TiD 2017
质量竞争力大会

大数据平台
数据质量管理实践

大数据平台定义

大数据 big data	互联网大数据：新闻、博客、招聘信息等 企业大数据：企业内部行为产生的数据 政府大数据：交通、医疗、诉讼等公开数据 个人大数据：个人信息、涉及隐私需要保护
数据平台 data platform	包含数据整合、数据处理、数据存储、数据分析、可视化等功能 帮助我们挖掘数据背后的业务逻辑，发现数据背后问题的一体化平台

大数据现状

- 数据质量参差不齐
- 唯一主体识别困难
- 海量数据瞬息万变

数据平台面对的问题

- 数据字段预估不充分
  --> 经常重跑数据，浪费人力物力

- 数据噪音、重复主体
  --> 数据质量难以保证，易被用户投诉

- 网站频繁改版导致模型失效
  --> 传统方法反应不及时，可能导致数据污染

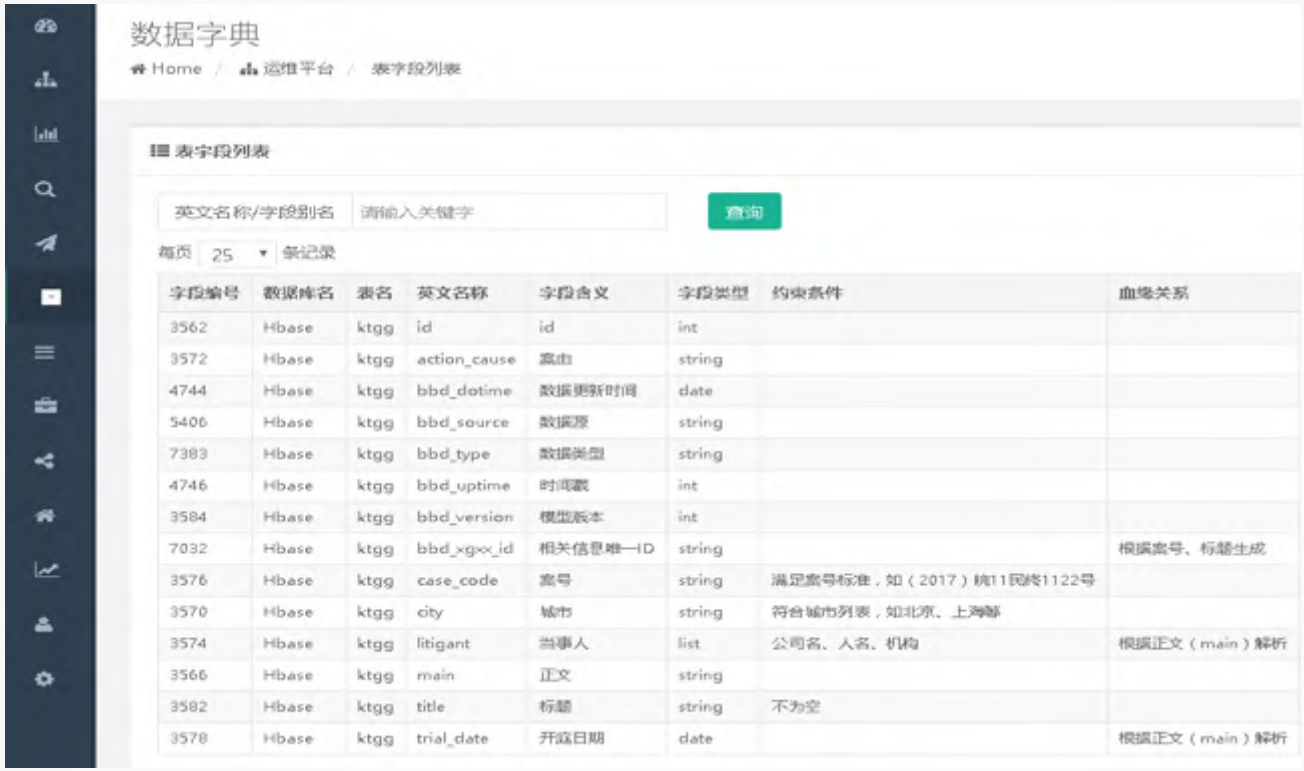
问题举例

假日酒店（中国）有限公司与北京昌信回龙园别墅有限公司、北京龙城丽宫国店债权人撤销权纠纷申请再审民事裁定书

发布时间：2014-11-13

解决之道——数据字段预估不充分

- 数据字典预研，构建数据模型
  - 字段统计分析
  - 定义数据结构、约束条件
  - 创建血缘关系图、分析可行性



数据字典

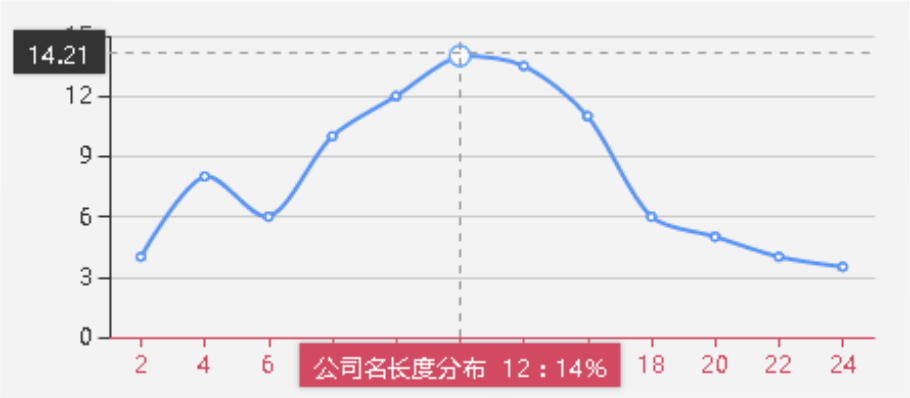
Home / 运维平台 / 表字段列表

表字段列表

英文名称/字段别名 请输入关键字 查询

每页 25 条记录

