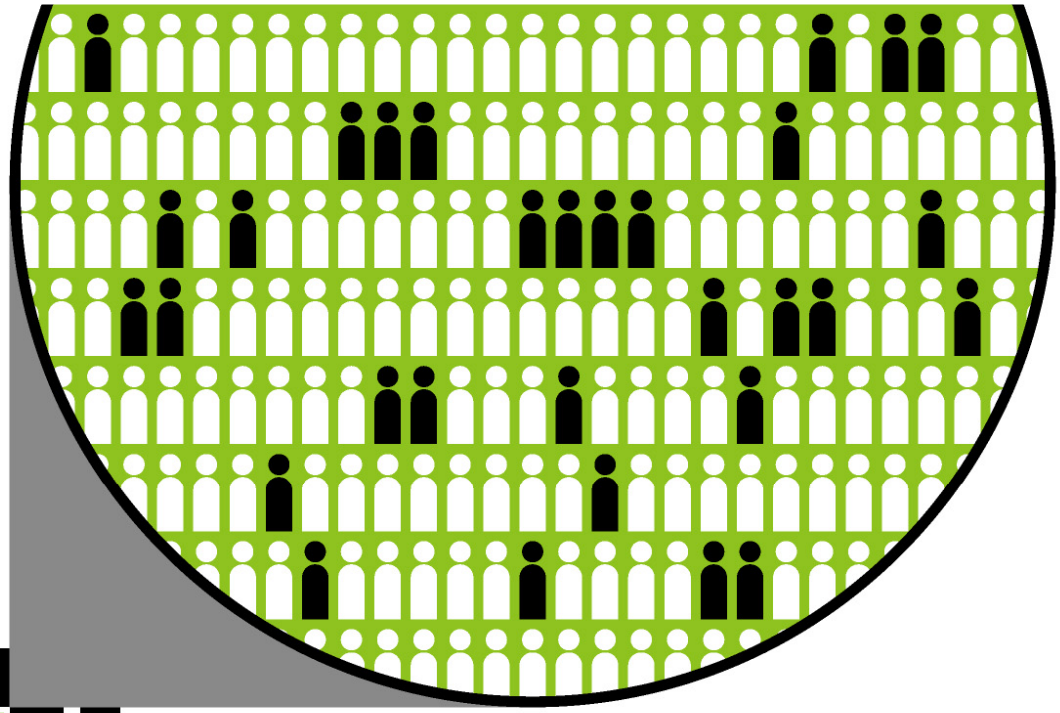
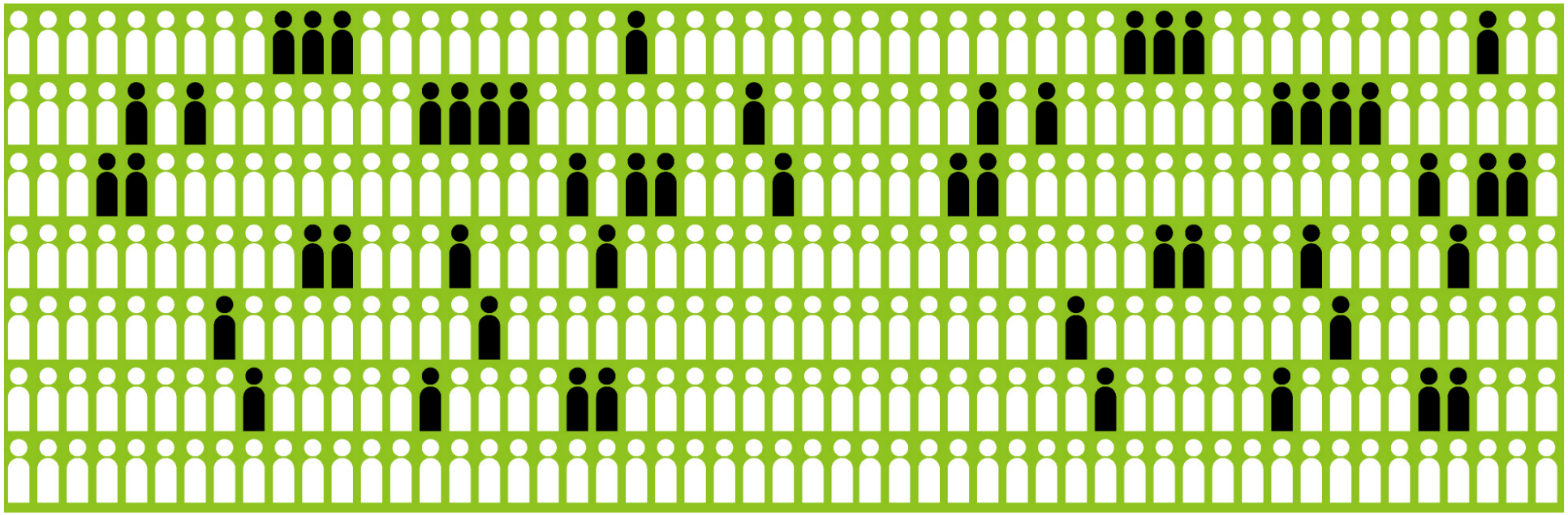


