ETL流程概述及常用实现方法

2015年09月05日 13:01:52 Cloud-g 阅读数：18831

ETL是英文Extract-Transform-Load 的缩写，用来描述将数据从来源端经过抽取（extract）、转换（transform）、加载（load）至目的端的过程。[illegible]数仓发中将数据由业务系统归集到数据仓库（DW）或者数据集市的过程。在ETL三个部分中，花费时间最长的是“T”(Transform，清洗、转换)的部分，[illegible]清况工作量是整个ETL的2/3。

1抽取作业

将从源数据库（通常为业务系统）获得数据的过程。

在做这一步的之前，往往要预先分析自己需要什么数据，划分好范围，确认具体的技术部门和业务部门。

1.1手工开发抽取作业时候的常用方法：

1.1.1当数据源和DW为同一类数据库时

一般情况下，DBMS(SQLServer、Oracle)都会提供数据库链接功能，可以在数据源（业务系统）和DW内建立数据库链接（如DB2的联邦数据库NICKNAME后在DW内直接SELECT访问。

优点是实现使用简单，逻辑简单；缺点是容易被滥用对源数据库造成较大的负载压力。

1.1.2当数据源和ODS为不同类型数据库时

o 将源数据库的数据导出为文本文件，利用FTP协议进行传输导入ODS区域。

优点是实现简单，对源系统压力较小。缺点是传输步骤增加了，处理需要的时间增加。

o 将部分数据库间能通过ODBC建立源数据库和目标数据库链接，此时也能直接使用SELECT获取数据。

优点是实现使用简单，逻辑简单；缺点是容易被滥用对源数据库造成较大的负载压力，且建立时较为复杂。

1.2更新数据的时间和数量的问题

1.2.1实时抽取数据

这类抽取方式在数据仓库中很少见到，因为一般来说数据仓库对数据的实时性要求并不高。实时抽取常见于BI中的CRM系统，比如在实时营销中，客户一旦某类操作就实时触发对应的营销行为。

o 时间戳方式

要求源表中存在一个或多个字段(时间戳),其值随着新纪录的增加而不断增加，执行数据抽取时，程序定时循环检查通过时间戳对数据进行过滤，抽取结束后录时间戳信息。

这种方式的优点是对源系统的侵入较小，缺点是抽取程序需要不断扫描源系统的表，对其 有一定压力。

o 触发器方式

要求用户在源数据库中有创建触发器和临时表的权限，触发器捕获新增的数据到临时表中，执行抽取时，程序自动从临时表中读取数据。

这种方式的优点是实时性极高，缺点是对源系统的侵入性较大，同时会对源数据库造成很大的压力（行级触发器），很可能影响源系统的正常业务。

o 程序接口方式

改造源系统，在修改数据时通过程序接口同步发送数据至目标库，发送数据的动作可以跟业务修改数据动作脱耦，独立发送。

这种方法的优点是对源系统的造成压力较小，实时性较强；缺点是需要对源系统的侵入性较强，需要源系统做较大的改造。

为了保证数据抽取时数据的准确性、完整性和唯一性，同时降低抽取作业对源数据库造成的压力，抽取作业的加载必须避开源数据的生成时间。这种方法一时性要求不高的数据。比如T+1或者每月1日进行抽取。

