



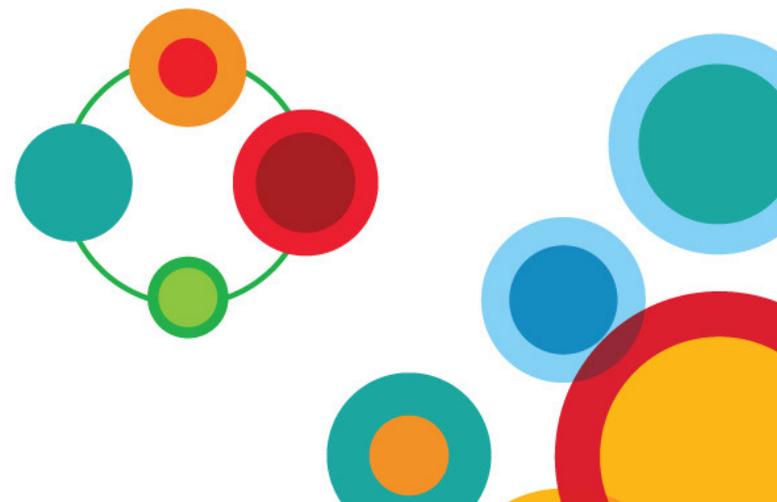
您的**信息** 您的**智慧**

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）



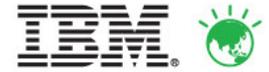


## 议程



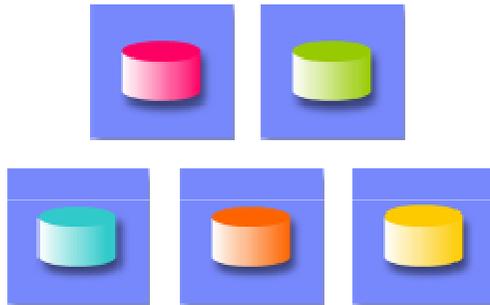
- 数据治理中的数据质量和元数据
  - ✓ 数据质量重要性
  - ✓ 元数据管理重要性
  - ✓ IBM 解决方案（Information Server）
- 数据质量和元数据管理 – 工具
  - ✓ 业务术语表（Business Glossary）
  - ✓ 洞察数据（Information Analyzer）
  - ✓ 规范开发（Fasttrack）
  - ✓ 清洗/转换数据（QualityStage/DataStage）
  - ✓ 元数据工作台（Metadata Workbench）





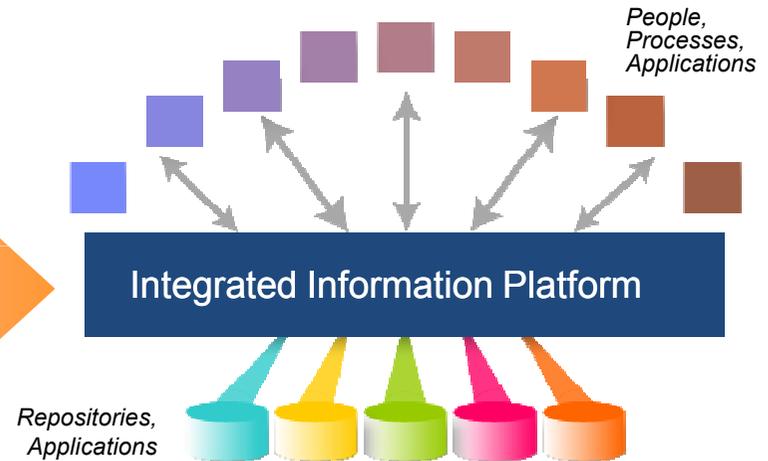
# 日新月异信息架构

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。

## 数据逻辑错误

- 违反业务规则
- 违反业务代码定义

数据  
不完整

数据  
不一致

数据质量表现

数据逻辑  
错误

数据  
冗余

## 数据不一致

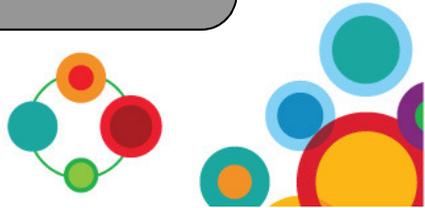
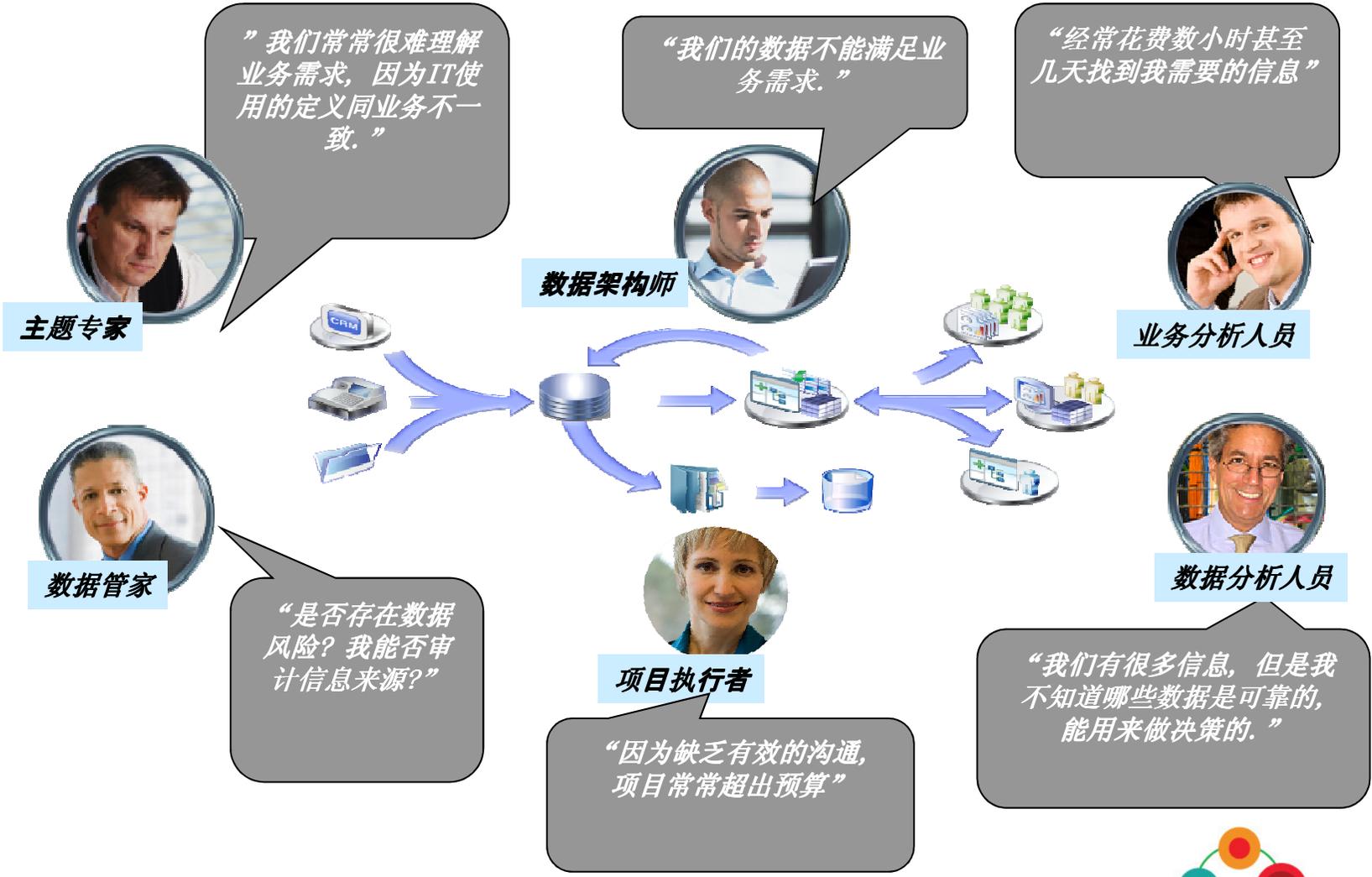
- 相关联业务系统数据不同步；
- 不同系统之间描述同一业务问题的数据定义存在差异。

## 数据冗余

- 重复数据记录
- 非法键值



# 数据治理中遇到的问题





# 危害

83% 数据集成项目  
需要重复实施甚至失败



无效和重复性工作  
增加运作成本



消费者缺乏信心

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



痛失商机

低劣数据质量严重地降低  
公司年收入

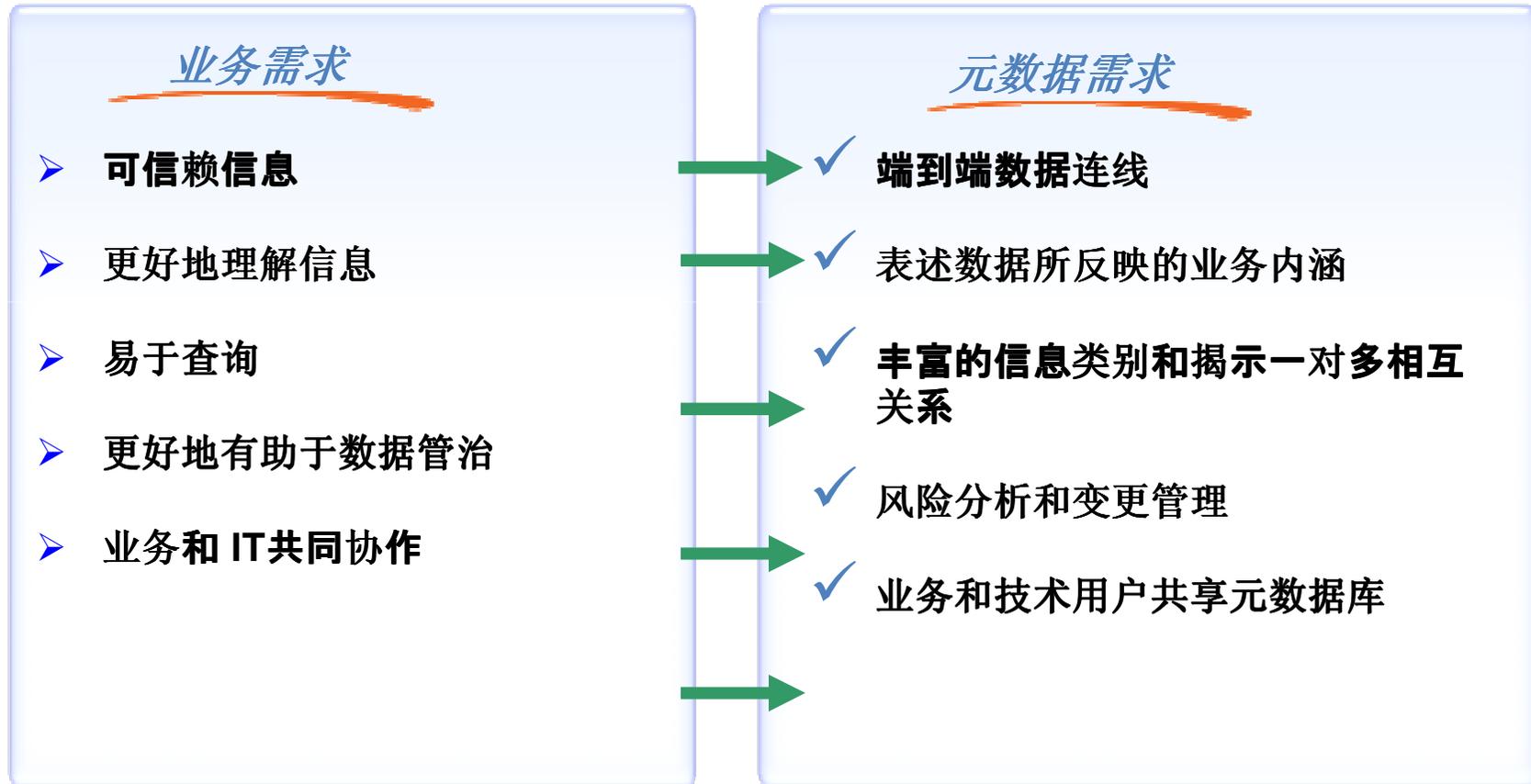
25% 时间浪费在  
辨别数据是否“坏数据”

