# BI项目中ETL设计与思考

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，目的是将企业中的分散、零乱、标准不统一的数据整合到一起，为企业的决策提供分析依据。ETL是[BI](http://mie168.com/BIS/index.htm)项目重要的一个环节。通常情况下，在BI项目中ETL会花掉整个项目的1/3的时间,ETL设计的好坏直接关接到BI项目的成败。

ETL的设计分三部分：数据抽取、数据的清洗转换、数据的加载。在设计ETL的时候我们也是从这三部分出发。数据的抽取是从各个不同的数据源抽取到ODS(OperationalDataStore，操作型[数据存储](http://mie168.com/DSM/index.htm))中——这个过程也可以做一些数据的清洗和转换)，在抽取的过程中需要挑选不同的抽取方法，尽可能的提高ETL的运行效率。ETL三个部分中，花费时间最长的是“T”(Transform，清洗、转换)的部分，一般情况下这部分工作量是整个ETL的2/3。数据的加载一般在数据清洗完了之后直接写入DW(DataWarehousing，数据仓库)中去。

　　ETL的实现有多种方法，常用的有三种。一种是借助ETL工具(如Oracle的OWB、SQLServer2000的DTS、SQLServer2005的SSIS服务、Informatic等)实现，一种是SQL方式实现，另外一种是ETL工具和SQL相结合。前两种方法各有各的优缺点，借助工具可以快速的建立起ETL工程，屏蔽了复杂的编码任务，提高了速度，降低了难度，但是缺少灵活性。SQL的方法优点是灵活，提高ETL运行效率，但是编码复杂，对技术要求比较高。第三种是综合了前面二种的优点，会极大地提高ETL的开发速度和效率。

**一、数据的抽取**

　　这一部分需要在调研阶段做大量的工作，首先要搞清楚数据是从几个业务系统中来,各个业务系统的数据库服务器运行什么DBMS,是否存在手工数据，手工数据量有多大，是否存在非结构化的数据等等，当收集完这些信息之后才可以进行数据抽取的设计。

　　1、对于与存放DW的数据库系统相同的数据源处理方法

　　这一类数据源在设计上比较容易。一般情况下，DBMS(SQLServer、Oracle)都会提供数据库链接功能，在DW数据库服务器和原业务系统之间建立直接的链接关系就可以写Select语句直接访问。

　　2、对于与DW数据库系统不同的数据源的处理方法

　　对于这一类数据源，一般情况下也可以通过ODBC的方式建立数据库链接——如SQLServer和Oracle之间。如果不能建立数据库链接，可以有两种方式完成，一种是通过工具将源数据导出成.txt或者是.xls文件，然后再将这些源系统文件导入到ODS中。另外一种方法是通过程序接口来完成。

　　3、对于文件类型数据源(.txt,.xls)，可以[培训](http://mie168.com/human-resource/peixunfazhan.htm)业务人员利用数据库工具将这些数据导入到指定的数据库，然后从指定的数据库中抽取。或者还可以借助工具实现，如SQLServer2005的SSIS服务的平面数据源和平面[目标](http://mie168.com/manage/zuzhi-mubiao.htm)等组件导入ODS中去。

　　4、增量更新的问题

　　对于数据量大的系统，必须考虑增量抽取。一般情况下，业务系统会记录业务发生的时间，我们可以用来做增量的标志,每次抽取之前首先判断ODS中记录最大的时间，然后根据这个时间去业务系统取大于这个时间所有的记录。利用业务系统的时间戳，一般情况下，业务系统没有或者部分有时间戳。

**二、数据的清洗转换**

　　一般情况下，数据仓库分为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，在从ODS到DW的过程中转换，进行一些业务规则的计算和聚合。

    1、数据清洗

　　数据清洗的任务是过滤那些不符合要求的数据，将过滤的结果交给业务主管部门，确认是否过滤掉还是由业务单位修正之后再进行抽取。不符合要求的数据主要是有不完整的数据、错误的数据、重复的数据三大类。

　　(1)不完整的数据：这一类数据主要是一些应该有的信息缺失，如供应商的名称、分公司的名称、客户的区域信息缺失、业务系统中主表与明细表不能匹配等。对于这一类数据过滤出来，按缺失的内容分别写入不同Excel文件向客户提交，要求在规定的时间内补全。补全后才写入数据仓库。

　　(2)错误的数据：这一类错误产生的原因是业务系统不够健全，在接收输入后没有进行判断直接写入后台数据库造成的，比如数值数据输成全角数字字符、字符串数据后面有一个回车操作、日期格式不正确、日期越界等。这一类数据也要分类，对于类似于全角字符、数据前后有不可见字符的问题，只能通过写SQL语句的方式找出来，然后要求客户在业务系统修正之后抽取。日期格式不正确的或者是日期越界的这一类错误会导致ETL运行失败，这一类错误需要去业务系统数据库用SQL的方式挑出来，交给业务主管部门要求限期修正，修正之后再抽取。

　　(3)重复的数据：对于这一类数据——特别是维表中会出现这种情况——将重复数据记录的所有字段导出来，让客户确认并整理。

　　数据清洗是一个反复的过程，不可能在几天内完成，只有不断的发现问题，解决问题。对于是否过滤，是否修正一般要求客户确认，对于过滤掉的数据，写入Excel文件或者将过滤数据写入数据表，在ETL开发的初期可以每天向业务单位发送过滤数据的邮件，促使他们尽快地修正错误,同时也可以做为将来验证数据的依据。数据清洗需要注意的是不要将有用的数据过滤掉，对于每个过滤规则认真进行验证，并要用户确认。

2、数据转换

　　数据转换的任务主要进行不一致的数据转换、数据粒度的转换，以及一些商务规则的计算。

　　(1)不一致数据转换：这个过程是一个整合的过程，将不同业务系统的相同类型的数据统一，比如同一个供应商在结算系统的编码是XX0001,而在[CRM](http://mie168.com/CRM/index.htm)中编码是YY0001，这样在抽取过来之后统一转换成一个编码。

　　(2)数据粒度的转换：业务系统一般存储非常明细的数据，而数据仓库中数据是用来分析的，不需要非常明细的数据。一般情况下，会将业务系统数据按照数据仓库粒度进行聚合。

　　(3)商务规则的计算：不同的企业有不同的业务规则、不同的数据指标，这些指标有的时候不是简单的加加减减就能完成，这个时候需要在ETL中将这些数据指标计算好了之后存储在数据仓库中，以供分析使用。

**三、ETL日志、警告发送**

　　1、ETL日志

　　ETL日志分为三类。一类是执行过程日志，这一部分日志是在ETL执行过程中每执行一步的记录，记录每次运行每一步骤的起始时间，影响了多少行数据，流水账形式。一类是错误日志，当某个模块出错的时候写错误日志，记录每次出错的时间、出错的模块以及出错的信息等。第三类日志是总体日志，只记录ETL开始时间、结束时间是否成功信息。如果使用ETL工具,ETL工具会自动产生一些日志，这一类日志也可以作为ETL日志的一部分。记录日志的目的是随时可以知道ETL运行情况，如果出错了，可以知道哪里出错。

　　2、警告发送

　　如果ETL出错了，不仅要形成ETL出错日志，而且要向系统管理员发送警告。发送警告的方式多种，一般常用的就是给系统管理员发送邮件，并附上出错的信息，方便管理员排查错误。

　　ETL是BI项目的关键部分，也是一个长期的过程，只有不断的发现问题并解决问题，才能使ETL运行效率更高，为BI项目后期开发提供准确的数据。