# 烟草数据中心 ETL 技术应用①

密　红，何利力，杨秀梅

(浙江理工大学，杭州 310012)

**摘　要**：简要介绍了 ETL 技术的主要功能和主要实现方法，并结合国内某烟草省公司数据中心营销主题域建设项目的具体需要，提出了一种适合该项目的具体 ETL 实现方法，该方法成功将市烟草公司业务系统中主数据和营销业务数据抽取到省公司数据仓库中，对整个数据中心项目的顺利完成起到了至关重要的作用，也对其他行业数据中心建设提供了相关经验。

**关键词**：ETL；数据仓库；增量抽取；商务智能；数据模型

## ETL Technology Applied in DC of Tobacco Company

MI Hong, HE Li-Li, YANG Xiu-Mei

(Zhejing Sci-Tech University, Hangzhou 310012, China)

**Abstract:** This article briefly introduces the main functions and main implementation methods of ETL, and proposes a method suitable for a sale subject DC project of a domestic provincial company of tobacco. This method can successfully extract basic data and transaction data from systems of city companies to the data warehouse of the provincial company. It plays an important role in the whole DC project and can give some relative experience to the DC project construction of other industries.

**Keywords:** ETL; data warehouse; incremental extraction; business intelligence; data model

经过多年的信息化建设，目前的烟草商业企业都有了各自的营销业务系统。为了充分挖掘信息资源的潜在价值，全面提升管理的科学性、及时性、有效性，提高省市两级的管理和决策能力，建立一个集中管理、安全规范、充分共享、全面服务的数据中心是非常必要的。但各个系统间关键性数据结构和编码标准存在不一致的现象，因此 ETL 过程就显得尤为重要,ETL 设计的好坏直接影响数据仓库中数据的质量。

## 1　ETL主要功能

ETL，是 Extraction-Transformation-Loading 的缩写，就是数据抽取、转换和装载的过程，是将分散的、标准不统一，编码不一致的数据整合到目标数据库中的技术。

### 1.1 数据抽取

数据抽取是从数据源中抽取数据仓库所需数据的过程，主要有以下两种抽取方式。

(1) 全量抽取

全量抽取类似于数据迁移或复制，它将数据源中的表或视图数据原封不动的抽取出来再进行后续操作。全量抽取比较简单，只用于最初填充数据仓库。

(2) 增量抽取

增量抽取只抽取自上次抽取以来数据库中要抽取的表中新增或修改的数据，用于数据仓库的维护，增量抽取比全量抽取应用更广泛。如何准确捕获变化的数据是增量抽取的关键。目前常用的捕获变化数据的方法有：触发器方式、时间戳方式、全表比对方式、日志比对方式等。

① 基金项目:国家高技术研究发展计划(863)(2008AA042304)
收稿时间:2010-09-03;收到修改稿时间:2010-10-11

### 1.2 数据清洗转换[1]

数据仓库一般分为 ODS、DW 两部分。通常的做法是从业务系统到 ODS 做清洗，将脏数据和不完整数据过滤掉，再从 ODS 到 DW 的过程中转换，进行一些业务规则的计算和聚合。

#### 1.2.1 数据清洗

数据清洗的任务是识别那些错误的数据，将之交给主管部门修正之后再进行抽取。错误数据的种类主要有不完整的数据、错误的数据、重复的数据三大类。

#### 1.2.2 数据转换

数据转换的目的是将源系统的数据格式转换成目标系统的数据格式。

(1) 不一致数据转换：将不同业务系统的相同类型的数据统一，比如供应商编码、卷烟编码可能存在不一致的现象，这就需要统一转换成相同的编码。

(2) 数据粒度的转换：业务系统通常存储的是明细数据，而数据仓库并不需要非常明细的数据来做分析。所以通常要将明细数据按照设计的粒度进行聚合。

(3) 业务规则计算：企业的业务规则和数据指标各有不同，有的指标的计算比较复杂，比较好的做法是在 ETL 过程中将指标计算好再存储到数据仓库中。

### 1.3 数据装载

数据装载[2]是把数据加载到数据仓库或数据集市的过程。加载数据主要有两种方式：刷新方式和更新方式。刷新方式主要用于数据仓库创建时填充数据仓库，更新方式用于维护数据仓库。具体加载方法可以使用专门的加载工具或使用 sql 命令。

## 2 烟草数据中心ETL应用

ETL 有多种实现方法，常用的主要有三种，第一种是借助 ETL 工具[3]，如 IBM 的 Datastage,Oracle 的 OWB、SQL server 2000 的 DTS、SQL Server2005 的 SSIS 服务等，第二种是 SQL 方式实现，第三种方法是将 ETL 工具与 SQL 相结合。第一种方法可以屏蔽复杂的编码任务，提高速度，降低难度，但是欠缺灵活性。第二种方法优点是灵活性好，能提高运行效率，但是编码复杂。第三种方法则综合了前面二种的优点，既能提高 ETL 的效率又增强了灵活性。

本项目中采用第三种方式，ETL 工具选择 IBM 公司的 DataStage。DataStage 可以从多个不同的业务系统中抽取数据，完成转换和清洗，装载到各种系统里面，并且提供专门的设计工具来设计转换规则和清洗规则等，能实现增量抽取、任务调度等多种复杂功能。

