IBM

您的信息 您的智慧

2011 IBM 信息管理与业务分析论坛

数据治理中的数据质量和元数据管理重要性

刘春霞

资深工程师

议程

- 数据治理中的数据质量和元数据
  - ✓数据质量重要性
  - ✓元数据管理重要性
  - ✓IBM 解决方案（Information Server）

- 数据质量和元数据管理 –工具
  - ✓业务术语表（Business Glossary）
  - ✓洞察数据（Information Analyzer）
  - ✓规范开发 （Fasttrack）
  - ✓清洗/转换数据（ QualityStage/DataStage）
  - ✓元数据工作台 （Metadata Workbench)

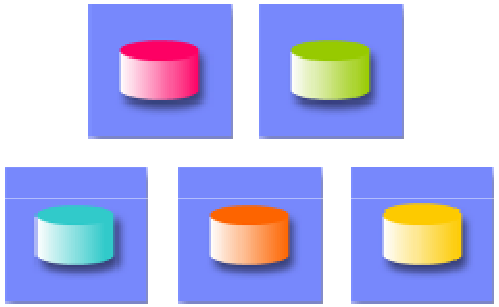
议程

- 数据治理中的数据质量和元数据
  - ✓数据质量重要性
  - ✓元数据管理重要性
  - ✓IBM 解决方案（Information Server）

- 数据质量和元数据管理 –工具
  - ✓业务术语表（Business Glossary）
  - ✓洞察数据（Information Analyzer）
  - ✓规范开发 （Fasttrack）
  - ✓清洗/转换数据（ QualityStage/DataStage）
  - ✓元数据工作台 （Metadata Workbench)

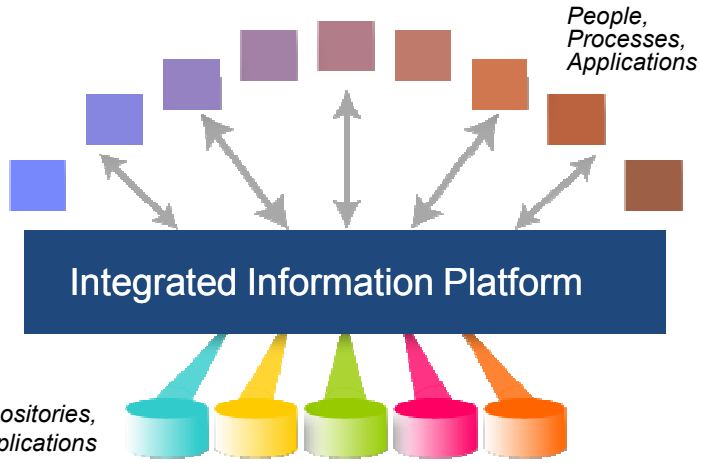
日新月异信息架构

相互没有关联的信息孤岛



Rich Standards, Flexible Architecture

动态提供整合的信息



70% of people's time can be spent finding relevant information

60%+ of CEOs say they need to do a better job leveraging information

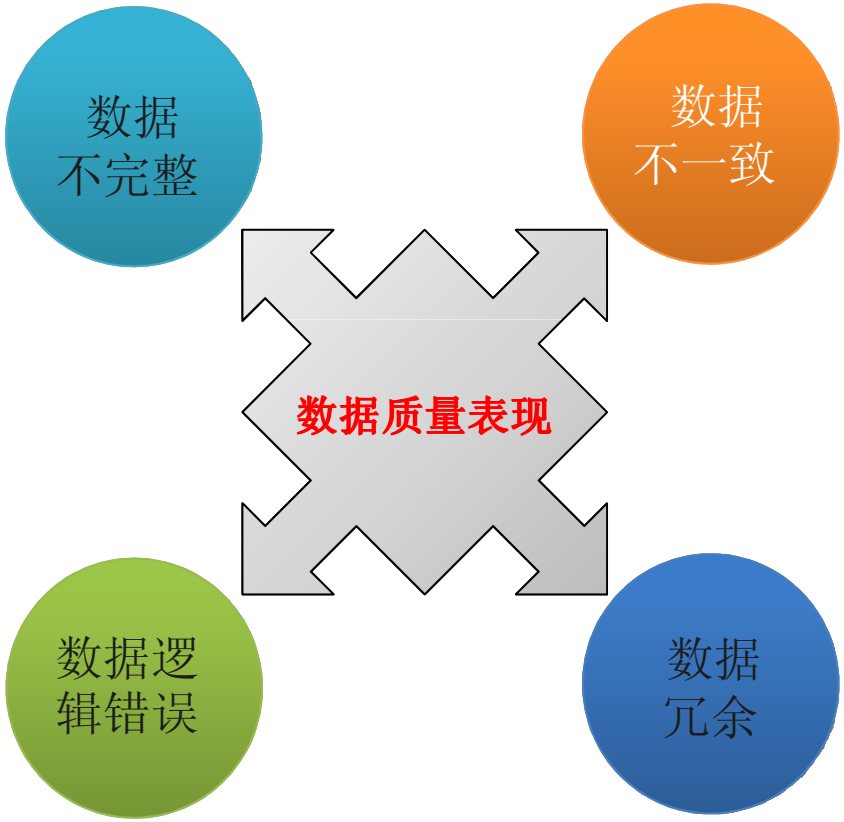
5X More **Value** creation by organizations effective at using information

Sources: IBM Attributes & Capabilities Study, 2005; Client Interviews 2004; IBM CFO Study, 2006

数据治理中遇到的问题

数据不完整

- 关键ID缺少，或者明显位数不符；
- 部分辅助信息的代码不规范很多是文本描述；
- 历史数据保留期限不一致。



数据不一致

- 相关联业务系统数据不同步；
- 不同系统之间描述同一业务问题的数据定义存在差异。

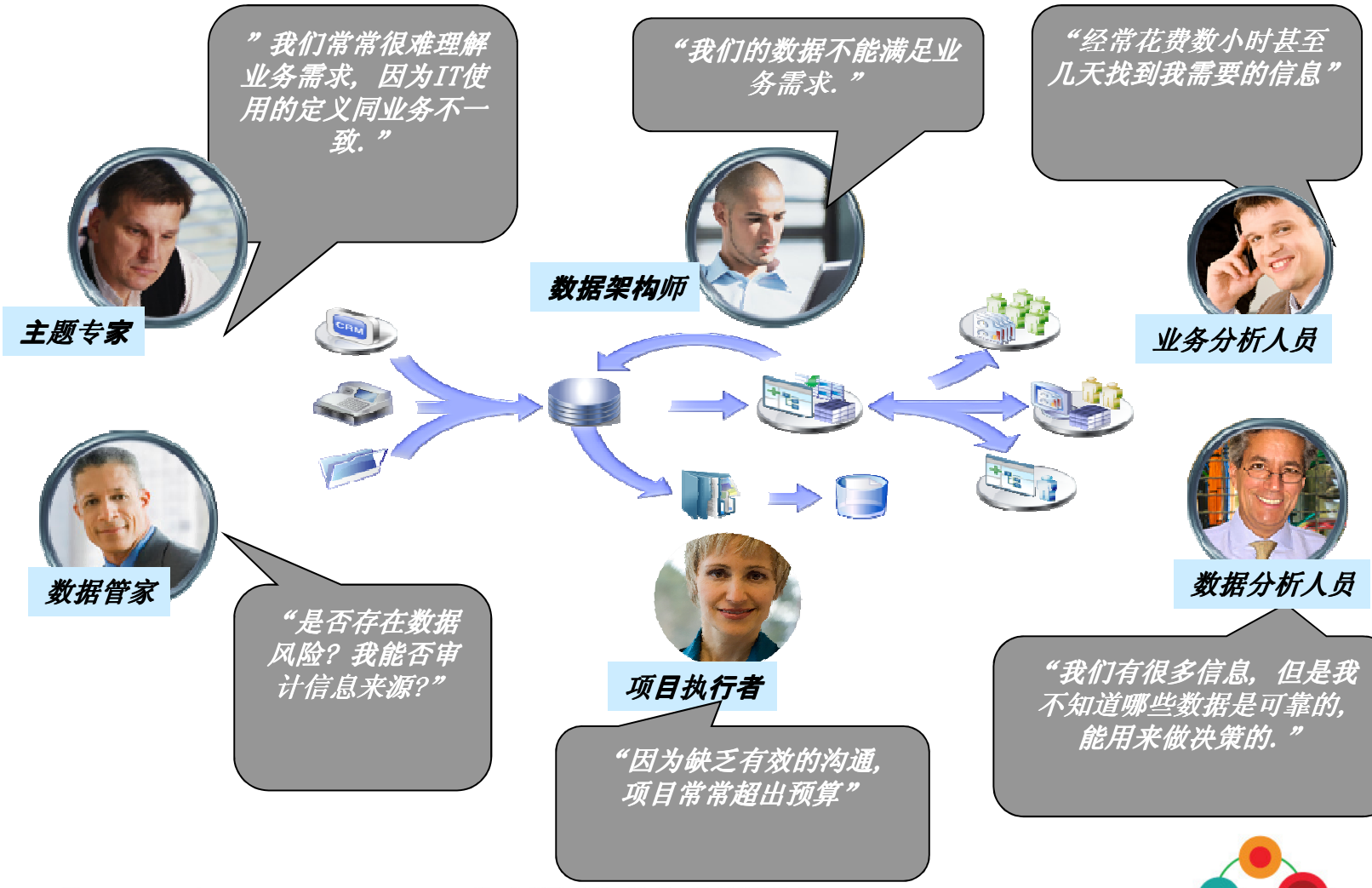
数据逻辑错误

- 违反业务规则
- 违反业务代码定义

数据冗余

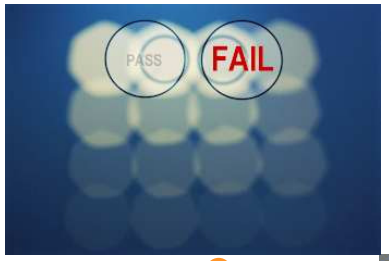
- 重复数据记录
- 非法键值

数据治理中遇到的问题


”我们常常很难理解业务需求，因为IT使用的定义同业务不一致.”
主题专家
“我们的数据不能满足业务需求.”
数据架构师
“经常花费数小时甚至几天找到我需要的信息”
业务分析人员
数据管家
“是否存在数据风险？我能否审计信息来源?”
项目执行者
“因为缺乏有效的沟通，项目常常超出预算”
数据分析人员
“我们有很多信息，但是我不知道哪些数据是可靠的，能用来做决策的.”

危害

83% 数据集成项目
需要重复实施甚至失败



错误或不完整数据导致
BI和CRM系统 不能正常
发挥优势甚至失效

25% 时间浪费在
辨别数据是否“坏数据 ”

无效和重复性工
作增加运作成本

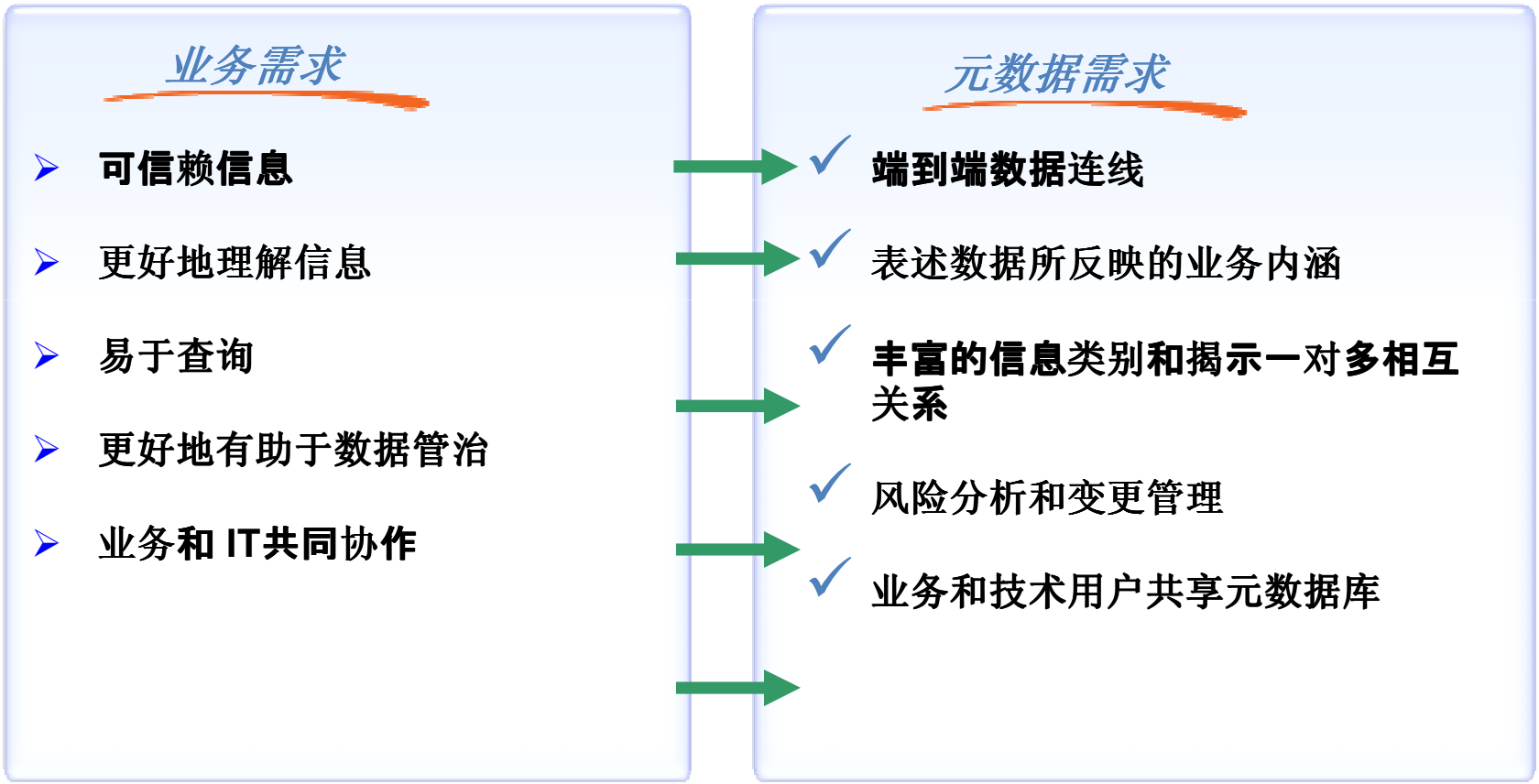
消费者缺乏信心

痛失商机

低劣数据质量严重地降低
公司年收入

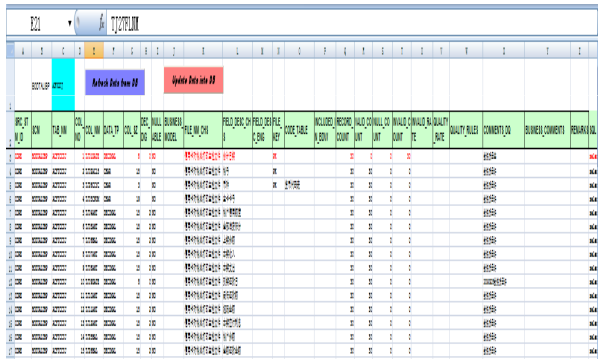
无法预测商机而造成损失，比事后
弥补将多达 10~100 倍

数据治理需求


业务需求
可信赖信息
更好地理解信息
易于查询
更好地有助于数据管治
业务和 IT共同协作
元数据需求
端到端数据连线
表述数据所反映的业务内涵
丰富的信息类别和揭示一对多相互关系
风险分析和变更管理
业务和技术用户共享元数据库

