大数据系统介绍

- 课程名称：**大数据开发技术Hadoop（Hadoop分布式计算系统)**
- 课时数：64课时（拟定）
- 上课时间：每周五下午、周六全天
- 授课起止时间：2021年11月12日~2021年12月31日
- 讲师：盛泳潘
- 班主任：王子暄（校内）、舒湛（校外）

课程概览

- 大数据的价值
- 大数据基础
- 大数据系统架构
- 数据生成
- 数据获取
- 数据存储
- 数据分析
- 大数据分析分类
- 大数据系统基准（benchmark）
- 大数据科学问题

本章目录概览

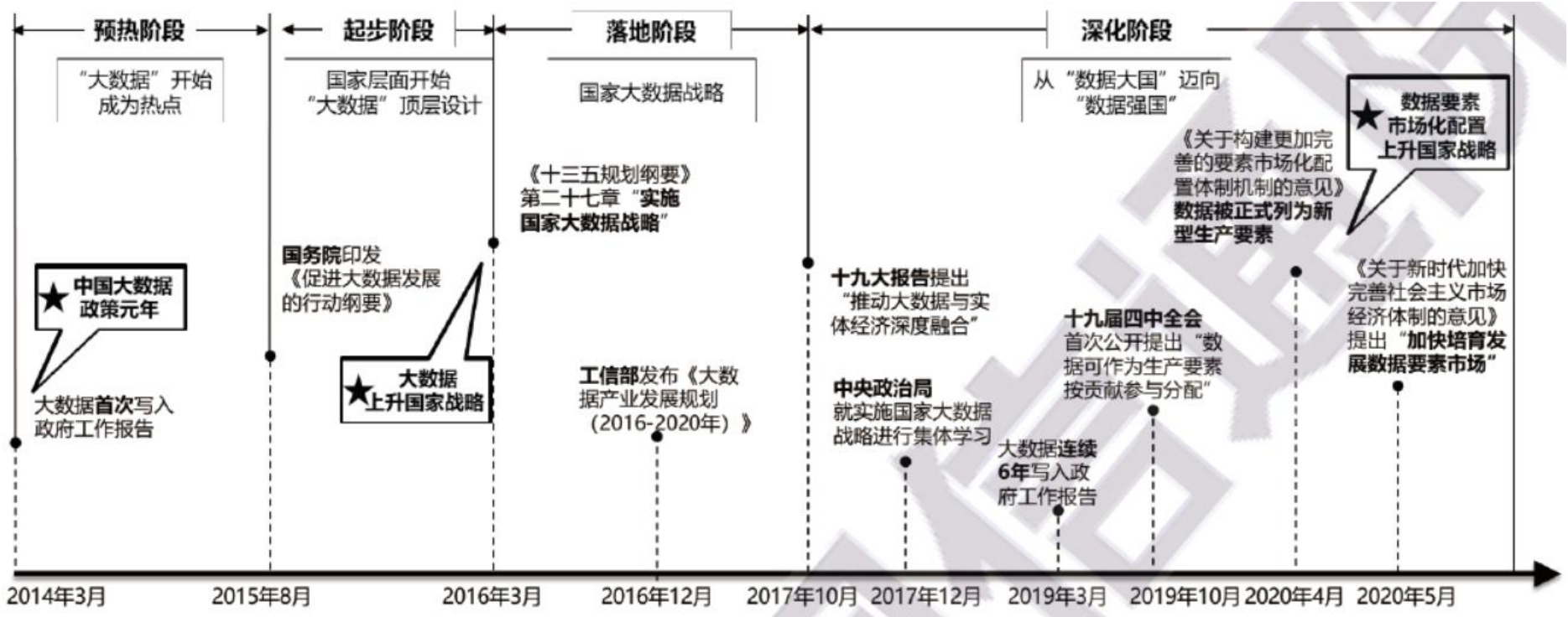
- **大数据的价值**
- 大数据基础
- 大数据系统架构
- 数据生成
- 数据获取
- 数据存储
- 数据分析
- 大数据分析分类
- 大数据系统基准（benchmark）
- 大数据科学问题

大数据的价值

- **在国家战略层面**
  - 2012年3月，美国总统奥巴马签署并发布了一个"**大数据研究发展创新计划（Big Data R&D Initiative）**"，投资2亿美元启动大数据技术和工具研发。
  - 这是继1993年美国宣布"信息高速公路"计划后的又一次**重大科技发展部署。**
  - 美国政府认为："**大数据是未来的新石油**"，将大数据研究上升为国家意志，认为大数据将对未来的科技与经济发展带来重大影响，一个国家**拥有数据的规模**和**运用数据的能力**将成为综合国力的重要组成部分，对数据的占有和控制也将成为国家间、企业间新的争夺焦点。
  - 英国、法国、德国、日本等发达国家也纷纷推出了相应的大数据发展战略计划。

