大数据浪潮下的数据仓库

吕永超

技术工程部-基础数据部

2008年4月	北京邮电大学，数据仓库与数据挖掘方向
2007.12-2011.2	中国雅虎、阿里巴巴B2B 数据仓库工程师
2011.2-2014.4	百度 全网用户数据整合 资深研发工程师
2014.4至今	数据仓库和产品组负责人、数据公会负责人

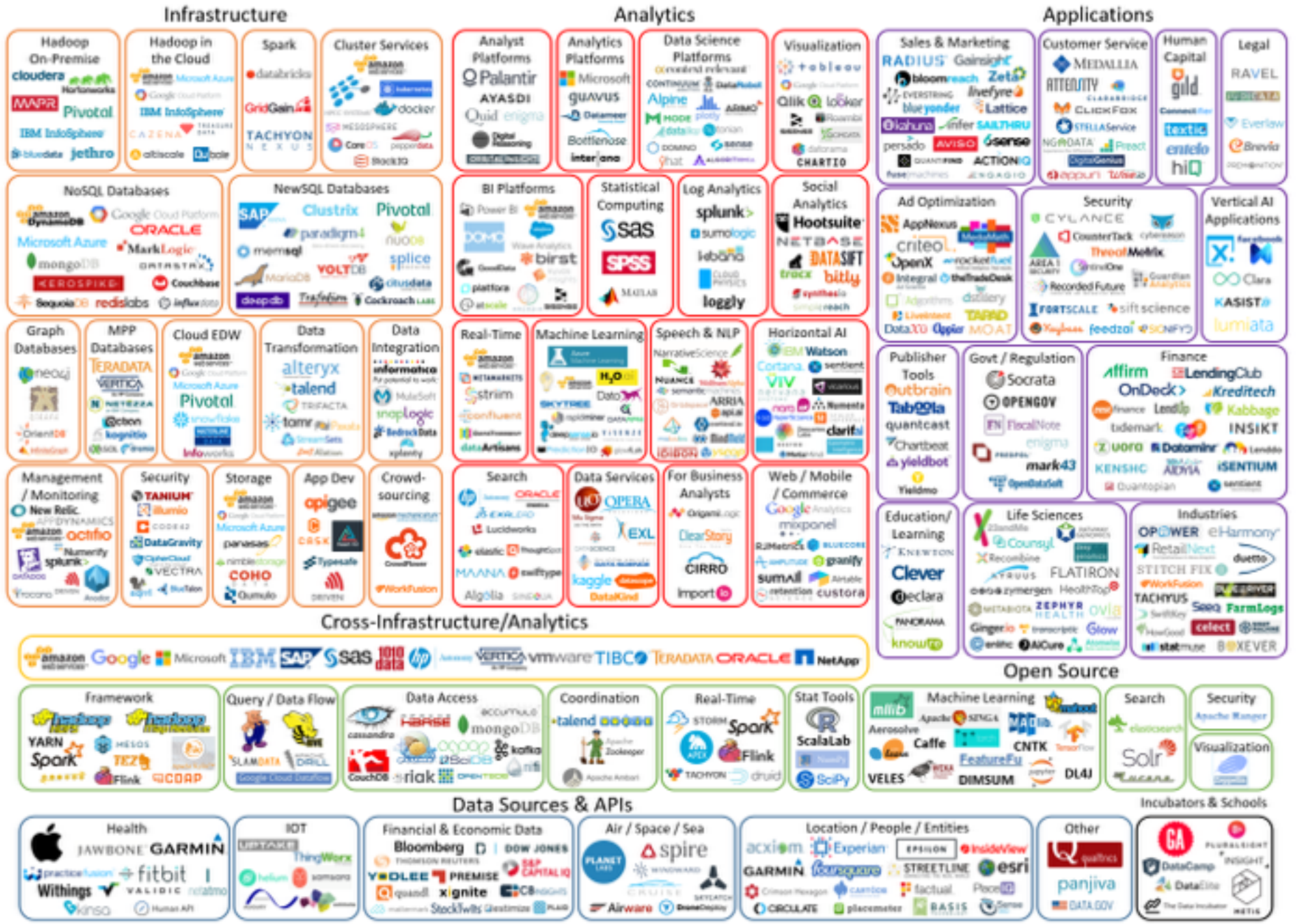
议题

- 大数据的特点、应用场景及其相关技术
- 数仓之本质
- 数仓的规划及成功因素
- 美大数仓体系结构
- 数仓研发工程师的成长之路

议题

- 大数据的特点、应用场景及其相关技术
- 数仓之本质
- 数仓的规划及成功因素
- 美大数仓体系结构
- 数仓研发工程师的成长之路

Big Data Landscape 2016 (Version 3.0)



Last Updated 3/23/2016

FIRSTMARK

大数据?

- 数据容量大、增长速度快、数据类型及来源丰富、潜在价值高(4V)等特征的数据集
- 对数量巨大、来源分散、格式多样的数据进行采集、存储和关联性分析的新一代信息系统架构和技术
- 数据思维——用数据说话，按理性思维的科学精神，并从信息社会海量数据中发现新知识、创造新价值的能力

大数据体系= 原材料 + 处理技术 + 使用方法

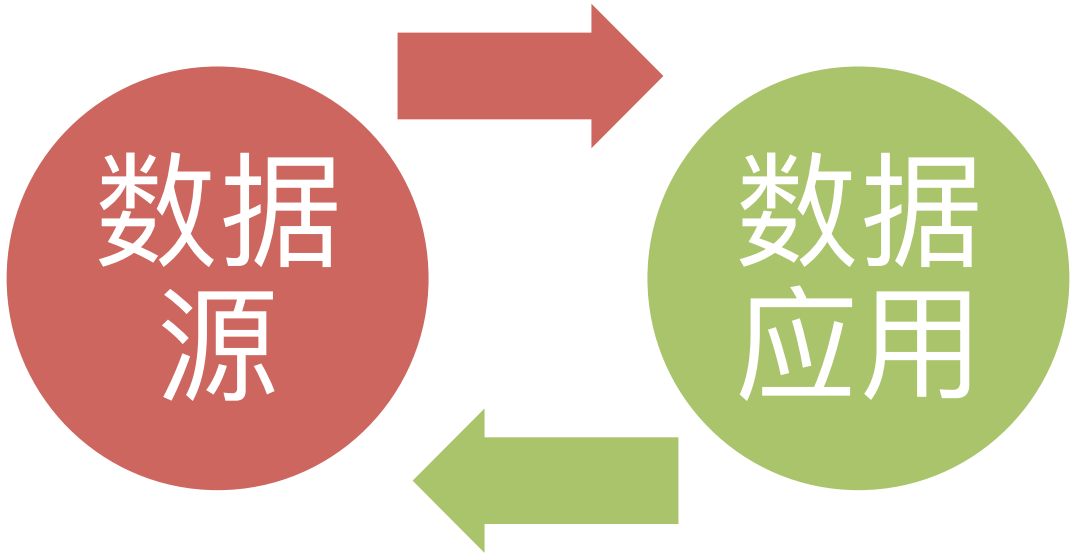
大数据的真正价值

- 互联网
- 传统企业
- 公共事务

大数据的真正价值在于创造，
在于填补无数个以前还未实现过的空白

当数仓遇到了大数据.....

