



西安交通大学管理学院
THE SCHOOL OF MANAGEMENT
XI'AN JIAOTONG UNIVERSITY

第3部分——大数据相关技术及应用

Part III: Big Data Related Techniques & Applications

刘跃文 博士 Dr. LIU, Yuewen

教授、博士生导师 Professor

liuyuewen@xjtu.edu.cn

西安交通大学管理学院

School of Management, Xi'an Jiaotong University

V2.0, 2023-Oct

Topic 7: 大数据的概念

Big Data Conception and Related Techniques

刘跃文 博士 Dr. LIU, Yuewen

教授、博士生导师 Professor

liuyuewen@xjtu.edu.cn

西安交通大学管理学院

School of Management, Xi'an Jiaotong University

V2.0, 2023-Oct

- 著名的数据仓库专家Ralph Kimball:

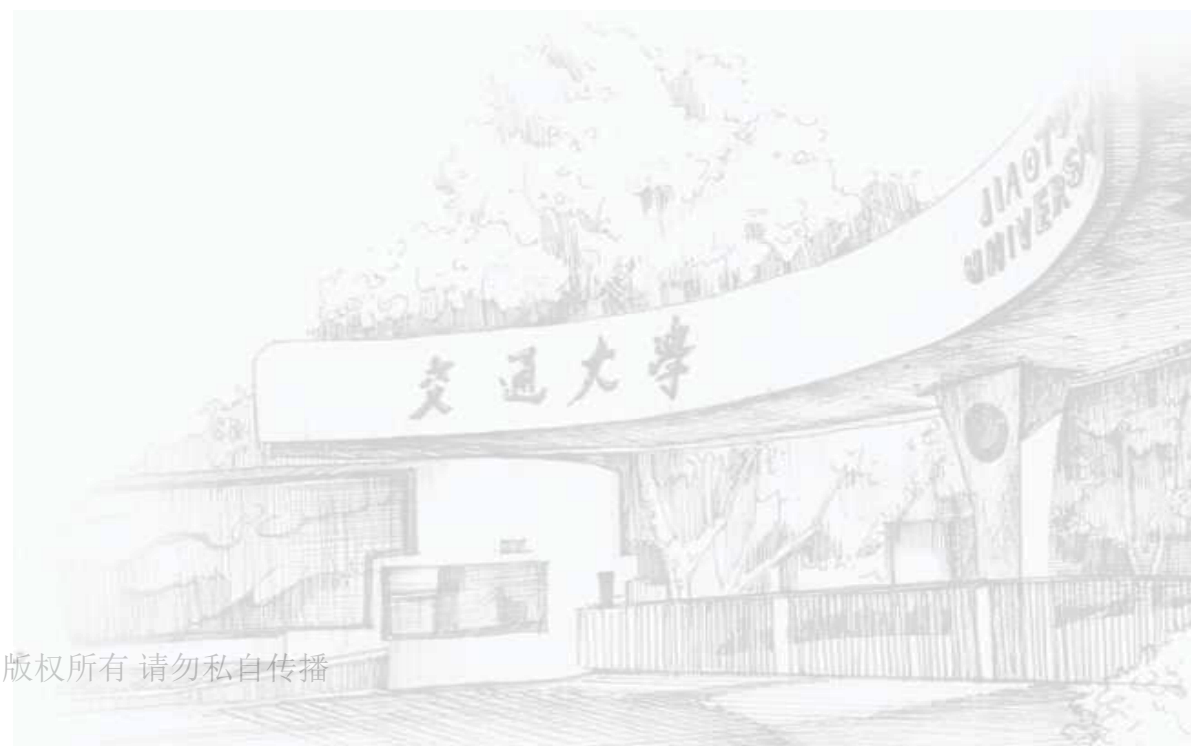
“我们花了二十多年的时间将数据放入数据库，如今是该将它们拿出来的时候了。”



提纲 Outline

1. 大数据概念的变迁
2. 分布式计算技术解决大数据问题
3. 大数据产生价值的2条路线
4. 大数据项目没有那么容易成功

1. 大数据概念的变迁

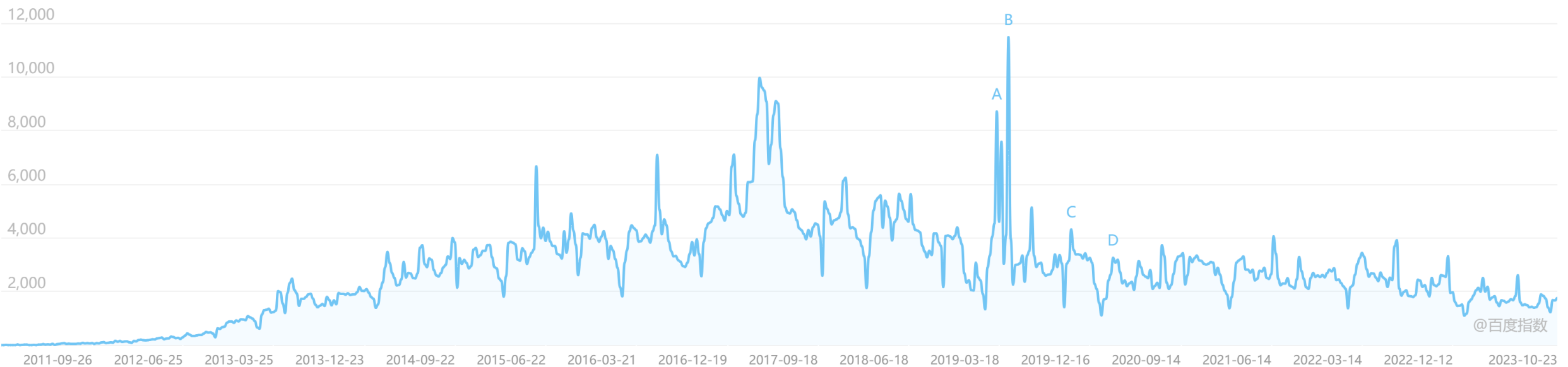


搜索指数 ?

对比时间段 | 2011-01-02 ~ 2023-10-23 | 自定义 | PC+移动 | 全国

大数据

新闻头条 平均值



Important Papers & Reports

- Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A. H. Byers (2011). Big data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute.
 - The most famous Big Data report
 - More than 150 pages
 - **Executive Summary**

奥巴马政府提出大数据规划

- U.S. Government. (2012). "Obama administration unveils "big data" initiative: Announces \$200 million in new r&d investments."
- http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf
- To make the most of this opportunity, the White House Office of Science and Technology Policy (OSTP)—in concert with several Federal departments and agencies—created the Big Data Research and Development Initiative to:
 - Advance state-of-the-art core technologies needed to collect, store, preserve, manage, analyze, and share huge quantities of data.
 - Harness these technologies to accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning
 - Expand the workforce needed to develop and use Big Data technologies.



-
- **NSF: National Science Foundation:**
 - http://www.nsf.gov/news/news_summ.jsp?cntn_id=123607
 - **HHS/NIH: National Institutes of Health – 1000 Genomes Project Data Available on Cloud:**
 - <http://www.nih.gov/news/health/mar2012/nhgri-29.htm>
 - **DOE: Department of Energy – Scientific Discovery Through Advanced Computing:**
 - <http://science.energy.gov/news/>
 - **DOD: Department of Defense – Data to Decisions:**
 - www.DefenseInnovationMarketplace.mil
 - **DARPA: Defense Advanced Research Projects Agency – the XDATA program**
 - <http://www.darpa.mil/NewsEvents/Releases/2012/03/29.aspx>
 - **USGS: US Geological Survey – Big Data for Earth System Science:**
 - <http://powellcenter.usgs.gov>

大数据总统

- 八卦： President in Big Data Era
- <http://edition.cnn.com/2012/11/07/tech/web/obama-campaign-tech-team>

How Obama's data crunchers helped him win

TIME

By Michael Scherer

November 8, 2012 -- Updated 1645 GMT (0045 HKT) | Filed under: Web



President Obama's campaign manager hired an analytics department five times as large as that of the 2008 operation.