大数据的价值

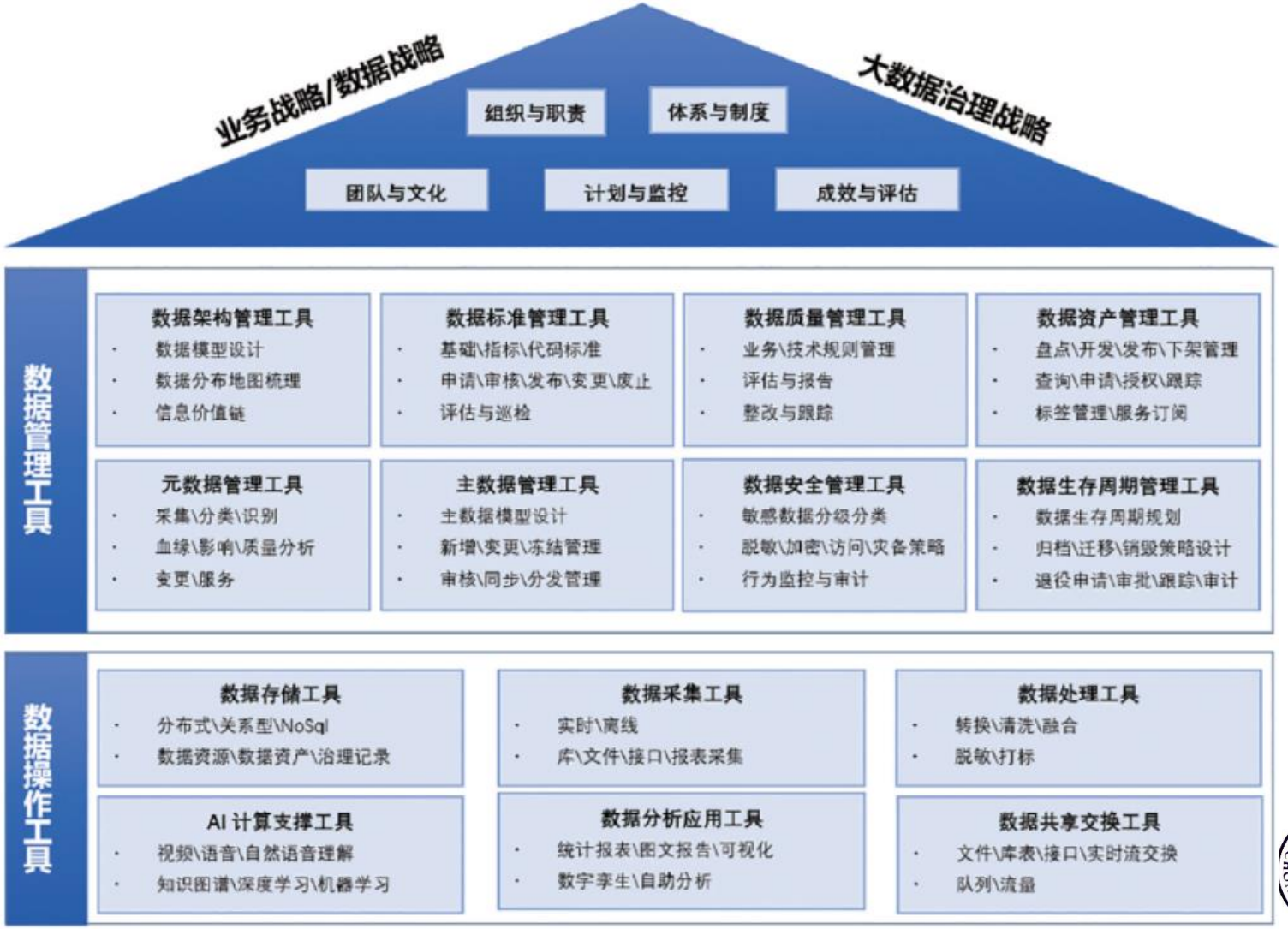
- **在国家战略层面**
  - 自2014年以来，我国的大数据战略的谋篇布局大致经历了4个不同阶段，正逐步从数据大国向数据强国迈进。



选自《大数据白皮书（2020）》

大数据的价值

- **数据治理工具全景图**



大数据的价值

- **在学术研究层面**
  - 2012年，中国计算机学会和通信学会都成立了**大数据专家委员会**（http://tc.ccf.org.cn/tfbd/zwjs/zzjg/qtcy/）
  - **学术活动**：CCF大数据学术会议（CCF Bigdata 2021，http://www.bigdata2021.net/ ）、中国大数据技术创新与创业大赛（http://www.cxcyds.com/cxcyds/dsxw/202007/89739afe494c4490b22d2f5f0b59da36.shtml ）、大数据分析与管理国际研讨会（ICBDMA 2021, http://www.icbdma.org/）、大数据科学与工程国际学术研讨会（BDSE 2021，http://bdse.xintongconference.com/Page ）、中国大数据技术大会（2021， https://www.ciiabd.org.cn/zt/21/bda/ ）、中国国际大数据大会（2021 http://science.china.com.cn/2021-04/02/content_41519716.htm ）...
  - **大数据中心**：国家大数据中心（百科词条，https://baike.baidu.com/item/%E5%9B%BD%E5%AE%B6%E5%A4%A7%E6%95%B0%E6%8D%AE%E4%B8%AD%E5%BF%83/19826915?fr=aladdin ），四川省大数据中心（http://www.scdsjzx.cn/ ）

大数据的价值

- **在产业层面**
  - 2021年中国大数据产业生态地图

课程概览

- 大数据的价值
- **大数据基础**
- 大数据系统架构
- 数据生成
- 数据获取
- 数据存储
- 数据分析
- 大数据分析分类
- 大数据系统基准（benchmark）
- 大数据科学问题

大数据基础-大数据定义

- **属性定义（Attributive definition）**
  - 国际数据中心IDC：”大数据技术描述了一个技术和体系的新时代，被设计于从大规模多样化的数据中通过高速获取、发现和分析技术提取数据的价值”。
  - 大数据的4个显著特点，即4Vs，Volume（容量）、Variety（多样性）、Velocity（速度）、Value（价值）。
- **比较定义（Comparative definition）**
  - 2011年，McKinsey公司的研究报告，将大数据定义为：”超过了典型数据库软件工具获取、存储、管理和分析数据能力的数据集”。
  - 主观定义，但包含了一种演化的观点（从时间和跨领域的角度）。
- **体系定义（Architecture definition）**
  - 大数据是指数据的容量、数据的获取速度或者数据的表示限制了使用传统关系方法对数据的分析处理能力，需要使用水平扩展的机制以提高处理效率。

大数据基础-大数据定义

- 大数据 VS 传统数据

表1　大数据和传统数据比较

	Traditional data	Big data
Volume	GB	Constantly updated (TB or PB currently)
Generated rate	Per hour, day, ...	More rapid
Structure	Structured	Semi-structured or un-structured
Data source	Centralized	Fully distributed
Data integration	Easy	Difficult
Data store	RDBMS	HDFS, NoSQL
Access	Interactive	Batch or near real-time

大数据基础-大数据处理方式

- 大数据处理方式：流式处理和批处理
  - 流式处理：

    数据的潜在价值是freshness；流携带了大量数据，只有小部分的数据被保留在有限的内存中；代表开源系统：Storm、S4、Kafka等（s/ms级别）
  - 批处理：