分析为道 Z 者见智

IBM主机商业分析(BA)高峰论坛

数据抽取转换加载ETL介绍

张学宇



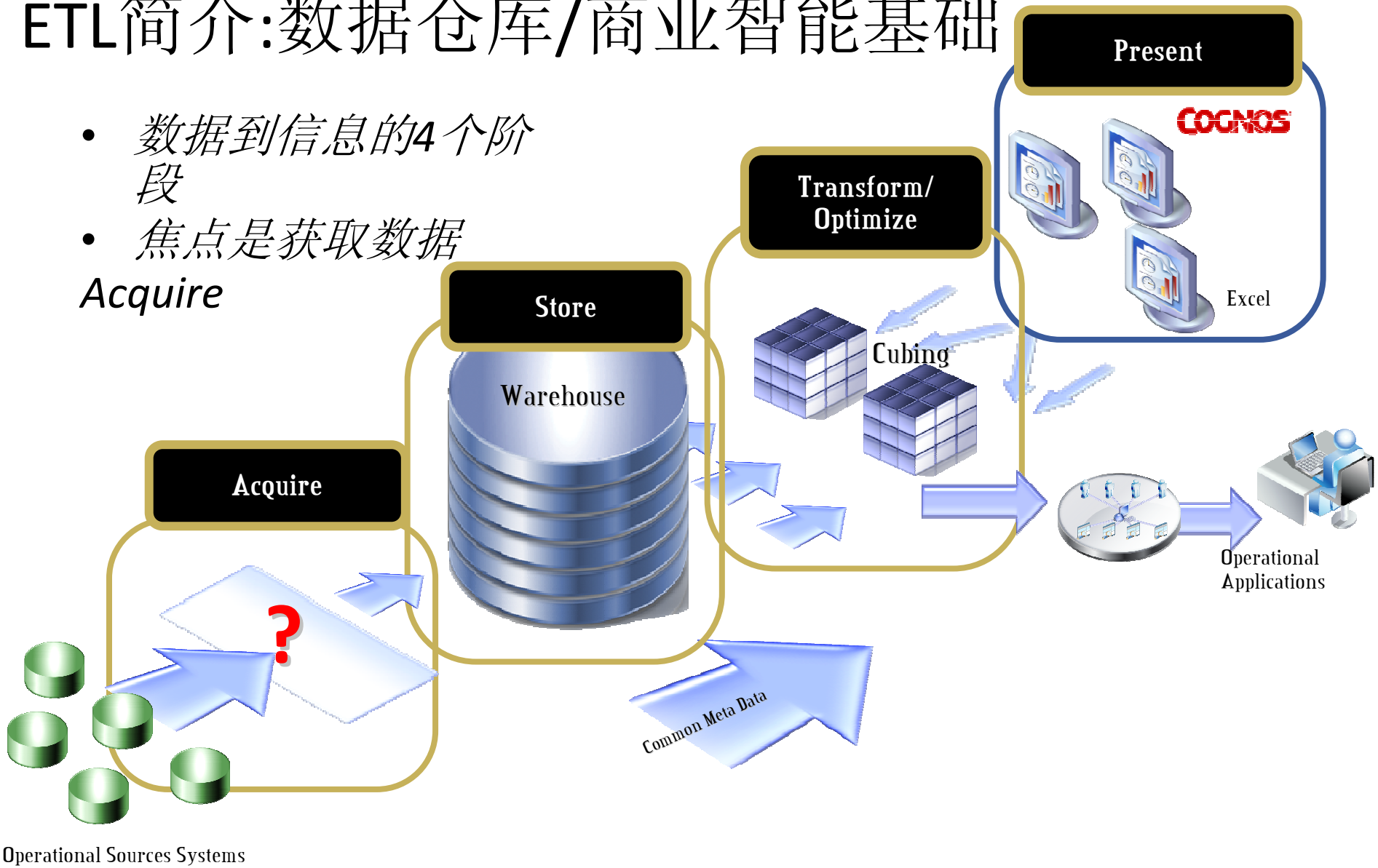
议程

- ETL简介和要点
- ETL解决方案
  - InfoSphere Information Server
  - 其他工具

ETL简介:数据仓库/商业智能基础

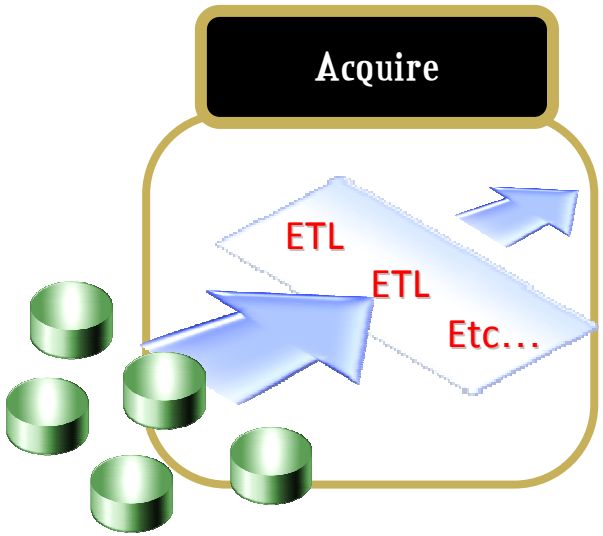
- *数据到信息的4个阶段*
- *焦点是获取数据*

Acquire



ETL简介:获取数据

- *如何?*
  - *传统的方法是批次抽取数据加工转换并加载*
- *问题*
  - *多种数据源、大量数据、转换规则复杂*
  - *对生产系统的影响*
  - *数据加载不及时*
  - *数据量增长对全量抽取的影响*



- *考虑…*
  - *更低系统影响*
  - *更及时（准实时）的数据仓库*
  - *Etc…*

ETL简介：考虑要点

- 数据源
  - 数据库
  - 文件
- 数据量
  - 全量
  - 增量
- 时间
  - 批次时间
  - 周期
  - 实时
- 转换规则复杂度
  - 技术转换
  - 业务转换
- 数据目标模型
  - 维度
  - 事实

- 运行管理
  - 监控
  - 作业调度
- 系统维护
  - 元数据
  - 通用服务
  - 初始化过程
  - 异常处理
  - 需求调整
- 项目管理
  - 开发
  - 维护
  - 变更

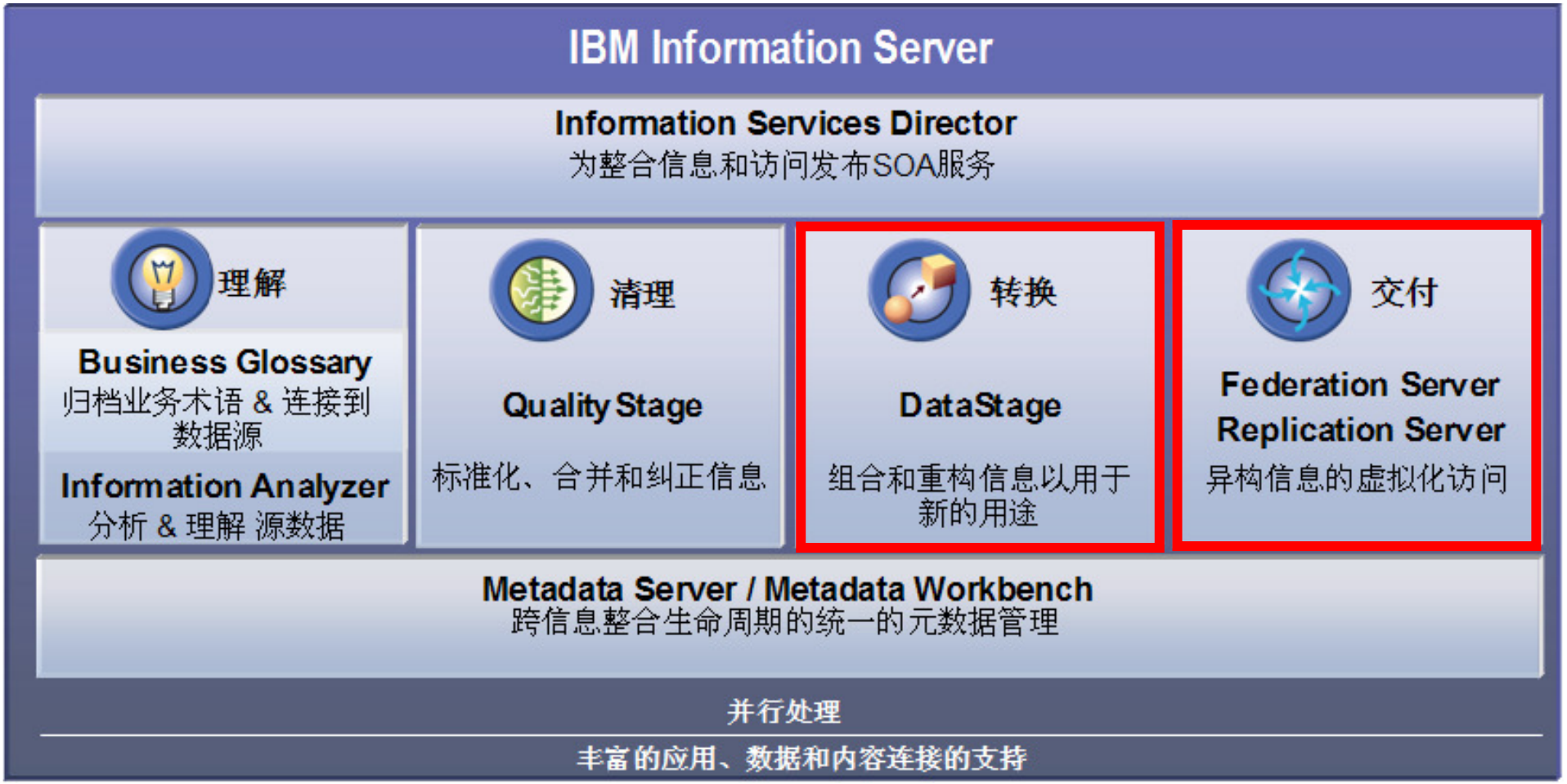
- 软件架构
  - 通用服务
  - 扩展能力
  - 元数据管理
- 硬件平台
  - 稳定性
  - 可用性
  - 可靠性
  - 扩展能力
    - I/O
    - CPU
    - NETWORK

