DOI:10.19641/j.cnki.42-1290/f.2019.14.011

基于数据仓库的行政事业单位收支管理内部控制评价

——以重庆海事局为例

程 平[1,2](博士生导师)，杨霁莞[1,2]

【摘要】数据仓库作为能够为内部控制评价提供各种类型海量数据支持的战略集合，可为行政事业单位收支管理内部控制的全面、科学、智能和可视化评价提供有效的技术支撑。以重庆海事局为例，依据财政部《行政事业单位内部控制规范(试行)》及收支管理相关规范和制度，设计面向收支管理内部控制评价的数据仓库体系结构，分析数据源和数据的ETL过程，详细阐述从主题域确定、数据粒度划分、事实表和维度表确定到模型设计的收支管理内部数据仓库构建过程，最后运用Hadoop Hive数据仓库的HiveQL语言和OLAP技术、结合案例分析基于数据仓库的收支管理内部控制评价的具体实施。

【关键词】行政事业单位；收支管理；数据仓库；内部控制；评价

【中图分类号】F233，C931　　【文献标识码】A　　【文章编号】1004-0994(2019)14-0092-6

一、引言

2012年11月29日，财政部印发的《行政事业单位内部控制规范(试行)》(以下简称《内控规范》)指出，对经济活动业务层面的收支管理进行风险评估时要重点关注以下内容：收入是否实现归口管理，是否按照规定及时向财会部门提供收入的有关凭据，是否按照规定保管和使用印章和票据等；发生支出事项时是否按照规定审核各类凭据的真实性、合法性，是否存在使用虚假票据套取资金的情形。这为行政事业单位的收支管理内部控制评价指明了方向。

数据仓库是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合，它能够把所有信息系统和外部的数据进行集成，保证数据的一致性，并且能反映出业务系统的变化，满足企业的深度分析需求[1]。其与数据库的区别主要在于数据库是以数据的增删改查为主，而数据仓库是以数据分析为主。在大数据时代，基于数据仓库构建的行政事业单位收支管理内部控制评价体系，能够对收支管理的效率和效果目标、相关信息真实完整目标、资产安全目标、合法合规目标等单个或整体控制目标的实现进行全面、科学、智能和可视化的评价。

从2012年开始，重庆海事局在持续推进财务云平台的建设过程中，对收支管理内部控制进行信息化优化，设计了基于财务云平台的收支管理内部控制审计指标体系[2]，并构建了审计实施流程，为基于数据仓库的收支管理内部控制评价研究奠定了良好的基础。鉴于此，本文以重庆海事局为例，设计了行政事业单位收支管理内部控制数据仓库体系结构，分析了数据源和数据的抽取、转换和加载过程，详细阐述了收支管理内部控制数据仓库的构建过程，最后探讨了HiveQL语言数据查询分析和OLAP多维分析在收支管理内部控制评价中的具体应用。

二、面向收支管理内部控制评价的数据仓库体系结构设计

1. 设计思路。《内控规范》从收入控制、支出控制、票据管理、归档控制、职责权限管理、债务管理等

【基金项目】国家社会科学基金项目(项目编号：17BGL194)；重庆海事局基于财务云平台的内部控制信息化系统建设项目(项目编号：2018Q60)

方面制定了七条具体的行政事业单位收支业务控制条例。依据收支管理相关规范进行内部控制评价，本质上是面向收支管理内部控制评价进行数据仓库体系结构设计的过程，其设计思路是收支业务相关数据首先通过数据抽取、转换和加载（ETL）到数据仓库中进行集中存储和管理，再按照星型模型或雪花模型组织数据建立若干数据集市，构建收支管理内部控制数据仓库，然后利用HiveQL查询语言或者OLAP工具从数据仓库中读取数据进行内部控制有效性分析和评价，进而形成面向收支管理内部控制评价的数据仓库体系结构。

