大型企业的数据仓库建设

张德奔

（铜陵有色金属集团控股有限公司，安徽 铜陵 244000）

摘 要 自国家提出加快数字经济发展，各企业尤其大型综合型公司都需要意识到信息化、数字化对自身的意义，大型企业长久存在的诸如数据质量问题、数据孤岛问题、数据多口径问题和数据及时性问题，导致对企业数据资产的利用困难，无法为日常生产经营和战略决策提供助力。而数据仓库技术则是解决上述困难的一把利器，本文就从数据仓库建模、数据仓库系统架构、大型企业存在异构数据仓库问题解决及相关应用方面阐述了数据仓库建设过程的思路经验。

关键词 数据仓库；数据库建模；异构融合

中图分类号：F49 **文献标识码**：A **文章编号**：1007-0745(2023)02-0052-03

2020年以来，随着《十四五规划和2035纲要》的出台，国家明确地提出了“加快数字化发展 建设数字中国”，持续地给我国数字化经济提供了支持和指导，在此基础上，至2021年数字经济规模达到45万亿元，占国内生产总值比重39.8%，相比10年前分别提升了309%和84.2%，数字产业化为产业数字化提供道路，同时后者也为前者提供推力，数字化转型已经成为企业的必经之路，尤其对于涵盖多行业、多组织且有一定信息化基础的大型企业，传统的业务信息系统已经无法满足企业对业务数据处理、分析、管理决策支持的需求[1]，具体表现有：

1. 数据量大、种类多，但是数据质量却不高，难以察觉发现无效或噪声数据，使得管理层难以准确地把握企业精确状态，从而某些管理或决策更依赖直觉或个人经验。

2. 各业务类型、各分子公司信息系统形成的信息孤岛，各系统之间的数据、业务流转、交互依靠人工手段处理，增加人力成本，且影响数据的时效性。

3. 信息化管理体系亟待提升，数据管理体系缺乏，各组织、各系统数据之间没有对应的标准化规范，数据格式、含义不兼容；业务流程体系的不通畅，流程节点繁多，分工划分不合理，都影响着企业的效率和成本。

以上问题的本质是大型企业在信息化、数字化发展中没有进行的统一的数据管理，没有规划、组织企业的大数据，针对此问题，提炼企业的数据资产，涵盖各业务模块、各公司组织数据信息，打通信息系统之间的壁垒，并提供切合需求的、高性能数据分析与存储，配合数据挖掘提炼数据价值，辅助管理层的企业决策，达到降本增效，而数据仓库就是实现这一目标的利器，它是一个面向主题的、集成的、稳定的、包含历史数据的数据集合，它用以支持经营管理中的决策制定过程。从它的定义就可以看出它与数据库的区别，数据仓库是以描述、分析事务为主，例如可以反映随时间某事物的变化趋势，而数据库一般是以处理、记录事务为主。所以数据仓库的设计和数据获取有其独特特征。[2]

1 数据仓库建模

设计过程可以分为两个阶段：第一步是调研分析阶段，需要对实际业务开展流程和源业务信息化系统运行方式进行探索研究，获取相关业务流程及源系统处理方式、数据规模等信息，并在此基础上，当前业务实际情况与用户需求转化为逻辑模型。第二步是物理建模设计和实现，包括业务流程的选择，即确认可以产生数据分析目标的业务流程，这是建模设计的基础和先决条件，例如采购业务、生产流程关系着公司生产经营的原料成本，库存的进出涉及公司的存货金额等，在完成业务流程选择后就要进行事实和维度的建模设计，这是数据仓库建模的主要也是核心步骤。

1.1 调研分析

这个过程需要对实际业务和源业务系统的调研探访，获取相关业务流程及源系统处理方式、数据规模等信息，在此基础上，需要业务人员和技术人员深度研讨，规划出不同数据分析主题域。本文以有色金属销售业务为例进行讨论，销售业务的由发起至收货付

款结束共涉及的部分系统及数据内容如下：ERP 系统（订单信息、产品库存信息、发运状态等）、财务共享（收款明细、客户信息、审批状态等）、质量计量系统（有色金属质量、品味、计量方式等）。