议程

- ETL简介和要点
- ETL解决方案
  - InfoSphere Information Server
  - 其他工具

z平台ETL解决方案

- InfoSphere Information Server
  - InfoSphere Information Server
  - 批次抽取转换加载DataStage
  - 实时增量复制Replication
  - 综合手段DataStage and Replication
- 其他工具



- IBM 信息服务器包含一组产品模块，用于解决各类商业问题。
- 信息验证、访问和处理规则适用于多个项目，提供高级别的一致性和严格的数据控制并能够提高 IT 项目的效率。

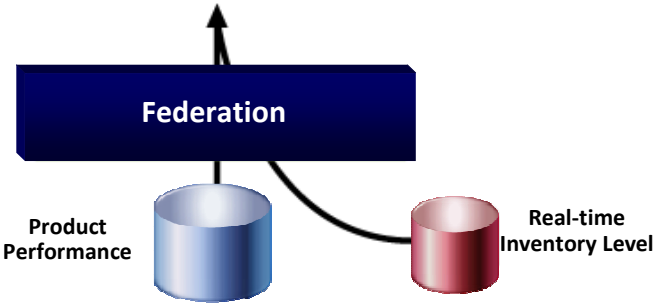
不同类型的数据整合解决方案

Federation

Federation Server
Classic Federation Server

Analytical & Reporting Tools

Web Applications



Consolidation (ETL/DQ)