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

#### Information Analyzer

分析 & 理解 源数据

#### 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

#### Change Data Capture

数据增量获取

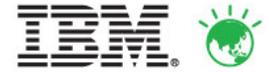
### Metadata Server / Metadata Workbench

跨信息整合生命周期的统一的元数据管理

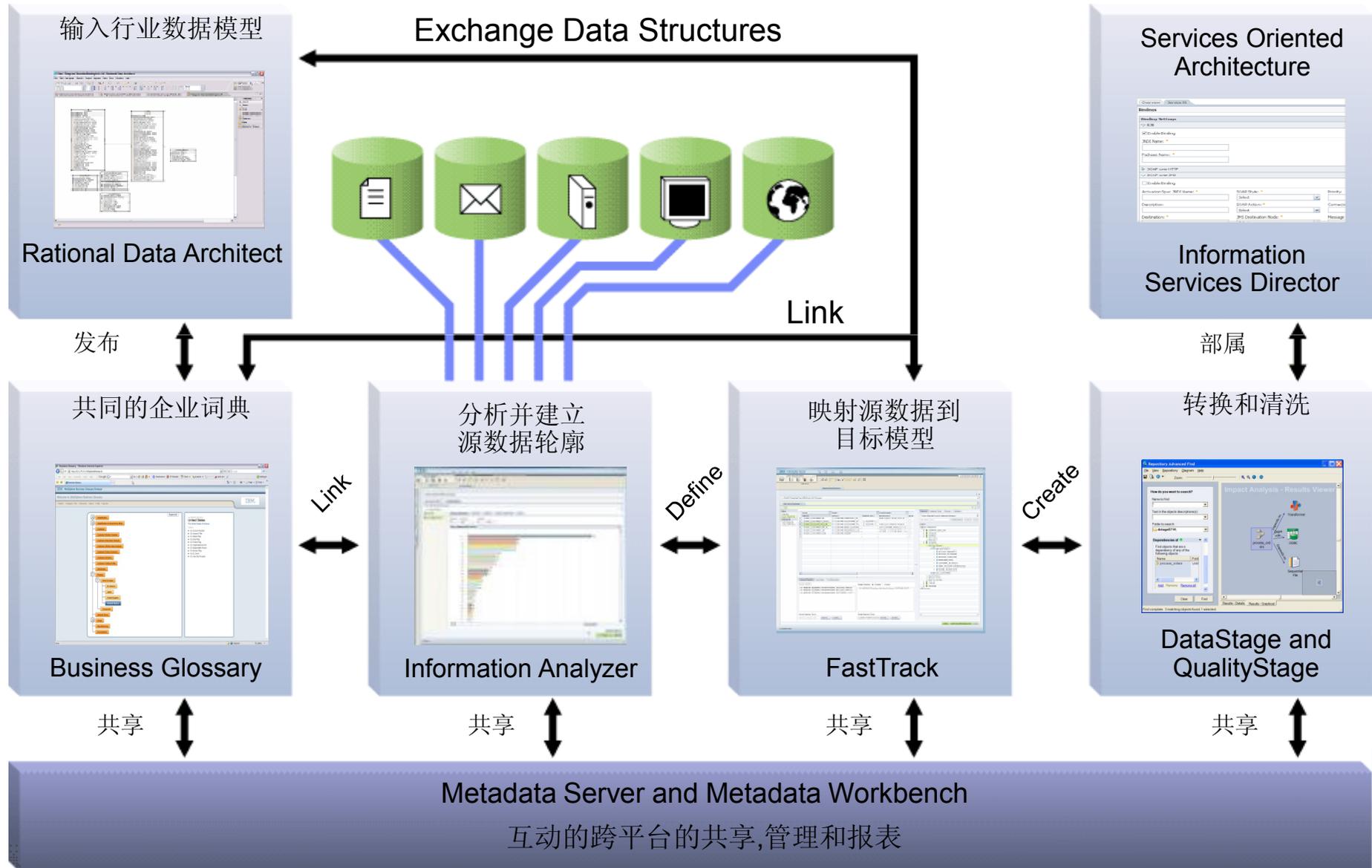
并行处理

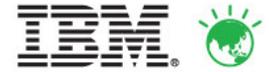
丰富的应用、数据和内容连接的支持





# 基于Information Server的数据管治架构图



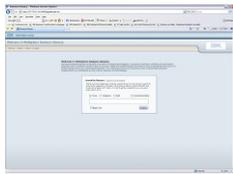


# 基于Information Server的元数据管理架构图

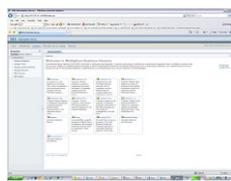
加强协作，让IT向业务看齐



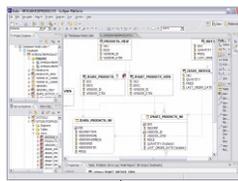
Business Users



Subject Matter Experts



Architects



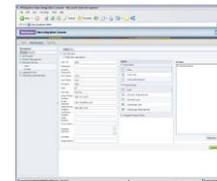
Data Analysts



Developers



DBAs



## 统一元数据管理



技术, 操作, 业务

- 易于集成
- 易于变更管理 & 重用
- 基于“可信赖”信息，更有信心使用信息
- 遵循业界规范和标准





## 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。



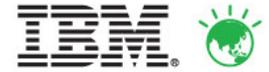


# 议程

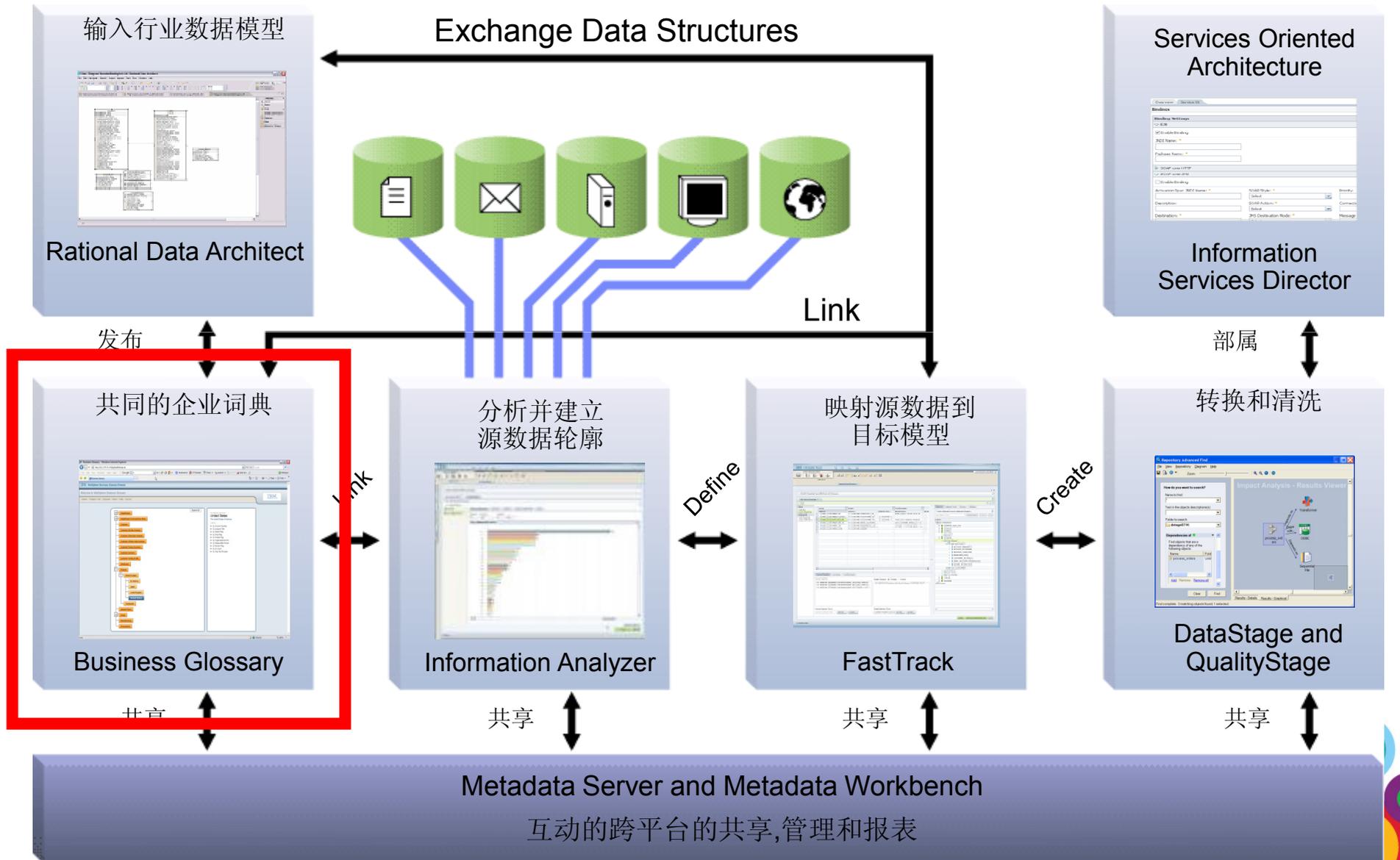


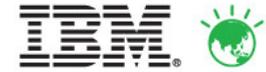
- 数据治理中的数据质量和元数据
  - ✓ 数据质量重要性
  - ✓ 元数据管理重要性
  - ✓ IBM 解决方案（Information Server）
- 数据质量和元数据管理 – 工具
  - ✓ 业务术语表（Business Glossary）
  - ✓ 洞察数据（Information Analyzer）
  - ✓ 规范开发（Fasttrack）
  - ✓ 清洗/转换数据（QualityStage/DataStage）
  - ✓ 元数据工作台（Metadata Workbench）





# 基于Information Server的数据管治架构图



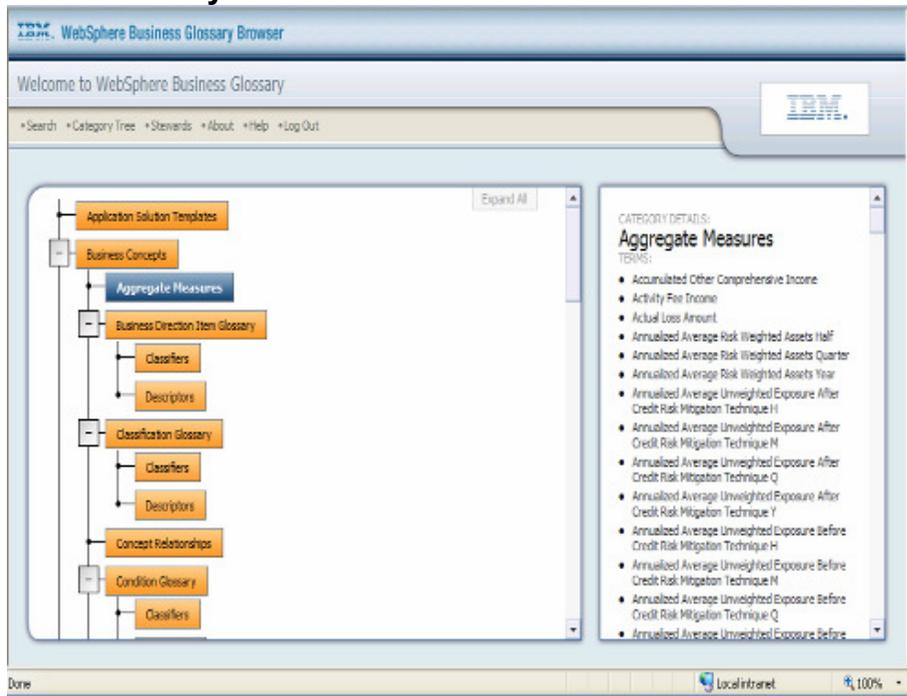


# InfoSphere Business Glossary



创建和管理业务词典和层级关系, 及相关的物理信息源

Business Glossary



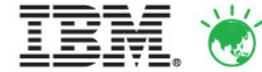
## 需求

- 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



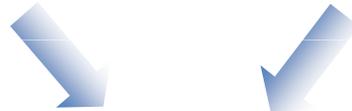


# 创建共享的业务词汇表

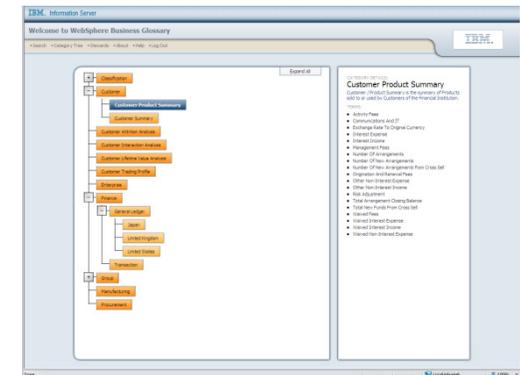
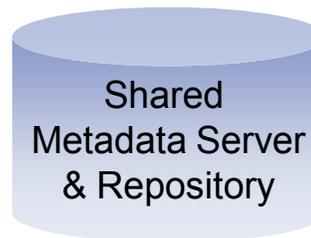
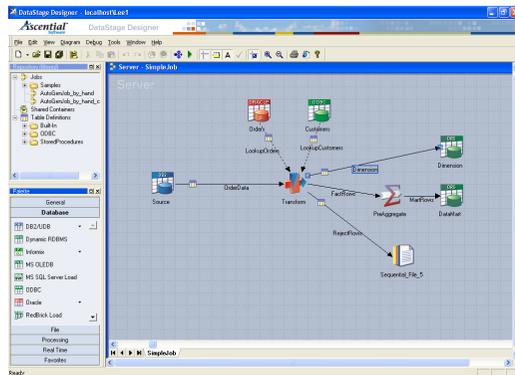
Database = DB2  
 Schema = NAACCT  
 Table = DLYTRANS  
 Column = TAXVL  
 data type = Decimal  
 (14,2)  
 Derivation: SUM(TRNTAXMT)