- Nature Special Issue on Big Data: September, 2008
 - Frankel, F. and R. Reid (2008). "Big data: Distilling meaning from data." Nature 455(7209): 30-30.
- Science Special Issue on Data Analysis: February, 2011
 - King, G. (2011). "Ensuring the data-rich future of the social sciences." Science 331(6018): 719-721.

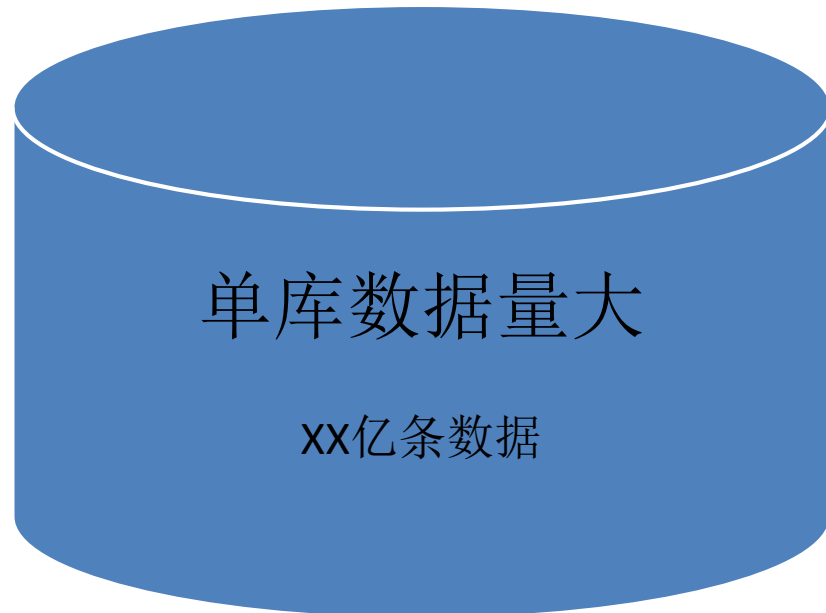


2. 早期的大数据概念

大数据 \neq 大量数据

数据体量 **大**

xx亿条数据



数据种类 **多**

xx个数据库，xx个维度



大数据 Big Data = 3V?



WHAT IS BIG DATA?

VOLUME

Large amounts of data.

VELOCITY

Needs to be analyzed quickly.






VARIETY

Different types of structured and unstructured data.

Key questions enterprises are asking about Big Data:

- How to store and protect big data?
- How to backup and restore big data?
- How to organize and catalog the data that you have backed up?
- How to keep costs low while ensuring that all the critical data is available when you need it?

WHAT ARE THE VOLUMES OF DATA THAT WE ARE SEEING TODAY?

-  **30 billion pieces of content** were added to Facebook this past month by 600 million plus users.
-  **Zynga processes 1 petabyte of content** for players every day, a volume of data that is unmatched in the social game industry.
-  **More than 2 billion videos** were watched on YouTube... yesterday.
-  **The average teenager sends 4,762 text messages** per month.
-  **32 billion searches** were performed last month... on Twitter.

WHAT DOES THE FUTURE LOOK LIKE?

Worldwide IP traffic will quadruple by 2015.

By 2015, nearly **3 billion people** will be online, pushing the data created and shared to nearly **8 zettabytes.**

HOW IS THE MARKET FOR BIG DATA SOLUTIONS EVOLVING?

A new IDC study says the market for big technology and services will grow from \$3.2 billion in 2010 to \$16.9 billion in 2015. That's a growth of 49% CAGR.

58% of respondents expect their companies to increase spending on server backup solutions and other big data-related initiatives within the next three years.

2/3rds of surveyed businesses in North America said big data will become a concern for them within the next five years.

Asigra.

3V, 4V, 6V?

Volume 量大

Velocity 更新速度快

Variety 种类多

Value 有价值

Veracity 准确性

Validity 正当性

Valence 连通性?

Visualization 可视化

早期的大数据概念

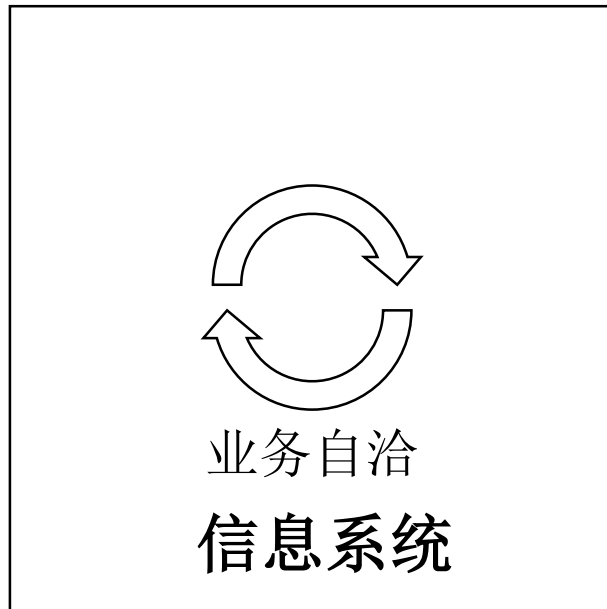
- Big data is a term used to refer to **the study and applications of data sets that are so big and complex** that traditional data-processing application software are inadequate to deal with them.
- **Big data challenges** include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy and data source.
- There are a number of concepts associated with big data: originally there were 3 concepts **volume, variety, velocity**.
- Other concepts later attributed with big data are veracity (i.e., how much noise is in the data) and value.

3. 现在的大数据概念： 数据分析挖掘价值。

- The term "big data" tends to refer to **the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data**, and seldom to a particular size of data set.
- There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem. Analysis of data sets can find new correlations to spot business trends, prevent diseases, combat crime and so on.
- Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data-sets in areas including Internet search, fintech, urban informatics, and business informatics.