    数据首先被存储，随后被分析；例如，MapReduce模型

表2　批处理和流处理比较

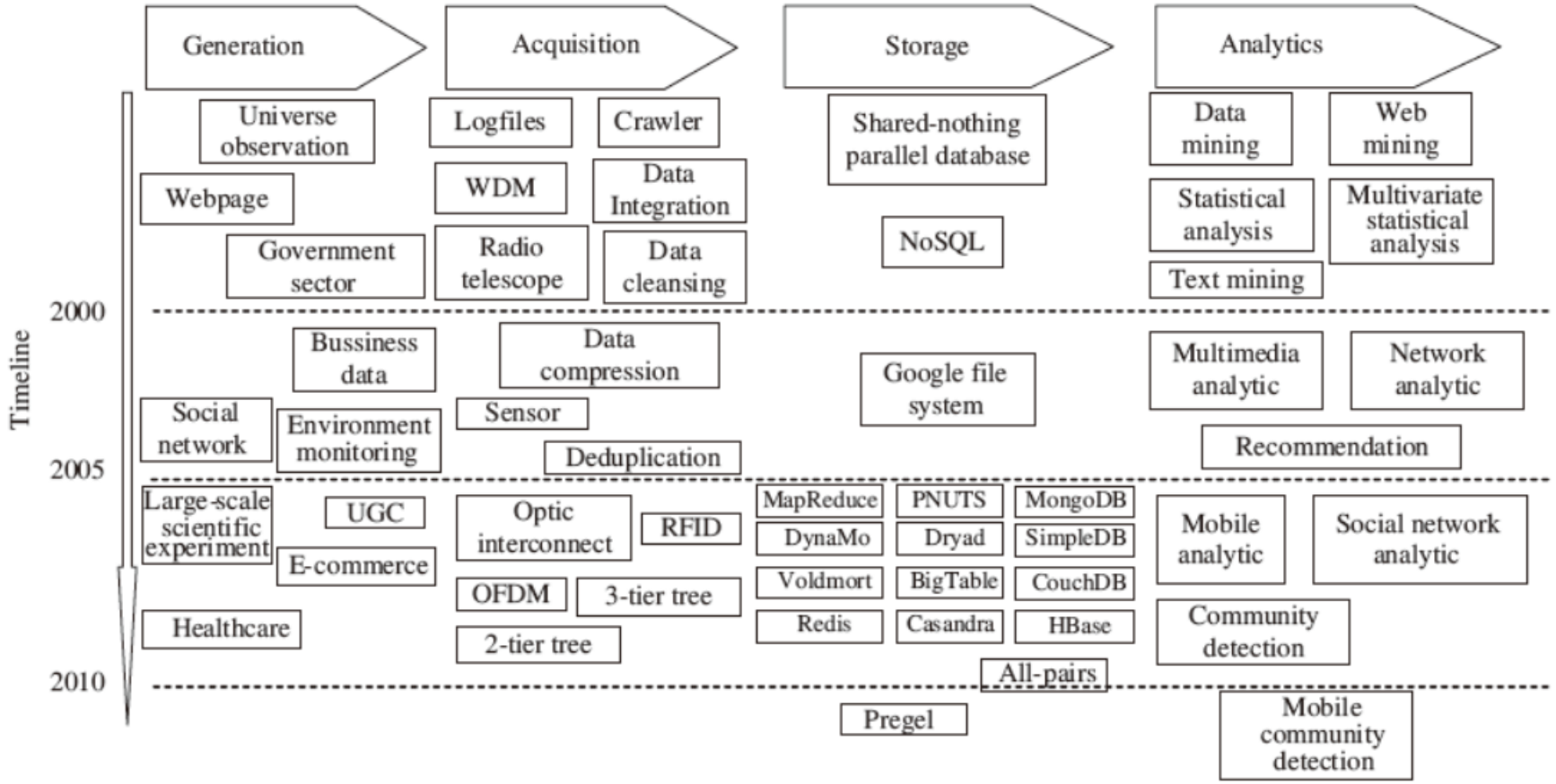
	Stream processing	Batch processing
Input	Stream of new data or updates	Data chunks
Data size	Infinite or unknown in advance	Known and finite
Storage	Not store or store non-trial portion in memory	Store
Hardware	Typical single limited amount of memory	Multiple CPUs and memory
Processing	A single or few pass(es) over data	Multiple rounds
Time	A few seconds or even milliseconds	Much longer
Applications	Web mining, sensor networks, traffic monitoring	Widely adopted in almost every domain

课程概览

- 大数据的价值
- 大数据基础
- **大数据系统架构**
- 数据生成
- 数据获取
- 数据存储
- 数据分析
- 大数据分析分类
- 大数据系统基准（benchmark）
- 大数据科学问题