2. *设计过程*。行政事业单位收支管理内部控制评价数据仓库体系结构的设计包括数据源分析与数据ETL、内部控制评价数据仓库构建、内部控制数据分析与评价、内部控制评价结果输出等四个过程。以重庆海事局为例，其收支管理内部控制评价数据仓库体系结构设计过程图参考《财会月刊》2019年第13期文章《基于数据仓库的行政事业单位单位层面内部控制评价——以重庆海事局为例》相关内容[3]。

在收支管理内部控制评价数据仓库体系结构设计过程中，收支管理内部控制评价数据源主要来自海事局收支业务管理和内部控制有关的制度规范、海事局财务云平台的收支业务数据、金蝶K3系统的会计凭证账表数据、各类报销标准、各种票据扫描影像等，为数据仓库的构建提供最基础的原始数据；原始数据经过Sqoop数据库ETL过程后，基于Hadoop架构的收支管理内部控制评价大数据平台，构建形成包括收入管理、支出管理、债务管理3大类共7个主题数据集市的收支管理内部控制评价数据仓库（Hive）；数据分析与评价则利用HiveQL语言对数据仓库的数据进行查询和分析，运用OLAP工具和数据挖掘算法（Mahout）对收支管理内部控制数据进行多维分析与智能评价；内部控制的分析可以通过饼状图、直方图、折线图等统计图形进行可视化展示，实时评价结果可以通过图文并茂的形式进行动态呈现和风险预警，并自动生成多粒度、多维度的收支管理内部控制评价报告。

三、收支管理内部控制评价数据源分析与ETL

收支管理内部控制评价数据仓库的构建，首先需要分析收支管理内部控制评价涉及的数据来源，再进行ETL数据抽取、转换和加载，才能为数据仓库构建做好最基础的数据准备。

1. *数据源分析*。重庆海事局收支管理内部控制评价的数据来源较多，数据量较大，包括结构化、半结构化和非结构化三种数据类型。结构化数据主要包括来自财务云平台中收支管理模型的付款记录表、报销记录表、借款记录表等数据库管理系统中的数据，来自金蝶K3财务软件系统的收付款凭证和账表数据，来自费收管理系统的业务数据，以及来自Excel的收支相关管理数据。非结构化数据主要包括海事局收支管理过程中相关的规章制度、报销标准、会议纪要、收支票据扫描影像、合同影像等数据结构不规则、不完整，不方便用数据库二维逻辑表直接存储的数据，主要以视频、音频、图像、文档、文本等形式存储。半结构化数据不同于关系型数据库或列表中具有特定数据模型结构的数据，但可以用数据标记或其他元素来分隔语义元素以及对记录和字段进行分层，主要包括邮件、HTML网页、资源库等。

2. *数据抽取、转换和加载*（ETL）。收支管理内部控制ETL是将数据通过抽取、清洗转换之后再加载到数据仓库的过程，目的是将分散、零乱、标准不统一的收支管理内部控制相关数据整合到一起，为分析评价提供依据。数据的抽取是从各个不同的数据源抽取到ODS（Operational Data Store，操作型数据存储）中，这个过程也可以做一些数据的清洗和转换，在抽取的过程中为了保证以及提高ETL的运行效率，需要根据数据源数据的特点，选用合理的抽取方法。数据清洗是对数据进行重新审查和校验的过程，过滤掉缺损的数据、错误的数据、冗余的数据等不符合要求的数据，保证数据的一致性。ETL的实现方法有多种，包括借助ETL工具实现、以SQL方式实现、以ETL工具和SQL相结合的方式实现等。

数据ETL工具Sqoop可以实现Hadoop和关系型数据库中数据的相互转换，将关系型数据库（如MySQL、Oracle、Postgres等）中的数据导入Hadoop的HDFS中。对于结构化数据，可以通过Sqoop ETL工具直接从海事局财务云平台和费收管理系统的SQL Server数据库中进行抽取，运用该工具进行转换、加载，解决数据冗余、语义矛盾、格式不统一等问题，最终形成能被Hive数据仓库存储和集成的，结构标准、格式规范的结构化数据，例如对于类型为文本格式的时间字段，需要将其转换为统一的时间类型格式，方便对其从时间维度进行分析。

