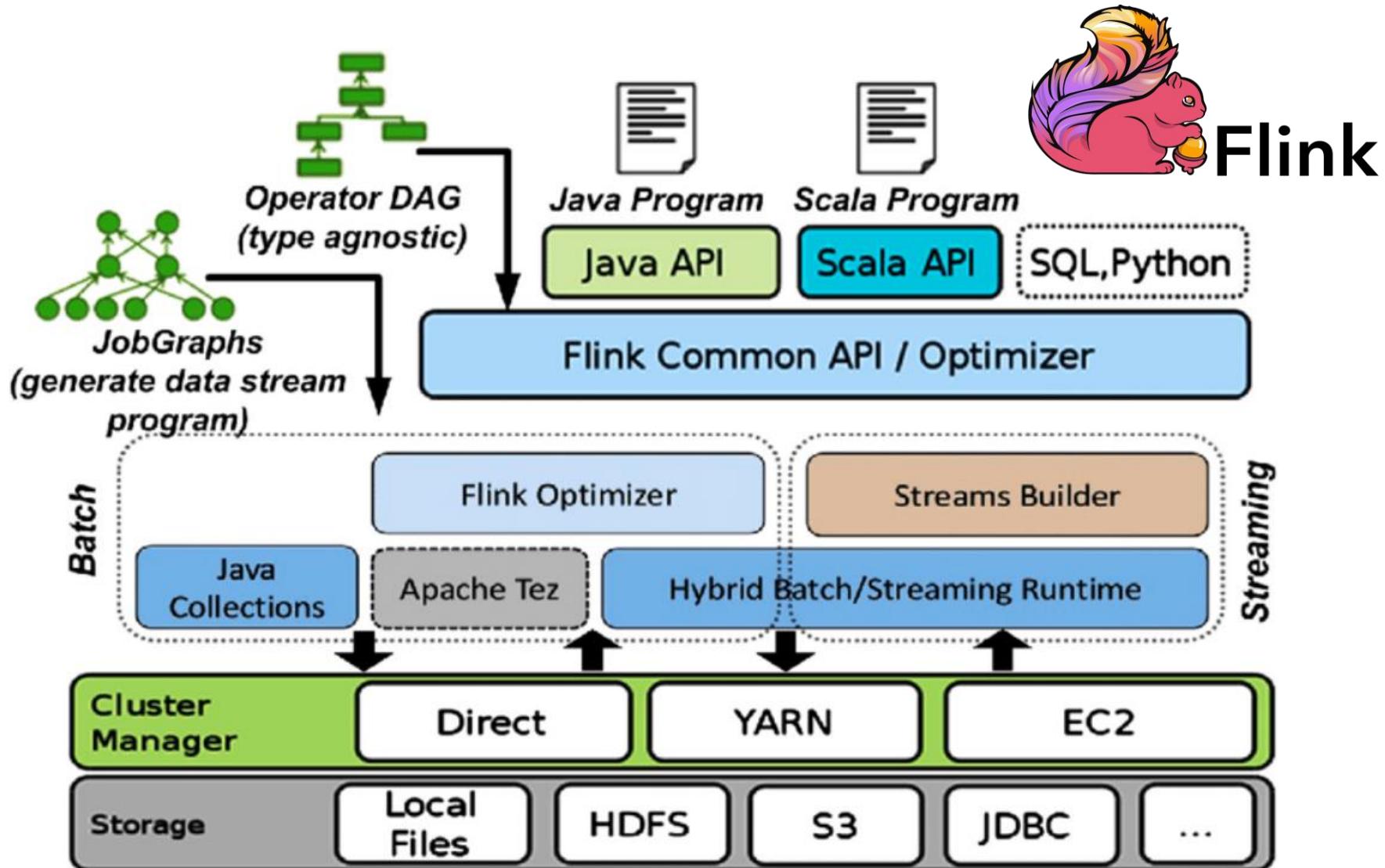
Python

Scala

Java

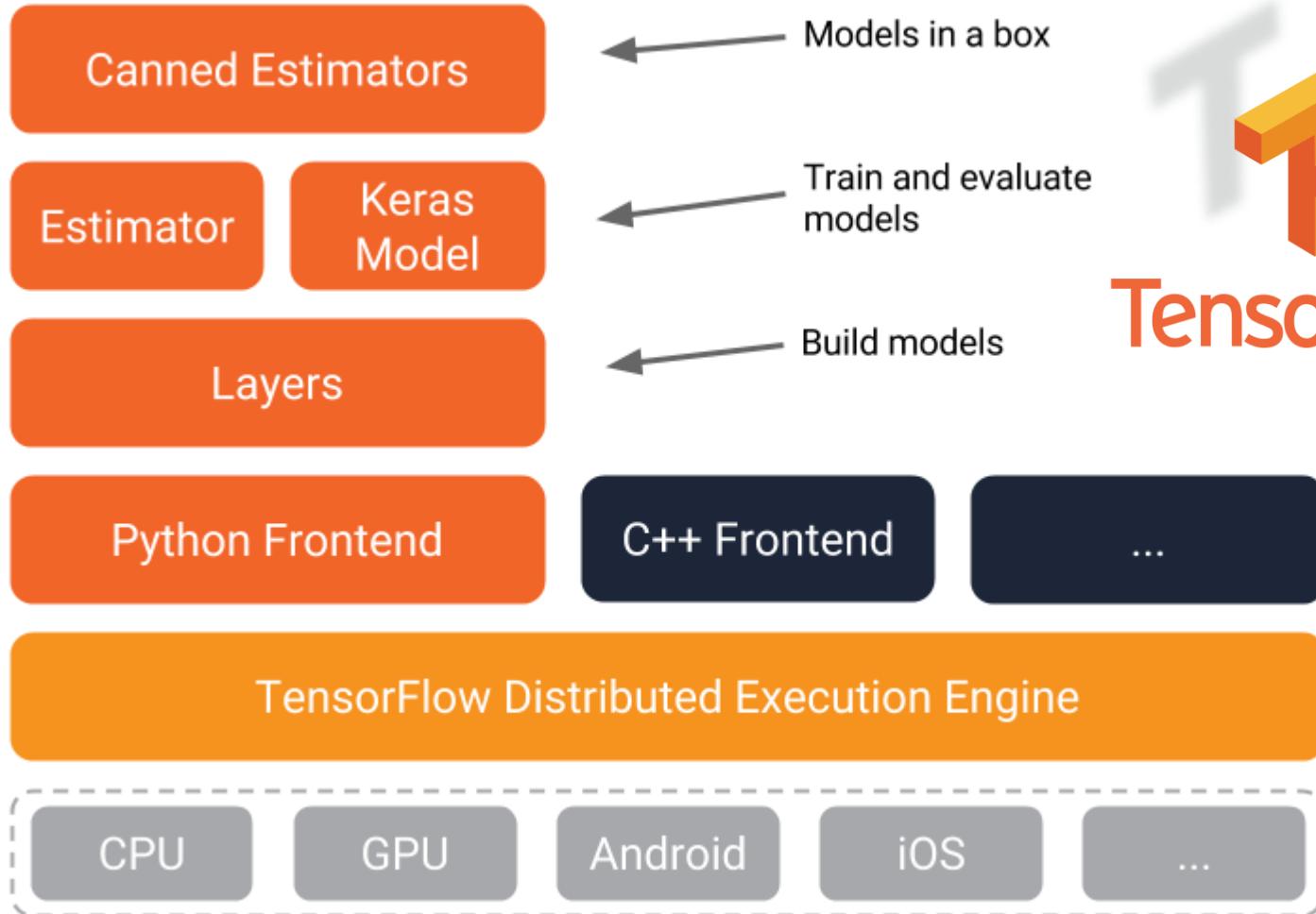


流式大数据系统:Flink





