Waha!Transformer技术白皮书

备实必（上海）软件科技有限公司

前 言

Waha!Transformer（简称：Waha，读音：哇哈）是一款专业的ETL工具，有着处理速度快、易操作、数据连接源种类多等特点，可以进行各种形式的数据清洗、挖掘等处理，功能强大。本白皮书是为了从技术上说明并指导用户进一步了解Waha，本书涉及到的Waha技术方面的内容及相应数据都是实际操作实验产生。

今后，如果遇到有软件版本的更新，请以最新的版本的白皮书为主。如果有Waha方面（不论产品本身还是应用技术层面）的问题，我们都建议用户与我们取得联系以获取帮助。

联系方式：

备实必（上海）软件科技有限公司

地址：上海市浦东新区商城路738号胜康廖氏大厦2107室

电话：+86 021- 5058-8268

电子邮箱: support@bsp-sh.cn

网站主页: <http://www.bsp-sh.cn>

 2016年8月 编

法律责任声明

本书是针对Waha 5.0的版本作技术说明，对于5.0以下的版本有着一定的阅读性，但并不推荐5.0以下的版本的用户阅读使用。但是，Waha的用户（无论正在使用哪个版本的Waha）如果遇到Waha的技术问题请联系我们。

本手册以“我们”这一人称代词代指备实必（上海）软件科技有限公司，以“用户”一词代指所有有着Waha合法使用权的操作人员或公司。

Waha!Transformer软件的一切权益以及本书的最终解释权都归备实必（上海）软件科技有限公司所有。

目 录

[第一章 概述 1](#_Toc14125)

[一、公司介绍 1](#_Toc29806)

[二、BSP在中国 1](#_Toc5262)

[第二章 什么是ETL？ 1](#_Toc21493)

[一、引言 1](#_Toc29)

[二、术语描述 1](#_Toc8064)

[三、ETL是数据仓库建立的核心过程 1](#_Toc27891)

[四、ETL过程的主要目的 1](#_Toc31799)

[五、ETL体系结构简介 1](#_Toc14404)

[第三章 为什么要使用ETL工具 1](#_Toc5247)

[一、数据现状 1](#_Toc9437)

[二、解决问题的必要性 1](#_Toc16897)

[三、ETL工具VS手工编码 1](#_Toc1932)

[第四章 Waha介绍 1](#_Toc2298)

[一、Waha背景 1](#_Toc21426)

[二、Waha的客户群 1](#_Toc30453)

[第五章 Waha基本概况 1](#_Toc29079)

[一、Waha的模块 1](#_Toc18524)

[二、 可以使用的情形 1](#_Toc21564)

[第六章 Waha功能 1](#_Toc4993)

[一、功能一览 1](#_Toc12978)

[二、Waha的使用环境 1](#_Toc27923)

[三、对数据源的支持 1](#_Toc10565)

[四、数据转换功能 1](#_Toc22812)

[五、管理和调度功能 1](#_Toc26884)

[六、集成和开放性 1](#_Toc13554)

[七、对源数据的管理 1](#_Toc4196)

[八、 和其他ETL的对比 1](#_Toc27084)

[九、性能参数 1](#_Toc24075)

[第七章 Waha用户 1](#_Toc29908)

[第八章 总结 1](#_Toc14102)

# 第一章 概述

## 一、公司介绍

备实必（上海）软件科技有限公司（英文名：BSP，中文简称：备实必）于1976年在日本东京成立，30多年来一直致力于大型机上的数据库以及数据仓库、中间件的研发、销售和技术支持。作为全球领先的企业数据集成解决方案提供商，BSP专注于数据集成软件和服务，以解决系统间的数据碎片问题，从而帮助企业获得了数据的更大商业价值。

自主ETL产品Waha!Transformer软件（简称为Waha）的使用，降低了企业IT的运营成本、加快了产出速度，可伸缩处理不同规模或不同复杂程度的数据集成项目。

