IBM

您的信息 您的智慧

2011 IBM 信息管理与业务分析论坛

如何获取全视角的商业智能

一商业智能领域中数据集成的重要性

丁朝阳

IBM SWG 高级信息集成咨询顾问

纠结：如何理解我们拥有的信息资产

“数据看来不正确” – 业务用户

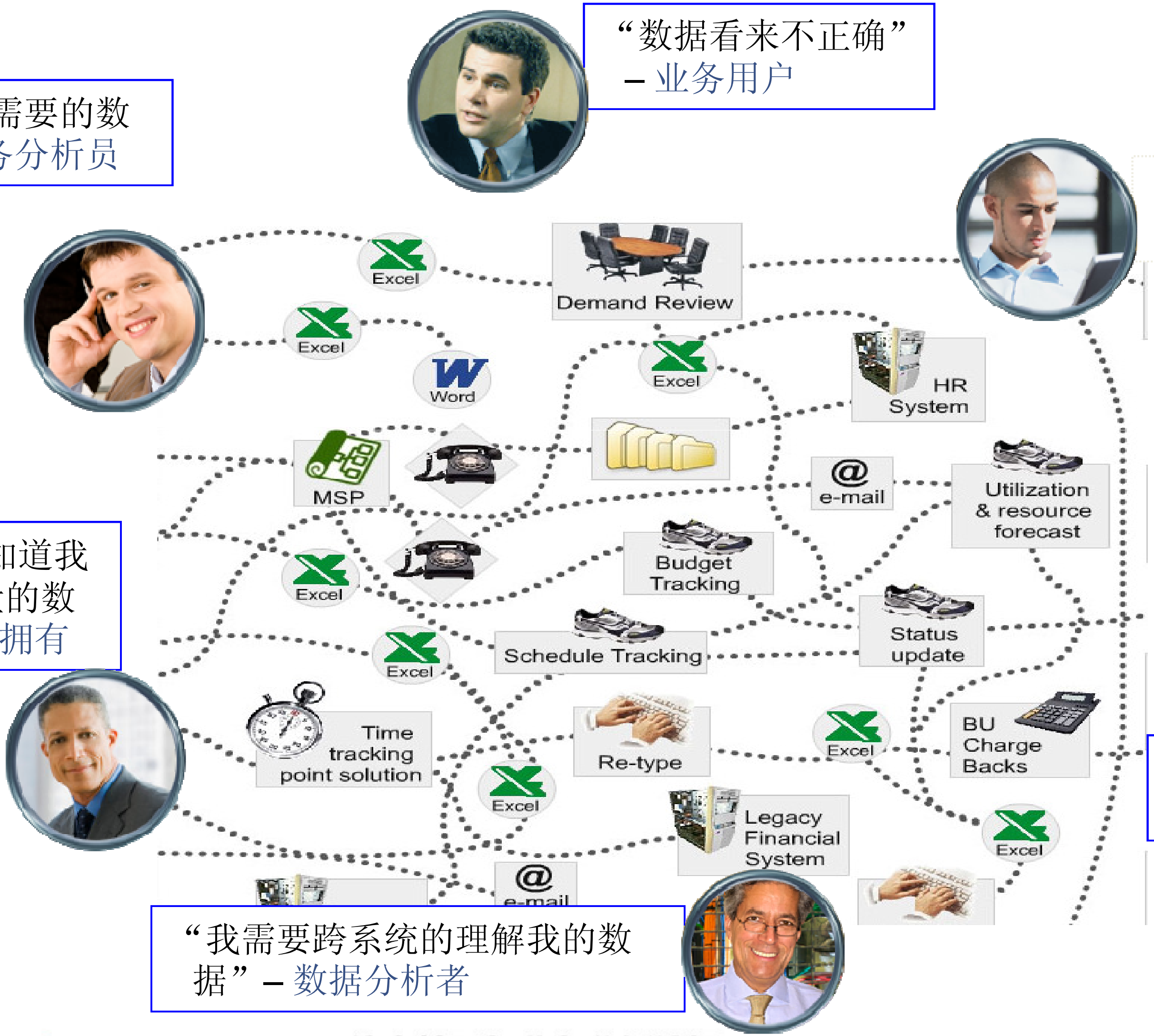
“我没有我需要的数据” – 业务分析员

“我们没有利用我们的信息” – 架构师

“我如何能知道我拥有高质量的数据” –数据拥有者

“我不了解业务人员要什么?” – 开发者

“我需要跨系统的理解我的数据” – 数据分析者



如果信息缺乏管理，会带给我们什么？

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

痛失商机

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

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

可信赖的信息是什么？

富有洞察（Insightful）

Derive meaning from information challenges

有内涵（In Context）

Real-time delivery of relevant information when and where it’s needed

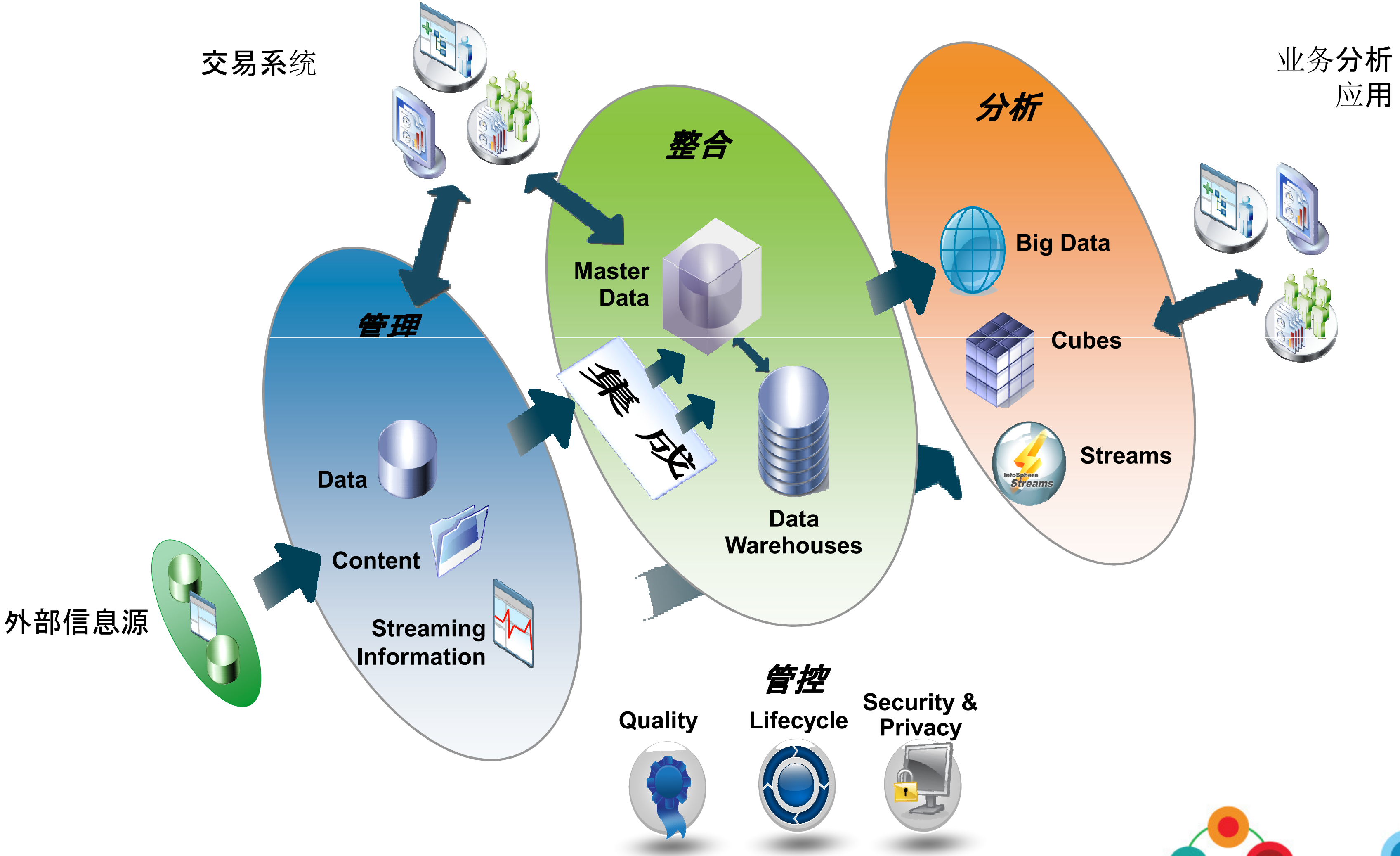
完整（Complete）

Related information reconciled into a single and holistic view

准确（Accurate）

Complex and disparate data transformed, cleansed and delivered

你需要一个对信息灵活管理，整合，分析的平台



对症下药：8个有用的良方（最佳实践）

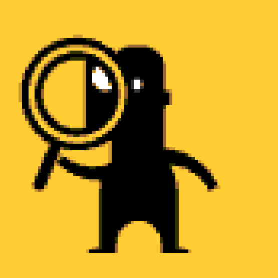
没有解决所有问题的万能药

需要从多方面对症下药

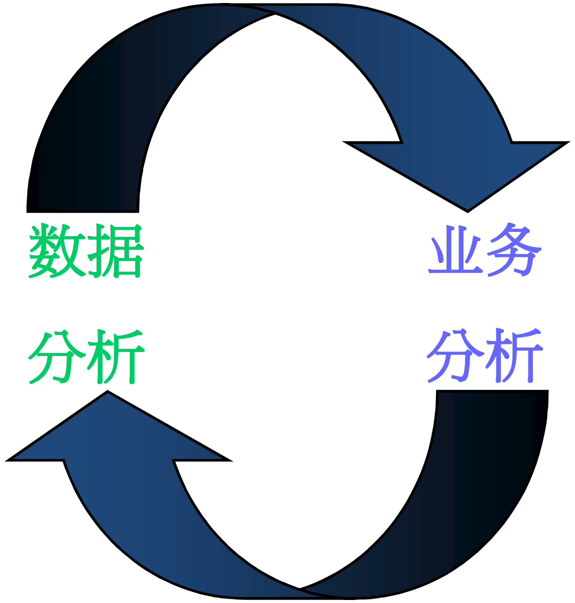
找出你最最痛心的问题

首先搞定它！

策略#1 －深刻了解源系统



1. 发现数据的实际特征



数据分析

业务分析

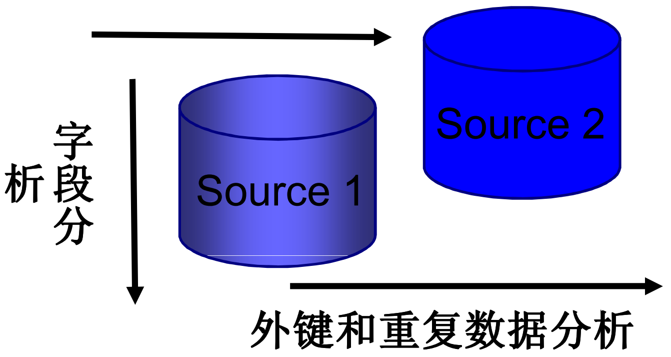
2. 确保数据能够符合已知的业务规则

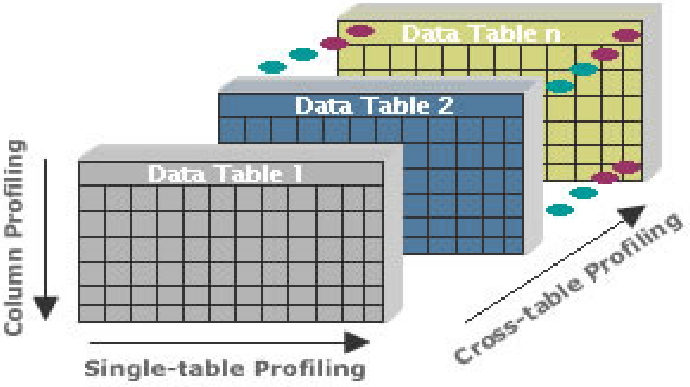
3. 报告当前的数据现状

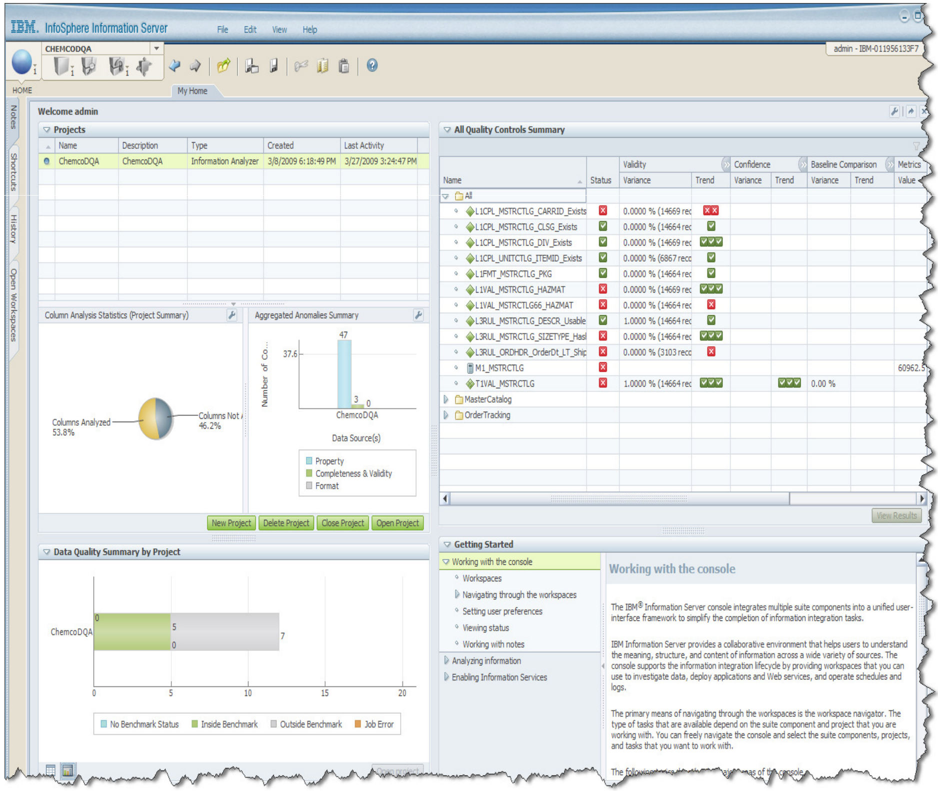
最佳实践：　自动的数据特征发现

忠告：　你没有时间和金钱以及足够的精力去手工检测数据

勿需编程







Welcome admin

Projects

Name	Description	Type	Created	Last Activity
ChemcoDQA	ChemcoDQA	Information Analyzer	3/8/2009 6:18:49 PM	3/27/2009 3:24:47 PM

