基于ETL技术的商业银行自动对账系统

广电运通金融电子股份有限公司　刘映辉　晏妮

目前，商业银行的各种业务在不断发展，传统业务的业务量不断增大，各种新业务也不断涌现。随着业务量的激增，人工对账调账的工作量也随之加大，对账工作产生差异的可能性也加大，因而给对账工作提出了更高的要求，银行业务人员面临更大的挑战。由银行业务人员根据报表进行手工对账，不但严重浪费人力资源，而且容易产生差错和业务风险。

为适应未来银行业务的发展，降低商业银行对账工作的难度，减少手工对账的工作量，增加对账的准确性，本文介绍一种代替人工对账方式的基于ETL技术的商业银行自动对账系统。

一、自动对账系统中的ETL过程

发现差异并进行调账的过程。这个过程可以用相对应的ETL过程来实现。

ETL的实现有多种方法，其中常用的有三种：第一种是借助ETL工具实现；第二种是通过SQL方式实现；第三种是ETL工具和SQL相结合来实现。本系统采用第三种，它综合了前面两种的优点，会极大地提高ETL的开发速度和效率。

1.数据的抽取

在数据的抽取阶段，需要确定所有内部数据源。对所有计算机平台和数据抽取的所有源文件进行详细说明。如果还有外部数据源，要决定内部数据结构和外部数据结构的兼容性，而且要指出数据抽取的方法。商业银行对账的数据源一般都是固定的数据报表，属于文件

而且容易产生差错和业务风险。

为适应未来银行业务的发展，降低商业银行对账工作的难度，减少手工对账的工作量，增加对账的准确性，本文介绍一种代替人工对账方式的基于ETL技术的商业银行自动对账系统。

一、自动对账系统中的ETL过程

ETL（Extraction，Transformation，Loading）即数据抽取、转换和加载。它是从各种原始的业务系统(异构多源)中提取数据，按照预先设计好的规则将抽取到的数据进行转换，最后将转换的数据按计划增量全部导入目标数据库或数据仓库中。从工程应用的角度来考虑，ETL过程按照物理数据模型的要求加载数据并对数据进行一些系列处理，其处理过程与业务及经验直接相关，同时这部分的工作直接关系到数据仓库中数据的质量，从而影响到联机分析处理和数据挖掘的结果的质量。

商业银行的对账工作，是将各种数据报表汇总，按照一定的业务规则和计算公式对这些数据进行处理，得到最终结果后再将此结果与目标结果进行比对，从而发现问题。

1.数据的抽取

在数据的抽取阶段，需要确定所有内部数据源。对所有计算机平台和数据抽取的所有源文件进行详细说明。如果还有外部数据源，要决定内部数据结构和外部数据结构的兼容性，而且要指出数据抽取的方法。商业银行对账的数据源一般都是固定的数据报表，属于文件类型数据源(.txt或.xls格式)，所以在此阶段自动对账系统所要做的就是抽取这些数据报表中有用的信息以便进行下一阶段的数据转换工作。数据抽取的过程就是读取相关数据报表文件并选择其中相应字段的过程。

2.数据的清洗和转换

此过程是将源数据变为目标数据的关键环节。由于源数据都有一定的缺点，所以，在源数据进入数据仓库或数据库之前，必须被加工处理。处理过程分为数据清洗和数据转换。

(1)数据清洗

数据清洗的任务是按照某种特定的条件过滤掉那些不符合要求的数据，将过滤的结果提交给下一步进行处

理。不符合要求的数据是相对下一步处理流程的需要而言的，其判断条件与业务规则密切相关。数据清洗需要注意的是不要将有用的数据过滤掉。每个过滤规则都必须经过用户确认，并认真进行验证。

(2)数据转换

商业银行自动对账系统的核心在于数据转换这一过程，而抽取和装载一般可以作为转换的输入和输出。数据转换的任务主要包括数据格式转换、数据类型转换、数据汇总计算、数据拼接等。

商业银行自动对账系统中进行数据转换的规则是依赖具体目标数据的，目标数据有多少字段，就有多少条规则，总结其中的基本类型如下。

①直接映射。就是对这样的规则，如果数据源字段和目标字段长度或精度不符，需要特别注意看是否真的可以直接映射还是需要做一些简单运算。

②字段运算。对数据源的一个或多个字段进行数学运算而得到目标字段。这种规则一般针对数值型字段而言。

③参照转换。在转换中通常要用数据源的一个或多个字段作为Key，到一个关联数组中去搜索特定值，而且交易是不需要进行对账处理的。但是无论怎样，对于可能有NULL值的字段，不要采用直接映射的规则类型，必须对其进行判断。

⑥日期转换。在数据库中日期值一般都是特定的，需要设计一些函数来处理，将日期转换为8位日期值、6位月份值等。

⑦日期运算。基于日期，我们通常会计算日差、月差、时长等。一般数据库提供的日期运算函数都是基于日期型的，而在商业银行自动对账系统中采用特定类型来表示日期的话，必须有一套自己的日期运算函数集。