半结构化和非结构化数据的ETL，针对PDF、Rtf、Word、Excel和PowerPoint等格式的文件主要是通过自然语言分词技术、针对图像格式的文件主

要是通过OCR光学字符识别技术、针对影音格式的文件主要是通过语义识别和自然语言分词技术，提取形成Excel和数据库等类型的结构化数据，然后通过Sqoop等ETL工具进行转换、加载到数据仓库中。例如，与报销有关的各项报销标准主要是通过Word格式存储，需要将其处理形成Excel格式文件或者数据库文件后才能导入数据仓库中。

四、收支管理内部控制评价数据仓库构建

行政事业单位收支管理内部控制评价数据仓库的构建主要包括确定主题域、划分数据粒度、确定事实表和维度表、模型设计等环节。

1. 确定主题域。主题域通常是联系较为紧密的数据主题的集合。根据《内控规范》对收支管理内部控制的具体规定以及收支管理内部控制审计指标设计[2]，再结合海事局收支业务处理的逻辑关系、数据勾稽关系以及关联关系，将收支管理内部控制评价划分为7大主题域，具体见表1。

表1　　主题域确定及其描述

序号	主题域名称	主题域描述	依据
1	收入归口管理执行性	该主题旨在考察是否有非财务部门处理收入业务的会计核算	《内控规范》第二十六条
2	票据保管规范性	该主题旨在考察是否设置票据管理员、是否建立票据台账、票据信息是否完整	《内控规范》第二十七、二十八条
3	票据使用范围合理性	该主题旨在考察票据的使用范围是否符合规定	《内控规范》第二十八条
4	报销审批流程完整性	该主题旨在考察单位是否设立完整的报销审批流程	《内控规范》第二十九条
5	支出范围遵循性	该主题旨在对比支出的范围是否符合对应标准	
6	支出核算处理及时性	该主题旨在对比支出业务结束时间与业务核算时间的差值是否超出规定期限	《内控规范》第三十条
7	债务控制流程完整性	该主题旨在考察债务内部管理制度是否完整	《内控规范》第三十一条

2. 划分数据粒度。数据粒度是数据仓库中存储数据的细化或综合程度。数据仓库中数据表包含的海量数据用何种粒度进行存储，将进一步影响查询分析和数据挖掘的效果，最终影响收支管理内部控制评价的效果。

数据仓库粒度划分主要考虑是采用单一粒度还是多重粒度，以及粒度的划分层级。粒度层级越低，细化程度越高，反之则反。例如，针对行政事业单位支出核算处理及时性的评价需求，时间维度数据粒度需要划分为“年”“月”“日”三个粒度层级。

3. 确定事实表和维度表。收支管理内部控制数据仓库由多个事实表和维度表组成，一个事实表必须与一个或多个维度表建立关联关系。事实表和维度表中的数据必须根据具体的主题域确定。

事实表用来存储事实的度量值和指向各个维度表的外键值，主要存储数字类型数据标记而非详细的描述性信息。根据确定的主题域，收支管理内部控制数据仓库可以建立收入归口管理执行性事实表、票据保管规范性事实表、票据使用范围合理性事实表、报销审批流程完整性事实表、支出范围遵循性事实表、支出核算处理及时性事实表、债务控制流程完整性事实表。

维度表用于描述维度层次及成员类别等元数据信息，用以记录事实表中特性描述以及事实记录信息，以便为内部控制评价提供有用的信息。例如，为了评价支出核算处理的及时性，可以通过建立时间维度表、报销类型维度表和人员维度表来分析和评价支出业务发生时间和核算时间的时间间隔。