Column Analysis Statistics (Project Summary)

Aggregated Anomalies Summary

All Quality Controls Summary

Name	Status	Validity		Confidence		Baseline Comparison		Metrics
		Variance	Trend	Variance	Trend	Variance	Trend	Value
All								
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 reco	☑					
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_SIZETYPE_Hasl	☒	0.0000 % (14664 rec	☑☑☑					
L3RUL_ORDHDR_OrderDt_LT_Ship	☒	0.0000 % (3103 reco	☒					
M1_MSTRCTLG	☒							60962.5
T1VAL_MSTRCTLG	☒	1.0000 % (14664 rec	☑☑☑		☑☑☑	0.00 %		
MasterCatalog								
OrderTracking								

Data Quality Summary by Project

Getting Started

Working with the console

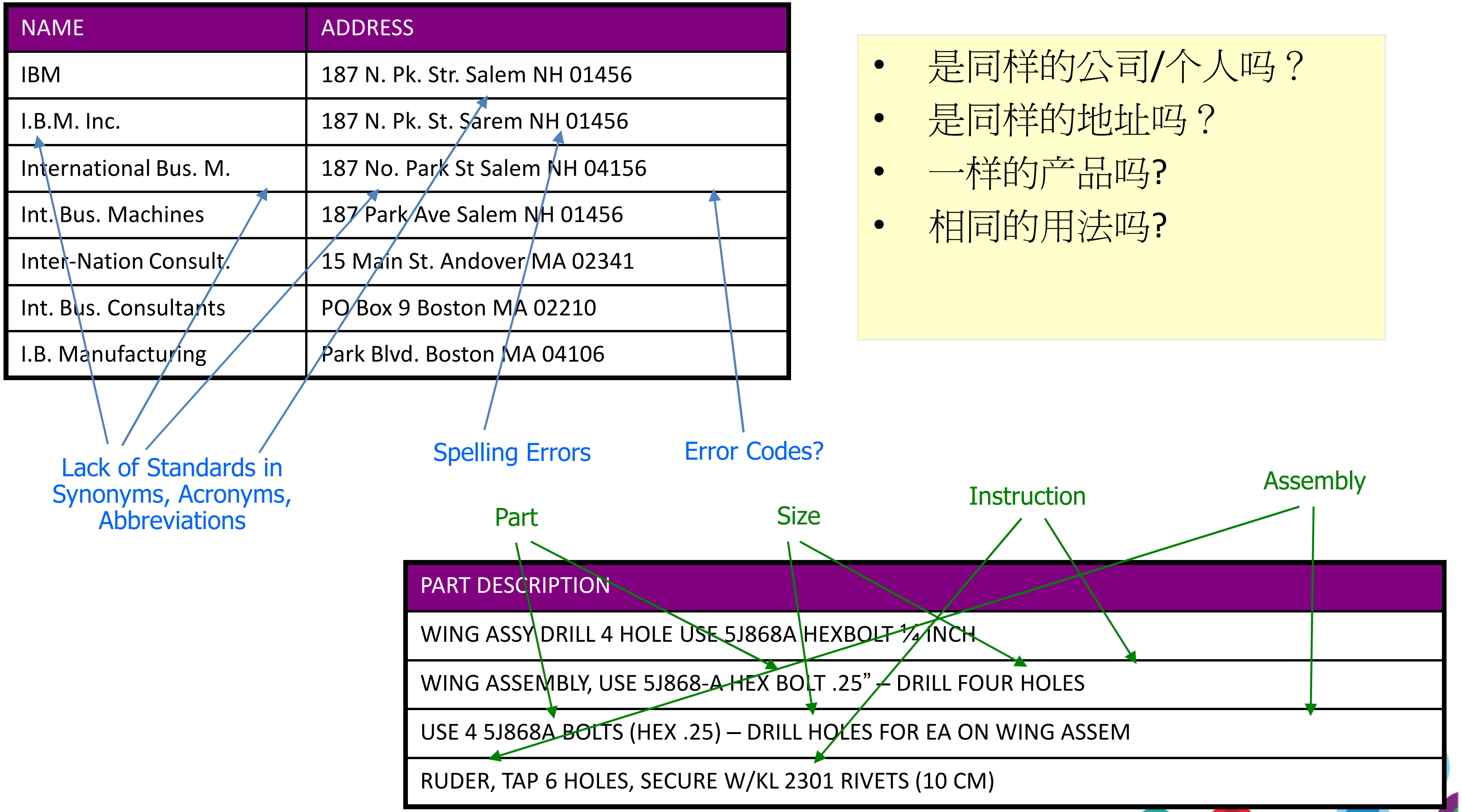
The IBM® Information Server console integrates multiple suite components into a unified user-interface framework to simplify the completion of information integration tasks.

IBM Information Server provides a collaborative environment that helps users to understand the meaning, structure, and content of information across a wide variety of sources. The console supports the information integration lifecycle by providing workspaces that you can use to investigate data, deploy applications and Web services, and operate schedules and logs.

The primary means of navigating through the workspaces is the workspace navigator. The type of tasks that are available depend on the suite component and project that you are working with. You can freely navigate the console and select the suite components, projects, and tasks that you want to work with.

The following [illegible] major areas of the console.

策略 #2 – 内部数据质量



NAME	ADDRESS
IBM	187 N. Pk. Str. Salem NH 01456
I.B.M. Inc.	187 N. Pk. St. Sarem NH 01456
International Bus. M.	187 No. Park St Salem NH 04156
Int. Bus. Machines	187 Park Ave Salem NH 01456
Inter-Nation Consult.	15 Main St. Andover MA 02341
Int. Bus. Consultants	PO Box 9 Boston MA 02210
I.B. Manufacturing	Park Blvd. Boston MA 04106

Lack of Standards in Synonyms, Acronyms, Abbreviations

Spelling Errors

Error Codes?

Part

Size

Instruction

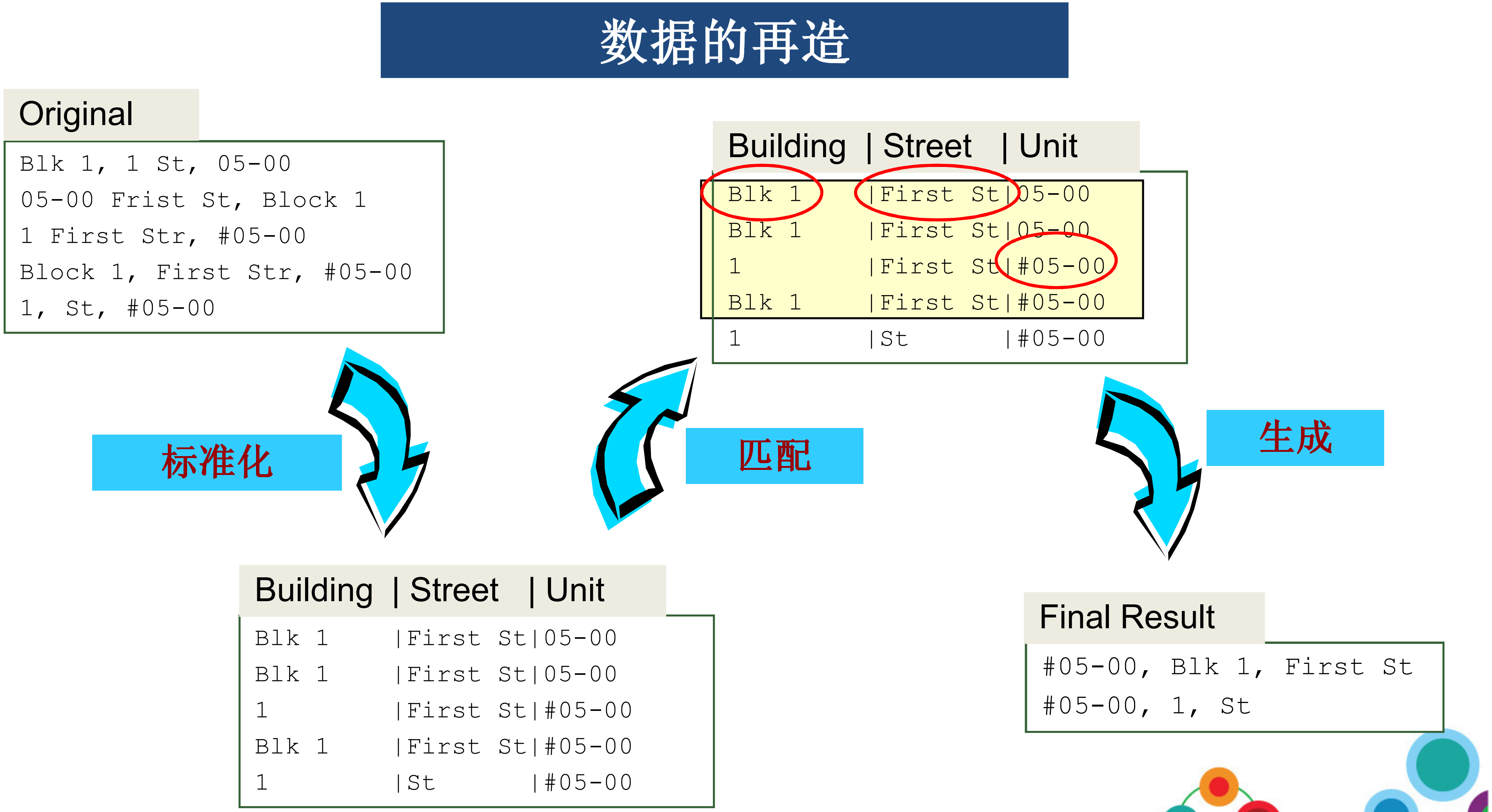
Assembly

PART DESCRIPTION
WING ASSY DRILL 4 HOLE USE 5J868A HEXBOLT ¼ INCH
WING ASSEMBLY, USE 5J868-A HEX BOLT .25" – DRILL FOUR HOLES
USE 4 5J868A BOLTS (HEX .25) – DRILL HOLES FOR EA ON WING ASSEM
RUDER, TAP 6 HOLES, SECURE W/KL 2301 RIVETS (10 CM)

- 是同样的公司/个人吗？
- 是同样的地址吗？
- 一样的产品吗?
- 相同的用法吗?

最佳实践：　数据清理

数据的再造



Original

```
Blk 1, 1 St, 05-00
05-00 Frist St, Block 1
1 First Str, #05-00
Block 1, First Str, #05-00
1, St, #05-00
```

标准化

Building	Street	Unit
Blk 1	First St	05-00
Blk 1	First St	05-00
1	First St	#05-00
Blk 1	First St	#05-00
1	St	#05-00

匹配

Building	Street	Unit
Blk 1	First St	05-00
Blk 1	First St	05-00
1	First St	#05-00
Blk 1	First St	#05-00
1	St	#05-00

生成

Final Result

```
#05-00, Blk 1, First St
#05-00, 1, St
```

策略 #3 – 公共元数据共享

From Data Model

Customer
CustomerNumber
Name
Address
Comments

The Identifier of customers that are tracked for ordering purposes. Corporate customer identifiers are assigned by the Sales Data Controller according to the corporate data description and naming policy for reference identifiers.

