贵州大學
GUIZHOU UNIVERSITY

大数据分析与应用

杜逆索
贵州省大数据产业发展应用研究院

课程内容

- 大数据概念
- 大数据计算体系
- 数据采集与建模
- 数据清洗与数据预处理
- 大数据分析算法
- 文本读写技术
- 数据处理技术
- 数据分析技术
- 数据可视化
- 大数据应用案例

教学模式

- 讲课 + 案例分析
- 阅读课本 + 课外资料
- 课堂讨论 + 课外交流
- 课后作业 + 课堂练习

评分标准

平时成绩 （随机点名、书面作业、课堂练习）	40%
期末考试	60%
Total	100%

教材与参考书

教材

汤羽、林迪等编著，**《大数据分析与计算》**，

清华大学出版社，2017年9月第1版

教材与参考书

参考书

（美）Rachel Schutt, Cathy O’Neil 著，《数据科学实战》，人民邮电版社，2015年3月第1版

魏琴，欧阳智，袁华，《数据未来——图解大数据+产业融合》，贵州人民出版社，2018年5月第1版

Lecture 1 大数据计算概论

1.1 大数据概念

1.2 大数据技术特征

1.3 标准与模式

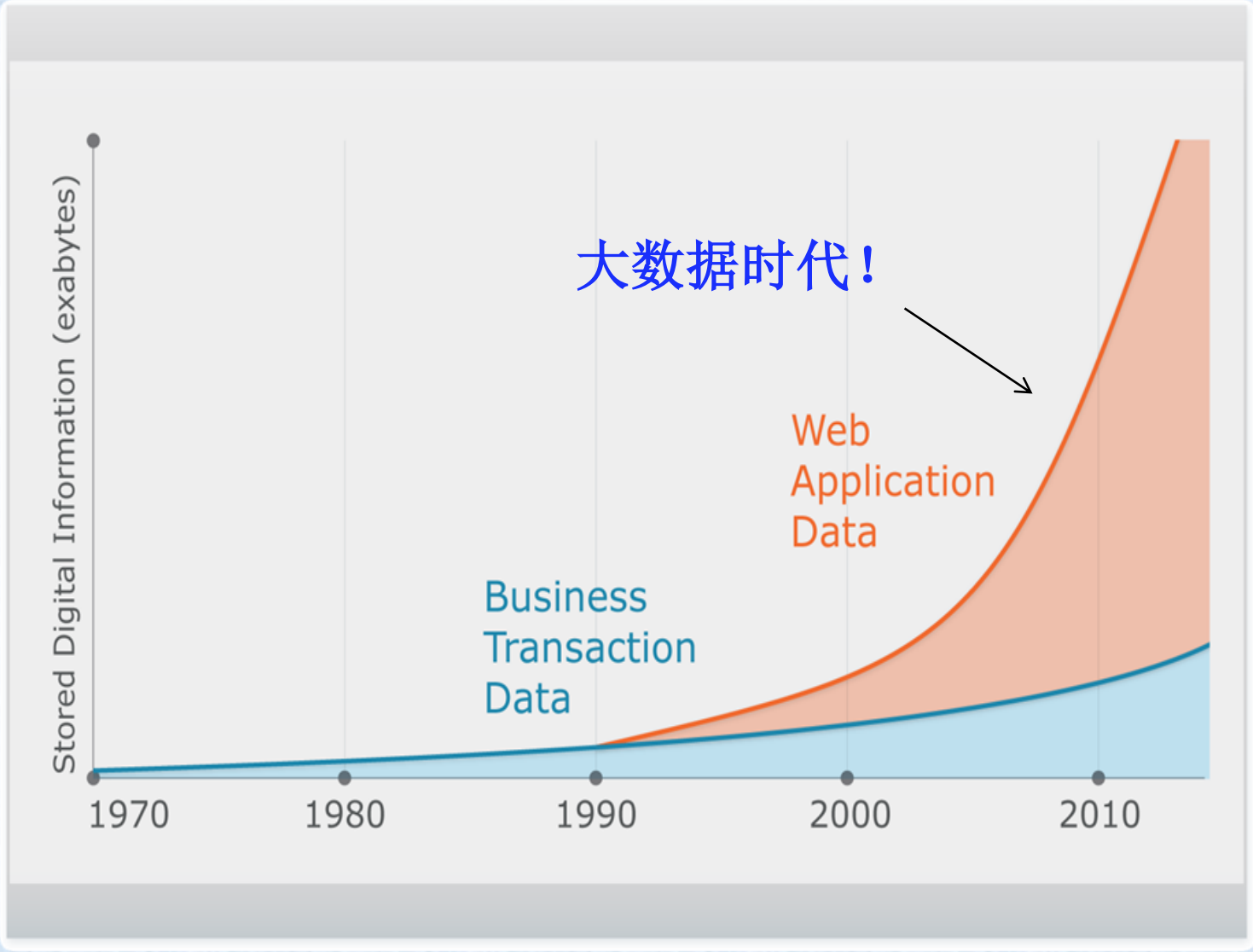