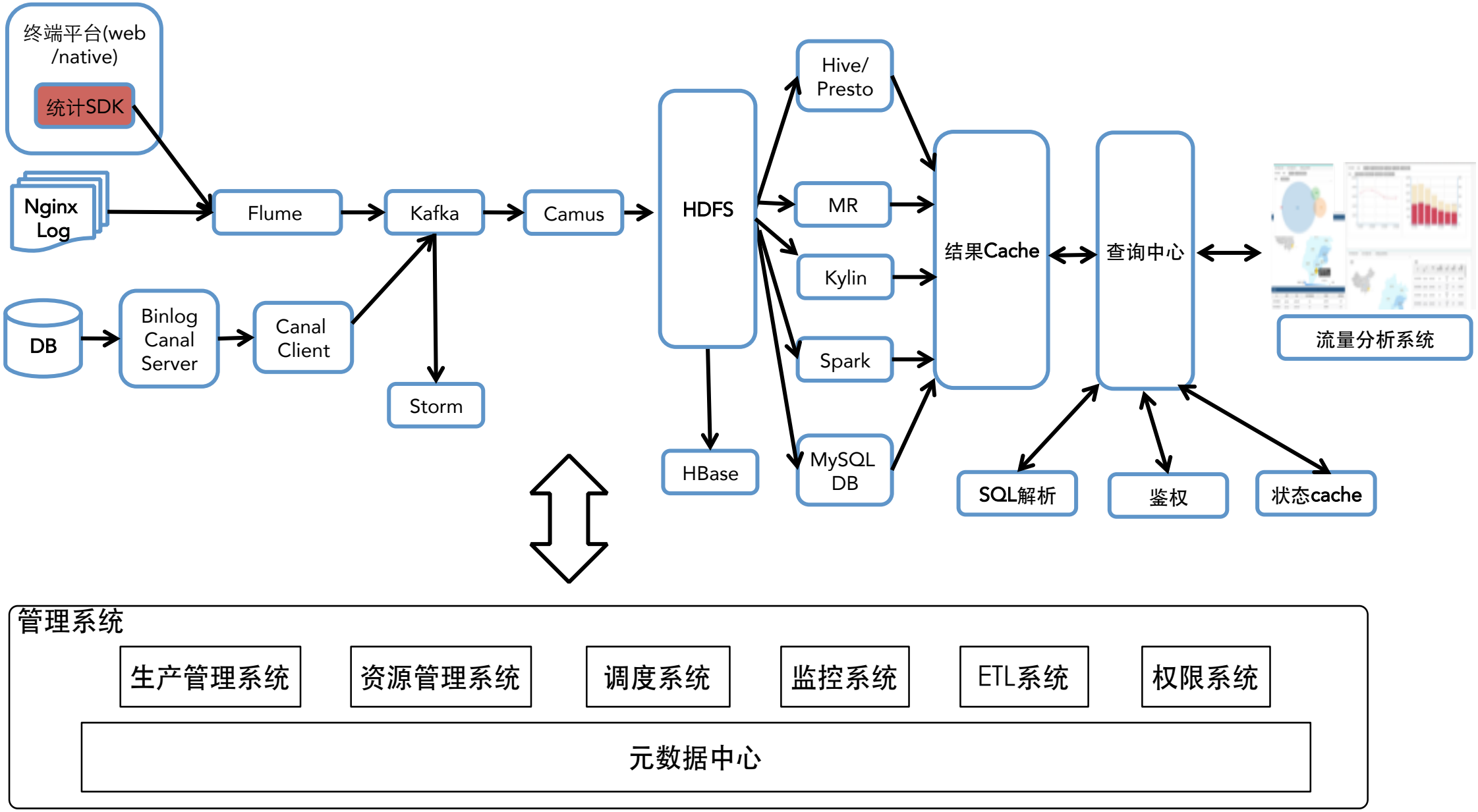
- 本企业的业务数据源外,加入了来自社交网络、传感器等方面的非关系型数据
- 数据规模膨胀，但采集、存储与计算的性能要求不变



- 应用模式愈加丰富：离线—>实时
- 应用范围愈加丰富：经营分析—>业务运营
- 应用人群愈加丰富：分析人员—>一线业务员

举例：流量数据建设

- 数据源
  - 多终端导致数据生成方式多样
  - 大规模数据的采集、存储与计算 110亿+条/日
- 数据应用
  - DAU—>用户行为分析
  - 活动效果转化 离线+实时
  - 个性化推荐，人群画像


终端平台(web
/native)
统计SDK
Nginx
Log
Flume
Kafka
Camus
HDFS
Hive/
Presto
MR
Kylin
Spark
MySQL
DB
HBase
结果Cache
查询中心
流量分析系统
DB
Binlog
Canal
Server
Canal
Client
Storm
SQL解析
鉴权
状态cache
管理系统
生产管理系统
资源管理系统
调度系统
监控系统
ETL系统
权限系统
元数据中心