Waha拥有遍布全球的客户关系网络，其出色的成功案例显示，它将帮助更多的企业和政府机构实现其企业数据的商业价值。在全球，现在有1400多家各个行业的客户都将Waha作为主要的数据处理使用，这也证明了Waha是一款功能强大，性能优秀的软件工具。

## 二、BSP在中国

BSP把中国列为其全球投资的最重要国家之一，2007年正式进军中国，凭借其先进的技术和完善的服务，在短短的时间中，很快就发展了包括金融、电信、制造、政府、保险、公众服务等多个领域的众多客户，帮助更多企业架构随时随地呈现正确而重要信息的数据整合平台。

BSP着眼于中国市场的开拓，把自身的技术和经验分享给国内的企业和合作伙伴，用户通过使用Waha，能帮助其加强信息化管理，最终也能为推动中国的信息化建设和经济建设做出贡献。

# 第二章 什么是ETL？

## 一、引言

在当下的高度信息化的社会，一个企业信息化程度的高低，反映着这个企业的生产力水平和管理水平，也决定着这个企业新世纪生存与发展的实力和地位。信息是现代企业的宝贵资源，是企业运用科学管理、决策分析的基础。目前，大多数企业花费大量的资金和时间来构建联机事务处理(OLTP)的业务系统和办公自动化系统，从而记录事务处理的业务数据。遗憾的是，业务系统产生的数据增长很快，一般情况下，两到三年的时间，数据量就会成倍增长，这些丰富的、蕴含巨大商业价值的数据，所关注的通常只占在总数据量的2%-4%。从而企业将会失去有价值的客户、失去提高客服质量的指引、无法制定出最佳的营销手段和策略、等等。无论何种性质的企业，没有最大化的利用已存在的数据资源，都将浪费更多的时间和资金，也失去制定关键商业决策的最佳契机。

商业智能(BI,Business Intelligence) 就是为企业把数据转换为信息、知识相应蕴育而出的IT技术。企业级BI的基础就是一个完整的、准确的、统一视角的数据平台，即DW(Data warehousing)，而本文则要详细论述构建DW关键过程ETL的简介及必要性。

## 二、术语描述

ETL：数据仓库技术，是英文单词Extract-Transform-Load的缩写，E是数据抽取（Extract）、T是转换（Transform）、L是装载（Load）的过程。

ETL是构建数据仓库的重要一环，用户从数据源抽取出所需的数据，然后进行数据清洗，并将无效的，冗余的数据进行筛选，所得的数据最后再加载到数据仓库中，该技术也正是Waha的核心所在。

## 三、ETL是数据仓库建立的核心过程

数据仓库系统是在业务系统的基础上发展而来的，其内部存储的数据来自于事务处理的业务系统和外部数据源。而企业内各源数据缺少统一的标准，因企业的业务系统是在不同时期、不同背景、面对不同应用、不同开发商等各种客观前提下建立的，其数据结构、存储平台、系统平台均存在很大的异构性。因而其数据难以转化为有用的信息，原始数据的不一致性导致决策时其可信度的降低。

ETL是BI/DW的核心和灵魂，按照统一的规则集成并提高数据的价值，是负责完成数据从数据源向目标数据仓库转化的过程，是实施数据仓库的重要步骤。在整个项目中最难部分是用户需求分析和模型设计，而ETL规则设计和实施则是工作量最大的，其工作量要占整个项目的60%-80%，这是国内外从众多实践中得到的普遍共识。

## 四、ETL过程的主要目的

就是以最小代价(包括对日常操作的影响和对技能的要求) 将针对日常业务操作的数据转化为针对数据仓库而存储的决策支持型数据。 同时这个过程要快，不少的企业数据量已经达到PB（1PB=1024TB，1TB=1024GB）的级别，这样的数据量在进行处理的时候往往需要相当长的时间，这对于其所用的ETL工具也是一个不小的挑战，相比于其他ETL工具来说，Waha的性能高，处理速度快，更能够胜任数据量大的数据处理。

## 五、ETL体系结构简介

1. 数据抽取：从源数据源系统抽取目的数据源系统需要的数据;
2. 数据转换：将从源数据源获取的数据按照业务需求，转换成目的数据源要求的形式，并对错误、不一致的数据进行清洗和加工。