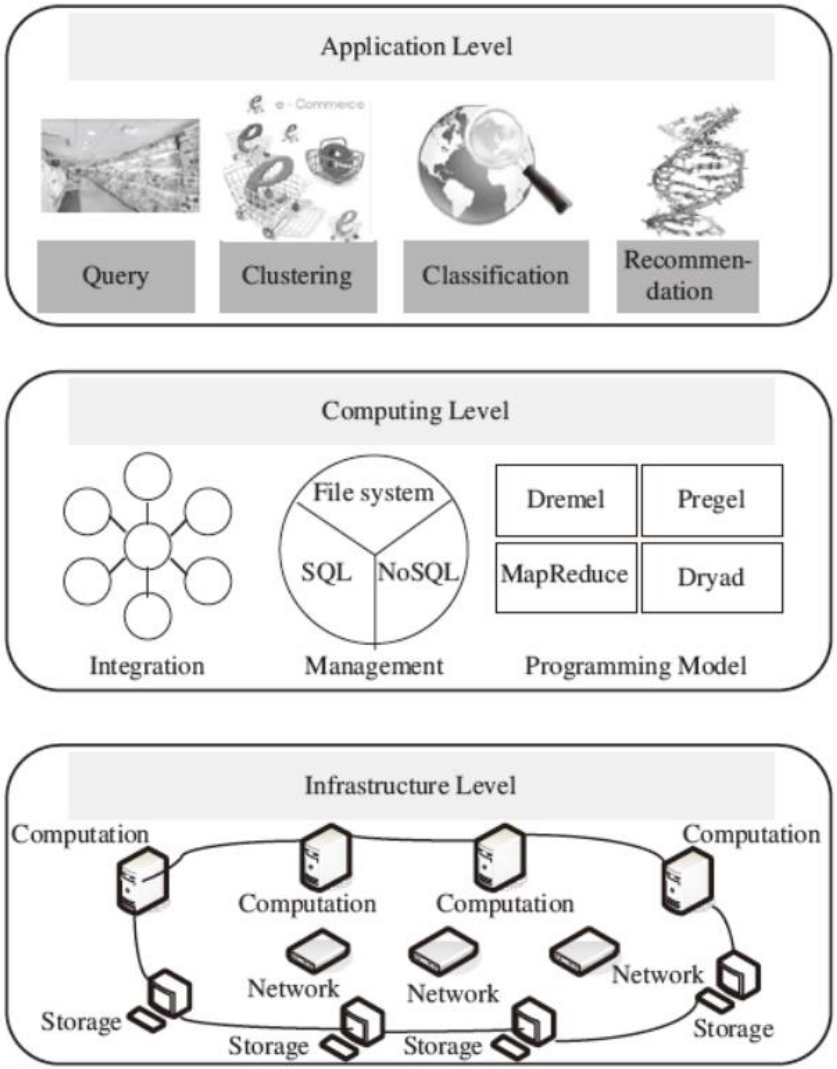
大数据基础-大数据系统架构

- 大数据系统：基于价值链的观点

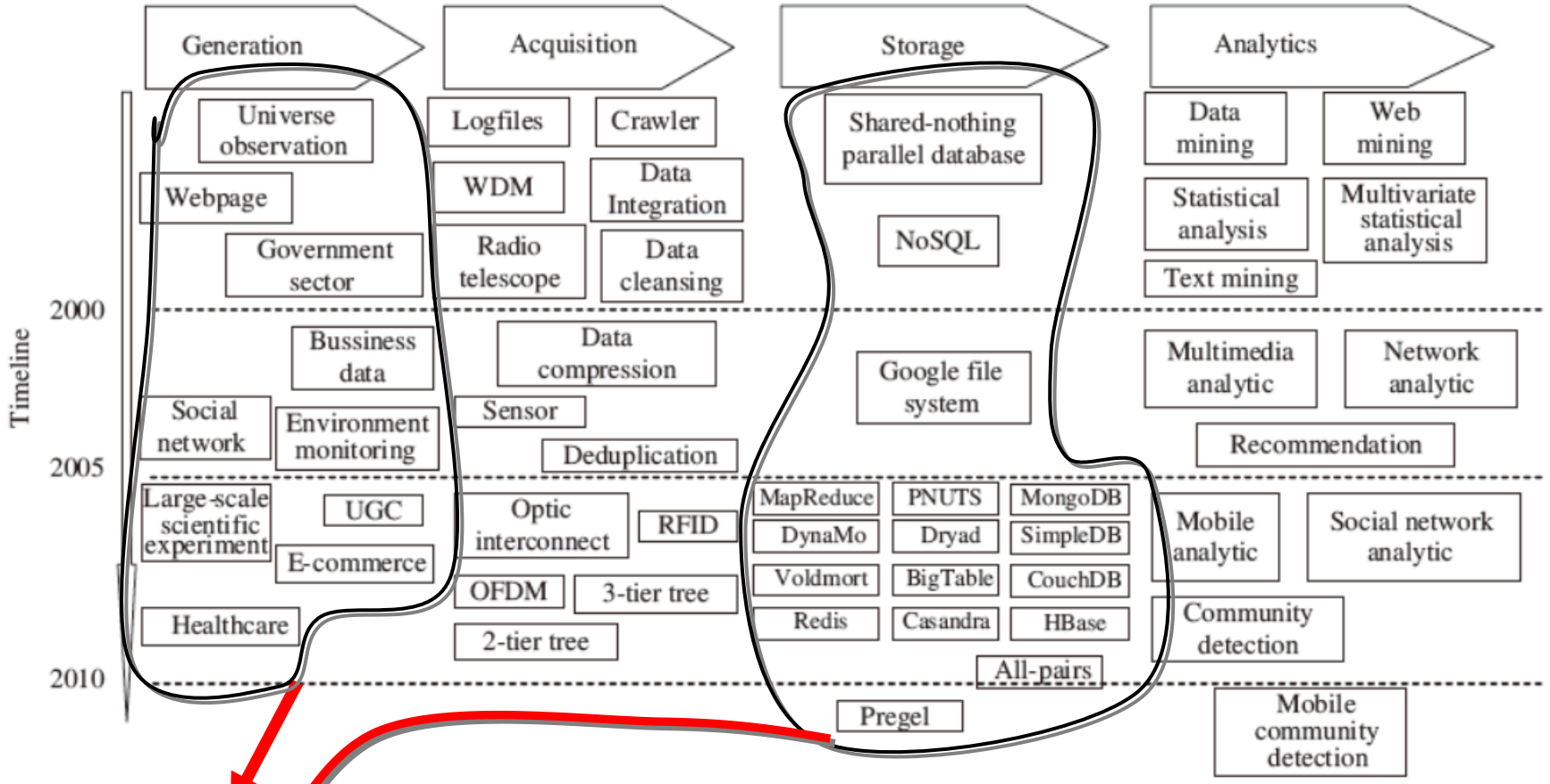


大数据基础-大数据系统架构

- 大数据系统：基于层次的观点

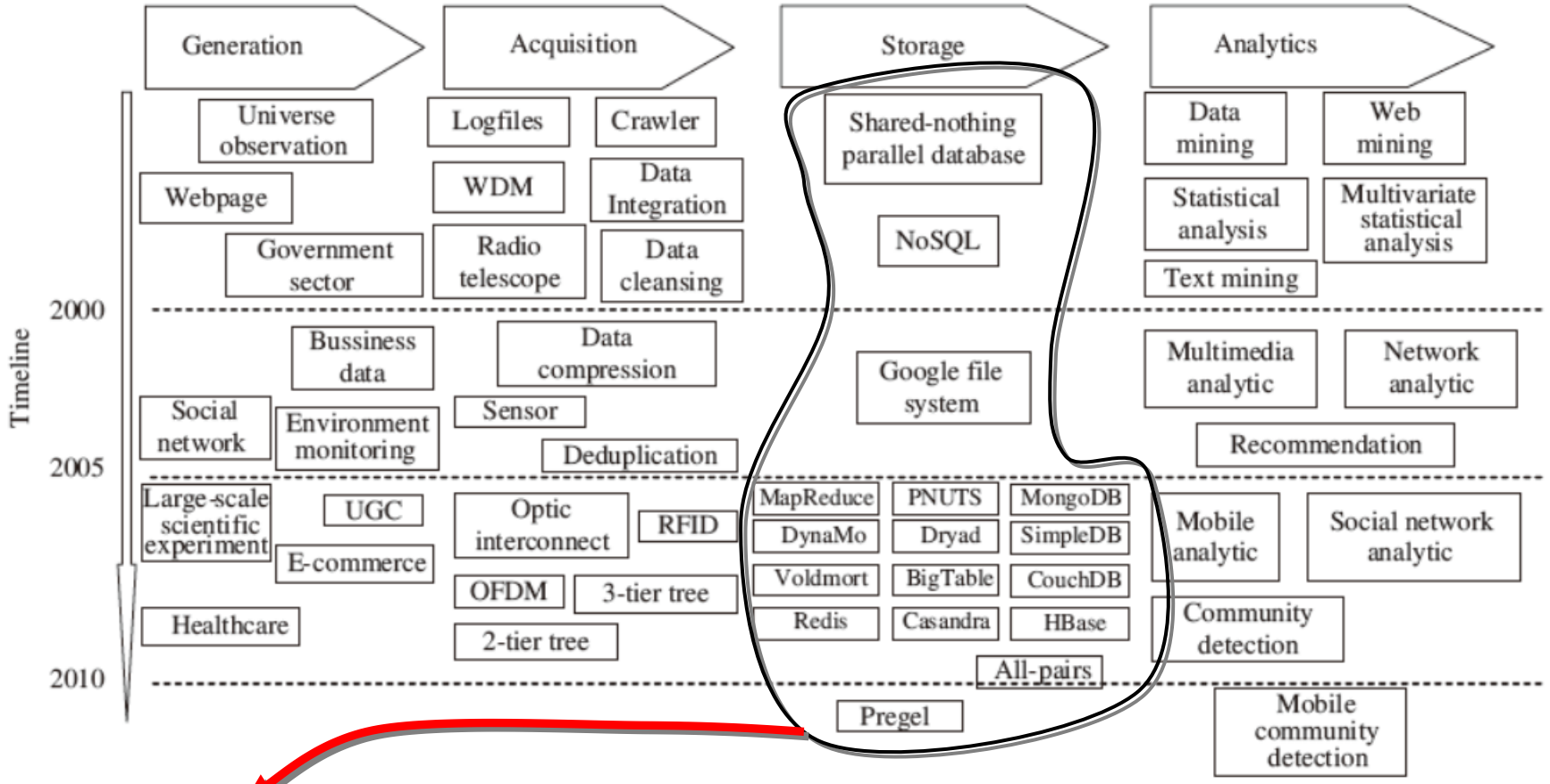