深度学习系统





异构计算引擎研究背景

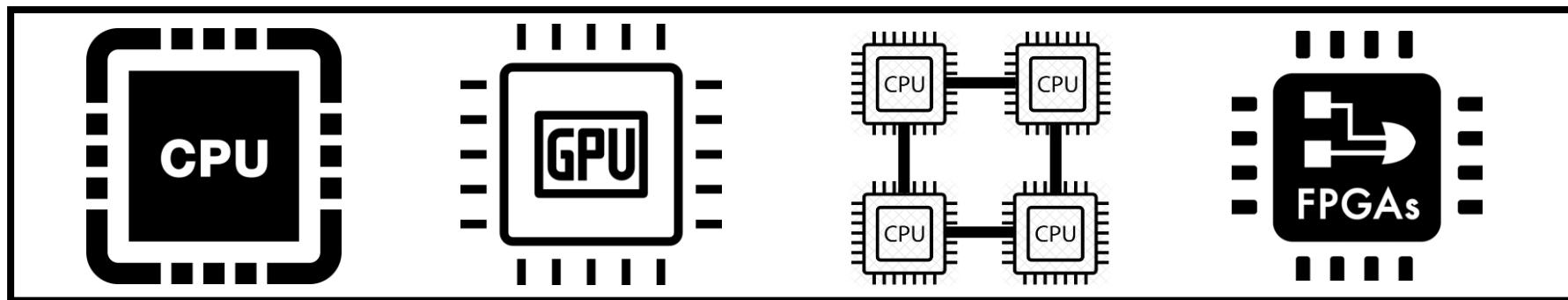


- ❖ **中央处理器:** CPU is the first-class citizen