(3)数据加载：将转换后的数据装载到目的数据源。

ETL原本是作为构建数据仓库的一个环节，负责将分布的、异构数据源中的数据如关系数据、平面数据文件等抽取到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集市中，成为联机分析处理、数据挖掘的基础。现在也越来越多地将ETL应用于一般信息系统中数据的迁移、交换和同步。一个简单的ETL体系结构如图2-1所示。



**临时数据**

DB

图2-1.简单的ETL体系

Waha所做的正是将这种专业的数据处理体系结构变得简单化，使得非专业开发人员在经过简单的教导之后也能够迅速上手，这也是很多知名企业都选择Waha的原因。

# 第三章 为什么要使用ETL工具

## 一、数据现状

在如今的企业中，企业的数据库系统由于存在着大量的脏数据，已有数据库中的数据也存在着不少的问题，比如有空值的数据，有着与表格类型不同的数据，这些数据都严重影响了工作中的数据处理，数据质量问题的原因有很多，由系统集成和历史数据造成的原因主要有以下一些：

滥用缩写词、惯用语、数据输入错误、重复记录、丢失值、拼写变化、不同的计量单位和过时的编码等；

业务系统不同时期系统之间数据模型不一致；

业务系统不同时期业务过程有变化；

旧系统模块在运营、人事、财务、办公系统等相关信息的不一致；

遗留系统和新业务、管理系统数据集成不完备带来的不一致性；

除了这些原因之外，还有一些其他的原因：

源系统缺少输入验证过程，不能阻止非法格式的数据进入系统；

可以验证但不能改正数据，验证程序不能发现格式正确但内容不正确的错误；

源系统不受控制的更改，而这种更改不能及时的传播到受影响的系统；

数据有多个交叉的访问界面，难以统一管理数据质量问题；

缺少参照完整性检查；

低劣的源系统设计；

数据转换错误，数据迁移过程造成的错误；

## 二、解决问题的必要性

即便是一个设计和规划良好的数据仓库系统，如果其中存在着大量的脏数据，那么这个系统也是没有任何意义的，因为“垃圾进，垃圾出”（garbage in, garbage out），系统根本就不可能为决策分析系统提供任何支持。为了清除脏数据，必须在数据仓库系统中进行数据清洗。或者需要对各个时期实施的系统进行数据迁移，各系统间数据传输、数据共享、备份等。这些问题都需要ETL技术去解决。

## 三、ETL工具VS手工编码

那么在数据处理的方法上，到底用ETL工具好还是手工编码实现ETL过程好？无论是在网络开发者的论坛还是企业的选择上，这个话题的争论可以说是永无休止，很显然这两种方式都有各自的优势和劣势，下面从几方面对两种方式作一个比较 。

|  |  |  |
| --- | --- | --- |
| 比较项目 | ETL工具 | 手工编程 |
| 设计界面 | 图形化 | 有赖于开发者的风格 |
| 数据库接口 | 自备各种数据库接口 | 依赖于重新开发 |
| 对应数据库 | 对应各种数据库 | 不对应，重新开发 |
| 元数据管理 | 提供元数据报告 | 不能提供元数据报告 |
| 灵活性 | 修改方便，灵活 | 不灵活 |
| 难易程度 | 相对容易 | 要求一定技术水平 |
| 管理和维护 | 容易 | 较难 |
| 性能和效率 | 数据调试、预览，触发调动执行 | 取决于编写者的水平 |
| 开发周期 | 很短 | 较长 |
| 工作量 | 较小 | 较大 |
| 业务逻辑变化 | 容易 | 较难 |
| 项目实施风险 | 很小 | 有一定风险 |
| 实施后的稳定性 | 稳定 | 1-2年后才能稳定 |

就目前各行业使用ETL的情况和企业发展的远景来看，选择并使用一款成熟的ETL工具制作各种流程处理数据，能使得效率更高、修改更方便、更能节省人力成本，这也是Waha的优势和特点所在。