4. 模型设计。根据事实表和维度表之间的关系，在设计具体模型时，根据评价需求，应当考虑是选用星型模型还是雪花模型进行数据的组织。

数据仓库用于为内部控制评价提供决策支持手段，而数据集市是数据仓库的子集，用于满足特殊应用需求的数据仓库通常包含较少的数据量和主题域。出于评价需求，根据主题域内容，将收支管理内部控制数据仓库划分为七个数据集市。

分析收支管理内部控制确定的主题域，各个主题均具有多维特性，且在不同的主题域中存在着相互关联与数据共享，事实表共享多个维度表，不同维度表之间也存在关联关系，基于此特性，收支管理内部控制评价数据仓库选用雪花模型进行数据组织，其模型描述及其数据组织见表2。

以支出核算处理及时性内部控制评价主题为例，采用雪花模型进行数据组织，主要包括报销类型、时间、人员维度表，其中，人员维度又关联部门维度和权限维度，具体如图1所示。

根据支出核算处理及时性的评价需求，分析事实表和维度表的数据特征，再进行事实表和维度表具体的物理结构设计，支出核算处理及时性事实表和维度表以及表中字段见表3。

表 2　　行政事业单位收支管理内部控制评价数据仓库模型及其数据组织

序号	主题域名称	模型	事实表	维度表	模型描述
1	收入归口管理执行性	雪花模型	收入归口管理执行性事实表NK_YW_SZ_SS_SRGKGLZXX	人员维度表NK_YW_SZ_WD_YY	该模型重点考察收入核算业务中单据的“审核人”是否属于财务部
				权限维度表NK_YW_SZ_WD_QX	
				部门维度表NK_YW_SZ_WD_BM	
				时间维度表NK_YW_SZ_WD_TIME	
2	票据保管规范性	雪花模型	票据保管规范性事实表NK_YW_SZ_SS_PJBGGFX	收入类型维度表NK_YW_SZ_WD_SRLX	该模型重点考察收入核算业务单据中的“票据编码”“票据银行”“票据类型”“付款账号”等字段是否完整，确保票据信息规范且完整
				票据银行维度表NK_YW_SZ_WD_YH	
				票据类型维度表NK_YW_SZ_WD_PJLX	
				付款账号维度表NK_YW_SZ_WD_FKZH	
				人员维度表NK_YW_SZ_WD_YY	
				部门维度表NK_YW_SZ_WD_BM	
				时间维度表NK_YW_SZ_WD_TIME	
3	票据使用范围合理性	雪花模型	票据使用范围合理性事实表NK_YW_SZ_SS_PJSYFWHLX	用途维度表NK_YW_SZ_WD_USE	该模型重点考察收入核算业务中单据的“用途”是否符合票据使用范围
				收入类型维度表NK_YW_SZ_WD_SRLX	
				人员维度表NK_YW_SZ_WD_YY	
				部门维度表NK_YW_SZ_WD_BM	
				时间维度表NK_YW_SZ_WD_TIME	
4	报销审批流程完整性	雪花模型	报销审批流程完整性事实表NK_YW_SZ_SS_BXSPLCWZX	适用对象维度表NK_YW_SZ_WD_SYDX	该模型重点考察现有的报销审批流程是否对不同的岗位层级、人员、报销金额、报销类型设置不同的报销方案
				报销类型维度表NK_YW_SZ_WD_BXLX	
				人员维度表NK_YW_SZ_WD_YY	
				部门维度表NK_YW_SZ_WD_BM	
				审核状态维度表NK_YW_SZ_WD_SHZT	
				时间维度表NK_YW_SZ_WD_TIME	
5	支出范围遵循性	雪花模型	支出范围遵循性事实表NK_YW_SZ_SS_ZCFWZXX	报销类型维度表NK_YW_SZ_WD_BXLX	该模型重点考察支出核算业务的报销单据中不同报销业务对应的报销金额是否符合对应的报销标准
				报销(金额)标准维度表NK_YW_SZ_WD_BXSTAND	
				用途维度表NK_YW_SZ_WD_USE	
				人员维度表NK_YW_SZ_WD_YY	
				部门维度表NK_YW_SZ_WD_BM	
				时间维度表NK_YW_SZ_WD_TIME	
6	支出核算处理及时性	雪花模型	支出核算处理及时性事实表NK_YW_SZ_SS_BXCLJSX	报销类型维度表NK_YW_SZ_WD_BXLX	该模型旨在对比支出业务的发生时间与支出核算时间的差值，即报销凭证录入中的“业务结束日期”和“报销日期”的时间间隔是否超出规定期限
				时间维度表NK_YW_SZ_WD_TIME	
				人员维度表NK_YW_SZ_WD_YY	
				部门维度表NK_YW_SZ_WD_BM	
				权限维度表NK_YW_SZ_WD_QX	
7	债务控制流程完整性	雪花模型	债务控制流程完整性事实表NK_YW_SZ_SS_ZWKZLCWZX	人员维度表NK_YW_SZ_WD_YY	该模型重点考察借款单中是否有“处长审核”“领导审核”，且“会计审核”“录入员”不能为同一人，保证债务控制流程完整且岗位权限分离
				部门维度表NK_YW_SZ_WD_BM	
				审核状态维度表NK_YW_SZ_WD_SHZT	
				时间维度表NK_YW_SZ_WD_TIME	

