数据仓库方案在企业工商数据治理的应用

王磊

中国经济信息社有限公司，上海 200000

摘 要：随着我国大数据领域相关建设的不断推进，市场上对于企业信息查询类软件产品的需求增加，基于企业征信与商业信息服务的大量依托商业大数据服务类的 IT 软件的不断蓬勃发展。由于竞争的不断加剧，一般的企业类信息在提高时效性和准确性之后，差异化不大，而基于企业信息的建模的特色数据服务，更加受到市场青睐，作者以企业工商数据为视角，企业基本数据为基础，根据数据在业务层面，技术层面遇到的主要问题为牵引，阐述通过数据仓库的设计逻辑以满足实际业务需求，解决企业数据建模与快速应对产品迭代的场景。

关键词：数据治理；商业；金融；数据仓库；企业数据 ; 数据应用；数据监控

一、企业信息类产品概述

2012 年 , “大数据” (Big Data) [1] 一词是个热门词汇。《纽约时报》称 , “大数据” 时代已经降临 , 在商业、经济及其他领域中 , 决策将日益基于数据和分析 , 而非基于经验和直觉。在这样的背景下，利用大数据技术治理企业信息，并推出对应的产品，为经济决策提供依据，就成了必经之路。

1.1 业务功能简介

常见的企业信息查询类产品，是以企业工商数据，财务数据，人员数据，金融市场数据等作为基础，以企业指标的特色算法为支撑，多功能聚合的企业类数据分析平台。

不同的场景应对不同的客户群体需求，既支持模块化的私有化输出，也支持提供整体的 SaaS 解决方案。为这些业务场景提供支撑的 , 是企业自身、产业链、供应链、产业园区、企业推荐与评价体系等不同的应用功能的组合，随着这些应用功能的补充，会不断丰富不同业务场景的应用需求，为客户提供 OLAP 级别的企业大数据服务。

1.2 技术体系简介

此类产品在技术上，采用标准的前后端分离架构，利用 TCP/IP 协议的 RESTful 标准定义 API。前端采用 Js 集成的 VUE 框架，提供快速的前端页面搭建。后端采用 J2EE 标准的开发体系，引入主流的 Springboot2.x 作为分布式微服务的基础，全面支持分部式的 Kubernetes 容器化部署与服务编排。

数据存储端采用 SQL 与 NoSQL 中间件相结合的方式，在国产信创的背景下依托分布式关系型数据库与分布式搜索引擎（如：Elasticsearch），提供海量数据的查询服务。

数据湖常采用主流 Hadoop 生态的 CDP（CDH）作为大数据算力集群方案，通过多种渠道获取外部数据，如 FTP 文件，ETL 工具拉取，面向对象程序部署访问等，完成结构化解析入库 CDP 的 HDFS 存储，Yarn 管理的 Hive 集群里完成数据治理的算法设计与计算，通过外部的 ETL 工具推送到后端微服务访问的分布式数据存储里，向前端应用展示。

1.3 数据治理简介

此类产品面临快速推向市场的压力，初期往往没有设定业务导向的数据标准，对于数据质量难有统一的管理，也忽视了数据的监控，对于数据同步的一致性，业务相互关联的一致性缺乏全局视角的考虑，数据字典对于库级，表级，字段级就更加难以标准化，业务含义分散在不同的库表中，为数据使用与算法的设计带来了较大的困难。

1.4 情况综述

从三个方面分析了此类产品的体系，业务市场面较广，不断的有新的功能和分析算法能进行产品迭代，应用端技术体系也能紧跟时代潮流，并融入主流的技术框架，兼顾一定的前瞻性和稳定性。主要的瓶颈在于数据治理方面。所以，不断提升数据治理的效能，对数据的范围，数据的标准，数据监控做到体系化的构建，不断优化其数据资产管理流程，是需要解决的问题。

二、数据仓库方案概况

利用数据仓库的技术，迎合了大数据时代的普遍特征。大数据主要回答是什么 , 而不是为什么的问题 , 通常有这样的回答就足够了。方案也围绕这企业数据治理，对纷繁复杂的企业数据进行管理，摸清规律，统一利用。

2.1 方案标准

一个好的数据治理方案，能囊括以下特征：（1）有数据，及，数据涉及的业务含义全面，没有明显的缺漏，质量可靠，数据时效性稳定，全量和增量补充及时，

能给与用户和开发团队稳定的预期，不会出现超过规定的延迟；（2）有标准，及，数仓字段的业务含义准确饱满没有歧义、名词标识统一且唯一，数据的事实明细体系，维度体系，标签体系，上下文关系体系健全，入库界定清晰；（3）有层级，及，结构化定义合理，存储类型合适，能贯彻分布式用“空间换时间”的基本精髓，算力分配合理，监控完备，干预点明确，且满足规划范围内的扩展要求，同时，表业务含义明确，抽取逻辑清楚；（4）有分析，及，能通过现有数据支持数据挖掘，支持算法设计。