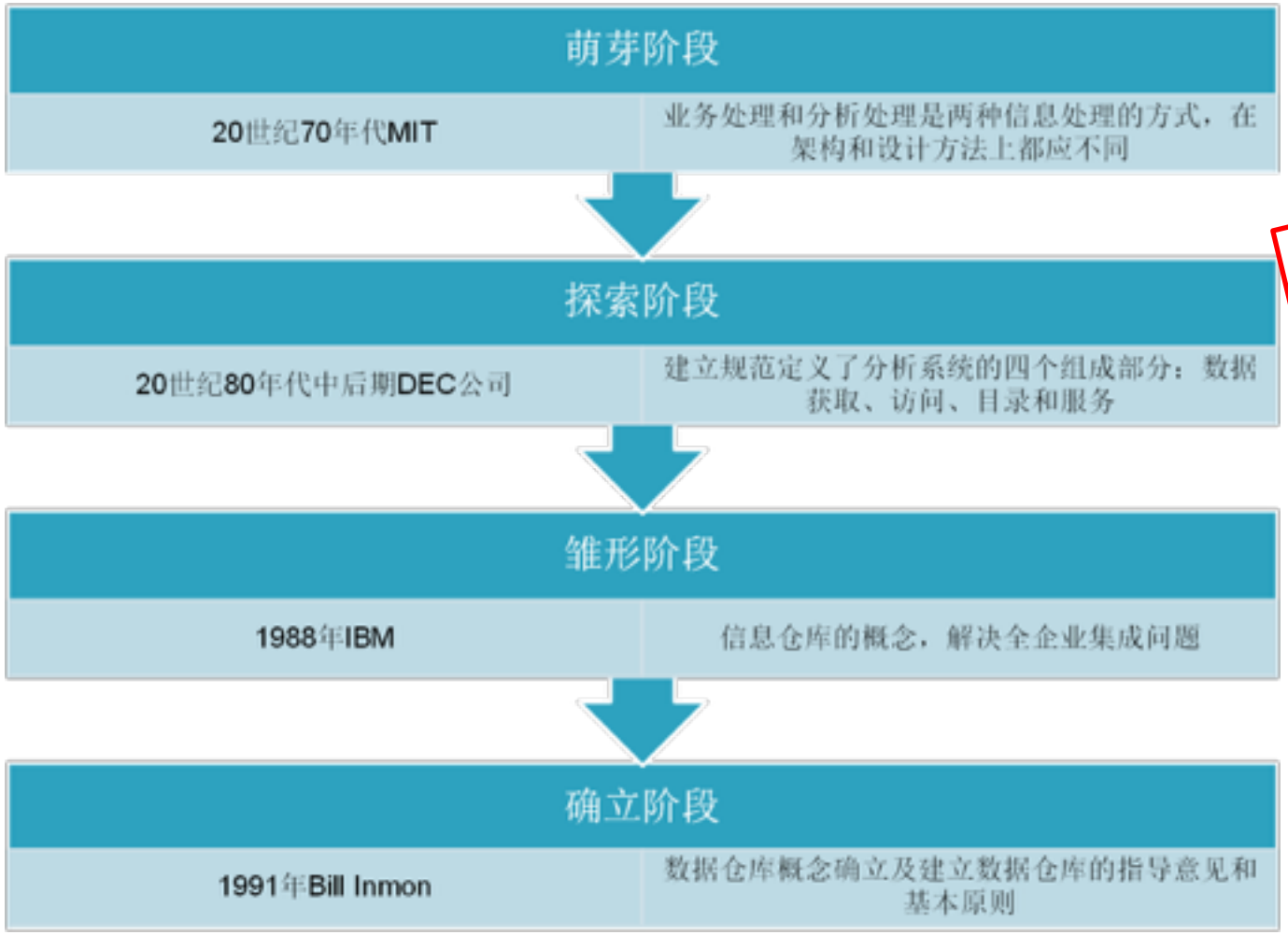
传统数仓 VS. 大数据技术下的数仓

	传统数仓	大数据技术下的数仓
行业	电信、金融、零售等传统行业	互联网等高新行业—>逐步渗透到传统行业
数据结构	结构化，基于关系型数据库(基于DB2/Oracle/Sybase)	结构化+半结构化+非结构化(基于HDFS/NoSQL)
硬件设备	大型机、小型机、工作站	廉价PC组成的分布式处理集群
数据收集	业务系统的DB	业务系统DB；移动终端/传感器/可穿戴设备等数据上报；外网数据抓取；数据开放API
数据规模	GB->TB	TB->PB
用途	经营分析、决策支持	经营分析、决策支持、用户行为分析、竞争情报、用户画像、个性化推荐等

议题

- 大数据的特点、应用场景及其相关技术
- 数仓之本质
- 数仓的规划及成功因素
- 美大数仓体系结构
- 数仓研发工程师的成长之路

数据仓库概念的演化



数仓的演化，是从满足分析需求应运而生的

分析的特点是什么?

1. 从业务问题出发
2. 注重逻辑性(维度+度量)
3. 探索问题答案，指导下一步行动

面向分析的系统与面向操作的系统的差异

	操作型	分析型
数据内容	当前值	存档数据，从操作环境中导出，包含汇总数据
数据结构	适于事务处理	适于复杂查询
连接频率	高	中低
连接类型	读取、更新、删除	读取
用途	可预知的、反复性的	特别查询，偶尔的，启发式的
响应时间	低于秒	几秒甚至数小时
用户量	大量	相关的少数

数仓是为分析而生的新型系统环境

- 为分析任务而设计的数据集
- 从多个业务系统数据源中采集数据
- 便利的数据分析交互形式
- 从总到分的数据组织
- 丰富的历史数据

针对已经存在的不同来源的数据，通过清洗、转化、整合，以商业维度和度量组织数据集，并以便利的方式供下游(分析者或业务程序)提取有用的商业信息，从而及时做出正确的决策的一项工程。

