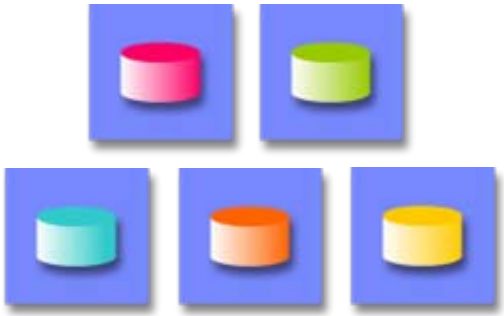
IOD 運用資訊威力

Information Management software

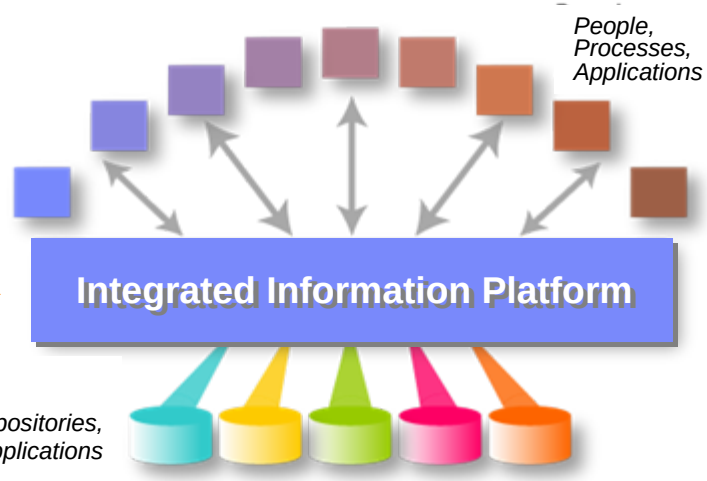
Information architecture is evolving

Disconnected Silos of Information



Rich Standards, Flexible Architecture

Dynamically Deliver Master Information



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

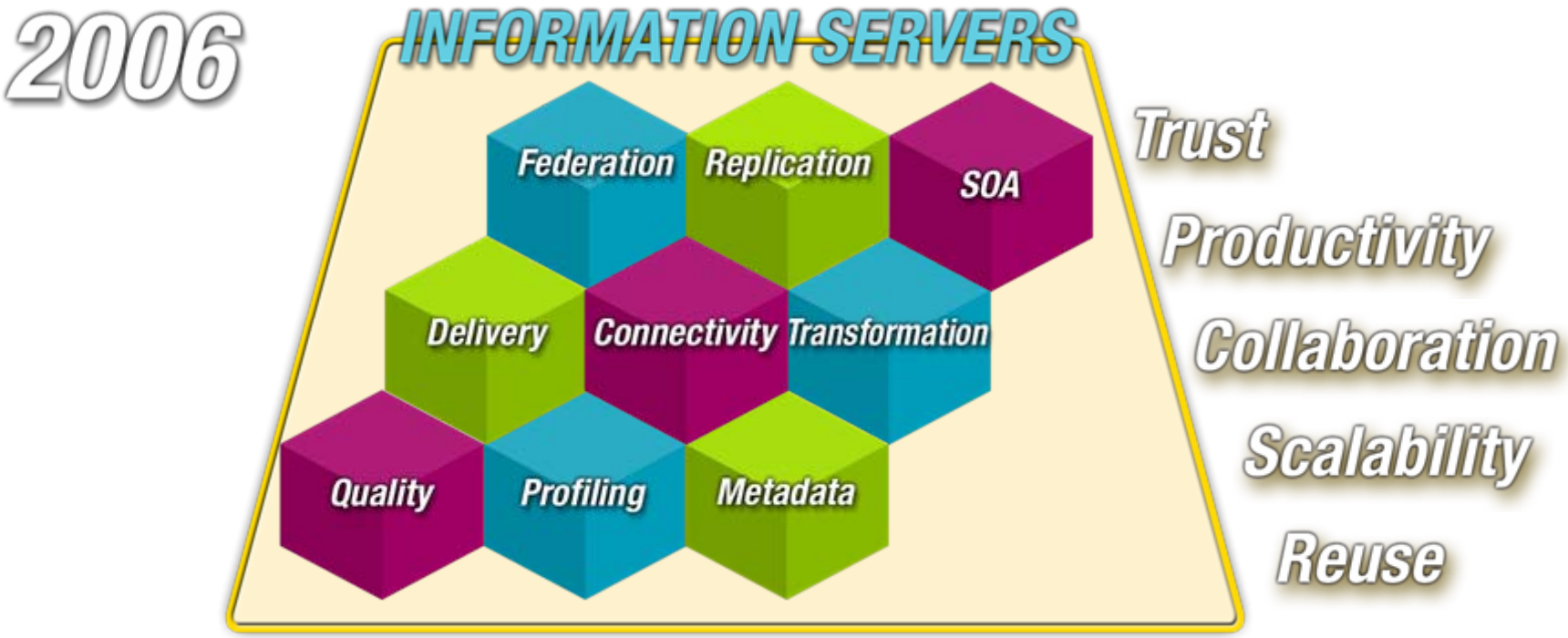
60%+ **of CEOs say they need to do a better job leveraging information**