From ETL Tool

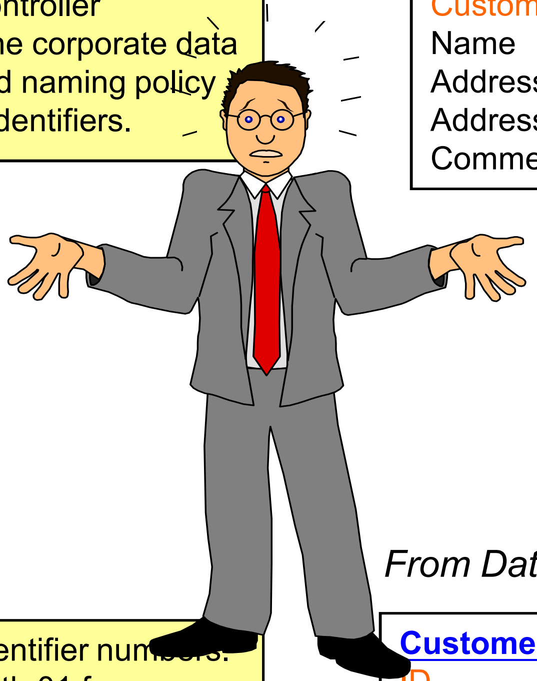
CustomerTbl
CustomerID
Name
Address
Address1
Comments

Unique identifier of customers that are tracked for ordering purposes. Values start with 02 for non-Corporate customers and 01 for Corporate customers.

哪个元数据信息是正确的？

哪个是当前正在使用的？

哪个是即将用到的？



From BI Tool

CustomerDetails
CustomerNumber
Name
Address
Remarks

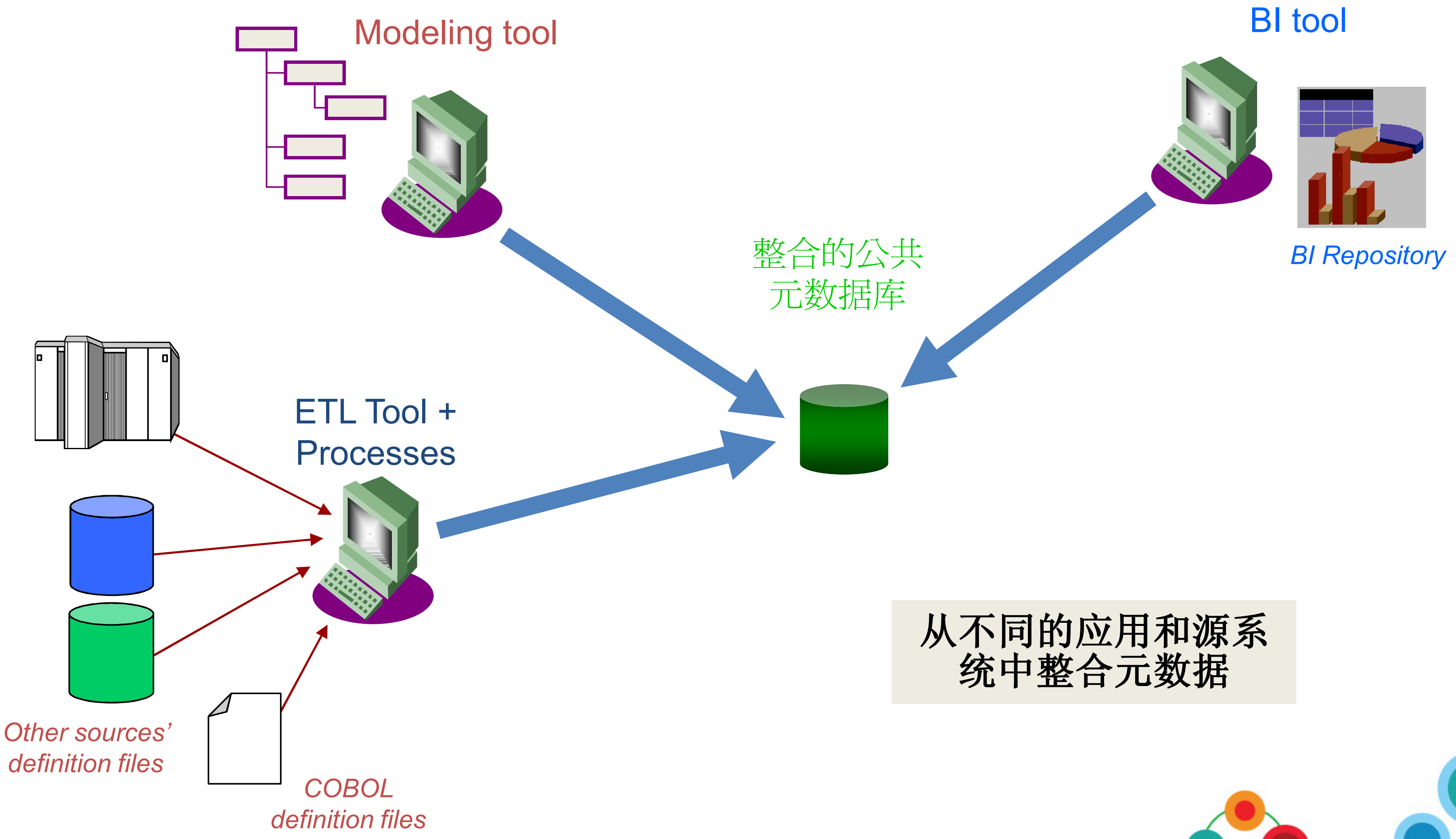
Customer's identifier numbers. Values start with 01 for Corporate customers, 02 for non-Corporate customers, 03 for overseas-based Customers.

From Database

Customer
ID
Name
Address1
Address2
Descr

<NULL>

最佳实践：　建立一个公共元数据库



从不同的应用和源系统中整合元数据

最佳实践:建立统一的业务术语

Database = DB2

Schema = NAACCT

Table = DLYTRANS

Column = TAXVL

data type = Decimal (14,2)

Derivation: SUM(TRNTXAMT)

Category: Costs

Term: Tax Expense

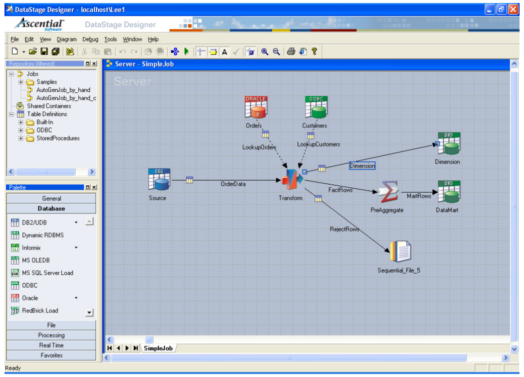
Full Name: Tax to be paid on Gross Income

“The expense due to taxes …..”

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

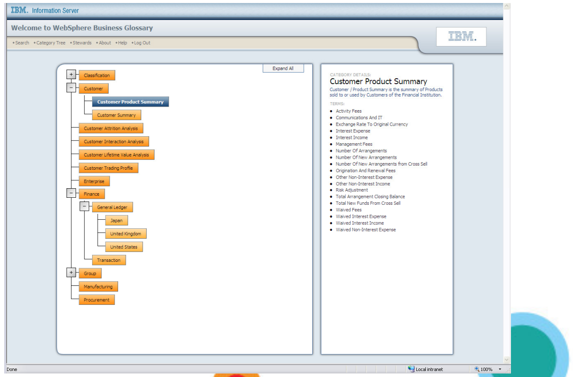
Status: CURRENT

在业务人员和技术人员之间建议一套统一的术语表!