狭义数据仓库 VS. 广义数据仓库

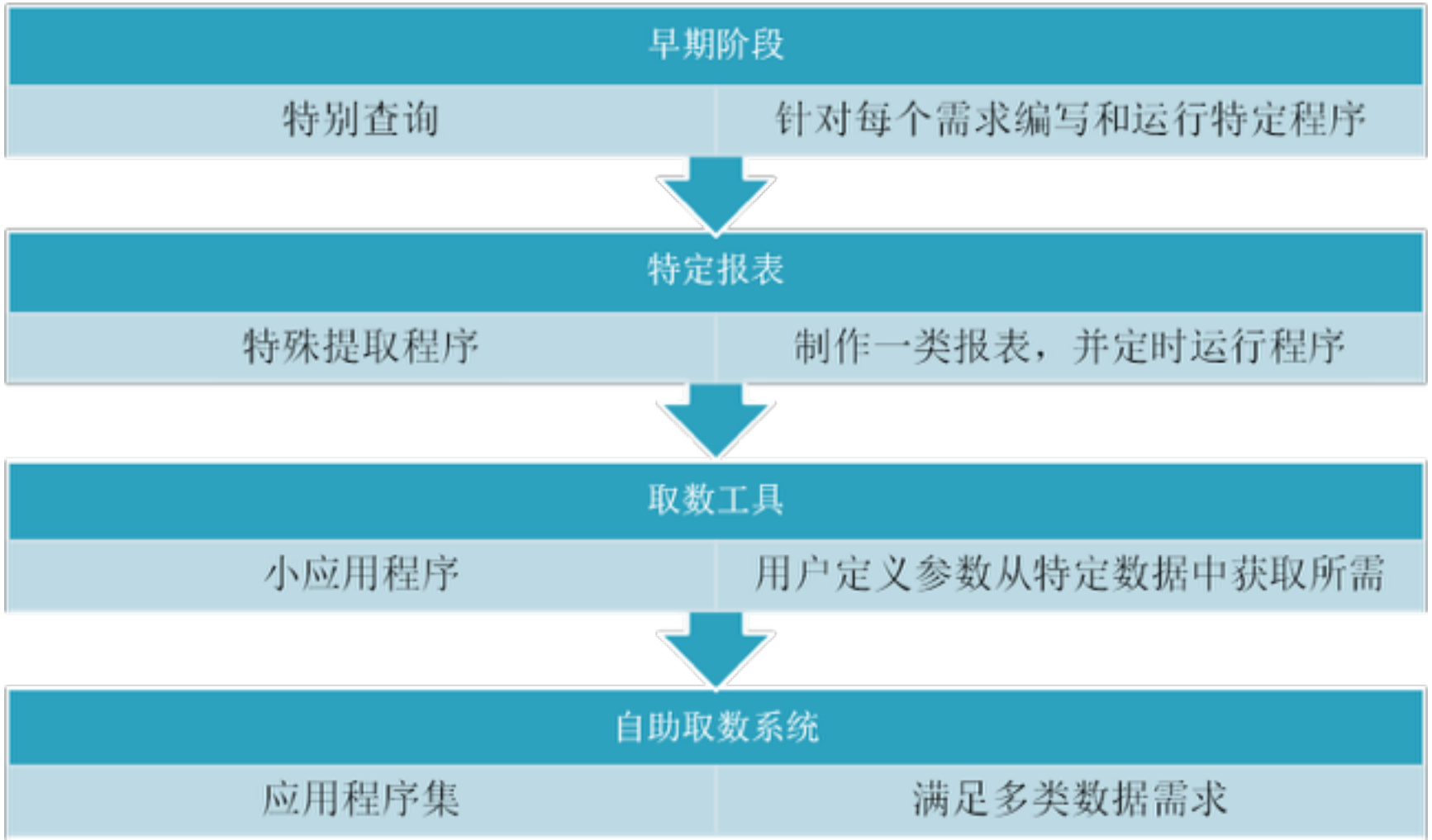
- 狭义：一种数据集合，侧重对数据仓库的结果
- 广义：一项工程，侧重对数据仓库构建的过程和迭代演化

在互联网的业务和技术快速迭代背景下，
更需要以工程的思想看待数据仓库

什么情况下应该谨慎决策建数仓?

- 处于满足取数的阶段
- 业务处于孵化阶段
- 不清楚从数仓中获取什么价值
- 缺乏高层管理者的关注与支持

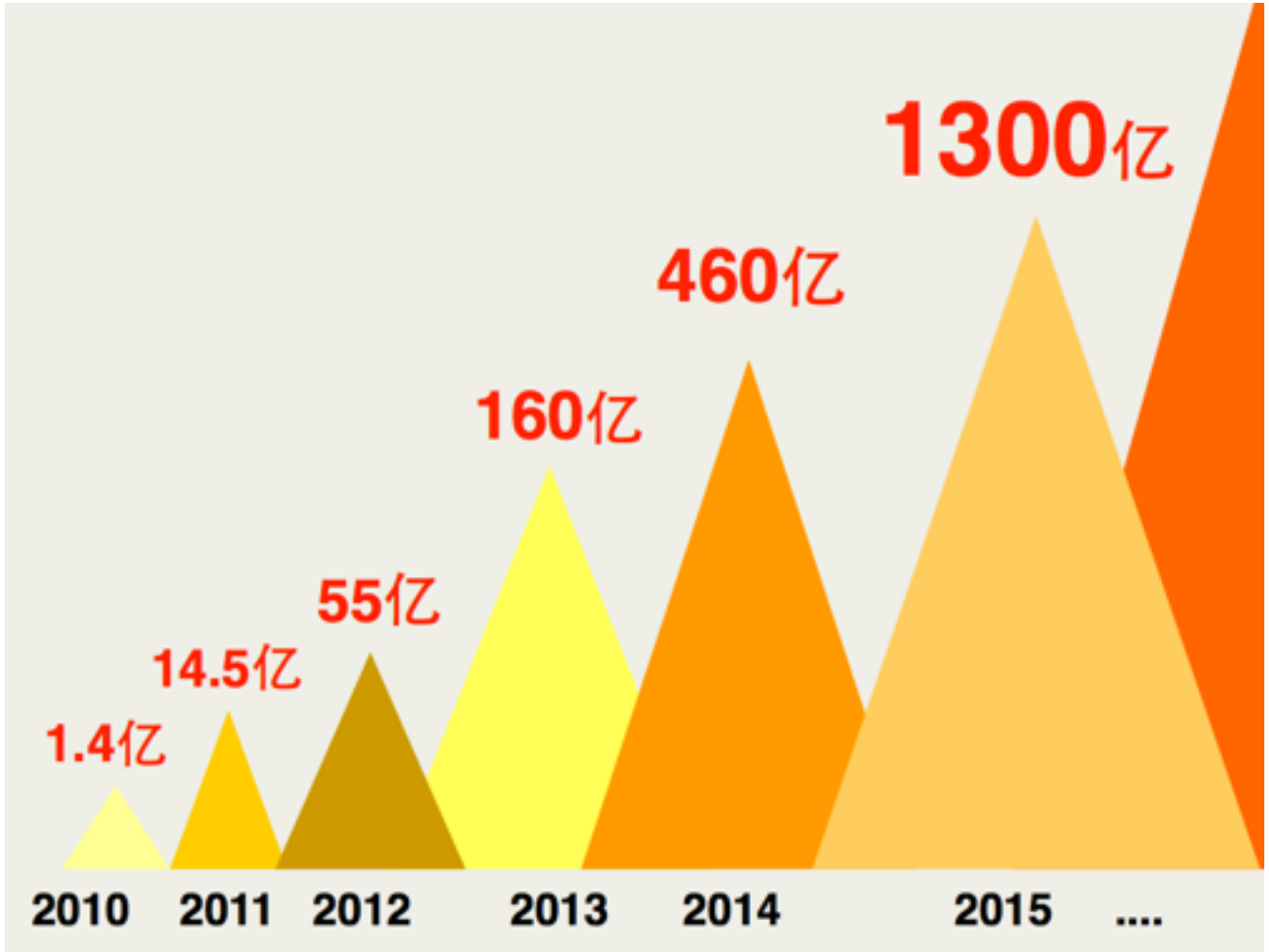
不同阶段的分析需求解决思路



不同阶段的解决思路

随着业务和应用的演化逐步深化

举例：美团侧数仓体系发展历程



举例：美团侧数仓体系发展历程

举例：美团侧数仓体系发展历程

2014年	
MDW2.0体系	面向BI的功能完善、公司级数仓建设
2015年	
各BG数据团队成立、业务BI系统落地	解决BG内业务快速发展的数据诉求
2016年	
流量中央数仓、BG数仓、交易联邦数仓、BI平台+专题分析系统	业务精细化运营对数据更深入的应用