⑧聚集运算。对于临时对账表中的度量字段，通常是通过数据源中一个或多个字段运用聚集函数得来的，这些聚集函数为SQL标准中已经包括的，例如sum,count,avg,min,max等。

⑨既定取值。这种规则和以上各种类型规则的差别就在于它不依赖于数据源字段，对目标字段取一个固定的或依赖系统的值。

3.数据的装载

数据的装载是根据用户定义的加载规则将数据缓冲区的数据直接送到数据库或数据仓库中对应的表中。在

运算而得到目标字段。这种规则一般针对数值型字段而言。

③参照转换。在转换中通常要用数据源的一个或多个字段作为Key，到一个关联数组中去搜索特定值，而且应该只能得到唯一值。这个关联数组使用Hash算法实现是比较合适也是最常见的，在整个ETL开始之前，它就装入内存，对性能提高的帮助非常大。

④字符串处理。从数据源某个字符串字段中经常可以获取特定信息，如交易类型，而且经常会有数值型值以字符串形式体现。对字符串的操作通常有类型转换、字符串截取等。但是由于字符类型字段的随意性也造成了核心数据的隐患，所以在处理这种规则的时候，一定要加上异常处理。

⑤空值判断。空值的处理是商业银行自动对账系统中一个常见问题。将它作为核心数据还是作为特定的一种维成员？这取决于业务规则，当某些记录的特定字段为空值时也许在数据清洗阶段就要将其过滤掉或筛选出来，例如银联文件中“冲正记录”字段为空值的就要筛选出来，不为空的就要过滤掉，因为有冲正记录的那笔的或依赖系统的值。

3.数据的装载

数据的装载是根据用户定义的加载规则将数据缓冲区的数据直接送到数据库或数据仓库中对应的表中。在商业银行自动对账系统中，数据经过清洗转换后得出的对账结果可以按照输出的需要分科目分类别写入数据库里不同的表中，然后由前端展示程序通过读取数据库来与用户交互，方便用户进行审核和调账处理等操作。

二、自动对账系统中的数据质量问题

商业银行自动对账系统中的ETL数据质量问题，首先是数据源的质量问题。如果在这个源头不能保障干净准确的数据，那么后面的分析功能的可信度就成问题。商业银行自动对账系统中的数据源来自银行服务器上的交易记录或银联机构提供的数据报表，于是数据源的质量有了充分的保障。然而，抛开数据源质量问题，自动对账过程中还可能存在以下因素对数据准确性产生重大影响。

(1)规则描述错误

设计人员对数据源系统理解的不充分，导致规则理

解错误。另一方面，如何无二义性地描述规则也是要探求的一个课题。规则是依附于目标字段的，但是规则总不能总是用文字描述，必须有严格的数学表达方式。

(2)ETL开发错误

即使规则很明确，ETL开发的过程中也可能发生一些错误，如逻辑错误、书写错误等。对于一个分段值，开区间闭区间是需要指定的，但是常常开发人员没注意，一个大于等于号写成大于号就会导致数据错误。

(3)人为处理错误

在整体ETL流程没有完成之前，为了图省事，通常会手工运行ETL过程，这其中一个重大的问题就是你不能按照正常流程去运行了，而是按照自己的理解去运行，发生的错误可能是误删数据、重复装载数据等。

对于自动对账过程中产生的质量问题，必须有保障手段，这个保障手段就是数据验证机制。它的目的是能够在自动对账过程中监控数据质量，产生报警。在商业银行自动对账系统中采取以下措施来避免可能出现的错误以保证对账结果的数据质量。

①提供前端。为银行业务人员提供友好的用户界面，可以即时看到每一步的对账结果，控制对账流程，

流程不规范。商业银行自动对账系统中在前端展示程序里可以很好的提示用户每一步的操作流程，避免出现人为处理错误。

三、自动对账系统的性能调优

商业银行自动对账系统中的整个ETL过程基本是通过控制采用SQL语句编写的存储过程和函数的方式来实现对数据的直接操作，SQL语句的效率将直接影响到数据仓库后台的性能。其主要特点是：面对海量的数据进行抽取；对大批量数据进行删除、更新和插入操作；面对异常的数据进行规则化的清洗；大量的分析和计算工作。所以，在商业银行自动对账系统中针对ETL过程的优化主要是结合其自身的特点，抓住需要优化的主要方面，针对不同的情况从如何采用高效的SQL入手来进行（以下的说明以Oracle为例，但其优化的方法和原理同样适合除Oracle之外的其他数据库）。

1.索引的正确使用

在海量数据表中，基本每个表都有一个或多个的索引来保证高效的查询，在ETL过程中的索引需要遵循以下使用原则。

够在自动对账过程中监控数据质量，产生报警。在商业银行自动对账系统中采取以下措施来避免可能出现的错误以保证对账结果的数据质量。

①提供前端。为银行业务人员提供友好的用户界面，可以即时看到每一步的对账结果，控制对账流程，并提示出错信息。

②提供框架。数据验证不是一次性工作，而是每次自动对账过程中都必须做的。因此，商业银行自动对账系统将每次验证结果数据都记录在表中，并且自动触发多维分析的数据装载、发布等。这样就提供了一个框架，自动化验证过程，并提供扩展手段，能起到规范化操作的作用。