五、基于数据仓库的收支管理内部控制评价应用

查询分析、OLAP联机分析处理和数据挖掘是基于数据仓库进行收支管理内部控制评价的重要方法。Hive是基于Hadoop的一个数据仓库，其通过HiveQL语言进行数据查询分析。OLAP联机分析处理允许在被称为多维数据集的多维结构中访问聚合并组织后的数据，使分析人员能快速、一致和交互地分析各个维度的信息，以达到获取分析数据价值的

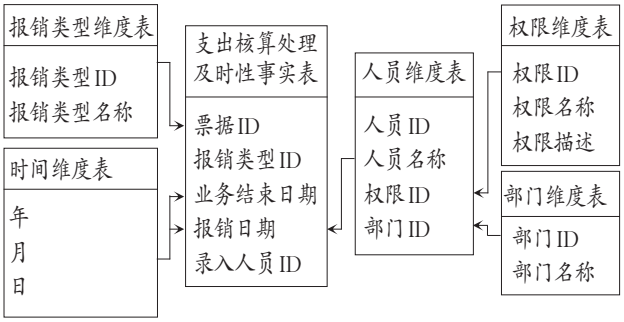

图 1 支出核算处理及时性评价雪花模型

表 3 支出核算处理及时性评价物理表结构

表名	包含字段
NK_YW_SZ_SS_BXJSX【支出核算处理及时性事实表】	VC_No【票据编码】、I_Bxlx【报销类型编码】、DT_Input【业务结束日期】、DT_Bx【报销日期】、I_Inputer【录入人员ID】
NK_YW_SZ_WD_YY【人员维度表】	ID【人员编码】、VC_Name【人员名称】、I_NodeID【权限编码】、I_Depot【部门编码】
NK_YW_SZ_WD_BM【部门维度表】	ID【部门编码】、VC_Depot【部门名称】
NK_YW_SZ_WD_QX【权限维度表】	ID【权限编码】、VC_NodeName【权限名称】、VC_NodeEsc【权限描述】
NK_YW_SZ_WD_BXLX【报销类型维度表】	ID【报销类型编码】、VC_Bxlx【报销类型名称(1车辆,2船舶,3差旅,4培训费,5接待,6会议,7通用)】
NK_YW_SZ_WD_TIME【时间维度表】	Year【年】、Month【月】、Day【日】

目的。数据挖掘通常是指通过决策树、随机森林等算法从大量数据中提取有价值的隐含信息的过程。因篇幅有限，下面主要阐述HiveQL数据查询和OLAP多维分析在支出业务内部控制评价中的具体应用。