##

# 第四章 Waha介绍

## 一、Waha背景

Waha!Transformer中的Waha一次取名于*Wa*re*h*ouse *A*utomation，至于Transformer一词则是实际的中文义，即“转换”，它的产生是由于数据仓库的需求浪潮的时代来临，作为一个数据仓库和其他数据系统连接的工具而开发完成的。

在20世纪八十年代，在BSP的很多使用大型机的客户当中，数据小型化的要求很多，如何把复杂的大型机环境下的数据抽取并投入到新的数据库中，这是一个很难的技术过程；而数据仓库的出现，需要把企业现有的很多系统中的数据投放到数据仓库里进行挖掘分析，等等需求。

鉴于以上几点，BSP应客户的要求，开发了ETL产品，提供给客户，大大方便了客户的数据利用。BSP自199年开始研发了Waha之后，到现在2016年已经开发完成了5.1的版本，这其中不断的推陈出新，增加了许多实用性高效性的功能，无论是读取速度，还是操作性，亦或是功能性等，都是行业领先水平。

|  |
| --- |
| Waha开发时间一览 |
| 1998年12月 | V1.0完成 |
| 2002年9月 | V2.0完成 |
| 2006年10月 | V2.2完成（从这个版本开始有了中、英、日三种操作界面） |
| 2009年9月 | V3.0完成 |
| 2012年5月 | V4.0完成 |
| 2015年3月 | V5.0完成 |

## 二、Waha的客户群

在出现商业化的ETL工具之前，很多客户曾经尝试着自己开发ETL软件来完成这一过程，这需要开发者熟悉各种平台，各种编程语言，各种数据源和目标数据库，而且数据格式统一等过程都很复杂，即使制作完成，经常性的维护和修改等都是很耗时耗力耗费的工程。

Waha从1998年开始研发与销售，是BSP结合三十多年的数据库和数据仓库研发、实施的经验，为解决数据整合的需求而开发研制的。至今，在以日本为主的国际市场上，有1400多家客户在不同的行业里使用，为客户的经济活动提供了很好的数据处理帮助。

作为全球领先的企业数据集成解决方案提供商，BSP一直致力于通过自身的产品和服务，提升企业的竞争优势。Waha拥有遍布全球的客户关系网络，客户范围涵盖银行、证券、保险、制造业、电信、消费品、能源、烟草和政府等行业。

#

# 第五章 Waha基本概况

一、Waha的模块

目前的5.1版本的Waha有着以下的几个模块，如图5-1所示。



图5-1.Waha模块

Designer：作业设计器，设计流程的图形化的ETL设计工具，也是Waha的主要模块；

Document Generator：文档生成器，把制作成的各种流程的设定以HTML形式记录下来，自动生成技术文档，对元数据进行管理；

Job Monitor：作业监视器，显示登录在各个作业库里流程，可以阅览并监视各流程的执行结果，时间调度；

Log View：日志查看器，可以看到流程执行之后的各种日志；

Mange Log Viewer：日志管理器，对于日志可以进行管理；

Online hlep：在线帮助，可以获得一个在线的帮助；

Repository Maintenance Tool：作业库管理工具对保存用户信息、作业流程信息的作业库进行管理；

System Setting Tool：系统设置工具，可以对Waha进行一些基本的设置；

Waha的服务器主要负责将业务系统数据库中的数据进行清洁，计算，标准化，并将处理好的数据装载到数据仓库/集市中。利用Designer，用户可以可视化地设计业务规则，这些规则将由Designer自动存储到Repository 中，并且由服务器来执行和处理这些规则。Designer可以设计和执行调度流程，Repository Maintenance Tool 用来管理用户信息、作业流程信息以及Repository 中的所有单元。Designer中作业的原理构造简单，一个简单的流程如下图5-2所示



图5-2.Waha作业的流程图

Waha的操作简单易懂，学习成本低，使用图形化界面处理，没有编程过程，这也使得非开发人员可以操作和使用，仅仅使用鼠标拖拉就可以实现流程设计，一个简单的流程设计如图5-3所示。