1.2.2.1常用实现

o 日志检查

需要源数据库生成数据完毕之后，在外部生成日志。抽取程序定时检查源系统的执行日志，发现完成标志后发起抽取作业。

这种方式优点是可靠性高，对源数据库造成的压力较小。缺点是需要源数据库配合生成可供检查的外部日志。

o 约定时间抽取

可以直接约定一个加载完毕同时对源数据库压力较小的时间（如每日凌晨2点），抽取程序建立定时任务，时间一到自动发起抽取作业。

这种方式优点是对源数据库的侵入性和造成的压力较小；缺点是可靠性不高，可能会发生数据未生成完毕也直接进行抽取的情况。

1.2.2.2根据下载时候对数据的筛选方式可以分为

o 全量下载

用于：

·源数据量较小，如维表。

·数据变化较大，比如90%的数据都产生了变化的表。

·变化的数据不能预期，无法标示，如账户表。

的时候。

优点在于下载较为简单且能容纳任何情况的数据变化；缺点是如果数据量较大，需要抽取相当长的时间，同时会占用大量的IO和网络资源。

o 增量下载

常用于数据只增不减的表，如交易明细表等。

·时间戳

源系统在修改或添加数据时更新对应的时间戳字段（如交易表的日期字段），抽取程序根据时间戳选择需要更新的数据进行抽取。

·触发器方式

要求用户在源数据库中有创建触发器和临时表的权限，触发器捕获新增的数据到临时表中，到执行抽取的时间时，程序自动从临时表中读取数据。占用资源建议使用。

优点是下载的数据较小，速度较快，占用资源少；缺点是使用限制较大，有时候需要源系统进行改造支持。

2转换作业

这一步包含了数据的清洗和转换。

2.1数据清洗

任务是过滤不符合条件或者错误的数据。

这一步常常出现在刚刚开始建立数据仓库或者源业务系统仍未成熟的时候，此时发现错误数据需要联系源业务系统进行更正，部分可预期的空值或者测试用过滤掉。

2.2数据转换

这一步是整个ETL流程中最为占用时间和资源的一步。

数据转换包含了简单的数据不一致转换，数据粒度转换和耗时的数据关联整合或拆分动作。这里可能存在各种各样千奇百怪的需求。对于核心数据仓库来说往是对数据进行按照主题划分合并的动作。同时，也会添加一些为了提升执行效率而进行反范式化添加的冗余字段。

根据实现方式的不同，可以区分为使用数据库存储过程转换和使用高级语言转换

o 使用数据库存储过程转换

使用SQL开发存储过程完成转换作业是很多银行常用的方法。

它的优点是开发简单、能支持绝大部分转换场景；缺点在于占用资源多且受制于单一数据库性能，无法做到横向扩展。

因此，除了业务的理解能力外，对SQL海量数据处理的优化能力在此也非常重要。比如:

· 利用数据库的分区性，选择良好的分区键。

· 建表时合理选择主键和索引，关联时候必须使用主键或索引进行关联。

· 关注数据库对SQL的流程优化逻辑，尽量选择拆分复杂SQL，引导数据库根据你选择流程进行数据处理

· 合理反范式化设计表，留出适当的冗余字段，减少关联动作。

具体的优化根据不同的数据库有着不同的处理方式，根据所选用的数据库不同而定。

o 使用高级语言转换

使用高级语言包含了常用的开发C/C++/JAVA等程序对抽取的数据进行预处理。

自行使用高级语言开发的优点是运行效率较高，可以通过横向扩展服务器数量来提高系统的转换作业处理能力；缺点是开发较为复杂，同时虽然能进行较为辑的开发，但是对于大数据量的关联的支持能力较弱，特别是有复数的服务器并行处理的时候。

3加载作业

转换作业生成的数据有可能直接插入目标数据库，一般来说，这种情况常见于使用数据库存储过程进行转换作业的方案。此时，ETL作业位于目标数据库上业只需要使用INSERT或者LOAD的方式导入目标表即可。此时转换作业和加载作业往往是在同一加工中完成的。

当使用高级语言开发时，ETL作业有着专门的ETL服务器，此时，转换作业生成的往往是文本文件，在转换作业完成后需要使用目标库特有的工具导入或者通INSERT入目标库。

同时，根据抽取作业的数据抽取方式的不同（全量、增量），对目标表进行替换或者插入动作。

4流程控制

抽取加载和转换作业需要一个集中的调度平台控制他们的运行，决定执行顺序，进行错误捕捉和处理。

较为原始的ETL系统就是使用CRON做定时控制，定时调起相应的程序或者存储过程。但是这种方式过于原始，只能进行简单的调起动作，无法实现流程依同时按步执行的流程控制能力也弱，错误处理能力几乎没有。只适合于极其简单的情况。

对于自行开发的较为完善的ETL系统，往往需要具有以下几个能力:

· 流程步骤控制能力

调度平台必须能够控制整个ETL流程（抽取加载和转换作业），进行集中化管理，不能有流程游离于系统外部。

· 系统的划分和前后流程的依赖

由于整个ETL系统里面可能跨越数十个业务系统，开发人员有数十拨人，必须支持按照业务系统对ETL流程进行划分管理的能力。

同时必须具有根据流程依赖进行调度的能力，使得适当的流程能在适当的时间调起。

· 合理的调度算法

同一时间调起过多流程可能造成对源数据库和ETL服务器还有目标数据库形成较大负载压力，故必须有较为合理的调度算法。

· 日志和警告系统

必须对每一步的流程记录日志，起始时间，完成时间，错误原因等，方便ETL流程开发人员检查错误。对于发生错误的流程，能及时通知错误人员进行错误复。

· 较高可靠性

5常用商业ETL工具

常用的ETL工具有Ascential公司的Datastage、Informatica公司的Powercenter、 NCR Teradata公司的ETL Automation等。