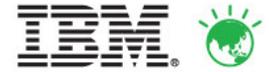


**Category: Costs**  
**Term: Tax Expense**  
**Full Name: Tax to be paid on Gross Income**  
 “The expense due to taxes .....”  
 (John Walsh is responsible for updates. 90% reliable source)  
**Status: CURRENT**



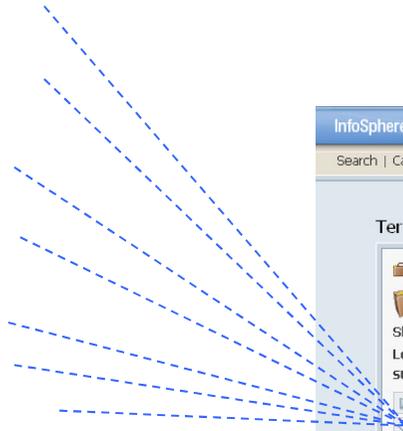
## Achieve a common vocabulary between business & technical users!





# 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...



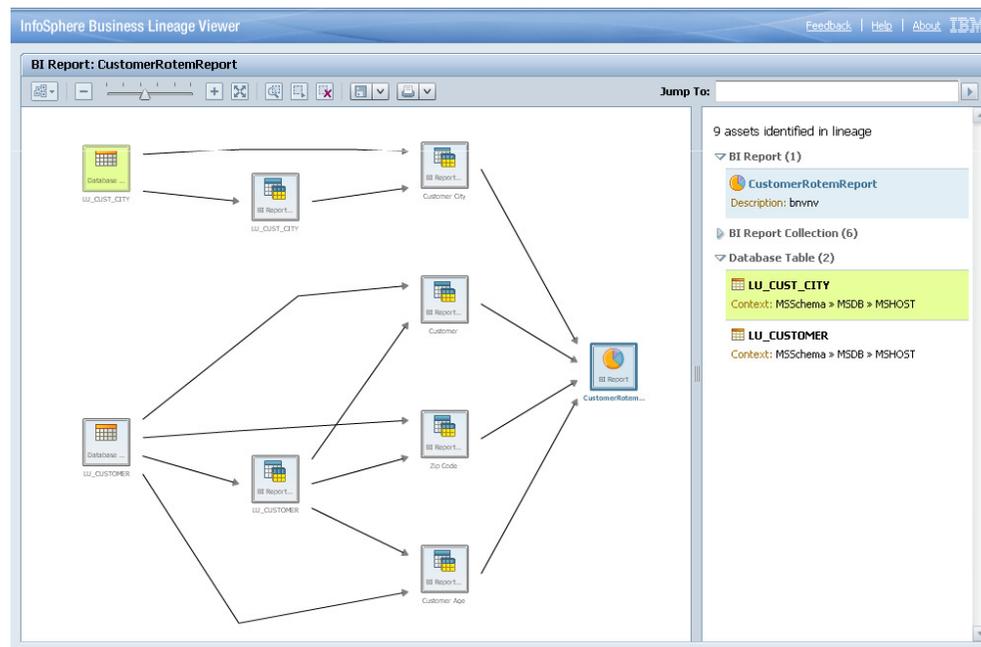
The screenshot displays the 'Term Details' page for 'GL Account Number' in the InfoSphere Business Glossary. The breadcrumb trail is 'Finance > General Ledger > United States'. The term's short description is 'Ten digit general ledger account number' and the long description is 'The ten digit account number. Sometimes referred to as the account ID. This value is of the form L-FIIIIVVVV'. The status is 'Standard'. The 'Assigned Assets' section lists three items: 1. Customers\_Table (Local Host > Northwind\_Database > Northwind\_Schema), 2. Marketing\_Campaign\_Email\_Message (WBM Task, Context: Marketing Campaign >> Mailing Members Process, Location: NE Process Server Node), and 3. SimpleReport1. There are also sections for 'Notes' and 'History'.





## 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





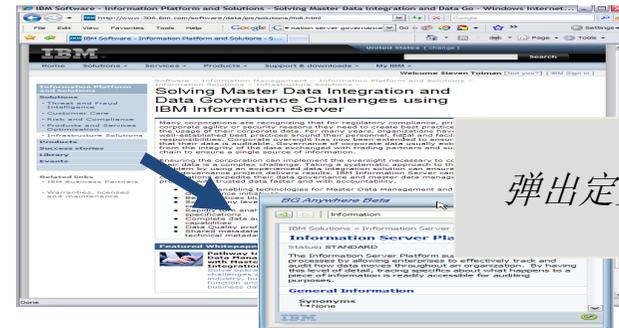
# InfoSphere Business Glossary Anywhere

查看术语内容, 全面理解信息含义



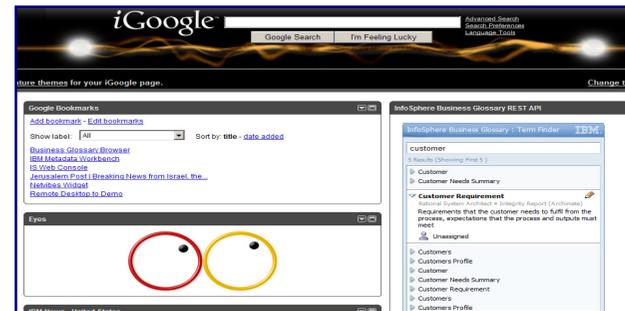
任何用户

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

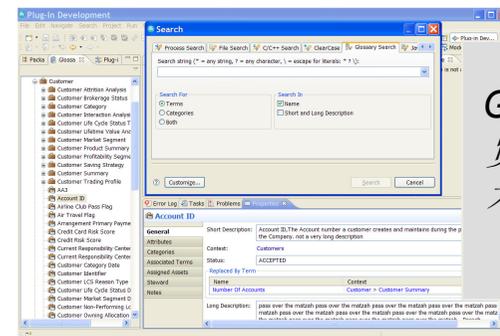


弹出定义!

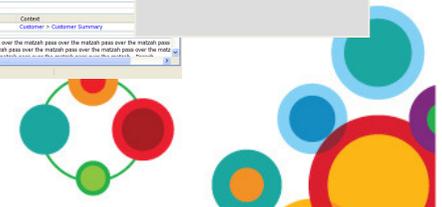
- 零足迹 (Zero-footprint) REST API, 可以将 Business Glossary 直接嵌入任何应用

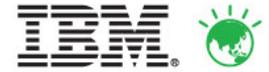


- 在Eclipse项目中访问授权信息源

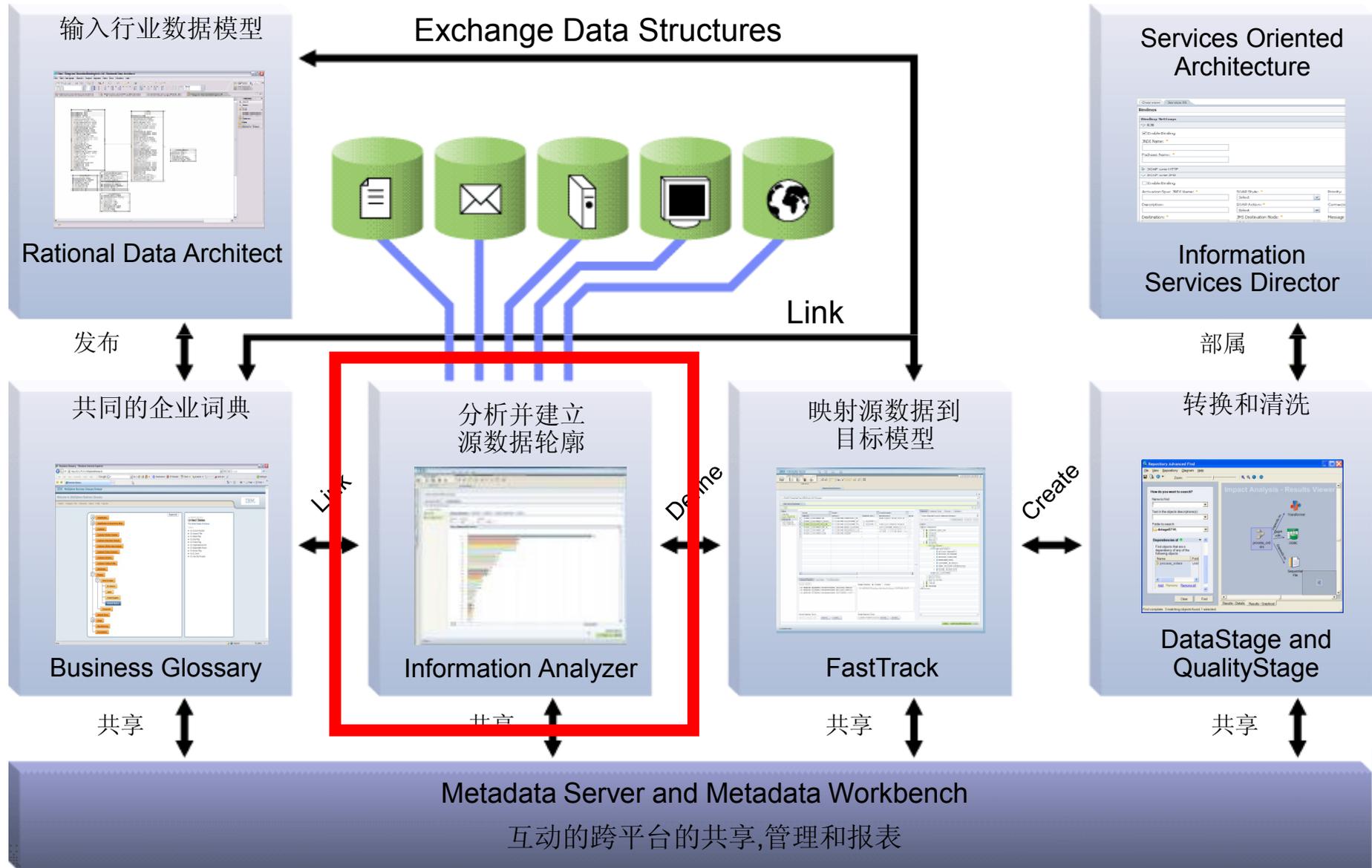


Glossary 浏览显示业务术语层次关系





# 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?

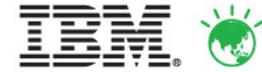
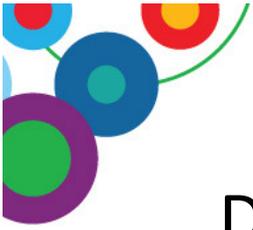
分析数据源，揭示信息的结构、内容和质量

- 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?