信息化到大数据的发展历程

1.0 信息化时代



2.0 数据汇聚时代

数据中心

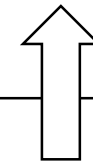


数据输出



3.0 数据赋能时代

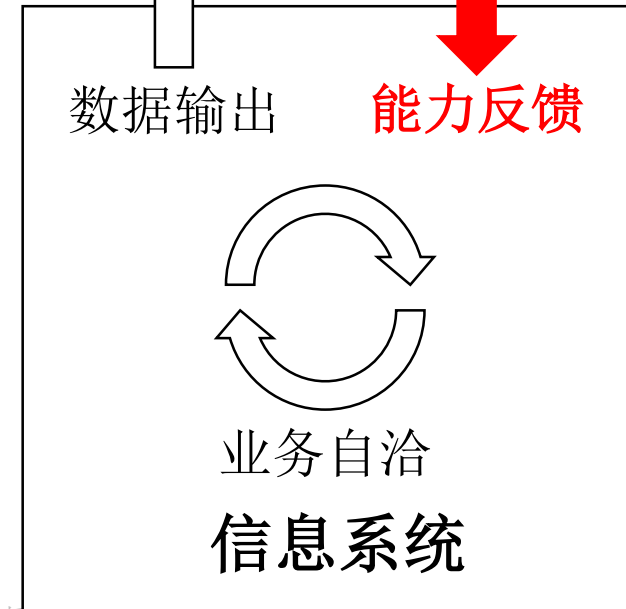
数据中心



数据输出



能力反馈

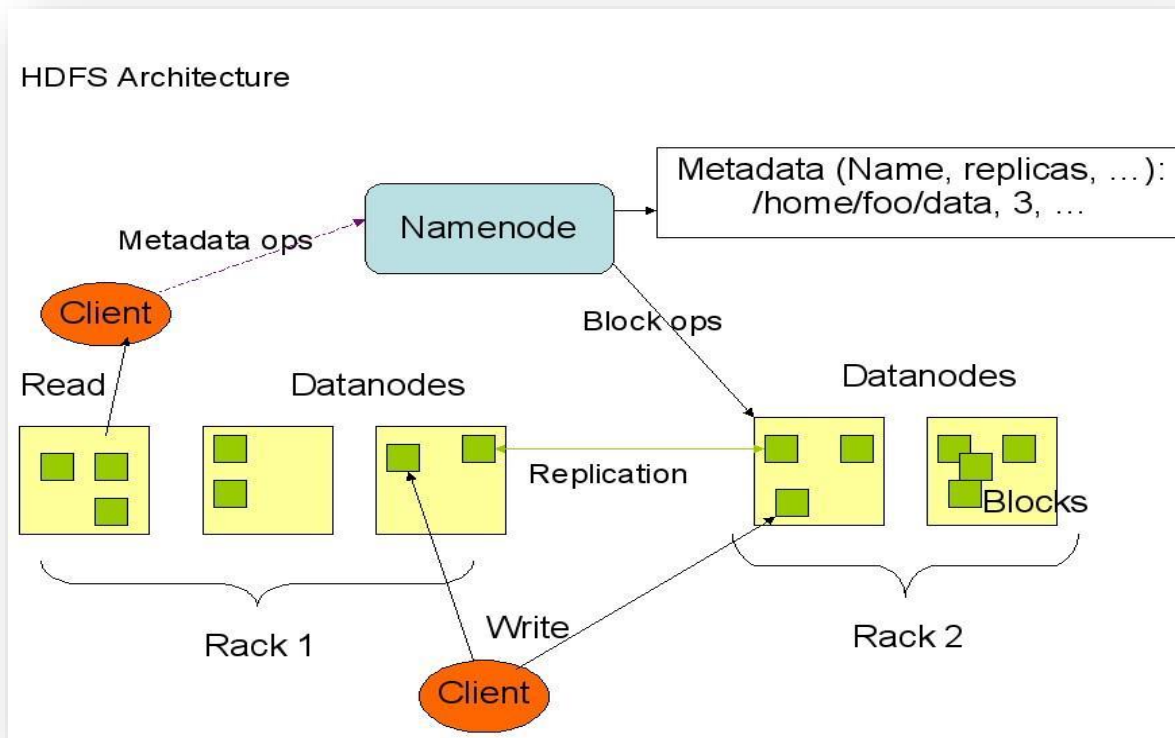


2. 分布式存储与计算系统

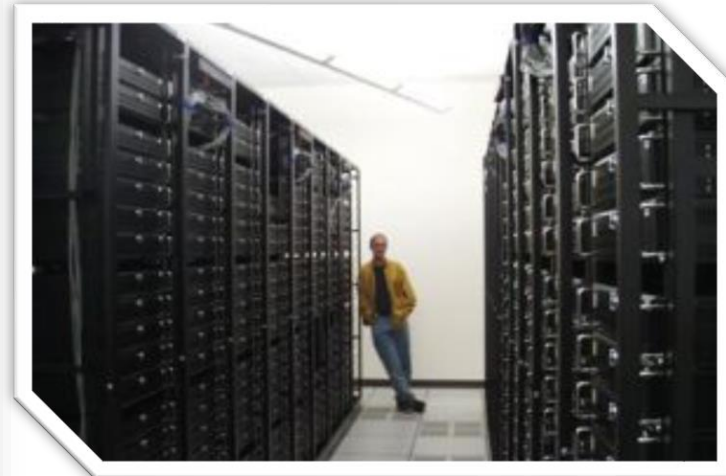


1. 分布式系统

- 分而治之：**Divide and Conquer**
- 冗余、容灾、可扩展

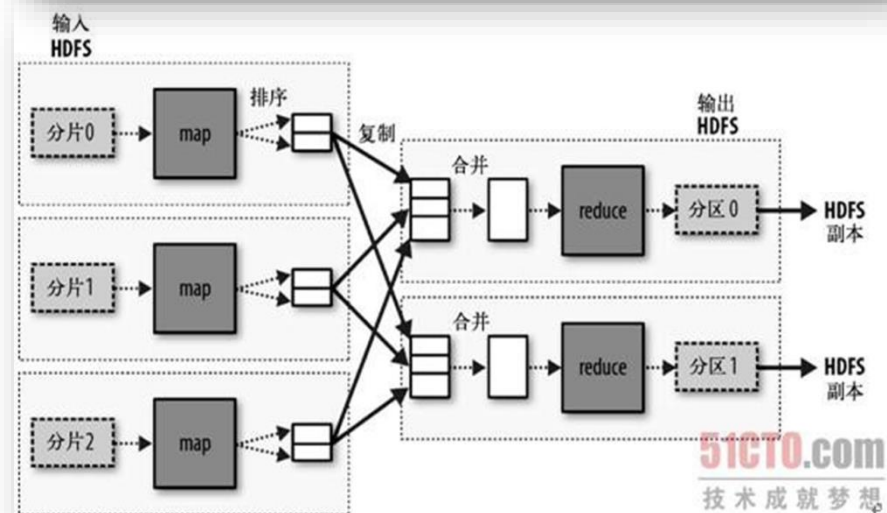
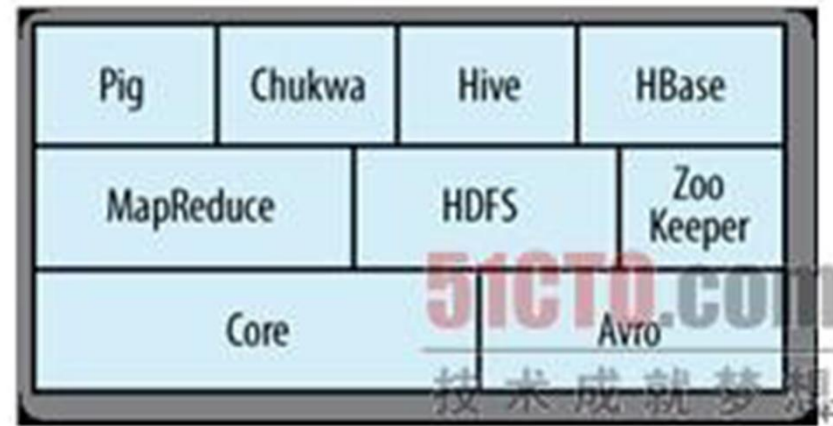