- ❖ **摩尔定律:**
 - 集成电路上可容纳的晶体管数目，约每隔18个月便会增加一倍，性能也将提升一倍。

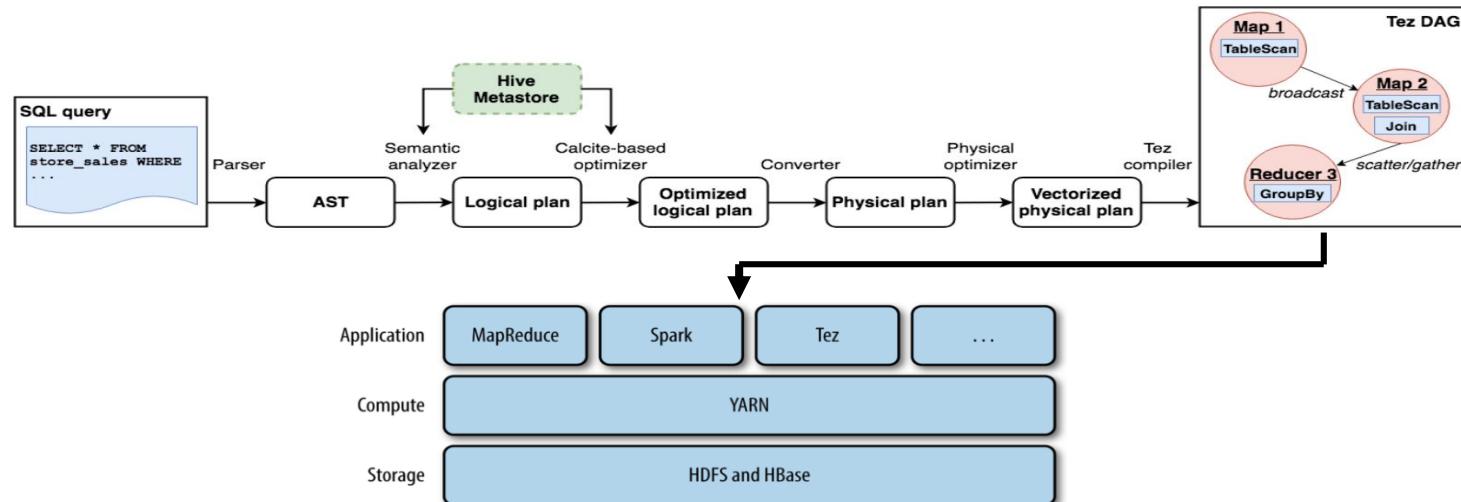


摩尔定律已经失效!