数据的建设诉求，是企业发展所驱动的

议题

- 大数据的特点、应用场景及其相关技术
- 数仓之本质
- 数仓的规划及成功因素
- 美大数仓体系结构
- 数仓研发工程师的成长之路

规划思路——自上而下

中央集权制

优点

1. 从整个企业的角度看数据
2. 体系结构完整
3. 保证数据一致性
4. 中央控制和集中管理，提高管理效率

缺点

1. 前期需要花更多的时间做架构和规划
2. 企业规模越大，失败风险越高
3. 费用很高，且短期收益不明显

规划思路——自下而上

从部门需求出发，建立单一的数据集市

联邦分权制

把各部门的数据集市加以组合，形成EDW

从EDW重新分发数据到各数据集市

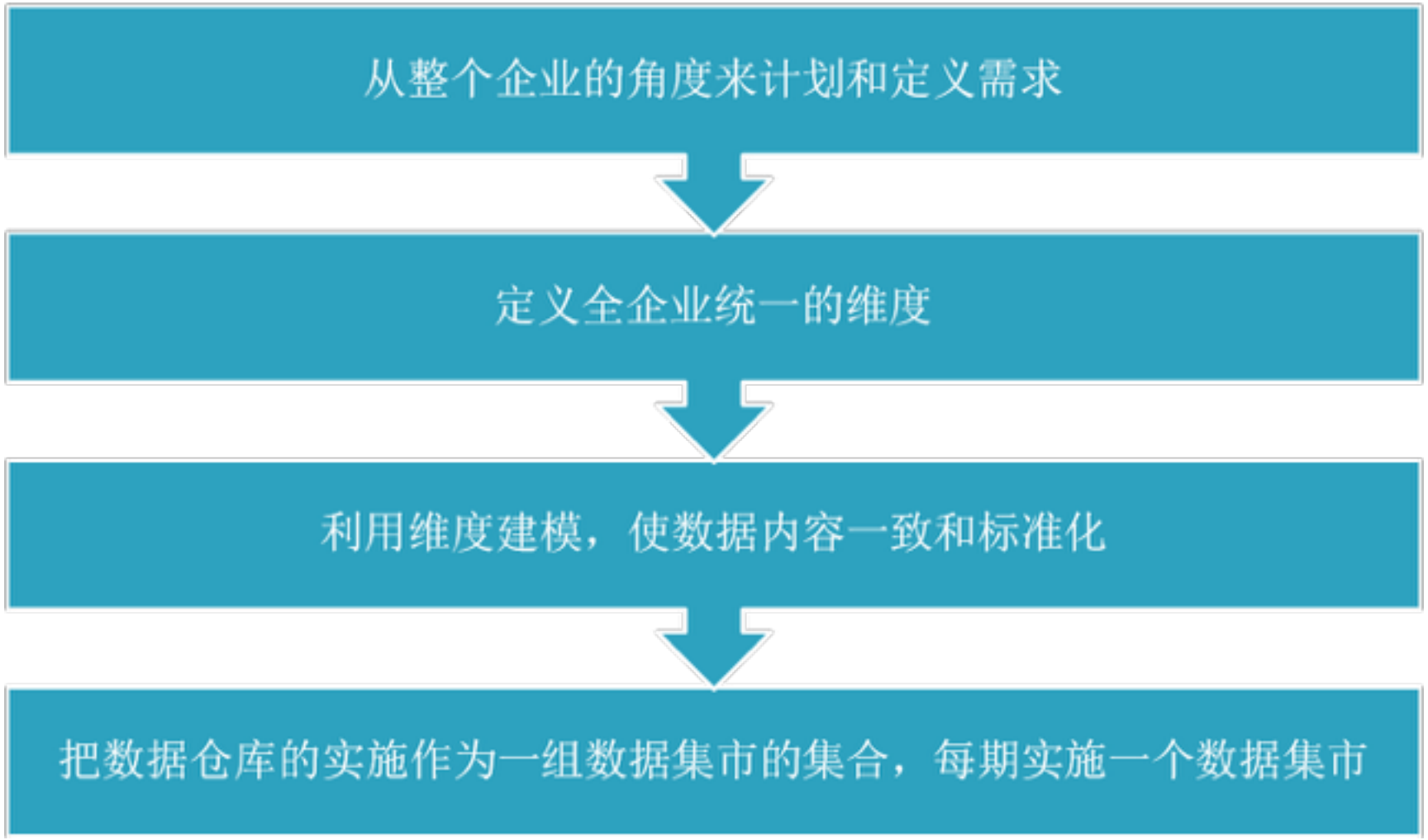
优点

1. 快速实施
2. 短期收益明显，投资回报快速体现
3. 螺旋上升，便于控制风险
4. 在资源有限情况下，从重要数据集市开展，便于重点解决重要问题
5. 项目团队可以在建设过程中学习和成长

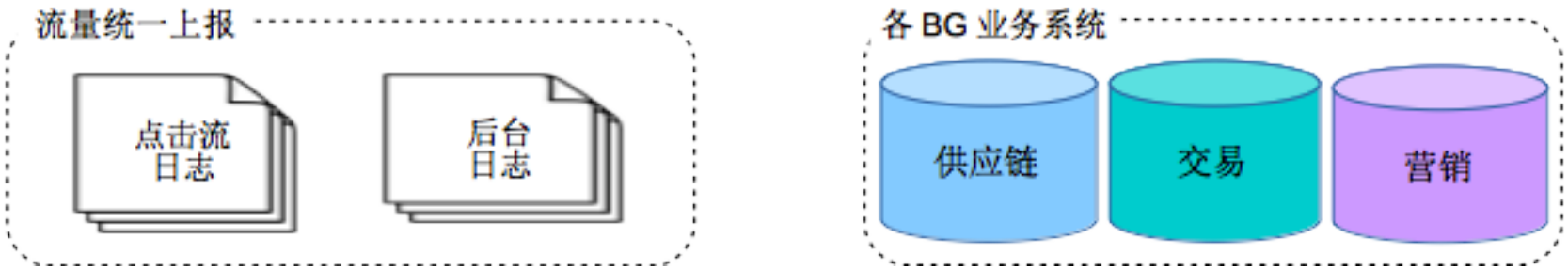
缺点

1. 每个集市对数据的视角都比较窄
2. 整合集市数据是，存在冗余数据，导致EDW的数据过于臃肿
3. 集市间容易出现不一致甚至相互矛盾的数据
4. 伴随集市本身也在持续升级，整合的EDW往往以失败告终

规划思路——妥协的办法


从整个企业的角度来计划和定义需求
定义全企业统一的维度
利用维度建模，使数据内容一致和标准化
把数据仓库的实施作为一组数据集市的集合，每期实施一个数据集市

美大做法?


