

商业银行数据仓库系统中ETL的设计与实现

周 飞, 刘梦娜, 张 晖, 李志刚, 曹 帅

(安徽建筑大学电子与信息工程学院, 安徽 合肥 230601)

摘 要: 本文基于国内某商业银行数据中心的新核心系统项目, 首先介绍近年来数据仓库的国内外发展状况, 其次根据商业银行各业务系统的特点, 分析出建设商业银行数据仓库的重大意义和存储规划。最后详细介绍了数据调度ETL系统的开发与设计, 从各个源系统中抽取数据到数据仓库的实现方法, 对银行数据仓库的设计与实现有一定的参考价值。

关键词: 数据仓库; ETL; 商业银行; 商业智能

中图分类号: TP311 **文献标识码:** A

Design and Implementation of ETL in the Commercial Bank Data Warehouse System

ZHOU Fei, LIU Mengna, ZHANG Hui, LI Zhigang, CAO Shuai

(School of Electronic and Information Engineering, Anhui Jianzhu University, Hefei 230601, China)

Abstract: Based on a new Data Warehouse project for a commercial bank's data center in China, the paper firstly introduces the development status of Data Warehouse in recent years. Secondly, according to the characteristics of commercial banks' business systems, it analyzes the storage plan of building a data warehouse system for the commercial bank. In the end, it describes in detail a dispatching system using Extract Transform and Load (ETL) system. The study provides a valid reference for the commercial banks to construct a Data Warehouse system in a complicated environment.

Keywords: data warehouse; ETL; commercial bank; business intelligence

1 引言(Introduction)

数据仓库是应用于银行业的重要技术。近年来, 随着商业银行各业务系统的不断发展, 产生了日益庞大的历史数据, 这些数据是银行十分重要的战略资源, 将这些重要的业务数据存储于数据仓库中统一管理, 进行数据分析, 挖掘出其中潜在的重要的商业信息, 可以很好的辅助银行领导层管理决策, 解决银行最紧迫的问题, 使银行的经济效益最大化, 从而使银行更好更快的发展。在大数据时代下, 基于数据仓库的数据挖掘是银行业之间竞争的重要工具^[1]。

在1990年Bill Inmon提出了一种新的数据库技术, 即数据仓库, 目的是用来存储和组织各业务系统的海量数据, 并对数据进行分析。ETL(Extract, Transform, Load)系统工具专用于业务数据到数据仓库中存储的处理过程, 在数据源和数据仓库之间起到了桥梁和纽带的作用, 直接影响数据仓库的建设和运行。对于ETL系统的研究和开发是建设数据仓库必须考虑的重要问题, 通过改进ETL技术提高数据的传输和存储效率, 是建设商业银行数据仓库永恒的目标, 能更好的促进银行业的发展。

2 ETL技术简介(Introduction of ETL technology)

完整的数据仓库系统如图1所示, 数据的ETL是建立数据仓库系统的第一步^[2,3], 是应用于数据仓库项目的重要步骤, 为后续的数据挖掘和决策支持做好准备。

数据仓库中数据的调度工作是通过ETL工具来实现的。ETL是数据调度系统三个阶段抽取(Extract)、转换(Transform)、装载(Load)的英文首字母缩写^[4,5]。下面我们来简要的介绍ETL的三个阶段。

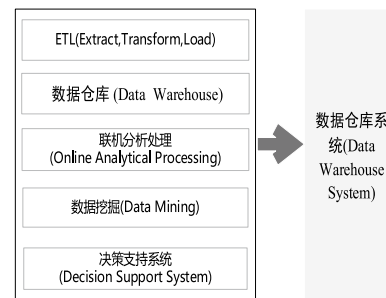


图1 数据仓库系统

Fig.1 Data warehouse system

2.1 数据抽取

首先分析从每个数据源获取数据的策略，分析从数据源系统中收集到的原始数据，存入ETL环境的物理存储磁盘上。每次执行完一次ETL调度后，源系统会产生新的数据，称这些新的数据为变化的数据，用变化的原因标记有变化的数据，抽取其中所有变化的数据。简要的抽取步骤为：