2.2 方案要解决的主要问题

根据前文对产品各个维度的总结，方案需要解决的数据治理问题如下：

(1) 全面梳理数据资产，对于数据的范围和支持的业务含义，数据供给的稳定性做到心中有数；(2) 产品端要考虑不同客户的差异化需求以及标品迭代的压力，对标品生产环境的数据仓库做对外隔离，让其他的数据开发需求，脱离标品环境，稳定产品的 SaaS 化服务；(3) 重新定义数据标准，从梳理业务功能，到库表的层级分配，维度与标签体系的建立，要进行全面整理，对数据进行分层处理，合理编排数据开发算力；(4) 建立数据治理的工具体系，架设资源可扩展的分布式数据逻辑平台，支持产品对于数据 ETL 的调度和补充当前 CDP 难以支撑的数据开发应用；(5) 建立数据监控运维体系，能通过 OLTP 的方式对数据进行干预，回馈用户关切，并保持下游数据的一致性。

2.3 方案涉及的工作内容

2.3.1 产品涉及的数据资产梳理

从产品的场景和功能端入手，梳理各功能点涉及到的库表清单，以及这些库表的上游源头情况，含数据源头供应商，采集情况，数据分发情况，数据存储情况等。可以先通过表格文档的形式进行资产化管理，请专人维护，后续，如有余力可以做一个资源共享平台，授权相关人员，进行产品功能相关的库表资源跟踪，并进行动态管理。

2.3.2 产品标品数据库隔离

作为产品标品，是对外提供服务的主要载体，从性能，安全等各个角度考虑，所访问的数据库需要与外界做物理隔离。

当前由于产品业务展开的需要，标品数据库常常要提供对外查询和对外数据调度，极大的影响了标品数据的并发性能和数据修正所带来的数据依赖问题。所以，在其他工作展开之前，需要对标品数据进行迁移，对于标品后端微服务所涉及的数据范围，要全量迁移至新的数据库服务集群，为含标品的多个输出端提供数据供给和数据查询服务。

2.3.3 建立数据标准

建立数据标准，实际是按照产品的业务体系，对数据结构进行重新建模的过程。主要分为三个维度，层级标准，表级标准，和字段标准。

（1）层级标准，及，对业务数据进行分层，从贴源层，明细与维度层，聚合层（也称服务层），应用层（展示层），分层对数据进行处理，遵循当前数据技术的条件下，用空间换时间的原则，合理分配不同业务的算力需求，形成结构清晰的数据仓库结构。

（2）表级标准，及，除去数据仓库的贴源层，各个表的数据能表示业务粒度最小化的事实或者维度，同时，对应不同业务含义的数据能存放在不同的表里，通过外键或者其他技术字段形成关联关系，确保不同的表存放的数据在业务含义上明确，唯一，具有业务理解层面的排他性，向范式设计靠拢，为后续扩展留有余地。

（3）字段标准，及，字段的业务含义唯一，排他，具有最小颗粒度，从数仓的明细层开始，字段的名字、长度，类型等在不同库表中，遵循一致，作为枚举值类型的字段，在相同业务含义的字段内进行添加。

（4）建立数据标准的主要目的，是为了通过梳理规范，加快数据开发效率，做到数据层面的所见及所得，降低沟通成本，赋能业务应用。

2.3.4 数据治理工具开发

数据治理的长期性和持续性，决定了，该项工作需要有能支持不同计算需求的数据迁移、调度工具给与支撑。

ETL 调度工具，可利用开源的海豚调度（DolphinScheduler）在进行配置改造后，在众多任务调度工具中，Dolphinscheduler 优势显著，尤其契合国产化趋势。

工具的开发主要分为三个部分，DolphinScheduler 集群，分布式部署，其他算力框架的接入，以及现有批处理任务的调度线上化。

2.3.4.1 DolphinScheduler 的分布式部署

已经部署的 DolphinScheduler（以下简称 DS），由于时间和资源关系，是按照伪分布式 Pseudo–Cluster 的方式部署。作为成熟的数据调度工具，较好的支持集群化的分布式部署，该模式也能更好的支持全面的生产调

度。

目前的发布版本大都支持 Kubernetes（以下简称 K8S）集群上进行部署，可通过 K8S 集群统一管理 DS 的镜像资源，更为便捷的实现服务编排的弹性伸缩以及服务的高可用，也为后续资源监控的扩展提供便利性。