· Datastage

是使用高级语言进行开发ETL服务器的代表。使用JAVA进行开发E/T/L的整个流程，同时支持平行添加服务器提升处理效率的方法。

· Automation

基于Teradata的TD数据库的ETL调度框架。其ETL流程是使用DSQL的存储过程进行开发，利用TD数据库的海量数据处理能力，也具有一定的平行扩展能力

ETL介绍与ETL工具比较

5.2万

本文转载自：http://blog.csdn.net/u013412535/article/details/43462537 ETL，是英文 Extract-Transfor… 来自： wl044090432的博客

相关热词 etl同步 etl的过程 etl流式处理 etl特殊字符 etl的英语

ETL讲解（很详细！！！）

7777

ETL讲解（很详细！！！） ETL是将业务系统的数据经过抽取、清洗转换之后加载到数据仓库的过程…

博主推荐

小小工匠 关注 597篇文章

青龙白虎米老鼠 关注 101篇文章

kowity 关注 109篇文章

ETL流程、数据流图及ETL过程解决方案

*详细原因: 取 消 提 交 ETL流程、数据流图及ETL过程解决方案 10 积分 立即下载 ...

BI开发流程和ETL介绍 - CSDN博客

BI开发流程和ETL介绍 ETL中的E->(ODS->SDE->SIL)(强大的ETL工具)ETL中的T-> ETL中的L->DW->BIEE(RPD物理)-> BIEE(RPD逻…

流行ETL数据传输解决方案 - CSDN博客

第2.1IBMDataStage解决方案 2.1.1IBMDataStage简介 IBM InfoSphereDataStage 是业界较为流行的ETL(Extract,Transform, Load)工…

ETL的过程原理和数据仓库建设 - CSDN博客

原文出处:http://home.dwway.com/home-space-uid-50388-do-blog-id-7366.html1.引言 数据仓库建设中的ETL(Extract, Transform, Load…

ETL的申请编写流程 - CSDN博客

【环境申请说明】1、测试环境写好源库目标库连接串信息,直接找架构组,修改架构图,派工单(有个工单号,在itsm上可以查,测试环境会很…

BI开发流程和ETL介绍

7957

BI开发流程和ETL介绍 ETL中的E->(ODS->SDE->SIL)(强大的ETL工具)ETL中的T-> ETL中的L->DW->… 来自： 狒狒也是程序猿的专栏

ETL的经验总结

6347

ETL的考虑 做数据仓库系统，ETL是关键的一环。说大了，ETL是数据整合解决方案，说小了，… 来自： 少年休闲海

流行ETL数据传输解决方案

5776

第 2.1IBMDataStage解决方案 2.1.1IBMDataStage简介 IBM InfoSphereDataStage 是业界较为流行的… 来自： lvzhuyiyi的博客

大数据处理过程之核心技术ETL详解

9391

ETL (数据转换)就是对数据的合并、清理和整合。通过转换，可以实现不同的源数据在语义上的一致… 来自： 数控小J 对大数据的…

Oracle BI基础之ETL数据增量抽取方案

9140

一篇好文见百度文库： ETL数据增量抽取方案 一、 ETL 简介 数据集成是把不同来源、格式和特点的… 来自： 苏南生的CSDN博客

关于数据迁移的方法、步骤和心得

6428

关于数据迁移的方法、步骤和心得 在项目中经常会遇到系统完全更换后的历史数据迁移问题，以示对… 来自： 语不惊人死不休

如何在不停机的情况下，完成百万级数据跨表迁移

2314

技术团队面临的困难总是相似的：在业务发展到一定的时候，他们总是不得不重新设计数据模型，以… 来自： kangbin825的专栏

ETL开发规范 2776

ETL规范概述 1.1 ETL含义：ETL是数据抽取（Extract）、转换（Transform）、装载（Loading）的缩… 来自： 技术积累

ETL工具-kettle安装部署 1889

ETL 来自： July_whj

ETL工具kettle源码编译 3605

kettle是一个开源项目，作为ETL工具，kettle提供了丰富的功能和简洁的图形化界面。 本文章详细介… 来自： verne_feng的博客

ETL工具Kettle的基本使用 9530

0.ETL简介ETL，是英文 Extract-Transform-Load 的缩写，用来描述将数据从来源端经过抽取（extract… 来自： embracejava