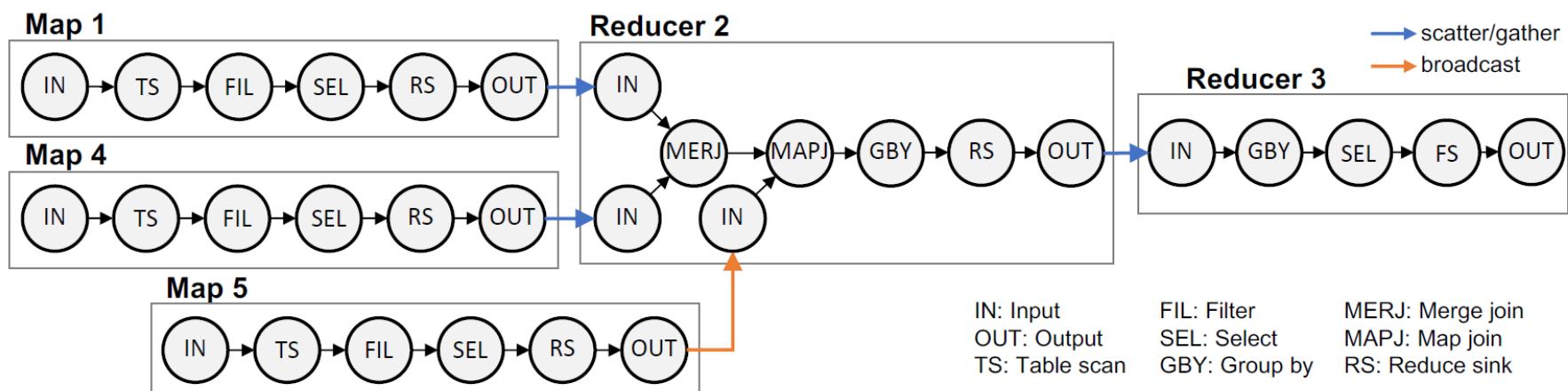
--- 英伟达CEO黄仁勋, CES2019



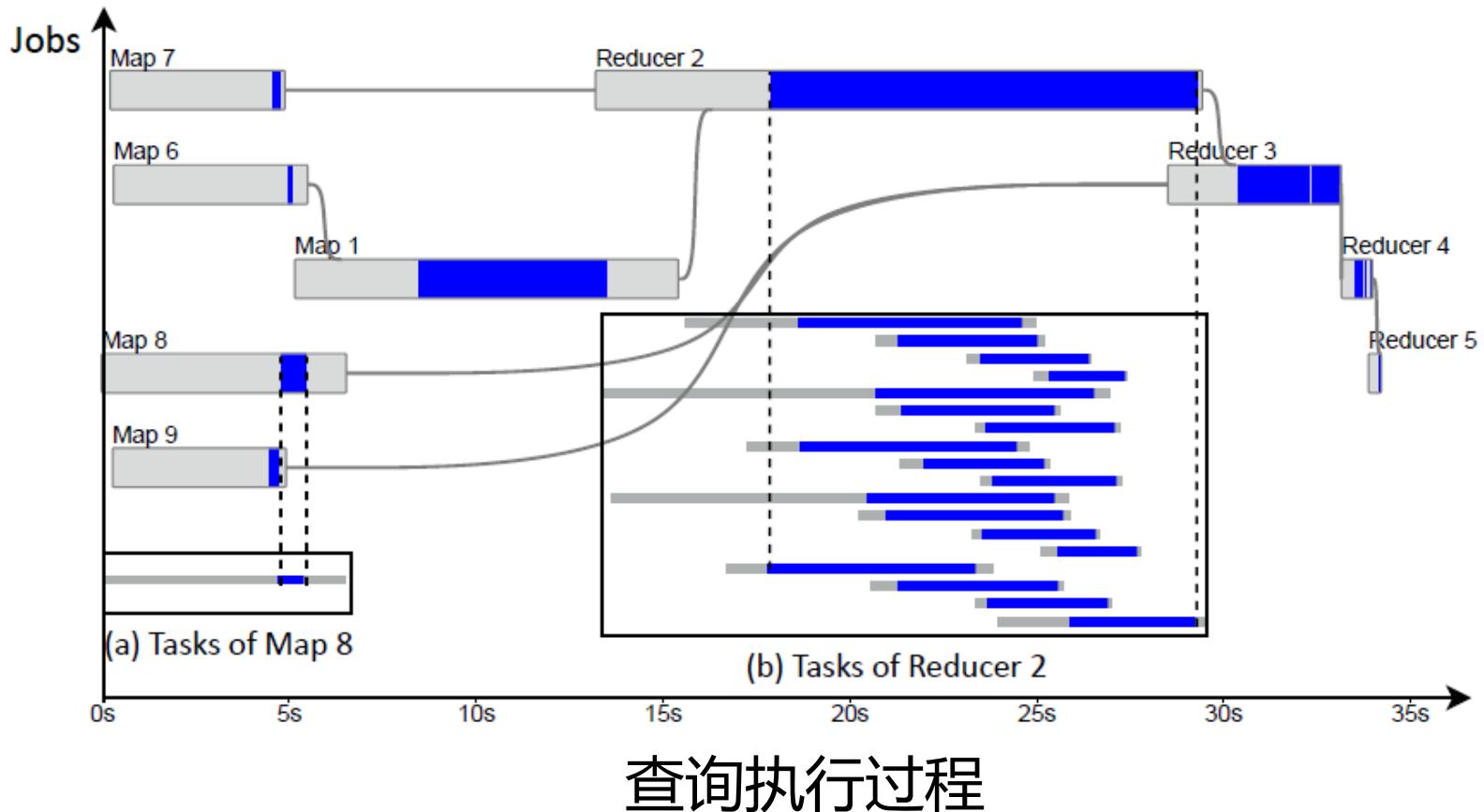
亟需架构异构计算引擎充分利用各种新型计算硬件性能！



Apache Hive上分析查询处理流程

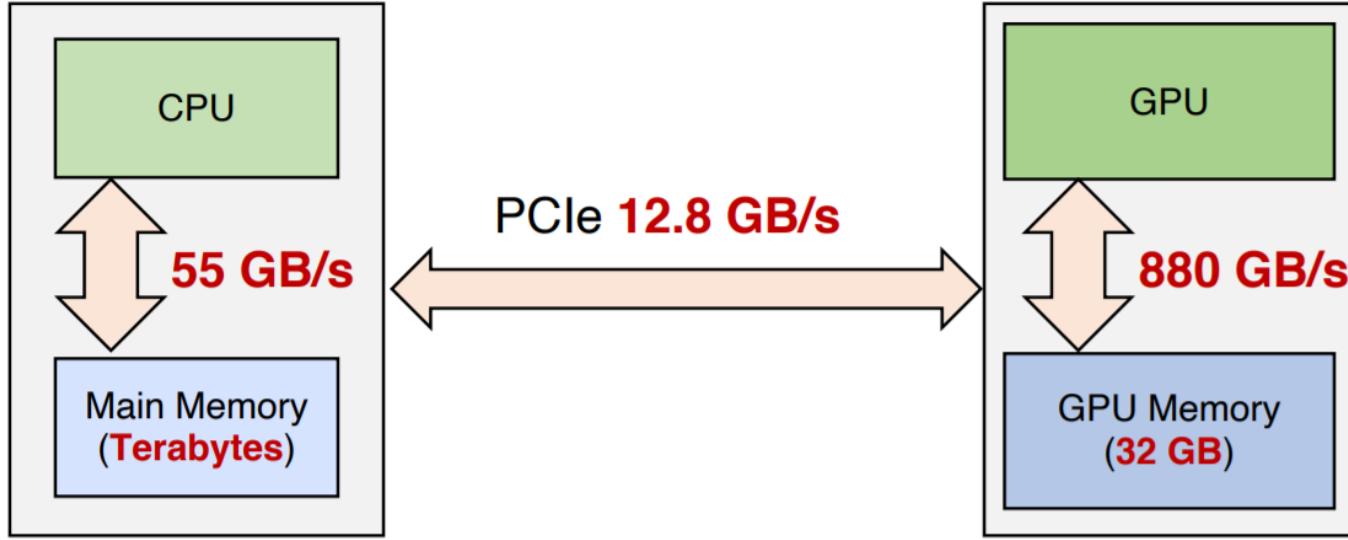


对应的MapReduce作业



查询执行过程

计算受限型 vs 内存受限型

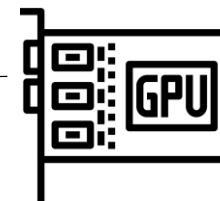
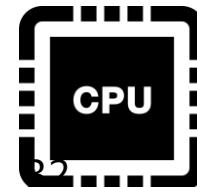
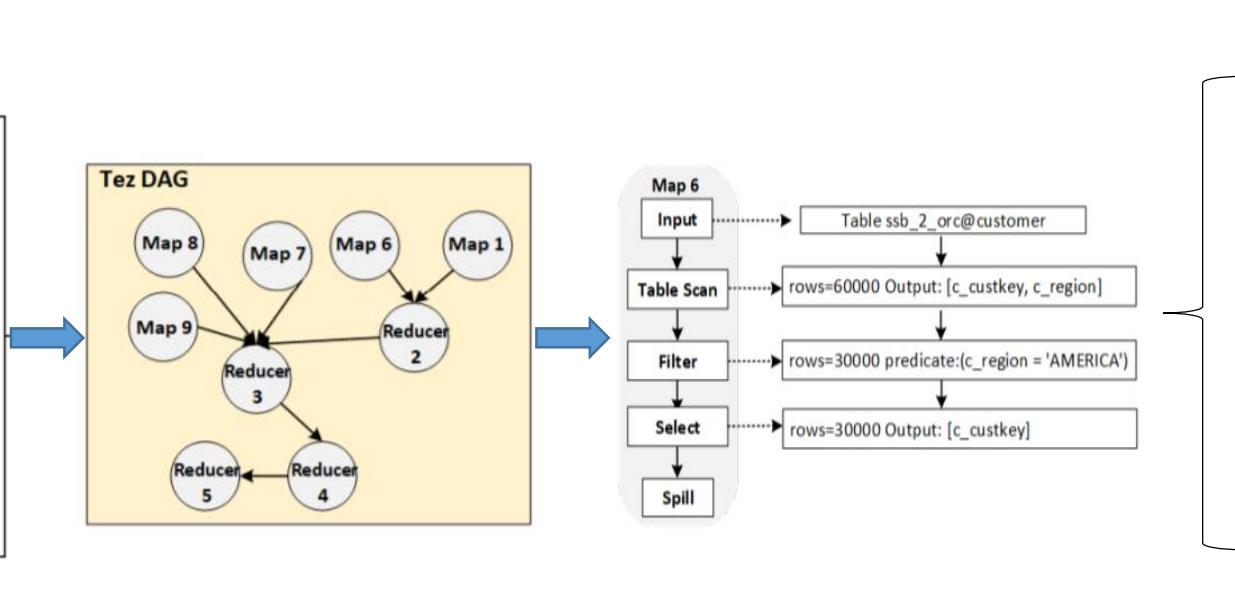


CPU vs GPU:

- GPU has **immense** computational power
- GPU memory has **high bandwidth**
- GPU memory has **small capacity**