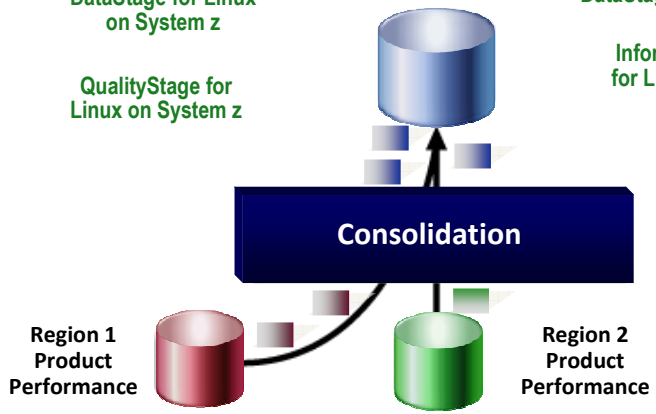
DataStage for Linux on System z

QualityStage for Linux on System z

DataStage MVS

Information Analyzer for Linux on System z

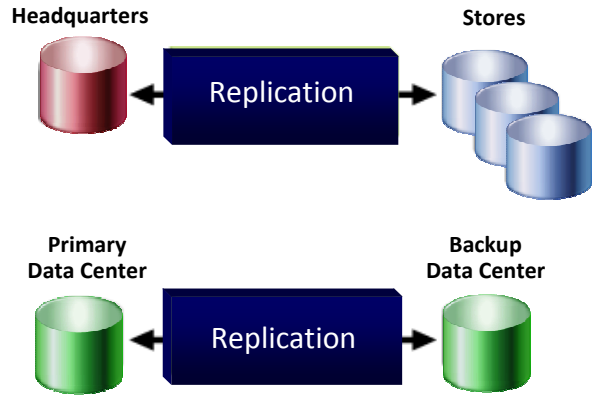
Business Glossary for Linux on System z



Replication

Replication Server
Classic Replication Server

IS Change Data Capture

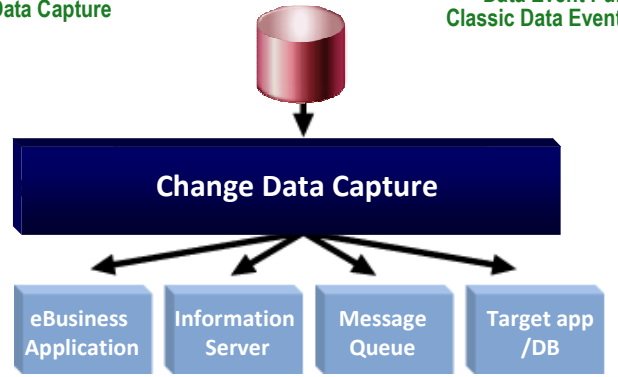


Change Data Capture

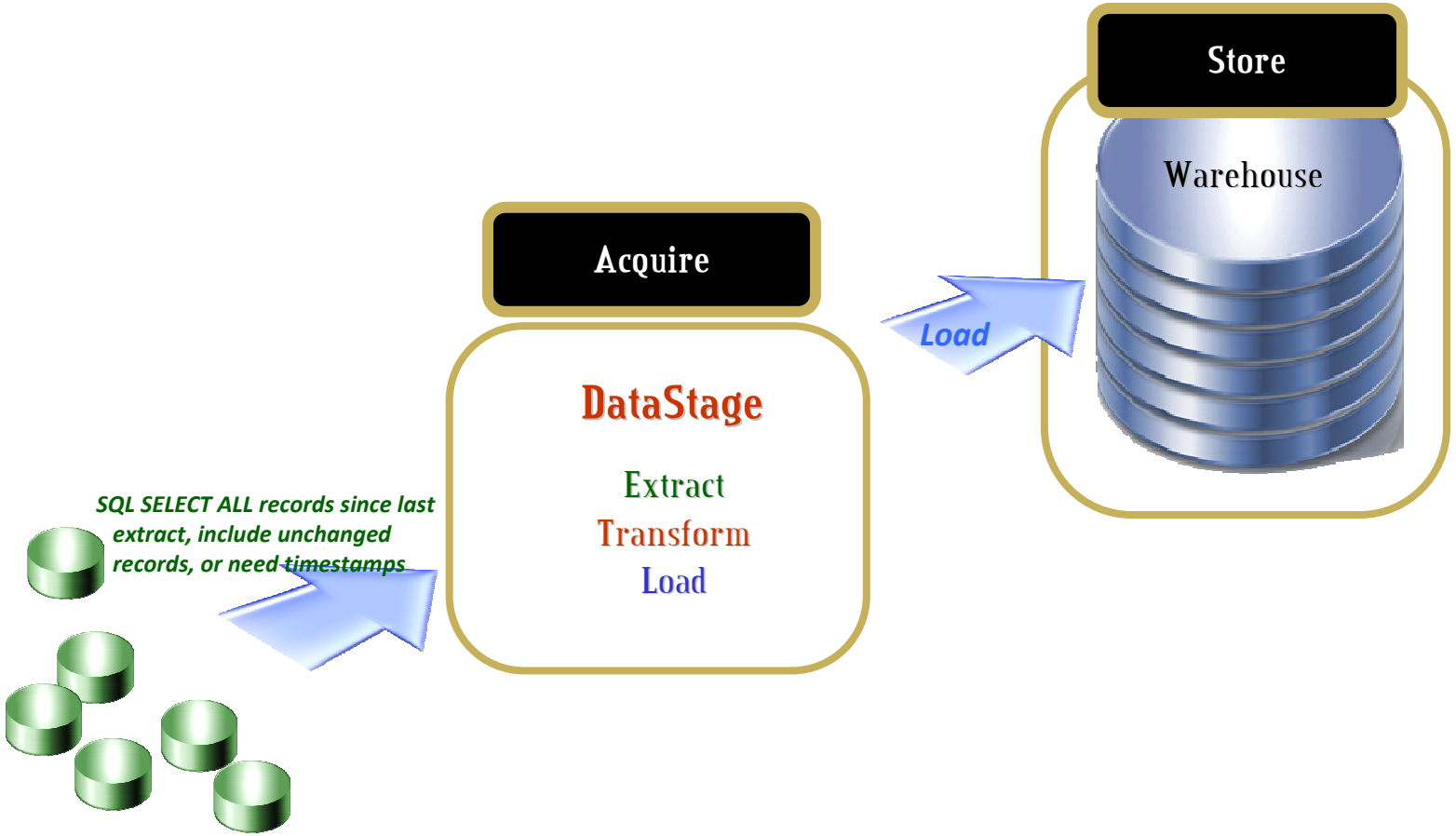
IS Change Data Capture

Database

Data Event Publisher
Classic Data Event Publisher

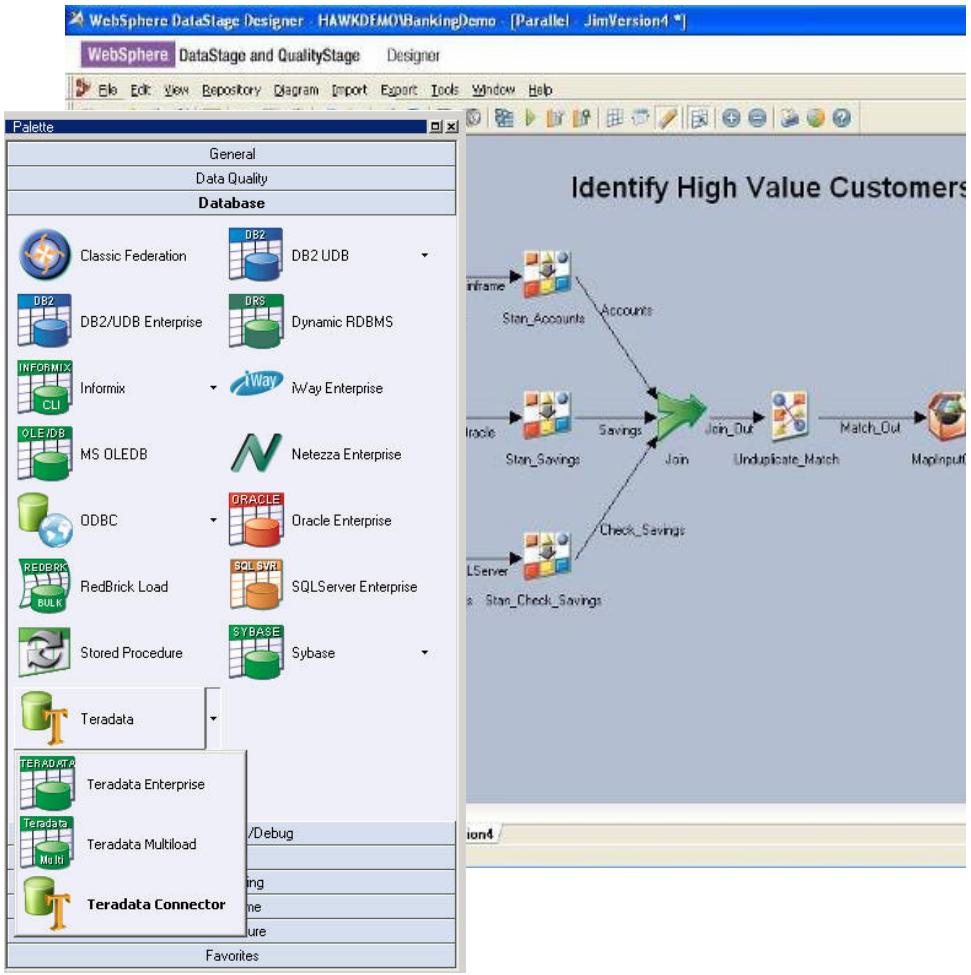


ETL --DataStage(Consolidation)