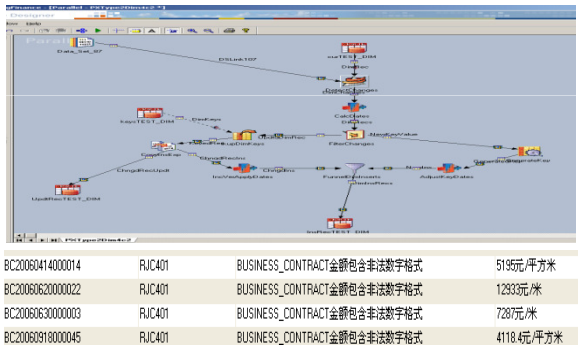
高效与完善的数据质量实施和管理体系

Data Profile



- Foundation to any success business projects
- Source System analysis
- Auditable Data Quality

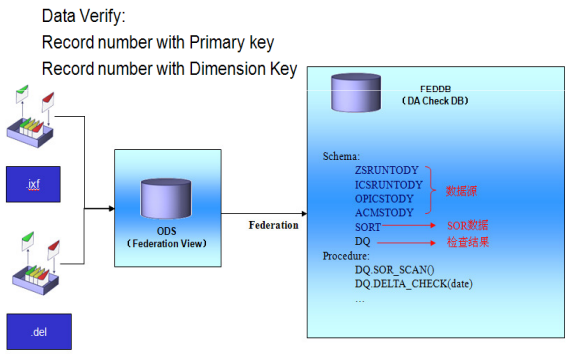
Data Clean



BC20060414000014	RJC401	BUSINESS_CONTRACT金额包含非法数字格式	5195元/平方米
BC20060620000022	RJC401	BUSINESS_CONTRACT金额包含非法数字格式	1293元/米
BC20060630000003	RJC401	BUSINESS_CONTRACT金额包含非法数字格式	7287元/米
BC20060918000045	RJC401	BUSINESS_CONTRACT金额包含非法数字格式	4118.4元/平方米

- DC rules will be stored in MDR
- DC rules will be implement in ETL
- DC result will be record

Data Verify



- Set up Data source view by federation
- Auto generate verify SP with MDR
- Verify result will be store with MDR

数据质量管理平台


业务用户


主题专家


架构师


数据分析师


开发人员


DBAs

IBM Information Server

Information Services Director
为整合信息和访问发布SOA服务

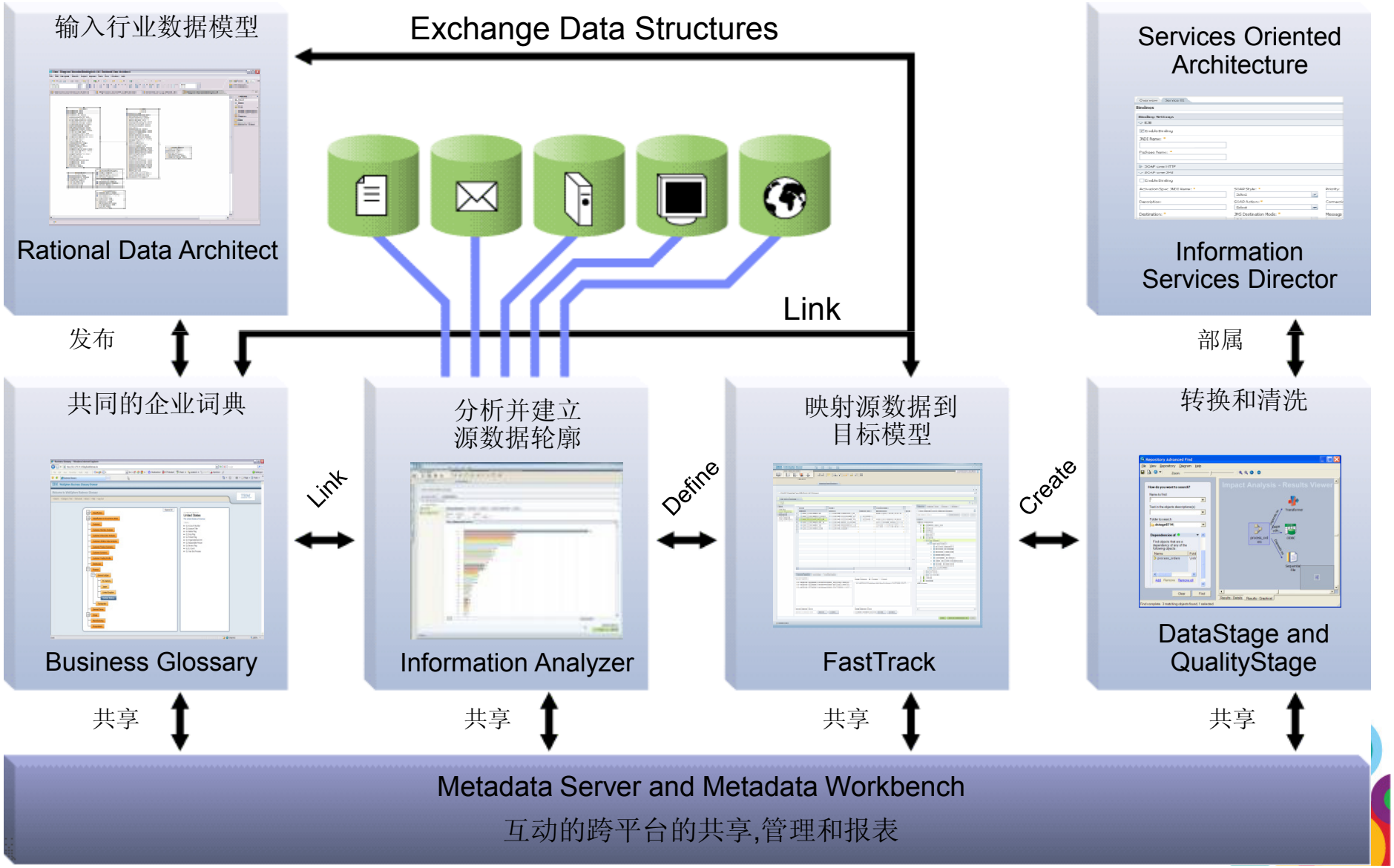
Business Glossary 归档业务术语 & 连接到数据源	QualityStage 标准化、合并和纠正信息	DataStage 组合和重构信息以用于新的用途	Federation Server 异构信息的虚拟化访问
Information Analyzer 分析 & 理解 源数据			Change Data Capture 数据增量获取

Metadata Server / Metadata Workbench
跨信息整合生命周期的统一的元数据管理

并行处理

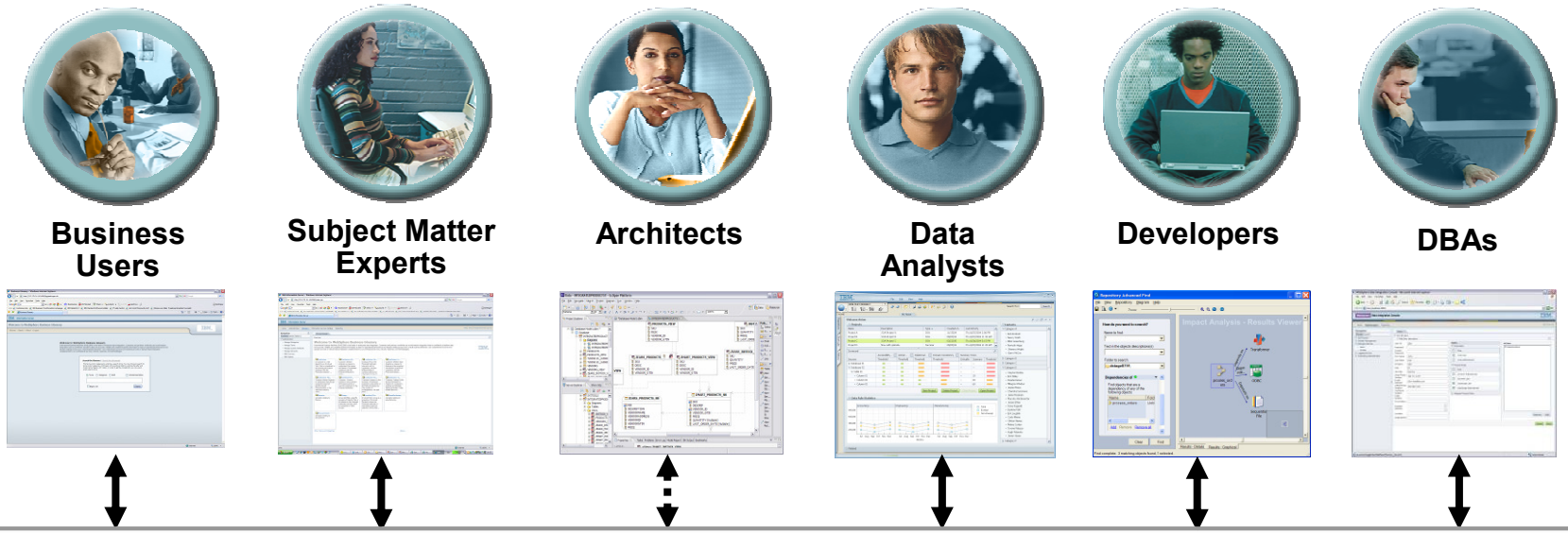
丰富的应用、数据和内容连接的支持

基于Information Server的数据管治架构图



基于Information Server的元数据管理架构图

加强协作，让IT向业务看齐



IBM 元数据管理目标

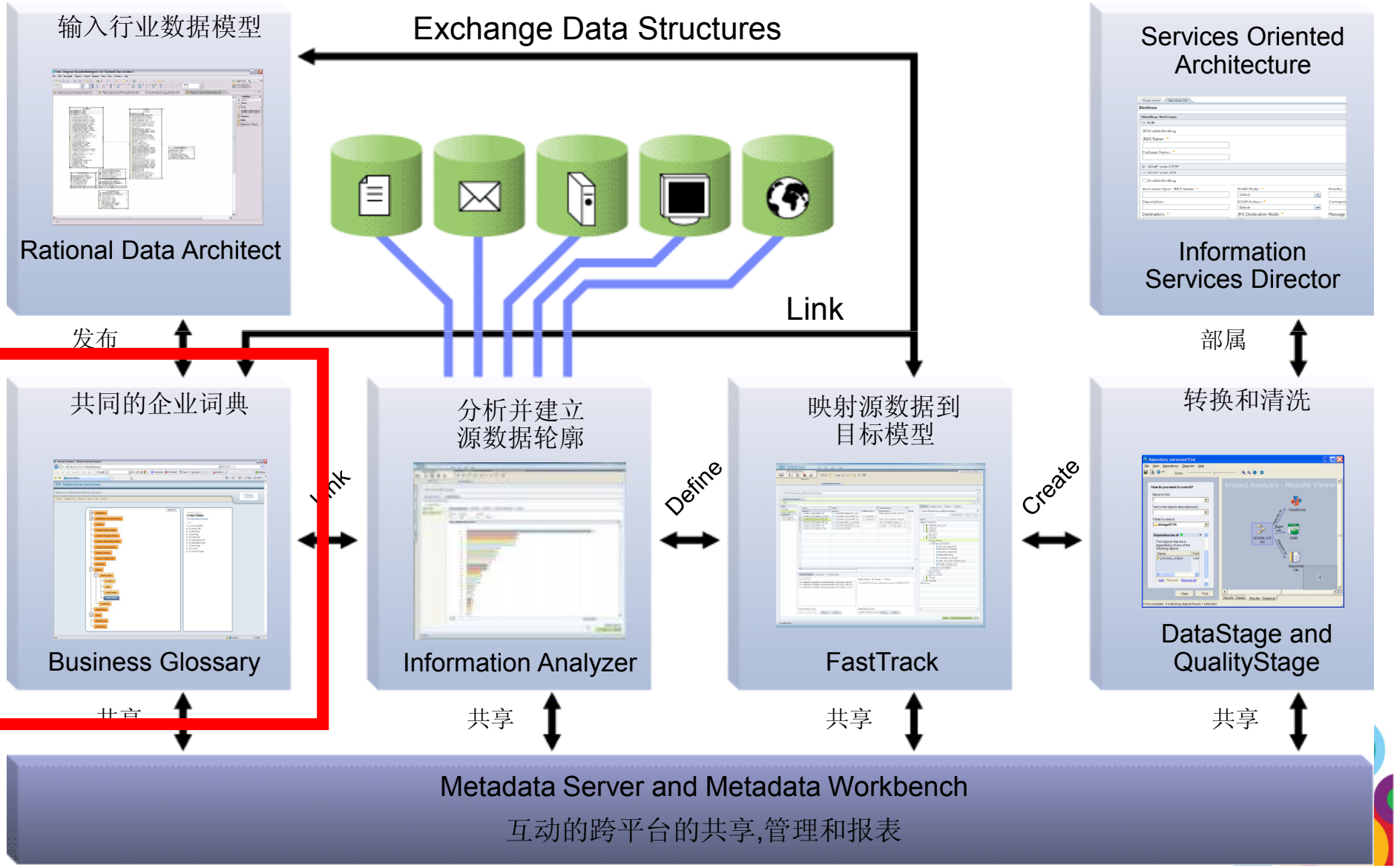
- 集合各个产品元数据管理到一个单一的，共享的元数据管理库中。
- 消除了元数据在不同工具之间的交换需要。
- 通过“统一模型”提供连续的元数据管理视图。
- 提供开放的体系架构允许额外的组件方便的扩展。
- 提供元数据管理的新的目标集合。
- 允许通过整个套件共享元数据分析。

议程

- 数据治理中的数据质量和元数据
  - ✓数据质量重要性
  - ✓元数据管理重要性
  - ✓IBM 解决方案（Information Server）

- 数据质量和元数据管理 –工具
  - ✓业务术语表（Business Glossary）
  - ✓洞察数据（Information Analyzer）
  - ✓规范开发 （Fasttrack）
  - ✓清洗/转换数据（ QualityStage/DataStage）
  - ✓元数据工作台 （Metadata Workbench)

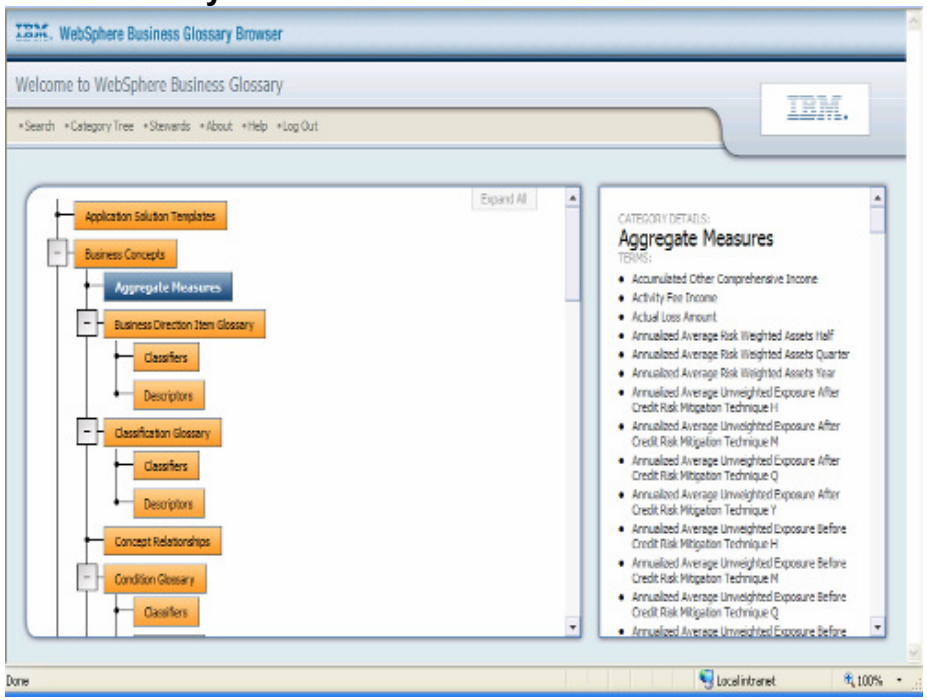
基于Information Server的数据管治架构图



InfoSphere Business Glossary

Business Glossary

创建和管理业务词典和层级关系, 及相关的物理信息源



需求

- 获取业务数据和类别
- 连接业务术语&类别到IT资产
- 识别数据管理员和类别访问管理

益处

- 信息的内容对每个人都是可以立即了解到的
- IT 项目同数据监管结合
- 促进业务和IT的紧密协作

创建共享的业务词汇表

Database = DB2

Schema = NAACCT

Table = DLYTRANS

Column = TAXVL

data type = Decimal (14,2)

Derivation: SUM(TRNTXAMT)

Category: Costs

Term: Tax Expense

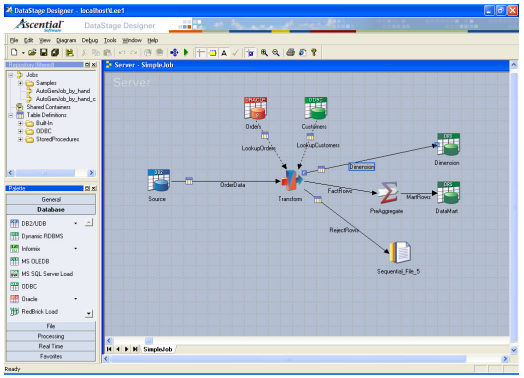
Full Name: Tax to be paid on Gross Income

“The expense due to taxes …..”