- (1)Dump：复制来源数据。
- (2)Capture：抽选所需资料。
- (3)Detect：数据变更获取。

2.2 数据转换

将业务数据以Push/Pull方式从源系统环境传送到数据仓库环境中。数据转换加工,完成的是从缓冲区到基础区的加工，以及基础区到汇总区的转换任务。在本项目中这类作业通过JAVA程序调用实现，JAVA程序调用SQL脚本，执行具有特定转化逻辑的SQL语句。简要的转换步骤为：

- (1)Staging Load：将数据加载暂存区。
- (2)Validation：检查数据并指定索引键。
- (3)Transformation：将来源文件的结构转换成目标仓储表格结构。
- (4)Merge：将多个数据源的数据融合。

2.3 数据装载

将源系统抽取转换的数据最终装载到数据仓库中，简要的装载步骤为：

- (1)Apply：将变更记录更新到数据仓库区。
- (2)Summarize and Mart：建立数据汇总和数据集市。
- (3)Cube：建立数据集市的多维分析。

3 ETL系统的架构设计(Framework design of the ETL system)

我们分析了国内A商业银行各业务系统的数据源，对数据仓库存储的数据做一个总体的预估，如表1所示。根据数据的存储情况，设计相应的ETL系统架构，把数据从数据源系统中抽取导出数据，加载至数据仓库中^[6]。由于在源系统开发中有着非常多的挑战，导致ETL系统开发存在着极高的难度。任何意外情况的发生都会导致ETL系统的开发延期，耽误项目整体进度，增加项目成本。因此ETL调度系统的架构设计一定要非常仔细谨慎。

表1 数据仓库存储规划

Tab.1 Data warehouse storage plan

项目	数据要求	存储空间计算	存储
文件缓存区	保存7天数据	7*0.1*3	≈0.2T
报表文件存储区	保存5年的报表数据	365*5*4.5*3	≈24T
历史数据存储	保存5年全部历史数据	365*5*10*3	≈53T
总计			≈77T

3.1 ETL系统的功能与特点

根据商业银行的业务特点和客户对产品的需求，此项目设计的ETL系统将实现以下功能：

- (1)ETL支持数据源多，异构平台的数据关联。支持系统的可视化、参数化。支持集群、负载均衡。

- (2)ETL支持事件触发、消息触发等多种触发机制。

- (3)ETL支持多并发处理方式，可以根据系统处理资源的情况调整并发程度，处理效率高。

- (4)ETL支持多节点处理，可以根据系统性能的要求，增加硬件功能来提高系统的生成效率。

3.2 ETL系统架构设计

结合国内某商业银行的业务特点，我们规划出ETL系统架构图如图2所示。

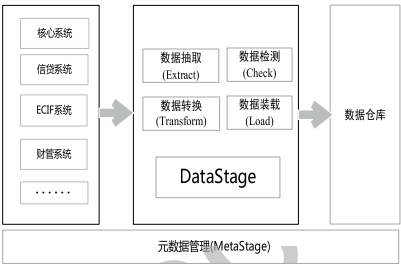


图2 ETL架构设计

Fig.2 ETL architecture design

对于详细的ETL系统架构，我们参照下面四个方面来设计^[7-9]。

- (1)设计全局规划

规划源数据，各系统业务逻辑，目标数据仓库三级设计架构。根据各个系统业务逻辑的不同，开发相应的ETL系统，解决当前数据调度存在的和还未解决的问题。

- (2)选择ETL工具

在数据仓库市场上存在多种ETL工具，本文采用调度抽取工具Datastage主动到各个源系统中获取数据。

- (3)开发默认策略

对ETL工具的基本需求进行整体考虑，开发默认策略。这些包括：从每个源系统获取数据，归档获取的数据或分级的数据，监管维度和特定事实的数据质量，维度属性变化的管理，确保数据仓库和ETL系统满足系统可用性需求，设计数据审计子系统，组织ETL过渡区。

- (4)按照目标表获取数据

开发完所有的ETL调度任务后，还需要深入研究详细的转换工作。确定数据仓库中数据来源的目标表，完成源数据到目标表的映射后，继续完成数据概要描述工作，全部理解每个表所需要的数据转换。