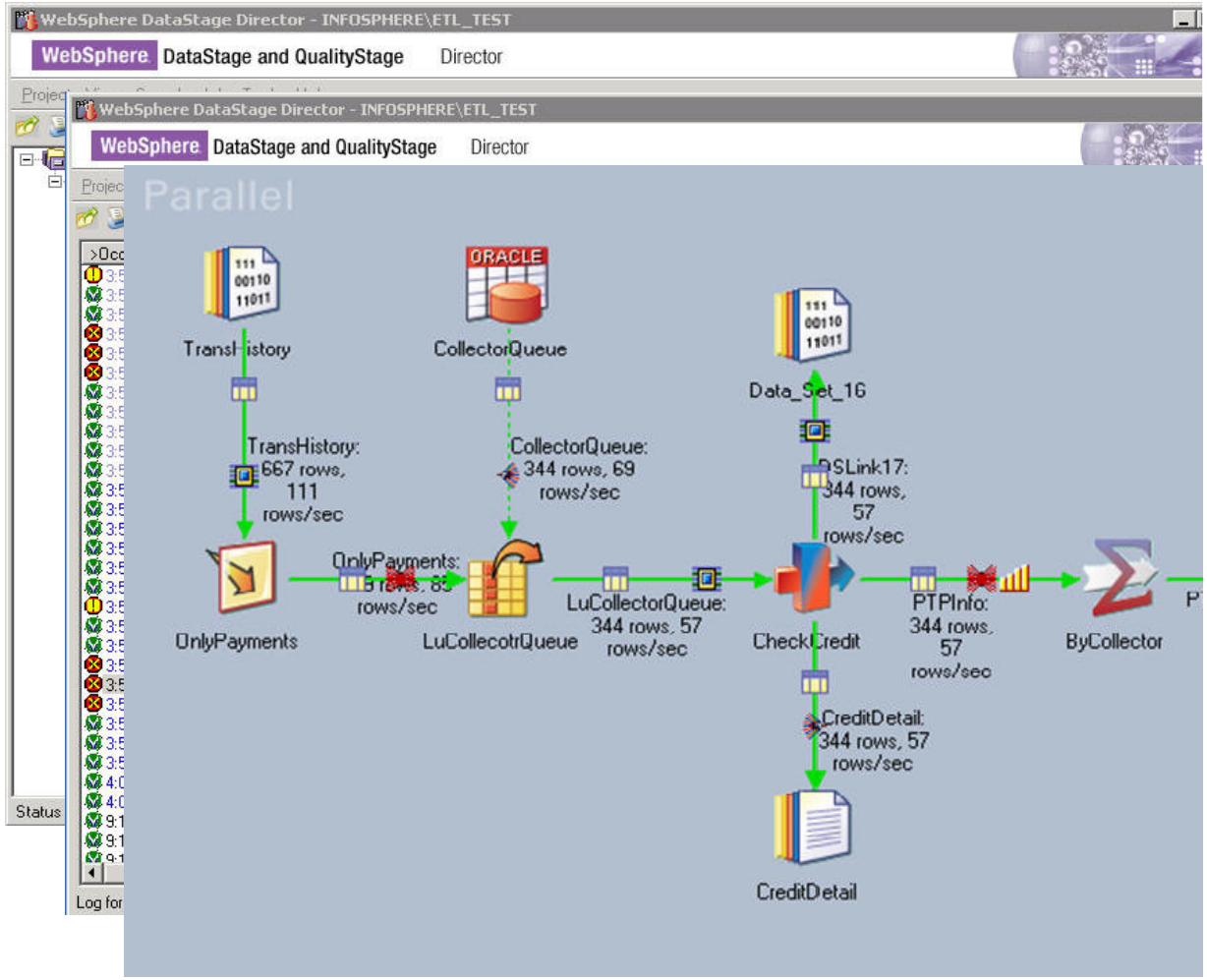
ETL--InfoSphere DataStage 开发

- 易于使用的图形化界面
  - 使开发、维护和调试变得容易
  - 只需要鼠标的点击即可完成数据整合
  - 基于组件的体系结构
  - 可重用性
  - 顺序开发, 并行执行
- 使复杂的转换变得容易
  - 复杂的转换规则通过使用 Transformer 可以很容易的完成
- 可视化的流程控制开发
  - 可视的任务控制无需脚本语言
  - 对处理条件的完全支持
  - 支持等待文件上传, 执行外部命令, Email 通知等
- 内建转换组件
  - 50多种内置转换和处理组件，能够迅速创建作业并提高开发效率
  - 大量内建函数使得实现复杂转换逻辑更加容易
  - 易用的组件编辑器，能够快速定义转换逻辑
- 整套支持各类数据库控件

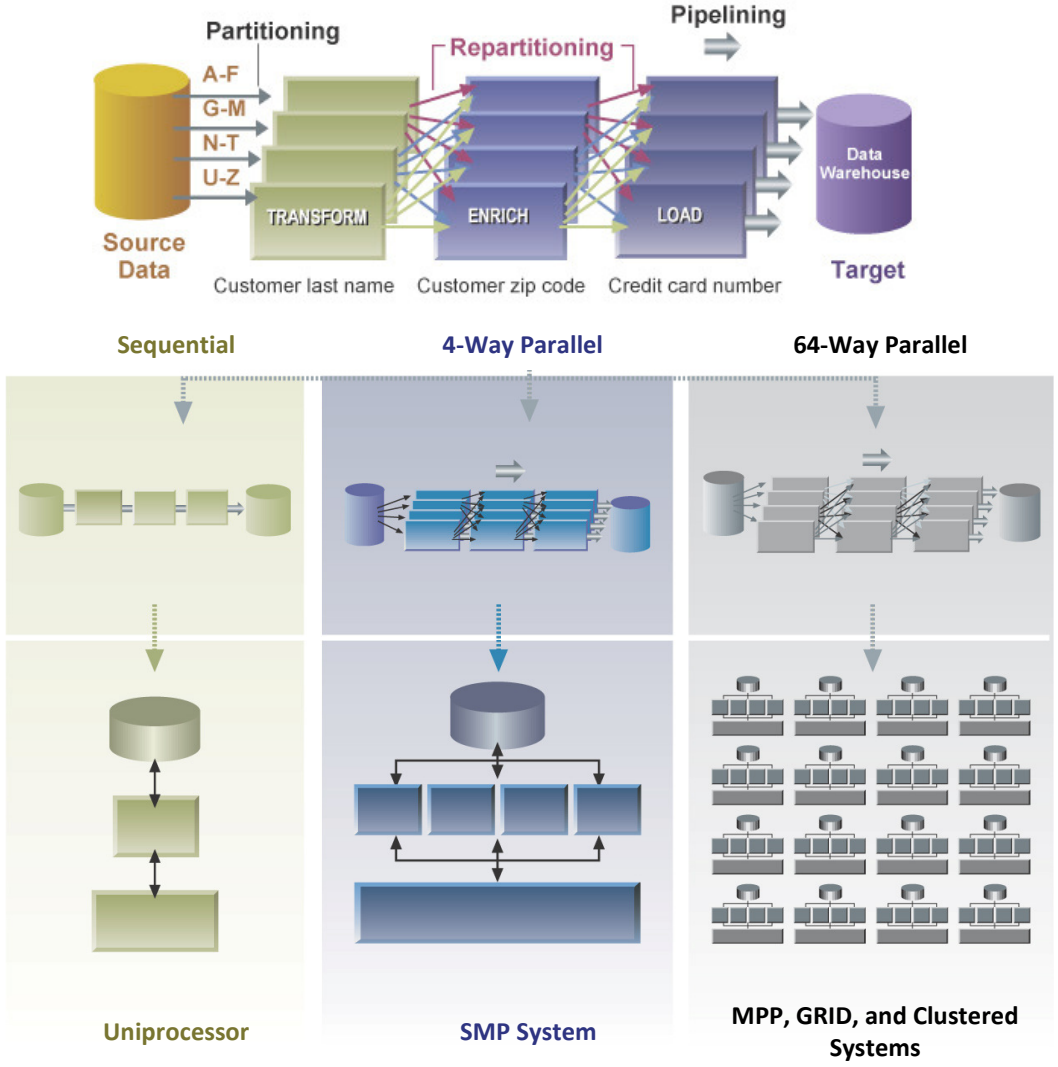


ETL--InfoSphere DataStage 工作

- 任务的调度
  - 灵活的安排作业运行的时间
- 任务运行的审查
  - 任务运行的详细日志
  - 提供了恢复和诊断的机制
- 图形化的监控
  - 监视ETL工作流程的进度和性能，帮助发现流程中的瓶颈