Google MapReduce
架构设计师
Jeffrey Dean



分布式数据处理技术

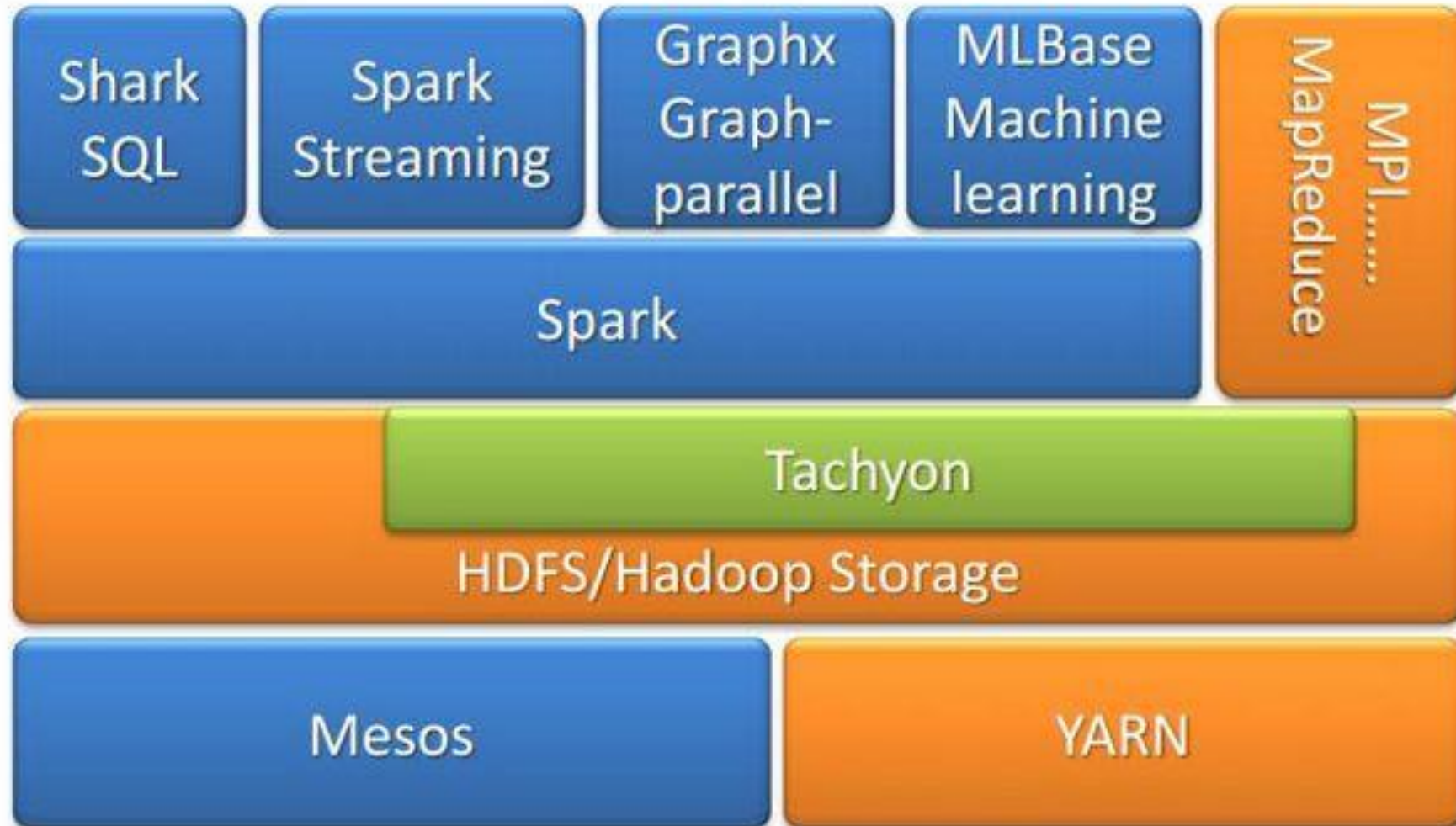
- Hadoop
- MapReduce
- 如何利用Hadoop对海量数据进行优化处理是Yahoo正在致力于工作的内容。以网络分析为例，Yahoo目前有超过100亿个网页，1PB的网页数据内容，2万亿条链接，每日面临这300TB的数据输出。“在应用Hadoop前，实施这一过程我们大概需要1个月的时间，但应用后仅需要1周时间”



2. Hadoop生态体系



3. Spark生态体系



4. 云平台与大数据平台

云平台：管理全部计算机

裸金属
服务器

虚拟机
服务器

大数据套件
服务器

Hive

HBase

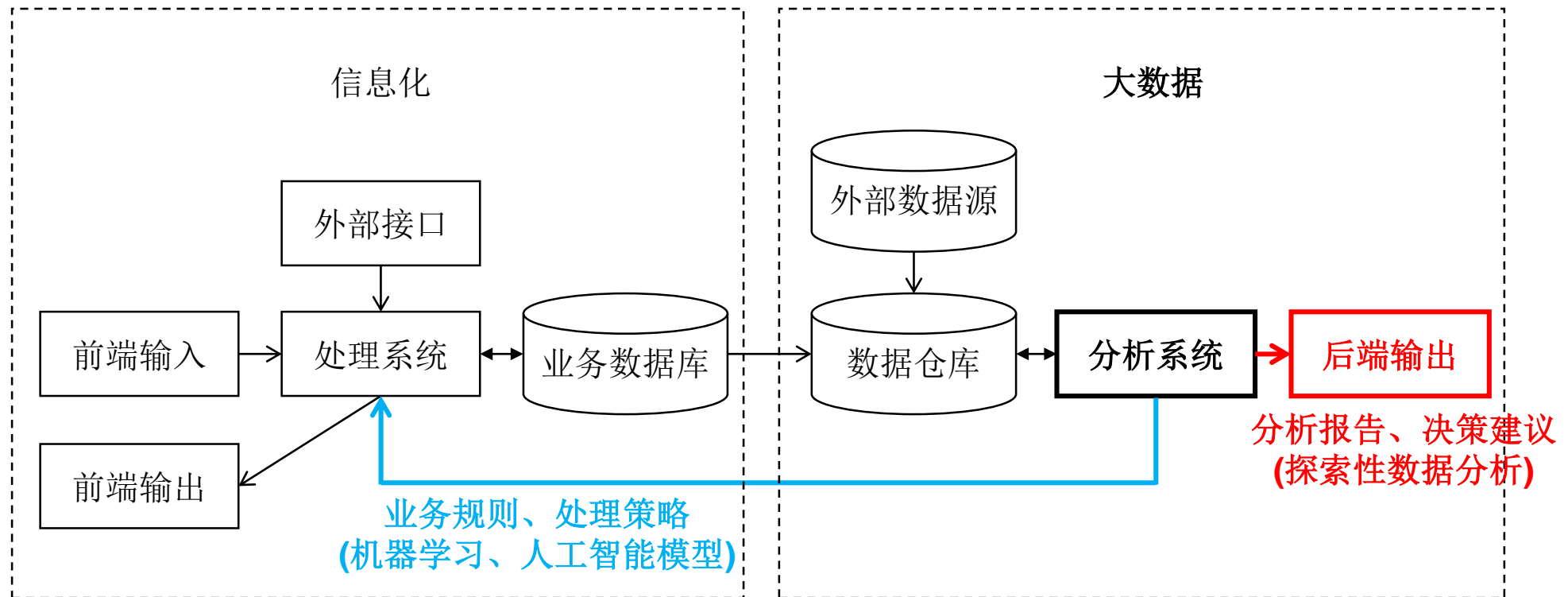
GraphX

.....