1. 基于HiveQL语言的内部控制情况查询分析评价。原始数据经过ETL过程构建数据仓库后，可以根据评价主题确定分析点，然后编写HiveQL语句进行数据查询分析。

（1）报销审批流程完整性。考察是否具有完整的报销审批流程，是否对不同的岗位层级、人员、报销金额、报销业务设置不同的报销方案。运用HiveQL语句，提取出报销审批流程完整性事实表需要分析的字段[“方案名称”“报销（金额）标准”“适用对象”等]进行如下详细分析。

查找报销审批流程是否涵盖所有报销类目，首先提取事实表中涉及的报销类型，与报销类型维度表数据进行比对，找出报销类型不完整的数据列，对应的HiveQL分析语句如下：

```
Hive>
SELECT DISTINCT I_Bxlx FROM NK_YW_SZ_SS_BXSPLCWZX
WHERE I_Bxlx NOT IN
(SELECT ID FROM NK_YW_SZ_WD_BXLX)
```

筛查每一条报销方案任何一个字段出现空值的列，找出流程内容不完整的数据列，对应的HiveQL分析语句如下：

```
Hive>
SELECT * FROM NK_YW_SZ_SS_BXSPLCWZX
WHERE VC_Title =''OR I_BusinessID =''OR VC_Business = ''OR VC_Note =''
OR M_Begin =''OR M_End = ''OR I_Role = ''OR VC_Role =''OR I_NAME =''OR VC_NAME = ''
```

筛查每一条报销方案的适用对象，查看相同业务类型下不同的报销方案是否覆盖所有的岗位层级（适用于任何岗位人员、一般部门人员、部长、局长），找出未能覆盖到的对象，对应的HiveQL语句如下：

```
Hive>
SELECT DISTINCT I_Role FROM NK_YW_SZ_SS_BXSPLCWZX
WHERE I_Role NOT IN (SELECT ID FROM NK_YW_SZ_WD_SYDX)
```

（2）支出范围遵循性。对比支出的报销内容下对应的金额范围是否遵循报销标准。运用HiveQL语句提取出支出范围遵循性事实表需要分析的字段，具体分析如下：

以差旅费报销为例，查找出每条报销记录中“人数”“往返地”“住宿费”“住宿标准”“往返交通费”“往返交通费标准”“伙食费”“伙食费标准”等字段，找出实际费用超出“人数”×对应报销标准的报销金额超标的数据列，对应的HiveQL语句如下：

```
Hive>
SELECT * FROM NK_YW_SZ_SS_ZCFWZXX a JOIN NK_YW_SZ_WD_BXSTAND b on (a.I_Bxlx=b.ID) WHERE b.VC_Bxlx='差旅费'
AND a.M_Food > a.I_People * b.M_FoodStand OR a.M_Hotle > a.I_People * b.M_HotelStand OR a.M_Cityin+a.M_Cityout > a.I_People * b.M_City
```

其他报销类型的内部控制评价数据分析参照差旅费报销，在此不做赘述。

（3）支出核算处理及时性。对比支出业务的发生

时间与支出核算时间的差值，即报销凭证录入中的“业务结束日期”和“报销日期”的时间间隔是否超出规定期限。运用HiveQL语句，提取出支出核算处理及时性事实表中的“业务结束日期”和“报销日期”，计算每一条记录的时间间隔是否超时（如按照有关规定支出核算业务处理天数不得超过半个月），对应的HiveQL语句如下：

```
Hive>
SELECT * FROM NK_YW_SZ_SS_BXCLJSX
WHERE DATEDDIFF(d,DT_Input,DT_Bx) >15
```