商业和数据分析师、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

The screenshot shows the IBM Information Server Column Analysis interface. The main window displays the 'GlobalCo\_Ord\_Dtl' data source. The 'View Details' section shows a table of data values for the 'QTYORD' column. The table includes columns for Data Value, Frequency (# and %), Value Flag, Data Type, Length, Format, Transform, and Value (Definition, Source, Type). The table shows 11 data values ranging from 0 to 11, with frequencies ranging from 0.49% to 6.01%. The 'QTYORD' column is selected in the left-hand pane.

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	Data	Numeric zero
1	384	6.01	Valid	DFLOAT	1	9		Data	Data	Data
2	314	4.92	Valid	DFLOAT	1	9		Data	Data	Data
3	316	4.95	Valid	DFLOAT	1	9		Data	Data	Data
4	254	3.98	Valid	DFLOAT	1	9		Data	Data	Data
5	447	7	Valid	DFLOAT	1	9		Data	Data	Data
6	442	6.92	Valid	DFLOAT	1	9		Data	Data	Data
7	287	4.49	Valid	DFLOAT	1	9		Data	Data	Data
8	415	6.5	Valid	DFLOAT	1	9		Data	Data	Data
9	348	5.45	Valid	DFLOAT	1	9		Data	Data	Data
10	223	3.49	Valid	DFLOAT	2	99		Data	Data	Data
11	31	0.49	Valid	DFLOAT	2	99		Data	Data	Data





# Data Profiling: Table Analysis (表分析)



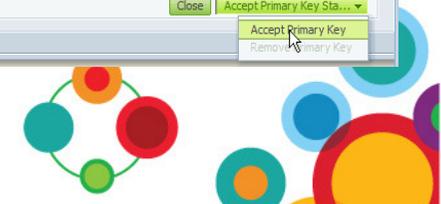
- Primary Keys (single or multi-column)
- Key Duplicates

The screenshot shows the IBM Information Server Primary Key Analysis tool. The main window displays a table with columns: Defined Primary Key, Selected Primary Key, Defined Foreign Key, Column, Data Class, Data Type, Length, Unique %, Null %, Duplicate %, and Candidate. The table lists various columns from the 'GlobalCo\_Ord\_Dtl' table, with 'ordIDitemNo' selected as the primary key. Below the table, there are sections for 'View Duplicate Check (ordIDitemNo)' and 'Duplicate Check Results View'. The 'Duplicate Check Results View' shows a summary of records: Unique (6383, 99.93737%), Duplicate (2, 0.03131361%), and Nulls (0, 0%). A 'Duplicates' table lists primary key values and their counts: 22347|2 and 27511|4.

Defined Primary Key	Selected Primary Key	Defined Foreign Key	Column	Data Class	Data Type	Length	Unique %	Null %	Duplicate %	Candidate
			ordIDitemNo	T	STRING	0	99	0	0	False
			ORDERID	Q	DFLOAT	8	20	0	79	False
			ITEMNO	C	DFLOAT	8	0	0	100	False
			STOCKCODE	C	STRING	8	0	0	99	False
			LISTPRICE	C	DECIMAL	19	0	0	99	False
			QTYORD	C	DFLOAT	8	0	0	100	False
			QTYSHIP	C	DFLOAT	8	0	0	99	False
			QTYDUE	C	DFLOAT	8	0	0	99	False
			VALORD	Q	DECIMAL	19	43	0	56	False
			VALSHIP	Q	DECIMAL	19	32	0	67	False
			VALDUE	C	DECIMAL	19	18	0	81	False
			COMPLETE	U	INT16	0	0	0	100	False

Total Records	Records	%
Unique	6383	99.93737
Duplicate	2	0.03131361
Nulls	0	0

Duplicates	Number of Records	%
Primary Key Value		
22347 2	2	0
27511 4	2	0





# Data Profiling: Cross Table Analysis (跨域分析)



• Foreign Key Relationships

• Referential Integrity

• Cross-Domain Relationships

• Data Redundancy

您的信息 您的智慧

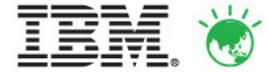
The screenshot displays the 'Foreign Key Analysis' window in the IBM Data Profiling tool. It shows a table of 'Foreign Key Candidate Pairs' and a 'Common Domain' visualization.

	Base Column	Paired Column
Column	CUSTOMER_ID	PARENT_CUST_ID
Table	WorldCo_BillTo	WorldCo_ShipTo
Source	GlobalCo	GlobalCo
Primary Ke	Yes	No
Foreign Ke	No	No
Data Class	Identifier	Code
Data Type	INT32	INT32
Length	0	0
Precision	0	0
Scale	0	0
Cardinality	1030	3717
Unique	No	No
Constant	No	No
Definition	No	No

The 'Common Domain #' section shows two overlapping circles, one labeled 'PK' (Primary Key) and one labeled 'FK' (Foreign Key), representing the relationship between the two domains.

2011 IBM 信息管理与业务分析论坛





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



•Current-to-Prior Comparison

•Content & Structural Variation

The screenshot shows the IBM Information Server interface. The main window is titled 'Baseline Analysis' and is connected to 'wb-gecko-xp:9080'. It displays a 'Select Data Source to Work With' dialog for 'WorldCo\_BillTo'. The left pane shows a tree view of fields under 'Common', with 'STATE\_ABBREVIATION' selected. The right pane shows a 'Differences' table comparing 'Checkpoint' and 'Baseline' data.

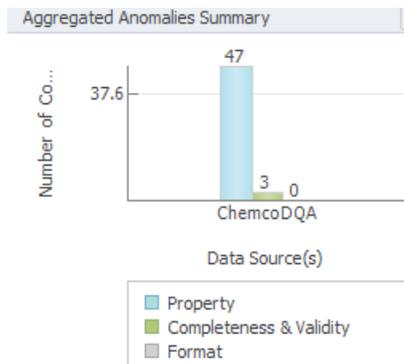
Value & Format Profile			Completeness & Validity Measures		
Name	Checkpoint	Baseline	Name	Checkpoint	Baseline
Cardinality	42	41	# Incomplete	3	3
# Distinct Values	1027	1026	% Incomplete	7.142857	7.317073
# Distinct Formats	2	2	# Invalid	0	0
Standard Deviation Value Frequency	0	0	% Invalid	0	0
Standard Deviation Format Frequency	0	0	# Format Violations	0	0
# Null	3	3	% Format Violations	0	0
% Nulls	7.142857	7.317073			





# 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。



**Overview** | **Result**

Select View

- By Record
- By Distribution
- By Rule
- By Pattern

**Distribution by Rules Not Met**

View By: # of Rules Not Met

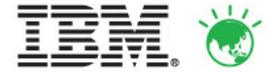
# of Rules Not Met	Run		Baseline	
	Record #	Record %	Record #	Record %
0	13894	94.7490 %	10966	74.7818 %

**Baseline Set Comparison**

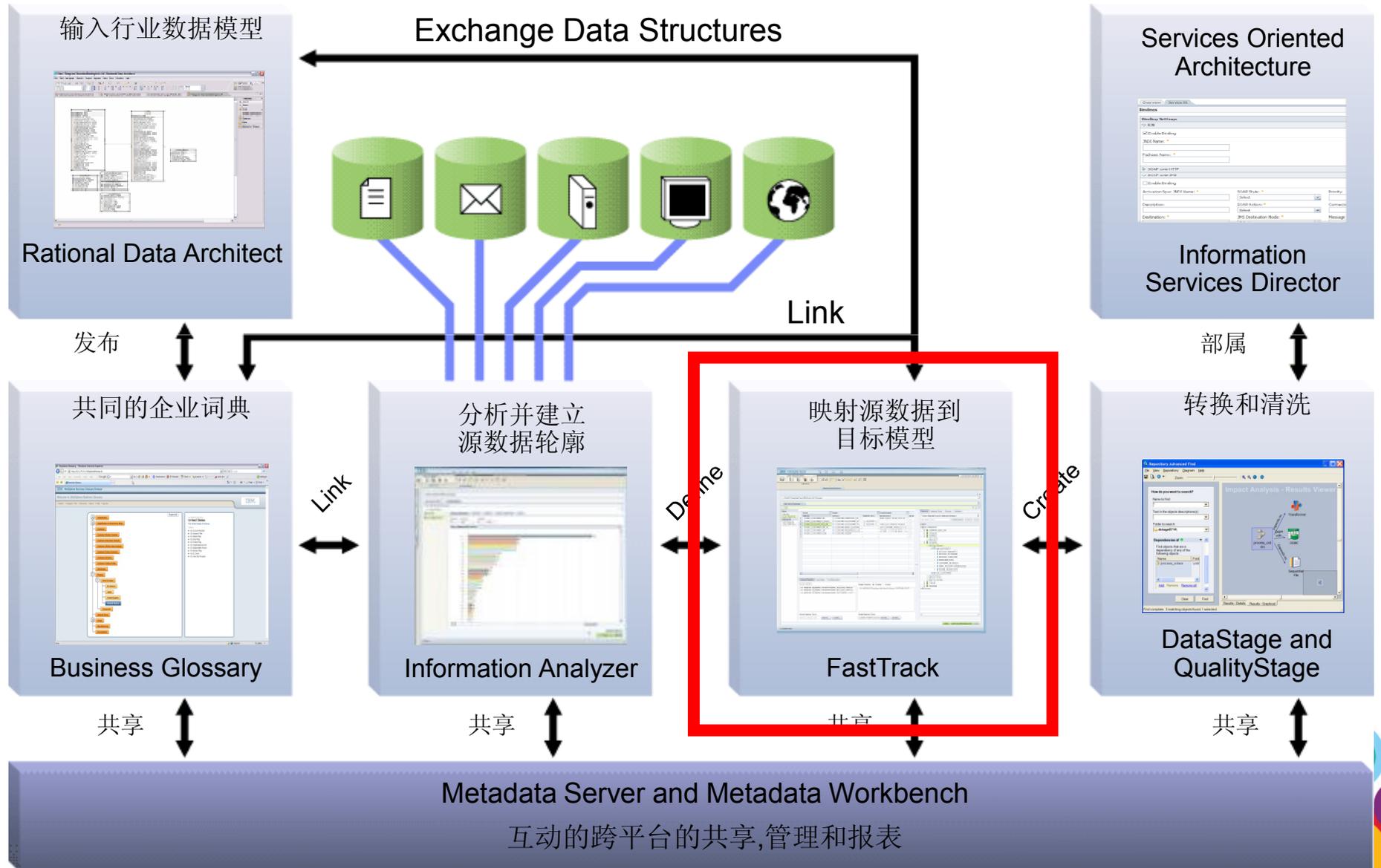
	Run	Baseline
Date/Time Executed	3/11/2009 12:45:50 PM	3/10/2009 4:57:45 PM
Total Records	14664	14664
Mean Rules Not Met	1.3281 %	6.4171 %
Standard Deviation	5.6750 %	11.1747 %
Similarity	62.6696 %	
Degradation	0.0000 %	

Name	Status	Validity		Confidence	
		Variance	Trend	Variance	Trend
◇ L1CPL_MSTRCTLG_CARRID_Exists	✘	0.0000 % (14669 rec)	✘✘		
◇ L1CPL_MSTRCTLG_CLSG_Exists	✔	0.0000 % (14664 rec)	✔		
◇ L1CPL_MSTRCTLG_DIV_Exists	✔	0.0000 % (14669 rec)	✔✔✔		
◇ L1CPL_UNITCTLG_ITEMID_Exists	✔	0.0000 % (6867 rec)	✔		
◇ L1FMT_MSTRCTLG_PKG	✔	0.0000 % (14664 rec)	✔		
◇ L1VAL_MSTRCTLG_HAZMAT	✘	0.0000 % (14669 rec)	✔✔✔		
◇ L1VAL_MSTRCTLG66_HAZMAT	✘	0.0000 % (14664 rec)	✘		
◇ L3RUL_MSTRCTLG_DESCR_Usable	✔	1.0000 % (14664 rec)	✔		
◇ L3RUL_MSTRCTLG_SIZEYPE_Has	✘	0.0000 % (14664 rec)	✔✔✔		
◇ L3RUL_ORDHDR_OrderDt_LT_Ship	✘	0.0000 % (3103 rec)	✘		
◇ M1_MSTRCTLG	✘				
◇ T1VAL_MSTRCTLG	✘	1.0000 % (14664 rec)	✔✔✔	✔✔✔	



