建立一个成功的数据仓库系统，依靠的是最佳实践而不是直觉。

三个简单的字母，E-T-L，很容易的让大家忽视了38个ETL子系统在数据仓库建设中的重要性。

抽取-转换-加载（ETL）系统，或者非正式的称为“后台系统”，在建立整个数据仓库系统中占据了70的工作量和时间。但是这还不足以说明ETL系统的复杂性。每个人都理解这三个字母的含义，E，从源系统中将数据取出来；T，对这些数据做处理；L，加载到最终用户访问的表中。

但是当我们问及如何来分解这三大步骤时，很多设计人员都会说，“具体问题，具体分析”。例如，这依赖于不同的数据源；这依赖于数据的特性；这依赖于脚本语言以及可以使用的ETL工具的情况；这依赖于员工的技术能力；这还依赖于最终用户使用的查询和报表工具。

“具体情况，具体分析”是一个很危险的事情，因为它很容易称为系统混乱的一个借口。伴随着几千个成功数据仓库项目的经历，我们整理出了一系列的最佳实践。

最近的18个月，我们一直在钻研ETL的实践和ETL的产品。我们标识出了在每一个数据仓库项目的后台部分都会涉及到的38个子系统。坏消息是ETL系统确实占据了数据仓库项目的大部分资源。好消息是如果你能掌握所有的这些子系统，你就可以很容易的使用你的经历来建立成功的数据仓库系统。

1.抽取系统（Extract System）

主要功能包括源数据的适配器，推/拖/搬运数据的工作调度，对源数据的过滤和排序功能，数据格式的转换，迁移到ETL环境后的数据暂存功能。

2.变化数据捕获系统（Change Data Capture System）

主要功能包括对源数据日志文件的阅读功能，源数据日期和序列号的过滤功能，基于CRC算法的记录比较功能。

3.数据概况分析系统（Data Profiling System）

主要功能包括字段属性分析，如参照域的分析；结构分析，如主外键关系分析；数据规则分析；值规则分析等。

4.数据清洗系统（Data Cleansing System）

主要功能包括一个典型的数据字典驱动的系统，用于解析个体和组织的名称、地址等信息，也用来解析产品、场所等内容；一个“De-duplication”系统，用于鉴别和移除个体和组织信息，也用于产品和场所；一个“Surviving”系统，使用特定的数据合并逻辑，用来保存特定数据源的指定字段，这个特定数据源的数据将成为数据仓库的最终版本；为所有的数据源维护后台数据的对应关系，如自然键和代理键对应关系等内容。

5.数据一致性处理系统（Data Conformer System）

主要功能包括标识和生成专用的一致性维度属性、一致性事实的度量属性，这两组属性作为数据整合工作的基础，用来支持跨多个数据源的数据集成工作。

6.审计维度生成系统（Audit Dimension Assembler System）

主要功能是将与事实表相关的元数据内容加载到一张审计维度表中，这样最终用户可以像查看普通维度一样查看与事实表相关的元数据。

7.数据质量过滤系统（Quality Screen Handler System）

主要功能是在ETL的处理过程中自动的检测所有的数据质量问题。检测的结果将进入错误事件处理系统（详见子系统8）。

8.错误事件处理系统（Error Event Hander System）

主要功能是全面的记录和报告在ETL处理中的所有的错误事件。包括各类错误的分枝处理逻辑，还包括对ETL处理中数据质量的实时监控。

9.代理键生成系统（Surrogate Key Create System）

主要功能是以一种鲁棒的机制生成流水的代理键，生成规则不依赖与任何维度，也不依赖与任何数据库实例，可以支持分布式系统。

10.缓慢变化维处理系统（Slowly Changing Dimension Processor，SCD）

主要功能是处理维度表的属性随时间变化的情况，处理方式为：类型1（直接覆盖），类型2（生成新行），类型3（添加新列）。

11．迟到维度处理系统（Late Arriving Dimension Handler）

主要功能是当维度数据的变化情况到达数据准备区的时间晚于对应的事实数据时，对维度数据的插入和更新策略。

12．固定层级结构生成系统（Fixed Hierarchy Dimension Builder）

主要功能是对维度表中各类多对一关系的层级结构进行数据有效性检查和维护。

13．可变层级结构生成系统（Variable Hierarchy Dimension Builder）

主要功能是对维度表中所有的层深可变的层级结构的的数据有效性检查和维度，例如组织的层级结构，零件的层级结构等。

14．多值维度桥接表生成系统（Multivalued Dimension Bridge Table Builder）

主要功能是建立和维护桥接表，用来描述维度间的多对多关系。

15．杂项维度生成系统（Junk Dimension Builder）

主要功能是将来自多个数据源的多个低基数的标志字段、状态字段等小型维度建立成一个杂项维度，并对之进行维护。

16. 交易粒度事实表加载系统（Transaction grain fact table loader）

主要功能是更新交易粒度事实表，包括对数据、索引和分区的处理。通常是用来处理增量数据，即最新的数据。需要使用代理键替换管道系统（详见子系统19）。

17. 周期快照事实表加载系统（Periodic snapshot grain fact table loader）

主要功能是更新周期快照事实表，包括对数据、索引和分区的处理。包括对当期数据的增量更新策略。需要使用代理键替换管道系统（详见子系统19）。

18. 累计快照事实表加载系统（Accumulating snapshot grain fact table loader）

主要功能是更新累积快照事实表，包括对数据、索引和分区的处理，同时更新维度外键和累积事实。需要使用代理键替换管道系统（详见子系统19）。

19. 代理键替换管道系统（Surrogate key pipeline）

主要功能是使用多线程技术将来到数据仓库数据的自然键替换为代理键。

20. 迟到事实处理系统（Late arriving fact handler）

主要功能是处理对迟到事实记录的插入和更新策略。