(John Walsh is responsible for updates. 90% reliable source)

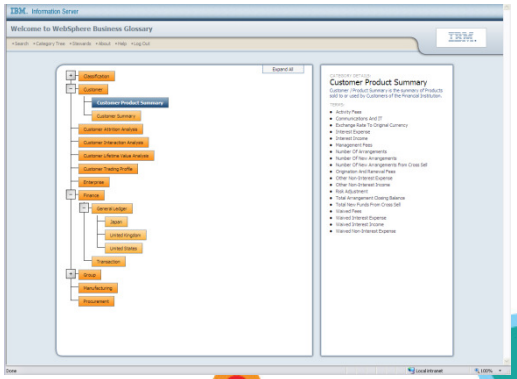
Status: CURRENT

Achieve a common vocabulary between business & technical users!



InfoSphere DataStage

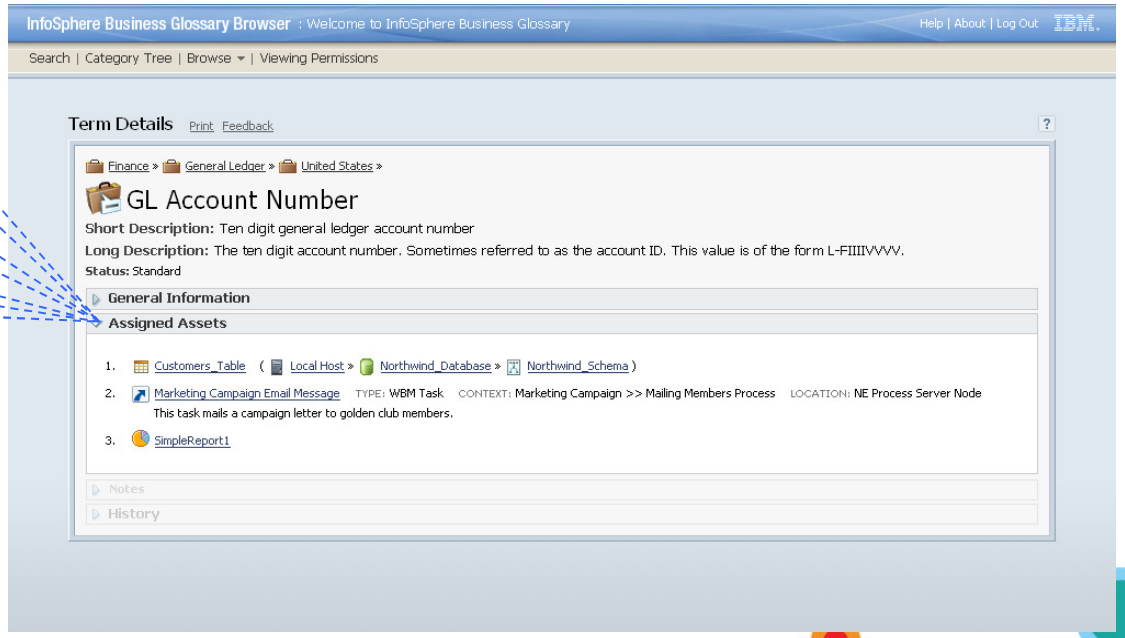
Shared Metadata Server & Repository



InfoSphere Business Glossary

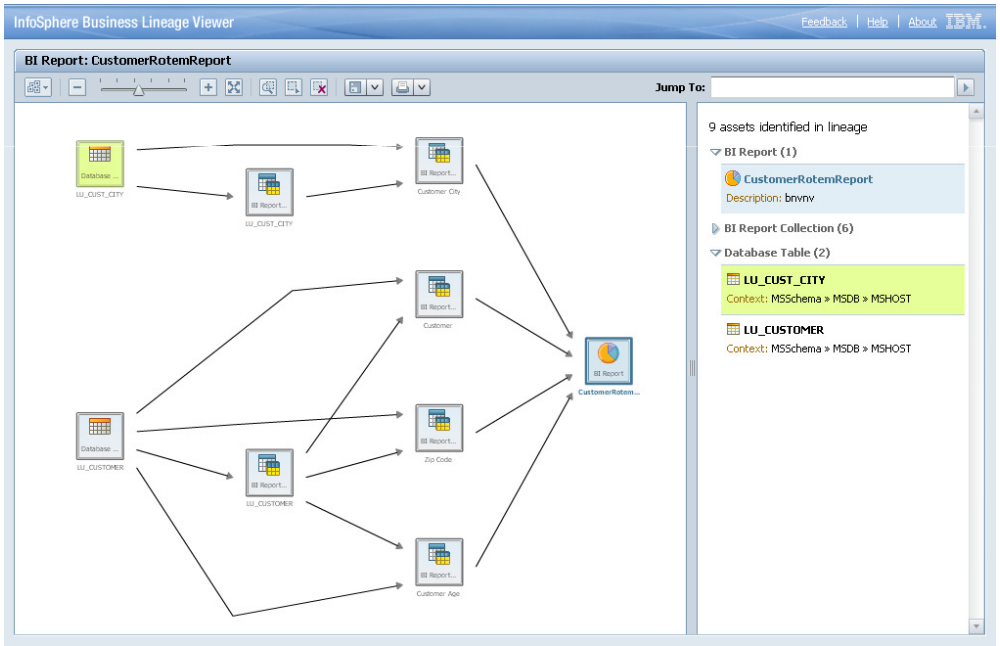
每一IT资产的业务含义

- 设计**信息监管项目的出发点**
- **“业务用户”** 通往异构IT环境的桥梁
  - ✓*数据源*
  - ✓*ETL 任务*
  - ✓*存储过程*
  - ✓*应用*
  - ✓*BI 报表*
  - ✓*业务流程*
  - ✓*数据模型*
  - ✓*Web Services*
  - ✓*更多…*



Business Lineage业务世系– 可信任的信息来源

- **给予“业务用户”** 对信息的信心和信任, 从而进行重要决策的能力
- 快速**理解**信息**来源哪里**
- **减少**跨不同部分或团队确认数据准确的**周期**



易于理解和使用,
无需培训!

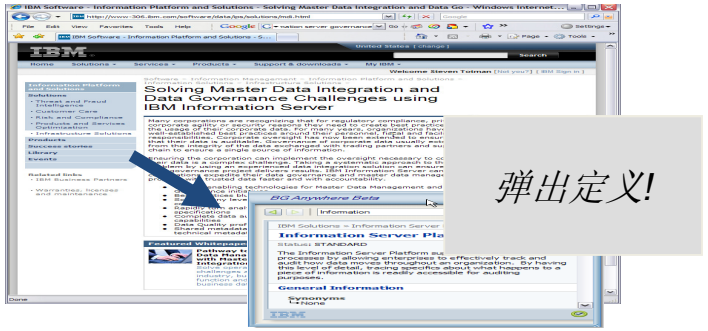
*只包含业务人员关注的**信息的关键路径***

InfoSphere Business Glossary Anywhere

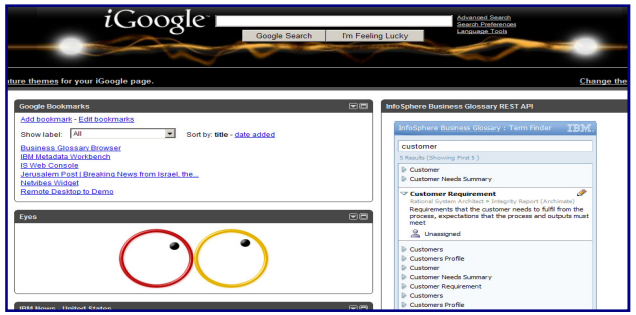
查看术语内容,全面理解信息含义

任何用户

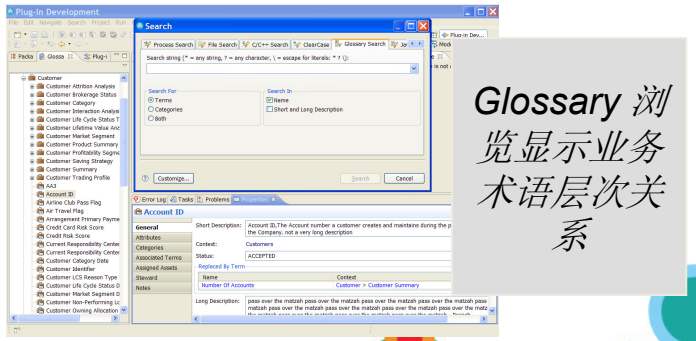
- **从任何应用实时访问**Business Glossary
  - 例如. BI报表



- **零足迹(Zero-footprint)**REST API,可以将 Business Glossary直接嵌入任何应用

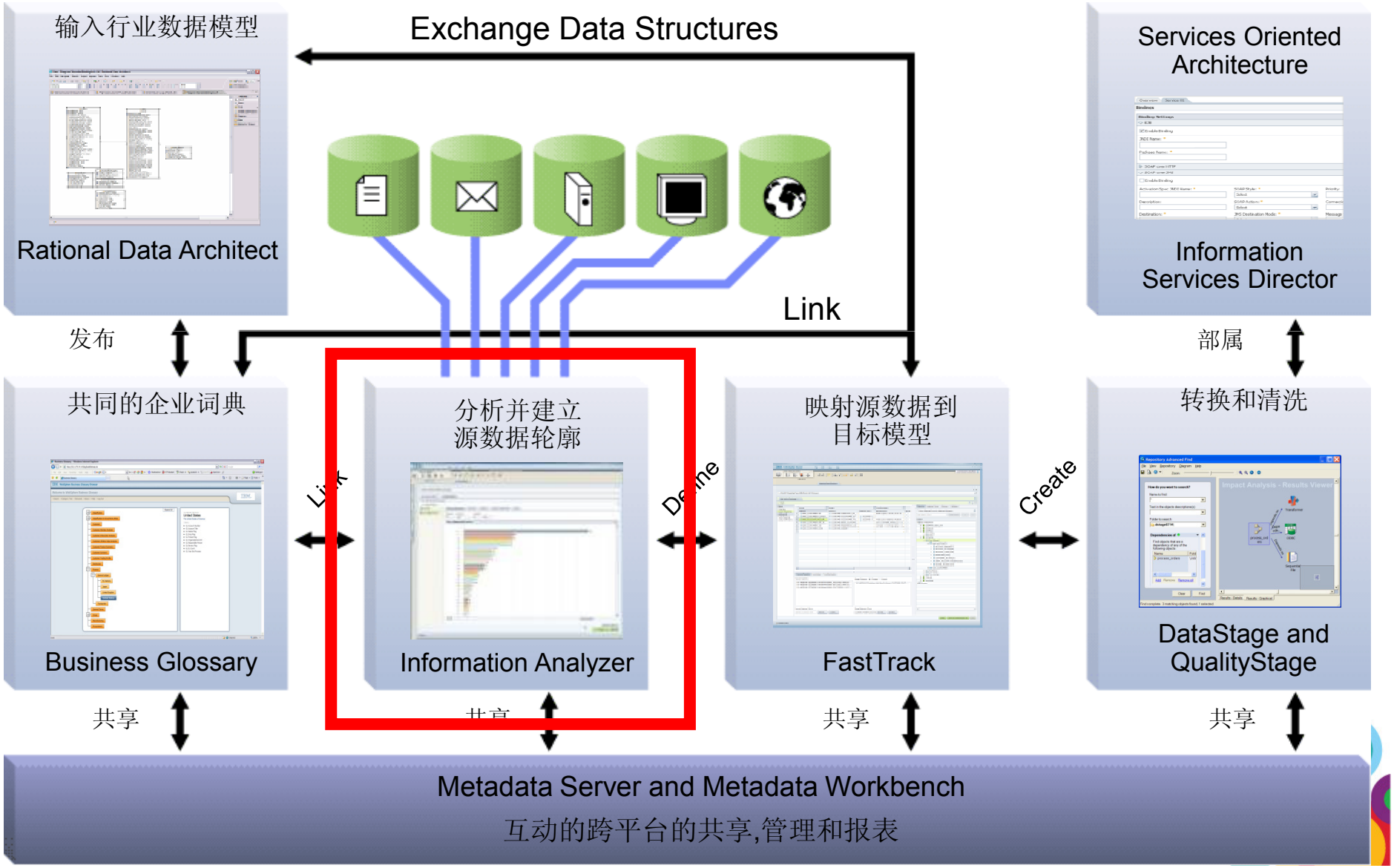


- 在**Eclipse项目**中访问**授权信息源**



Glossary 浏览显示业务术语层次关系

基于Information Server的数据管治架构图



Information Analyzer

What is it?

用于企业数据源的数据剖析、分析和监控工具

- 数据剖析
- 数据质量监控

What does it do?