### 2.1 省公司 ETL 重点、难点

省公司 ETL 的重点是抽取所属所有市公司营销系统中的基础数据和业务数据。各个市营销系统中数据的编码、类型和数据结构可能不一致，因此前期的调研和数据分析阶段工作量比较大，工作的质量将直接影响整个 ETL 的效果。通过分析规划需要的数据源及数据定义，整理出待抽取的基础数据表和业务数据表目录。制定抽取的各种策略，尽量降低因数据抽取对市公司营销系统产生的影响及系统风险，尽量不要求市公司营销系统变更其设计。

### 2.2 ETL 实施方案

数据抽取过程主要分为两个阶段，第一阶段，是从所有市公司营销系统中获取所需的原始数据，将它们抽取集成到数据仓库的 ODS（操作数据存储）部分，为数据分析做好准备。为降低对市营销系统运行性能的影响，抽取工作设计在晚上进行，抽取频率为每天晚上抽取一次。

第二阶段则是根据分析主题和数据模型的要求，将 ODS 层的数据抽取到数据仓库层，形成按主题组织的分析型数据结构,数据仓库中主要有三种表类型，分别是：维度表、事实表和聚合表，抽取时应按照维度表、事实表、聚合表的顺序进行。不同的表类型采用不同的抽取策略。

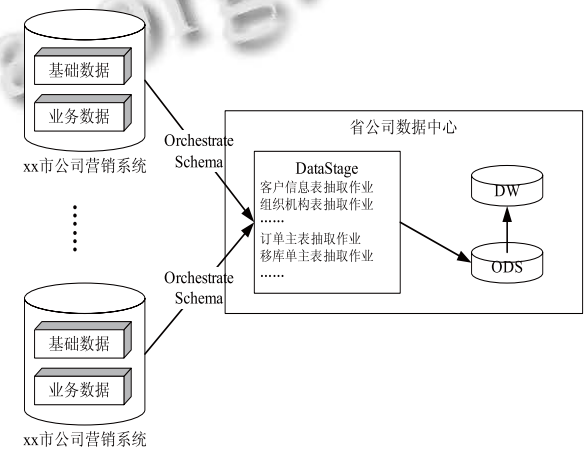


图 1 ETL 实施方案

#### 2.2.1 数据抽取方式

对于基础数据表，考虑其数据量较小，抽取其所有源数据，保存到缓冲区后，对每条数据进行目标表

比对，判断记录的增、删、改，更新情况，并记录更新时间，实现数据的递增加载。

对于业务数据表，由于其数据量大，采用增量抽取的方式，并基于市公司进销存数据判断数据变动的时间。

(1) 抽取市公司一段时间至今的进销存数据，保存在缓冲区中。

(2) 根据主键值，缓冲数据与目标表中的相应记录作字段比较，判断市公司数据发生变动的最早日期 T。

(3) 删除目标表中在日期 T 之后的所有业务数据记录。

(4) 将日期 T 之后的所有业务数据抽取到目标表中，中间可作转换处理。

2.2.2 数据抽取步骤如下：

(1) 设置数据源连接信息，包括地址、访问用户名与密码。

(2) 使用 Orchestrate Schema 导入数据表定义，选择所需的字段。没用的字段可以略去。

(3) 设计 DataStage 作业（Job）。依据设计好的数据抽取方案，设计进行抽取、缓冲、比较、转换、加载等操作的控件，形成作业。设置作业参数，如运行时间，出错处理等。

(4) 编译、运行 ETL 作业，访问数据库，获取数据。

从数据源中抽取数据时，DataStage 可以从表、生成的 SQL SELECT、或用户定义的 SQL 中读取。增量抽取可充分利用 SQL SELECT 方式来抽取所需数据。

| 序号 | 字段名 | 字段描述 |
|---|---|---|
| 1 | DateID | 日期 |
| 2 | CorpID | 公司标识 |
| 3 | ProductID | 卷烟标识 |
| 4 | BuyAmount | 购入数量 |
| 5 | BuyMoney | 购入不含税金额 |
| 6 | SellAmount | 销售数量 |
| 7 | SellMoney | 销售不含税金额 |
| 8 | RestAmount | 期末数量 |
| 9 | RestMoney | 期末金额 |

图 1　T_CORPDAYREPORT

### 2.3 ETL 作业示例

每个要抽取的表对应一个作业。下面以县公司月报生成为例，基于省公司数据库 T_CORPDAYREPORT(县公司进销存日报表)表对最近一个月内县公司日报进行汇总，更新信息到省公司数据库 T_CORPMONTHREPORT(县公司进销存月报表)表。下图为县公司进销存日报表的表结构。