下载 **数据采集系统ETL工具** 10-26

数据采集系统主要解决了药品经销商业务系统的进销存数据的采集、上报的问题。医药渠道数据的采集需要连接多种数据源，实现异构数据之间灵活导数据。本系统兼容了SQL Ser…

小白学习大数据测试之ETL 290

之前发布过一篇关于ETL的文章，无奈被人说太简单。。。。唉，小编也是刚接触啊，自然不能那么… 来自： 测试帮日记

OLAP之全过程介绍（ETL过程） 925

经过多年来企业信息化建设，大部分都拥有了自己的财务，OA，CRM 等软件。这些系统都有自己的… 来自： lhy55040817的专栏

下载 **ETL概述及部分工具比较** 03-12

ETL概述及部分工具比较，基本点etl介绍

下载 **2018年新能源汽车行业概述及产业链全分析** 07-29

2018年新能源汽车行业概述及产业链全分析，内部资料2018.7

下载 **ETL概述及部分工具比较.rar** 09-10

ETL概述及部分工具比较; olap 专业工具介绍；报表工具介绍

本月试用各大杀毒软件的感想 986

上个月台湾地震导致海底光缆断裂，使得国内使用外国杀毒软件的用户升级不便，所以国内各大… 来自： 悠悠思故乡

ETL方法与过程讲解 1.4万

1 ETL基本概念和术语1.1 ETLExtract-Transform-Load的缩写，数据抽取（Extract）、转换（Transfor… 来自： 望月思灵

下载 **ETL流程、数据流图及ETL过程解决方案** 08-11

ETL流程、数据流图及ETL过程解决方案

ETL流程标准化思路 1757

根据培训计划，以下是整理的每个ETL开发维护的每个阶段的输入和产出物。 ETL流程可以分成五个… 来自： zjw00417236的专栏

ORACLE EBS常用表及查询语句（最终整理版） 1521

建议去看参考二 参考一: call fnd_global.APP... 来自： Jane

ETL工具——kettle插件开发（基础篇） 4545

在我们做ETL工作的时候，在某些项目中往往会遇到一些特别的流程任务，kettle原有的流程处理节点… 来自： 天将降大任于是人

etl kettle 执行日志输出到数据库 1934

1.右键进入转换设置页面，选择日志表要放在哪个数据库几日志表名称 2.设置好后，点击下方的SQL… 来自： 邓轩

ETL学习之八：添加日志记录 1282

Microsoft SQL Server 2005 Integration Services (SSIS) 包含日志记录功能，这些功能使您可以通过提… 来自： 俞欣力's 技术博客

ETL AUTOMATION介绍 3857

分类： 数据仓库与数据挖掘 /**********/ 目录： 第一部分：ETL Automation简介 … 来自： ding_shan的专栏

SqlServer ETL 数据抽取工具SSIS之环境搭建 7089

SSIS 是Microsoft SQL Server integration Servers 的简称，是数据集成的解决方案，它包含数据提取… 来自： lucius_yu00的专栏

接口的概述和讲解 114

/* 接口的特点： A:接口用关键字interface表示 interface 接口名 {} B:类实现接口用implements表示 cla… 来自： dwq_5678的博客

Java语言概述、环境搭建及新增功能介绍 96

目录1、概述 1、概述 Java语言是一门非常纯粹的面向对象编程语言，它吸收了C++语言的各种优点，… 来自： 疯人愿

反渗透技术工艺特点与工作原理 1361

反渗透设备工程工作原理 ReverseOsmosis 逆渗透为现有科技中最有效的水处理方式之一，他能… 来自： 反渗透设备

Cloud-g 关注

原创	粉丝	喜欢	评论
4	5	1	1

等级： 博客 2 访问： 2万+

积分： 280 排名： 30万+

最新文章

《数据仓库》读书笔记：第二章

《数据仓库》读书笔记：第一章

海量数据处理的SQL性能优化

个人分类

数据仓库	4篇
数据库	4篇
ETL	2篇
SQL	3篇

归档

2017年2月	2篇
2015年9月	2篇

热门文章

ETL流程概述及常用实现方法
阅读量：18809

海量数据处理的SQL性能优化
阅读量：2446

《数据仓库》读书笔记：第一章
阅读量：761

《数据仓库》读书笔记：第二章
阅读量：515

最新评论

《数据仓库》读书笔记：第二章
dengjian1169：更新啊

联系我们

扫码联系客服　官方公众号

QQ客服　kefu@csdn.net
客服论坛 (以上工作时间8:00-22:00)
400-660-0108 (工作时间8:00-19:00)

关于我们　招聘　广告服务　网站地图