- Loading data from main memory is **slow**

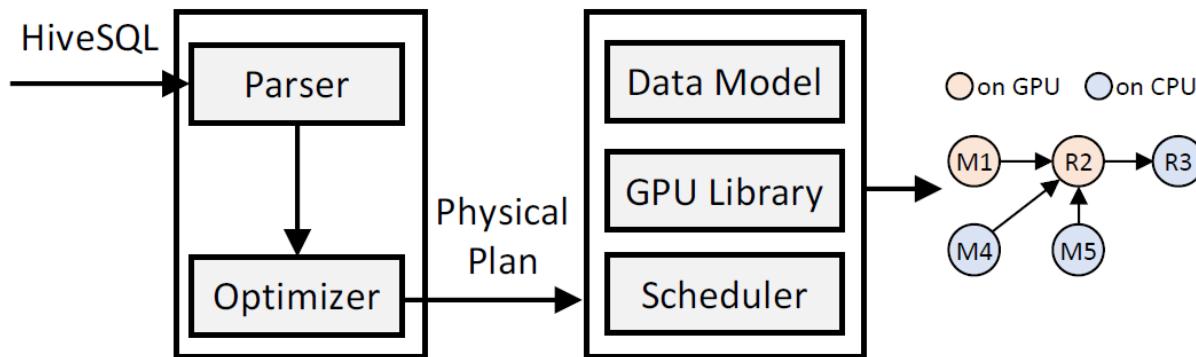
- ❖ 什么样的作业该被GPU执行?
- ❖ 如何在GPU上执行这些工作
 - ❖ 数据和执行计划如何迁移到GPU上?
 - ❖ GPU上如何执行这些算子?

```
select
    d_year, s_city, p_brand1,
    sum(lo_revenue - lo_supplycost) as profit
from
    dates, customer, supplier, part, lineorder
where
    lo_custkey = c_custkey
    and lo_suppkey = s_suppkey
    and lo_partkey = p_partkey
    and lo_orderdate = d_datekey
    and c_region = 'AMERICA'
    and s_nation = 'UNITED STATES'
    and (d_year = 1997 or d_year = 1998)
    and p_category = 'MFGR#14'
group by
    d_year, s_city, p_brand1
order by
    d_year, s_city, p_brand1;
```



GHive: 面向CPU-GPU异构硬件的新型计算引擎

系统开发历时3年, 研究成果发表于2022年SoCC / SIGMOD demo
Github开源链接: <https://github.com/DBGroup-SUSTech/GHive/>



创新点

- ❖ 【新模型】提出适应CPU-GPU异构计算的**新型数据模型gTable**
 - JAVA/C++跨语言、按需传输
- ❖ 【新技术】设计面向大数据系统的SQL查询**GPU算子库Panda**
 - 高效、易扩充、支持UDF
- ❖ 【新策略】建立算子执行开销预估模型**实现硬件感知调度**
 - 精准估计、感知调度