流量统一上报
点击流
日志
后台
日志
各 BG 业务系统
供应链
交易
营销

成功因素——项目管理层面

- 除非你们公司已经准备好了，否则不要上马数据仓库。
- 选择最好的执行发起人。保证持续、长期和忠诚的支持。
- 强调项目的业务方面，而不是技术方面。选择熟悉业务的项目经理。
- 从整个企业角度看需求。
- 有一个实际的，分段实施计划。
- 和用户交流实际的期望，信守承诺。
- 不要超出成本预算和预期投资汇报率。
- 建立合适有效的交流手段。
- 整个项目生存期，将项目作为 IT 人员和用户之间的结合点。

高层领导的强有力支持是首要成功要素

成功因素——技术层面

- 采用被证明过的技术；避免过分超前的技术
- 知道数据质量的重要性。
- 不要忽视从外部源获得的潜在数据。
- 不要低估花在数据解析、转换和加载（ETL）功能上的时间和精力。
- 选择适合于你环境的架构；数据仓库不是一个万能的提议。
- 架构第一，技术其次，然后才是工具。

议题

- 大数据的特点、应用场景及其相关技术
- 数仓之本质
- 数仓的规划及成功因素
- 美大数仓体系结构
- 数仓研发工程师的成长之路

数仓体系结构

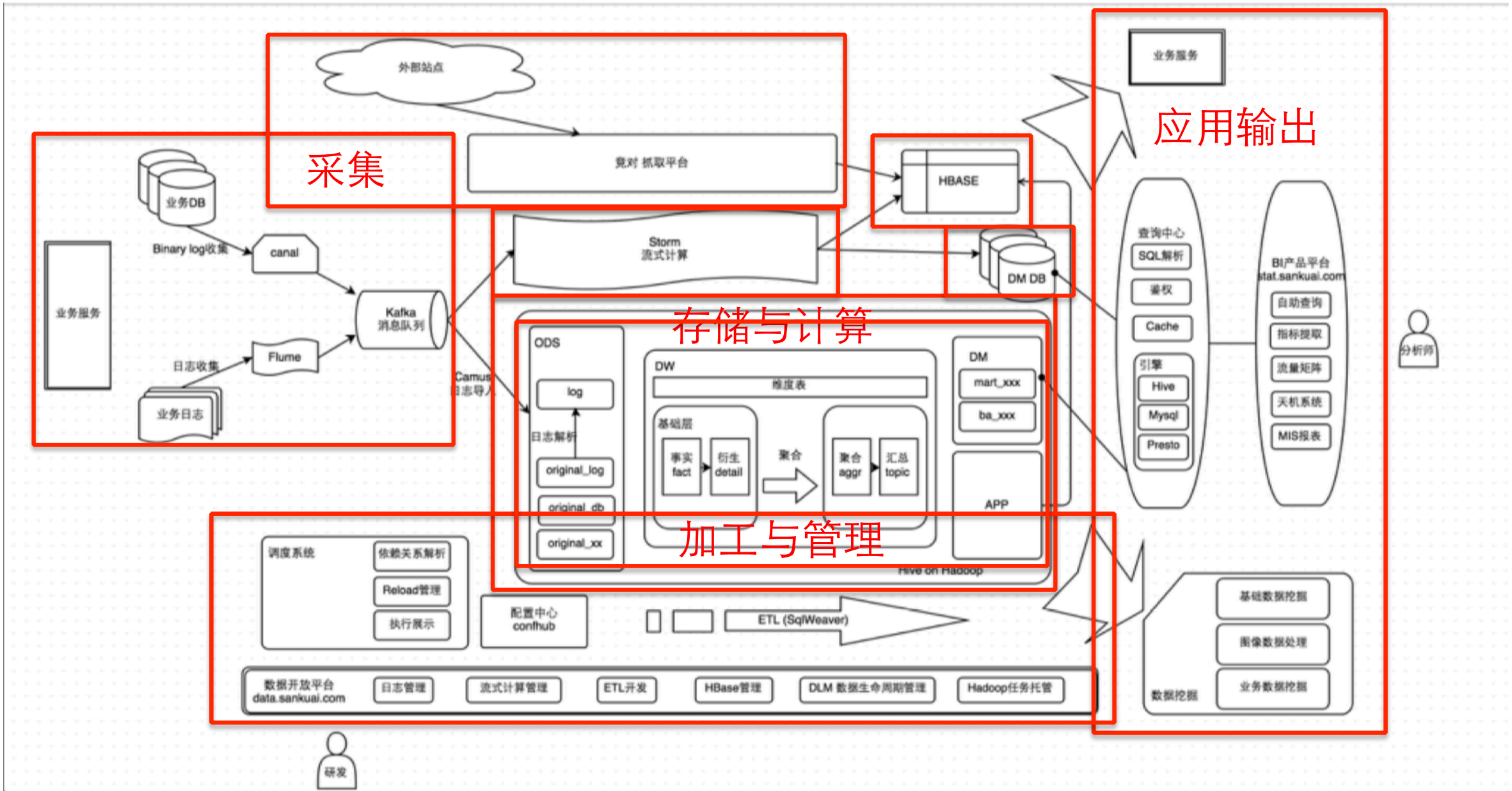
把技术上下文数据转化为业务上下文数据的处理管道

数仓体系结构

- 内容加工与管理
  1. 源数据
  2. 内容加工
  3. 元数据管理
- 能力供给
  1. 基础设施(采集、存储与计算)
  2. 信息传递系统
  3. 管理和控制系统