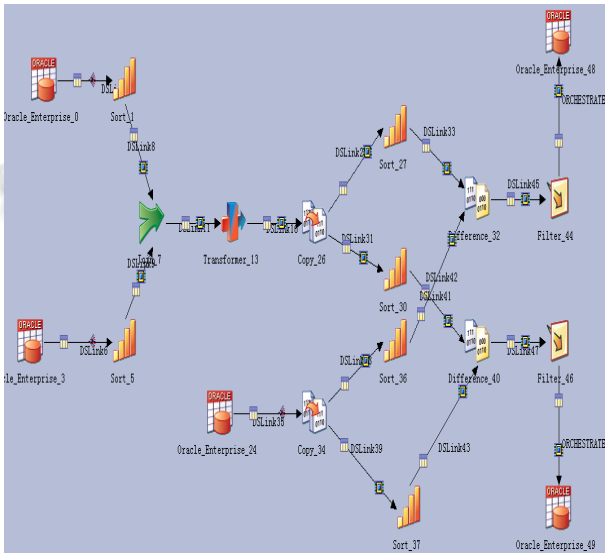


图 2　ETLjob

(1) Oracle_enterprise_0：连接省公司数据库(oracle)，对数据表 T_CORPDAYREPORT 最近一个月内的购入和销售数据按月进行汇总。所用 SQL 语句如下：

```
select
b.yearmonthnumber,a.corpid,a.productid,sum(a.buyamount) as
buyamount,sum(a.buymoney) as buymoney,sum(a.sellamount)
as sellamount,sum(a.sellmoney) as sellmoney from
t_corpdayreport a,t_ datedimesions b where a.dateid=b.dateid
and b.yearmonthnumber>=to_number(
to_char(add_month(sysdate,-1),'yyyymm'))
group by b.yearmonthnumber,a.corpid,a.productid
order by 1,2,3;
```

(2) Oracle_enterprise_3：连接省公司数据库，从数据表 T_CORPDAYREPORT 获取期末数。所用 SQL 语句如下：

```
select n.yearmonthnumber as monthid,m.corpid as
corpid,m.productid,m.restamount from
```

```
t_corpdayreport m,
 (select b.yearmonthnumber,a.corpid,max(b.dateid) as
maxdateid from t_corpdayreport a,
t_ datedimesions b where a.dateid=b.dateid
and b.yearmonthnumber>=to_number(
to_char(add_month(sysdate,-1),'yyyymm'))
group by b.yearmonthnumber,a.corpid
) n
where m.dateid=n.maxdateid and m.corpid=
n.corpid order by 1,2,3;
```

(3) Sort_1、Sort_5、Sort_27、Sort_30、Sort_36、Sort_37：都是排序阶段，依次按三个字段 MONTHID，CORPID，PRODUCTID 升序排列。

(4) Join_7：连接阶段，将从 Oracle_enterprise_0 阶段和 Oracle_enterprise_3 阶段的数据连接起来，主键字段为 MONTHID，CORPID，PRODUCTID。

(5) Transformer_13：数据转换阶段，对于期末数据，如果为 NULL 则转换为 0。

(6) Copy36：复制阶段，复制从 Transformer_13 阶段所得数据。

(7) Oracle_enterprise_24：连接省公司数据库，从数据表 T_CORPMONTHREPORT 获取最近一个月的县公司月报记录。

(8) Copy_121：复制阶段，复制从 Order_enterprise_24 阶段所得数据。

(9) Difference_32：比较阶段，从 Order_enterprise_24 阶段所得数据为旧数据，以 Transformer_13 阶段所得数据为新数据，所有非键字段参与比较。只输出 MONTHID，CORPID，PRODUCTID，DIFF 四个字段。其中 DIFF=0 表示新插入的数据，DIFF=1 表示删除数据，DIFF=2 表示复制数据，DIFF=3 表示修改数据。

(10) Difference_40：比较阶段，从 Transformer_13 阶段所得数据为旧数据，以 Order_enterprise_24 阶段所得数据为新数据，所有非键字段参与比较。输出所有字段并添加 DIFF 字段，DIFF 字段值含义同上。

(11) Filter_44：过滤 Difference_32 阶段所得差异数据，只留下 DIFF=1 即删除了的记录。

(12) Filter_46：过滤 Difference_40 阶段所得差异数据，只留下 DIFF=1 或 3，即删除或修改过的记录。

(13) Oracle_enterprise_48：连接省公司数据库，在 T_CORPMONTHREPORT 表中删除与 Filter_44 输出记录的键值相同的记录。

(14) Oracle_enterprise_49：连接省公司数据库，在 T_CORPMONTHREPORT 表中插入 Filter_46 输出的记录，并更新表中与 Filter_46 输出记录的 MONTHID，CORPID，PRODUCTID 键值相同的记录。

## 3 结论

ETL 技术的成功应用，降低了项目因为数据源异构，数据结构不一致等原因带来的难度，提高了数据的质量，是省烟草公司数据中心项目成功的关键。随着商务智能技术的兴起发展，ETL 技术必将受到越来越多的关注。

## 参考文献

1 姚家奕.数据仓库与数据挖掘技术原理与应用.北京:电子工业出版社, 2009.66−73.

2 陈志泊.数据仓库与数据挖掘.北京:清华大学出版社, 2009.30−36.

3 王丽珍,周丽华,陈红梅,肖清.数据仓库与数据挖掘原理及应用.第 2 版.北京:科学出版社, 2009.21−37.