3. 大数据分析的两条路线：探索性分析与机器学习



1. 大数据的两条路线



2. 案例：东航的安全运行研究院

- 案例来源：https://www.sohu.com/a/412449304_115926
- 东航安全运行研究院的成立，必须承认东航的高层管理者看得很远。领导者所考虑的是，十年、甚至二十年之后航空公司将面临怎样的威胁？新的机遇又在哪儿？没有人能保证公司几十年下来还依然存在，尤其民航业，是变化极为剧烈的一个行业。就算现在有利润，一旦外部环境发生变化，我们该靠怎样的核心竞争力去生存？
- 要赶上世界民航的发展趋势，必须有一个专门的‘情报小组’。对于东航来说，研究院就扮演着‘情报小组’的角色：不断解读新信息、引入新方法、新理论、新思路、新技术，为公司发展添柴加油。



- 现在东航拥有一万一千多名飞行员，未来还将继续增长，这要求我们的管理方式一定要产生颠覆性的变化。譬如，以前我们很强调SOP，一抓SOP，事故率立马下降；之后，我们开始做人因管理、文化管理，事故率又降了一些；而现在，公众对于飞行安全要求更高了，我们该怎么继续改善事故率？
- 现在的飞机已经相当先进了，在飞行中会产生大量的数据，都积累在飞行部、运营部等生产部门。之前我们不用他，这些数据就是‘死的’，但这其实是一个大‘宝藏’，靠这个，我们还能从飞行安全这个“柠檬”里再榨点汁出来，让安全表现更好一些。

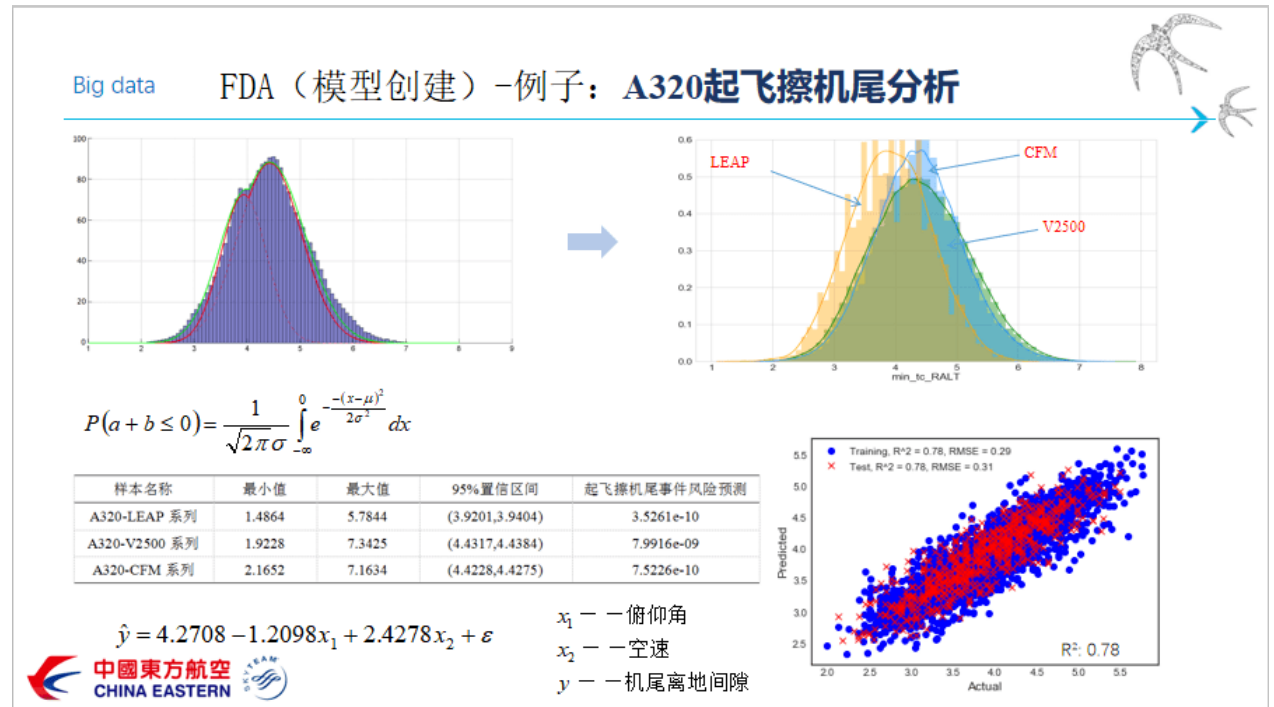


- 了解现状：
- 2015年，中国民航发生了一起飞机在高原机场降落时爆胎的事故，我们了解情况后，就用三年的数据回滚，研究东航机队在高原运行的情况。一位研究员做了一组分析起飞数据的PPT，我看到其中一页的图示后发现，飞机只用了不到跑道的一半就起飞了，产生了“跑道浪费”——而一般来说，我们会在跑道的一半多一点起飞。这意味着飞机发动机“劲儿”用大了。但发动机是有寿命的，长期这样使用对其损耗是很大的。
- 于是接下来，研究院将各个分公司、各个机场的数据分类汇总，再进行分析。我们发现**其中一个分公司的发动机推力用得比别的分公司都要猛一些。**

-
- 摒除了其他可能有关的因素之后，我们通过公司领导召集了一次会议，大家讨论问题原因在哪儿。
 - 这个分公司表示，他们减推力减得少，是因为他们飞机性能表中假设温度最高只有50度，而其他分公司，有的是60度，有的是70度。
 - 也就是说，就在这一项上，东航的各个分公司拿到的数据完全不一样，但每个分公司都默认自己肯定和其他公司都一样。发现了各个分公司性能表的差异后，有关部门迅速处理了这个问题，补上了漏洞。

- 识别异常：
- 之前大家都只用QAR告警的情况来说明飞行品质，这个告警的本质是发现飞行数据中的‘奇点’，之后大量的数据就被浪费掉了。而我們希望能用大数据的思维来做飞行品质分析。
- 以A320擦机尾专项分析为例：对于A320来说落地姿态不能超过 11° 。譬如飞行员A的飞行角度日常是 10° ，但A的标准差很小，也就是说很稳定，那么其实QAR就不应该告警A；而飞行员B的落地姿态一会儿是 5° ，一会儿是 10° ，虽然在要求内，但B不稳定，这才应该是注意的问题。

- 在做飞行的品质分析的时候还有个意外的收获——通过大数据，我们发现一架A320的飞行姿态不正常：其他飞机平均飞两度半，这架飞机平均飞六度。
- 虽然没有造成任何事故或者事故征候，但这肯定有问题。后来东航工程部和空客联合起来，发现是一次大修中襟翼安全角度不对，之后空客重新装配了襟翼，解决了这个问题。



节约 ¥7500万

- 预测未来（分析报告、决策建议）：
- 东航某分公司向某发动机OEM选购发动机，这个OEM提供了两种推力的发动机供选择，两型发动机的价格不一样，大推力的发动机更贵一些。
- 我们基于航线、机型、机场等多方面的大数据分析，在两天内就出具了一份报告，说明较小推力的发动机已经足够用，被公司采纳了。在这一笔交易上，研究院就为公司节省了7500万余元人民币。
- 以数据为基础的决策，能够为航司指明更加安全、高效的方向，毕竟“数字从不说谎”。

3. 案例：UPS 不许左转弯

- UPS
 - 快递业务：
 - 10万辆运营车辆；自有飞机237架，租用412架
 - 43.5万名员工
 - 供应链及货运业务：
 - 5599辆货运卡车；
 - 19884辆拖车
- 识别异常：
 - UPS安全带扣上比例：**98%**
 - 目标：**99.8%**
 - 如何达成？



不要左转弯

- **2001年**配备了更好的追踪系统后，快递商们好好研究了下卡车在运送包裹时候的表现。作为一家有着96000辆卡车，几百架飞机的物流公司，UPS急需优化一系列问题来提升业绩，如，减少汽油使用，节省时间，更高效的利用空间。(UPS停车场里的车都停的后视镜靠后视镜，以节省空间。)
- UPS为了使总部能在车辆出现晚点的时候跟踪到车辆的位置和预防引擎故障，它的货车上装有**传感器、无线适配器和GPS**。

- 发现规律（业务规则、执行策略）：
- UPS工程师们发现，左转弯是高效率的主要妨碍。与车流对着干，会导致在左转弯区长时间停留，浪费时间和汽油，并且会导致一系列事故。通过绘制了一系列右回环路线，UPS既提高了利润和安全性，又很好的宣传了它的口号和环境友好政策。
- **2004年，UPS对它的司机们宣布：到达任何目的地的正确方法是避免左转弯。**即便像是这个路线，一名UPS司机对一个不信的记着说：
 - “我们要在135大街上右转到西大街，然后再右转到139大街。在139大街上右转，走过一个街区后再右转。”
- 自从**UPS使用软件规划路线**后，它能让司机在交通繁忙路段右转，而在方便快捷的时候破例让他们左转。

