DAT 337
使用 SQL Server Integration Services 装载数据仓库

吕科
技术咨询顾问
上海星移软件有限公司

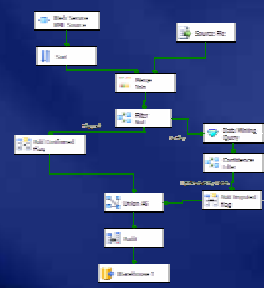
Microsoft Tech·Ed 2005中国

创新 · 远见 · 分享 · 协作

Microsoft SQL Server 2005
商业智能的平台

Integration Services 简介

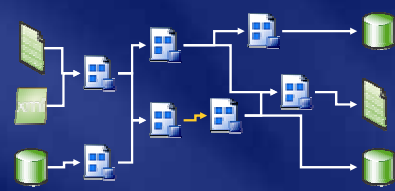
- 新的 SQL Server 商业智能应用程序
- DTS的升级版本
- 新一代的高性能数据整合平台



演示

SSIS Overview

SSIS 能做什么？



- 支持多种数据源，文本、Xml 、OLE DB、 ODBC
- 支持复杂的数据转换流程，包括多路、循环、条件执行
- 数据可以按规则分发
- 强大的错误处理机制，强大的事件提醒功能
- 支持多路输入及输出

聚合和排序

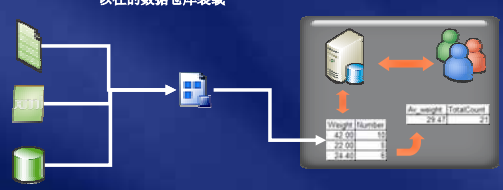
- 数据转换的重要操作
- 在数据流中支持多种聚合函数
  - Group By, Sum, Count Distinct 等
- 在数据流中排序
- SSIS 支持真正的 ETL，不是 ELT
  - 聚合性能甚至超过数据库

关联

- 清洗数据的重要操作
- Lookup
  - 较大的事实表和较小的维度表，如分类维度
- Merge Join
  - 较大的事实表和较大的维度表，如客户维度

适应新的数据仓库架构

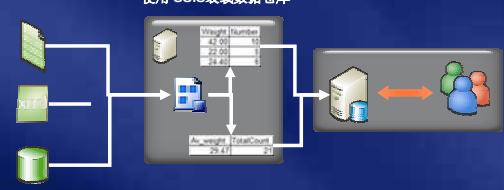
以往的数据仓库装载



- 在以往的场景中，只能够把数据原封不动的抽取到数据库服务器上
- 由数据库来执行聚合、排序和其他操作
- 影响用户访问的性能
- 这种解决方案就不适合于大容量的数据或复杂的逻辑

适应新的数据仓库架构(续)

使用 SSIS装载数据仓库



- SSIS 抽取数据还是和以前一样
- 但是，由 SSIS 来执行聚合和排序操作，然后装载到数据库中
- 它释放了数据库服务器的压力，使其可用于用户查询
- 支持 64 位的环境，可以更好的处理大数据量和复杂的逻辑
- 就算在 32 位 的环境下，也可以更好的分散服务器的压力

特殊的功能

- 时间维的生成
  - 通过 Analysis Services
  - 指定日期范围来生成纬度成员
- Pivot 和 Unpivot 操作
- 更多的功能

演示 – UnPivot

From Currency	To Currency	Exchange Rate	EUR
GBP	1.0000	0.4048	0.6643
GBP	2.4714	1.0000	1.6407
GBP	1.5058	0.6100	1.0000

演示

Core Features

数据清洗

- Fuzzy lookup
  - 模糊查找，找到最合适的匹配值
- Fuzzy duplicates
  - 模糊分组，例如 “Windows XP,” “WinXP,”
- 由微软中国研究院完成

数据挖掘

- 在 ETL 过程中使用挖掘模型
  - 使用输入数据来训练挖掘模型
  - 使用挖掘模型来处理数据
- 基于模式的 data quality
  - 使用挖掘模型来预测背离值
- 在运行时根据数据来选择最佳算法

演示

Data Quality

Analysis Services

- 设计时和运行时的集成
- Analysis Services 作为数据目的

演示

Analysis Services 集成

装载到 Analysis Services

- Analysis services 作为数据目的
- 一步到位，不需要中间存储
  - 以往
    二维数据 → SQL Server → AS
  - SSIS
    二维数据/XML/Web Service/... → AS

对Analysis Services进一步的支持

- Execute DDL task
  - 就像执行 SQL 命令一样
  - 建立多维数据集、分区等
- 共享的项目结构
  - 一起管理 AS 和 ETL 项目（包括 RS 和 VS）
  - 只需一次定义数据源信息

缓慢变化维

- 自动建立一类和二类缓慢变化维
- Fixed Attribute
  - 当改变是个错误
- Changing Attribute
  - 不记录历史的更改，如名称
- Historical Attribute
  - 记录历史痕迹的更改，如所属机构
  - 会引发新增维度成员
- Inferred member
  - 迟到的维度数据
  - 先装载事实数据

缓慢变化维(续)

- Business key
  - 来自于源数据的key
- Surrogate key
  - 可来自于其它字段
  - 通常是自动增长

演示

缓慢变化维向导

其他的改变

- SQL Task – 可执行存储过程和其他 SQL 语句
- 可扩展性
  - 可以编写自己的 SSIS 组件
  - 可以在script任务中调用.NET 代码
- 可调用性
  - Web service
  - XML
  - .NET

可靠性和可恢复性

- Error flows
  - 在任务中处理错误数据
- Event handling
  - 在执行包时响应事件
- Package restart
  - 设置失败的检查点

操作

- 日志记录
  - 丰富的logging providers，如：文本、数据库、XML、Windows event log 或自定义的 log provider
  - 可以给单个或多个任务设置独立的日志
- 性能监视器
- 整个生命周期的支持
  - 运行时设置属性和变量
  - 可通过 XML，注册表，环境变量，SQL 进行配置
  - 发布功能

总结

- 更快的速度！
  - 以数据流的方式，能够高效的处理复杂的逻辑。
- 更强的功能！
  - 新增的Data flow
  - 增加的Task和Transformation
  - 与其它微软BI产品的结合
- 真正的ETL！
  - 在流程中清洗数据

社区资源

- 微软SQL Server社区
  - microsoft.public.sqlserver.server
  - microsoft.public.cn.sqlserver
- 我的社区
  - http://www.dev-club.com
  - http://blog.joycode.com/luke
- 中文 SQL Server 主页：http://www.microsoft.com/china/sql
- 下一个会议
  - DAT239：SQL Server 2005数据仓库新功能

社区资源

- 其它英文资源
  - http://msdn.microsoft.com/SQL/sqlwarehouse/SSIS/default.aspx
  - http://www.sqlis.com

问题

- 什么是SQL Server Integration Services？
- ETL和ELT相比有什么好处？
- 和Analysis Services的集成带来什么好处？
- 缓慢变化维有什么用处？