2. 基于OLAP的内部控制多维分析评价。在数据仓库中存储的数据不仅能像关系数据库中存储二维关系数据，通常还需要反映数据多维的信息甚至查看某一维度下一级维度的信息，例如收支管理内部控制评价数据仓库中支出核算处理及时性主题对应的多维数据集（报销类型维—时间维—人员维）可以表示为：（报销类型维度，时间维度，人员维度，3），根据多维数据子集构建的多维立方体是逻辑上的数据组织形式，是进行OLAP的分析主体，支出核算处理及时性OLAP评价模型如图2所示。

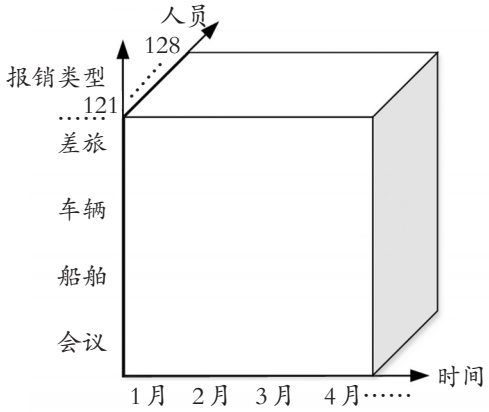

图 2 支出核算处理及时性OLAP评价模型

通过扩展Hive的OLAP驱动建立OLAP多维数据分析模型后，可以通过对多维模型的切片、切块、上卷、下钻等操作进行收支管理内部控制数据分析和评价。

（1）切片。数据切片操作用于选取和定义收支管理内部控制评价主题多维数组的一个二维子集，例如，可以从报销类型、时间、人员三个维度组织建立的三维数据结构中选取报销类型维度和时间维度进行二维切片。

（2）切块。数据切块是将多个切片进行叠加，进而形成特定维度区间的操作。例如，可以从报销类型、时间和人员组成的三维数据结构中选取报销类型、时间和人员维度进行三维切块。

切片和切块后的模型如图3所示。

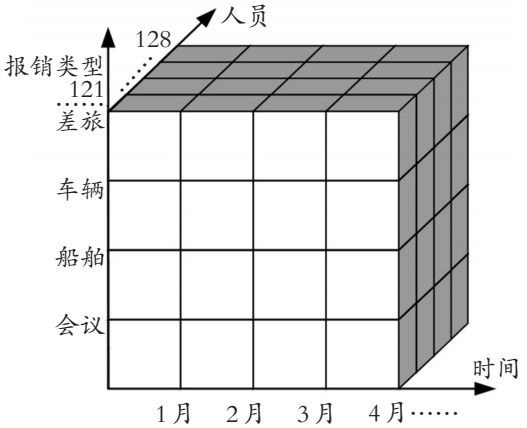

图 3 三维数据子集切片、切块过程

时间维度：如按日、月、年对数据进行分组汇总，不仅可以按不同长度时间点查看数据，还可以统计不同时间段的数据，从而更好地从短期到长期对报销数据进行统计分析，以及评价其中可能存在的风险。

报销类型维度：按不同报销类型对报销数据进行分组汇总，按类别统计，可以对平时风险较多的报销类型进行重点分析。

人员维度：根据人员ID对报销数据进行分组汇总统计，重点关注存在较多风险的报销核算处理数据的录入人员。

（3）钻取。钻取包括上卷和下钻两种操作。上卷是从明细级数据向上汇总到高级数据视图；下钻是从高级数据向下钻取到明细级数据视图。对数据钻取的能力取决于数据结构以及粒度层级。

主要参考文献：

[1] 王珊，王会举，覃雄派等. 架构大数据：挑战、现状与展望[J]. 计算机学报，2011(10)：1741～1752.

[2] 程平，尹赤. 数据视角下基于财务云平台的收支管理内部控制审计——以重庆海事局为例[J]. 财会月刊，2018(2)：129～134.

[3] 程平，范洵. 基于数据仓库的行政事业单位单位层面内部控制评价——以重庆海事局为例[J]. 财会月刊，2019(13)：71～76.

作者单位：1.重庆理工大学会计学院，重庆400054；2.重庆海事局内部控制信息化课题组，重庆400054