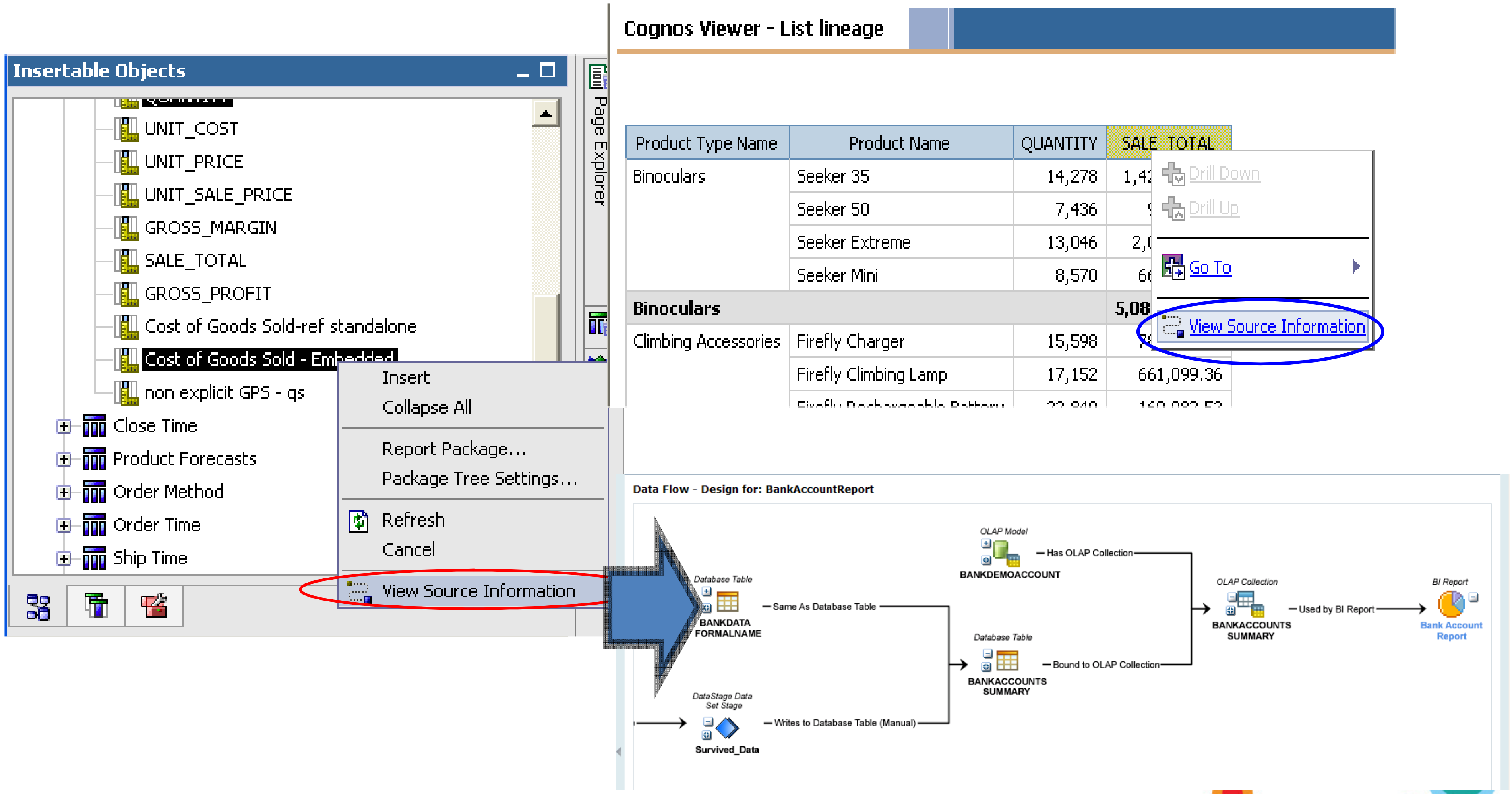
InfoSphere DataStage

共享元数据



InfoSphere Business Glossary

建立对数据血统的追踪



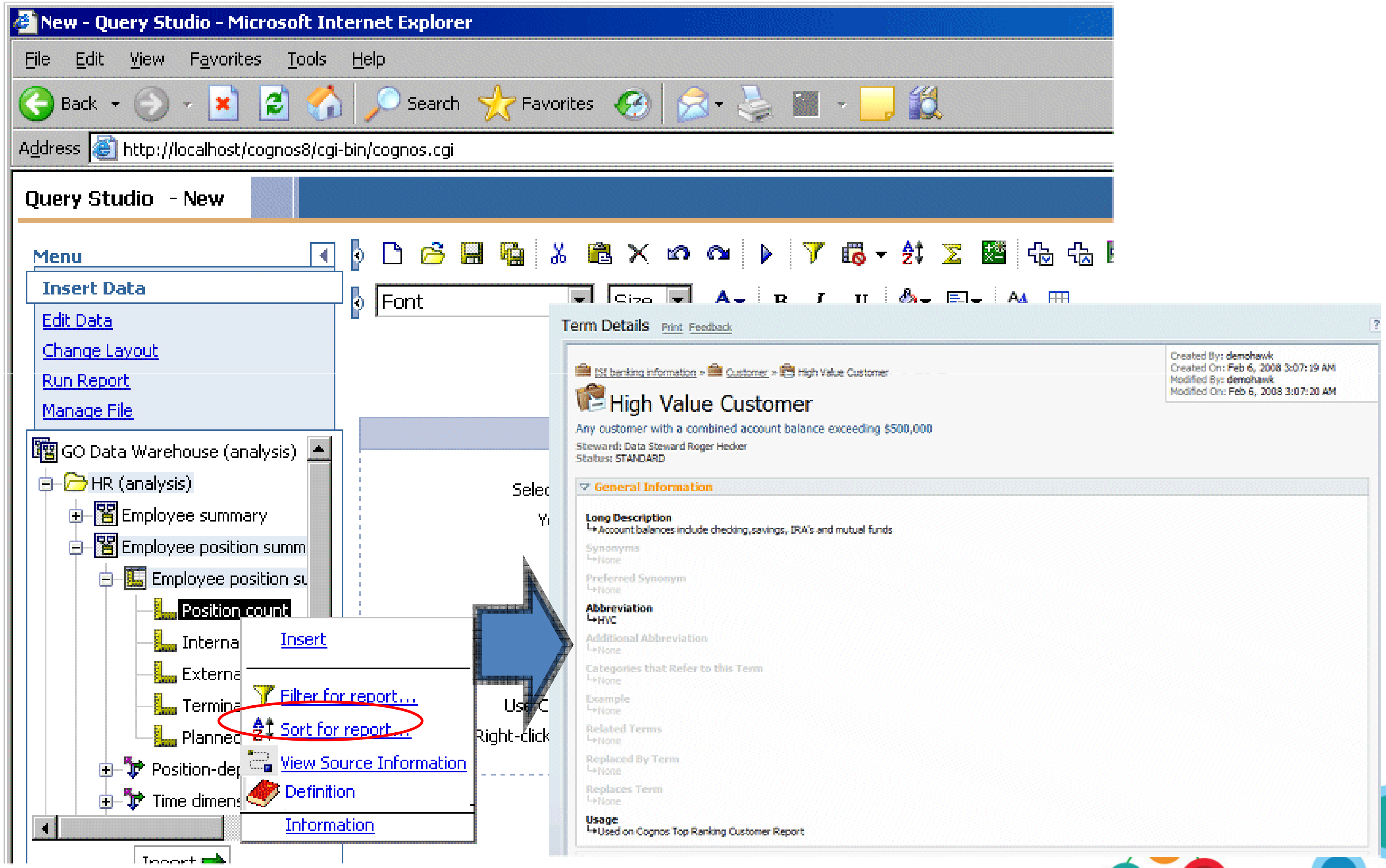
Insertable Objects

Cognos Viewer - List lineage

Product Type Name	Product Name	QUANTITY	SALE_TOTAL
Binoculars	Seeker 35	14,278	1,4
	Seeker 50	7,436	
	Seeker Extreme	13,046	2,(
	Seeker Mini	8,570	6(
Binoculars			**5,08**
Climbing Accessories	Firefly Charger	15,598	7(
	Firefly Climbing Lamp	17,152	661,099.36

Data Flow - Design for: BankAccountReport

在BI应用中访问业务元数据



Term Details Print Feedback

ISI banking information » Customer » High Value Customer

Created By: demohawk
Created On: Feb 6, 2008 3:07:19 AM
Modified By: demohawk
Modified On: Feb 6, 2008 3:07:20 AM

High Value Customer

Any customer with a combined account balance exceeding $500,000

Steward: Data Steward Roger Hecker
Status: STANDARD

General Information

Long Description
↳Account balances include checking,savings, IRA's and mutual funds

Synonyms
↳None

Preferred Synonym
↳None

Abbreviation
↳HVC

Additional Abbreviation
↳None

Categories that Refer to this Term
↳None

Example
↳None

Related Terms
↳None

Replaced By Term
↳None

Replaces Term
↳None

Usage
↳Used on Cognos Top Ranking Customer Report

策略#4 – 与任何地方的任何系统相连

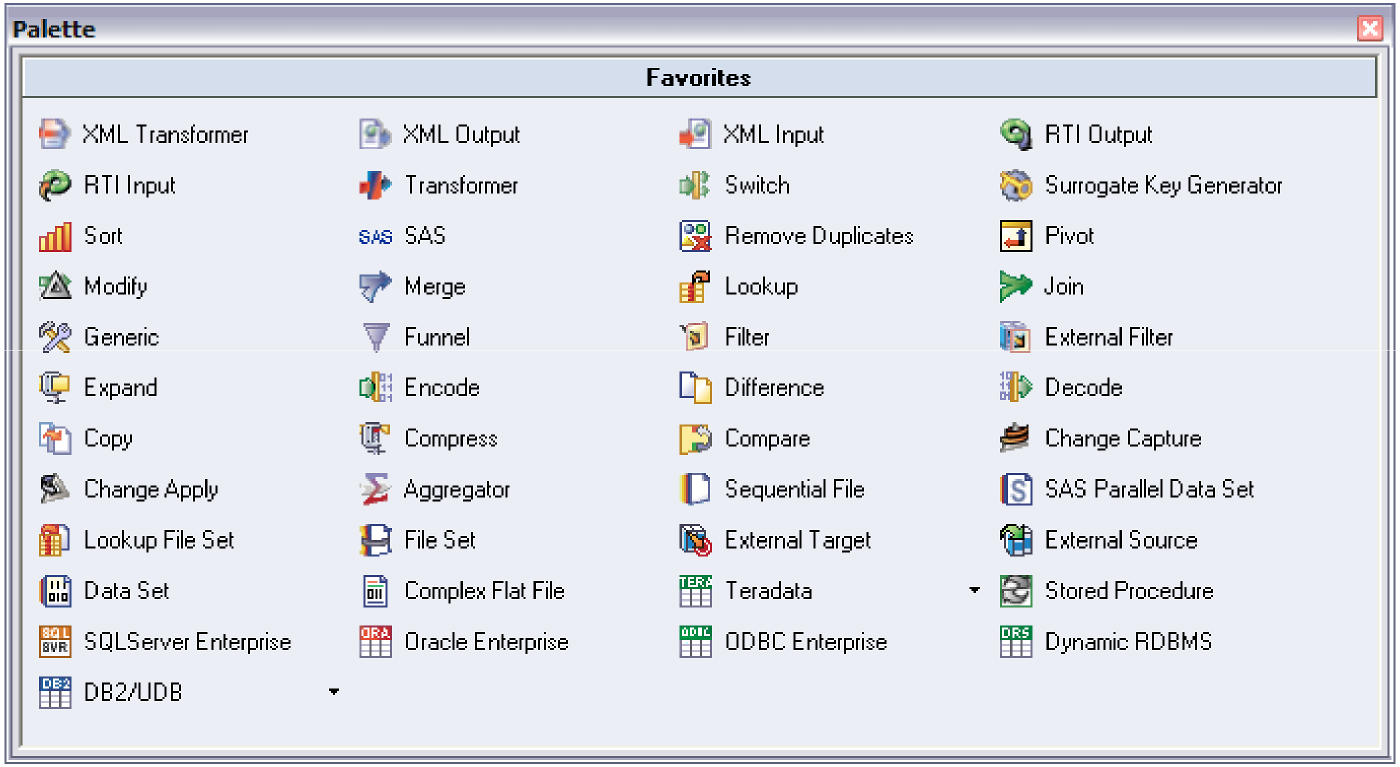
DB2, Informix, Netezza, ODBC, Oracle, Red Brick, SAS, Sybase, Teradata, etc

Adabas, Allbase/SQL, Datacom/DB, DB2/400, DB2/OS390, Essbase, FOCUS, IDMS/SQL, IMS, NonStopSQL, RDB, VSAM, etc

WebSphere MQ, SeeBeyond, JMS, XML, EJB, Web Services, EXML, XMLS, EDI, SWIFT, etc

Oracle Applications, PeopleSoft, SAP R/3, SAP BW, Siebel

最佳实践：　利用有丰富连接的工具



Palette

Favorites

- XML Transformer
- XML Output
- XML Input
- RTI Output
- RTI Input
- Transformer
- Switch
- Surrogate Key Generator
- Sort
- SAS
- Remove Duplicates
- Pivot
- Modify
- Merge
- Lookup
- Join
- Generic
- Funnel
- Filter
- External Filter
- Expand
- Encode
- Difference
- Decode
- Copy
- Compress
- Compare
- Change Capture
- Change Apply
- Aggregator
- Sequential File
- SAS Parallel Data Set
- Lookup File Set
- File Set
- External Target
- External Source
- Data Set
- Complex Flat File
- Teradata
- Stored Procedure
- SQLServer Enterprise
- Oracle Enterprise
- ODBC Enterprise
- Dynamic RDBMS
- DB2/UDB

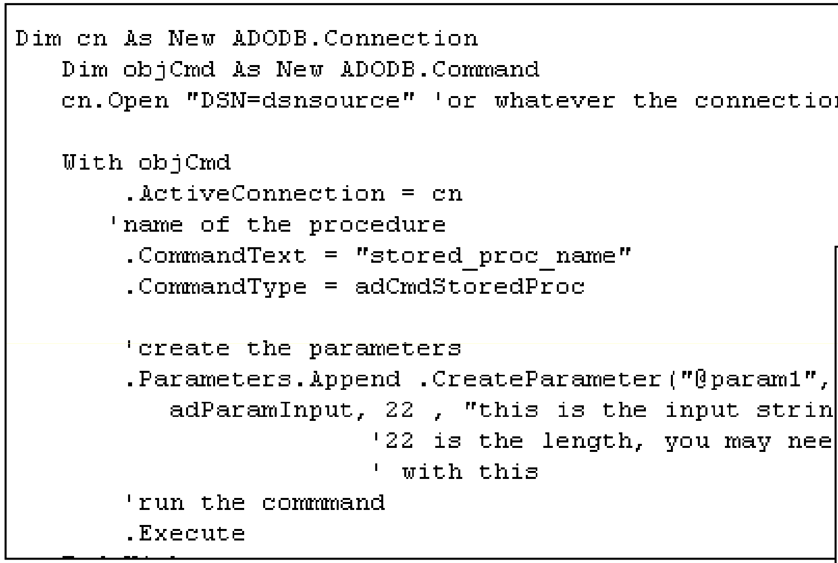
建议：

**用预建的连接器，
而不是用手工**

你希望总是为你的下一个应用和数据库的连接而担忧吗？

策略 #5 – 拒绝手工编码

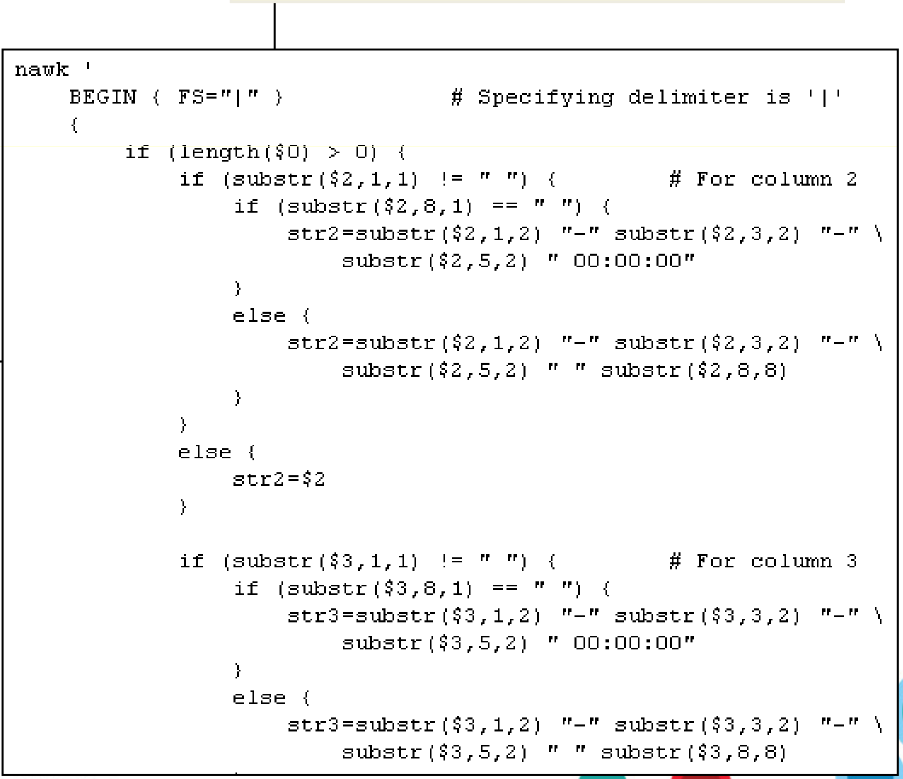
这些 Visual BASIC, Java, C++, UNIX 程序的开发成本低而且能够保证功能满足 …



```
Dim cn As New ADODB.Connection
   Dim objCmd As New ADODB.Command
   cn.Open "DSN=dsnsource" 'or whatever the connection string

   With objCmd
       .ActiveConnection = cn
      'name of the procedure
       .CommandText = "stored_proc_name"
       .CommandType = adCmdStoredProc

       'create the parameters
       .Parameters.Append .CreateParameter("@param1",
          adParamInput, 22 , "this is the input strin
                       '22 is the length, you may nee
                       ' with this
       'run the commmand
       .Execute
```