节约7000万人民币

- UPS的过程管理总监杰克·莱维斯（Jack Levis）认为这个分析项目效果显著。**2011年，UPS的驾驶员们少跑了近4828万公里的路程，节省了300万加仑的燃料并且减少了3万公吨的二氧化碳排放量。**
- 2012年，右转规则和其它改进——出于各种因素考虑，UPS不说——节省了大约一千万加仑汽油，相当于减少了**7000**辆汽车一年的排放量。



4. 大数据分析的“十六字箴言”

Keywords of Big Data Analytics



4. 大数据项目没有那么容易成功

1. 对大数据认识的误区

- **与我无关型：**

- 我们是传统农业/工业/制造业，大数据/人工智能和我们没什么关系。
- 大数据/人工智能是高科技，主要是互联网企业在做，和我无关。

- **过高期望型：**

- 有了大数据/人工智能，就可以大量地替代人工/大幅度降低成本。
- 搭建一套大数据系统，我们公司/组织就从此进入了大数据时代。
- 一套最先进的大数据系统能够让我们公司/组织的各个方面都大幅度提升。
- 大数据/人工智能可以自动发现各种市场机会，帮我们做决策，大幅度提高我们的盈利能力和水平。

• 花钱搞定型：

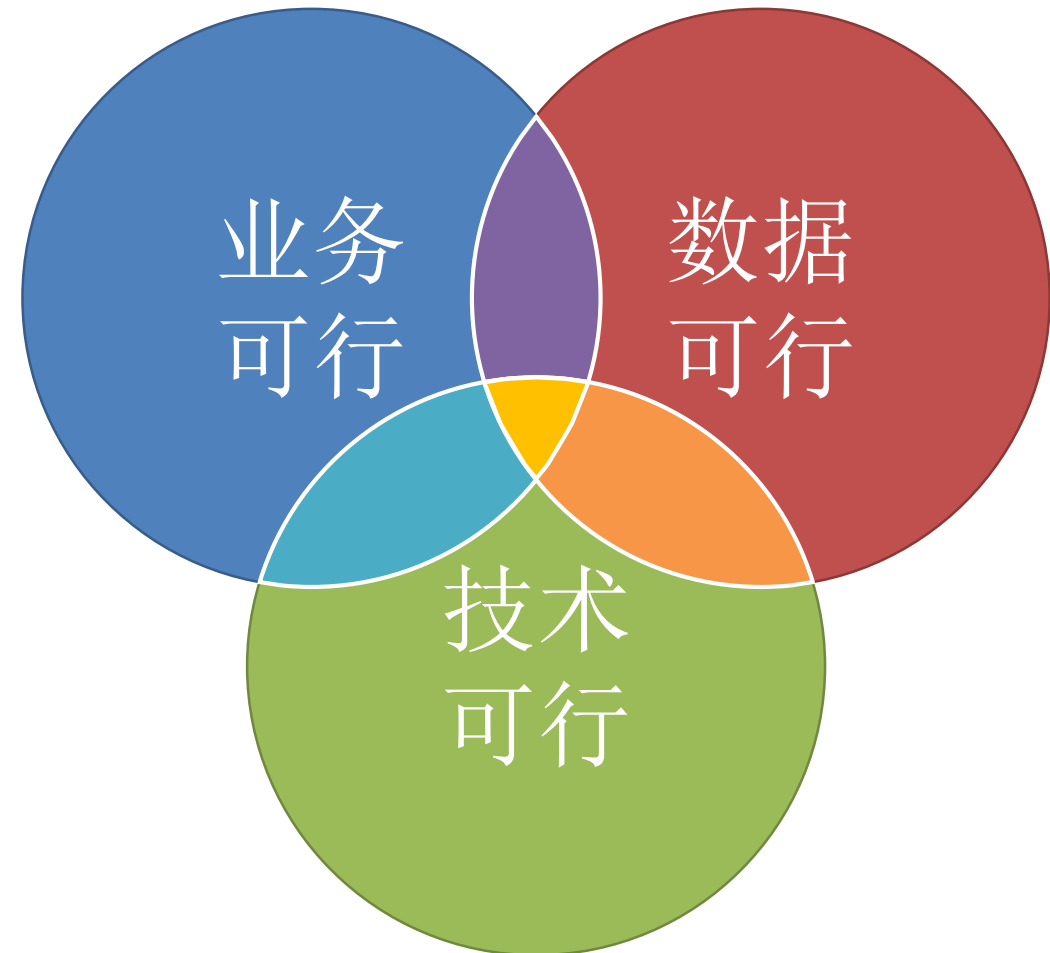
- 我们有很多数据，能否帮我们做一些大数据模型？
- 找知名的大数据公司/专家，就能做出水平很高的大数据系统，给我们公司/组织各方面带来巨大价值。
- 大数据/人工智能是高科技，我们公司人员素质不行，要聘请高水平专家来完成。

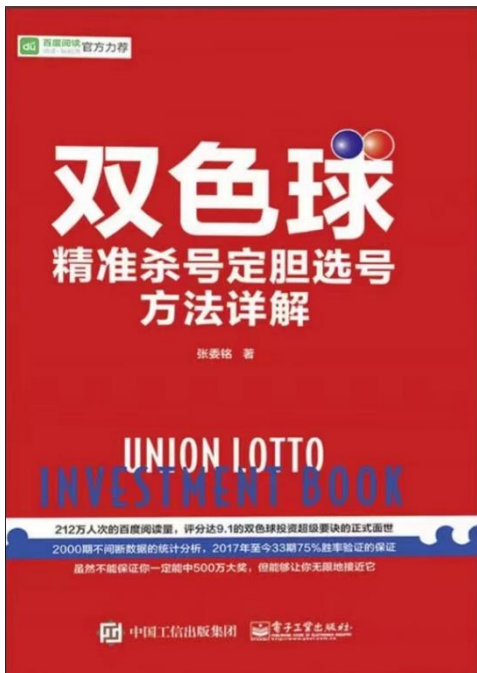
• 积累数据型：

- 现在我们的数据量积累还不够，还没办法建设大数据系统。
- 我们制定了数字化转型战略，从现在开始花3年时间做数字化转型，采集大量数据；到3年后，我们的数据就会成为巨大的资产，产生巨大的价值。
- 我们建设了数字孪生系统，能够产生大量数据。

2. 大数据并不是万能的！

- 真正可以付诸大数据/人工智能实践，并产生商业价值的部分，只有业务、数据、技术均可行的区域。
 - 业务本身具有一些客观存在的较为简单的规律尚待发掘。
 - 存在相匹配业务的数据，能够满足发现规律的需要。
 - 已有的大数据/数据分析/人工智能技术能达到要求。
- **选题比解题更困难。**





格,但有些彩票类书籍通篇都是表格,要么是统计数据,要么是电脑程序,有的还将历年开奖数据附上以扩充篇幅,这纯粹是浪费资源,也没什么意义。本书除保留一些必要的表格之外,大部分内容都属于举例、分析、验证、归纳、说明等。本书中表格也占有一定篇幅,但绝对不像有些彩票类书籍那样占有大量篇幅甚至占有90%以上的篇幅。一句话,本书有实实在在的内容和方法。

另外,本书也没有高深的数学理论、易学知识,没有任何电脑程序,更没有晦涩难懂的学说,也不会引入复杂的数学模型和假想实验。本书只有具体、明确而又易于掌握、便于操作的杀号、定胆和选号方法与技巧。一句话,本书让每个人都能看得懂。