1.1 大数据概念

- 数据是什么？
- 数据科学是什么？
- 大数据基本属性是什么？

全球视野下的大数据：机遇与挑战

“黄河之水天上来”！

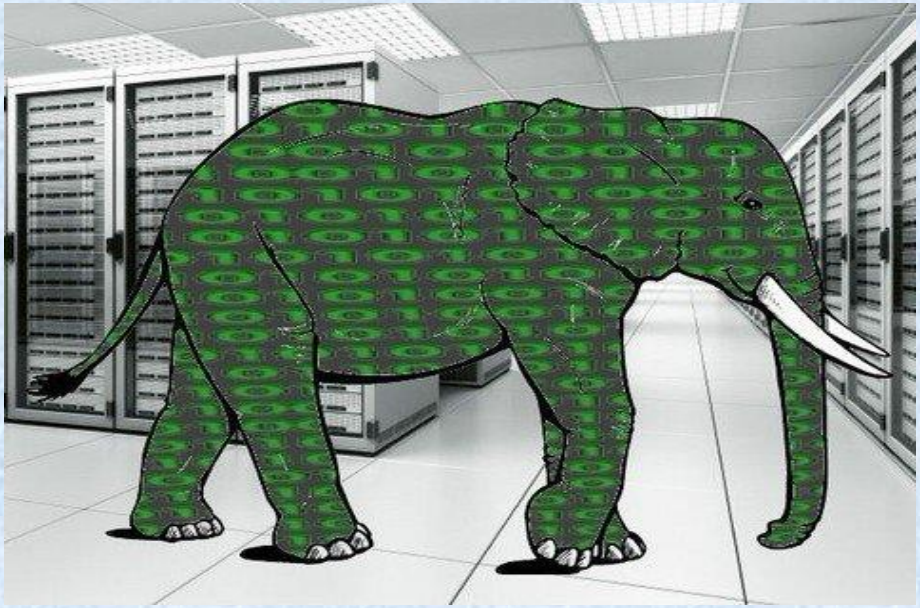
- Facebook每天处理80亿条信息
- Google每天完成10亿次查询
- 全世界的信息量以每两年翻番的速度增长
- 2011年全球数据量为1.8ZB，IDC预测2015年达到8ZB, 2020年更将达到35ZB！



1 EB = 10^3 ZB = 10^6 PB = 10^9 TB = 10^{12} GB

■ 什么是大数据（Big Data）？

- Volume:　数据量异常庞大，一般达到PB量级
- Variety:　数据呈异构化，数据来源呈多样性
- Velocity: 数据处理要求时效性
- Value:　单个数据无价值，但大规模数据拥有巨大价值



■ 什么是大数据？（续）

- 数据种类的多样性：文字、语音、图片、视频、信息等
- 数据对象的多样性：个人信息、个人数据、商业服务数据、社会公共数据、自然界数据、物质世界的数据
- 数据来源的多样性：在数据层面打破现实世界的界限，多家公司的共享替代一家公司的数据