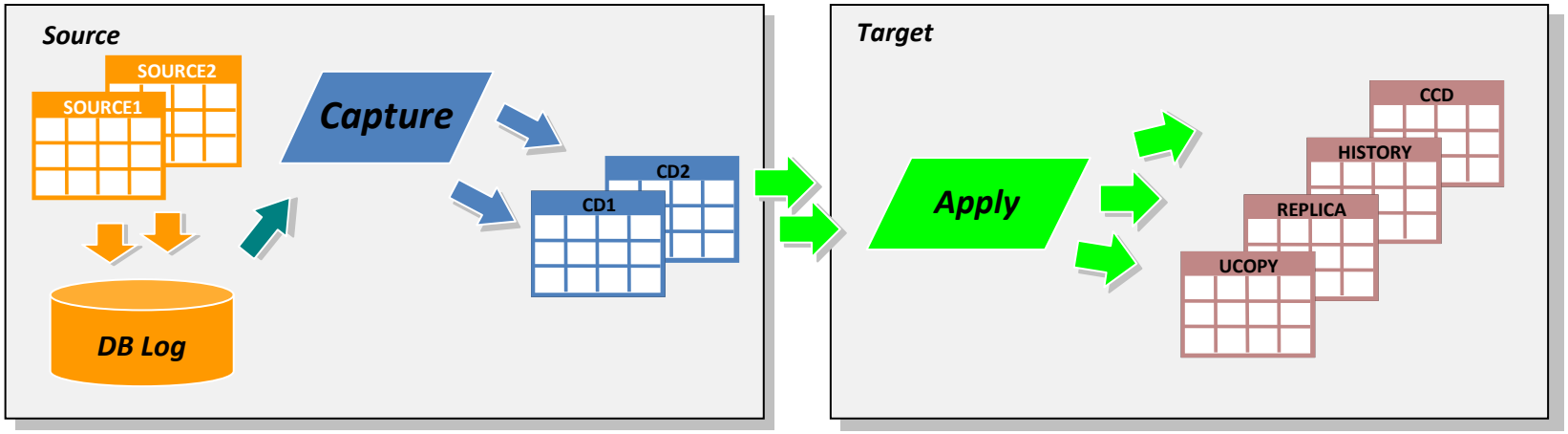
ETL--性能：并行执行



- 设计数据整合流程而不用担心庞大的数据量和时间的限制
- 能够充分利用数据库分区技术提高数据处理的性能
- 通过简单的步骤就能在数据处理时定义数据的分区
- 通过一个简单的配置文件就能增加处理器和其他硬件
- 不需要编写大量代码，并能充分利用多个处理器进行处理
- 支持对称式多处理器(SMP)，集群(Clustered)，网格(Grid)和大量信息并行处理机(MPP)体系结构

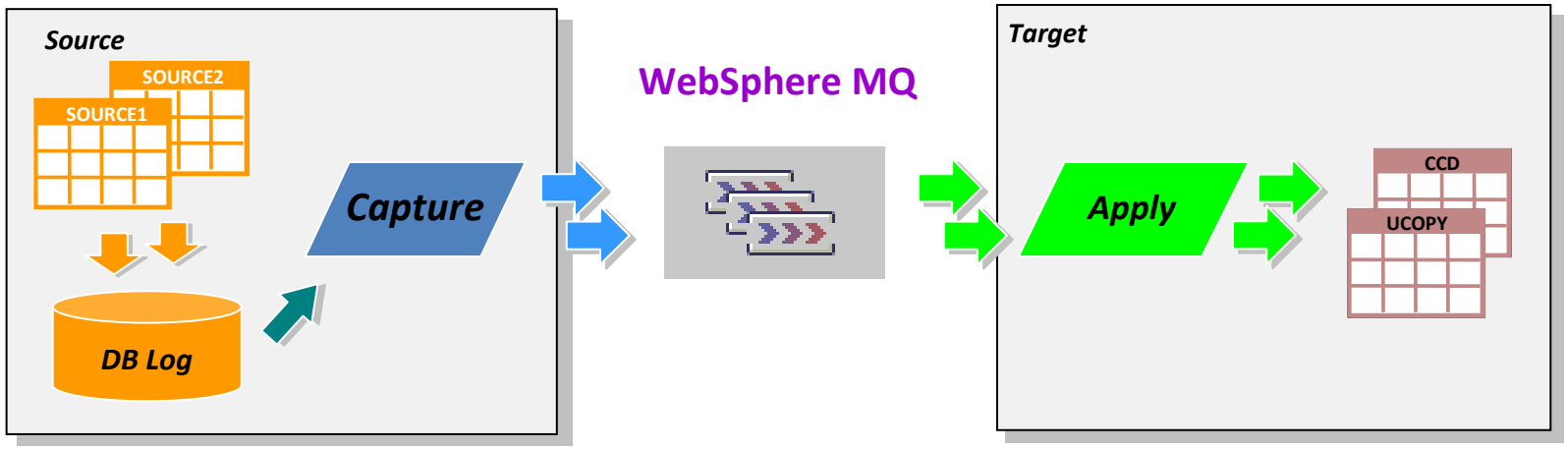
复制技术：SQL Replication

- 从DB2 LOG获得增量数据
- 增量数据落地在DB2表（CD表）中
- 增量数据驱动Apply program或Apply program获取
  - 通过DB2 SQL，基于 DB2 client-server架构实现