数仓体系结构——百年老字号版

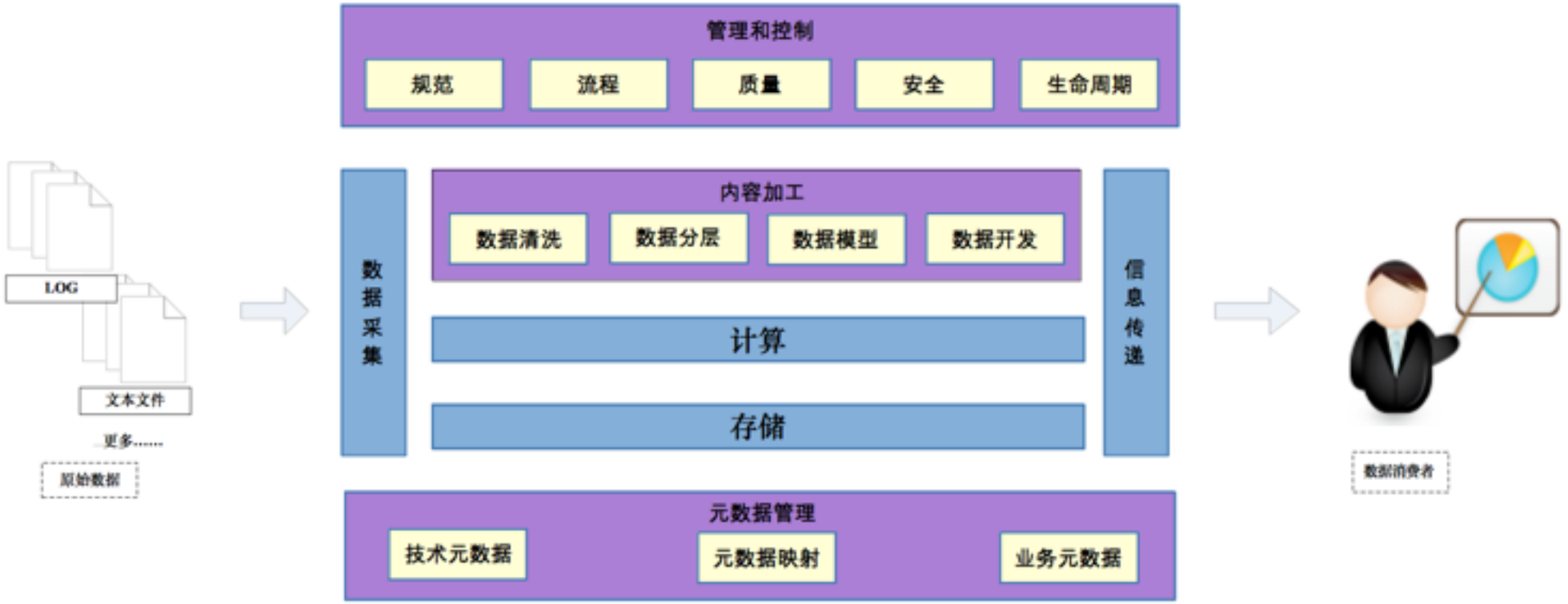
- 内容加工与管理
  1. 源数据 —>原材料
  2. 内容加工 —>材料加工
  3. 元数据管理 —>库存/物件记录、菜谱等的标准化管理
- 能力供给
  1. 基础设施(采集、存储与计算) —>炊具等固定资产
  2. 信息传递系统 —>传菜
  3. 管理和控制系统 —>饭店管理流程


采集
外部站点
竞对 抓取平台
业务DB
Binary log收集
canal
业务服务
Kafka
消息队列
日志收集
Flume
业务日志
Storm
流式计算
HBASE
DM DB
存储与计算
ODS
log
日志解析
original_log
original_db
original_xx
DW
维度表
基础层
事实
fact
衍生
detail
聚合
聚合
aggr
汇总
topic
DM
mart_xxx
ba_xxx
APP
Hive on Hadoop
加工与管理
调度系统
依赖关系解析
Reload管理
执行展示
配置中心
confhub
ETL (SqlWeaver)
数据开放平台
data.sankuai.com
日志管理
流式计算管理
ETL开发
HBase管理
DLM 数据生命周期管理
Hadoop任务托管
研发
应用输出
业务服务
查询中心
SQL解析
鉴权
Cache
引擎
Hive
Mysql
Presto
BI产品平台
stat.sankuai.com
自助查询
指标提取
流量矩阵
天机系统
MIS报表
分析师
基础数据挖掘
图像数据处理
业务数据挖掘
数据挖掘

议题

- 大数据的特点、应用场景及其相关技术
- 数仓之本质
- 数仓的规划及成功因素
- 美大数仓体系结构
- 数仓研发工程师的成长之路

数据仓库研发子通道?