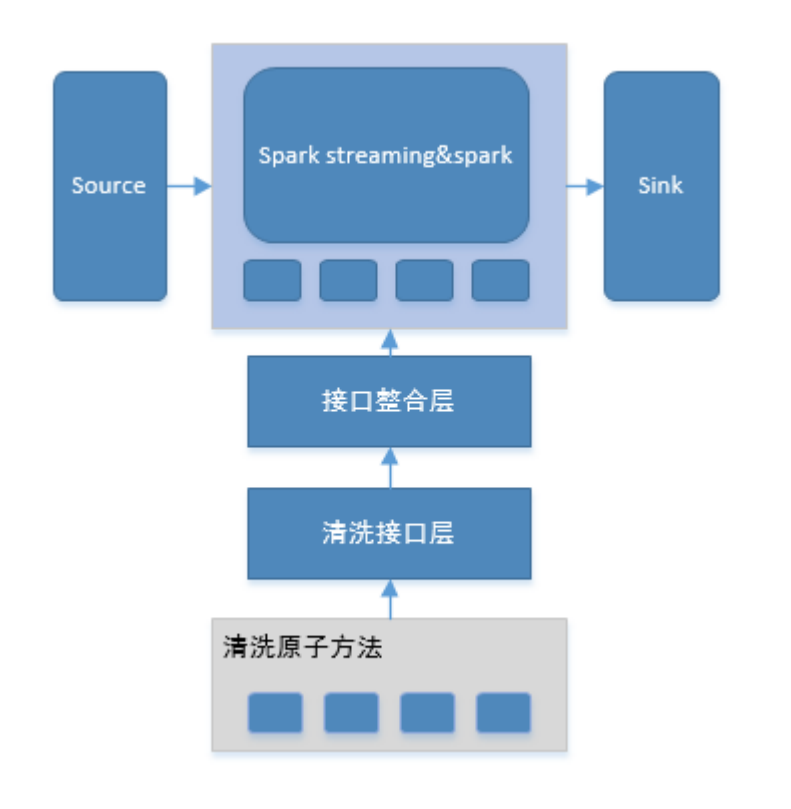
字段编号	数据库名	表名	英文名称	字段含义	字段类型	约束条件	血缘关系
3562	Hbase	ktgg	id	id	int		
3572	Hbase	ktgg	action_cause	案由	string		
4744	Hbase	ktgg	bbd_dotime	数据更新时间	date		
5406	Hbase	ktgg	bbd_source	数据源	string		
7383	Hbase	ktgg	bbd_type	数据类型	string		
4746	Hbase	ktgg	bbd_uptime	时间戳	int		
3584	Hbase	ktgg	bbd_version	模型版本	int		
7032	Hbase	ktgg	bbd_xgxx_id	相关信息唯一ID	string		根据案号、标题生成
3576	Hbase	ktgg	case_code	案号	string	满足案号标准，如（2017）皖11民终1122号	
3570	Hbase	ktgg	city	城市	string	符合城市列表，如北京、上海等	
3574	Hbase	ktgg	litigant	当事人	list	公司名、人名、机构	根据正文（main）解析
3566	Hbase	ktgg	main	正文	string		
3582	Hbase	ktgg	title	标题	string	不为空	
3578	Hbase	ktgg	trial_date	开庭日期	date		根据正文（main）解析



