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：

“We spent over 20 years figuring out how to build data warehouses, and only 20 minutes thinking about how to use them.”

“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。”

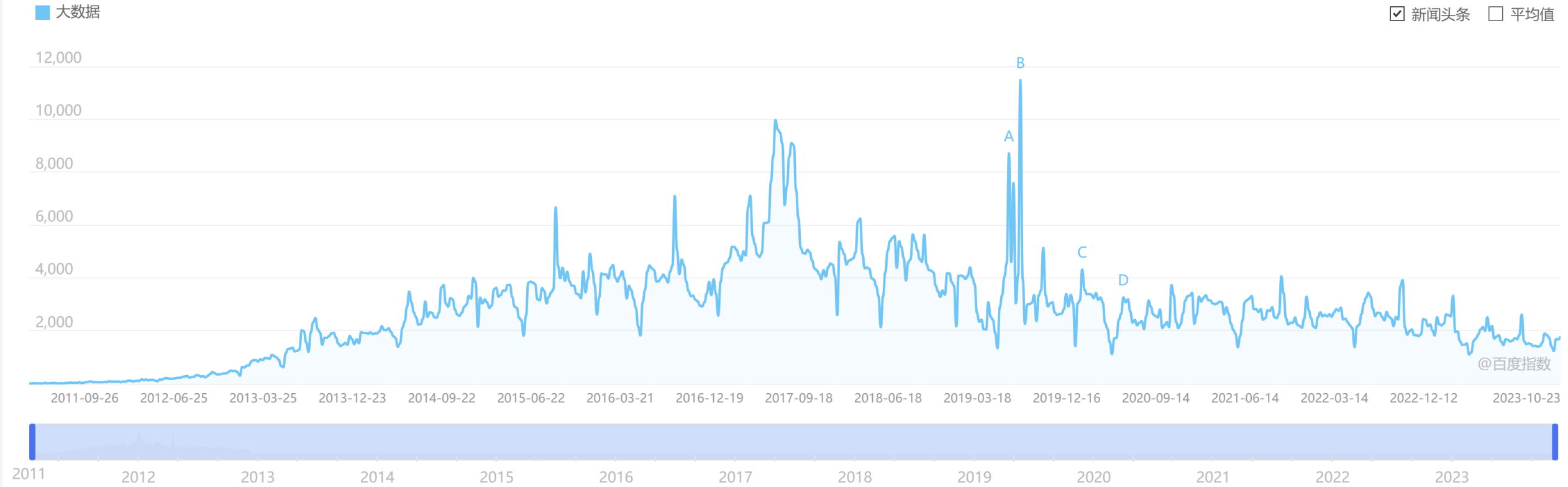
提纲 Outline

1. 大数据概念的变迁 The Evolution of the Big Data Concept
2. 分布式计算技术解决大数据问题 Solving Big Data Issues with Distributed Computing Technology
3. 大数据产生价值的2条路线 Two Pathways to Unlocking Value from Big Data
4. 大数据项目没有那么容易成功 The Challenges of Achieving Success in Big Data Projects

1. 大数据概念的变迁
The Evolution of the Big Data Concept

搜索指数

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



重要论文与报告
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 最有名的大数据报告
  - More than 150 pages 多达150页
  - **Executive Summary 执行摘要**

Obama Administration Proposes Big Data Plan

- 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（国家科学基金）:**
  - http://www.nsf.gov/news/news_summ.jsp?cntn_id=123607
- HHS/NIH: **National Institutes of Health（国家卫生研究院）– 1000 Genomes Project Data Available on Cloud**:
  - http://www.nih.gov/news/health/mar2012/nhgri-29.htm
- DOE: **Department of Energy（能源部）– Scientific Discovery Through Advanced Computing**:
- http://science.energy.gov/news/
- DOD: **Department of Defense（国防部）– Data to Decisions**:
  - www.DefenseInnovationMarketplace.mil
- DARPA: **Defense Advanced Research Projects Agency（国防高级研究计划局）– the XDATA program**:
  - http://www.darpa.mil/NewsEvents/Releases/2012/03/29.aspx
- USGS: **US Geological Survey（美国地质调查局）– 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

  《自然》大数据特刊

  - 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

  《科学》杂志数据分析专题

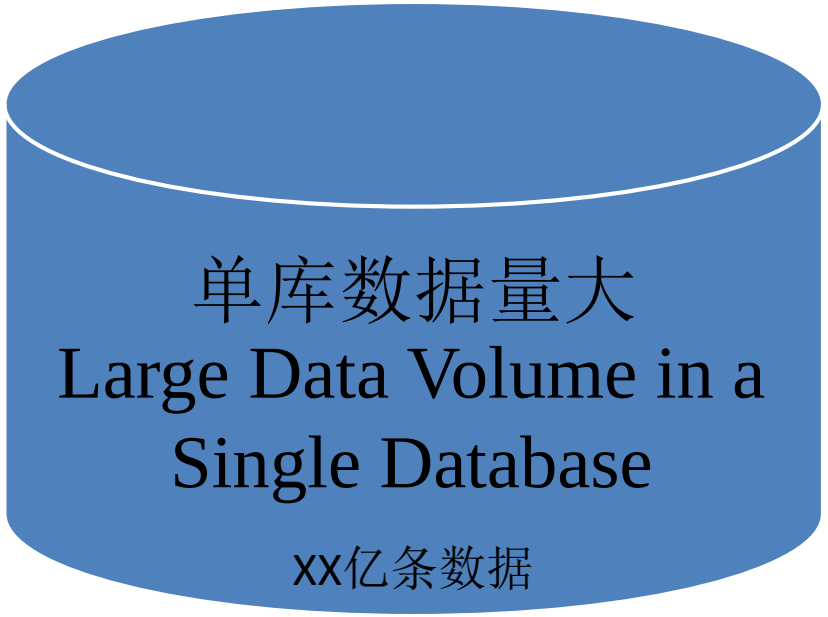
  - King, G. (2011). "Ensuring the data-rich future of the social sciences." Science 331(6018): 719-721.