本书简介
本书共分8章,主要是通过双色球2000多期历史开奖数据进行统计,找出一些经得起检验的杀号、定胆、选号方法与技巧。

第1章 彩票研究与分析方法入门。主要内容包括:彩票起源与规则、彩票投注方法6大注意事项、彩票是否能够预测。

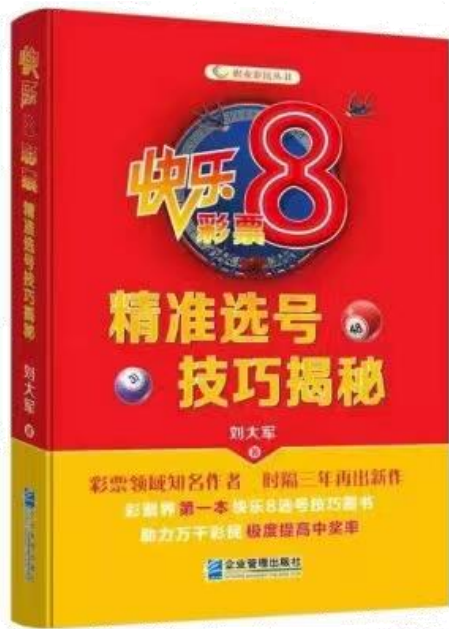
第2章 双色球投注技巧分析。本章主要结合双色球游戏规则,重点从投注方式、奖级设置和中奖规则等方面为大家详细解析双色球,帮助大家更深入地了解双色球。

第3章 前区精准杀号方法。本章对双色球前区860多种杀号方法,依据双色球从2003001期至2016105期共2004期的开奖数据,进行了系统而又精确的统计、整理、测试、对比和分析,得出了双色球前区各种杀号方法的胜率,并对其胜率较高的杀号方法进行了50期(从第2016106期至2017002期)的验证,找出了最有效、最经得起检验的双色球前区杀号方法。

第4章 后区精准杀号方法。本章对双色球后区630多种杀号方法,依据双色球从2003001期至2016105期共2004期的开奖数据,进行了系统而又精确的统计、整理、测试、对比和分析,得出了双色球后区各种杀号方法的胜率,并对其胜率较高的杀号方法进行了50期(从第2016106期至2017002期)的验证,找出了最有效、最经得起检验的双色球后区杀号方法。

第5章 前区精准定胆方法。本章对胆码的概念、定胆的意义、定胆成功率等进行了说明,重点介绍了定一个胆码、定两个胆码的方法,对定多个胆码的方法进行了特别解释,并对前区两个号码伴生(同时出现)现象进行了深入研究,为读者朋友们找出了最有效、最精准的定胆方法。

第6章 前区技术指标详解。本章对网上流传的技术指标进行了详细解析,包



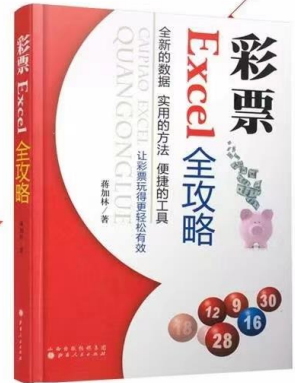
Excel 彩票全攻略

最新技术指南

01 全新的数据
CHAPTER ONE

02 实用的方法
CHAPTER TWO

03 便捷的工具
CHAPTER THREE



本书特色



电脑彩票
EXCEL
基础知识

开奖号码
数理统计

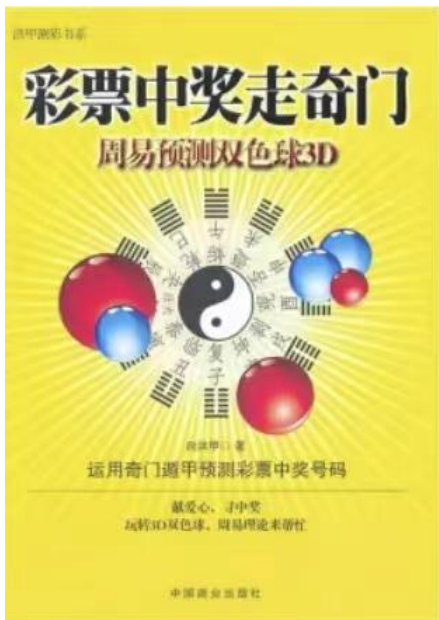
选择号码
旋转矩阵

号码组合
筛选过滤

基本信息

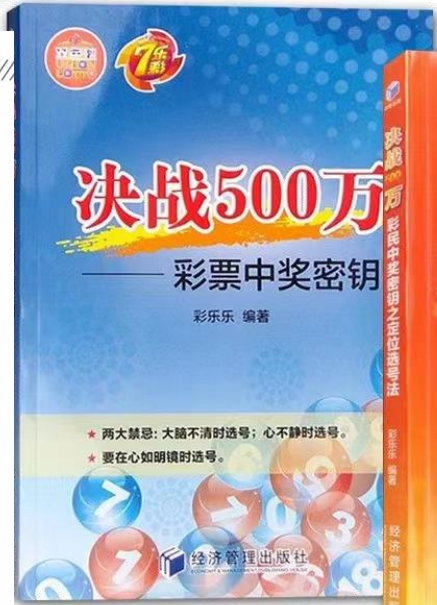
产品名称:彩票EXCEL全攻略
是否是套装:否
定价:38.00元
出版社名称:山西人民出版社
出版时间:2015年1月