解决之道——数据噪音、重复主体

- 清洗数据、验证数据、挖掘数据间关系
  - 抽象数据清洗原子方法，使之服务化
  - 根据数据模型，通过离线或流式计算，使用清洗方法和数据验证约束条件
  - 提供多维排重算法服务，解决主体识别问题

数据清洗设计、多维排重设计

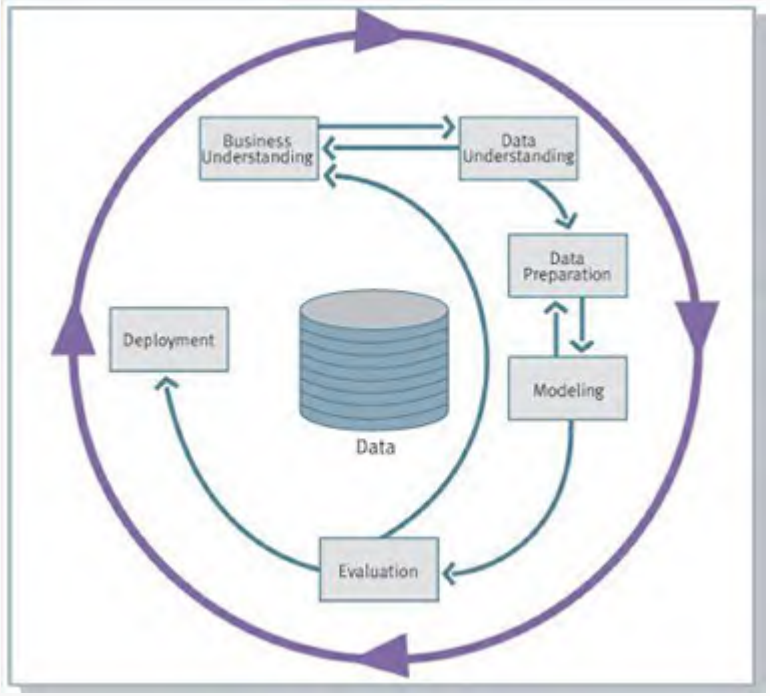


解决之道——网站频繁改版导致模型失效

- 模型版本控制，监控模型失败率
  - 模型失败率实时告警，触发数据处理熔断机制
  - 在线升级模型、主动推送生效
  - 根据模型版本、处理时间戳追溯数据

数据平台之数据建模系统

- 基于CRISP-DM模型
- 解决数据质量问题
- 形成数据处理标准流程
- 构建数据质量反馈体系
- 利用大数据方法提升数据处理效率与性能

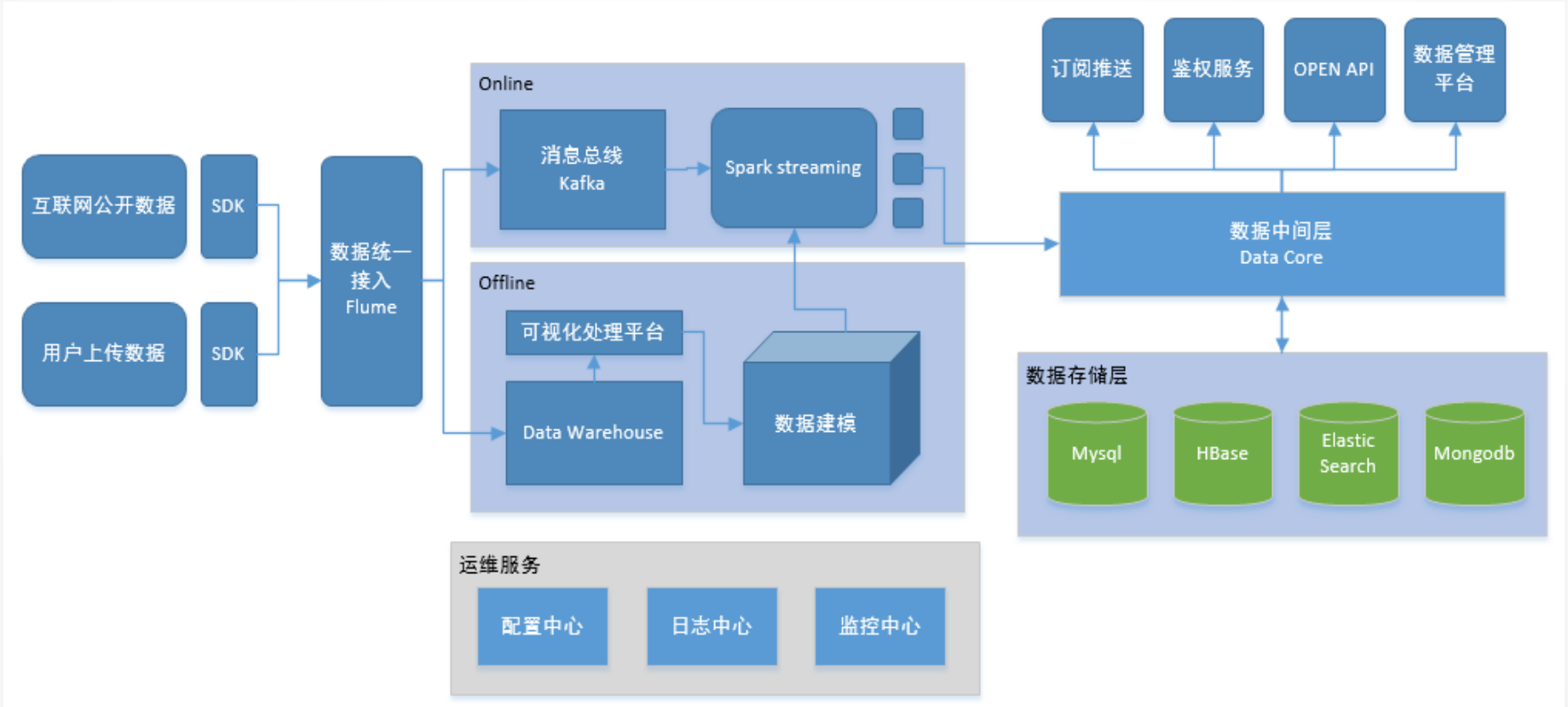


数据建模流程：
商业理解（business understanding）
数据理解（data understanding）
数据准备（data preparation）
建模（modeling）
评估*（evaluation）
部署（deployment）