2. 早期的大数据概念
Early Concepts of Big Data

大数据 ≠ 大量数据

数据体量大

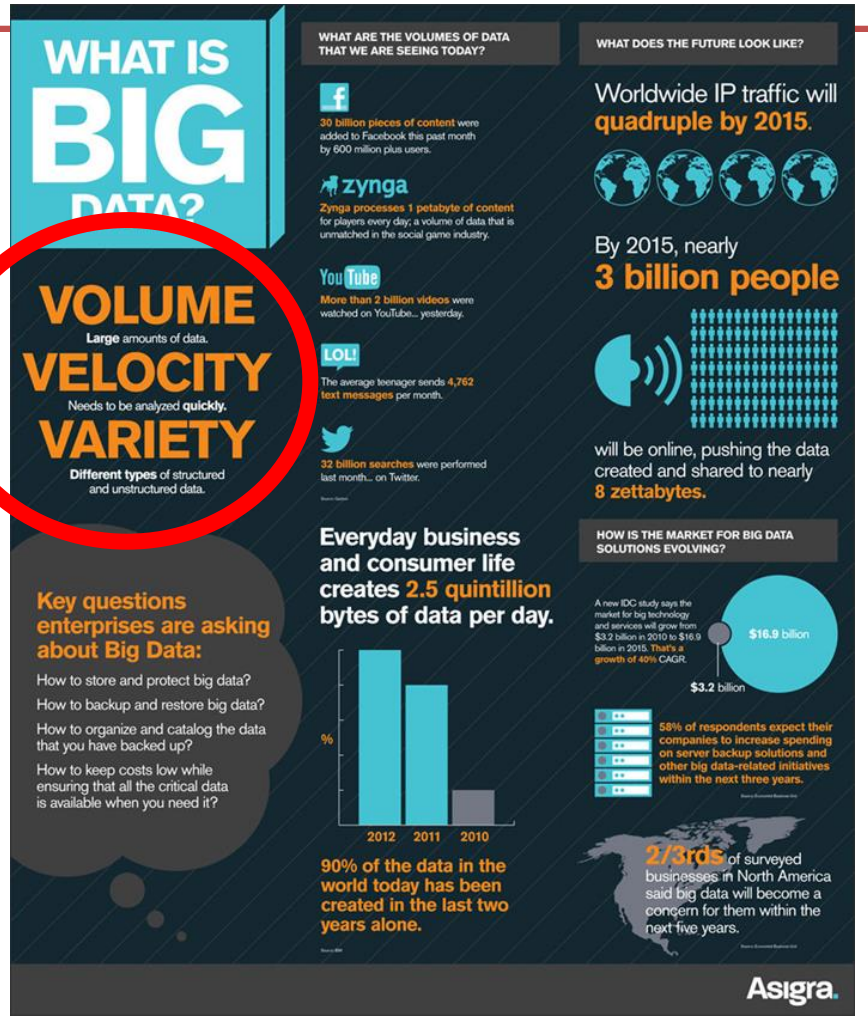
Large Volume of Data



数据种类多

Multiple Data Types

大数据 Big Data = 3V?



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

Early Concepts of Big Data

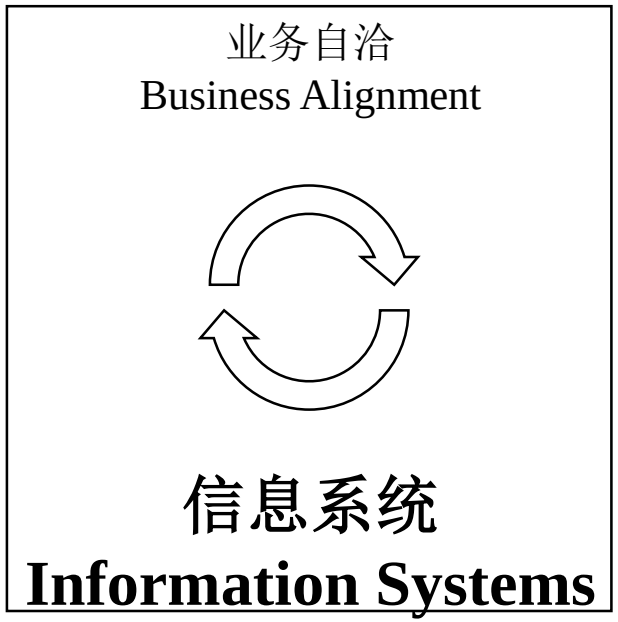
- Big data is a term used to refer to **the study and applications of data sets that are so big and complex** that traditional data-processing application software are inadequate to deal with them.
- **Big data challenges** include capturing data, data storage, data analysis, search, sharing, transfer, 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. Current Big Data Concept: Uncovering Value through Data Analysis

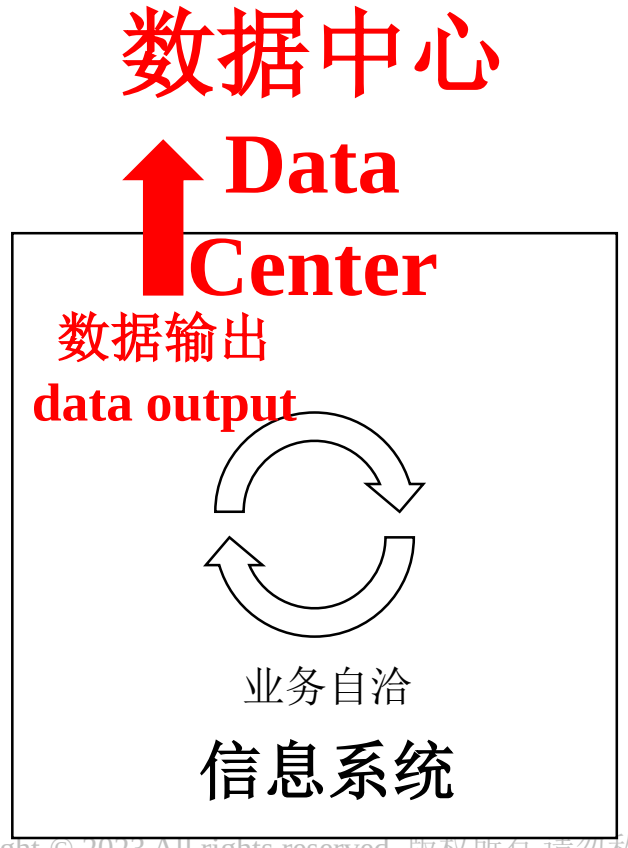
- The term "big data" tends to refer to **the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data**, and seldom to a particular size of data set.
- There is little doubt that the quantities of data now available are indeed 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
The Evolution from Informatization to Big Data