分析数据源，揭示信息的结构、内容和质量

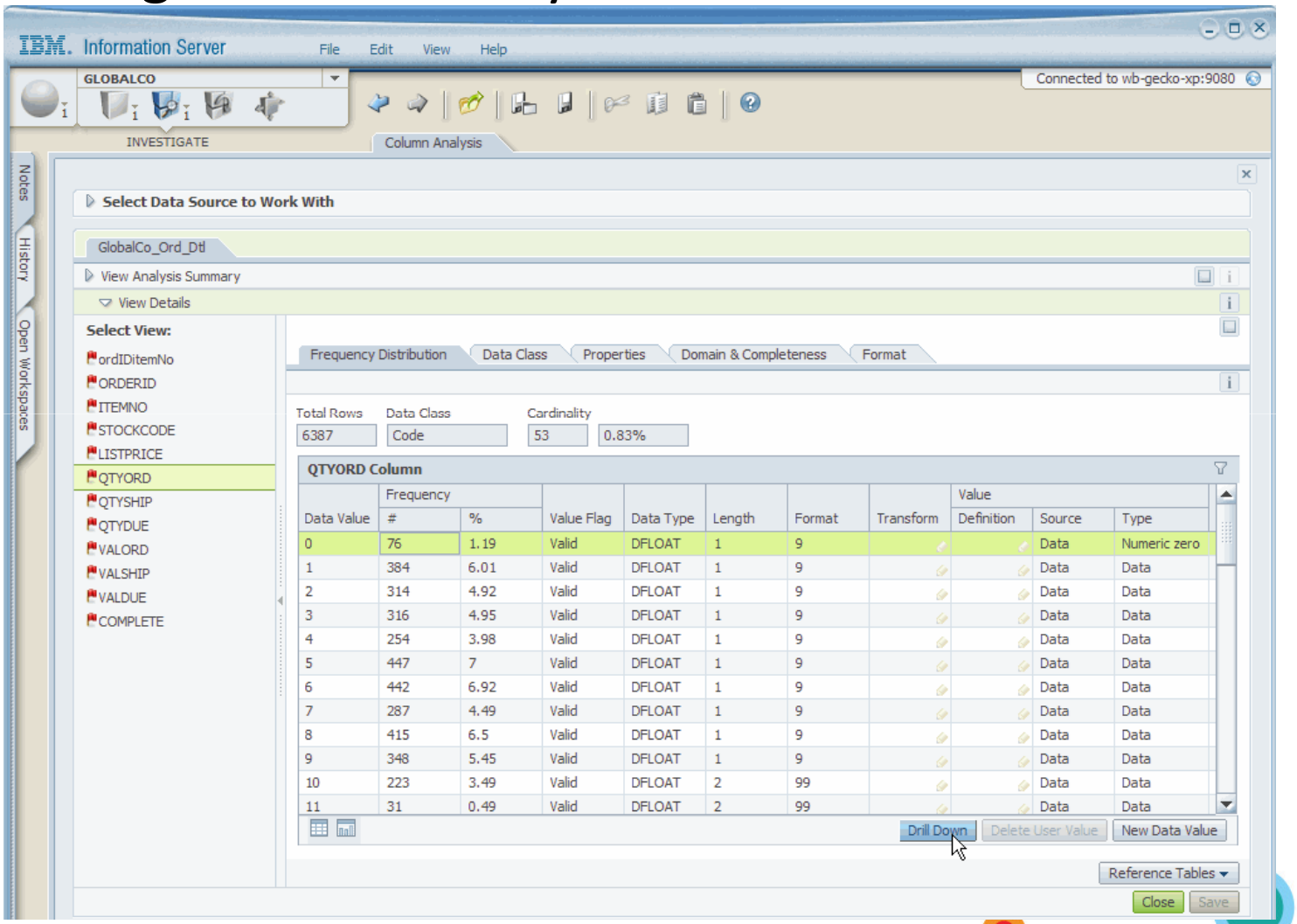
- 发现缺失、不准确和不一致的数据
- 贯穿生命周期监控数据质量

Who uses it?

商业和数据分析员、数据质量专员、数据架构师和数据管理员、数据集成项目领导和实施人员。

Data Profiling: Column Analysis （字段分析）

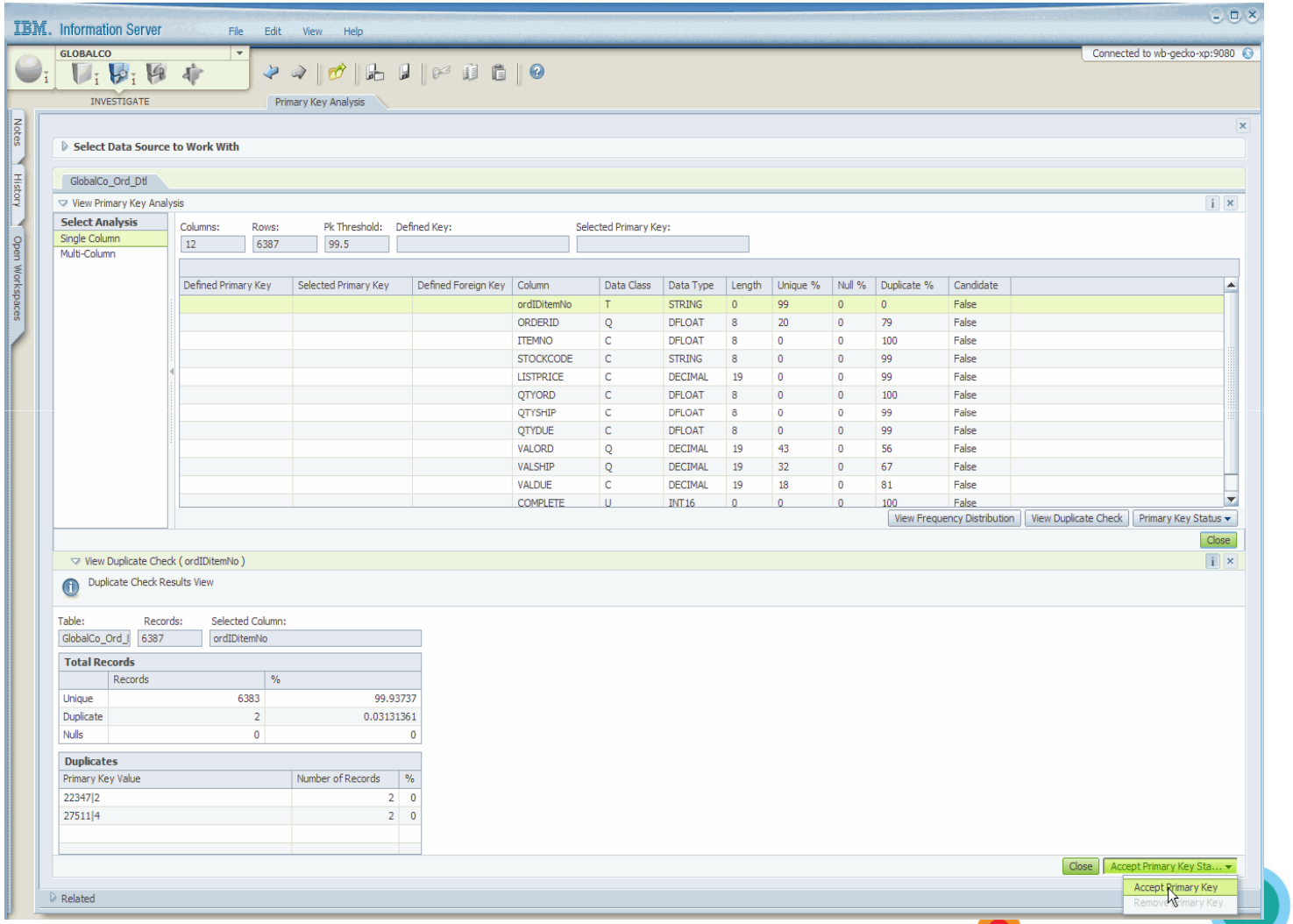
- Domain Values & Validation
- Data Classification
- Data Properties
- Formats



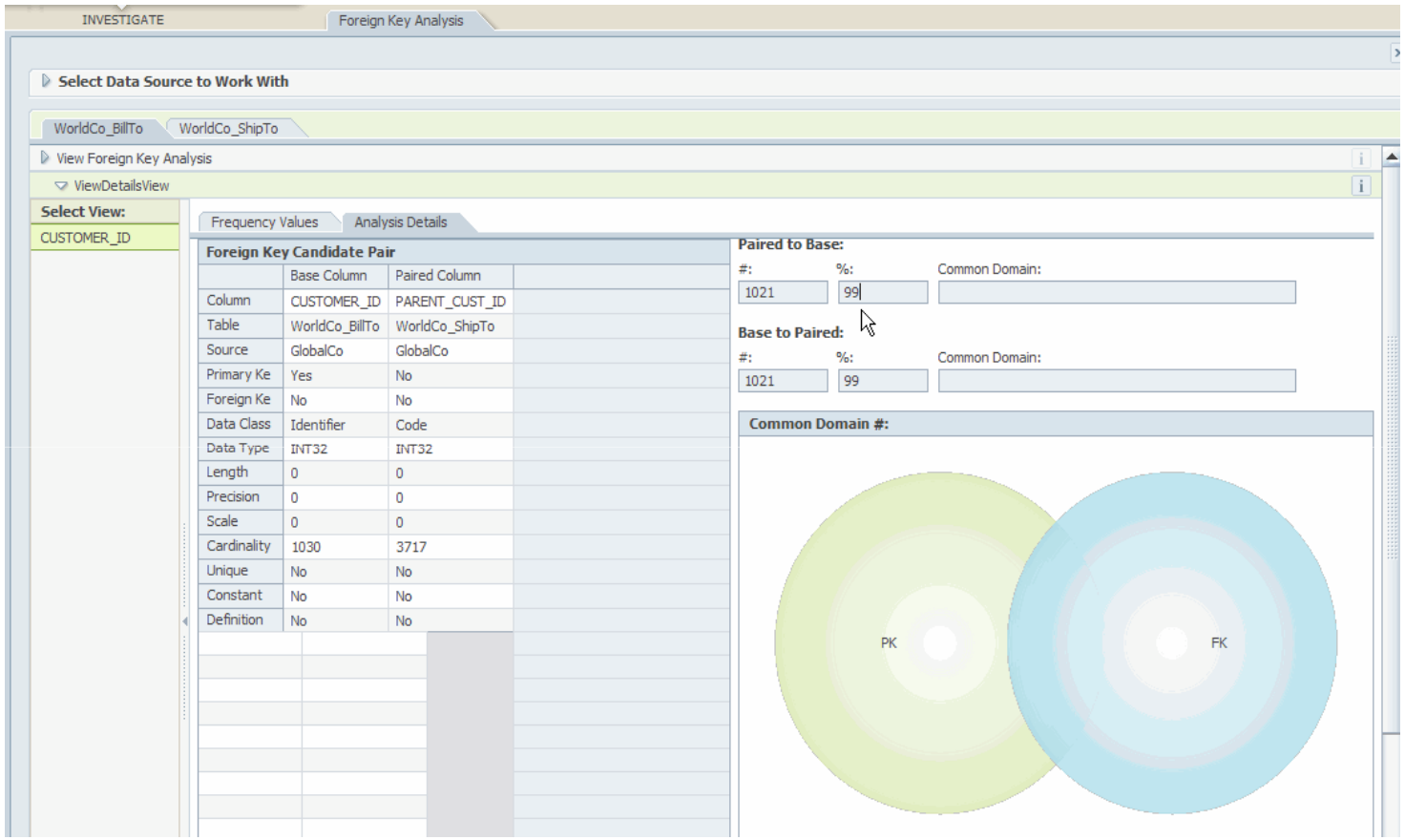
Data Value	Frequency #	Frequency %	Value Flag	Data Type	Length	Format	Transform	Value Definition	Value Source	Value Type
0	76	1.19	Valid	DFLOAT	1	9			Data	Numeric zero
1	384	6.01	Valid	DFLOAT	1	9			Data	Data
2	314	4.92	Valid	DFLOAT	1	9			Data	Data
3	316	4.95	Valid	DFLOAT	1	9			Data	Data
4	254	3.98	Valid	DFLOAT	1	9			Data	Data
5	447	7	Valid	DFLOAT	1	9			Data	Data
6	442	6.92	Valid	DFLOAT	1	9			Data	Data
7	287	4.49	Valid	DFLOAT	1	9			Data	Data
8	415	6.5	Valid	DFLOAT	1	9			Data	Data
9	348	5.45	Valid	DFLOAT	1	9			Data	Data
10	223	3.49	Valid	DFLOAT	2	99			Data	Data
11	31	0.49	Valid	DFLOAT	2	99			Data	Data

Data Profiling: Table Analysis （表分析）

- Primary Keys (single or multi-column)
- Key Duplicates



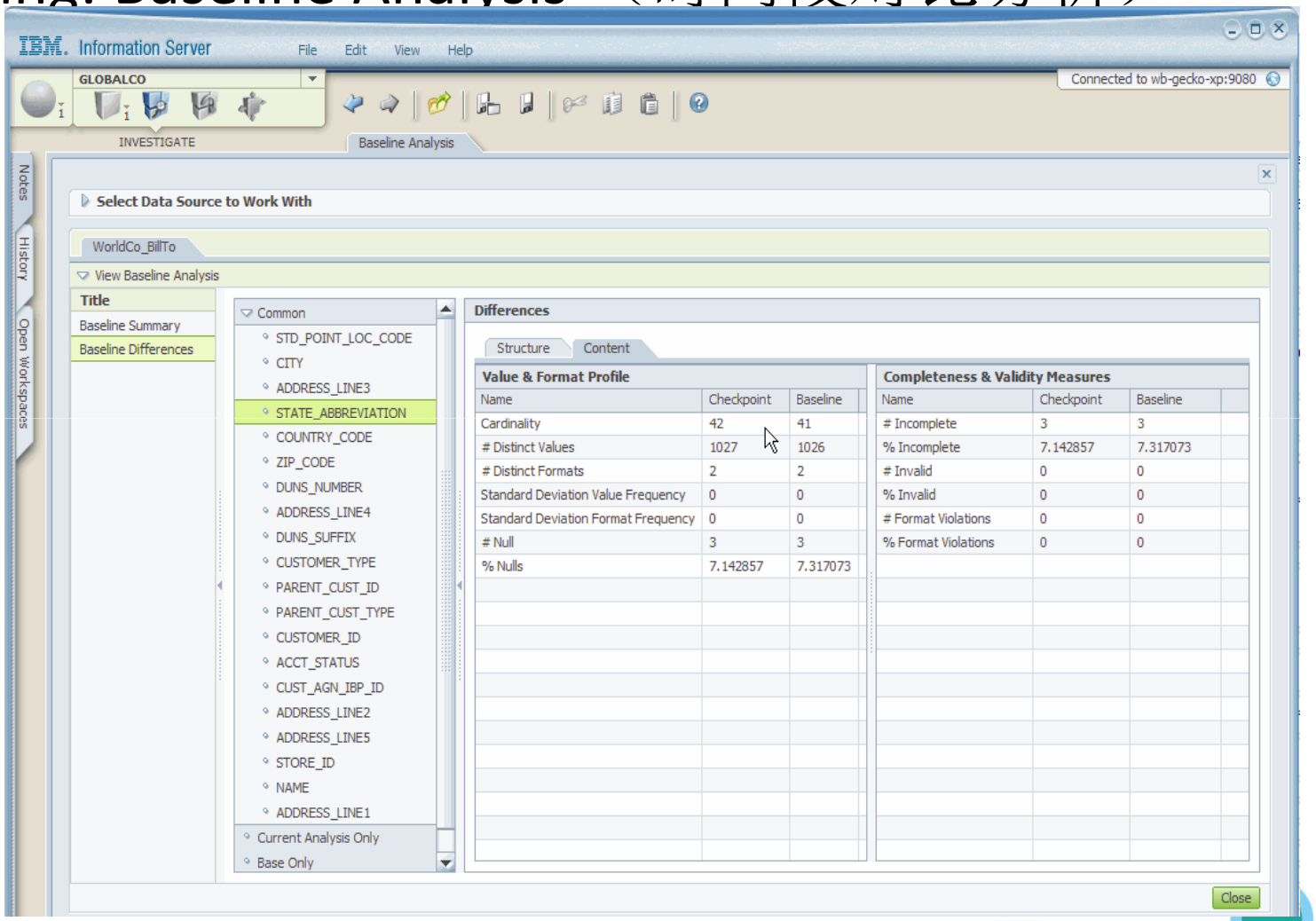
Data Profiling: Cross Table Analysis （跨域分析）