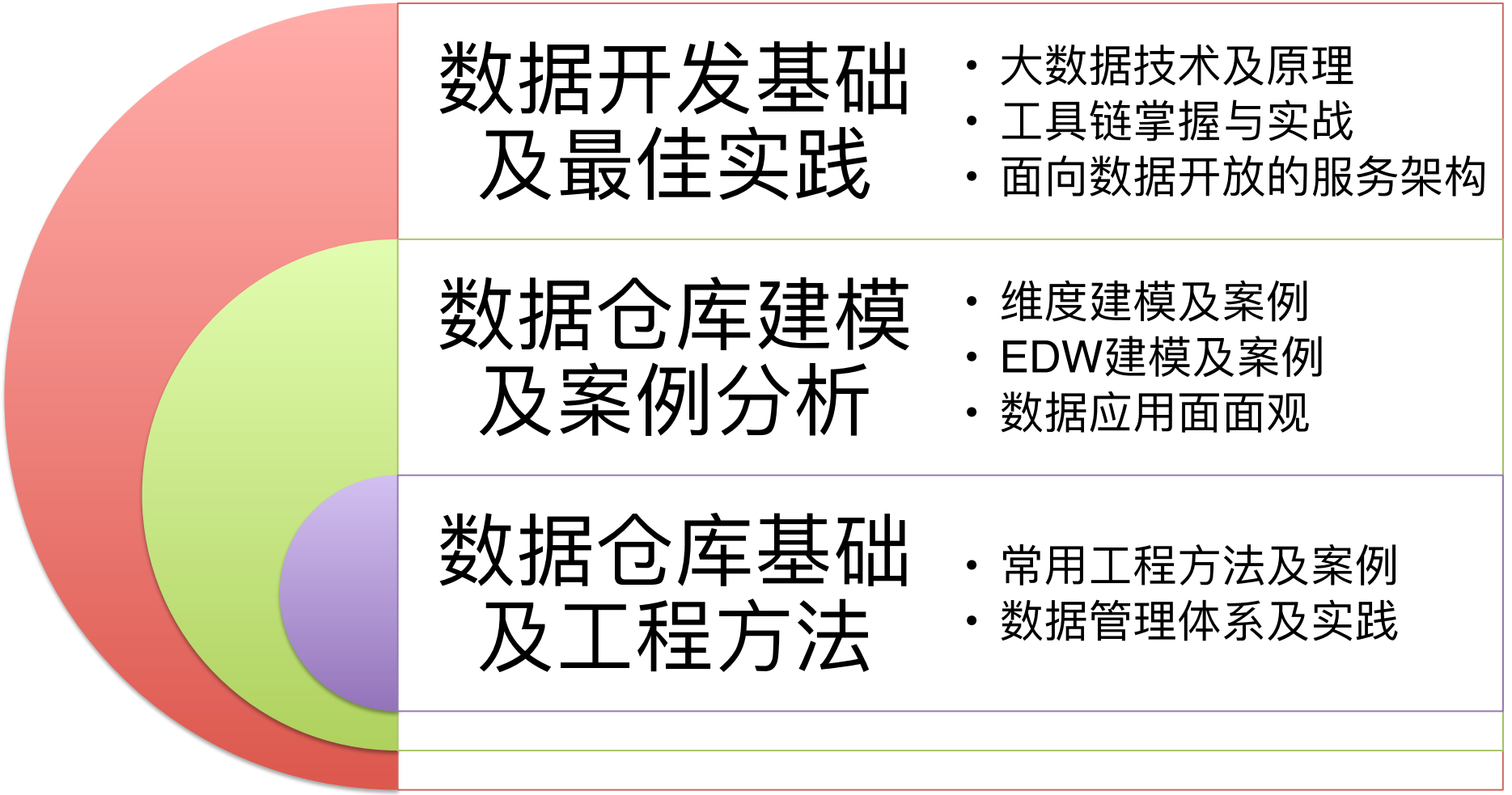
保证内容可用性

保证内容可用性?

未来，随着技术的成熟、云计算和数据思维的普及，“大数据”将是空气一样的存在，以至于会被**熟视无睹**

我们的价值在哪里?

培训体系


数据开发基础及最佳实践
• 大数据技术及原理
• 工具链掌握与实战
• 面向数据开放的服务架构
数据仓库建模及案例分析
• 维度建模及案例
• EDW建模及案例
• 数据应用面面观
数据仓库基础及工程方法
• 常用工程方法及案例
• 数据管理体系及实践

敬请期待(12月).....

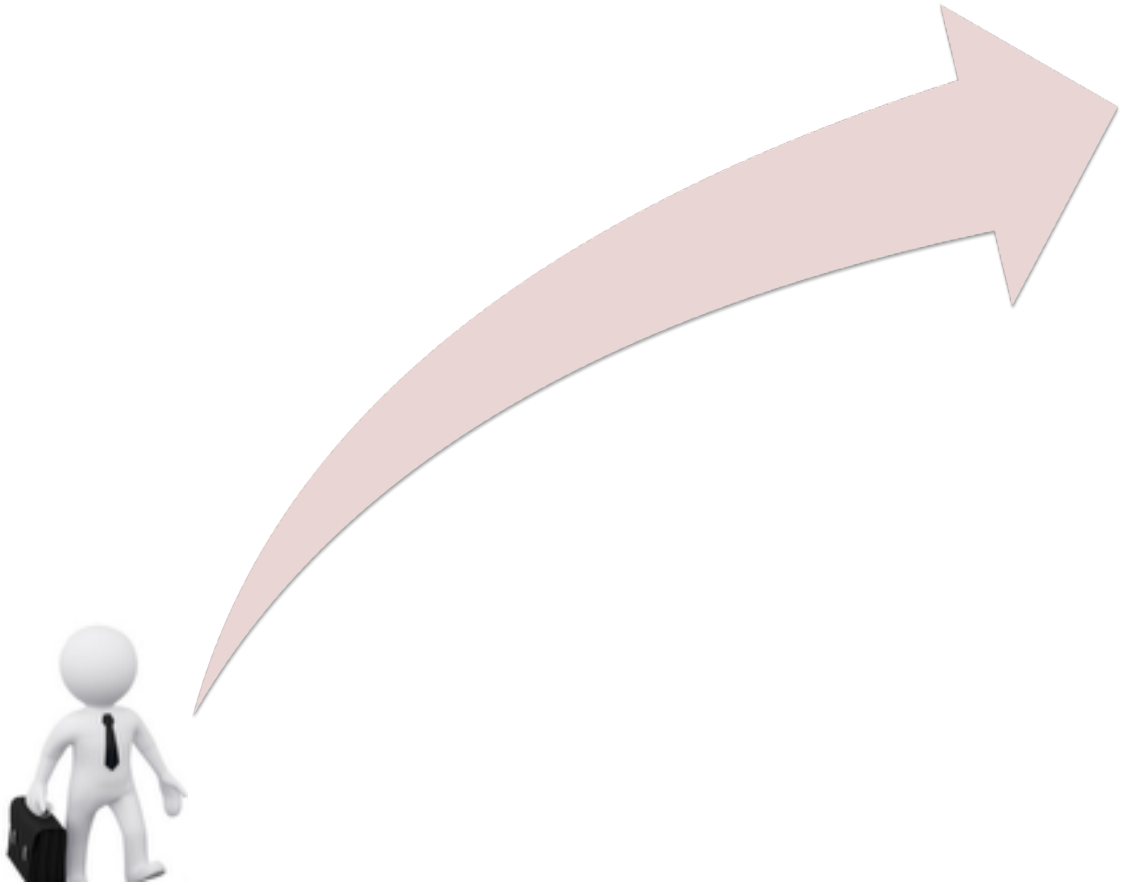
- 《初识数据仓库建模》——from 数据仓库与产品组
- 《大数据生态基本技术及原理》——from 数据平台组
- 《公司数据采集工具及实战》——from 数据平台组
- 《公司数据开发工具及实战》——from 数据平台组

总结

- 大数据与数据仓库
  - 互联网快速迭代的业务和急速膨胀的数据量，需要结合大数据技术解决分析问题
  - 数据仓库要解决的问题没变，只是解决问题的“手段”有了更丰富的选择
  - 大数据的门槛将越来越低，在此趋势下我们需建立自己的核心竞争力
- 什么是数据仓库
  - 综合多种技术的一项工程，而不是一门单一技术
  - 体系结构可分为内容加工管理和能力供给两部分
  - 数仓建设需要随时关注所处的业务和技术环境，在合适的时候选择合适的方案解决分析问题

谢谢大家

成长之路



行政总厨

饭店老板