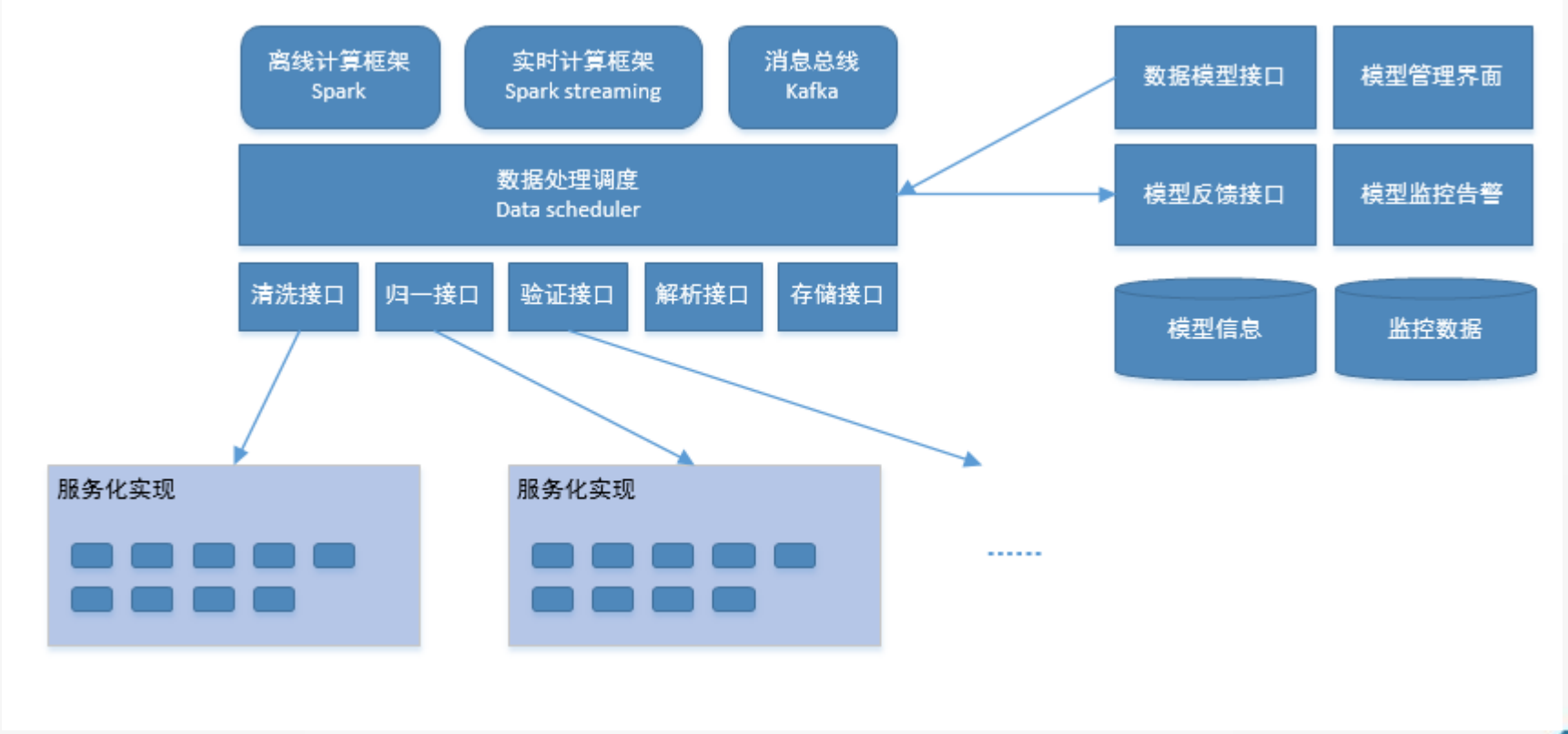
数据质量常见评估维度

序号	数据质量维度	定义
1	数据规范性 (Data Specification)	用于检验数据的定义和取值是否满足相关规范要求，如数据类型、数据精度、字符长度、数据格式、取值范围、编码等方面是否满足公司相关规范要求。
2	数据完整性 (Data Integrity)	用于检验业务所需的数据项是否在系统中有定义。完整性用于度量哪些数据丢失了，信息是否完整等。
3	数据准确性 (Data Accuracy)	用于检验数据值是否真实反映业务情况，或数据是否被准确记录。准确性用于度量哪些数据和信息是不准确的。
4	数据及时性 (Data Timeliness)	用于度量哪些数据和信息是不及时的，是否在规定的期限内获取、录入、更新、加工、删除最新的数据。
5	数据一致性 (Data Consistency)	用于检验不同系统或同一系统内不同表单的相同数据项取值是否一致，关联数据之间的逻辑关系是否正确和完整。

以数据建模为核心的数据平台架构



数据建模与大数据处理


离线计算框架
Spark
实时计算框架
Spark streaming
消息总线
Kafka
数据处理调度
Data scheduler
清洗接口
归一接口
验证接口
解析接口
存储接口
数据模型接口
模型管理界面
模型反馈接口
模型监控告警
模型信息
监控数据
服务化实现
服务化实现
.....

建模技术可行性分析