图5-3.Waha作业流程图

数据预览：在ETL设计的过程中，使用者对每个流程都可以进行预览，这样有利于数据的确认和设计的进行，如图5-4所示



图5-4.视表预显

流程监控：通过Job Monitor，可以监控各个流程的执行情况，及时发现异常情况。



图5-5.作业监视工具

查询日志：能够查询作业执行的情况，如下图5-6所示



图5-6.日志浏览工具

## 可以使用的情形

Waha能够胜任各种各样的场合，不单单为用户提供了数据清洗的功能，甚至在于其他的地方，Waha也能够无缝连接和使用。

Waha在以下的场合中同样能够适用，如图5-7所示：

1. 数据迁移：用户企业的服务器之间进行数据迁移的时候，可以使用Waha来完成，并且Waha能够进行批处理，十分的快捷与方便。
2. 数据连接：对于不同数据库的数据或是不同文件夹中的数据进行相互连接。
3. 主数据管理：用户进行主数据管理的时候，Waha能够很好的用到。
4. 数据整合：对于不同类型的数据，比如数据库数据或是本地文件中的数据进行整合使得能够得到单一类型的数据以输出。
5. 数据同步：目前来说云服务器开始普及，越来越多的用户开始将自己的数据备份至云服务器端，在这样的数据同步的过程上，Waha也有着相当程度的适用性。



图5-7.Waha适用的场合

#

# 第六章 Waha功能

## 一、功能一览

Waha的功能强大，能够胜任所有的主流的数据处理操作，从最基本的“增删改查”到高级的批处理，批运行的工作，Waha都能运行，甚至，从一些开源的数据处理软件的角度来看，Waha还能做一些这些开源软件无法处理的工作，比如数据类型以及字码的转换，还有高级功能批处理的操作等等，这都是Waha的强项。这里列出Waha的功能：

数据增量，数据拆分，数据清洗；

数据类型转换，字码转换，排序，匹配，参照，差分抽取，统计，结合，分割，调度；

处理、后处理、异常处理时的记录；

不需要编码制作作业流程，作业导入导出，多线条流程高速处理，并行数据加工；

定时刷新，作业监控，日志，自动文档生成；

多平台数据库系统，支持FTP，和其它软件接口；等等。

还有一些其他功能如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
| MF | AS400 | UNIX | NT |
| 字码转换功能 | EBCDIC假名(IBM931) | ○ | ○ |  |  |
| EBCDIC英英小文字(IBM930)  | ○ | ○ |  |  |
| EBCDIK（日立） | ○ |  |  |  |
| ASCII |  |  | ○ | ○ |
| IBM汉字 | ○ | ○ |  |  |
| 富士通JEF | ○ |  |  |  |
| 日立KEIS | ○ |  |  |  |
| ShfitJIS |  |  | ○ | ○ |
| 日语EUC |  |  | ○ |  |
| 对应汉字、非汉字、假名的混合文字列 | ○ | ○ |  |  |
| EBCDIC GB HOST(IBM836) | ○ | ○ |  |  |
| GB18030 |  |  | ○ | ○ |
| BIG5(CP950) |  |  | ○ | ○ |
| Unicode Big Endian |  |  | ○ | ○ |
| Unicode Little Endian |  |  | ○ | ○ |
| UTF-8 |  |  | ○ | ○ |
| 处理数据类型 | Pack | ○ | ○ |  |  |
| Unpack | ○ | ○ |  |  |
| 整数型 | ○ | ○ | ○ | ○ |
| 文字型 | ○ | ○ | ○ | ○ |
| 浮動小数点 | ○ | ○ |  |  |
| 二进制 | ○ | ○ |  |  |
| 处理文件类型 | 固定长文件（ＴＸＴ） | ○ | ○ | ○ | ○ |
| 可变长文件（ＣＳＶ） | ○ | ○ | ○ | ○ |
| 二进制文件 | ○ | ○ | ○ | ○ |
| 用FTP直接连接 | ○ | ○ | ○ | ○ |

## 二、Waha的使用环境

Waha操作平台：Windows操作系统的各个版本（Win XP/7/8/8.1/10等）、Windows server的各个版本（2008/2012等）、Linux、Unix等；