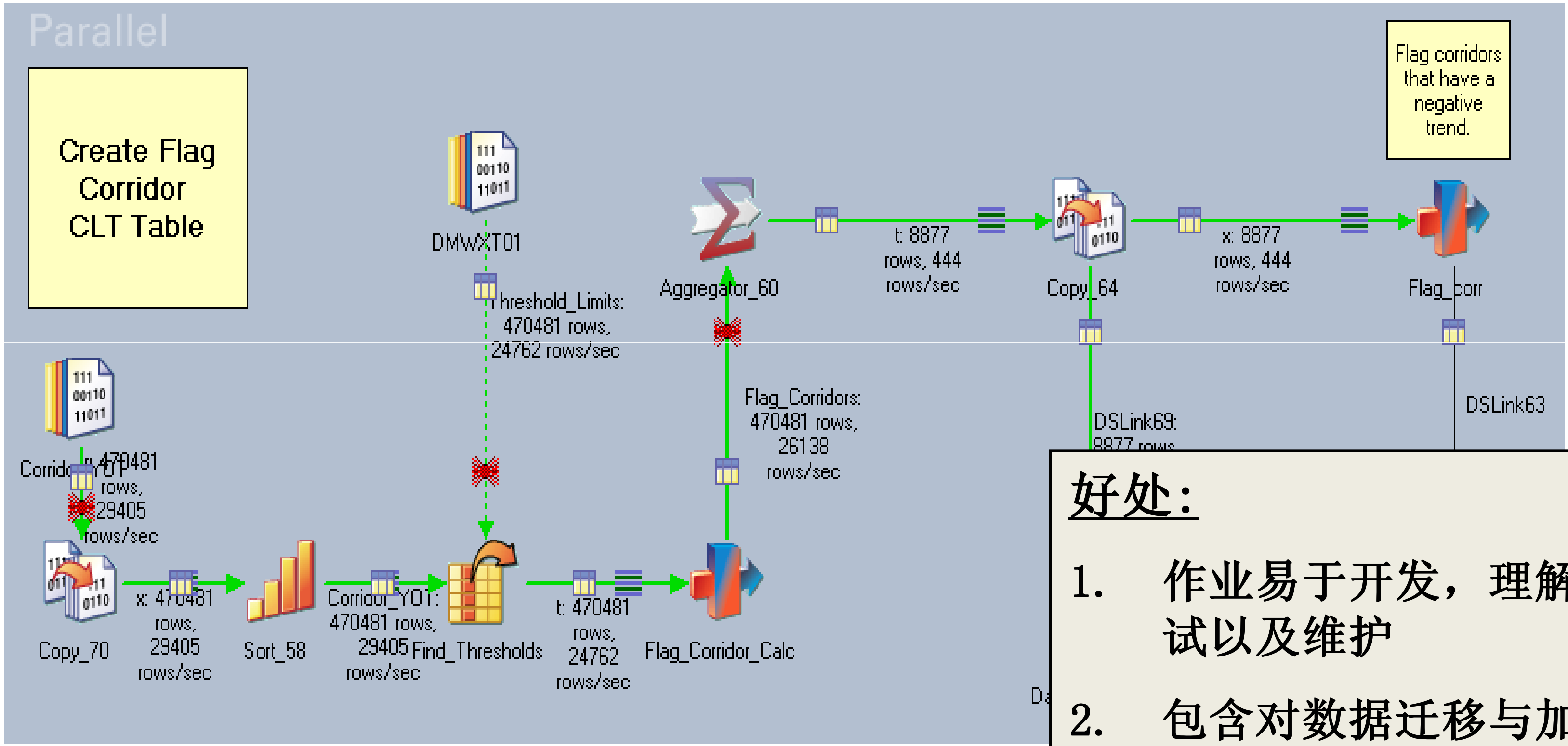
```
nawk '
    BEGIN { FS="|" }              # Specifying delimiter is '|'
    {
        if (length($0) > 0) {
            if (substr($2,1,1) != " ") {          # For column 2
                if (substr($2,8,1) == " ") {
                    str2=substr($2,1,2) "-" substr($2,3,2) "-" \
                        substr($2,5,2) " 00:00:00"
                }
                else {
                    str2=substr($2,1,2) "-" substr($2,3,2) "-" \
                        substr($2,5,2) " " substr($2,8,8)
                }
            }
            else {
                str2=$2
            }

            if (substr($3,1,1) != " ") {          # For column 3
                if (substr($3,8,1) == " ") {
                    str3=substr($3,1,2) "-" substr($3,3,2) "-" \
                        substr($3,5,2) " 00:00:00"
                }
                else {
                    str3=substr($3,1,2) "-" substr($3,3,2) "-" \
                        substr($3,5,2) " " substr($3,8,8)
```

… 但如果有新增的需求怎么办呢？

还有便宜而且运行良好吗？

最佳实践：　图形化的开发工具

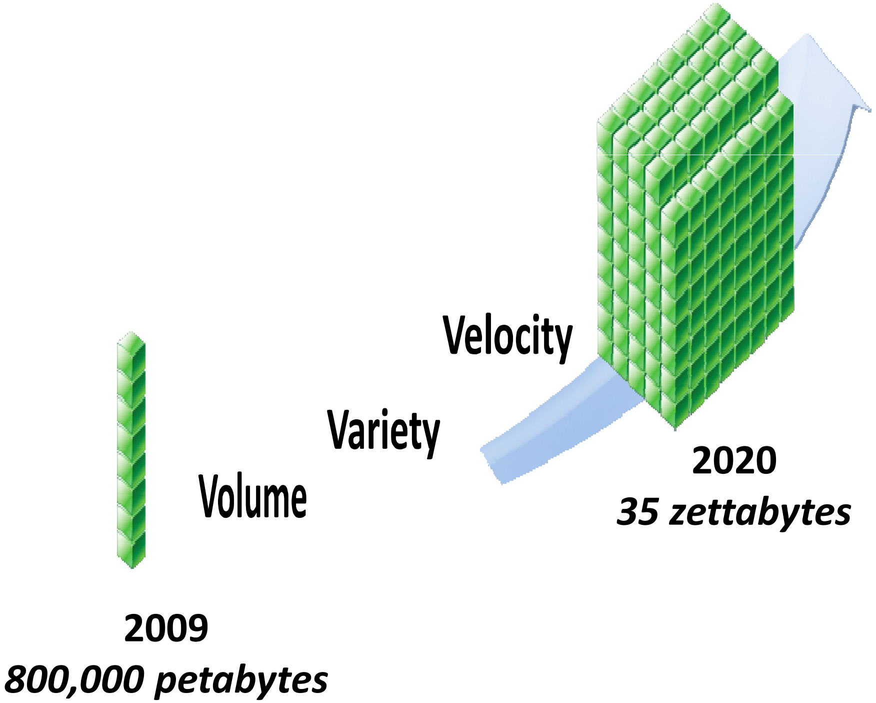

Parallel
Create Flag Corridor CLT Table
Flag corridors that have a negative trend.
DMWXT01
Threshold_Limits: 470481 rows, 24762 rows/sec
Aggregator_60
t: 8877 rows, 444 rows/sec
Copy_64
x: 8877 rows, 444 rows/sec
Flag_corr
Flag_Corridors: 470481 rows, 26138 rows/sec
DSLink69: 8877 rows
DSLink63
Corridor_Y01
r: 470481 rows, 29405 rows/sec
Copy_70
x: 470481 rows, 29405 rows/sec
Sort_58
Corridor_Y01: 470481 rows, 29405 rows/sec
Find_Thresholds
t: 470481 rows, 24762 rows/sec
Flag_Corridor_Calc
Da