- 基于Hadoop生态圈
- 基于分布式基础服务
- 基于数据预处理平台
- 为数据管理、数据存储、数据可视化提供服务

数据建模不足与规划

- 数据模型配置繁琐
  - 根据正确数据自适应配置模型
- 验证失败数据需要大量人工审核
  - 智能调整“质”与“量”的平衡点

数据建模之于大数据分析

数据模型是将数据元素以标准化的模式组织起来，用来模拟现实世界的信息框架蓝图。

高质量的数据由下至上，最终变成智慧实现决策。

DT时代我们可以做的更多

- 数据质量与数据模型相辅相成
- 定义数据模型通用开发标准
- 定义通用数据模型“宽表”
- 数据私有，模型开源

参考文献：

[1] https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

[2] 田雪松，张树功. 数据质量分析评估模型的设计与实现. 计算数据 2007

[3] 吴永欢. 数据质量管理模型及应用研究. 电力信息化. 2013

[4] 中国科学院数据应用环境建设与服务项目组. 数据质量评测方法与指标体系. 2009

谢谢！

2017中国企业敏捷实施情况调查

中国敏捷实施现状如何？
企业敏捷实施过程中会遇到哪些困难？
哪些实践是比较普遍的？

如果你已经开始实践敏捷，请扫码参与调查。
同ThoughtWorks一起，为中国敏捷行业打造这份权威报告。

您需要一份权威报告！

深入了解敏捷趋势、最佳实践和经验教训，
帮助你成功实现敏捷转型。