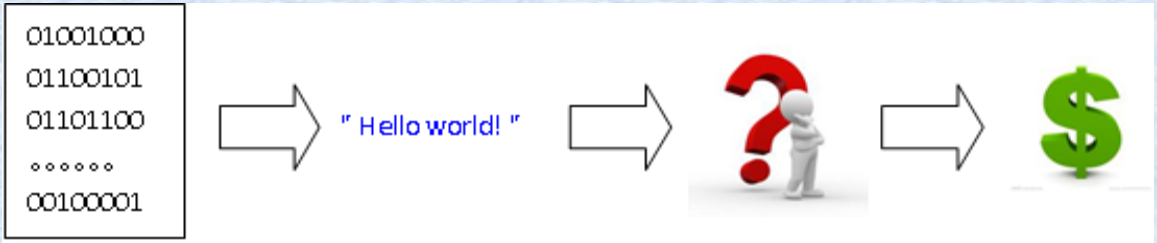
■ 大数据已上升到21世纪国家战略的高度

2012年3月美国奥巴马政府宣布推出“*大数据的研究和发展计划*”，包括

- 美国国家科学基金（NSF）
- 美国国家卫生研究院（NIH）
- 美国能源部、美国国防部
- 美国国防部高级研究计划局、美国地质勘探局等6个联邦政府部门

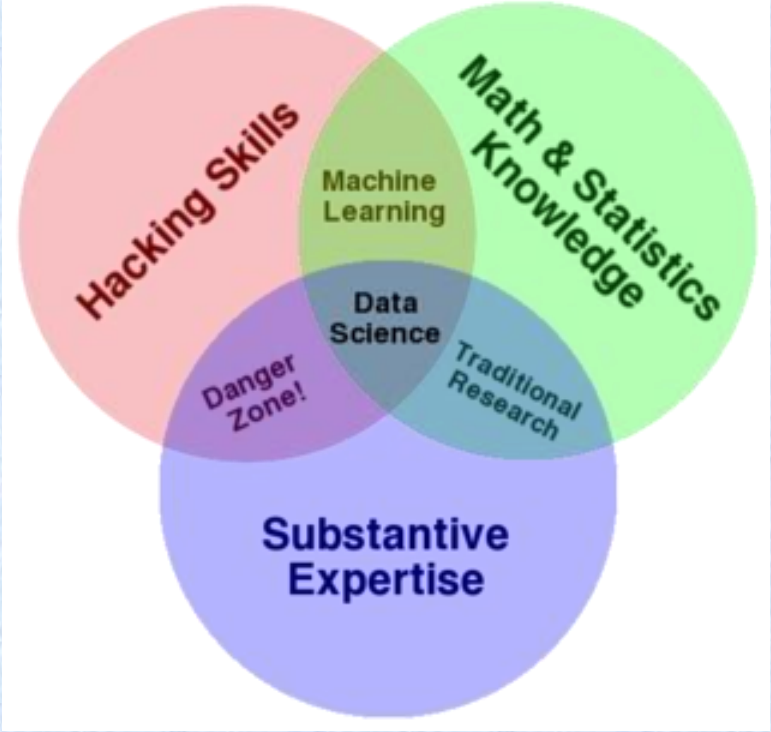
1.1 大数据概念——数据的定义

- 数据的定义
  - 数据的基本定义
  - 计算机学科中数据的定义
- 数据的多样化
  - 数据的形式多样化
  - 数据的来源多样化
  - 数据的范围多样化
- 数据转换过程
  - 数据-信息-知识-价值转换模型



1.1 大数据概念—数据科学

- 数据科学基本理解
- 数据科学六大研究方面
- 数据科学整体知识结构



1.1 大数据概念——基本属性

- Volume：大数据的超大规模
  - 规模体现
  - 带来的影响：
    - 数据存储架构：
      - 基于行-键表格存储格式的关系型数据库？
      - 基于分布式文件系统的分布式数据库！
    - 计算模型：
      - 离线批处理计算框架（MapReduce）
      - BSP图并行计算框架（Pregel、Hama）
      - 交互式计算模型
      - 大内存计算系统

1.1 大数据概念——基本属性

- **Variety：大数据来源多样性与异构性**
  - 大数据类型划分：
    - 依结构特征划分
    - 依时效性划分
    - 依关联特性划分
    - 依数据类型划分
    - 依数据来源划分
  - 带来影响：
    - 数据存储、管理和快速查询异常困难

1.1 大数据概念——基本属性

- Value：价值低密度特性
  - 区别于传统数学统计学方法的关键之处

	传统数学统计学	大数据分析计算方法
处理对象	局部数据或数据子集	以数据整体或完整数据集作为处理对象
处理方法	基于抽样调查的随机分析方法	机器学习方法 通过数据的积累来训练和改进算法和计算程序
结果正确性	取决于随机抽样模型产生的数据集的代表性	处理数据量越大，计算结果越越优化