好处：

1. 作业易于开发，理解，调试以及维护
2. 包含对数据迁移与加工的最佳实践

策略 #6 – 高可扩展的功能

44x 未来十年数据的增长速度

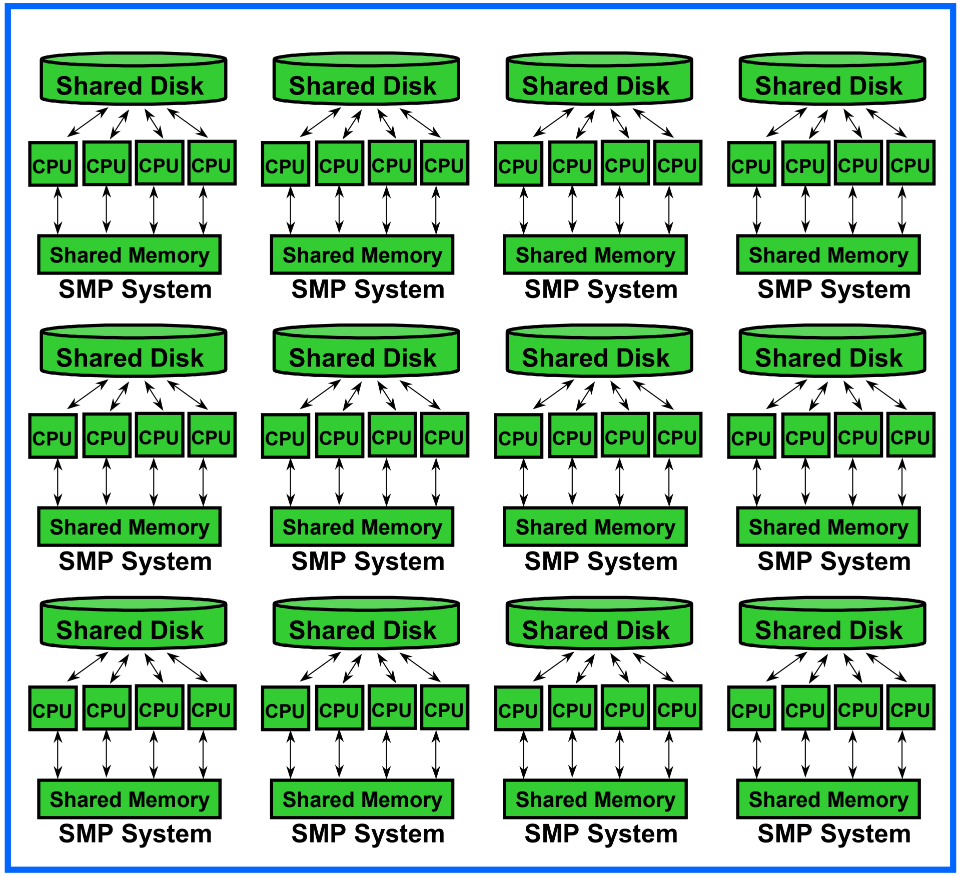


预言：

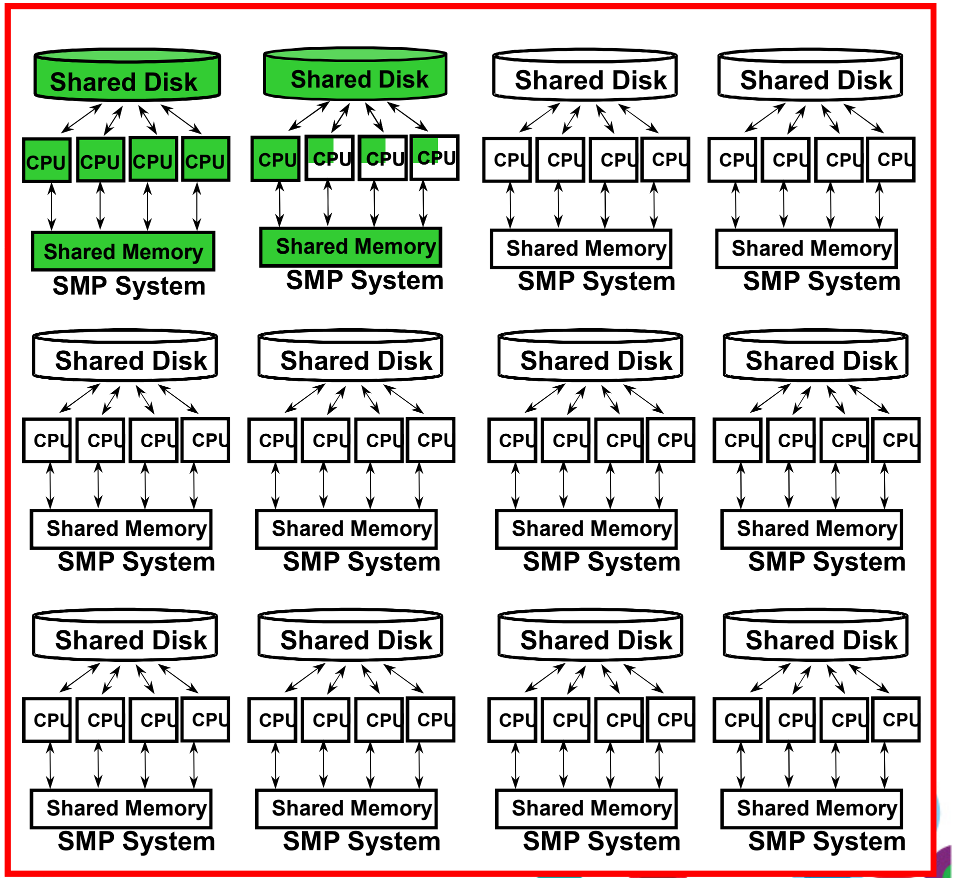
你的数据不可能会越来越小

最佳实践： 并发机制

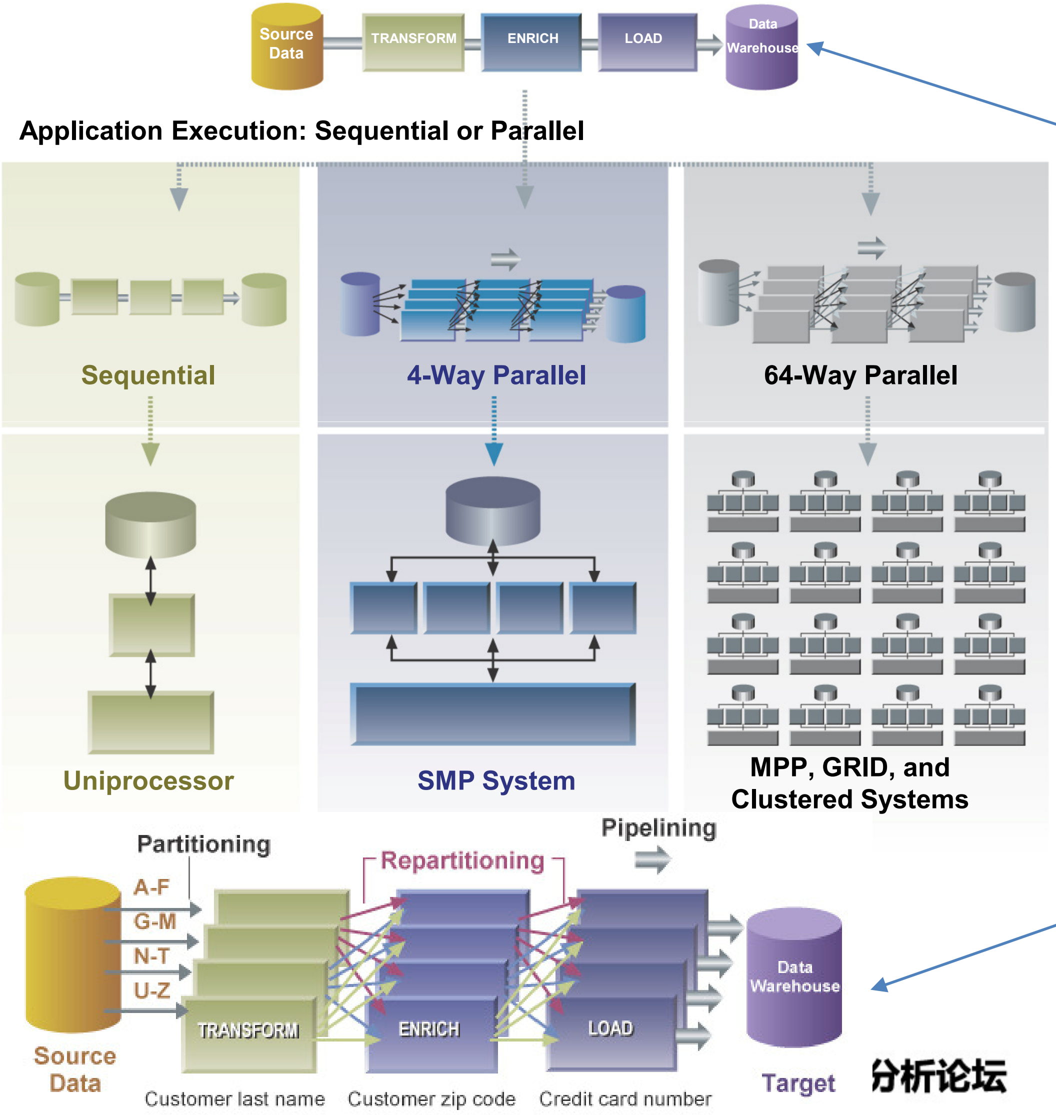
你需要要的是这样的



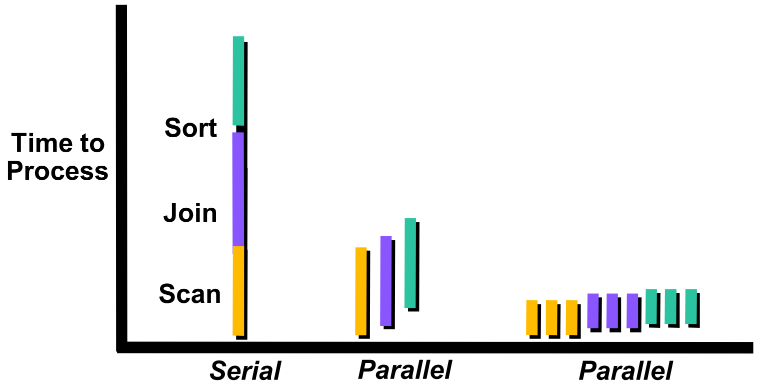
而不是这样的



最佳实践:并发机制



顺序的开发



在运行时自动的并行处理

策略 #7 – 建立“实时”的架构

动态数据仓库 & 商业智能 实时报表

- *昨天的数据已经过时，不足以满足今天的决策的需要*

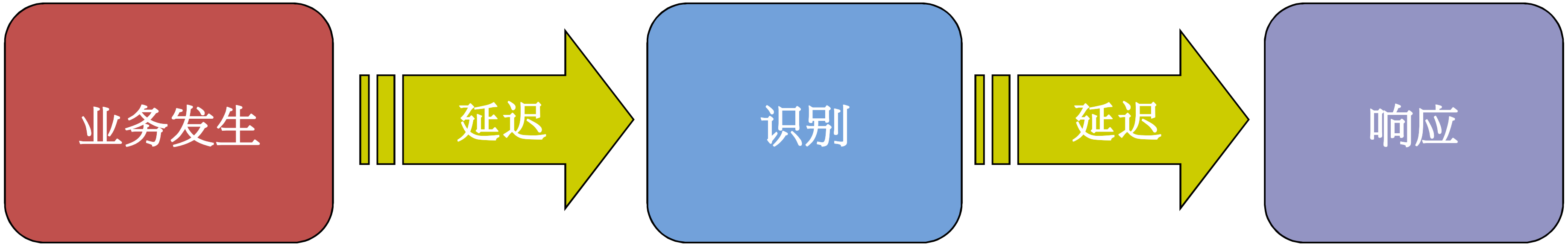
生产数据与电子商务数据整合

- *我们希望从网页上看到最新的数据*

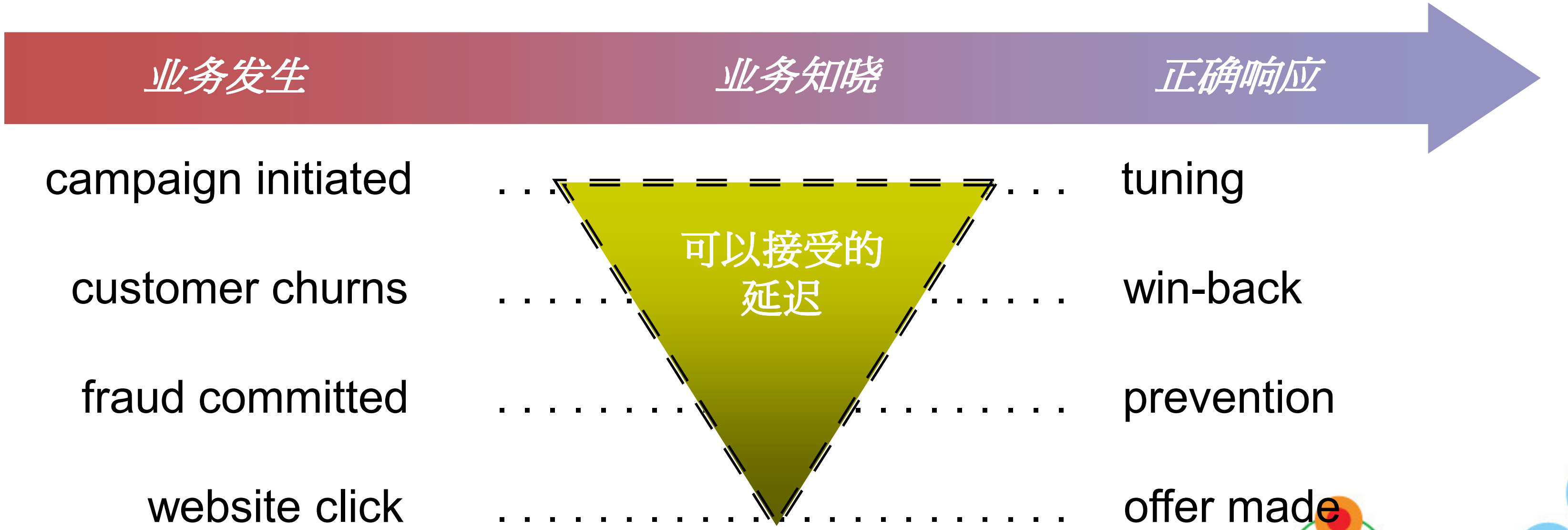
实时事件检测

- *我们希望前摄性的监控和响应业务的变化*

最佳实践：　实时机制



***Latency** is defined as the elapsed time between when an event occurs and when an appropriate response or action is made*