4 ETL在BI项目中的应用(ETL application in the Business Intelligence project)

下面在A商业银行BI(Business Intelligence)项目，我们通过ETL调度系统来完成数据的转换和存储。在遵循ETL逻辑架构的基础上，为各个源数据系统中的每一张业务表设计一个调度作业，表中数据的开发就是对调度作业数据的ETL操作过程。完成调度作业中数据的ETL阶段代码编写、ETL调度开发，实现整个ETL数据存储到数据仓库中^[10,11]。

进行ETL调度系统开发，首先要搭建开发环境，包括：导入初始化数据、设计调度程序、设计监控程序。其中调度程序的设计分为ETL任务设计和ETL调度设计，此外还要开发一个数据检测程序SendMessage作为调度程序的辅助工具。监

控程序的设计就是ETL监控设计。

4.1 ETL任务设计

首先要分析整理出调度作业的任务信息,包括数据源系统、各个系统中所有调度作业的编号、调度作业的阶段信息、调度作业之间的前后驱关系等。具体设计步骤如下:

(1)按照数据源系统和作业处理阶段收集任务信息,包括任务的执行周期、任务类型、任务渠道、任务阶段等。把收集到的所有数据源系统依次编号,例如现在已经收集到了核心系统、信贷系统、ECIF系统、财管系统,依次将核心系统编号为10000、信贷系统编号为20000、ECIF系统编号为30000、财管系统编号为40000,后续出现的数据源系统以10000为单位编排下去。将作业每个调度阶段进行编号,如数据检测阶段编号为1000,数据导出阶段编号为2000,数据传输阶段编号为3000,数据装载阶段编号为4000,数据备份阶段编号为5000,其他阶段编号为9000等。

(2)对收集到的所有作业进行任务编码,例如核心系统共100个作业,每个作业都经过检测装载备份三个阶段,第一个作业检测阶段编号11001,其中万位数的1代表核心系统10000,千位数的1代表检测阶段1000,后三位001代表第一个作业。第一个作业生成装载阶段的作业编号14001,备份阶段的作业编号15001,其他作业依次按顺序生成相应的检测,装载,备份作业编号。统计所有的作业编号,阶段信息和渠道信息,生成作业信息表。

(3)根据作业信息表中的任务信息,分析其所属的阶段和数据源,确定各任务之间的前后驱关系。一个ETL调度系统一般都是由上万个不同的任务共同协作完成,任务之间的关系决定整个数据调度过程是否能够顺利进行。ETL调度系统中处理作业是按顺序来的。先执行的作业被称为前驱任务,用来决定其他作业是否执行,后执行的作业被称为后驱任务。ETL任务中只有前一个作业处理执行成功才会执行后一个作业,所以作业之间的前后驱关系很重要。每个数据源系统设有一个首作业,首作业的后驱作业是所有的检测作业,检测作业的后驱作业是装载作业,装载作业的后驱作业是备份作业。梳理完成所有作业的前后驱关系,整合到调度系统中,在作业信息表中添加阶段开始和结束的标志性任务,生成作业流程表。

(4)将前面整理好的作业任务转化为SQL语句程序,在ETL工作站上部署该程序,系统启动后,调度生成程序根据作业信息表和作业流程表生成作业调度表。

4.2 ETL调度设计

由于数据源系统非常多,所需加载的数据也非常多,导致ETL调度程序的设计非常复杂,因此设计一个完善且易于维护的ETL调度程序,对数据仓库的正常工作运行是非常重要的^[12]。ETL调度流程如图3所示。结合业务系统的特点,设计出高效的ETL调度程序,具体设计步骤如下:

(1)作业调度表生成完毕,开始执行调度任务程序,根据作业调度表中作业信息,作业调度执行的情况,生成调度执行情况表。

(2)根据作业优先级关系向各个ETL工作站分配相应数量的作业任务。如果某ETL工作站执行作业已经完成,则将新的作业分配给该ETL工作站,并更新调度执行情况表。

(3)假如某个调度任务执行出错,则根据作业错误处理级别对作业进行相应处理,如等待人工处理、自动跳过错误任务、重复执行等。如果需要人工处理调度任务,则人工对调度任务进行相应处理,并记录日志。