数据源的硬件平台：PC，Unix、AS/400，大型机（Mainframe）；

数据库的软件平台：Oracle、SQL server、MySQL等主流数据库软件，一些专业数据软件也能连接，比如Hyperion、Sybase等；

其他：支持FTP方式数据读取数据源，支持区域网；

## 三、对数据源的支持

Waha可以对各种平台下的数据进行处理，它的数据连接通过各种接口实现。针对主流数据库有专用接口，对一般的数据库通过ODBC连接，文本文件也可以读取，并在此基础上实现各种后续设计。数据库接口Loader有对主流关系型数据库的直接访问的功能，主流关系型数据库包括Oracle，IBM DB2， Microsoft SQL Server，Essbase，Sybase，Excel。此外，也可以采用ODBC，OLE-DB 等通用接口抽取或加载各种关系型数据库或多维数据库中的数据。

文件型数据可以直接抽取本地的平面文件，XML 文件中的数据，也可以通过FTP，HTTP 等方式抽取远程的数据。在抽取平面文件时，可以采用多线程并行抽取的方式来提高抽取效率。支持将转换处理好的数据加载到分隔符文本文件，定长文本文件以及XML 文件中。对应的数据接口如下图6-1所示。

支持多种数据库：

专用接口连接：Oracle，Essbase，Sybase，Excel，CSV，html等。

ODBC连接：DB2，MySQL，PostgreSQL，FoxPro，Foxbase，Dbase，Informix，TXT等

JDBC连接：SQL server等



图6-1.Waha所能对应的数据接口

## 四、数据转换功能

 Waha对数据进行的转换和计算，是通过使用九个视表过滤器（简称为过滤器）连接数据源视表和目标视表而实现的。如图所示，将过滤器放在两个视表中间，在过滤器内设定数据处理的条件，然后手工或自动执行，就可以完成这一流程，如图6-2所示。



图6-2.一个简单的作业流程

以下就是九个过滤器一览表，在新建视表过滤器时，该就个过滤器会自动跳出供选择，如图6-3所示。



图6-3.新建过滤器示意图

标准过滤器：具有表格栏移动、计算、整形、排序、记录的压缩等基本功能。其它的过滤器也具有标准过滤器所具有的基本功能，如图6-4所示。



图6-4.标准视表过滤器

结合过滤器：相当于SQL的结合。指定数个键值，将数个输入数据结合起来。不仅是内部连接，也可以外部连接，如下图6-5所示。



图6-5.结合视表过滤器

归并过滤器：可以将不同视表中的数据合并到一个目标数据表中。除了单纯的数据统合外，在指定主键出现同一数据时，在输入数据的优先顺序和更新日等表格栏中、可以选择要采用的记录，如图6-6所示。



图6-6.归并视表过滤器

统计过滤器：设定多个集团化，可以进行合计、平均、最大值、最小值、件数的统计。也可以用于一些独特的数据抽取。统计键可以按照年龄等一定的间隔和存款额的指定，以一定的幅度进行增量计算。



图6-7.统计视表过滤器

分割过滤器：指定数个分割条件，将记录分割成数个文件。以部门编码划分输入的数据，在销售额的范围内进行赛选，如下图6-8所示。



图6-8.分割视表过滤器

参照过滤器：将原视表中的数据对照参照表中的数据。比如，以地区代码为基准，从地区总表中参照符合条件的数据，可以取得或置换。可以只选择符合参照数据的输入数据，如下图6-9所示。



图6-9.参照视表过滤器

纵展开过滤器：把横向显示的数据转换为纵向显示。可以用于分布式数据库的保存格式向RDB正规化数据展开时使用。



图6-10.纵向展开过滤器

更新SQL过滤器：根据删除视表的内容将数个记录删除，根据输入视表的主键值以字段为单位对数据库进行更新等等，可以简单地实现对复杂数据库的更新控制，如下图6-11所示。