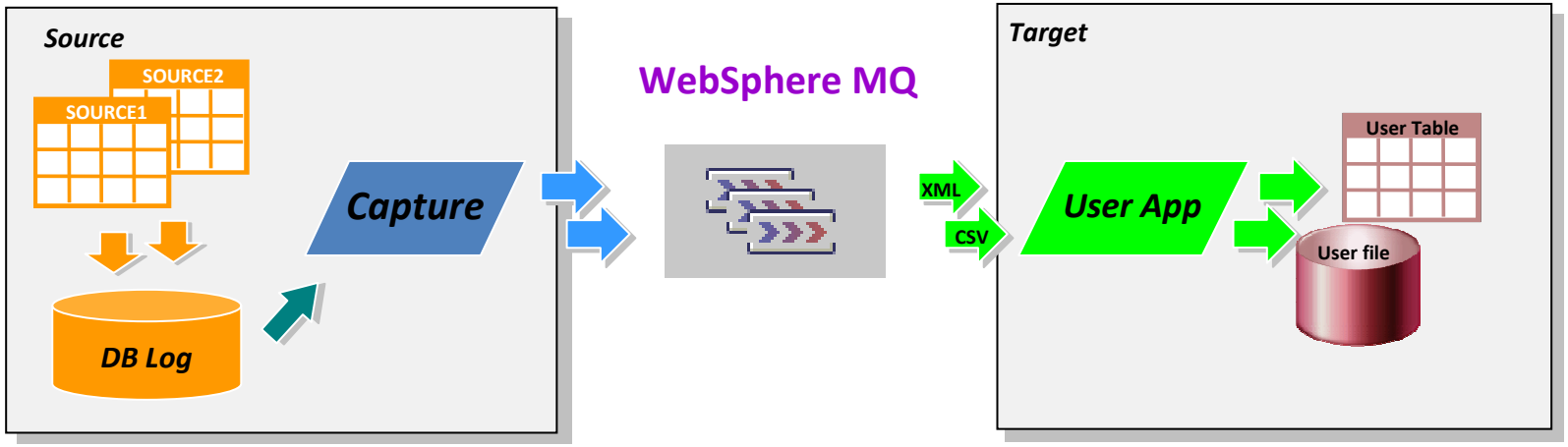
复制技术：Q Replication

- 从DB2 LOG获得增量数据
- 通过MQ 发布数据，不需要落地在DB2表中
- Apply 程序从队列获取数据
- 高性能，连续高可用性

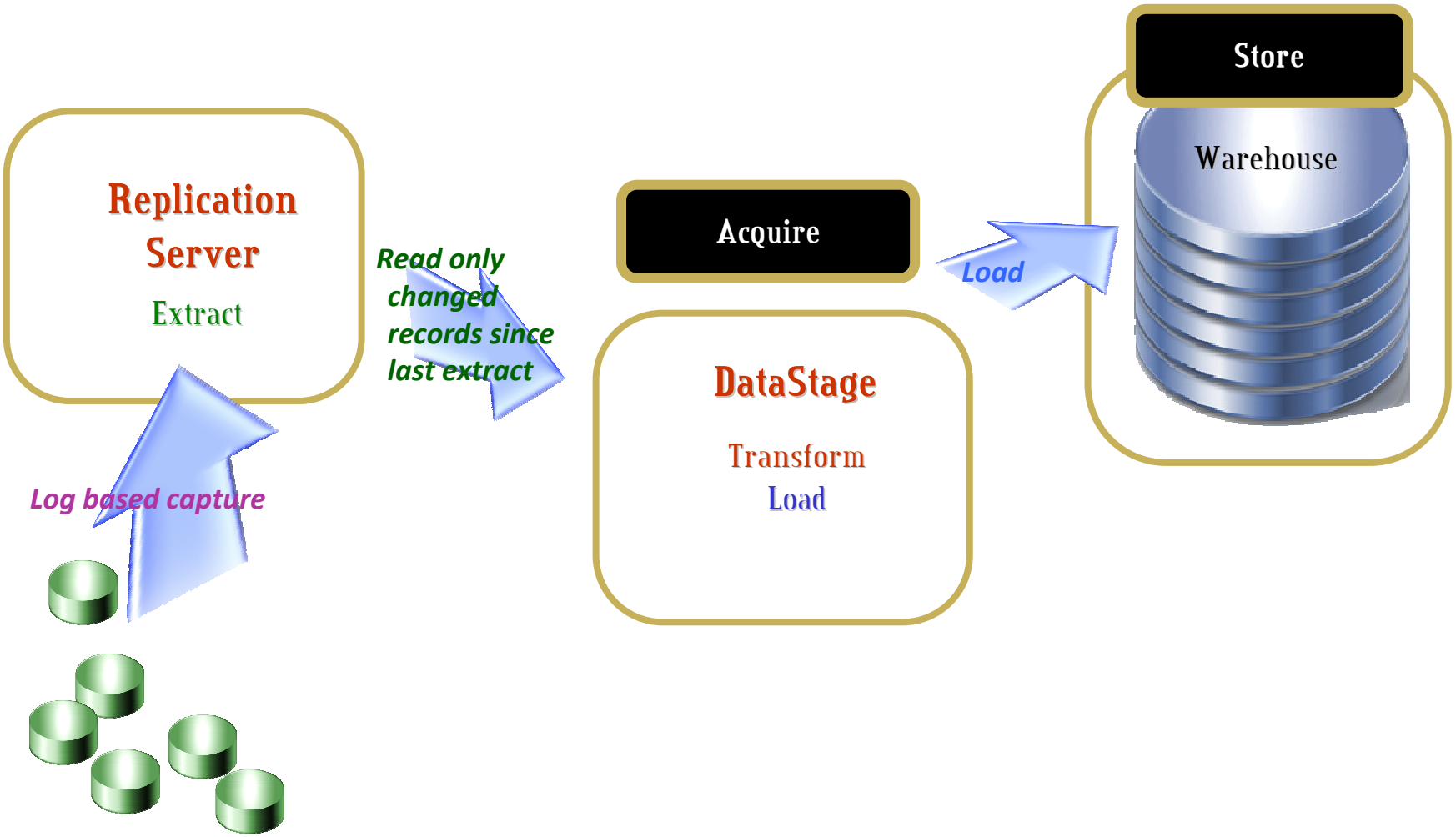


Event Publisher

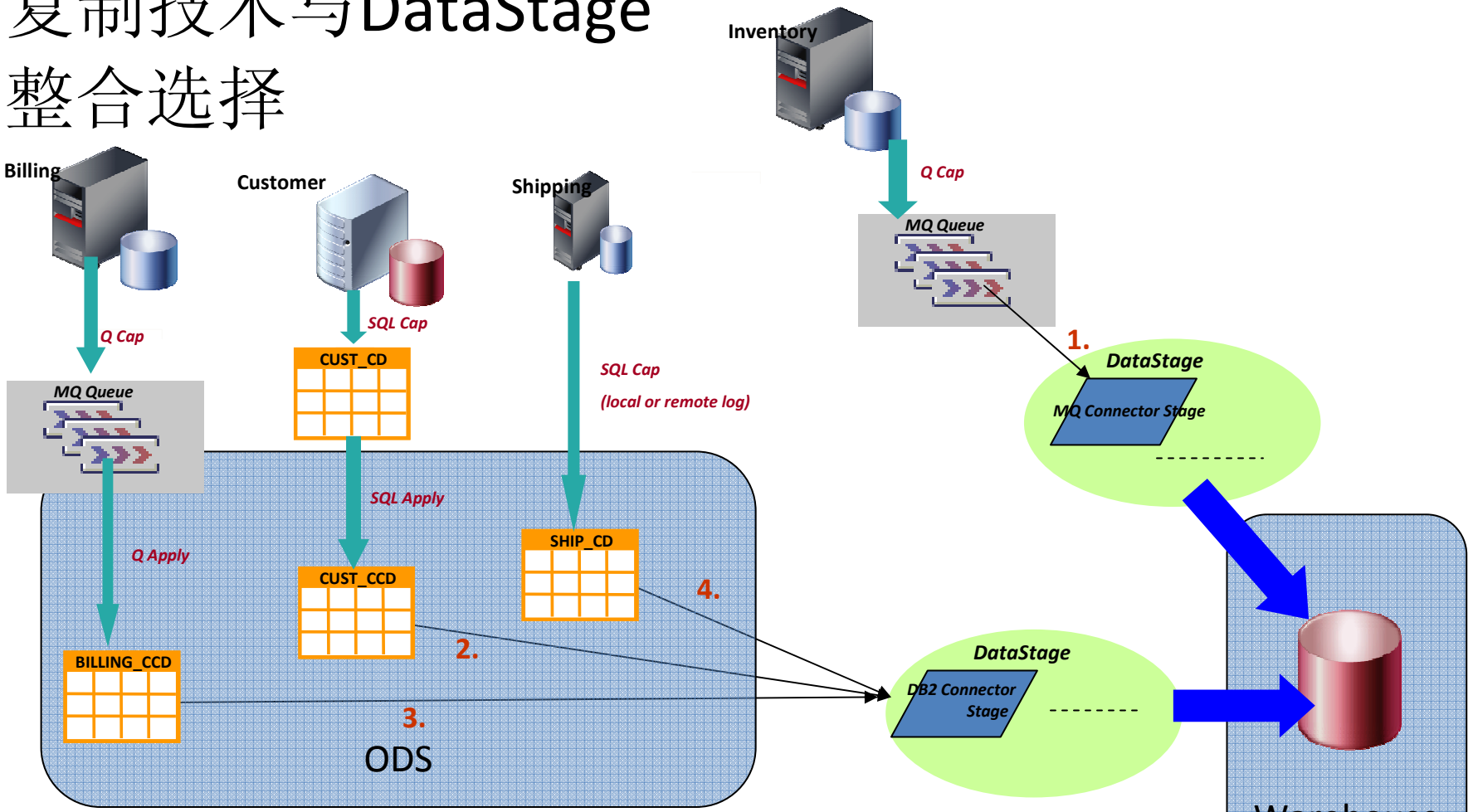
- 从DB2 LOG获得增量数据
- 通过MQ 发布数据，不需要落地在DB2表中
- 目标应用从队列获取数据，XML或者CSV格式
- 高性能，连续高可用性，应用灵活



复制技术与DataStage 整合



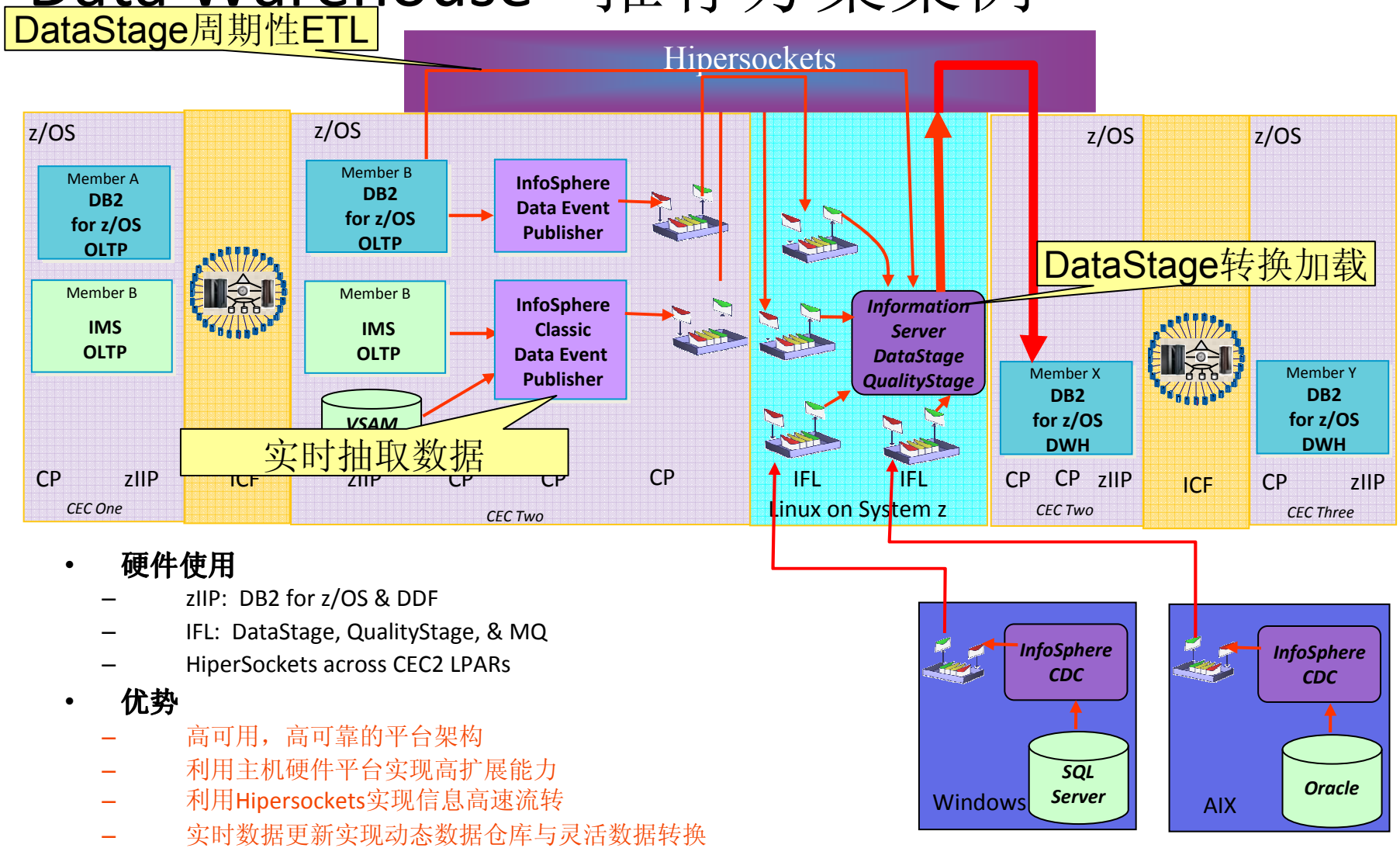
复制技术与DataStage整合选择



1. *Inventory: Event Publishing with Q Capture only, no Q Apply*
2. *CUST_CCD: SQL Replication with CCD*
3. *BILLING_CCD: Q Replication with CCD*
4. *SHIP_CD: SQL Replication with CD directly, no SQL Apply. DS must run source machine.*

at *Cognos* *Excel*

Data Warehouse - 推荐方案案例



- **硬件使用**
  - zIIP: DB2 for z/OS & DDF
  - IFL: DataStage, QualityStage, & MQ
  - HiperSockets across CEC2 LPARs
- **优势**
  - 高可用，高可靠的平台架构
  - 利用主机硬件平台实现高扩展能力
  - 利用Hipersockets实现信息高速流转
  - 实时数据更新实现动态数据仓库与灵活数据转换能力相结合

其他工具

- SQL Warehouse Tool(SQW)
- DB2 High Performance Unload
- File Manager
- Optim Test Data Management Solution
- Optim Data Growth Solution
- …


Thank
You