- Foreign Key Relationships
- Referential Integrity
- Cross-Domain Relationships
- Data Redundancy

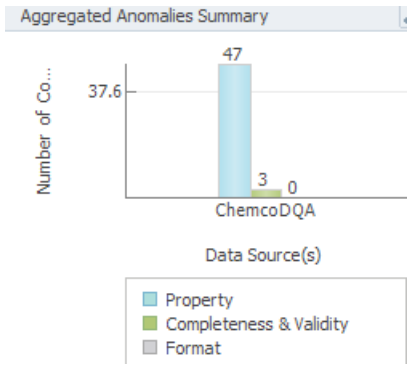
Data Profiling: Baseline Analysis （时间段对比分析）

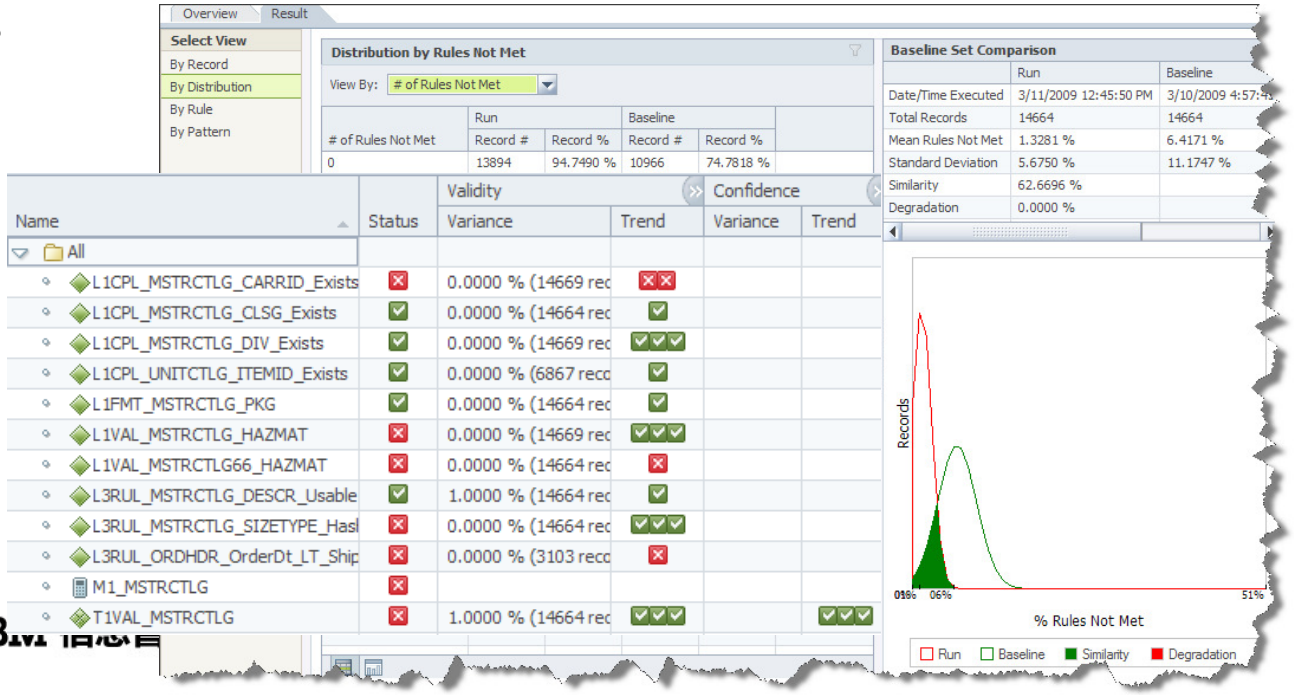
- Current-to-Prior Comparison
- Content & Structural Variation



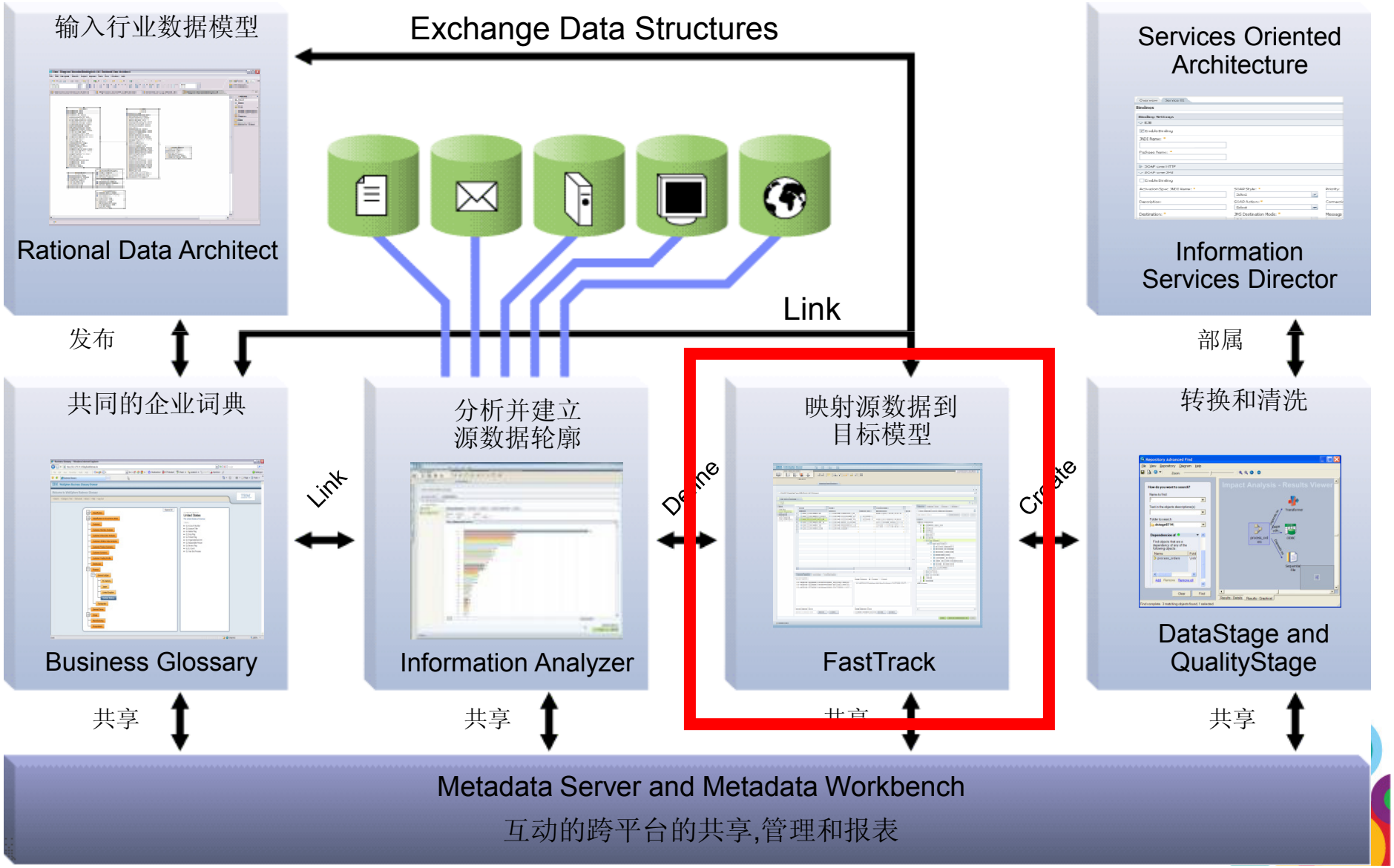
基于业务规则的数据质量分析

- ***结合业务规则作分析*** 建立关键数据规则，以为开发、部署和评估提供依据。
- ***以业务为动力的规则定义*** 可定义/修改/重用的规则适用多个数据源，提高使用价值 。
- ***整体上多种-层面的规则评估*** 以规则、记录和数据源等层面上洞察潜在的数据质量问题。
- ***仪表盘和报告*** 支持规则分析。
- ***轮廓分析增强功能*** 支持对flat file文件定义，指定目标数据细分内容进行分析，和执行作业调度运行。





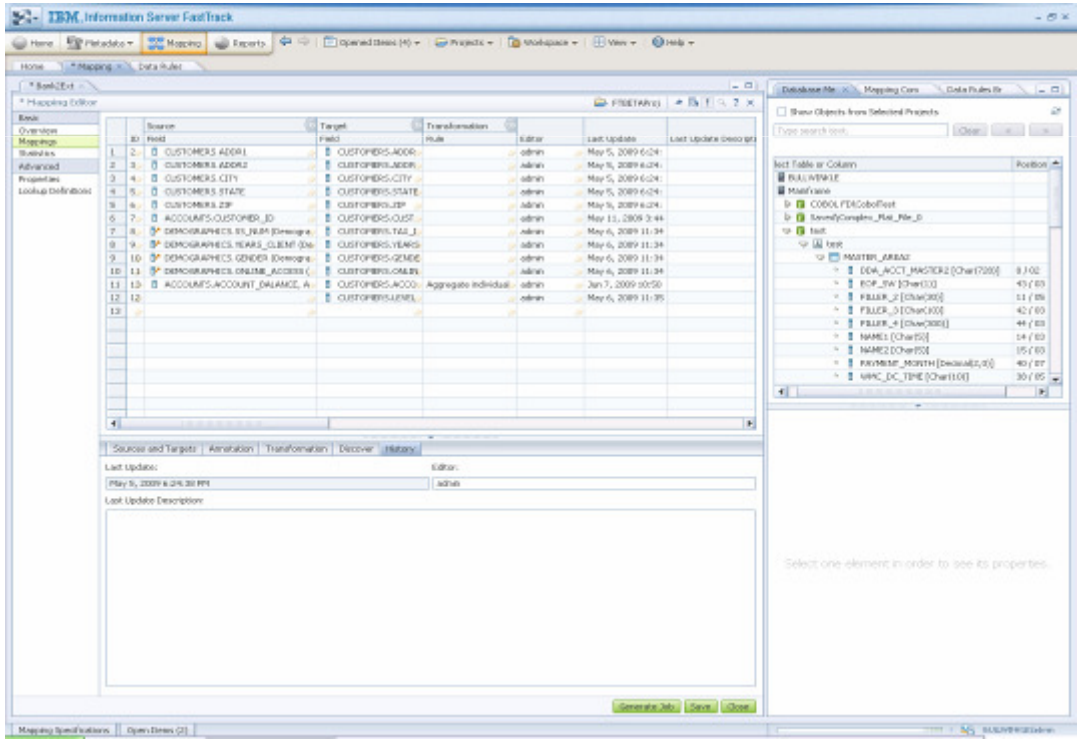
基于Information Server的数据管治架构图



InfoSphere FastTrack

FastTrack

获取设计规范书并且加快从规范书到数据集成项目的转换



功能

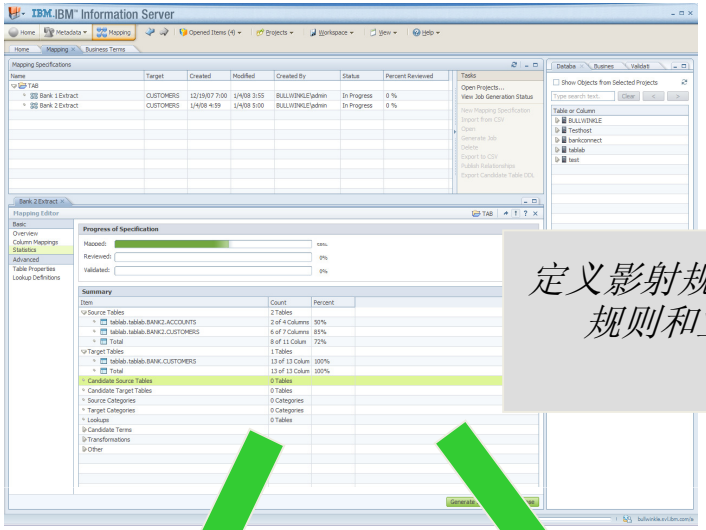
- 获取业务需求用以数据源到目标的影射
- 额外提供数据源的分析和对应的业务术语
- 自动生成ETL样本作业

益处

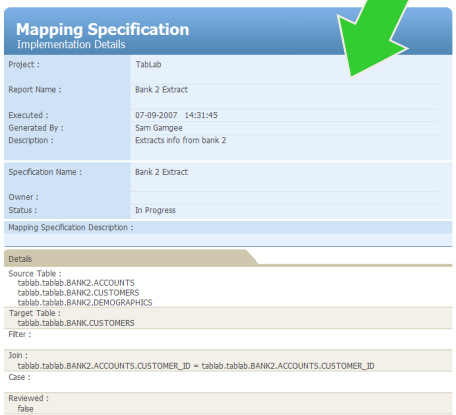
- 加快数据集成的开发
- 集中管理客户的需求规范
- 可在日后审计设计决策

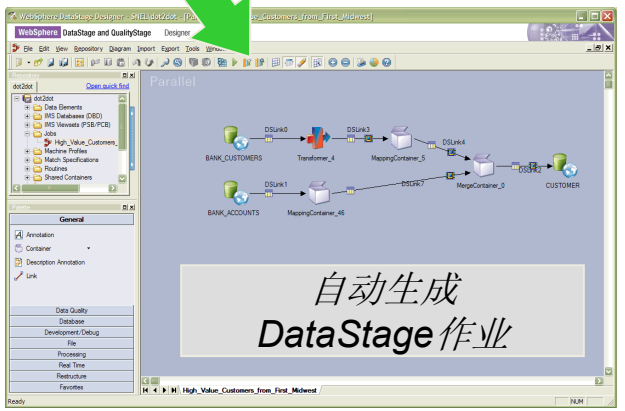
跟踪从业务需求到应用开发的过程

- 单一，集中管理的架构跟踪从业务需求到开发的整个过程
- 可以输入Excel格式的影射规范文件
- 可以定义业务术语并且将其连接到相应的物理元数据
- 自动生成DataStage作业并且将复杂的转换的描述通知开发员
- 生成历史文档用于审计