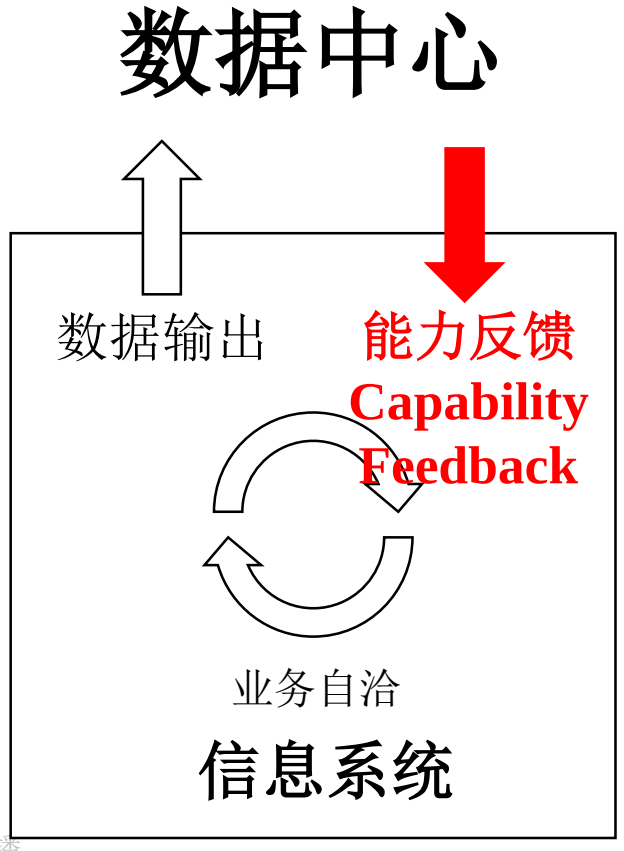
1.0 信息化时代
The Information Era



2.0 数据汇聚时代
The Era of Data Convergence



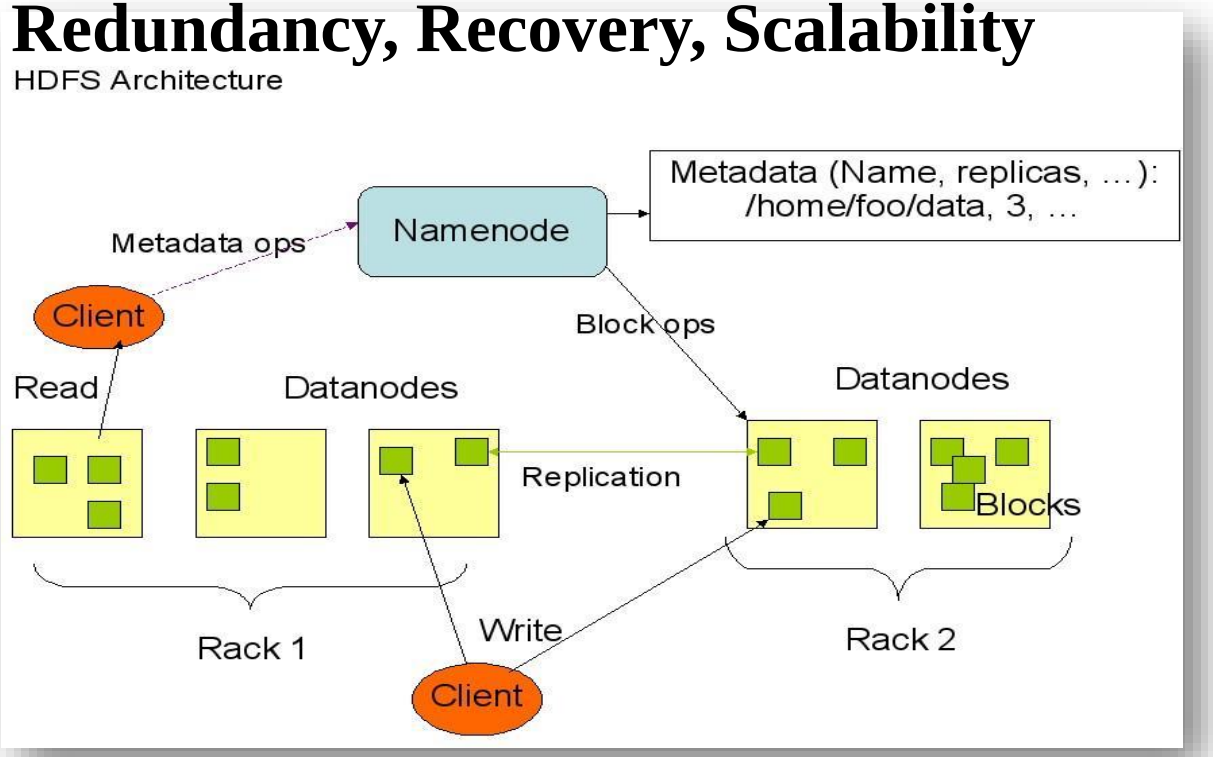
3.0 数据赋能时代
The Era of Data Empowerment

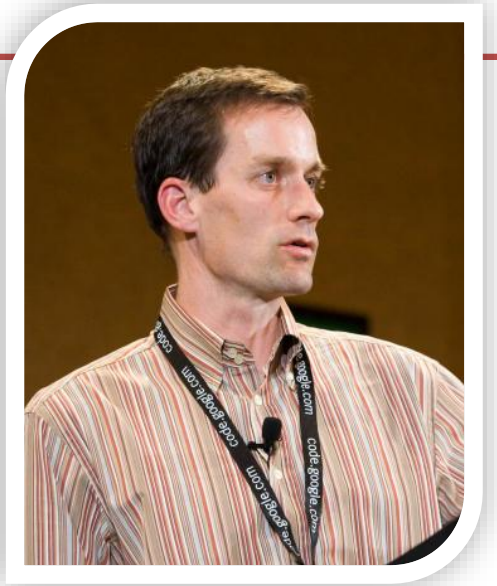


2. 分布式存储与计算系统
Distributed Storage and Computing System

1. 分布式系统
Distributed Systems

- 分而治之：**Divide and Conquer**
- 冗余、容灾、可扩展：
- **Redundancy, Recovery, Scalability**