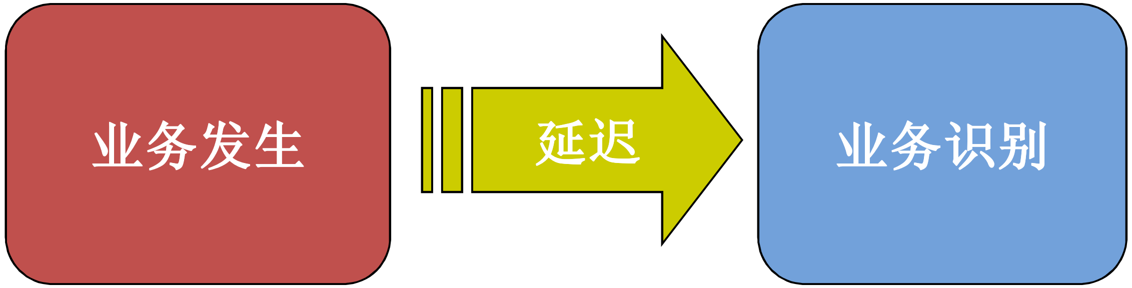
campaign initiated . . . . tuning

customer churns . . . . win-back

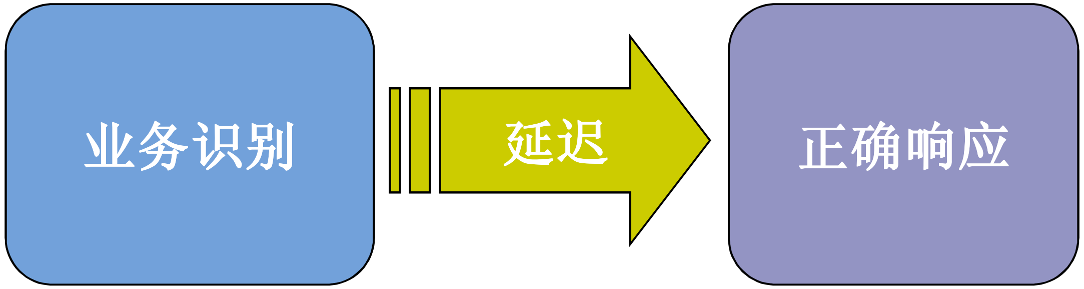
fraud committed . . . . prevention

website click . . . . offer made

最佳实践：实时机制



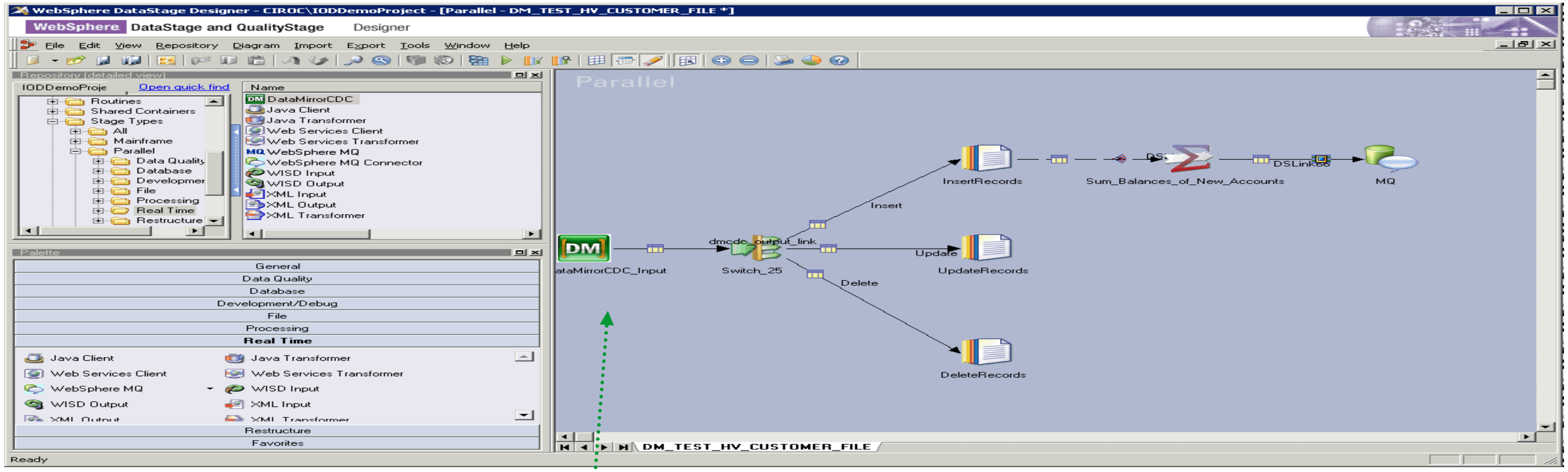
1. 提升对业务事件的识别能力

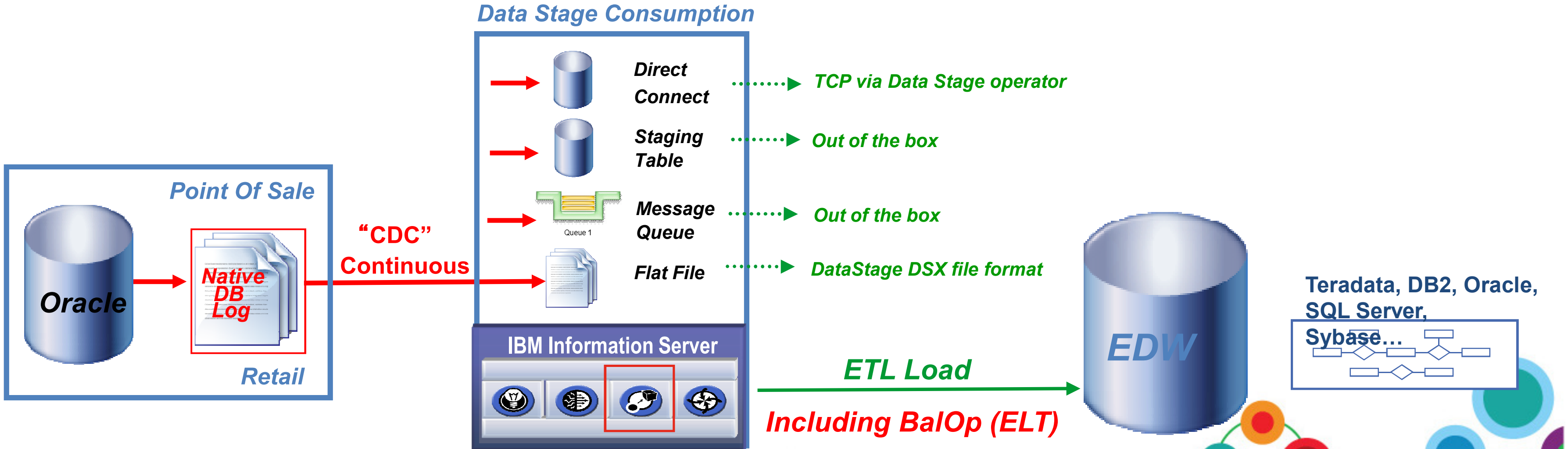


2. 提升对事件的响应能力

实时的变化数据捕获与ETL工具的结合

Information Server Change Data Capture



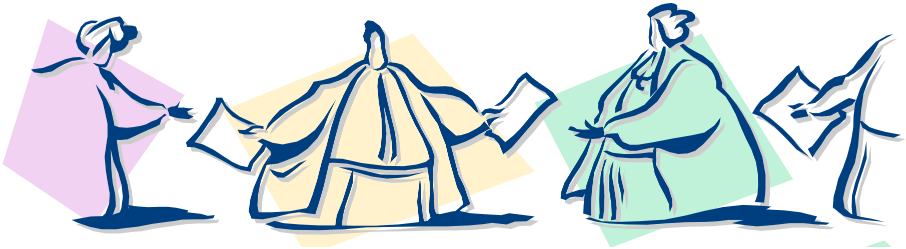


策略 #8 – 确保能够相互协作的整合架构

目标 →

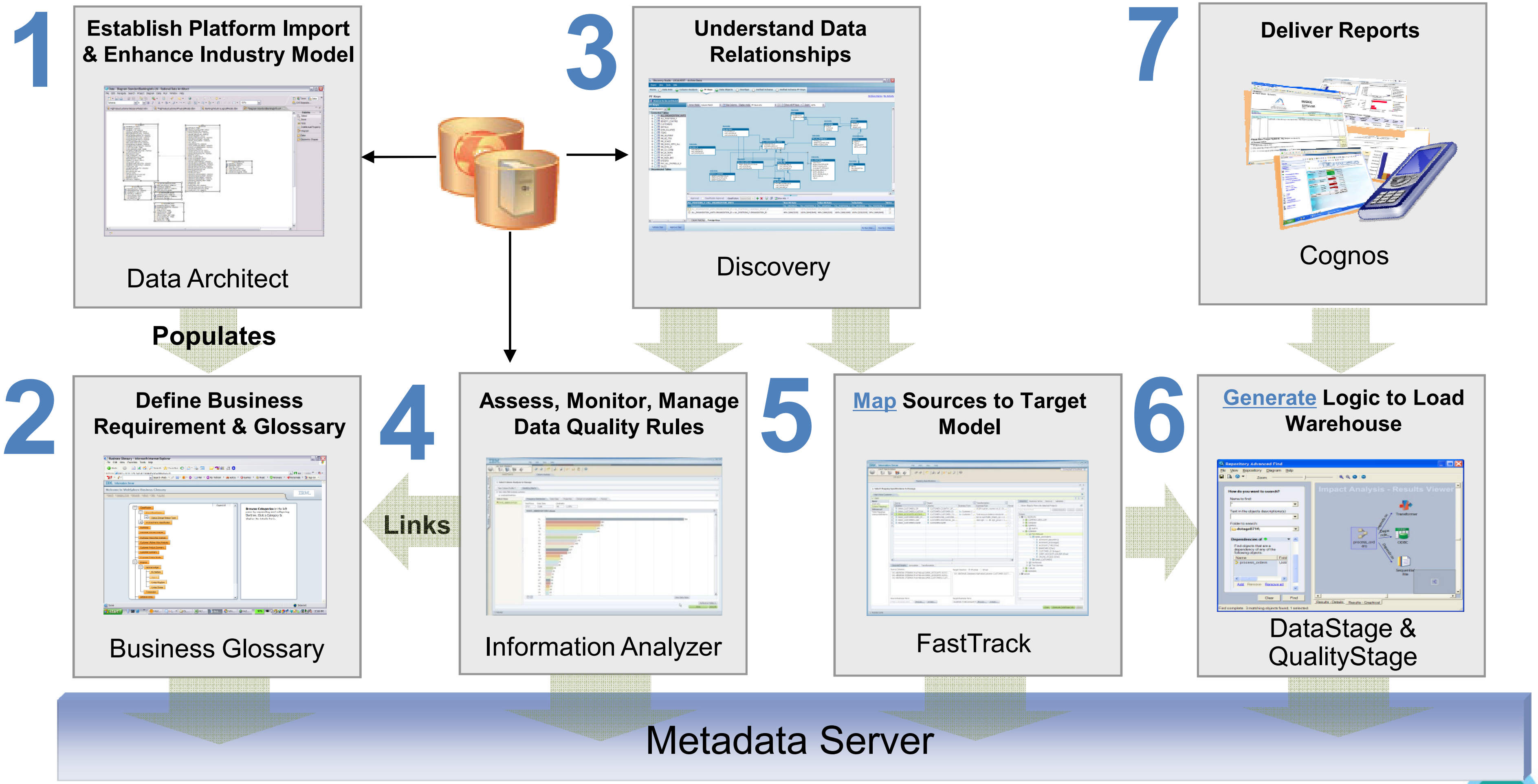
互通的，整合的，无缝的

现实 →



分散的，独立的，缺乏沟通

最佳实践： 整合的工具套件



1 Establish Platform Import & Enhance Industry Model

Data Architect

Populates

2 Define Business Requirement & Glossary

Business Glossary

3 Understand Data Relationships

Discovery

4 Assess, Monitor, Manage Data Quality Rules

Information Analyzer

Links

5 Map Sources to Target Model

FastTrack

6 Generate Logic to Load Warehouse

DataStage & QualityStage

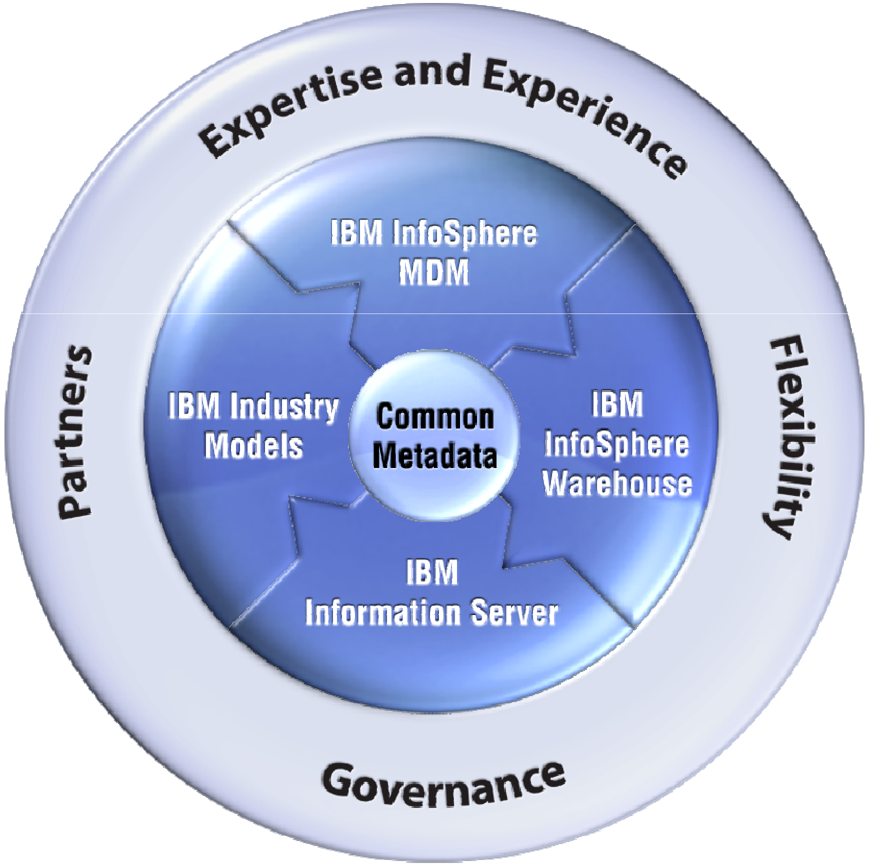
7 Deliver Reports

Cognos

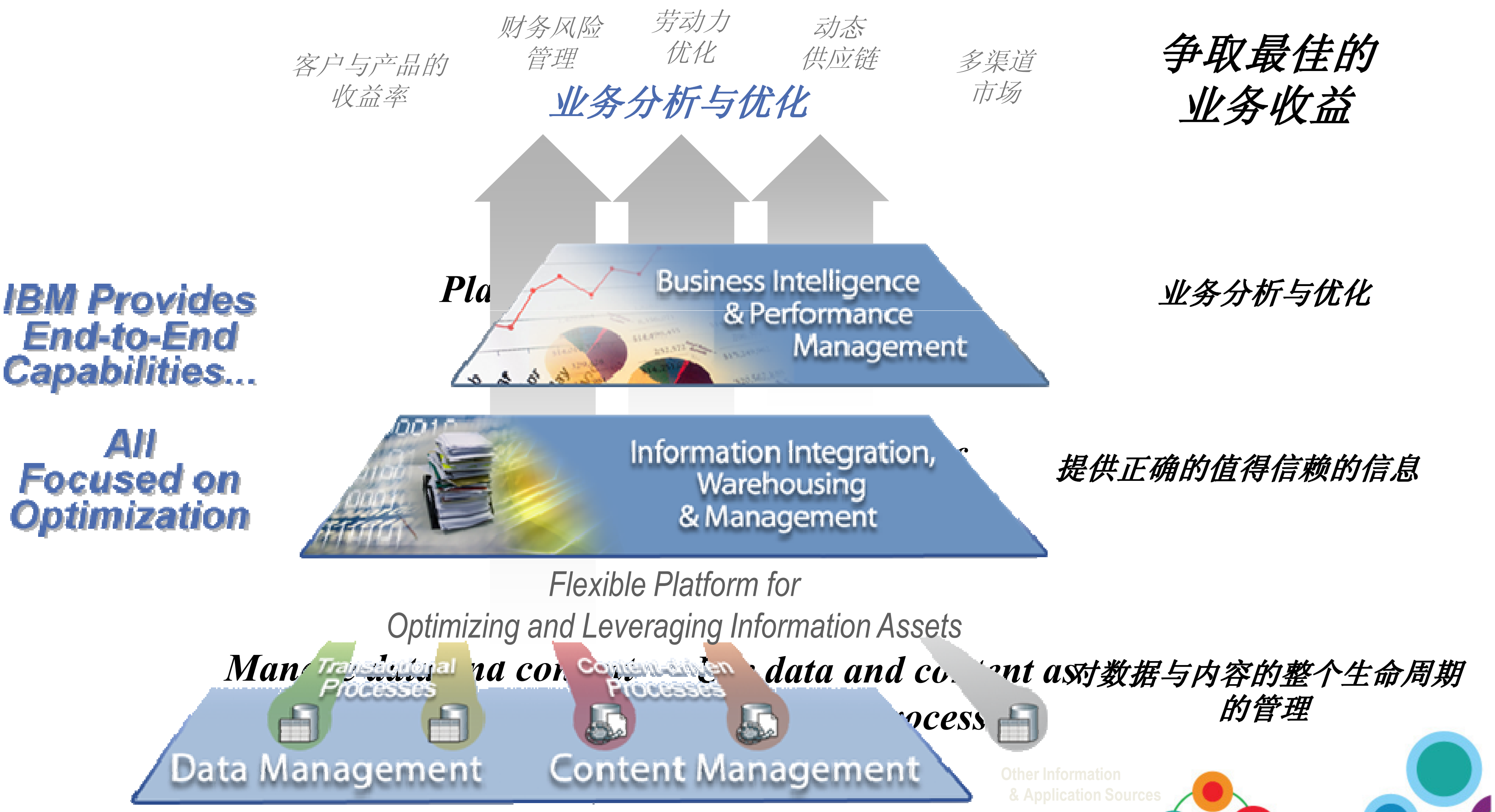
Metadata Server