深度挖掘业务需求之后，最终规划设计各主题区域分析。销售主题区域包括：客户余额分析、产品销售分析、退货分析、收入成本分析等。

1.2 数据仓库物理建模

业务流程的选择的意义在于确认业务活动中的产生的有价值信息如何保存分析及处理这些信息时所需要遵守的业务规则，事实和维度的建模是确认业务下的事实表数据颗粒度的粗细，例如某类产品硫酸的生成过程、成本原料、销售，也可以是 98 硫酸或 925 硫酸的生产销售数据，这两个就维度粒度粗细不同，越是明细、具体的数据需要的运算量更高和空间更大，这就需要在成本和需求之间找到一个合适的平衡点。

1.2.1 事实表设计模型

在数据仓库中，我们把事务的某一指标数值称为度量值，如收入、成本等；而指标的属性称为维度，如年份、公司部门等，同一类型的维度可能存在不同的粗细大小称为粒度，如时间维度可以分为年份、月份不同粒度，不同粒度的划分由最终分析的需求而定。这样我们把一个带有维度的某个指标的度量值表叫做事实表（fact），把某一属性不同粒度的维度集合为维度表。根据需求不同，数据仓库中的事实表可分为事务事实表（transaction）和快照（snapshot），前者用来描述原子级的事务活动，后者用来描述事务按周期的总结或是持续累计变化，在事务事实表中，描述源系统发生的每一单独事务，以某铜板带销售业务中客户订购单数据为例，包括订购日期、订购产品、编码、数量、库存数量、订单行状态等。

事务表包含多个维度外键的表，多个维度表示此事务发生的时间、地点等信息，同时事务表内部也可能有一些退化的维度列。一般来说这种事实表是不会更新删除的，它会随着 OLTP 系统里业务的增加不断产生新的记录。而在快照表一般可以分为周期型快照和累计型快照，前者用以描述一个特定周期内（一天、一个月、半年或一年）某一类业务数据的其他各维度的汇总，例如企业的每个月的资产负债表、利润表、主营业务收支表（图）等，它们的共同点是有规律周期下的统计意义和需求，每个周期都会产生快照，即使没有任何事务发生也同样会产生，而后者一般用来描述单个独特业务从开始到结束的全过程，这种业务可能没有固定的节点和规律，例如一名员工从入职到离职的工作信息，员工入职便会在事实表里插入信息，包含员工的职位、职级、岗位、组织的各种信息和该状态下的起始日期，在该员工岗位调动后可能职位、组织都会发生变化，则会修改上一条信息的结束日期，并重新生成一条信息的工作信息及起始时间，直至该员工离职在系统中完成出库，这个事实表就是上面所说的累计快照，它在节点流程事件发生时是会有更新操作的，整个表的稀疏都不像周期快照那么规律的稠密，而是取决于业务事件的发生频率。[3]

1.2.2 维度设计建模

数据仓库的事实表 – 维度一般是星形结构和雪花型结构。其中星型结构简单说就是一张事实表里包含多个维度表的外键，从而共同决定了各维度下的事实度量值。

二者的区别在于雪花结构是星型结构的拓展，即某一维度也是一个星形结构，简单来说星形结构会更多产生冗余数据，查询时更少的表连接也会提高性能，而雪花结构的数据空间更少，并且对层次逻辑分析有较高需求的场景更为适合。

在数据仓库中，维度表的更新策略比事实表更为重要，也更加复杂，对不同业务含义背景下会有不同的维度更新策略：

1. 覆盖：即新的维度信息数据直接覆盖旧的，这种策略简单高效，但是无法重现历史业务信息，对需要历史分析的业务是灾难性的，一般对历史维度追溯需求很小或没有时采用。

2. 增加新行：在维度值有修改发生时新增维度行，某维度列变更，新增后生成有效期的时间戳或有效性 flag，形成历史拉链表，这里需要注意可能形成拉链交叉或断链。这种策略功能强大，但可能会影响系统性能。

3. 增加新列：维度某列变化后增加列记录历史列数据，可能会使维表变得很宽、很稀疏，但是某些特定场景需要展现前后维度值对比时性能比第二种方法好。

4. 增加微型维：对方法 2 的改进，把频繁变化的维值合并到一个范围，如公司部门人员统计信息，在分析管理需求满足条件下可以新建微型维，把员工年龄合并到 20–30 岁，30–40 岁类似区间内，统计分析公司员工可以展现为各部门、各职务级别等年龄区间人数而不是统计具体年龄的人数，这样可减少维度发生变化的数量以提升性能，缺点是区间范围有变化需求