1.2 大数据技术特征

- 大数据算法特性
- 大数据计算系统特性
- 大数据开发技术特性

1.2.1 大数据算法特性

	大数据计算	传统统计学	优势
样本空间	整个数据集	基于独立同分布原理抽取样本集	避免样本失真
计算方法	机器学习方法	按照固定数学模型进行预测	预测结果的精度改进是一个动态过程

1.2.2 大数据计算系统特性

	大数据计算系统	传统数据库系统	优势
基础模型	分布式文件系统 NoSQL非关系型数据库	关系型模型	支持非结构化或异构数据的存储和处理 支持分布式系统部署 支持超大规模数据集完成快速查询操作
存储格式	基于键值对的列存储格式	基于主键的行存储格式	更优的查询效率 更好的对计算模型的支持

1.2.2 大数据计算系统特性

某大学学生总数
N=30000

数据库中每个学生相关值域数量m=50

从数据库中搜出并计算某一专业学生（含不同年级）某一门课的平均成绩？

关系型数据库：

从数据库总表中搜出满足上述条件的学生记录，操作次数是O(N)量级

对搜出的每一条学生记录完成该门课程成绩的读取，操作次数是O(m)

总操作次数为O(N) * O(m) 量级，最坏情况下需要操作 30000 x 50 = 1500000次！

1.2.2 大数据计算系统特性

某大学学生总数
N=30000

数据库中每个学生相关值域数量m=50

从数据库中搜出并计算某一专业学生（含不同年级）某一门课的平均成绩？

NoSQL数据库：

所有学生的成绩都存入树状结构的某一分枝

搜索进入该门课的分枝
（最坏情况下查询次数2000）

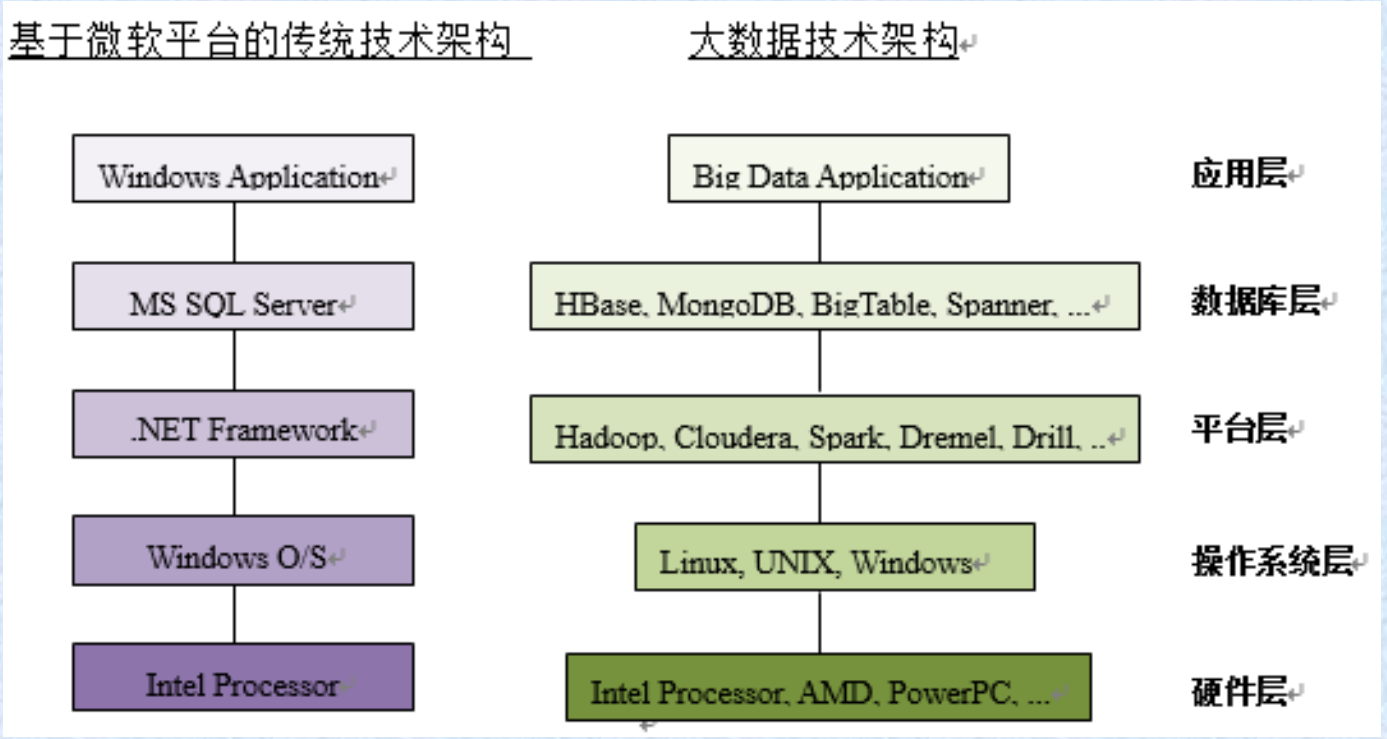
在该分枝内搜索该专业
（最多查询次数100）

完成符合条件的学生成绩的读取（最多读取1000次）

总的操作次数为：
2000 + 100 + 1000 = 3100次

1.2.3大数据开发技术特性

大数据计算系统	传统数据库系统	优势