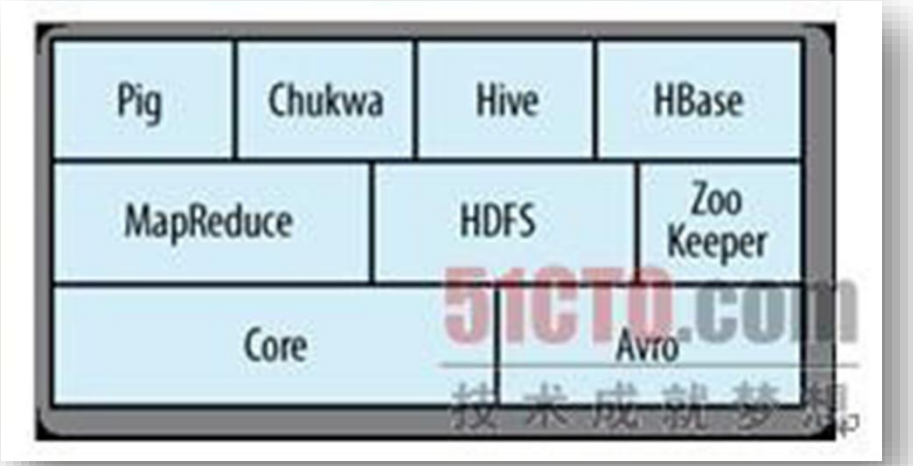


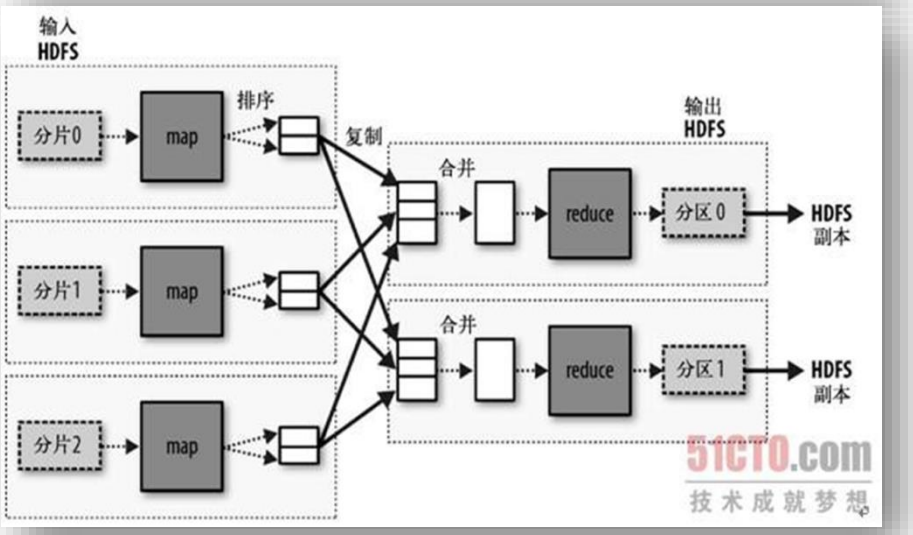
Google MapReduce
架构设计师
Jeffrey Dean

分布式数据处理技术
Distributed Data Processing Technologies

- Hadoop
- MapReduce

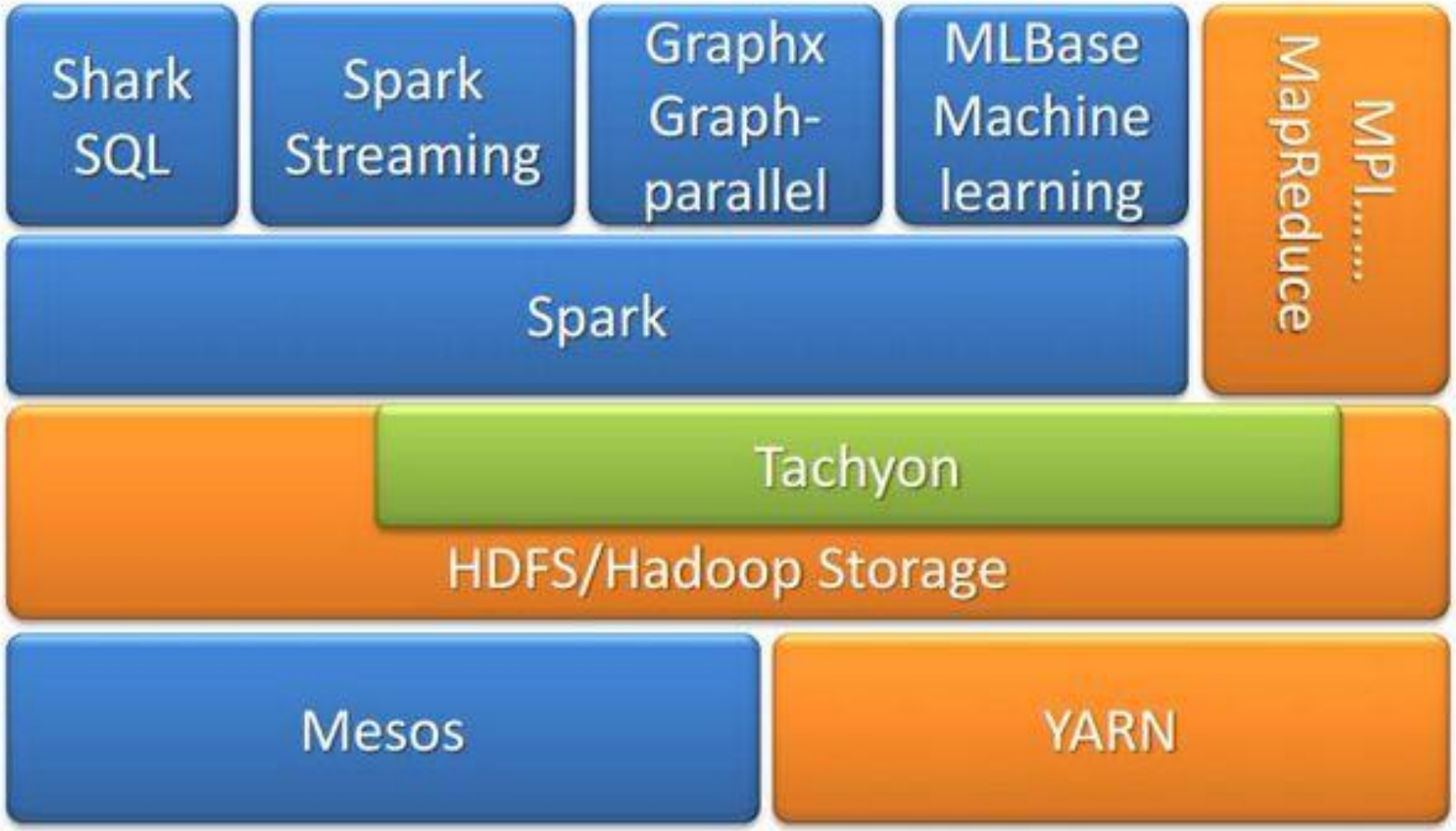
- 如何利用Hadoop对海量数据进行优化处理是Yahoo正在致力于工作的内容。以网络分析为例，Yahoo目前有超过100亿个网页，1PB的网页数据内容，2万亿条链接，每日面临这300TB的数据输出。“在应用Hadoop前，实施这一过程我们大概需要1个月的时间，但应用后仅需要1周时间”Yahoo is currently working on optimizing the processing of massive data using Hadoop. Taking network analysis as an example, Yahoo has over 100 billion webpages, 1 petabyte of webpage content, and 200 trillion links. They face a daily data output of 300 terabytes. "Before implementing Hadoop, this process would take approximately one month, but after its application, it now only takes one week."