(4)当作业调度表中任务全部执行完成后,暂停调度程序,等待下一批次调度任务,ETL工作站也处于等待状态。

4.3 ETL监控设计

ETL作业任务运行过程中,需要实时监控所有的作业执行情况,其中包括ETL单个作业的状态监控等信息,ETL监控的主要作用是监控任务的执行情况及支持对错误任务的手工处理。

接收调度执行终端的反馈结果并进行相应的处理,反馈结果有任务成功,失败状态。接收数据检测消息发送器的消息并做相应的处理,如接收到核心系统渠道2018-03-01的数据准备好,调度则开始跑批对应的任务。ETL监控相关流程如图4所示。

4.4 ETL调度系统的实现

在上述设计的ETL调度系统中,所有的操作过程都是通过SQL程序语句来实现的。开发外部SQL脚本程序,调用该SQL程序来实现ETL系统的数据调度功能。SQL脚本逻辑可以根据不同的需求分别编写,实现其功能,这样做的好处是配置作业比较容易。

根据调度作业信息表的结构创建一个临时作业表INIT_JOB, INIT_JOB表起到一个调度作业中转的作用。INIT_JOB表结构如图5所示。该临时表可以记录SQL脚本程序调用作业所需的大部分信息,例如:作业类型、作业名称、所属阶段和渠道、调用脚本及参数等。这样做的好处是可以方便的初始化作业信息。向INIT_JOB表中添加新增作业信息,通过INIT_JOB临时表完成新增作业的信息存储,调用SQL脚本完成作业的数据开发过程,最终将表中的业务数据存储备份至数据仓库中。当需要重新调度新的作业时,初始化INIT_JOB表为空,插入新增作业信息,根据新作业的业务特点,修改SQL脚本逻辑,实现新增作业的调度功能。

当作业调度开发结束后,通过查询调度作业日志,查看数据调度是否成功,调度成功则数据成功存储至数据仓库中,如图6所示。若调度失败,则查看日志,找到调度失败原因并修改,重新执行调度任务,直至调度作业执行成功。

5 结论(Conclusion)

本论文以国内某商业银行的BI项目为背景,设计开发了

商业银行数据仓库的ETL调度系统，实现了银行各业务数据到数据仓库的集中存储，实现全行数据的统一化、标准化、高质量、高效率，极大的提升了商业银行数据信息化水平。在大数据时代下，数据挖掘是银行业之间竞争的重要工具，基于数据仓库的数据分析，挖掘出其中潜在的重要的商业信息，可以很好的辅助银行领导层管理决策，解决银行最紧迫的问题，使银行的经济效益最大化，从而使银行更快更好的发展。