图6-11.更新SQL视表过滤器

作业执行过滤器：该过滤器是Waha5.0的版本所新增的，是为了能够将以设计好的作业进行批处理或是预处理，也可以设置自动处理，值得说明的是，这个过滤器可以单独存在于一个作业中并成功执行，这个过滤器可以看成是一个已存在的且被调用的作业，因此可以单独存在预作业中并能执行，甚至也能够直接连接到另一个作业执行过滤器，以实现一个调用的功能，如下图6-12所示。



图6-12.作业执行过滤器

## 五、管理和调度功能

在Windows系统中里有数据处理自动执行的日程管理功能，事件触发功能，在“控制面板”->“管理工具”中能够找到。将Windows自有的这个自动执行功能配合Waha的批处理和调用，就能够在夜间执行一些无人的操作，很大程度的提升了工作效率。在数据处理过程中，如果某一段流程出现故障，可以设定自动报警，同时在流程图上显示出红色的旗子，以示此处有故障。如果流程执行正常，则为蓝色，警告完成时则为黄色。

Waha用户可以根据实际的需要，设定处理日程，按照设定时间自动启动执行数据处理任务，编者这里使用的是Windows10的操作系统，该系统中的自动执行功能如下图6-13所示。



图6-13(A).自动执行功能



6-13(B).自动执行功能的设置

## 六、集成和开放性

Waha的集成，主要是和其它的产品打包到一起，可以和其他系统结合，可以调用外部程序，也可以为其他工具提供接口供被调用。Waha有着处理速度快、性能优良等优点，得益于这些优点，目前已有多个软件厂商将Waha集成到其产品中，并为其带来经济效益。

## 七、对元数据的管理

Waha以独特的方式，对元数据进行管理。通过文档自动生成器，可以把每个流程的概要，部分的相互关系，数据来源，数据形式及数据性质等等，都以图形自动记录下来。



## 和其他ETL的对比

|  |  |  |  |
| --- | --- | --- | --- |
| 产品名/其厂商 | PowerCenter/Informatica | DataStage/IBM | WahaTransformer/BSP |
| 在美国市场的排名 | 2 | 1 | × |
| 在日本市场的排名 | 3 | 2 | 1 |
| 在日本国内的客户 | 未公布 | 120家 | 1400家 |
| 数据抽取 | ○ | ○ | ○ |
| 数据转换 | ○ | ○ | ○ |
| 数据清洗 | ○ | ○ | ○ |
| 数据装载（批处理） | ○ | ○ | ○ |
| 数据装载（ODBC） | ○ | ○ | ○ |
| 输入输出（1:1） | ○ | ○ | ○ |
| 输入输出（1:N分割） | ○ | ○ | ○ |
| 输入输出（N:1结合） | × | ○ | ○ |
| 各种变换函数 | ○ | ○ | ○ |
| 字码变换(MF IBM) | ○ | ○ | ○ |
| 字码变换(UNIX UNICODE big endians) | ○ | ○ | ○ |
| 字码变换(UNIX UNICODE Little endians) | ？ | ？ | ○ |
| 字码变换（UNIX UTF-8) | ○ | ○ | ○ |
| 字码变换（UNIX SJIS) | ○ | ○ | ○ |
| 字码变换（UNIX GB18030) | ○ | ○ | ○ |
| 字码变换（Windows UTFｰ8) | ○ | ○ | ○ |
| 字码变换（Windows SJIS) | ○ | ○ | ○ |
| 字码变换(Windows GB18030) | ○ | ○ | ○ |
| 外码对应 | ○ | ○ | ○ |
| 实时更新 | △ | △ | ○ |
| 并行体系构造 | ○ | ？ | ○ |
| Security/User管理 | ○ | ○ | ○ |
| Remote | ○ | ○ | ○ |
| M/F | × | × | × |
| UNIX | ○ | ○ | × |
| NT | ○ | ○ | ○ |
| Source对象是NT | ○ | ○ | ○ |
| Source对象是UNIX | ○ | ○ | ○ |
| Source对象是大型机 | × | × | ○ |
| 错误/故障报告 | ○ | ○ | ○ |
| Roolback | ○ | ○ | ○ |
| Logfile | × | ○ | ○ |
| 来自外部AP,API,DLL的调用 | ○ | ○ | ○ |