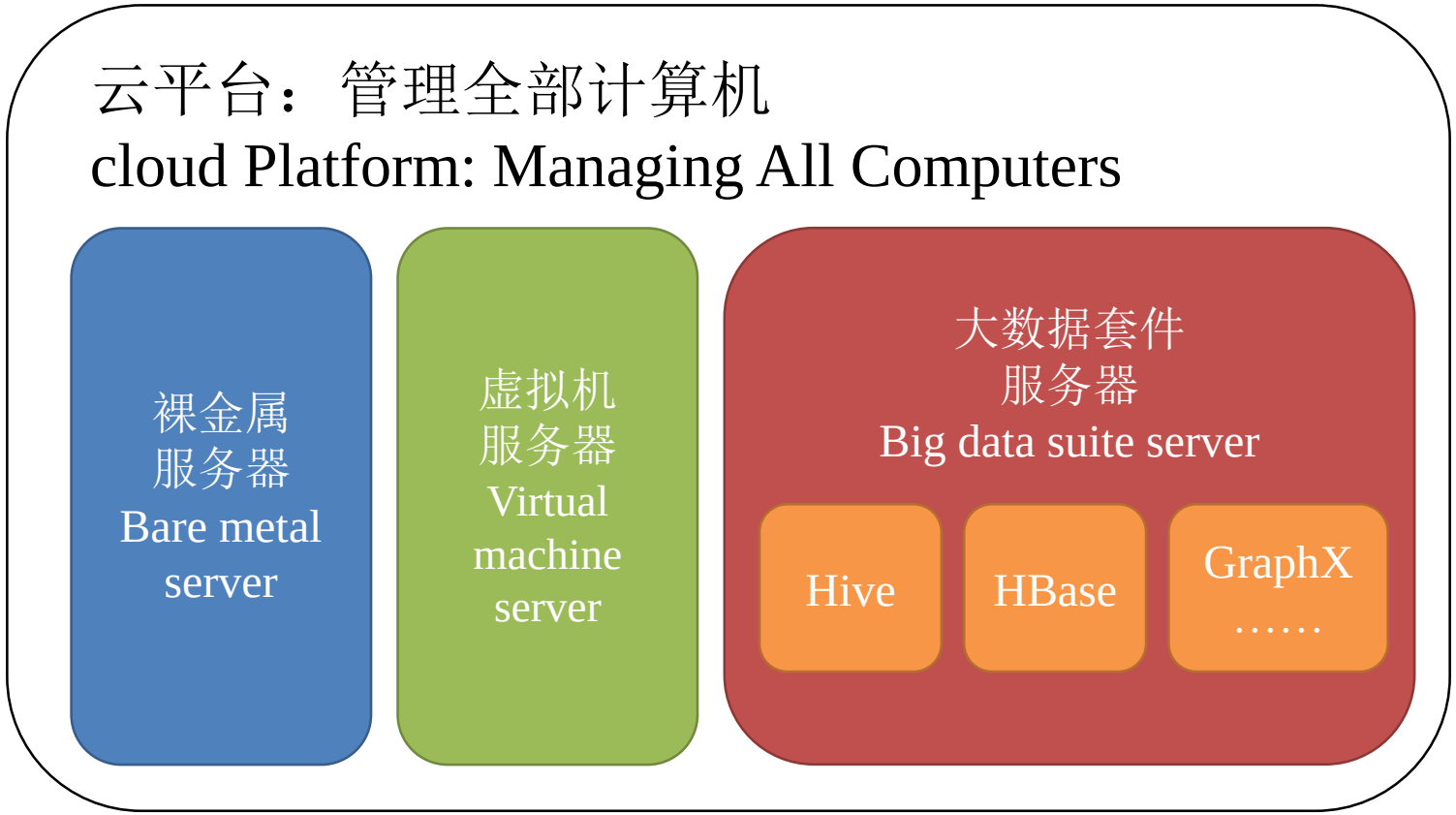


2. Hadoop生态体系
Hadoop Ecosystem

3. Spark生态体系
Spark Ecosystem

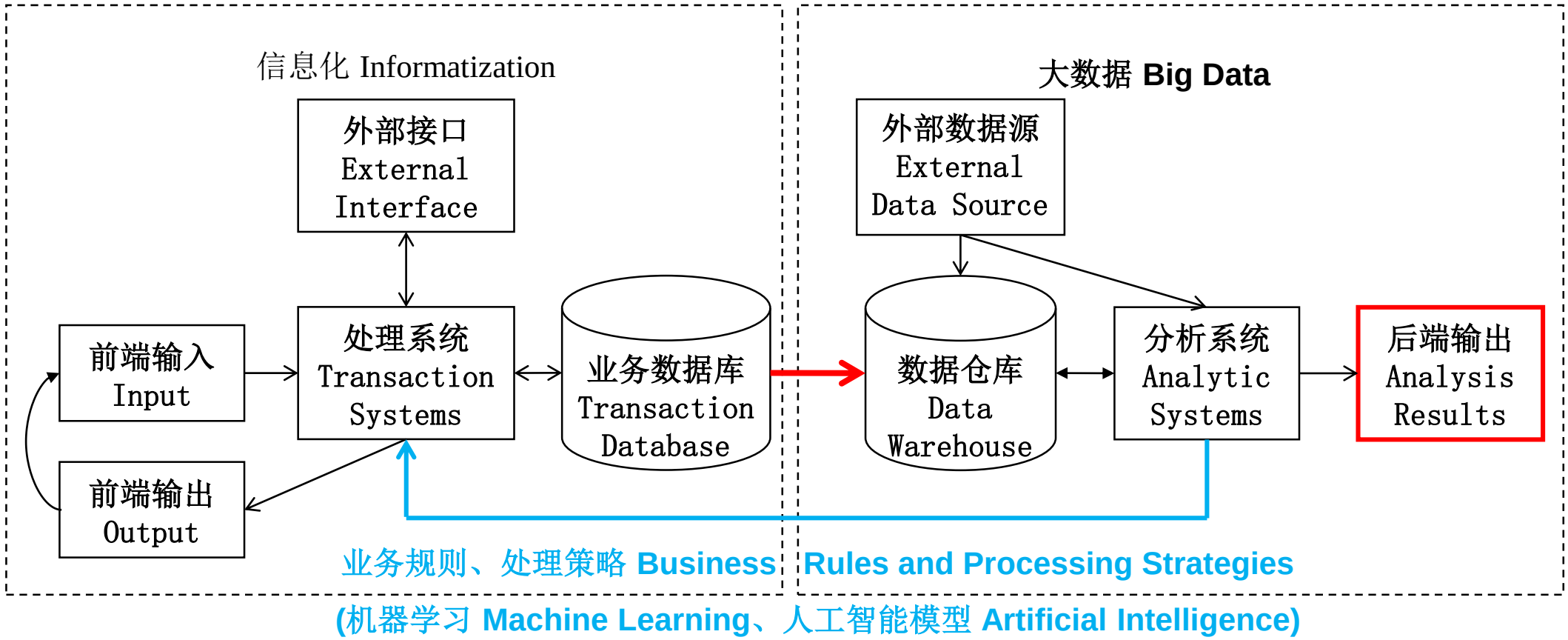


4. 云平台与大数据平台
Cloud Platform and Big Data Platform



3. 大数据分析的两条路线：探索性分析与机器学习
The Two Paths of Big Data Analysis: Exploratory Analysis and Machine Learning

1. 大数据的两条路线
Two Paths of Big Data Analysis



2. Case: China Eastern Airlines' Safety Operations Research Institute

- Source of the Case: https://www.sohu.com/a/412449304_115926
- The establishment of China Eastern Airlines' Safety Operations Research Institute must be acknowledged as a far-sighted move by the airline's top management. What the leaders were contemplating was, what kind of threats would the airline company face in ten, or even twenty years from now? Where would the new opportunities lie? No one can guarantee the company's existence for decades, especially in the aviation industry, which is known for its rapid changes. Even if there is profitability now, what core competencies should we rely on to survive once the external environment undergoes changes?
- To keep up with the global aviation industry's development trends, it's essential to have a dedicated "intelligence team." For China Eastern Airlines, the Research Institute plays the role of this "intelligence team": constantly interpreting new information, introducing new methods, theories, ideas, and technologies, and adding momentum to the company's growth.