时则会有较高的处理成本。

2 数据仓库架构

数仓中数据的获取，这个过程称为ETL（Extract-Transform-Load），作用主要是整合数据来源的维度、清除与分析需求无关的无用数据、按照统一的维度转换事实数据、根据分析和性能的需求对数据进行计算聚合。数据仓库中数据来源一般为企业内部各个业务系统数据或辅以部分企业外部数据，这个流程一般按先后划分为四个数据层：STAGE（初始缓冲层），ODS（数据操作层），DW（数据仓库层），ADS（表示应用层）。具体来说，STAGE层通常用来获取节点业务系统的实际静态数据，一般来说生命周期可以只限于ETL当次处理过程；ODS层用来存放上一层处理后的数据，在这里，数据类型与业务系统高度类似，但是其定义、维度、质量是经过整理转换统一的，数据的含义是明确且唯一的，是整个数据仓库的原子基础数据，一般情况下会长期保存；DW层可以明细分为DWM（数仓中间层）和DWS（数仓服务层），前者是对ODS层的数据进行少量的聚合统计操作，组织成新的结果表和中间表，后者是以面向主题为目的，对ODS和DWM数据进行高度聚合加工，形成关于各业务的独立完整信息和知识，实际应用中，DWS提供用户需要的80%数据，剩余的20%绝大部分可以由DWM提供，这也提高了用户层面的性能感观。ADS层则是为最终面向各个不同类型的用户（数据分析师、业务部门、企业决策者）提供定制化服务，满足对数据内容和格式等的不同需求。

3 大型企业数据仓库的异构融合

大型综合性企业由于涉及的行业多种多样，各公司业务差异可能非常大，因此在数据仓库的设计部署时有非常大的差异，如：

1. 制造行业的公司供应链、财务数据均以物料的事务为原子级别，而房地产、信息服务行业的公司则是以工程项目核算。

2. 例如合并报表等财务系统是以报表实体来划分公司组织，会出现与ERP等业务系统的划分方式不一致的情况，难以进行数据对接统一。

3. 类似产成品在不同公司的生产工艺可能存在很大差异，成本、绩效的分析模型无法兼顾所有公司。

对以上困难问题的深度分析后，也制定出不同的应对策略，对于不同行业数据的融合问题采用抽象化，建立虚拟维度，工程项目型业务也可以进行虚拟化产品采购/销售/入库等流程，对不同公司管理生产方法、不同问题采用参数化模型和ETL过程，根据不同实际业务情景采取不同的定制化预处理，再进行数据的融合，而作为用户或前端应用则不会感受到任何区别。

4 数据仓库应用

4.1 风险预警

数据仓库由于其数据模型结构有着面向主题、保存历史快照等特点，更具有分析预测优势。以有色加金属行业的铜产品为例，一般来说春秋季节为消费市场高峰，这两季的铜销售额对全年销售额有较大影响，在系统建立预测模型，根据旺季和淡季的销售额、成本费用等指标的对比，再综合考虑同比数据等因素可以对当年销售收入预算完成情况做出判断和预警，以供相关业务主管部门参考。

4.2 分析数据获取速度提高

在企业业务数据量达到一定量级后，在分析型数据检索方面数据仓库有非常大的优势，如某一供应商的两年内的单一原料供应价格波动，如果在业务系统检索，需要从采购单据逐个原料采购再关联总账付款信息，而数据仓库可以在采购主题的供应商分析里建立采购金额、采购量等指标，在ETL过程完成大部分的预计算，前端只需进行简单的聚合分类即可得到结果。

4.3 可视化和数据挖掘应用

数据仓库作为大数据分析的基础，其架构的ADS层可以轻松提供数据挖掘和可视化工具的结构，进一步提升企业数据的价值和应用效果，更好地辅助企业管理层完成战略决策。

5 结论

本文介绍了大型企业在信息系统多、数据量大的背景下建立数据仓库的必要性，详细讨论了数据仓库建模设计和架构设计的相关技术，提出大型企业背景下异构数据融合的部分困难和解决方法，最后结合相关实践分析，阐述了建立数据仓库给企业带来的收益。

参考文献：

[1] 程志强.关于大数据时代的数据仓库建设研究[J].长江信息通信,2022,35(07):156-158.

[2] 郑权.融合架构下数据仓库建设研究实践[J].金融科技时代,2021,29(12):52-56.

[3] 陈逸伦,周萃銮,温新叶,等.基于数据仓库建设的营配调数据管理[J].农电管理,2019,286(09):41-42.