定义影射规范以及业务规则和业务术语



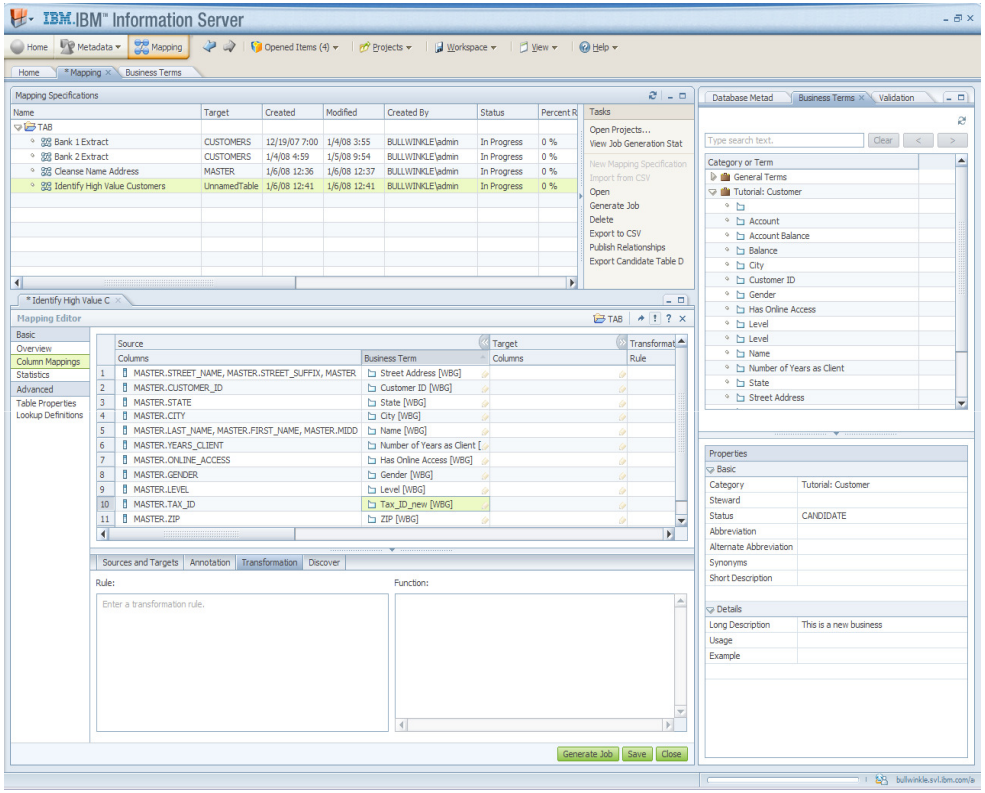


自动生成 DataStage作业

灵活的报表功能和审计

InfoSphere FastTrack – 业务的联系

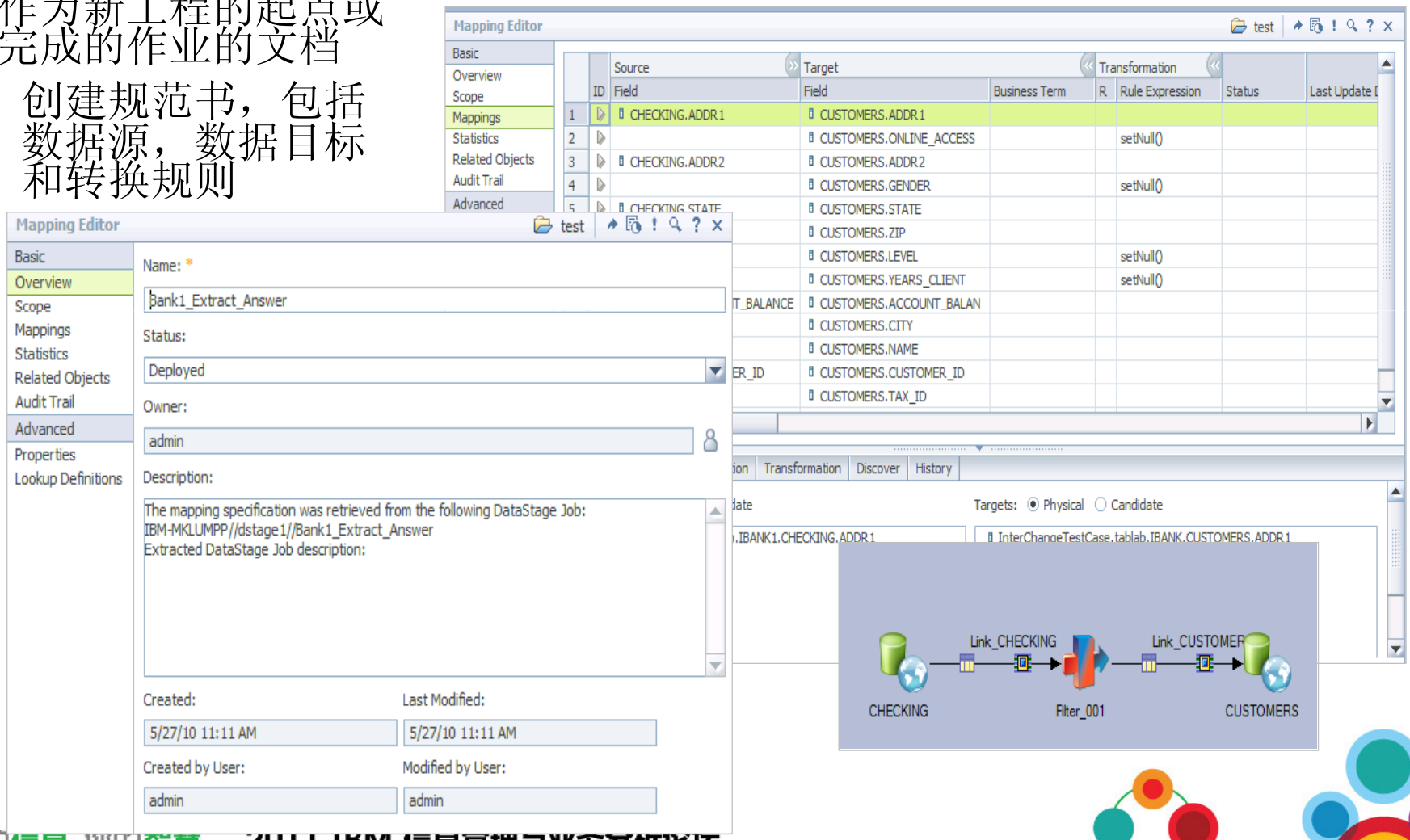
- 通过利用已存在或生成新的业务术语，并将其联系到相应的物理元数据，用以增强业务规范
- 通过将数据源影射到业务术语以生成新的目标表和DDL



提供集中的，可管理的业务方法论，项目规范书和需求以及它们之间的联系 支持企业的数据治理

InfoSphere FastTrack –*DataStage 作业的反向工程化*

- 可作为新工程的起点或已完成的作业的文档
  - 创建规范书，包括数据源，数据目标和转换规则

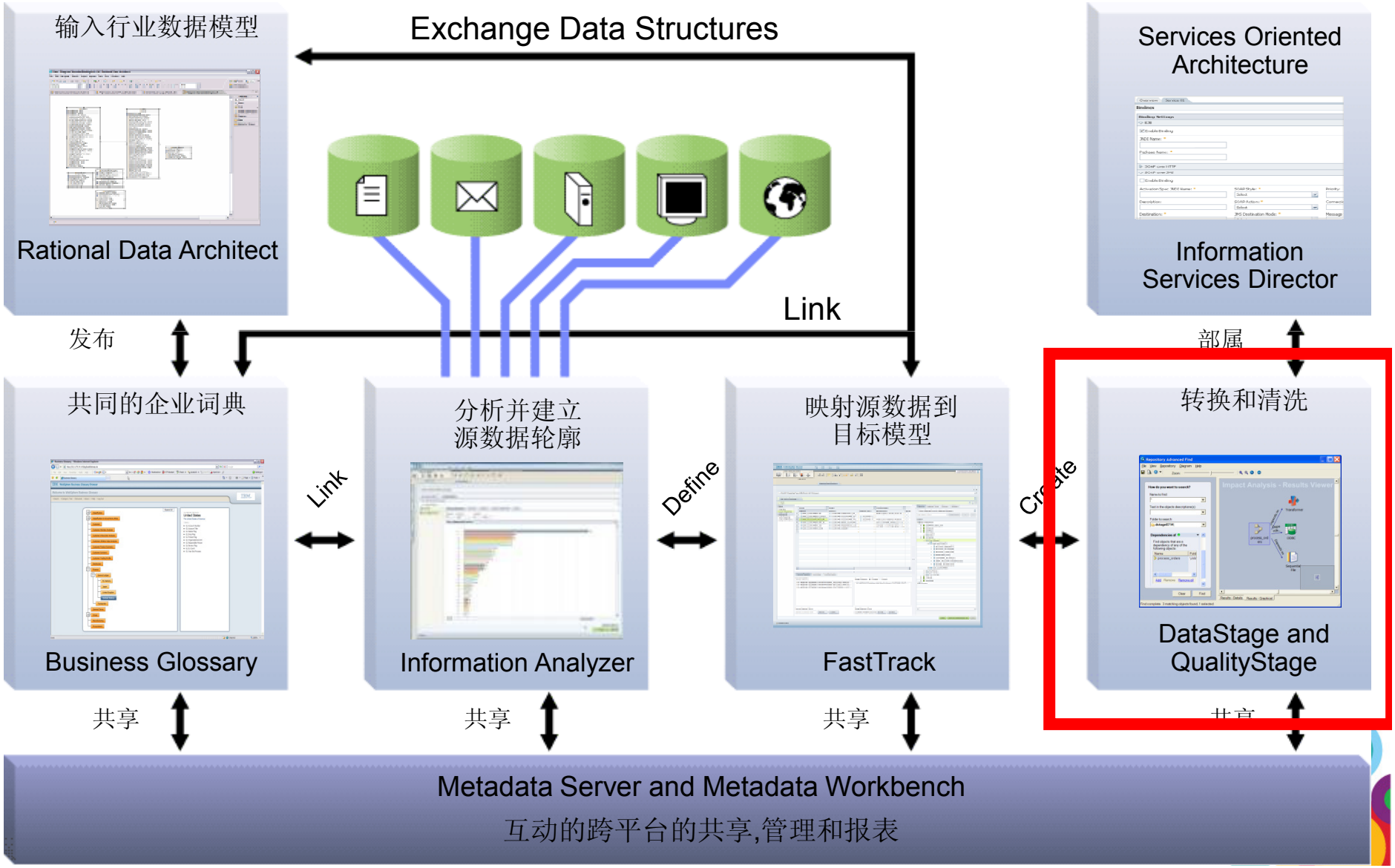


业务价值最大化

客户定义的在机构流程中FastTrack最主要的3项影响

快速实现价值	提高生产率	更强的数据治理
• 通用的格式 支持需求和文档 • 在项目规范中去除歧义 • 跟踪历史规范用语审计 • 改善团队成员之间的沟通	• 通过转换业务逻辑到代码直接开始ETL • 同步不同开发团队的工作 • 提高off-shore ETL 开发模式的成功 • 在分析员和开发员之间共享通用的元数据，而不增加额外工作量	• 集中管理从项目开始到结束的生成，存储，维护和审计 • 加强了规范标准和文档化 • 揭示了转换规则和业务需求之间的关键联系 • 减少了Excel文件存储和隐藏的规则

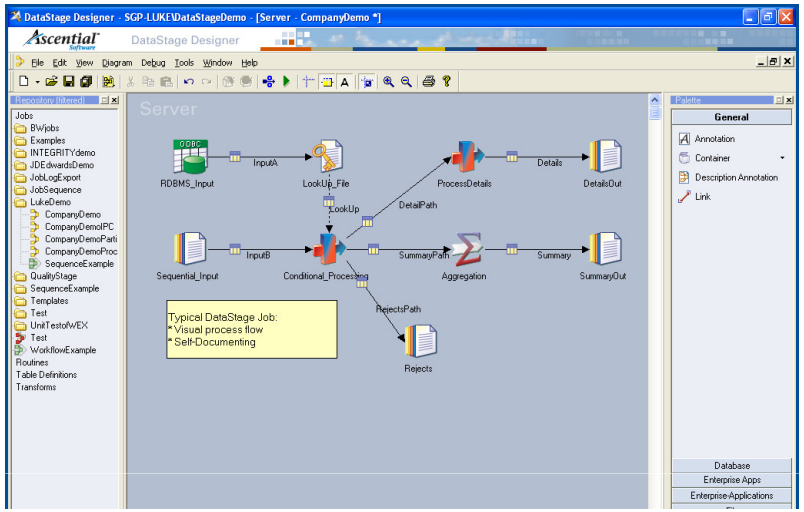
基于Information Server的数据管治架构图



数据处理流程设计

完全图形化的设计工具：

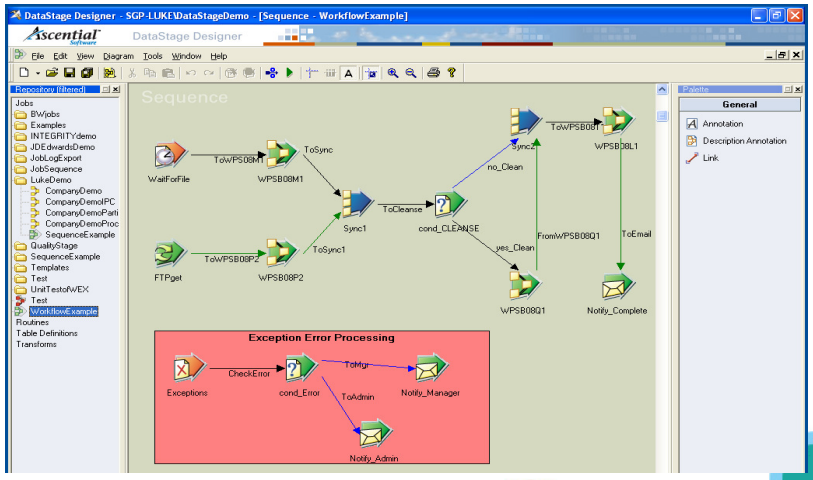
1. 作业容易开发、理解、调试和维护。
2. 强大的、被验证的最好的数据转移和抽取工具 。



易于设计、易于管理、易于维护

工作流程控制：

1. 图形化的工作流，而非代码。
2. 支持条件路径和错误处理
3. 支持EMAIL通知



数据清理： QualityStage

- 与 DataStage 无缝集成的专用数据质量功能
- 通过可视化的界面，定义复杂的匹配和留存逻辑
- 确保干净、标准化和不重复的信息
- 得到事实的单一版本
- 用于客户地址，电话号码，传真，电子邮箱等