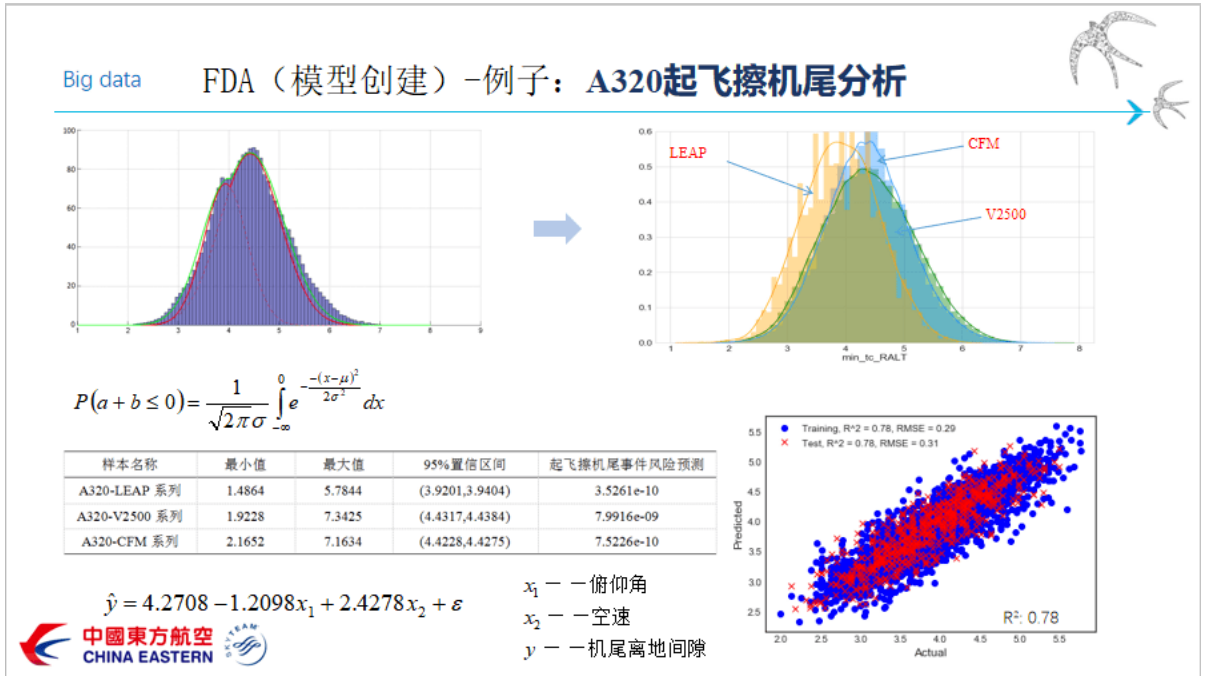
- Currently, China Eastern Airlines has more than 11,000 pilots, and this number is expected to continue growing in the future. This demands a disruptive change in our management approach. For instance, in the past, we emphasized Standard Operating Procedures (SOP), and focusing on SOPs led to an immediate decrease in the accident rate. Later, we introduced human factors management and cultural management, which further reduced the accident rate. Now, with higher public expectations for flight safety, how can we continue to improve the accident rate?
- Today's aircraft are quite advanced, and during flights, they generate a substantial amount of data, which is accumulated in various departments, such as the Flight Operations Department and the Operations Department. Previously, when we didn't utilize this data, it was essentially "dead," but in reality, it's a valuable "treasure." By harnessing this data, we can extract more juice from the "lemon" of flight safety, thus improving safety performance even further.

- **Understand Present Situation:**
- In 2015, a civil aviation incident occurred in which an aircraft had a tire burst during landing at a high-altitude airport. After learning about this incident, we performed a three-year data analysis rollback to study China Eastern Airlines' operations at high-altitude airports. One of the researchers created a set of PPT slides analyzing takeoff data, and when I saw a particular chart, I noticed that the aircraft used less than half of the runway for takeoff, resulting in "runway waste." Typically, we take off with a bit more than half of the runway. This meant that the aircraft's engines were working harder than necessary. However, engines have a lifespan, and prolonged usage in this manner results in significant wear and tear.
- Following this, the Research Institute categorized and compiled data from various subsidiaries and airports for further analysis. We found that **one of the subsidiaries was using engine thrust more aggressively than others.**

- After ruling out other potentially relevant factors, we convened a meeting led by company executives to discuss the root causes of the issue.
- This particular subsidiary explained that they were reducing thrust less because their aircraft's performance data assumed a maximum temperature of only 50 degrees, whereas other subsidiaries had set it to 60 or 70 degrees.
- In other words, in this specific aspect, the data received by various subsidiaries of China Eastern Airlines were completely different, but each subsidiary had assumed they were the same as the others. Upon discovering the disparities in performance data among the subsidiaries, the relevant departments promptly addressed this issue and filled the gap.

- **Identify Anomalies：**
- Previously, everyone primarily used QAR (Quick Access Recorder) alerts to assess flight quality, but the essence of these alerts is to detect "anomalies" in flight data, leading to a substantial waste of data. However, we aspire to employ a big data mindset for flight quality analysis.
- Taking the example of the A320 tail scrape analysis: For the A320 aircraft, the landing attitude should not exceed 11 degrees. For instance, Pilot A maintains a daily flight angle of 10 degrees, and Pilot A exhibits low standard deviation, indicating stability. In such cases, QAR should not generate alerts for Pilot A. In contrast, Pilot B's landing attitude may vary between 5 and 10 degrees. Although it falls within the specified range, Pilot B's instability should be a cause for concern.

- During the process of conducting flight quality analysis, we also had an unexpected discovery. Through the use of big data, we identified an abnormal flight attitude in one A320 aircraft: while other aircraft typically maintained an average flight angle of around two and a half degrees, this particular aircraft had an average flight angle of six degrees.
- Even though it didn't result in any accidents or precursor signs of accidents, there was undoubtedly an issue. Later, China Eastern Airlines' Engineering Department and Airbus collaborated and identified that during a major maintenance, the flaps' safe angles were incorrectly set. Afterward, Airbus reassembled the flaps, resolving the issue.



Saved￥75million

- **Predicting the Future (Analytical Report and Decision Recommendations):**
- One of China Eastern Airlines' subsidiaries was considering purchasing engines from a particular engine OEM (Original Equipment Manufacturer). This OEM offered two types of engines with different levels of thrust, with the higher thrust engine being more expensive.
- Based on extensive big data analysis considering factors such as routes, aircraft types, and airports, we generated a report in just two days, concluding that the lower thrust engine was sufficient for the company's needs. This recommendation was adopted by the company, resulting in cost savings of over 75 million Chinese yuan in this transaction.
- Data-driven decision-making can guide airlines toward safer and more efficient directions, as 'numbers never lie' after all."

3. 案例：UPS不许左转弯
Case: UPS Prohibits Left Turns

- UPS
  - **快递业务 Courier Business：**
  - Operating vehicles: 10万辆运营车辆；
  - Self-owned airplanes: 自有飞机237架
    Leased airplanes: 租用412架
  - Employees: 43.5万名员工
  - **供应链及货运业务 Supply Chain and Freight Business：**
  - Cargo trucks: 5599辆货运卡车；
  - Trailers: 19884辆拖车
- **识别异常 Identify Anomalies：**
  - UPS安全带扣上比例 Seat Belt Fastened Ratio：**98%**
  - 目标 Goal：**99.8%**
  - 如何达成 How？