大数据基础-大数据系统面临的挑战



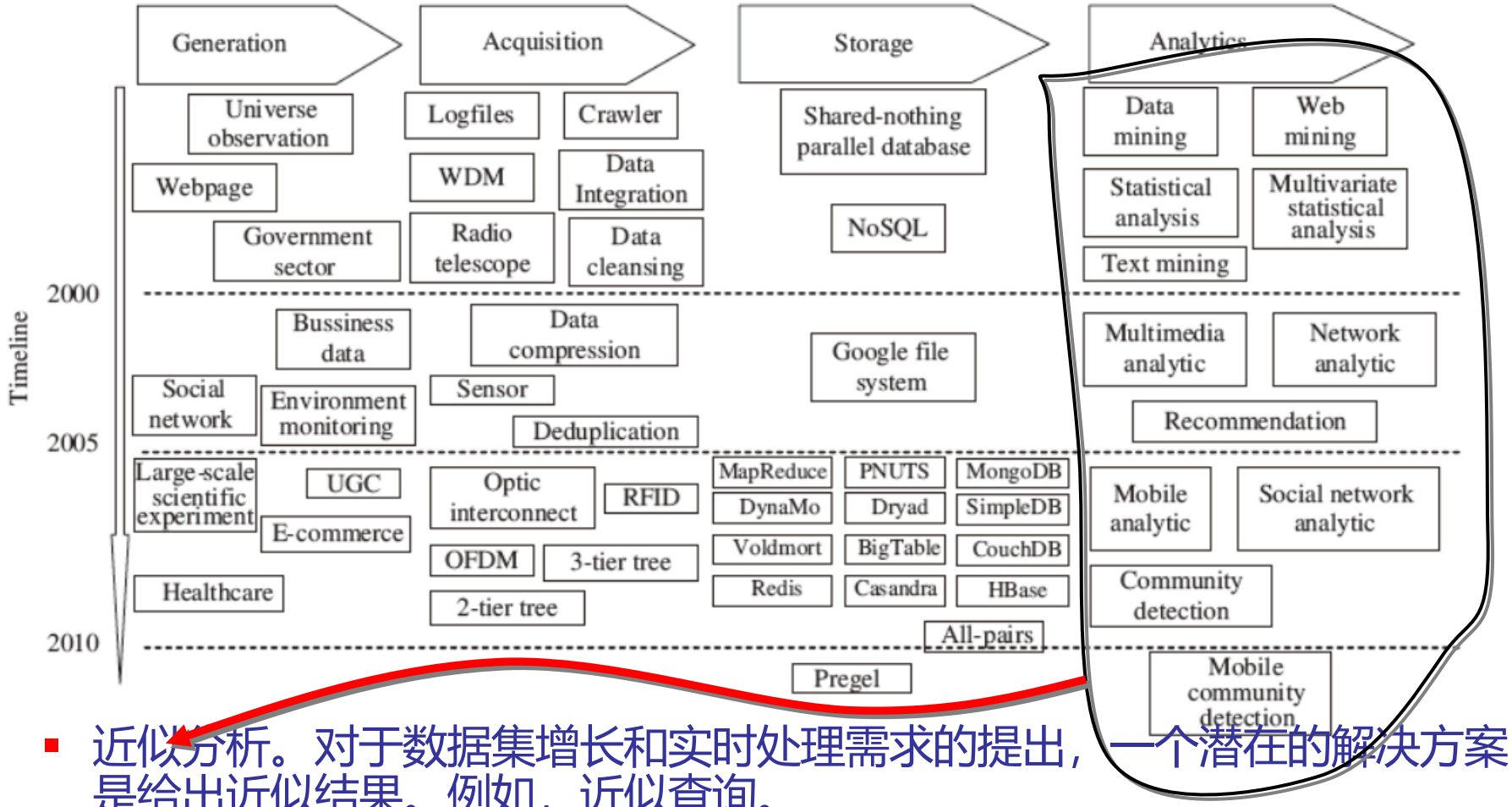
- 数据表示。数据集在类型、结构、语义、组织、粒度、可访问性等方面是异质的；通过**集成技术**实现跨数据集的有效操作。
- 冗余缩减（Redundancy reduction，RR）和数据压缩。不损毁数据价值的RR和数据压缩是减少系统整体开销的有效方法。