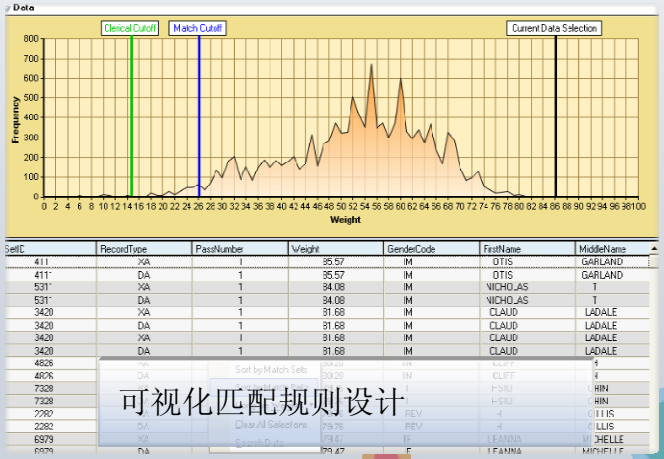
主题专家

数据分析师

清理

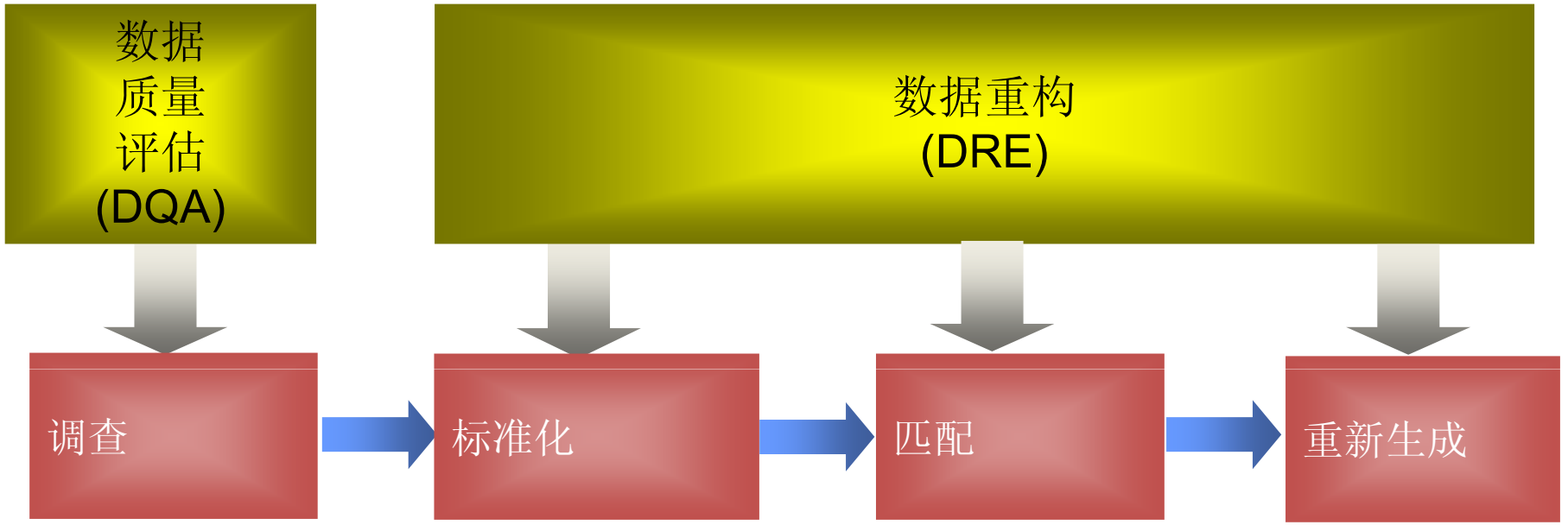
InfoSphere QualityStage™

标准化和纠正源数据字段，把不同来源的记录匹配在一起，以创建单一视图



可视化匹配规则设计

QualityStage的实施方法



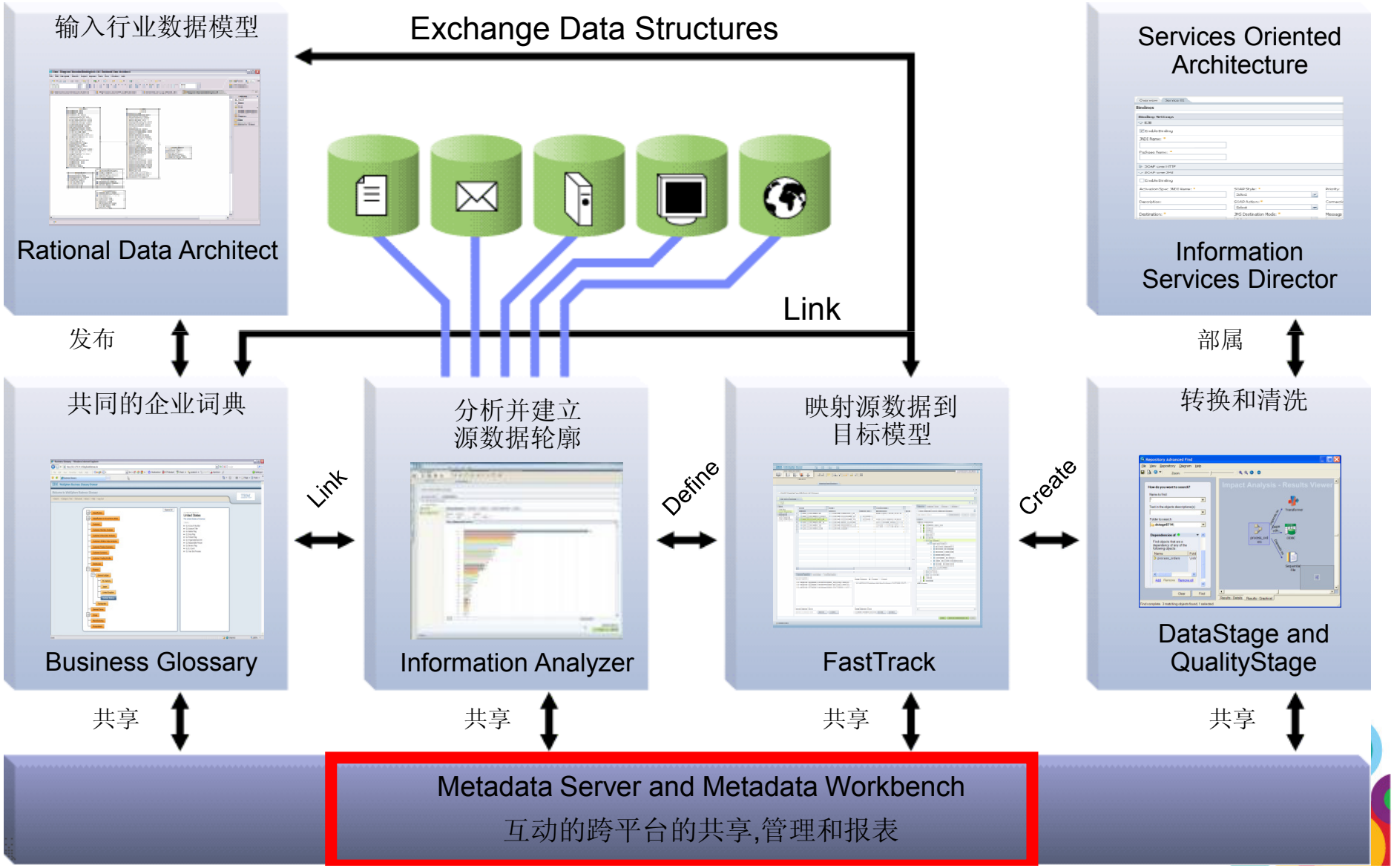
上海市南京西路1266号38楼02A室
上海南京西路1266号3802A室
上海市南京路（西）1266号38A02
南京西路1266/38/02A

上海市 | 南京西路 | 1266号 | 38楼 | 02A室
上海 市| 南京西路 | 1266号 | 38楼 | 02A室
上海市 | 南京西路 | 1266号 | 38楼 | 02A室
南京西路 | 1266号 | 38楼 | 02A室

上海市 | 南京西路 | 1266号 | 38楼 | 02A室
上海 | 南京西路 | 1266号 | 3802A室
上海市 | 南京路（西） | 1266号 | 3802A
南京西路 | 1266/ | 38/ | 02A

上海市南京西路1266号38楼02A

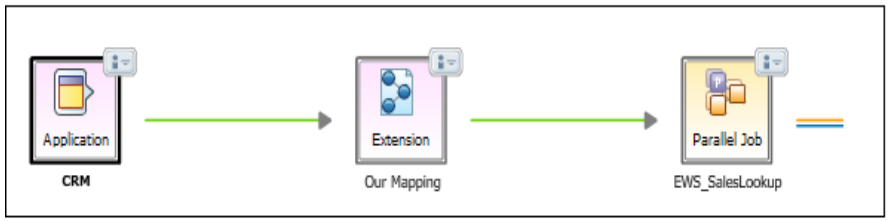
基于Information Server的数据管治架构图



InfoSphere Metadata Workbench

Metadata Workbench

支持信息监管, 可追踪数据移动,数据模型和BI应用



满足需求

- 理解对信息环境进行改造带来的影响
- 跨企业范围的图形展示和追踪信息流
- 访问和报告操作性元数据

带来益处

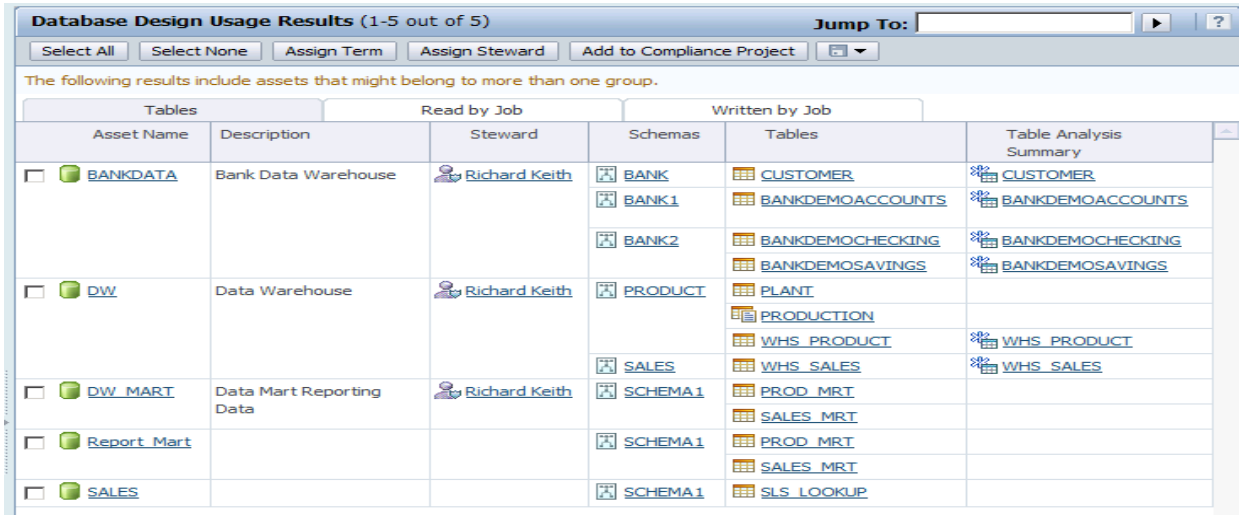
- 避免系统断层
- 为数据监管提供审计信息
- 构建LOB 用户的信心

元数据关系的编目和报告

Window into Your Data Integration World

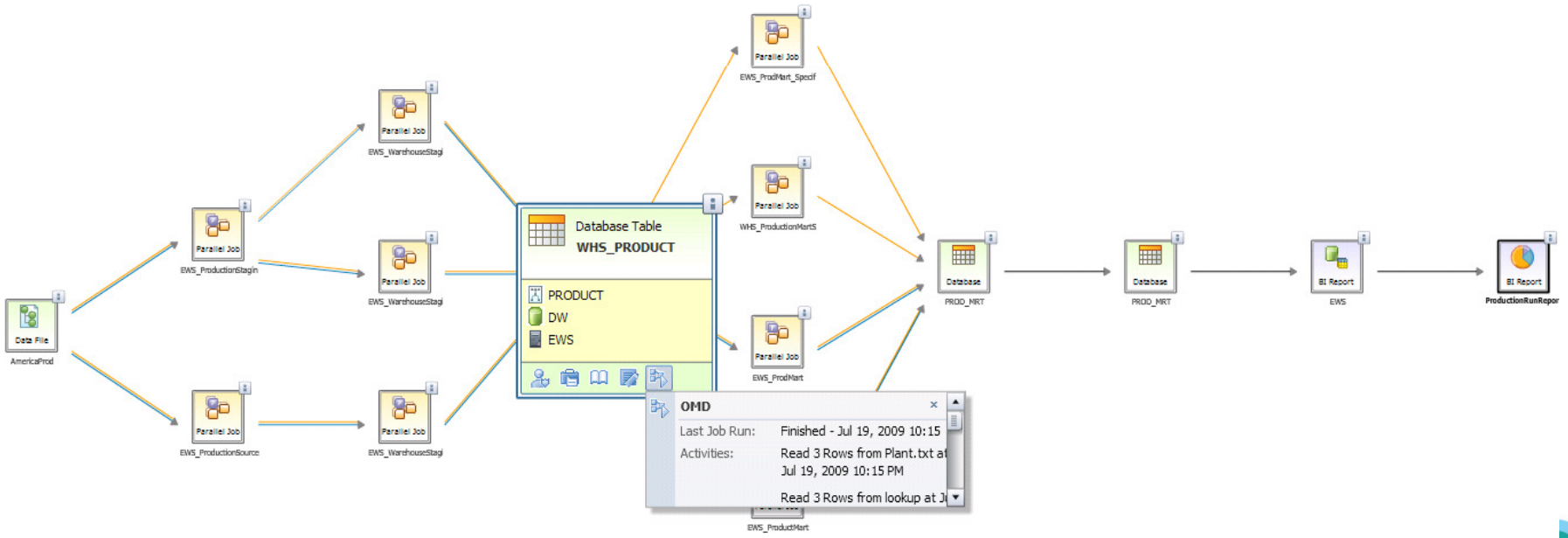
- ✓业务规则
- ✓数据质量规则
- ✓业务数据和定义
- ✓数据管理员
- ✓运行和操作统计信息