③规范流程。前面提到有一种ETL数据质量问题是由于人工处理导致的，其中最主要原因还是

适合除Oracle之外的其他数据库）。

1.索引的正确使用

在海量数据表中，基本每个表都有一个或多个的索引来保证高效的查询，在ETL过程中的索引需要遵循以下使用原则。

(1)当插入的数据为数据表中的记录数量10%以上时，首先需要删除该表的索引来提高数据的插入效率，当数据全部插入后再建立索引。

(2)避免在索引列上使用函数或计算，在WHERE子句中，如果索引列是函数的一部分，优化器将不使用索引而使用全表扫描。

(3)避免在索引列上使用NOT和“!=”，索引只能告诉什么存在于表中，而不能告诉什么不存在于表中，当数据库遇到NOT和“!=”时，就会停止使用索引执行全表扫描。

(4)索引列上用≥替代>。

2.游标的正确使用

当在海量数据表中进行数据的删除、更新和插入操作时，用游标处理的效率是最慢的方式，但它在ETL过程中的使用又必不可少，而且使用有着及其重要的地

够在自动对账过程中监控数据质量，产生报警。在商业银行自动对账系统中采取以下措施来避免可能出现的错误以保证对账结果的数据质量。

①提供前端。为银行业务人员提供友好的用户界面，可以即时看到每一步的对账结果，控制对账流程，并提示出错信息。

②提供框架。数据验证不是一次性工作，而是每次自动对账过程中都必须做的。因此，商业银行自动对账系统将每次验证结果数据都记录在表中，并且自动触发多维分析的数据装载、发布等。这样就提供了一个框架，自动化验证过程，并提供扩展手段，能起到规范化操作的作用。

③规范流程。前面提到有一种ETL数据质量问题是由于人工处理导致的，其中最主要原因还是

适合除Oracle之外的其他数据库）。

1.索引的正确使用

在海量数据表中，基本每个表都有一个或多个的索引来保证高效的查询，在ETL过程中的索引需要遵循以下使用原则。

(1)当插入的数据为数据表中的记录数量10%以上时，首先需要删除该表的索引来提高数据的插入效率，当数据全部插入后再建立索引。

(2)避免在索引列上使用函数或计算，在WHERE子句中，如果索引列是函数的一部分，优化器将不使用索引而使用全表扫描。

(3)避免在索引列上使用NOT和“!=”，索引只能告诉什么存在于表中，而不能告诉什么不存在于表中，当数据库遇到NOT和“!=”时，就会停止使用索引执行全表扫描。

(4)索引列上用≥替代>。

2.游标的正确使用

当在海量数据表中进行数据的删除、更新和插入操作时，用游标处理的效率是最慢的方式，但它在ETL过程中的使用又必不可少，而且使用有着及其重要的地

WHERE子句的末尾。

(2)删除全表时用TRUNCATE替代DELETE

当DELETE删除表中的记录时，有回滚段(rollback segments)用来存放可以被恢复的信息，而当运用TRUNCATE时，回滚段不再存放任何可被恢复的信息，所以执行时间也会很短。同时需要注意TRUNCATE只在删除全表时适用，因为TRUNCATE是DDL而不是DML。

(3)尽量多使用COMMIT

商业银行自动对账系统中同一个过程的数据操作步骤很多，只要有可能就在程序中对每个DELETE、INSERT和UPDATE操作尽量多使用COMMIT，这样系统性能会因为COMMIT所释放的资源而大大提高。

(4)用EXISTS替代IN

在商业银行自动对账系统中，为了满足一个条件往往需要对另一个表进行联接，因此经常需要关联多个维表，在这种情况下，使用EXISTS而不用IN将提高查询的效率。

(5)用NOT EXISTS替代NOT IN

在子查询中，NOT IN子句将执行一个内部的排序

基于ETL技术的商业银行自动对账系统融合了先进的数据处理技术，能够对商业银行相关账目的对账、调账及清算处理全过程进行自动处理。该系统不但操作方便，计算正确，而且对账速度快，能帮助商业银行解决业务量不断增大带来的对账处理困难和潜在的人力成本问题。目前，该系统已经在中国银行山东分行和深圳平安银行投入使用，效果显著，银行业务人员反映良好。随着银行业务的不断扩大和信息技术的不断发展，商业银行的对账工作自动化程度也会不断提高，人工对账方式终将退出历史舞台，被计算机自动对账方式所取代。商业银行自动对账系统必将在降低业务风险、，提升商业银行账务处理的水平与效率等方面发挥越来越大的作用。 FCC

栏目编辑： 杨洁 yangjie@fcc.com.cn