简化和包罗万象：减少项目的周期，风险，成本！

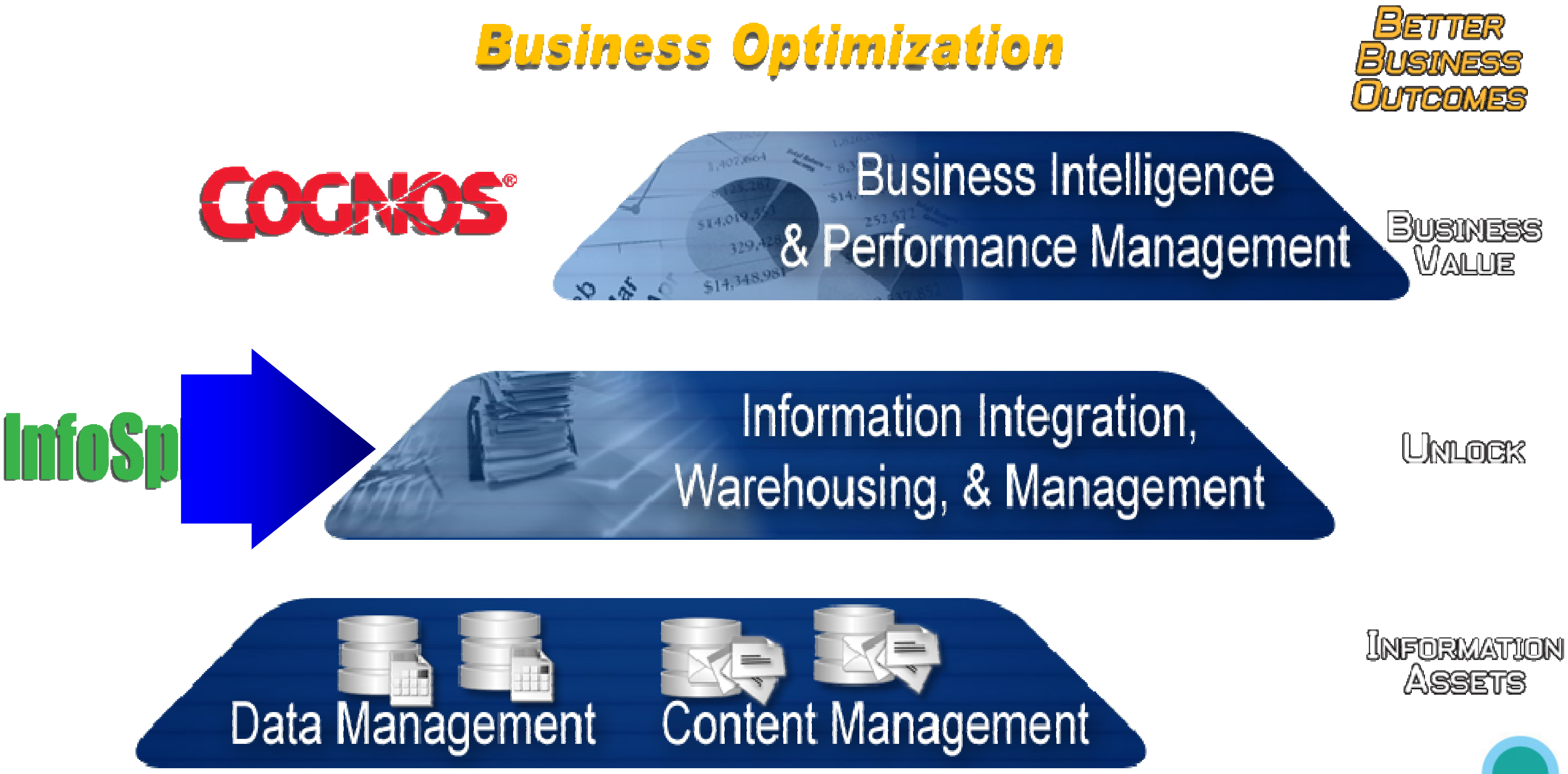
IBM的数据集成解决方案：InfoSphere



Information On Demand—释放信息的业务价值

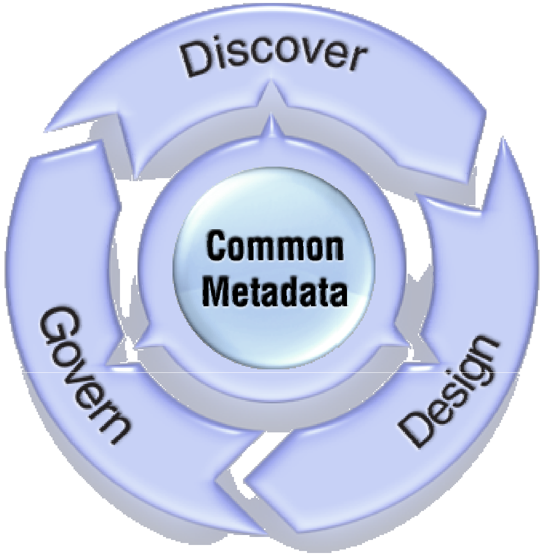


Information On Demand一释放信息的业务价值



把信息转换成为可以信息的企业核心资产

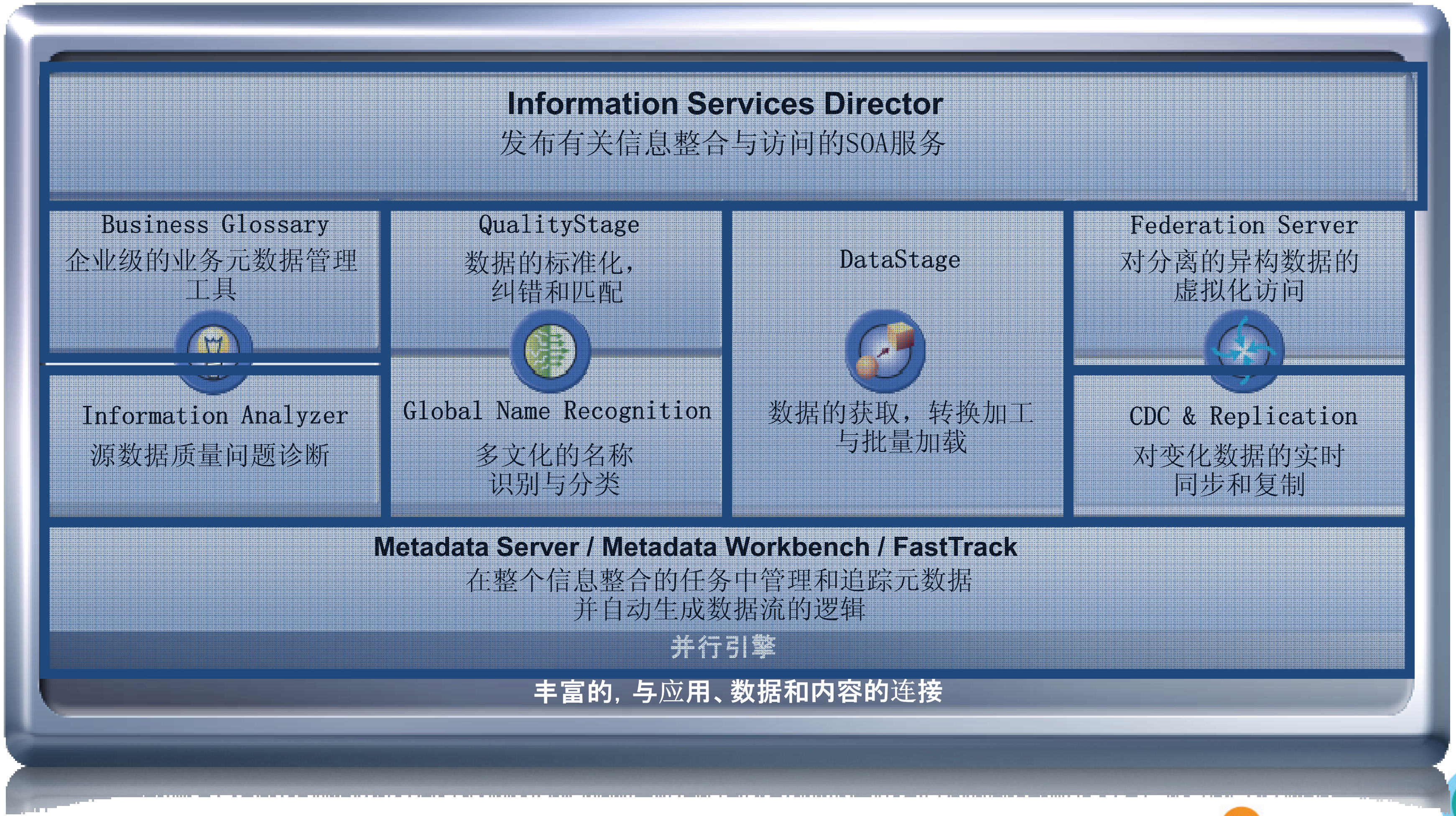
在信息的全面定义和管理方面，IBM投入了大量的精力，拥有整体的解决方案



- 业务术语
- 数据关系
- 数据质量的合规
- 数据模型和映射
- 业务规则
- 信息来源

- *发现* 与理解异构系统中的数据
- *设计* 为了业务优化而需要的可信数据
- *管控* 随着时间而变化的信息

InfoSphere Information Server：为您提供可以信赖的信息



Information Services Director
发布有关信息整合与访问的SOA服务

Business Glossary
企业级的业务元数据管理工具

Information Analyzer
源数据质量问题诊断

QualityStage
数据的标准化，纠错和匹配

Global Name Recognition
多文化的名称识别与分类

DataStage
数据的获取，转换加工与批量加载

Federation Server
对分离的异构数据的虚拟化访问

CDC & Replication
对变化数据的实时同步和复制

Metadata Server / Metadata Workbench / FastTrack
在整个信息整合的任务中管理和追踪元数据并自动生成数据流的逻辑

并行引擎

丰富的，与应用、数据和内容的连接

您的选择…

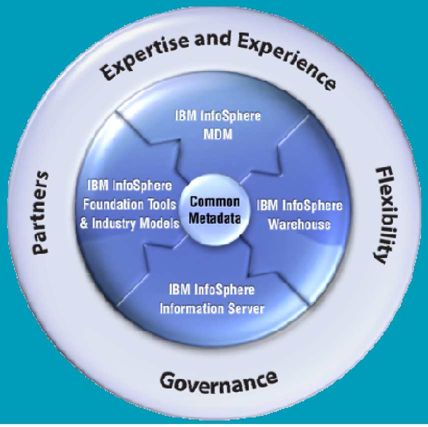
单点的产品

? + HARTE HANKS TRILLIUM SOFTWARE + INFORMATICA + siperian + ORACLE + Business Objects + ?

Models	Cleansing	ETL	MDM	Warehouse	BI	Mashups

IBM + IBM + IBM + IBM + IBM + IBM + IBM

整合的平台



Thank You！