- ✓源和目标系统
- ✓表，字段，历史文件
- ✓源到目标的映射
- ✓BI 报表，集市，数据模型
- ✓ETL 任务，FTP 流程，脚本



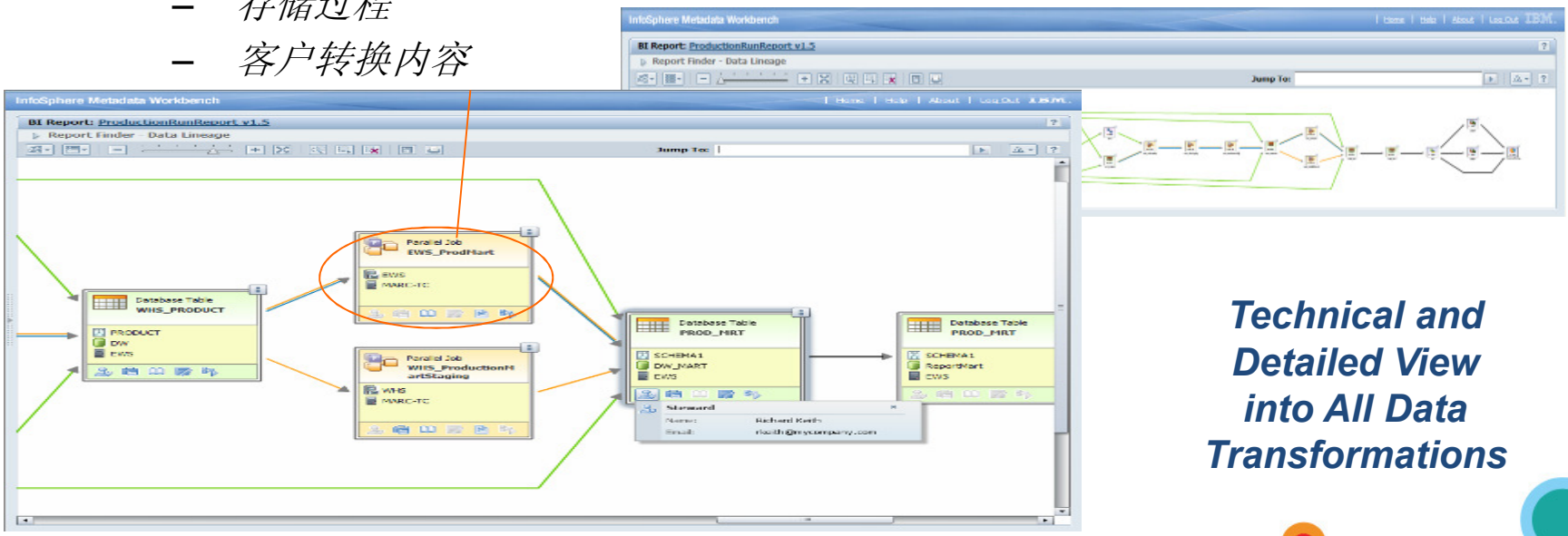
跨工具冲突分析

- 评估改造的冲突,降低分险
- 展示对后续应用和BI报表的影响
- 冲突区域导航和下钻



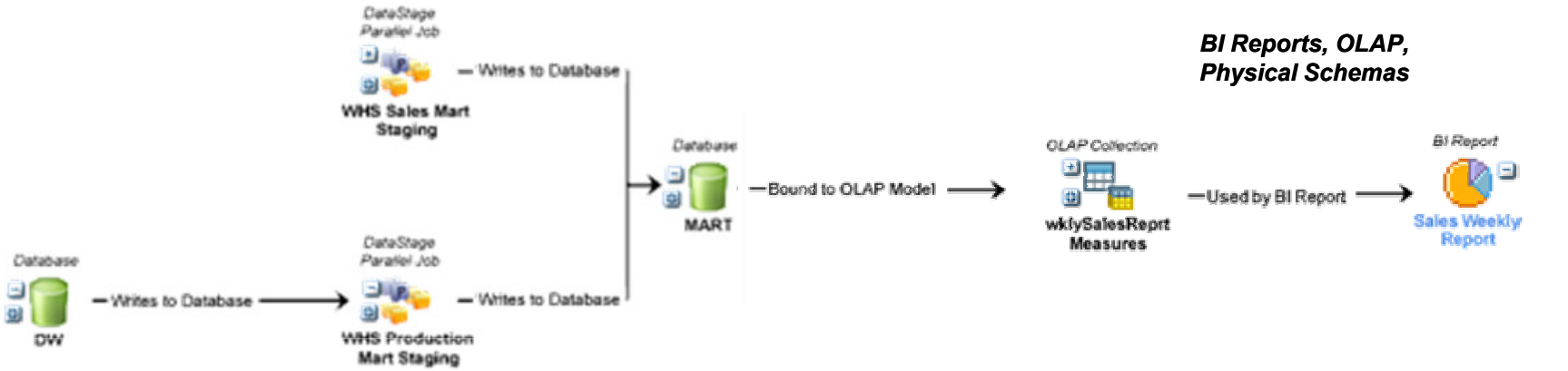
数据世系

- **可视化扩展到不同种类的**数据整合流程 — 包含**Information Server** 和 **3rd Party**
- 通用业务使用案例
  - *第三方ETL工具和应用*
  - *主机COBOL程序*
  - *外部脚本, Java 程序, 或者 web services*
  - *存储过程*
  - *客户转换内容*



Technical and Detailed View into All Data Transformations

端到端的数据血缘分析



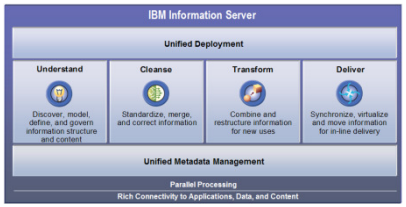
Data Warehouse

DataStage & QualityStage Job

Data Mart

OLAP Collection

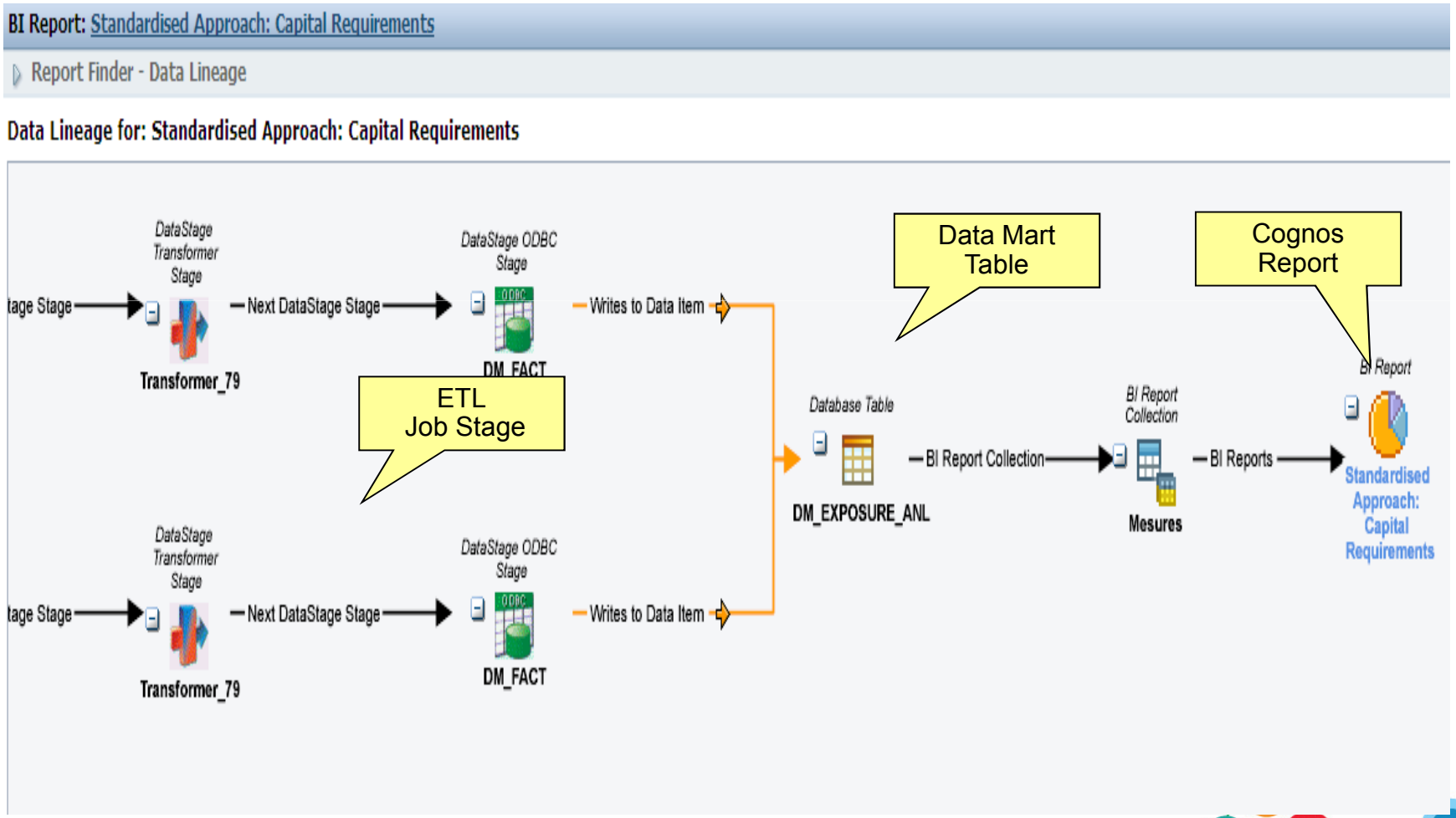
Report



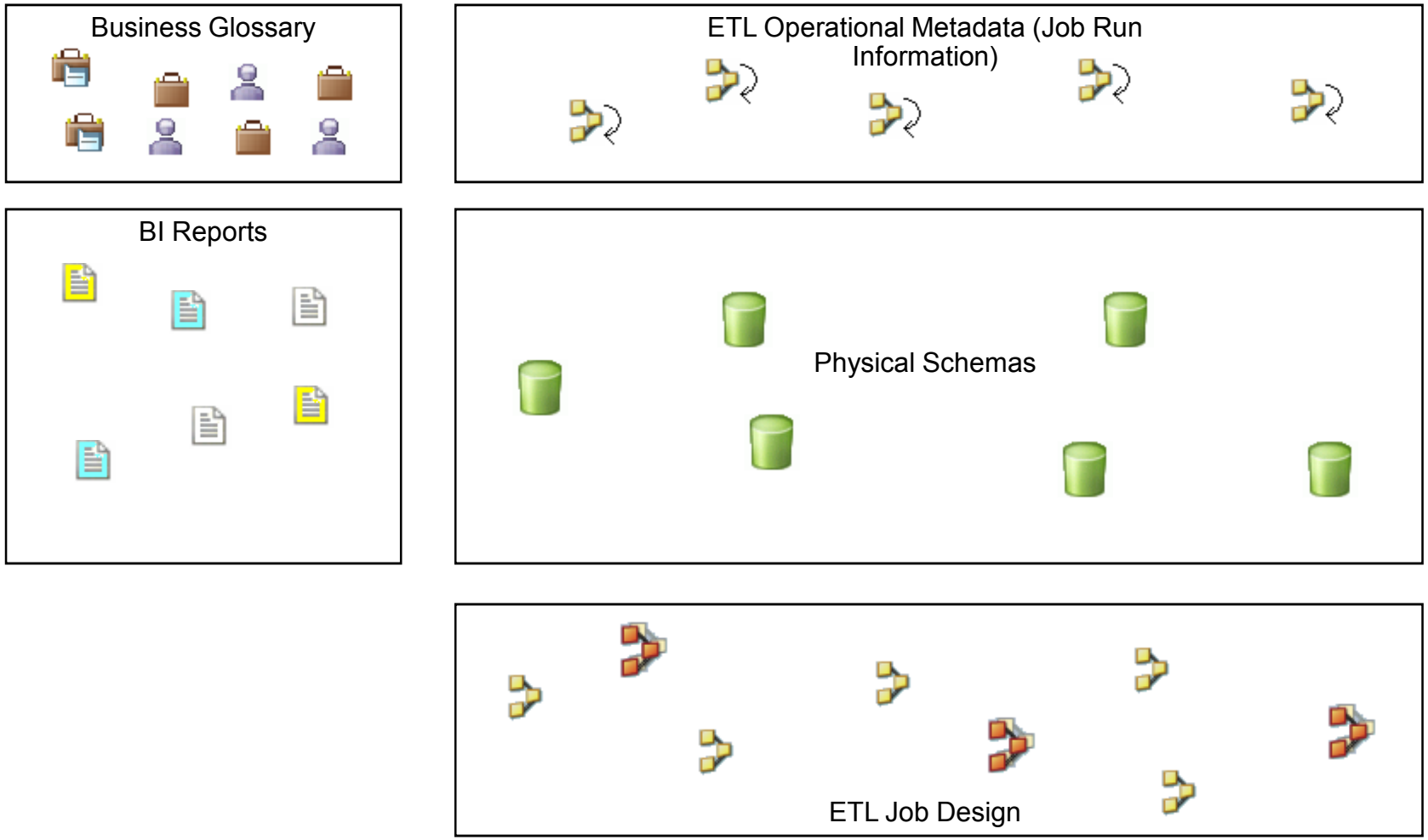
IBM InfoSphere Information Server

COGNOS AN IBM COMPANY

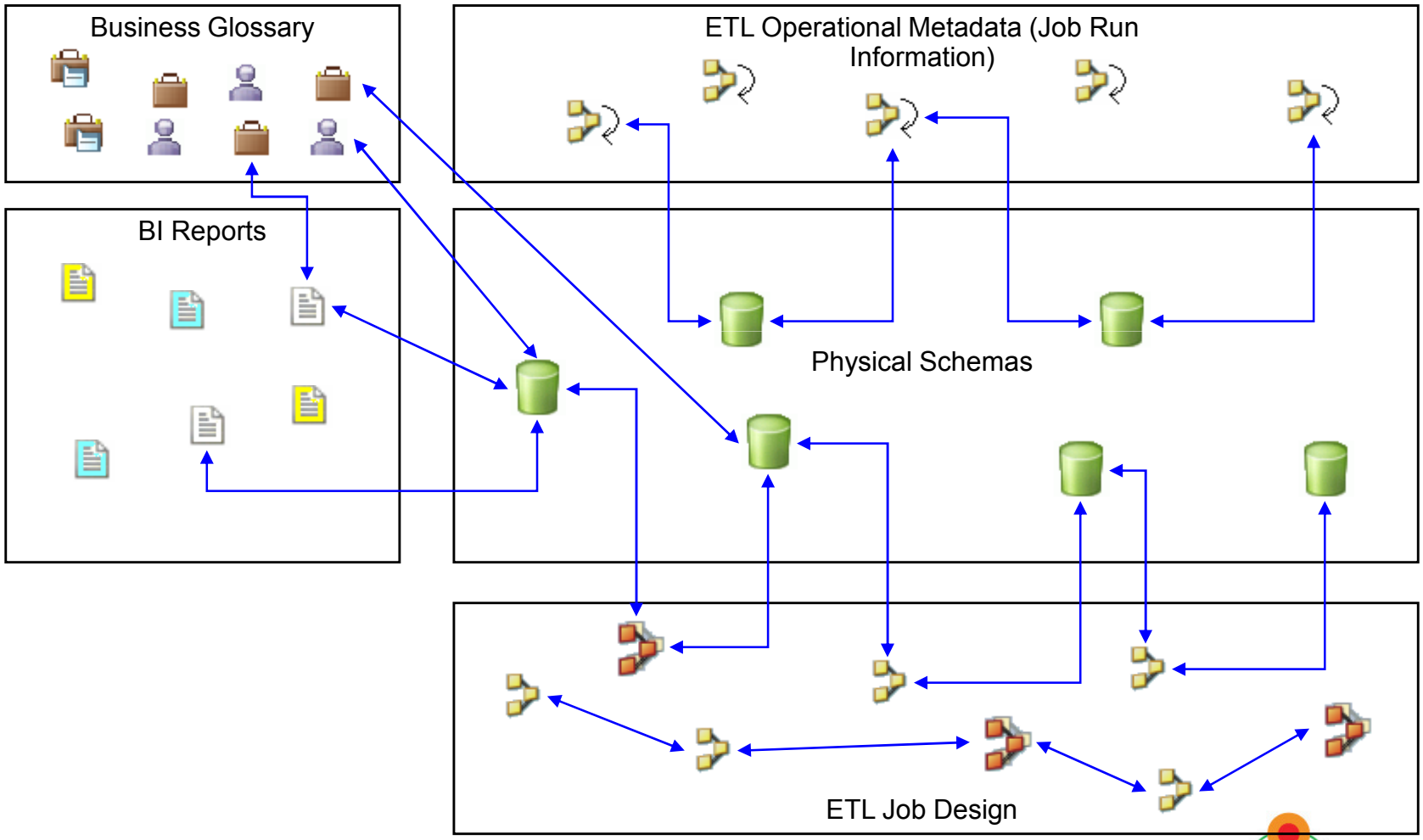
Metadata Workbench图形化世系图 (部分Snapshot)



元数据区域


Business Glossary
ETL Operational Metadata (Job Run Information)
BI Reports
Physical Schemas
ETL Job Design

元数据互动


Business Glossary
ETL Operational Metadata (Job Run Information)
BI Reports
Physical Schemas
ETL Job Design

IBM信息服务器帮助完善数据治理

- *在项目实施早期揭示数据质量和非规范化问题，有助于提高成功率。*
- *确保数据项目提供可信赖信息，并把由于使用错误数据而带来的风险减至最低。*
- *在数据整合中提供多种组件灵活的清理数据，提升数据质量。*
- *提供独特的模糊匹配算法，提供语义层面的数据清洗。*
- *基于元数据的整合为整合信息和丰富信息提供了具有突破意义的生产效率和灵活性*
- *基于元数据管理，可以了解和信任信息的血缘世系，满足法律遵从性和审计要求*
- *集中管理从项目开始到结束的生成，存储，维护和审计，强化数据治理。*

InfoSphere™
software

IBM

谢谢!

Trusted Information