## 九、性能参数

测试环境：

|  |
| --- |
| 硬件配置 |
| CPU | Interl(R) Xeon(R) CPU E5630 @ 2.53GHz 2.53GHz(双核) |
| 内存 | 18.0 GB |
| 磁盘 | 900 GB |
| 操作系统 | 64 位操作系统 |
|  软件配置 |
| 系统 | Windows server 2008 R2 Enterprise |
| DB環境 | mysql-5.6.17-winx64 |
| ETL | Waha! Transformer V5.0 |

测试结果：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据类型 | 测试用例 | 过滤器 | 500万条 | 4000万条 | 1亿条 |
| MysqlDB | 往DB中插入 | 标准过滤器 | 00:16:39 | 02:14:20 | 05:26:25 |
| 从DB中读取 | 标准过滤器 | 00:00:47 | 00:06:33 | 00:18:00 |
| 从DB中读取然后插入到DB中 | 标准过滤器 | 00:17:24 | 00:30:37 | 05:46:03 |
| CSV文件 | 从CSV读取到CSV写入 | 标准过滤器（使用了replace参数） | 00:01:51 | 00:12:59 | 00:31:26 |

# 第七章 Waha用户

Waha在各行业中的用户分布状况如下图所示。用户行业覆盖面广，用户数量大，其中不乏500强企业，主要的用户行业为制造业、流通业、IT行业、金融业。其中金融行业的数据量大，单个用户每日产生的数据达到TB级，但是Waha也能够完全胜任，这也是为什么Waha备受青睐的原因。

Waha的用户比例如下图7-1所示。

图7-1.Waha用户比例分布

# 总结

Waha作为一款专业的数据处理软件，在处理速度和性能上都有着不俗的表现，这些成绩也是成为了用户选择Waha的原因之一。作为数据处理软件，我们和其他平台的软件都做了纵向和横向的对比，详情可以参阅我们的《Waha!Transformer测试与对比报告》。除了处理速度快的优点，我们还支持相当多的连接源以进行连接Waha，这些连接源涵盖了主流的和小众的类型，使得用户能够在不同的数据源的情况下都能完美的使用Waha以进行数据处理，这也是为什么Waha备受青睐的原因。

我们总结了选择Waha的六个理由：

1. 丰富的数据加工和清洗功能：Waha有9个过滤器和100多个函数，可以快速对应系统间的复杂数据加工和数据连接。
2. 数据处理快：Waha作业会根据CPU的数量和处理的复杂性，为大量数据的处理提供高速的处理性能。
3. 学习成本低，易上手：针对于非开发人员，以流程图的形式操作，简洁易懂，任何人员都能迅速上手。
4. 维护工作量少：流程图的作业形式使得工程简洁易懂，易于理解，可灵活的变更系统的数据格式，可自动生成文档方便数据的查询与更改，可实现迅速的数据导入与开发。
5. 文字编码多：支持多种文字编码，同样的也有着中、英、日三种语言的操作界面，不用担心文档乱码或是国外操作人员不懂中文的问题。
6. 多种连接源：能够匹配和连接多种的数据格式和数据库，基本涵盖了主流的数据库和小众化的数据库，同时也涵盖了多种格式的本地文件。

Waha是我们自成立公司以来一直所专注研发和升级的产品，我们的对于产品的理念一直都是“尽善尽美”，Waha从最初的1.0版本到现在的5.0版本，从最初的单个语言版本到现在支持三个国家的语言版本，这中间离不开用户的支持和我们开发人员的精心开发过程，在过去以及未来还有现在的日子里，我们会一如既往的对Waha进行不断的改善，用户的满意就是我们的成绩。我们所编写的这一技术白皮书是为了用户能够更多了解我们的Waha，我们也同样的抱着“所见即所得”的理念去编写每一篇文档，这样，用户才能信任Waha，信任我们。

非常感谢您选择我们的产品！

备实必（上海）软件科技有限公司