大数据基础-大数据系统面临的挑战



- 数据生存周期管理。现有的存储系统难以容纳海量数据，而数据的潜在价值与数据新鲜度有关；设置和隐藏价值相联系的**数据重要性原则**。
- 数据隐私和安全。了解需要提供什么样的系统级别隐私保护机制至关重要。

大数据基础-大数据系统面临的挑战



- 近似分析。对于数据集增长和实时处理需求的提出，一个潜在的解决方案是给出近似结果。例如，近似查询。
- 连接社交媒体。通过连接领域间的数据和社交媒体，应该能够获得更高的精确性。
- 深度分析。大数据的一个令人兴奋的研究动机是**期望获得新的领悟**。

大数据基础-大数据系统面临的挑战

- 大规模并行处理系统通常面临的几个共同问题：
  - 能量管理：

    随着数据量和分析需求的增长，数据传输、存储和处理无疑将消耗更多的能量。因此，在大数据系统中需要提供系统级的能量控制和管理机制。
  - 可扩展性：

    大数据系统中的所有组件都应能够扩展，以解决复杂数据集的日益增长等需求。
  - 协作性：

    大数据系统作为一个综合基础设施，允许不同领域的科学家和工程师访问多样的数据，并应用各自的专业知识，协作完成数据分析任务。

课程概览

- 大数据的价值
- 大数据基础
- 大数据系统架构
- **数据生成**
- 数据获取
- 数据存储
- 数据分析
- 大数据分析分类
- 大数据系统基准（benchmark）
- 大数据科学问题

数据生成-数据源

- 数据生成的模式可分为三个阶段：

阶段一：*20世纪90年代*

代表事件：数字技术和数据库系统的广泛使用。

阶段二：*20世纪90年代末，2000年初*

代表事件：半结构化和非结构化的数据；Web 2.0，从在线社交网络中产生大量**用户创造内容**。

阶段三：*移动设备的普及*

代表事件：为**个人为中心**和上下文相关的数据。

- **商业数据**

  全球所有公司商业数据每1.2年会翻番。例如，Amazon每天要处理几百万的后端操作和来自第三方销售超过50万的查询请求。

- **网络数据**

  互联网、移动网络和物联网；Google在2008年每天要处理20PB的数据，Facebook每天需存储、访问和分析超过30PB的用户创造数据；2010年全球已有40亿人持有手机；在物联网领域，有超过3000万的传感器工作在运输、汽车、公用事业部门并产生数据。

- **科学研究数据**

  学科：光学观测和监控、计算生物学、天文学、高能物理。

数据生成-数据源

- 典型大数据源

表2　典型大数据源

Data source	Application	Data scale	Type	Response time	Number of users	Accuracy
Walmart	Retail	PB	Structured	Very fast	Large	Very high
Amazon	e-commerce	PB	Semi-structured	Very fast	Large	Very high
Google search	Internet	PB	Semi-structured	Fast	Very large	High
Facebook	Social network	PB	Structured, unstructured	Fast	Very large	High
AT&T	Mobile network	TB	Structured	Fast	Very large	High
Health care	Internet of Things	TB	Structured, unstructured	Fast	Large	High
SDSS	Scientific research	TB	Unstructured	Slow	Small	Very high

数据生成-数据属性

- NIST提出了大数据的5种属性
  - 容量：数据集的大小
  - 速度：数据生成速率和实时需求
  - 多样性：结构化、半结构化和非结构化的数据形式
  - 水平扩展性：合并多数据集的能力
  - 相关限制：包含特定的数据形式和查询。数据的特定形式包括：时间数据和空间数据；查询则可以是递归或其它方式

课程概览

- 大数据的价值
- 大数据基础
- 大数据系统架构
- 数据生成
- **数据获取**
- 数据存储
- 数据分析
- 大数据分析分类
- 大数据系统基准（benchmark）
- 大数据科学问题

数据获取-数据采集

- 3种常用的数据采集方法：
  - 传感器
  - 日志文件
  - Web爬虫

表2　三种数据采集方法的比较

Method	Mode	Data structure	Data scale	Complexity	Applications
Sensor	Pull	Structured or unstructured	Median	Sophisticated	Video surveillance, Inventory management
Log file	Push	Structured or semi-structured	Small	Easy	Web log, click stream
Web crawler	Pull	Mixture	Large	Median	Search, social networks analysis