No Left Turn

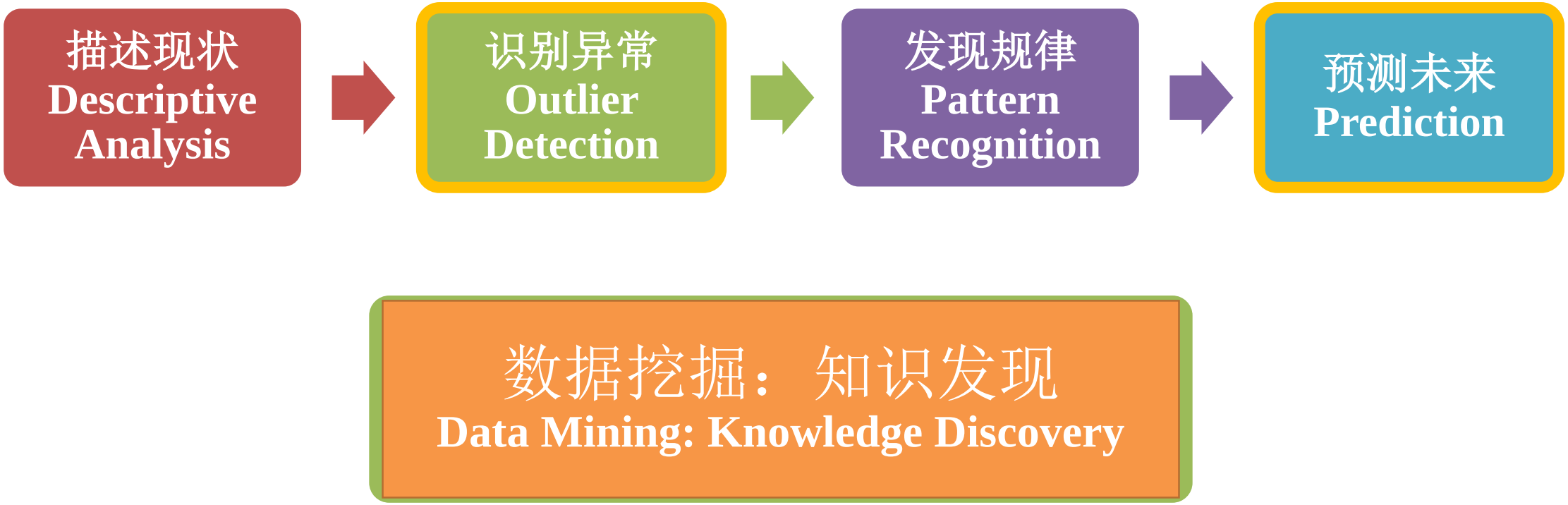
- In 2001, after implementing a better tracking system, courier companies took a closer look at how their trucks performed when delivering packages. As a logistics company with 96,000 trucks and hundreds of airplanes, UPS needed to optimize various aspects of its operations to improve performance, such as reducing fuel consumption, saving time, and maximizing space efficiency (UPS trucks in parking lots are parked with their mirrors against each other to save space).
- To ensure that the headquarters could track the location of vehicles in case of delays and prevent engine failures, UPS equipped its trucks with sensors, wireless adapters, and GPS.

- **Discovering Patterns (Business Rules, Execution Strategies):**
- UPS engineers discovered that left turns were a major impediment to efficiency. Turning left, against the flow of traffic, led to extended waits in left-turn zones, resulting in time and fuel wastage and an increased risk of accidents. By charting a series of right-turn loops, UPS not only improved profitability and safety but also effectively promoted its slogan and environmentally friendly policies.
- **In 2004, UPS announced to its drivers that the correct way to reach any destination was to avoid left turns.** Even in cases like this route, a UPS driver would say to a skeptic:"We'll make a right onto West Street from 135th Street, then another right onto 139th Street. Make a right on 139th, go a block, and make another right."
- Since **UPS started using route-planning software**, it has enabled drivers to make right turns in congested areas and make exceptions for left turns when they are more convenient and efficient.

Saved￥70million

- Jack Levis, UPS's Director of Process Management, believes that this analytical project has shown significant results. **In 2011, UPS drivers traveled nearly 48.28 million fewer kilometers, saving 3 million gallons of fuel and reducing carbon dioxide emissions by 30,000 metric tons.**
- In 2012, the right-turn rule and other improvements, for various reasons UPS didn't disclose, saved approximately 10 million gallons of gasoline, equivalent to reducing the emissions of 5,300 cars in a year.

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
The Challenges of Achieving Success in Big Data Projects

1. Misconceptions About Big Data

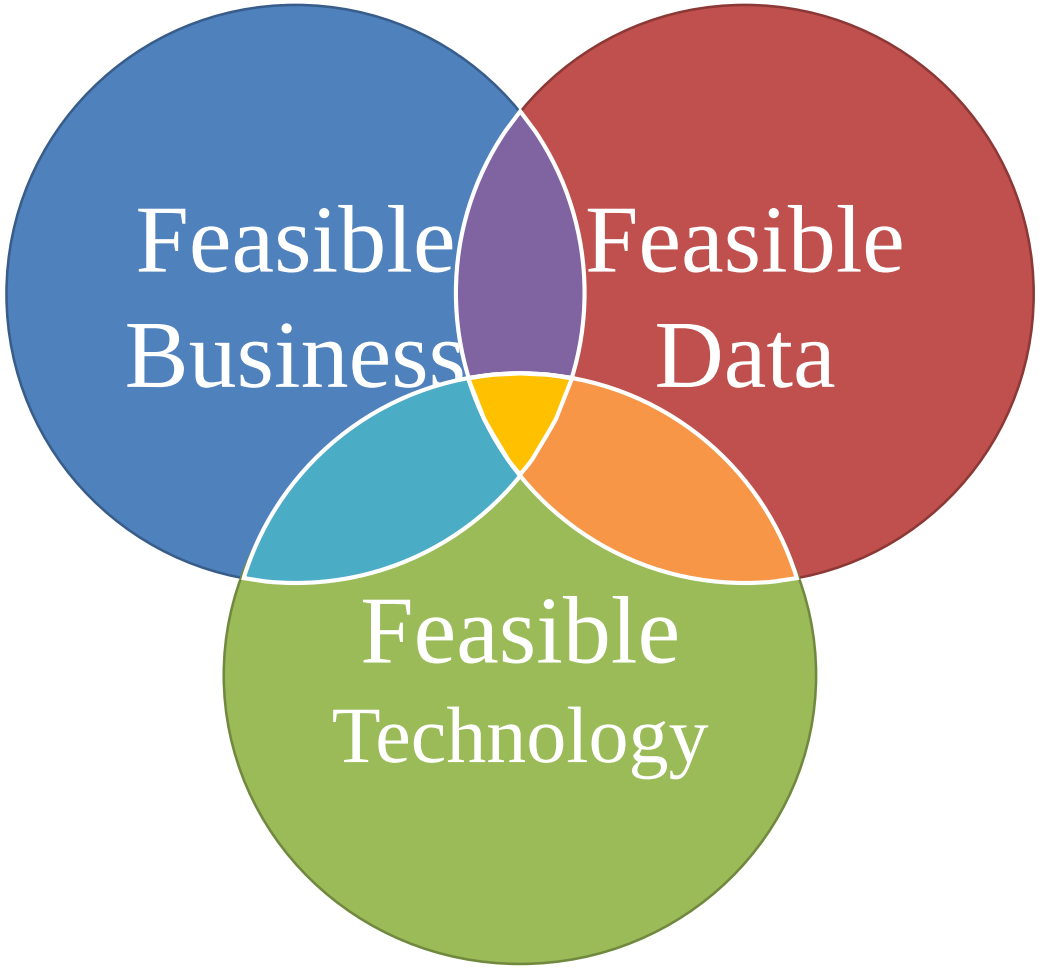
- **Irrelevance to Me:**
  - "We're in Traditional Agriculture/Industry/Manufacturing, Big Data/AI Isn't Relevant to Us.
  - Big Data/AI is High-Tech, Mainly for Internet Companies, Not Related to Us."

- **Excessive Expectations:**
  - Believing that big data/AI can completely replace human work or drastically reduce costs.
  - Thinking that implementing a big data system immediately thrusts a company or organization into the big data era.
  - Assuming that the mere installation of cutting-edge big data systems will significantly enhance all aspects of a company or organization.
  - Believing that big data/AI can automatically discover market opportunities, make decisions for us, and substantially boost profitability and performance.