分享提纲



- ✓ **大数据是什么**
 - 关于大数据的三两趣事
- ✓ **大数据的研究**
 - 从研究初探到未来挑战
- ✓ **大数据的未来**
 - 让大数据赋能生产生活



下一个时代



极致的数据分析能力

Databricks

Oracle

AlayaDB

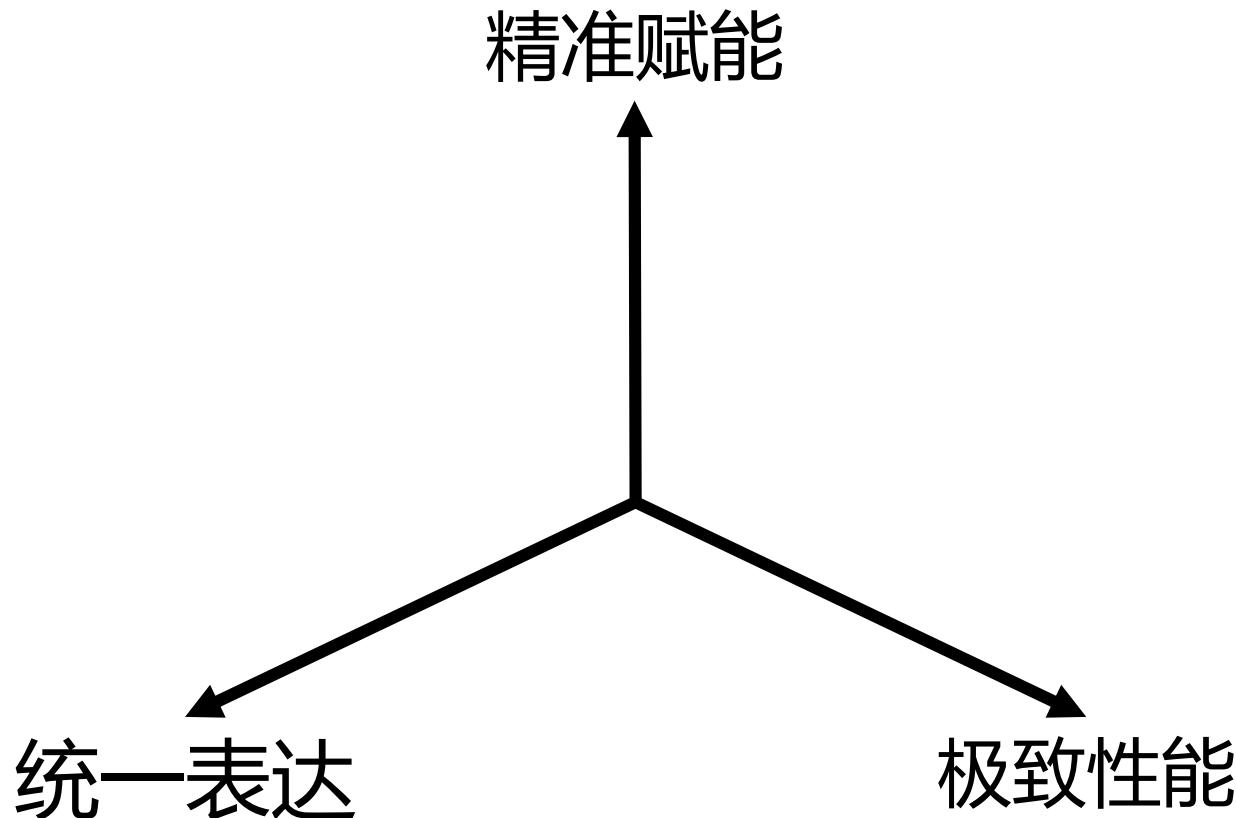
极致的数据管理能力

极致的数据生成能力

大数据时代划分



未来挑战





大数据有多大？



题西岭壁

[宋] 苏轼

横看成岭侧成峰，

远近高低各不同。

不识庐山真面目，

只缘生在此山中。



大数据有多大？



✓ 不要盲人摸象！



- ✓ 挖掘数据真实有效价值！
- ✓ 让数据赋能各行各业！





谢谢！

DBGroup @ SUSTech

Dr. Bo Tang (唐博)

tangb3@sustech.edu.cn