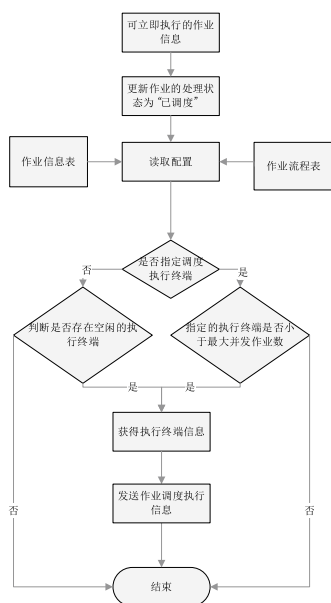


图3 调度作业的处理流程图

Fig.3 Flow chart of scheduling operation

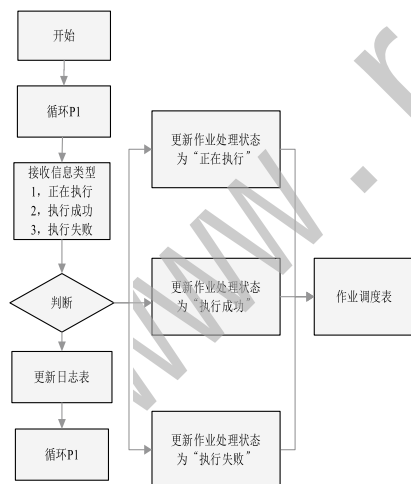


图4 ETL监控流程图

Fig.4 ETL monitor flow chart

A	B	C	D	E	F	G	H	I	J	K	L
ID	TYPE	JOBNAME	JOBDSRIBE	JOBPRIOITY	STAGE	CHANNEL	OBJCLE	APPLCL	PARAMETER	JOBVALUE	JOBLENORE
1	2	1 [script[]]		1	10000	30000000 D			diagjob/bin/script; data_csh	0	1
2	2	2 [script[]]Change the ELT status in fun system		1	1200000	120000000 D			diagjob/bin/script; \$1TTTTT0000	jsd	1
3	2	3 [script[]]End of ELT execution, change fun system :		1	1200000	120000000 D			diagjob/bin/script; \$1TTTTT0000	jsd	1
4	2	4 [script[]]Change the ELT status in pl system :		1	1200000	120000000 D			diagjob/bin/script; \$1TTTTT0000	jsd	1
5	2	5 [script[]]End of ELT execution, change pl system :		1	1200000	30000000 D			diagjob/bin/script; \$1TTTTT0000	jsd	1
6	2	6 [script[]]Change the ELT status in sop system :		1	130000	100000000 D			diagjob/bin/script; \$1TTTTT0000	jsd	1
7	2	7 [script[]]End of ELT execution, change sop system :		1	130000	100000000 D			diagjob/bin/script; \$1TTTTT0000	jsd	1
8	2	8 [stat system; _execinfo[]]		999	50000	100000000 D			diagjob/bin/script; \$1TTTTT0000	wh	1
9	2	9 [diag_bin; _execinfo[]]		999	40000	100000000 D			diagjob/bin/script; \$1TTTTT0000	jsd	1
10	2	10 [di_bin; _execinfo[]]		999	40000	100000000 D			diagjob/bin/script; \$1TTTTT0000	jsd	1

图5 INIT JOB临时作业表结构

Fig.5 INIT_JOB table structure

[illegible]

图6 调度作业执行成功

Fig.6 Scheduled job execution succeeded

参考文献(References)

- [1] Ren S,Sun Q,Shi Y.Customer Segmentation of Bank Based on Data Warehouse and Data Mining[C].The IEEE International Conference on Information Management and Engineering, IEEE,2010:349-353.
- [2] L Baldacci,M Golfarelli,S Graziani,et al.QETL:An Approach to On-Demand ETL from Non-Owned Data Sources[J].Data & Knowledge Engineering,2017.
- [3] V Theodorou,A Abelló,M Thiele,et al.Frequent Patterns in ETL Workflows: An Empirical Approach[J].Data & Knowledge Engineering,2017:112.
- [4] Dupon S,Jovanovic V.An approach to conceptual modelling of ETL processes[C].International Convention on Information and Communication Technology,Electronics and Microelectronics. IEEE,2014:1485-1490.
- [5] 徐俊刚,裴莹.数据ETL研究综述[J].计算机科学,2011,38(4): 15-20.
- [6] 常艳莉,李一凡,赵怀慈.基于银行数据仓库的ETL系统的设计与实现[J].微计算机信息,2006, 22(24):21-23.
- [7] 沈琦,陈博.基于大数据处理的ETL框架的研究与设计[J].电子设计工程,2016,24(2):25-27.
- [8] 张同杨.一种ETL数据整合架构设计与实现[J].价值工程,2016,35(31):88-91.
- [9] 张靖,雷航,唐雪飞,等.ETL应用优化设计与实现研究[J].微电子学与计算机,2012,29(4):134-137.
- [10] 林尤惠,项颖,张凤祺.ETL在银行数据中心系统中的应用[J].自动化与信息工程,2009,30(2):17-20.
- [11] Waas F,Wrembel R,Freudenreich T,et al.On-Demand ELT Architecture for Right-Time BI:Extending the Vision[J].International Journal of Data Warehousing & Mining,2013,9(2):21-38.
- [12] 宋旭东,刘晓冰.数据仓库ETL任务调度模型研究[J].控制与决策,2011,26(2):271-275.

作者简介:

周 飞(1989—),男,硕士生.研究领域:机器学习,数据挖掘.
刘梦娜(1992—),女,硕士生.研究领域:机器学习,数据挖掘.
张 晖(1994—),男,硕士生.研究领域:机器学习,数据挖掘.
李志刚(1994—),男,硕士生.研究领域:机器学习,数据挖掘.
曹 帅(1993—),男,硕士生.研究领域:机器学习,数据挖掘.