2.3.4.2 不同算力框架的接入

已经部署的 DS，仅支持通过传统 SQL 调度 HIVE 算力平台，并通过配置调用系统 Shell 任务完成对 DataX 任务的控制，流程长，开发难度大，实施速度慢，调试便利性不足。改进的方向是，打通与 CDP 集群的 Kerberos 认证体系，通过 DS 可视化配置 DataX 调度任务，完成项目工作流对任务的直接管理，缩减开发时间与数据逻辑的调试成本。也为后续可能另外部署 Hadoop 类算力集群提供 DS 打通的经验支撑。

在离线批处理数据治理模式的基础上，考虑到部分产品功能以及部分客户对数据时效性的要求，可以引入即席实时计算的 Flink 算力框架，通过 Flink–SQL 后者程序包导入调用，让 DS 实现统一管理。

2.3.5 数据监控平台开发

2.3.5.1 数据资源监控

前序章节提到了 DS 基于 K8S 部署的改造，那么数据平台的资源监控同样可以依托 K8S 进行。其提供了丰富的监控工具（如 Prometheus、Grafana），可以实时监控 DS 的运行状态和性能指标。同时，能集中日志管理，K8S 支持如 ELK Stack、Fluentd 的日志管理组件，可以方便地收集和分析 DS 的日志，对数据处理过程中的异常和流量进行全面监控，便于问题排查和性能优化。

2.3.5.2 数据流程监控

数据流程监控，主要关注数据工作流和数据任务的监控，以及数据增量更新的状态，数据量的准确性等，这些指标，部分可以依托 DS 进行，部分需要开发 WEB 端界面，提供数据监控功能。

2.3.5.3 数据质量监控

数据质量监控，这是整个数据监控平台的核心，是业务的主要关注点，监控的主要内容是数据的缺失与错漏，所以，监控的重点在于发现后的人工干预与数据改进，监控平台要进行基于 DS 的外部 API 调用开发一套适配业务体系的任务调度框架，用以方便数据在进行人工干预后，能及时反馈到数据流的正确流程中，同时，也需要支持不同数据层级的数据修正与数据调整。

三、方案效果监控

3.1 资源投入检核

引入项目管理的思想，从成本、范围、进度铁三角的方式去实施是值得推荐的方式。通过将实施方案，进行可执行的任务包分解，在基本确定工作编排总量的情况下，逐步细化与调整，渐进明细。

总体来说，由于数据治理方案的特殊情况，前期对于标准部分的工作量较大，涉及对于数据源头，产品功能，技术实现路径的综合考量，可能存在资源投入快于实际任务达成的情况，这点要在时间与时间商给与一定幅度的宽限，但是标准一旦评审通过，对于数据仓库的搭建，清洗，逻辑的治理，就不能接受不合理的逾期，同时，后续在数据监控平台搭建的过程中，由于数据逻辑的复杂性，也可能出现资源消耗过快的情况，这点要根据业务要求，做细化处理，单独安排资源。

3.2 方案效果检核

效果检核的主要目的是决定方案是否要继续推进。实际上是对方案目标达成的监控，明确的界定方案整体目标和阶段性目标，对目标本身进行阶梯化标定，可以从更全面的视角审核方案效果。

不同于产品规划的 MVP（最小可行产品）路径，数据类方案，难以短期产生立竿见影的效果，那么对于目标进行阶梯化检核，并判断多目标的组合是否有效就成了重中之重，这往往需要参与检核的专家团队，有足够的数据项目经验，深厚的市场业务理解，扎实的技术框架基础。设定合理的目标划分方式，合理的目标组合方式，是必不可少的。目标可以设定为，以下几个方面：(1) 数据标准制定目标，阶梯划分，业务主体域的制定目标，主体域表结构目标，主体域表关联目标，表字段定义目标；(2) 数据仓库制定目标，阶梯划分，数据仓库层级与关系目标，数据仓库存储目标，数据仓库调度目标；(3) 数据开发目标，阶梯划分，数据逻辑目标，数据开发工作量目标；(4) 数据质量目标，阶梯划分，数据一致性的目标，数据时效性的目标，数据值的偏差目标。

参考文献：

[1][英] 维克托·迈尔 – 舍恩伯格 . 大数据时代 [M]. 杭州 : 浙江人民出版社 , 2013.

[2] 罗巍，刘功总 . 基于大数据的数据仓库研究现状 [J]. 中国新技术新产品，2020(17):38–39.

[3] 贺晓松 . 大数据背景下的数据仓库架构设计及实践研究 [J]. 中国新技术新产品 ,2022,(19):22–25.

[4] 陈氢，张治 . 融合多源异构数据治理的数据湖架构研究 [J]. 情报杂志，2022,41(5):139–145.