多层次的分层结构	基于某一平台和某一标准的线性结构	在同一平台上尽可能多的兼容或集成不同的软件开发工具



1.3 技术标准与模式

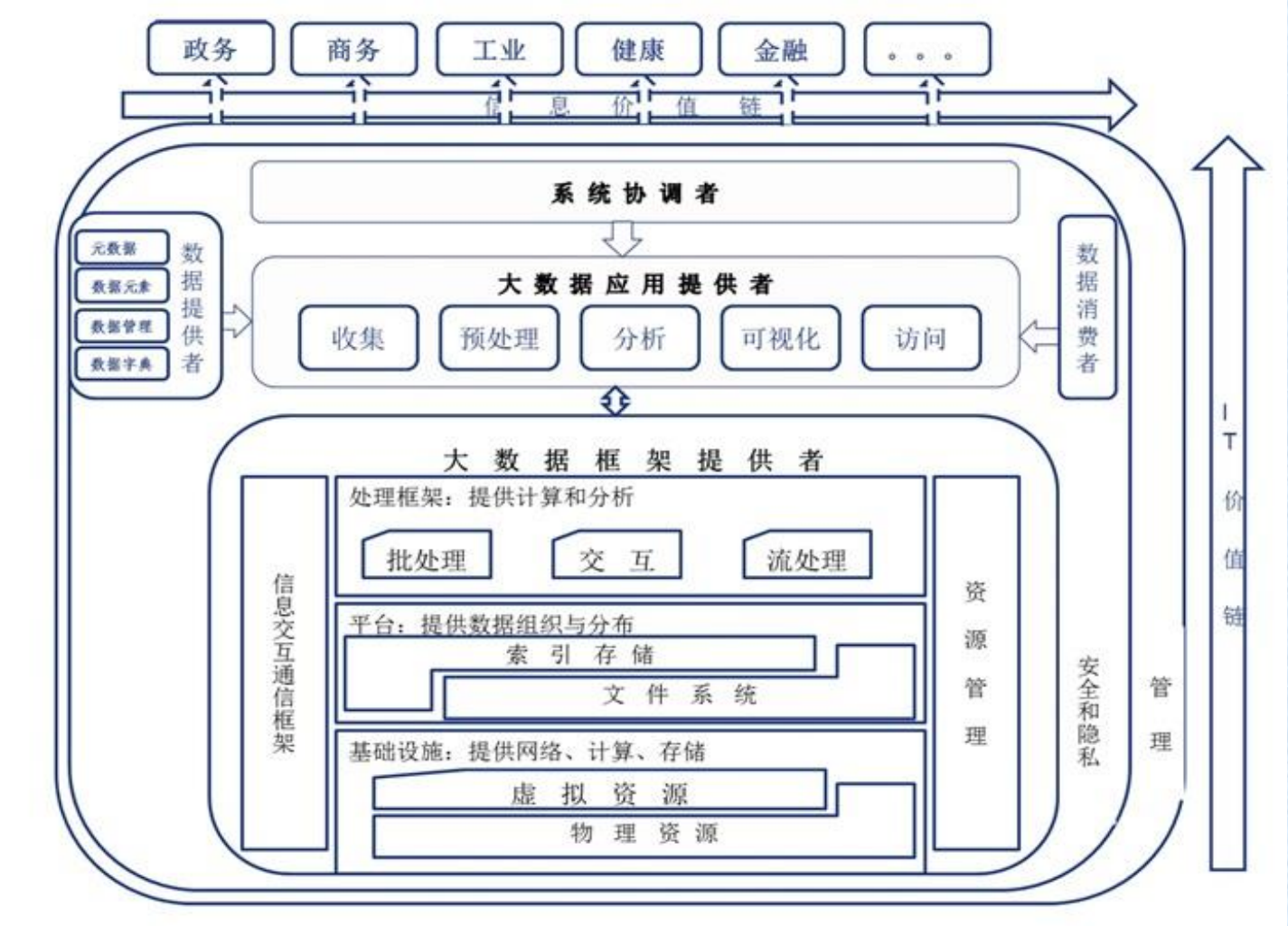
大数据计算技术标准

- 大数据技术架构参考模型
- 大数据计算体系主要角色
- 大数据标准体系框架

大数据计算模式

- 主要计算模式
- 各计算模式特性与优劣
- 大规模并行处理模式

1.3.1 大数据计算技术标准



大数据技术架构参考模型

1.3.1 大数据计算技术标准

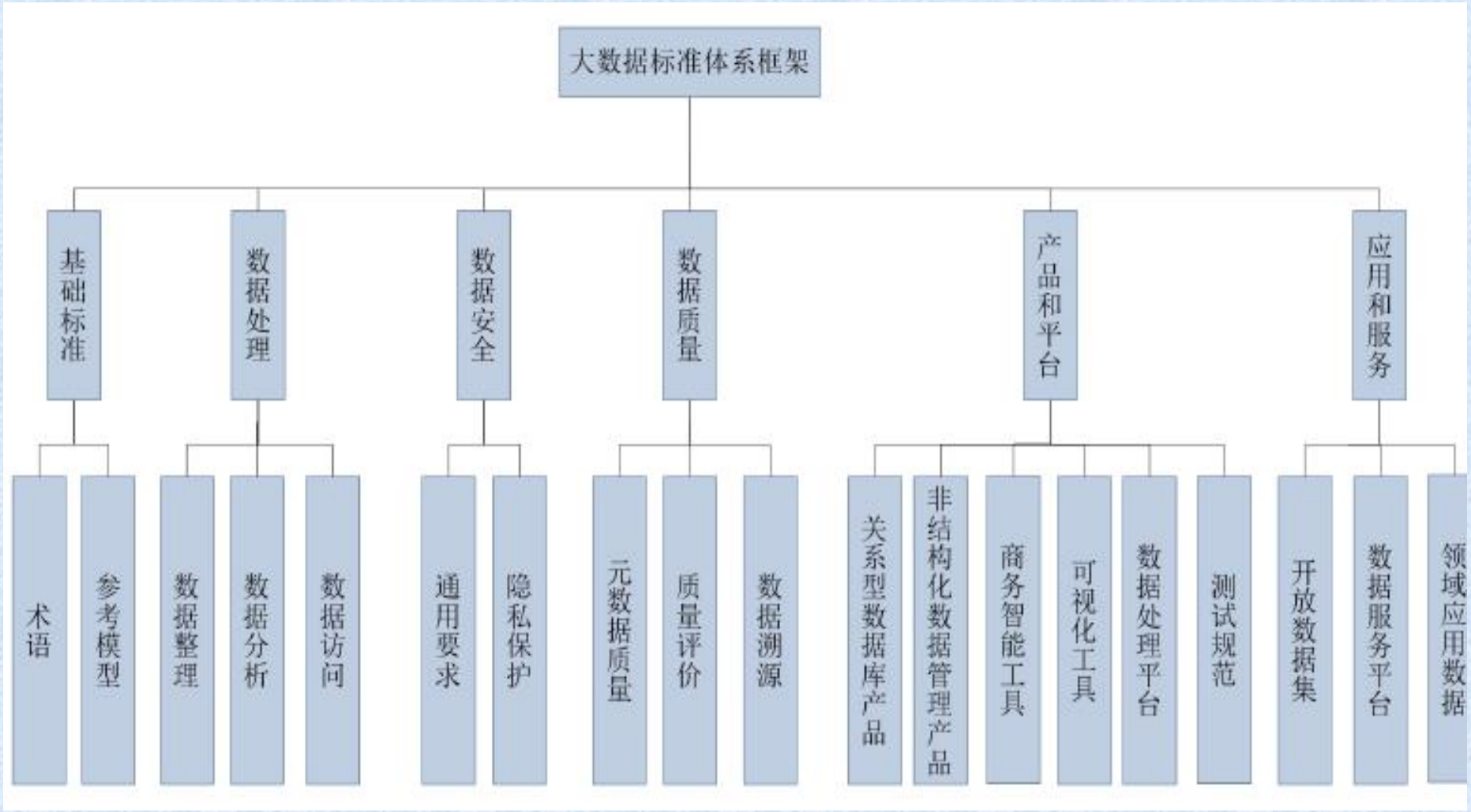
- 大数据技术架构参考模型基于两个维度组成：信息链（垂直方向）和价值链（水平方向）
- 信息链维度：通过数据采集、集成、分析、使用结果来实现价值
- 价值链维度：通过为大数据应用的实施提供拥有或运行大数据的网络、基础设施、平台、应用工具以及其他IT 服务来实现价值