# 基于Information Server的数据管治架构图



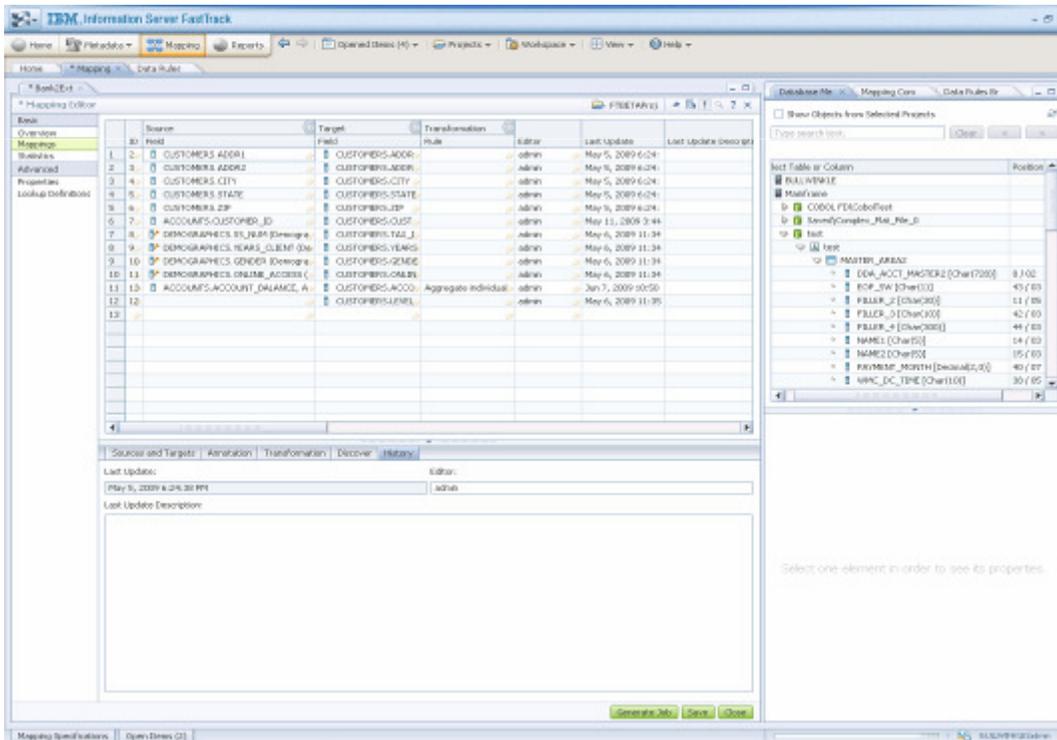


# InfoSphere FastTrack



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



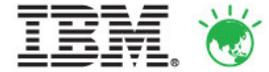
## 功能

- 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

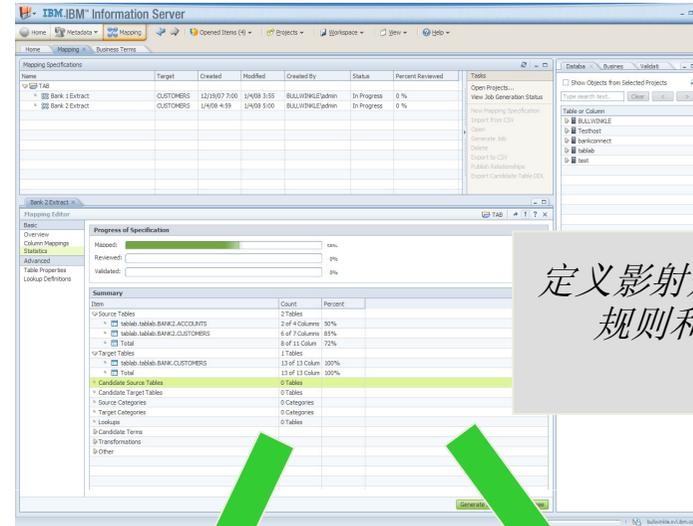
- 加快数据集成的开发
- 集中管理客户的需求规范
- 可在日后审计设计决策





# 跟踪从业务需求到应用开发的过程

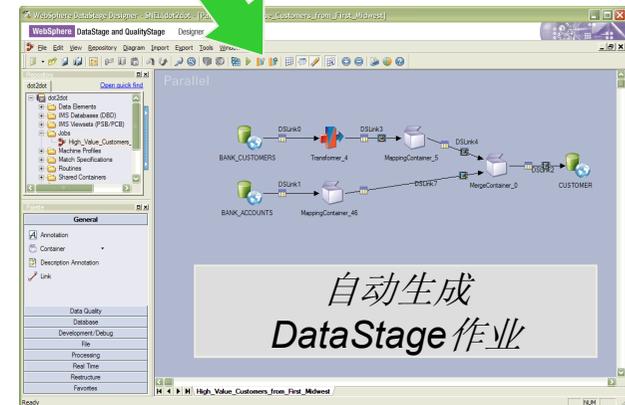
- 单一，集中管理的架构跟踪从业务需求到开发的整个过程
- 可以输入Excel格式的影射规范文件
- 可以定义业务术语并且将其连接到相应的物理元数据
- 自动生成DataStage作业并且将复杂的转换的描述通知开发人员
- 生成历史文档用于审计



定义影射规范以及业务规则和业务术语



灵活的报表功能和审计



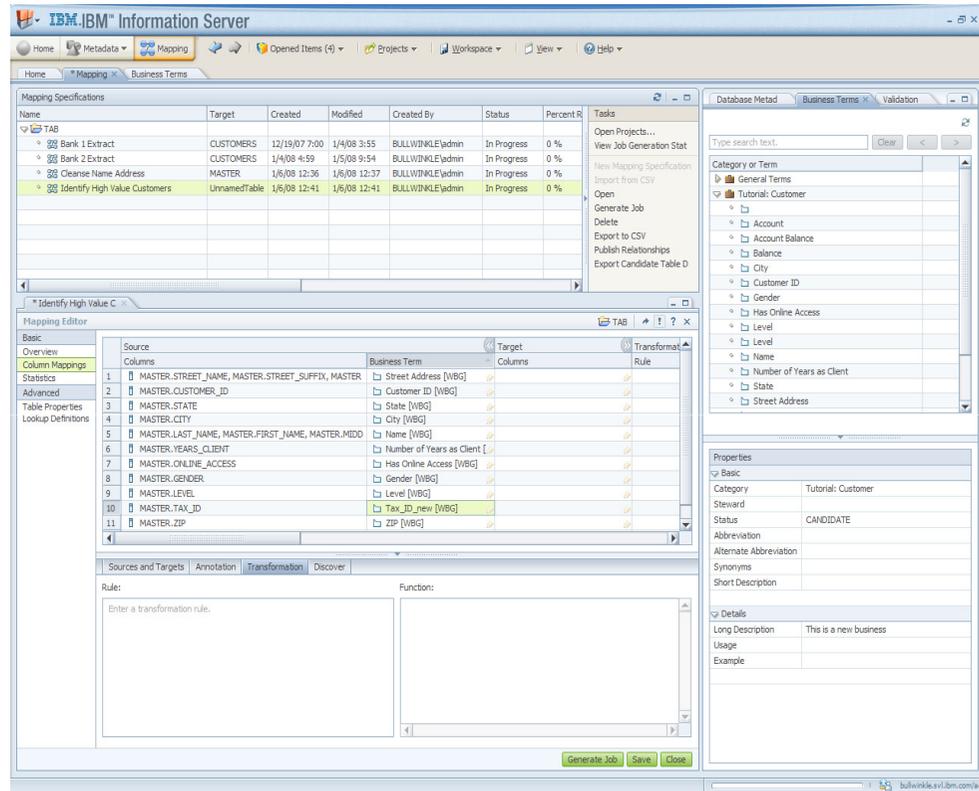
自动生成 DataStage 作业





# 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

Mapping Editor

Source		Target		Transformation			
ID	Field	Field	Business Term	R	Rule Expression	Status	Last Update
1	CHECKING.ADDR.1	CUSTOMERS.ADDR.1					
2		CUSTOMERS.ONLINE_ACCESS			setNull()		
3	CHECKING.ADDR.2	CUSTOMERS.ADDR.2					
4		CUSTOMERS.GENDER			setNull()		
5	CHECKING.STATE	CUSTOMERS.STATE					
		CUSTOMERS.ZIP					
		CUSTOMERS.LEVEL			setNull()		
		CUSTOMERS.YEARS_CLIENT			setNull()		
		CUSTOMERS.ACCOUNT_BALANCE					
		CUSTOMERS.CITY					
		CUSTOMERS.NAME					
		CUSTOMERS.CUSTOMER_ID					
		CUSTOMERS.TAX_ID					

Mapping Editor

Name: \*

Scope: Bank1\_Extract\_Answer

Status: Deployed

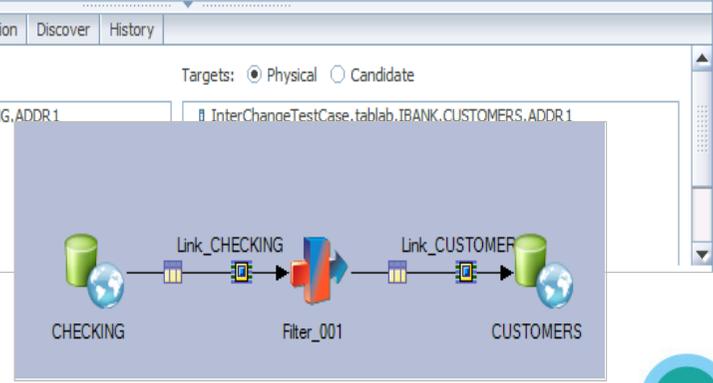
Owner: admin

Description:

The mapping specification was retrieved from the following DataStage Job:  
 IBM-MKLUMPP//dstage1//Bank1\_Extract\_Answer  
 Extracted DataStage Job description:

Created: 5/27/10 11:11 AM      Last Modified: 5/27/10 11:11 AM

Created by User: admin      Modified by User: admin





# 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

- 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

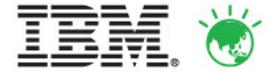
## 提高生产率

- 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

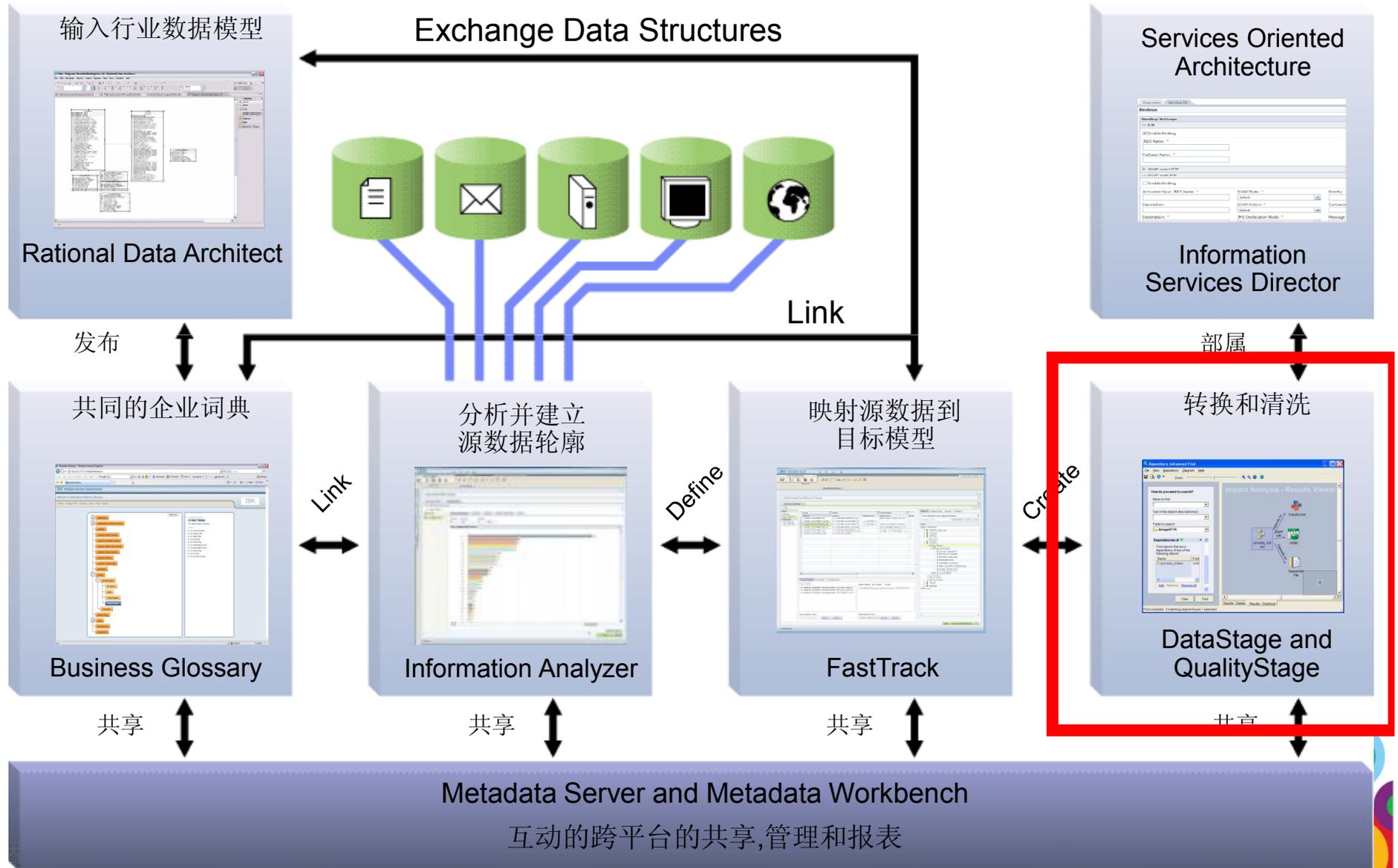
## 更强的数据治理

- 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

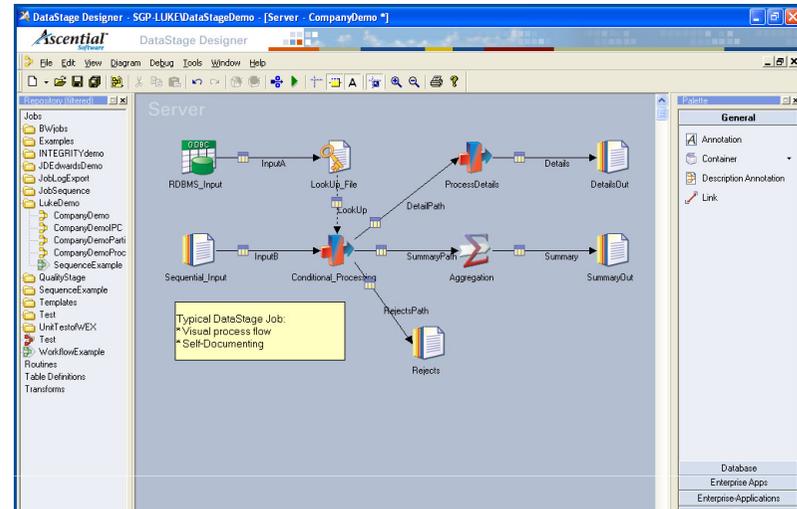




# 数据处理流程设计

## 完全图形化的设计工具:

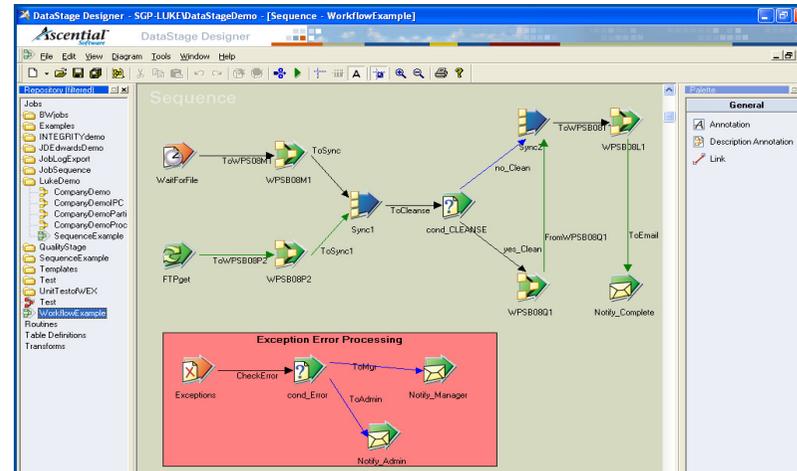
1. 作业容易开发、理解、调试和维护。
2. 强大的、被验证的最好的数据转移和抽取工具。



## 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



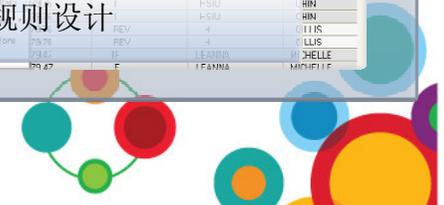
### InfoSphere QualityStage™

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



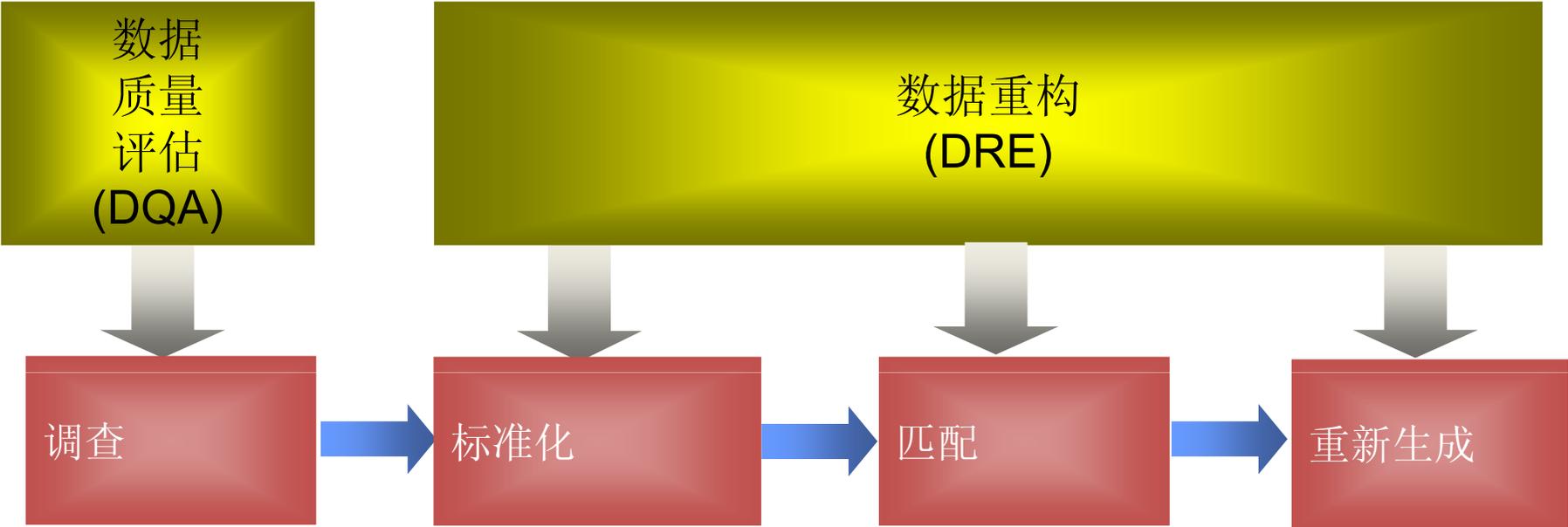
SelfC	RecordType	PassNumber	Weight	GenderCode	FirstName	MiddleName
411	DA	1	35.57	IM	OTIS	GARLAND
411	DA	1	35.57	IM	OTIS	GARLAND
531	DA	1	34.08	IM	NICHO_AS	T
531	DA	1	34.08	IM	NICHO_AS	T
3420	DA	1	31.68	IM	CLAUD	LADALE
3420	DA	1	31.68	IM	CLAUD	LADALE
3420	DA	1	31.68	IM	CLAUD	LADALE
3420	DA	1	31.68	IM	CLAUD	LADALE
4658	DA	1	31.68	IM	CLAUD	LADALE
4658	DA	1	31.68	IM	CLAUD	LADALE
7328	DA	1	31.68	IM	CLIFF	I
7328	DA	1	31.68	IM	CLIFF	I
7328	DA	1	31.68	IM	CLIFF	I
7328	DA	1	31.68	IM	CLIFF	I
2282	DA	1	31.68	IM	REV	LILLIS
2282	DA	1	31.68	IM	REV	LILLIS
6579	DA	1	31.68	IM	HELENE	MICHELLE
6579	DA	1	31.68	IM	HELENE	MICHELLE
6579	DA	1	31.68	IM	HELENE	MICHELLE
6579	DA	1	31.68	IM	HELENE	MICHELLE

可视化匹配规则设计





# QualityStage的实施方法



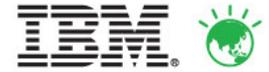
上海市南京西路1266号38楼02A室  
上海南京西路1266号3802A室  
上海市南京路（西）1266号38A02  
南京西路1266/38/02A

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

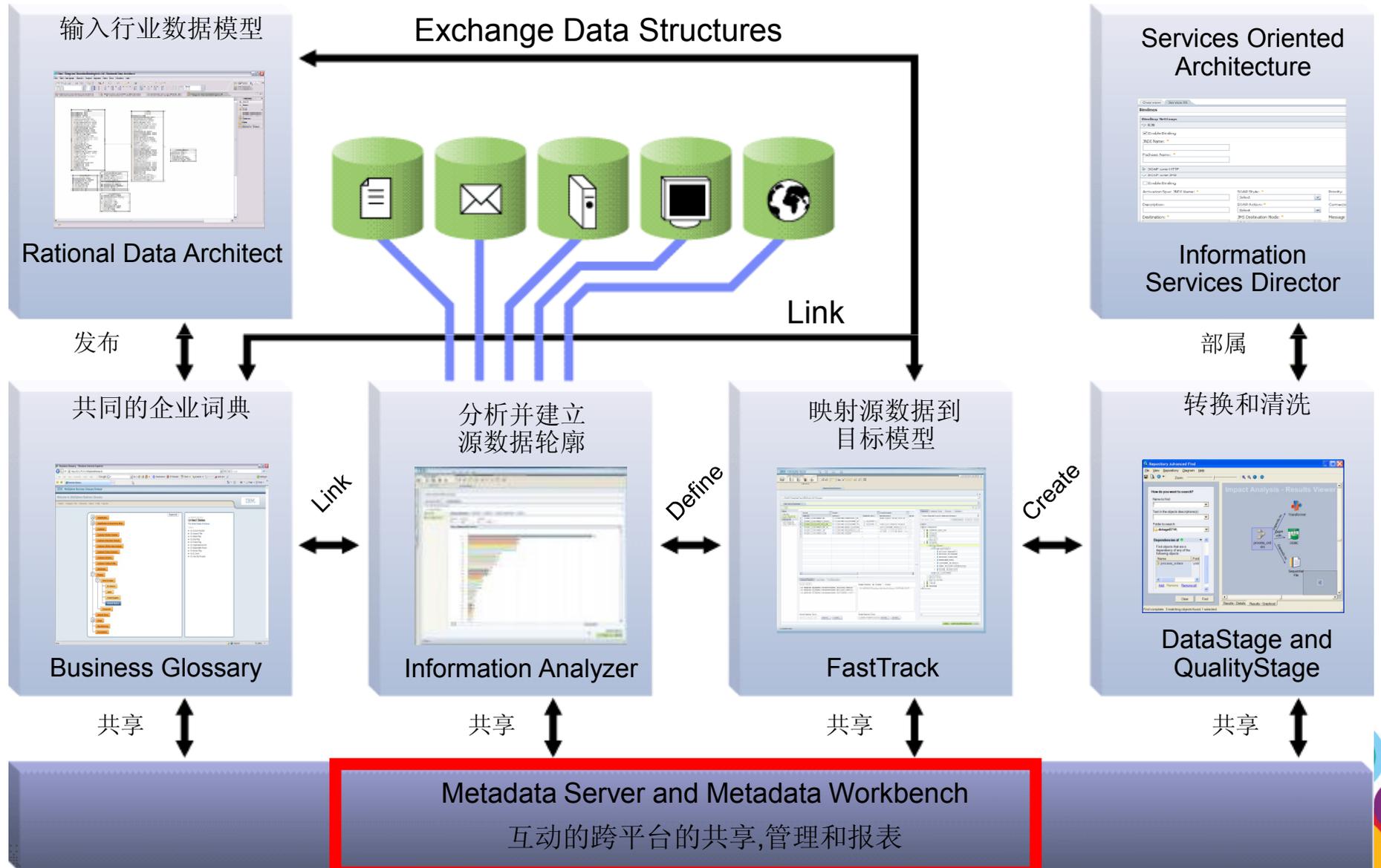
上海市 | 南京西路 | 1266号 | 38楼 | 02A室  
上海 | 南京西路 | 1266号 | 3802A室  
上海市 | 南京路（西） | 1266号 | 3802A  
南京西路 | 1266/ | 38/ | 02A