数据获取-数据传输

- 大数据传输：IP骨干网络-数据中心传输

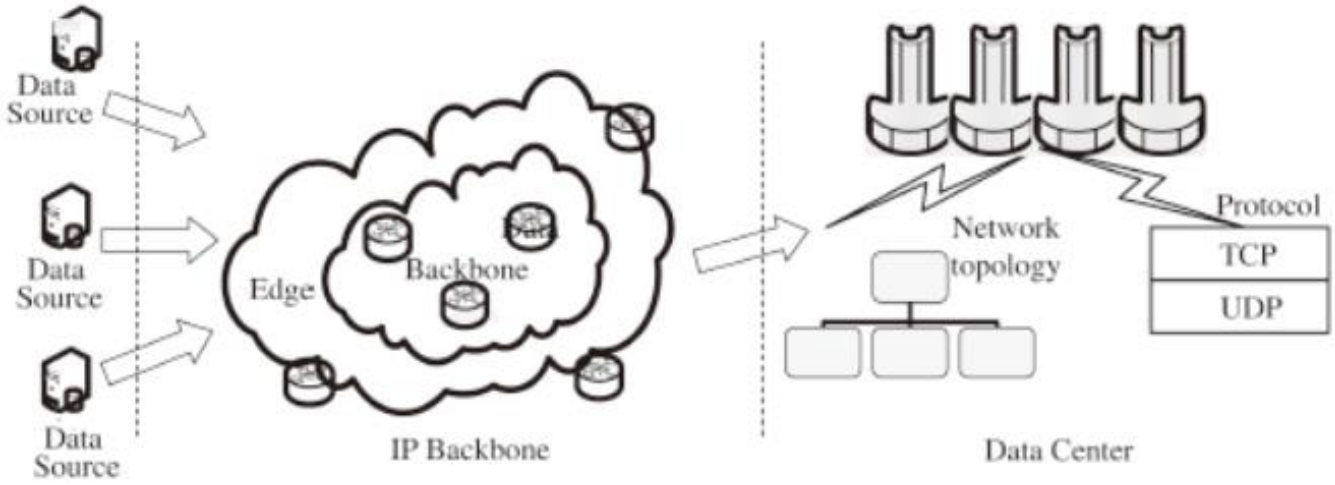


图1　大数据传输过程

数据获取-数据预处理

- 3种主要的数据预处理技术：
  - 数据集成（Data integration）：

    随着数据量和分析需求的增长，数据传输、存储和处理无疑将消耗更多的能量。因此，在大数据系统中需要提供系统级的能量控制和管理机制。
  - 数据清洗（Data cleansing）：

    大数据系统中的所有组件都应能够扩展，以解决复杂数据集的日益增长等需求。
  - 冗余消除（Redundancy elimination）：

    大数据系统作为一个综合基础设施，允许不同领域的科学家和工程师访问多样的数据，并应用各自的专业知识，协作完成数据分析任务。

课程概览

- 大数据的价值
- 大数据基础
- 大数据系统架构
- 数据生成
- 数据获取
- **数据存储**
- 数据分析
- 大数据分析分类
- 大数据系统基准（benchmark）
- 大数据科学问题

数据存储-数据管理框架

- 从层次的角度将数据管理框架划分为3层：文件系统、数据库技术和编程模型

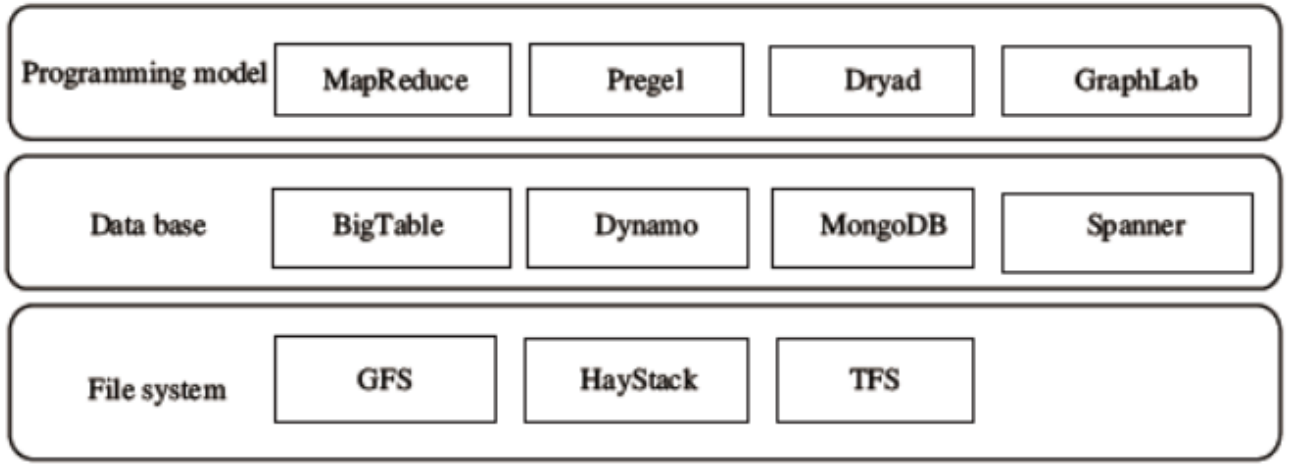


图1　数据管理技术

数据存储-数据库技术

- 键值存储数据库：数据以键值对的形式存储，键是唯一的。可扩展性和持久性主要依赖于两个关键机制：分割和复制、对象版本管理。
- 列式存储数据库：列存储架构，主要适合于批量数据处理和实时查询。
- 文档数据库：能够支持比键值存储复杂得多的数据结构。不同于文档存储的区别在于：数据复制和一致性机制方面。
- 其它NoSQL和混合数据库。

表2　NoSQL存储系统设计

Data model	Name	Producer	Data storage	Concurrency control	CAP option	Consistency
Key-value	Dynamo	Amazon	Plug-in	MVCC	AP	Eventually consistent
	Voldemort	LinkedId	RAM	MVCC	AP	Eventually consistent
	Redis	Salvatore Sanfilippo	RAM	Locks	AP	Eventually consistent
Column	BigTable	Google	GFS	Locks+stamps	CP	Eventually consistent
	Cassandra	Facebook	Disk	MVCC	AP	Eventually consistent
	HBase	Apache	HDFS	Locks	CP	Eventually consistent
	HyperTable	HyperTable	Plug-in	Locks	AP	Eventually consistent
Document	SimpleDB	Amazon	S3	None	AP	Eventually consistent
	MongoDB	10gen	Disk	Locks	AP	Eventually consistent
	CouchDB	Couchbase	Disk	MVCC	AP	Eventually consistent
Row	PNUTS	Yahoo	Disk	MVCC	AP	Timeline consistent

数据存储-数据管理框架

- 编程模型

表3　编程模型的特点

	MapReduce	Dryad	Pregel	GraphLab	Storm	S4
Application	General purpose parallel execution engine	General purpose parallel execution engine	Large scale graph processing	Large scale machine learning and data mining	Distributed stream processing	Distributed stream processing
Programming model	Map and Reduce	Directed acyclic graph	Directed graph	Directed graph	Directed acyclic graph	Directed acyclic graph
Parallelism	Concurrent execution within map and reduce phases	Concurrent execution of vertices during a stage	Concurrent execution over vertices within a superstep	Concurrent execution of non-overlaping scopes, defined by consistency model	Worker processes and executors	Worker processes and executors
Data handling	Distributed file system	Various storage media	Distributed file system	Memory or disk	Memory	Memory
Architecture	Master-slaves	Master-slaves	Master-slaves	Master-slaves	Master-slaves	Decentralized and symmetric
Fault tolerance	Node-level fault tolerance	Node-level fault tolerance	Checkpointing	Checkpointing	Partial fault tolerance	Partial fault tolerance