作者:蒋加林
开本:16开
ISBN编号:9787203086697
产品市场分析



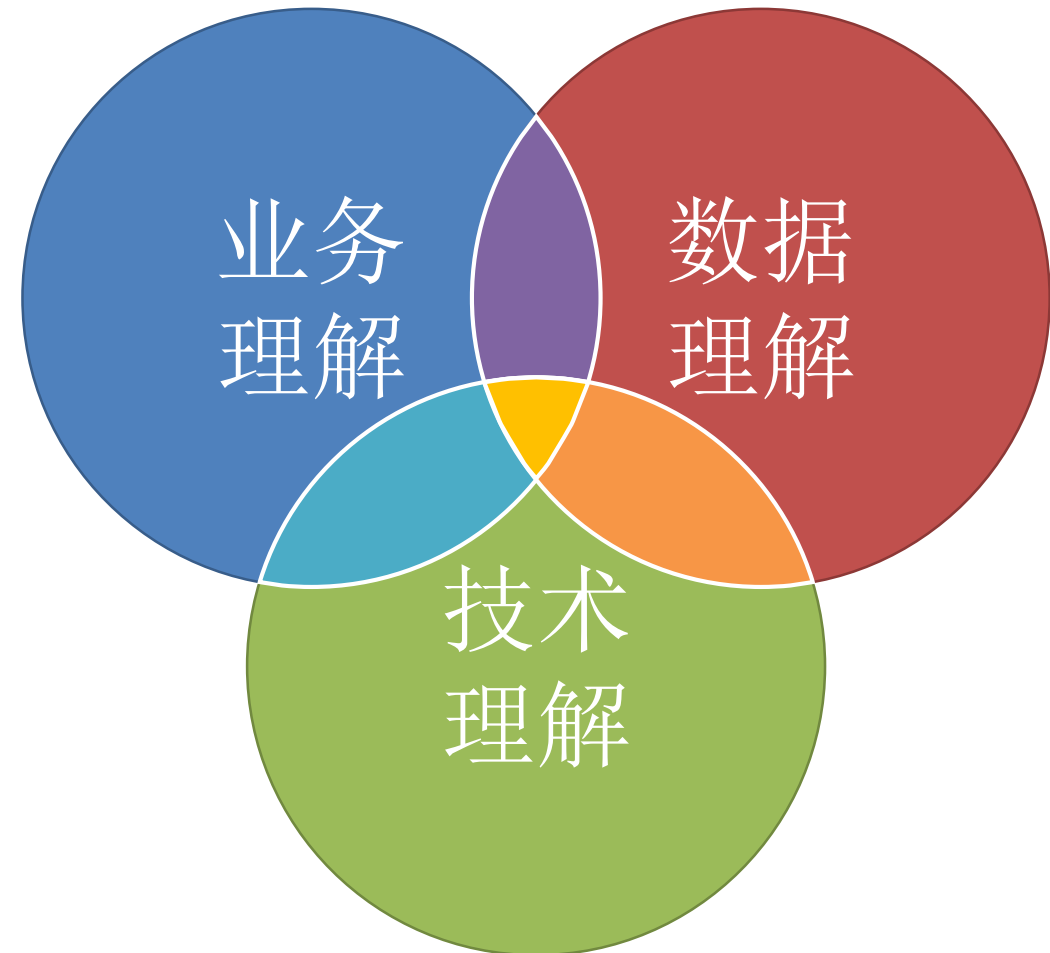
目录

- 第一章 奇门盘中的彩票中奖号码
 - 第一节 奇门遁甲让我中奖了
 - 第二节 如何用奇门遁甲排列彩票号码
- 第二章 奇门遁基础知识
 - 第一节 什么是奇门遁甲
 - 第二节 天干地支
 - 第三节 打开奇门遁甲的钥匙
- 第三章 如何用奇门遁甲预测彩票中奖号码
 - 第一节 奇门遁甲的排盘方法
 - 第二节 奇门遁甲吉凶格局的启示
 - 第三节 奇门遁甲预测彩票中奖号码的方法
- 第四章 奇门遁甲预测彩票中奖号码案例分析
 - 第一节 双色球中奖号码案例分析
 - 第二节 福利彩票3D中奖号码案例分析

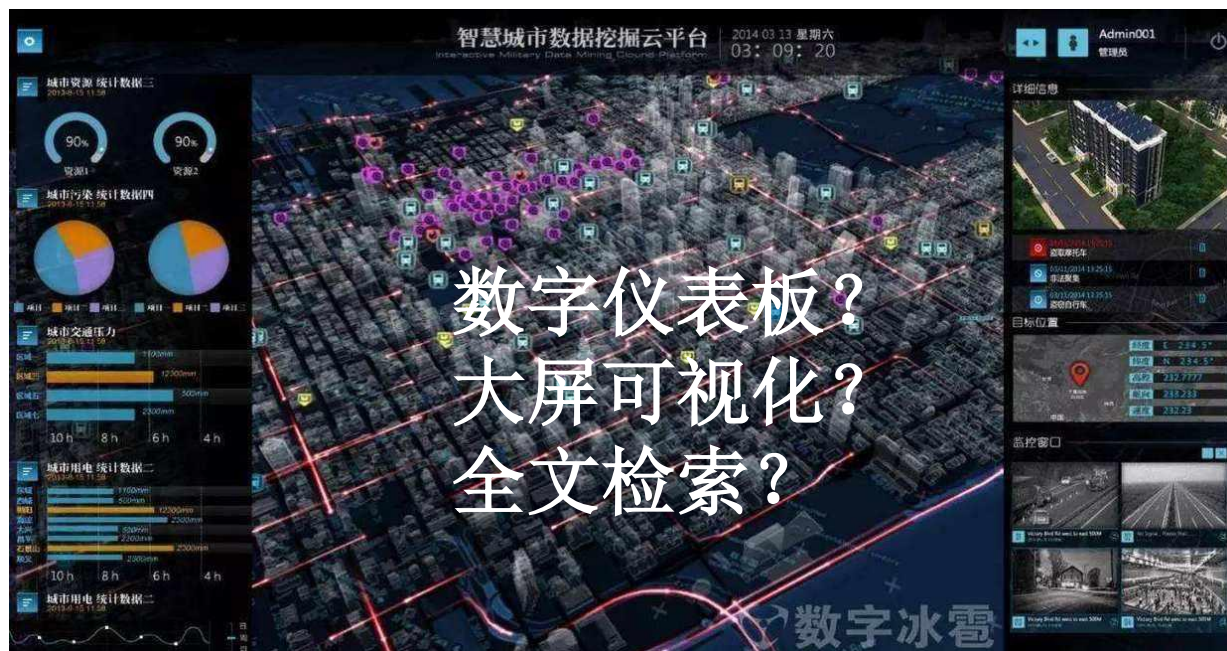
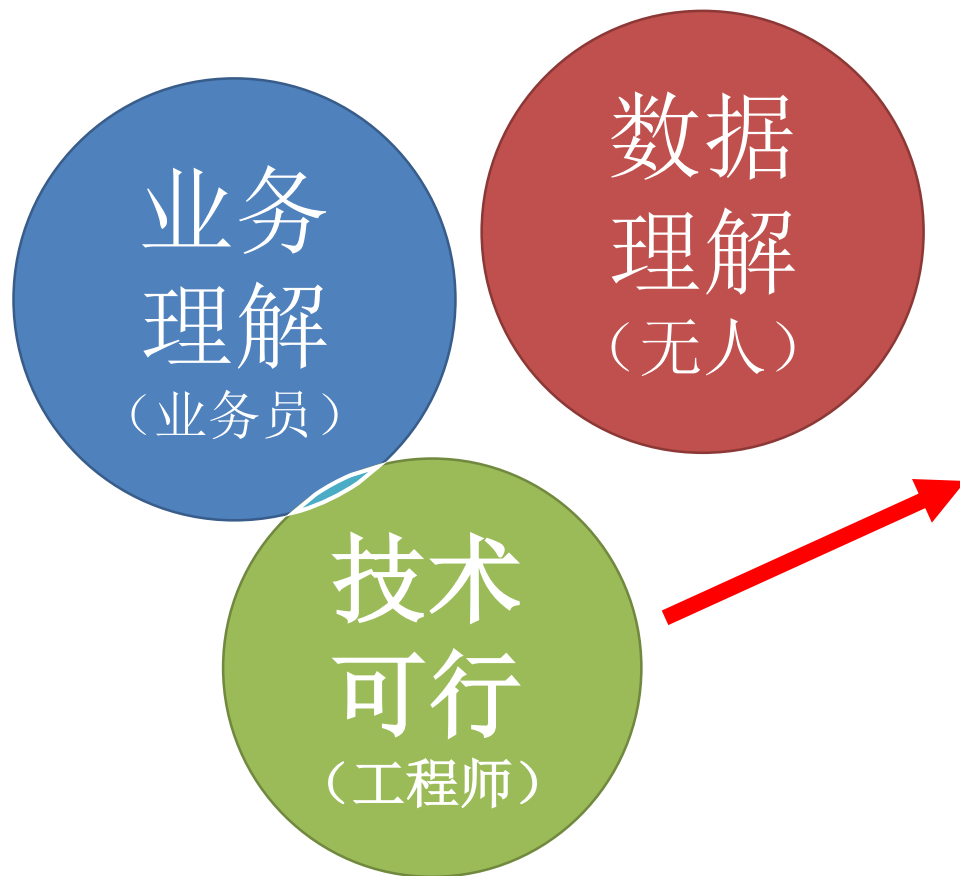


3. 大数据项目不容易成功！

- 要分析数据并产生价值，必须基于深刻的业务理解、数据理解以及技术理解。
 - 理解业务规律，并能根据业务规律提取相应的特征。
 - 理解产生的数据，对数据的内容及质量有准确的评估。
 - 理解所使用的技术，将合适的技术准确地应用于问题中。



真实的大数据项目



智慧隧道是如何失去智慧的？

- 智能城市、智慧隧道



<https://weibo.com/ttarticle/p/show?id=2309404663255657480861>

谢谢！
Thank you for your attention.

liuyuewen@xjtu.edu.cn