上海市南京西路1266号38楼02A





# 基于Information Server的数据管治架构图



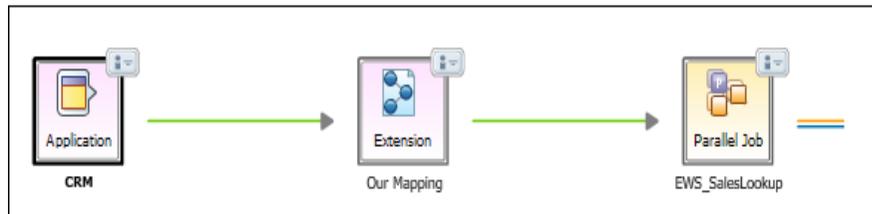


# InfoSphere Metadata Workbench



Metadata Workbench

支持信息监管, 可追踪数据移动, 数据模型和BI应用



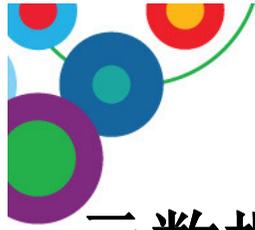
## 满足需求

- 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

Database Design Usage Results (1-5 out of 5) Jump To:

Select All | Select None | Assign Term | Assign Steward | Add to Compliance Project

The following results include assets that might belong to more than one group.

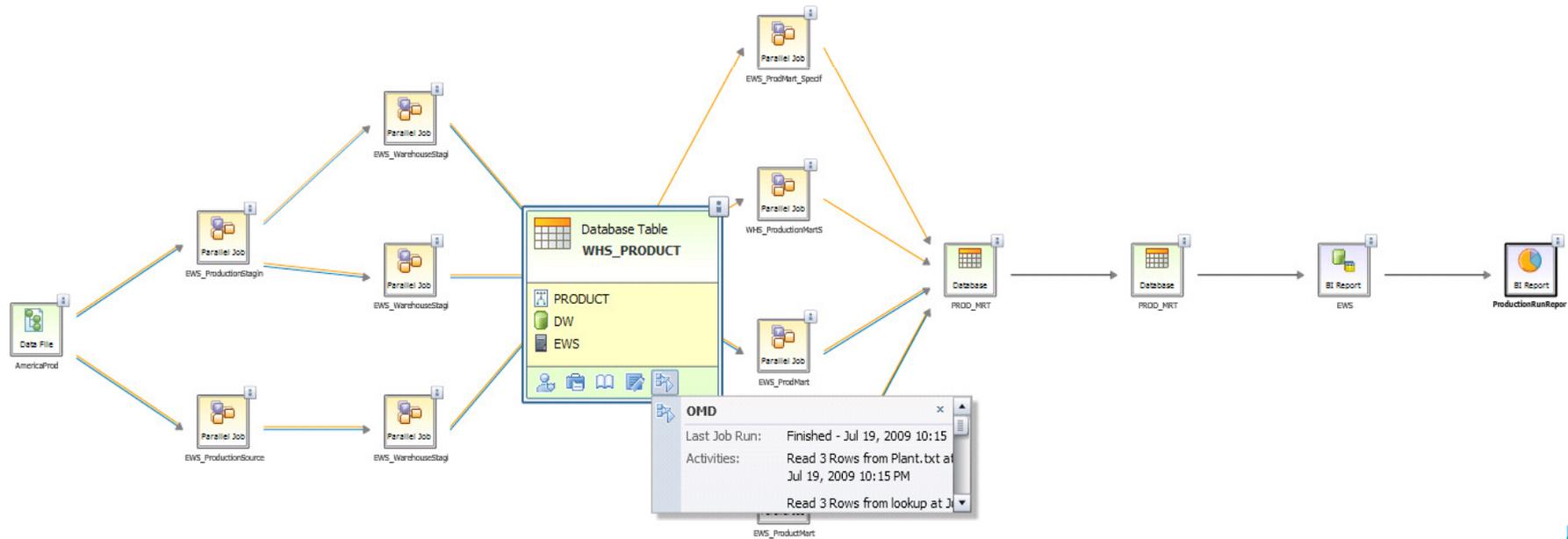
Tables		Read by Job	Written by Job		Table Analysis Summary
Asset Name	Description	Steward	Schemas	Tables	
<input type="checkbox"/> BANKDATA	Bank Data Warehouse	Richard Keith	BANK BANK1 BANK2	CUSTOMER BANKDEMOACCOUNTS BANKDEMOCHECKING BANKDEMOSAVINGS	CUSTOMER BANKDEMOACCOUNTS BANKDEMOCHECKING BANKDEMOSAVINGS
<input type="checkbox"/> DW	Data Warehouse	Richard Keith	PRODUCT SALES	PLANT PRODUCTION WHS_PRODUCT WHS_SALES	WHS_PRODUCT WHS_SALES
<input type="checkbox"/> DW_MART	Data Mart Reporting Data	Richard Keith	SCHEMA1	PROD_MRT SALES_MRT	
<input type="checkbox"/> Report_Mart			SCHEMA1	PROD_MRT SALES_MRT	
<input type="checkbox"/> SALES			SCHEMA1	SLS_LOOKUP	





# 跨工具冲突分析

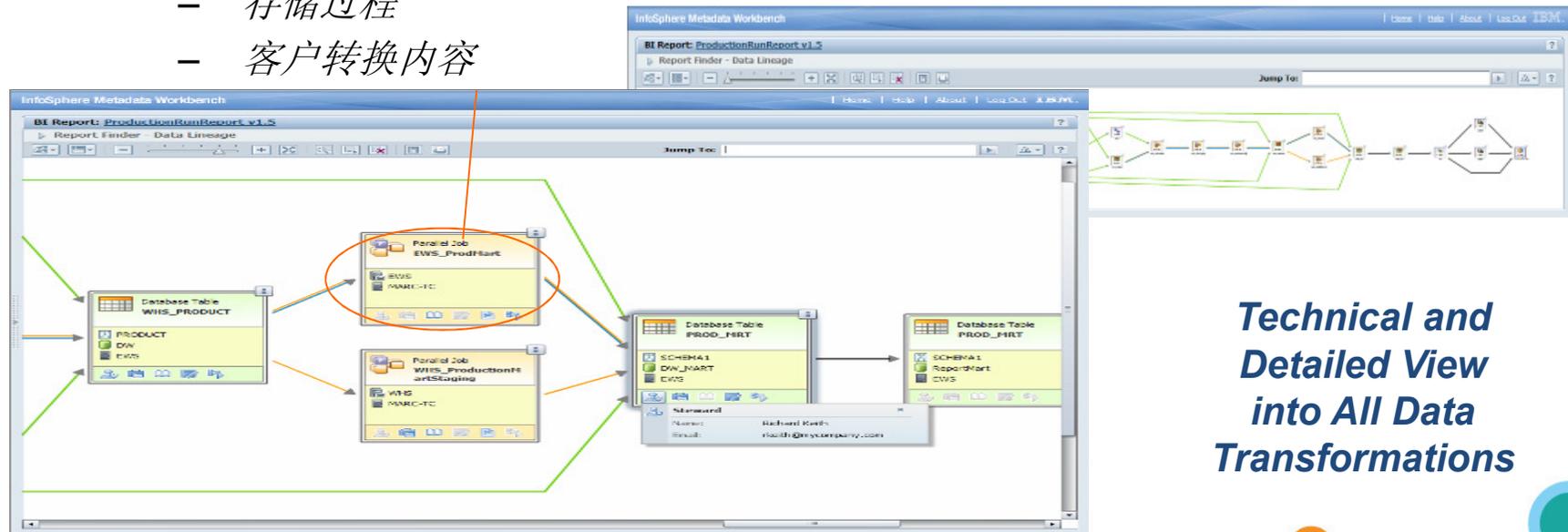
- 评估改造的冲突,降低分险
- 展示对后续应用和BI报表的影响
- 冲突区域导航和下钻





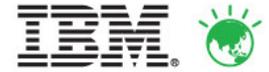
# 数据世系

- 可视化扩展到不同类型的**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

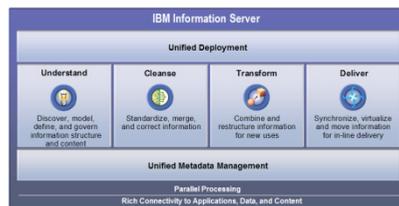
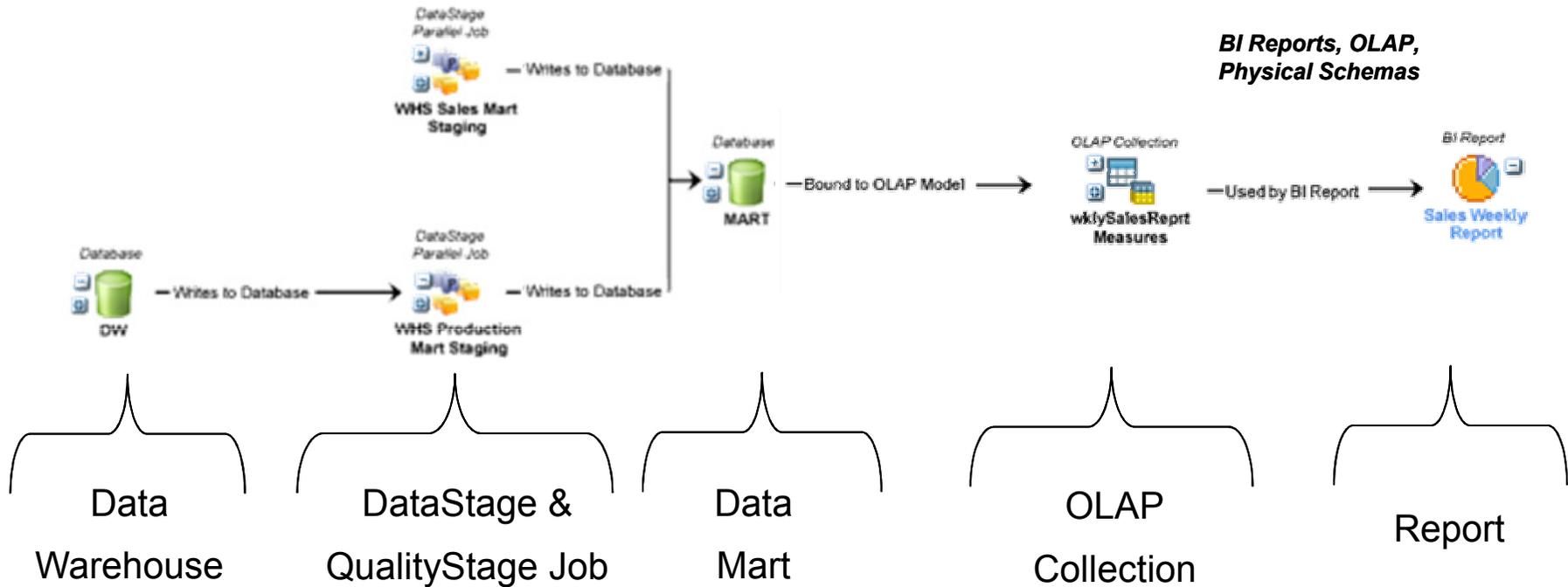