1.3.1 大数据计算技术标准

大数据计算体系主要角色：

- 系统领导者
- 数据提供者
- 安全和隐私角色
- 大数据应用提供者
- 大数据基础框架提供者
- 数据消费者
- 管理角色
- 安全及隐私管理角色

1.3.1 大数据计算技术标准



大数据标准体系

1.3.1 大数据计算技术标准

大数据标准体系框架组成

- 基础标准
- 数据处理标准
- 数据安全标准
- 数据质量标准
- 产品和平台标准
- 应用和服务标准

1.3.2 大数据计算模式

主要计算模式

- 批处理模式（MapReduce）
- 图计算模式（BSP）
- 流计算模式（流计算模型）
- 内存计算模式（大内存计算）
- 大规模并行处理模式（NUMA）

1.3.2 大数据计算模式

MapReduce（侧重吞吐量）

- 优：
  - 基于现有廉价商业硬件和成熟技术
  - 成本低
  - 在可处理超大规模数据集时有计算优势
  - 吞吐量大
- 劣：
  - 计算耗时长
  - 无法支持在线快速智能分析这类运用

1.3.2 大数据计算模式

内存计算模式（侧重处理时延）

- 特点：
  - 将DRAM内存集群作为主存储介质，构成大规模集中式内存结构（如内存云），计算数据一次装载入内存
- 优：
  - 计算速度快
  - 非常适宜于低时延要求的实时在线分析
- 劣：
  - 成本高
  - 持久性和可靠性尚未得到验证
  - 内存云受到外部网络速度迟缓的限制

1.3.2 大数据计算模式

图计算模式（侧重数据吞吐量）

- 优：
  - 处理数据量大
  - 优化图计算问题的处理
- 劣：
  - 不支持在线实时处理

1.3.2 大数据计算模式

流计算模式（侧重处理时延）

- 特点
  - 流数据针对的是动态数据流的实时处理，其一个计算任务（或一次循环）处理的数据量并不大
  - 计算时延短，针对流数据（stream data）

1.3.2 大数据计算模式

交互式计算模式

- 特点
  - 采用现有的分布式系统架构（Google的GFS/BigTable，开源社区的Hadoop/HDFS/Hive），
  - 通过改造数据存储结构和算法创新（如列存储结构,数据本地化，提高内存驻存率等）来降低计算耗时
- 优
  - 避免了物理大内存技术的高昂成本
  - 在计算架构和网络接口方面与现有体系能更好地集成
  - 可靠性也更可信

1.3.3大规模并行处理模式 (Massively Parallel Processing, MPP)

组成结构

- 系统由多个松耦合的处理单元组成
- 每个单元内的 CPU都有自己的本地资源如总线，内存，硬盘等
- 在每个单元内都有操作系统和数据库系统

结构特性

- 不共享资源(shared nothing)

1.3.3大规模并行处理模式MPP

MPP模式特征：

- 任务执行并行化
- 数据分布式存储(本地化)
- 分布式计算架构
- 计算节点私有资源
- 横向扩展性好（易于加入新的处理节点）
- Shared Nothing架构

1.3.3大规模并行处理模式MPP

MPP数据库特征：

- 具备ACID特性：满足原子性、一致性等要求
- 支持关系型模型，支持基于关系模型的数据库设计
- 使用SQL标准接口（支持ODBC和JDBC），易于开发，应用迁移方便
- Share Nothing架构特点使其可以横向扩展数百个节点，支撑PB级别的数据处理
- 特别擅长处理结构化数据，有明显的星型和雪花模型结构，便于进行OLAP分析和多维分析
- 可部署于开放架构的X86服务器，平台建设成本低