课程概览

- 大数据的价值
- 大数据基础
- 大数据系统架构
- 数据生成
- 数据获取
- 数据存储
- **数据分析**
- 大数据分析分类
- 大数据系统基准（benchmark）
- 大数据科学问题

数据分析-数据分析目的和分类

- 数据分析的主要目标：
  - 推测或解释数据并确定如何使用数据
  - 检查数据是否合法
  - 给决策制定合理建议
  - 诊断或推测错误原因
  - 预测未来发生的事情
- 数据分析分类（数据分析深度）：
  - **描述性（descriptive）分析**

    基于历史数据描述发生了什么，例如， 利用回归技术从数据中发现简单的趋势。
  - **预测性分析**

    用于预测未来的概率和趋势，例如，使用线性和对数回归等统计技术发现数据趋势。
  - **规则性分析**

    解决决策制定和提高分类效率，例如，使用优化技术在给定条件下得到最优解决方案。

数据分析-应用演化

- 不同时期典型大数据领域中具有高影响力的大数据应用的发展：
  - 商业应用演化
  - 网络应用演化
  - 科学应用演化

数据分析-常用分析方法

- 三种类型的常用数据分析方法：
  - 数据可视化

    图表、地图可以帮助人们快速理解信息。Tabusvis是一个轻量级的可视化系统。
  - 统计分析

    可分为描述性统计和推测性统计。描述性统计技术对数据集进行摘要或描述，而推测性统计则能够对过程进行推断。例如，回归、聚类和判别分析等。
  - 数据挖掘

    2006年ICDM会议上总结了影响力最高的10种数据挖掘算法，包括C4.5、K-means、SVM、Apriori、EM、PageRank、Adaboost、KNN、朴素贝叶斯和CART，涵盖了分类、聚类、回归和统计学习等方向。

课程概览

- 大数据的价值
- 大数据基础
- 大数据系统架构
- 数据生成
- 数据获取
- 数据存储
- 数据分析
- **大数据分析分类**
- 大数据系统基准（benchmark）
- 大数据科学问题

大数据分析分类

- 大数据分析分类
  - 结构化数据分析
  - 文本分析
  - Web数据分析
  - 多媒体数据分析
  - 社交网络数据分析
  - 移动数据分析

Analysis domains	Sources	Characteristics	Approaches
Structured data analysis	Customer transactions	Structured records	Data mining [152]
	Scientific data	Less volume and real time	Statistical analysis [151]
Text analysis	Logs	Unstructured	Document presentation [153]
	Email	Rich textual	NLP [154]
	Corporate documents	Context	Information extraction [155, 156]
	Government regulations	Semantic	Topic model [157]
	Text content of webpages	Language dependent	Summarization [158]
	Feedback and comments		Categorization [159]
			Clustering [160]
			Question answering [161]
			Opinion mining [162]
Web analytics	Various webpages	Text and hyperlink	Web content mining [163]
		Symbolic	Web structure mining [164~166]
		Metadata	Web usage mining [167]
Multimedia analytics	Corporation and user	Image, audio, video	Summarization [168], Annotation [169]
	Generated multimedia	Massive	Indexing and retrieval [170]
	Surveillance	Redundancy	Recommendation [171, 172]
	Health and patient media	Semantic gap	Event detection [173]
Social network analytics	Bibliometric	Rich content	Link prediction [174~176]
	Sociology network	Social relationship	Community detection [177, 178]
	Social networks	Noisy and redundancy	Network evolution [179~182]
		Fast evolution	Influence analysis [183, 184]
			Key words search [185]
			Classification [186], Clustering [187]
			Transfer learning [188, 189]
Mobile analytics	Mobile apps	Location based	Monitoring [190~192]
	Sensors	Person specific	Location based mining
	RFID	Fragmented information	

课程概览

- 大数据的价值
- 大数据基础
- 大数据系统架构
- 数据生成
- 数据获取
- 数据存储
- 数据分析
- 大数据分析分类
- **大数据系统基准（benchmark)**
- 大数据科学问题

大数据系统基准-挑战与研究现状

- **面临的挑战**
  - 系统复杂性：大数据系统通常由多个模块和组件构成，功能各异，相互耦合，对整个系统建模以及为所有模块提供一个统一的框架并不容易。
  - 应用多样性：一个好的基准应该反映大数据系统的典型特征。由于大数据系统的多样性，使得提取显著特征较为复杂。
  - 数据规模：大数据的数据量巨大并且不断增长，需要考虑一种有效的方式测试小数据的产品。
  - 系统演化：针对大数据增长和日益变化的需求，大数据基准也要迅速变化。

- **组件级别的基准和系统级别的基准**
  - 系统级别的基准通常提供端对端系统测试框架
  - 组件级别的基准也被称为微基准（micro benchmark），用于评价独立组件的性能，主要包括三类：TPC基准、NoSQL基准、Hadoop基准。

课程概览

- 大数据的价值
- 大数据基础
- 大数据系统架构
- 数据生成
- 数据获取
- 数据存储
- 数据分析
- 大数据分析分类
- 大数据系统基准（benchmark）
- **大数据科学问题**

大数据科学问题

- **大数据基础平台**
  - Hadoop集群，Spark集群，海量数据的分布式存储管理平台，基于SDN的大数据平台。
- **处理模式**
  - 现有的批处理模式难以适应海量数据实时处理的需求，需要设计新的实时处理模式。
  - 在处理模式上需要考虑新的因素，算法、传输、存储、可视化等。涉及大数据处理复杂性问题、并行化机器学习/深度学习算法、异构数据融合、基于海量数据低价值密度采样问题、高维海量降维问题等。
- **大数据应用**
  - 社会网络分析、媒体数据检索、自然语言处理等。
- **大数据隐私**
  - 增强访问控制。
  - 在增强访问控制与数据处理的便利性之间找到平衡。

大数据科学问题

- **”无限”数据**
  - 面对不断演化的”无限”大数据，需要研究新型增量学习方法，从而动态自适应地进行预测并确保模型的准确率，这或许会是大数据未来发展需要解决的重要问题。