- **Throwing Money at It：**
  - We have a lot of data. Can you create some big data models for us?
  - Hiring renowned big data companies/experts will allow us to develop highly advanced big data systems, bringing immense value to all aspects of our company.
  - Big data/AI is a high-tech field, and our staff doesn't have the necessary skills; We need to hire high-level experts to get the job done."
- **Data Accumulation Oriented:**
  - Our data volume is insufficient, and we cannot yet establish a big data system.
  - We've formulated a digital transformation strategy. Over the next three years, we will embark on this journey, collecting a significant amount of data. By the end of this three-year period, our data will become a substantial asset, generating immense value.
  - We've developed a digital twin system capable of producing a large amount of data.

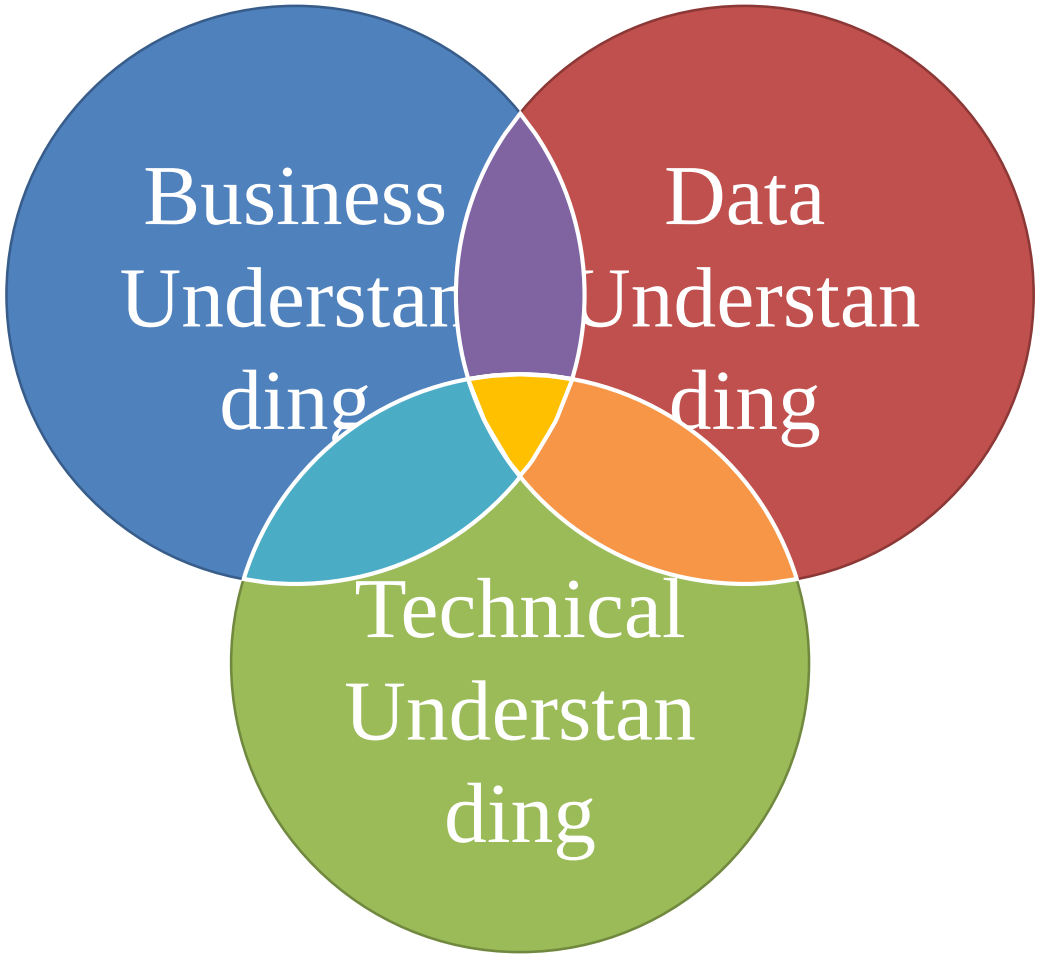
2. Big data is not a panacea !

- The real areas where big data/AI can be put into practice and generate business value are those where the business, data, and technology are all feasible.
  - The business itself has some yet-to-be-discovered relatively simple underlying patterns.
  - There is data available that matches the business needs for discovering patterns.
  - The existing big data/data analysis/AI technologies meet the requirements
- **Choosing the right problem is more difficult than solving it.**

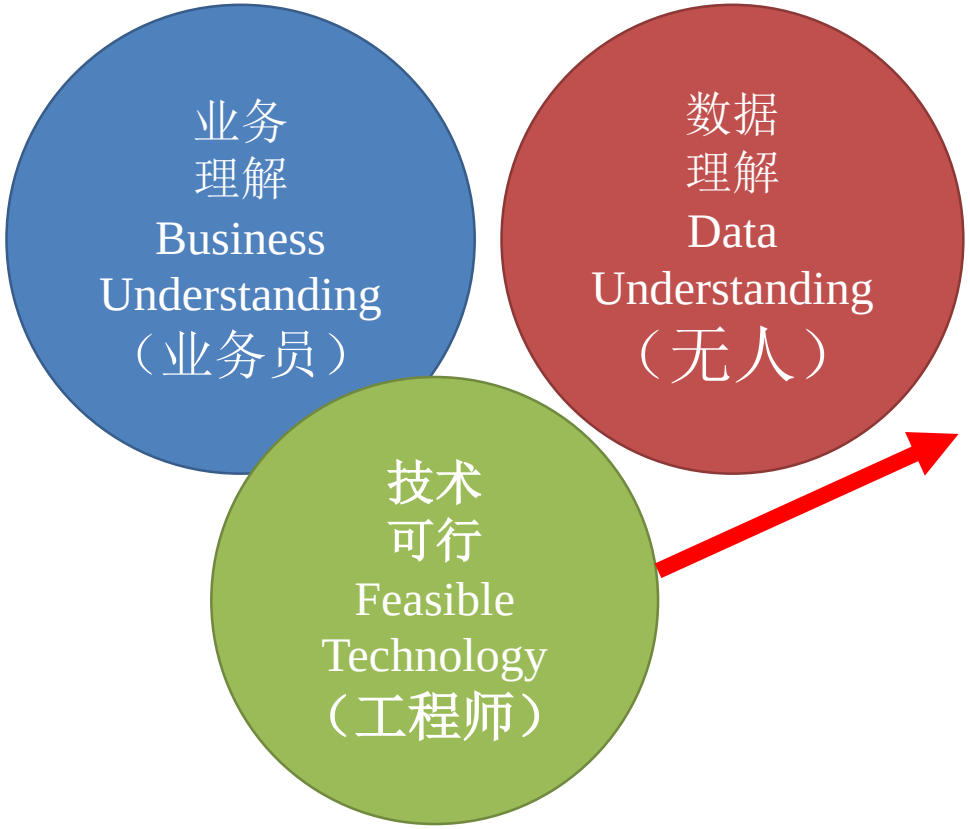


3. Big data projects are not easy to succeed!

- To analyze data and generate value, a profound understanding of the business, data, and technology is essential.
  - Understanding business patterns and being able to extract relevant features based on those patterns.
  - Understanding the generated data and accurately assessing its content and quality.
  - Understanding the technology being used and applying the right technology accurately to the problem.



真实的大数据项目
A Real Big Data Project



智慧隧道是如何失去智慧的？
How did the smart tunnel lose its intelligence?

- 智能城市、智慧隧道(Smart City, Smart Tunnel)

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

谢谢！

Thank you for your attention.

liuyuewen@xjtu.edu.cn