**Technical and Detailed View into All Data Transformations**

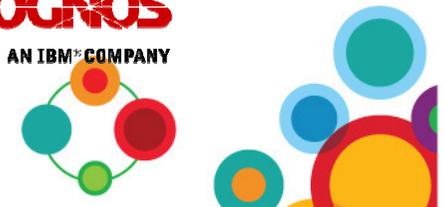




# 端到端的数据血缘分析



IBM InfoSphere Information Server



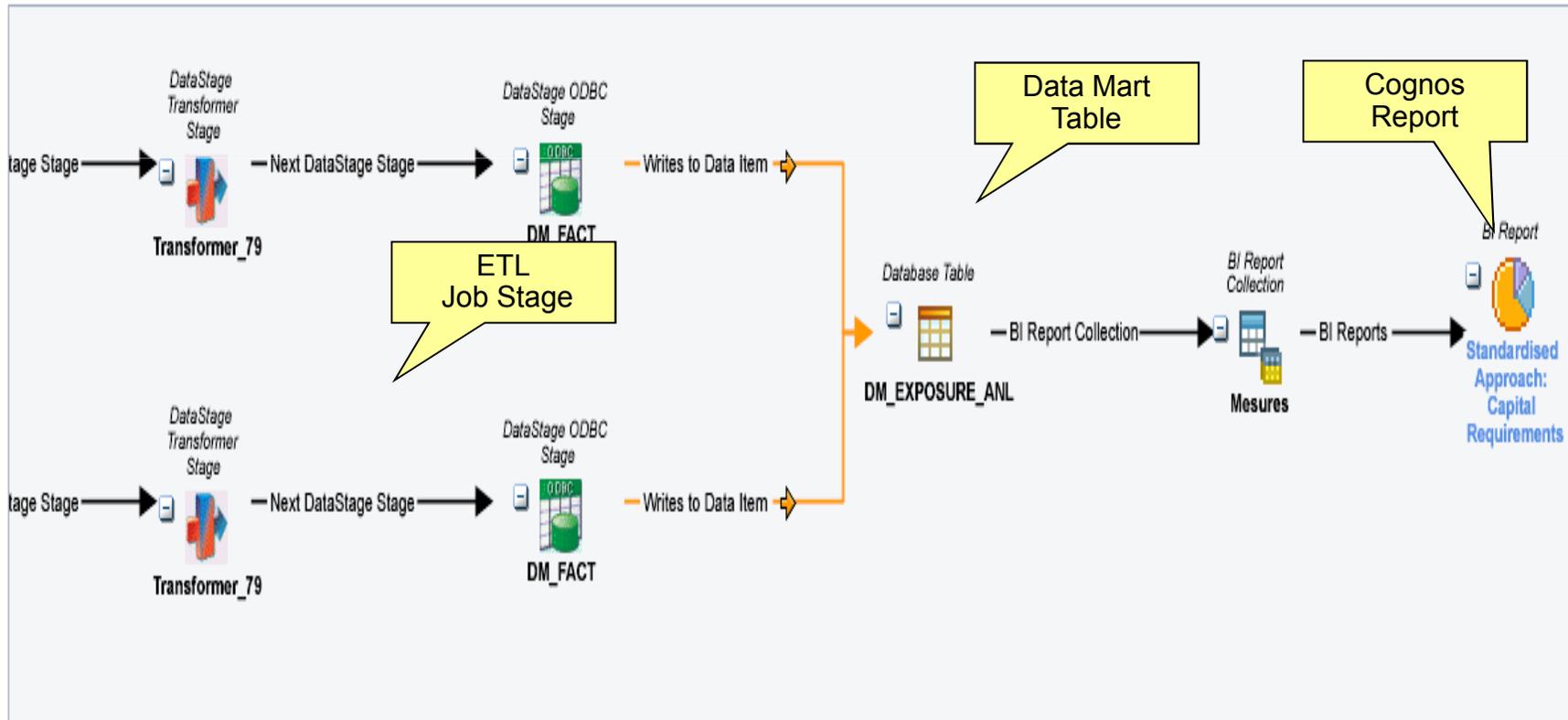


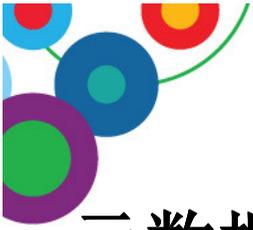
# Metadata Workbench 图形化世系图 (部分Snapshot)

BI Report: Standardised Approach: Capital Requirements

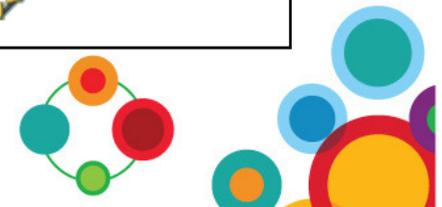
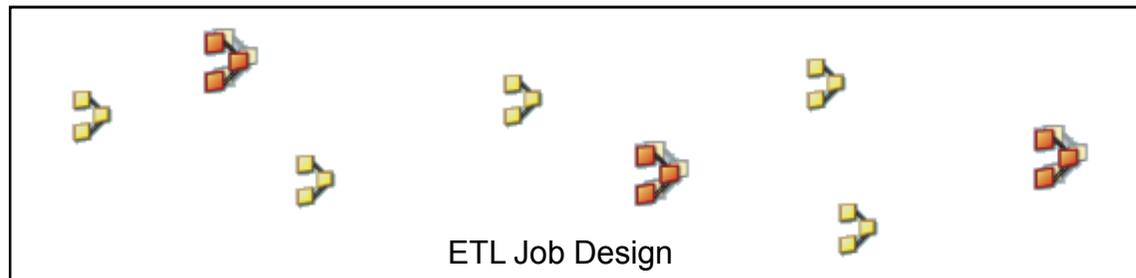
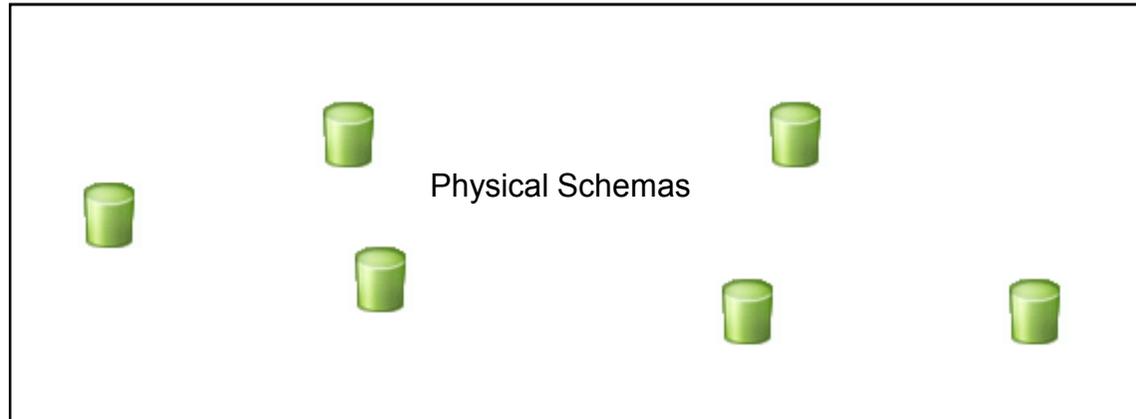
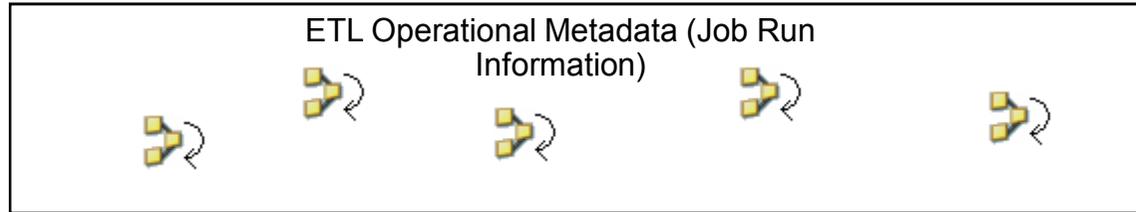
Report Finder - Data Lineage

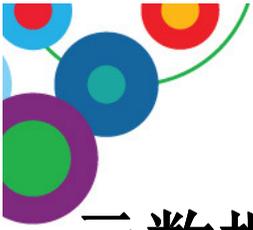
Data Lineage for: Standardised Approach: Capital Requirements



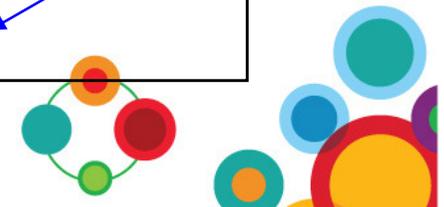
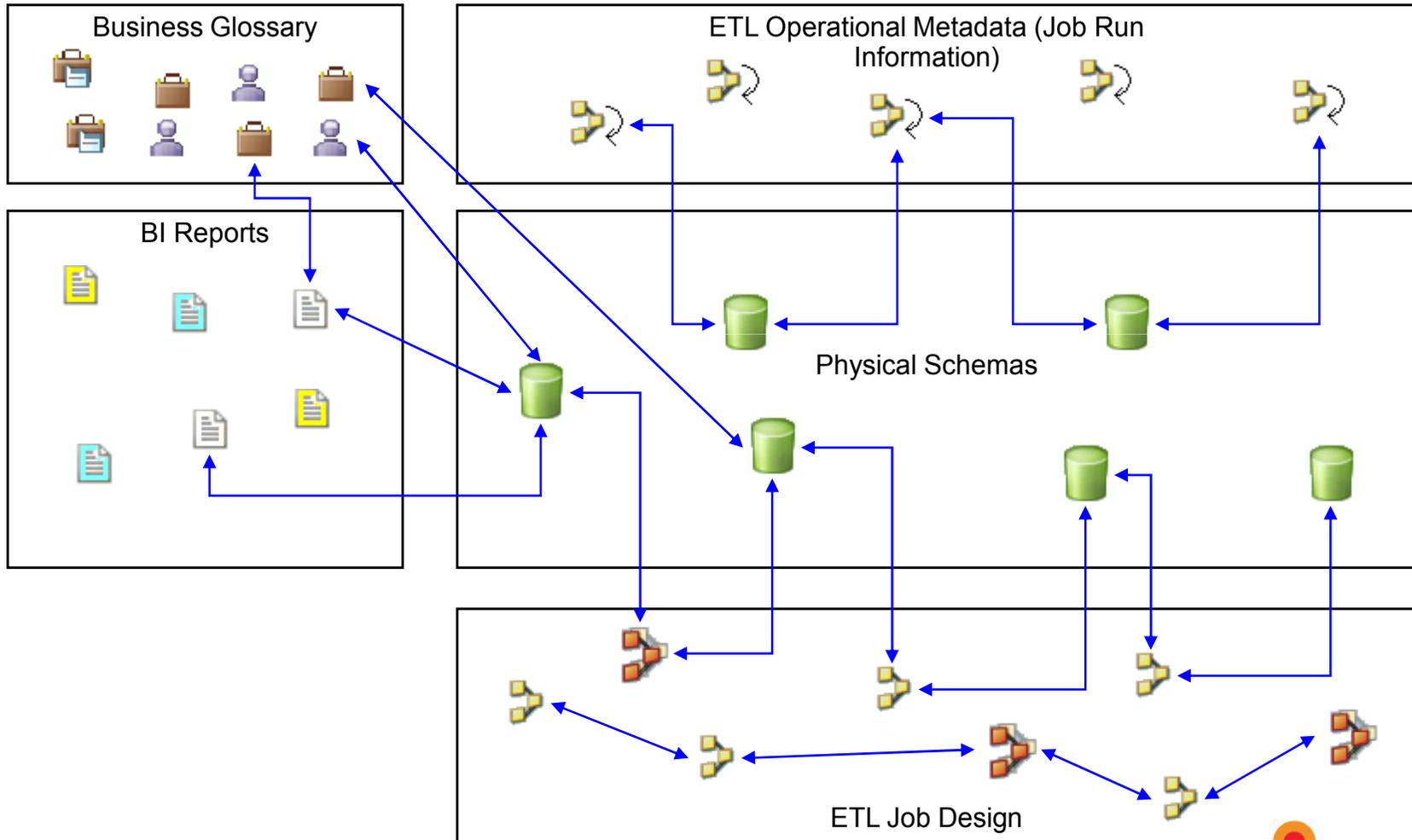


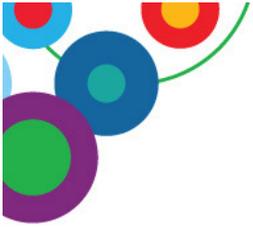
# 元数据区域





# 元数据互动





## 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**<sup>™</sup>  
software

谢谢!

*Trusted Information*