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

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

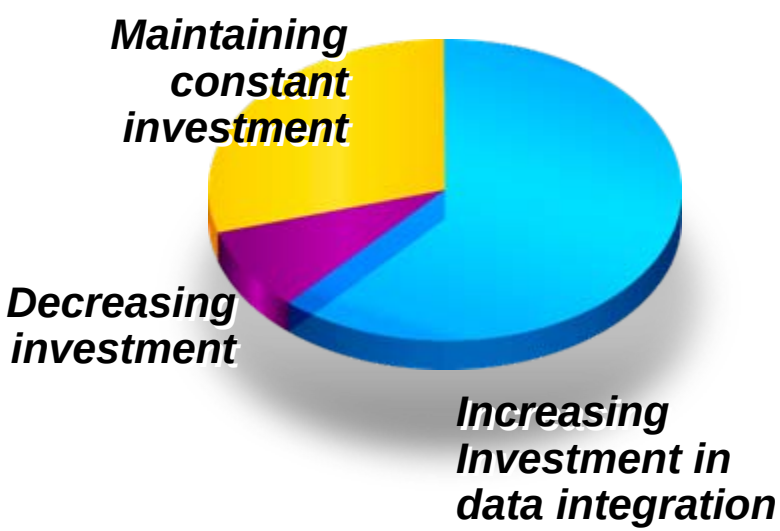
An historic inflection point

Today's inflection point

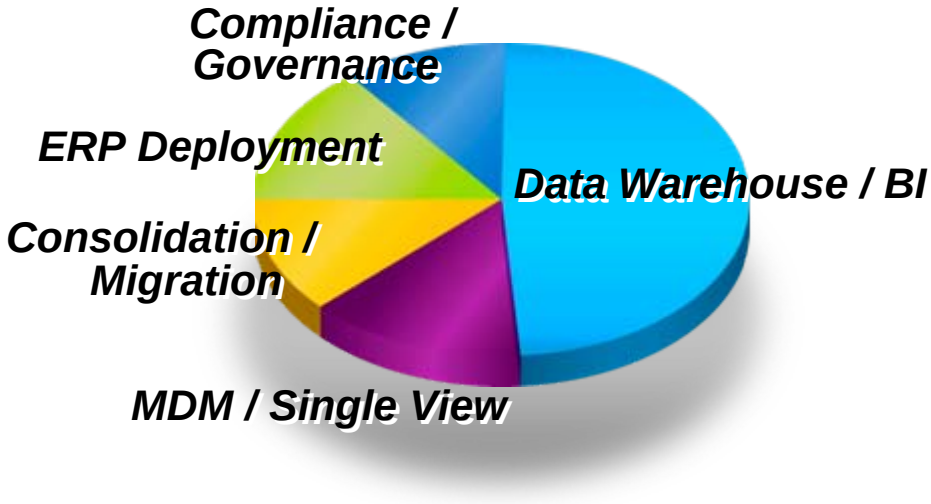


Businesses Are Responding to Market Demands

Investments in data integration are increasing



Driven by strategic initiatives



BI applications are the #1 technology priority
Business process improvement is the #1 concern

Source: Gartner 2006
"Gartner Study on Data Integration Identifies Key Usage Trends"

Source: IBM 2006
IBM Survey of 1,600 CIOs

Customer Business Issues

- **Too much information and not knowing what's important**
  - Not using demand signals to drive supply chain
  - Not using customer analysis to tailor marketing and sales
  - Not leveraging valuable unstructured information

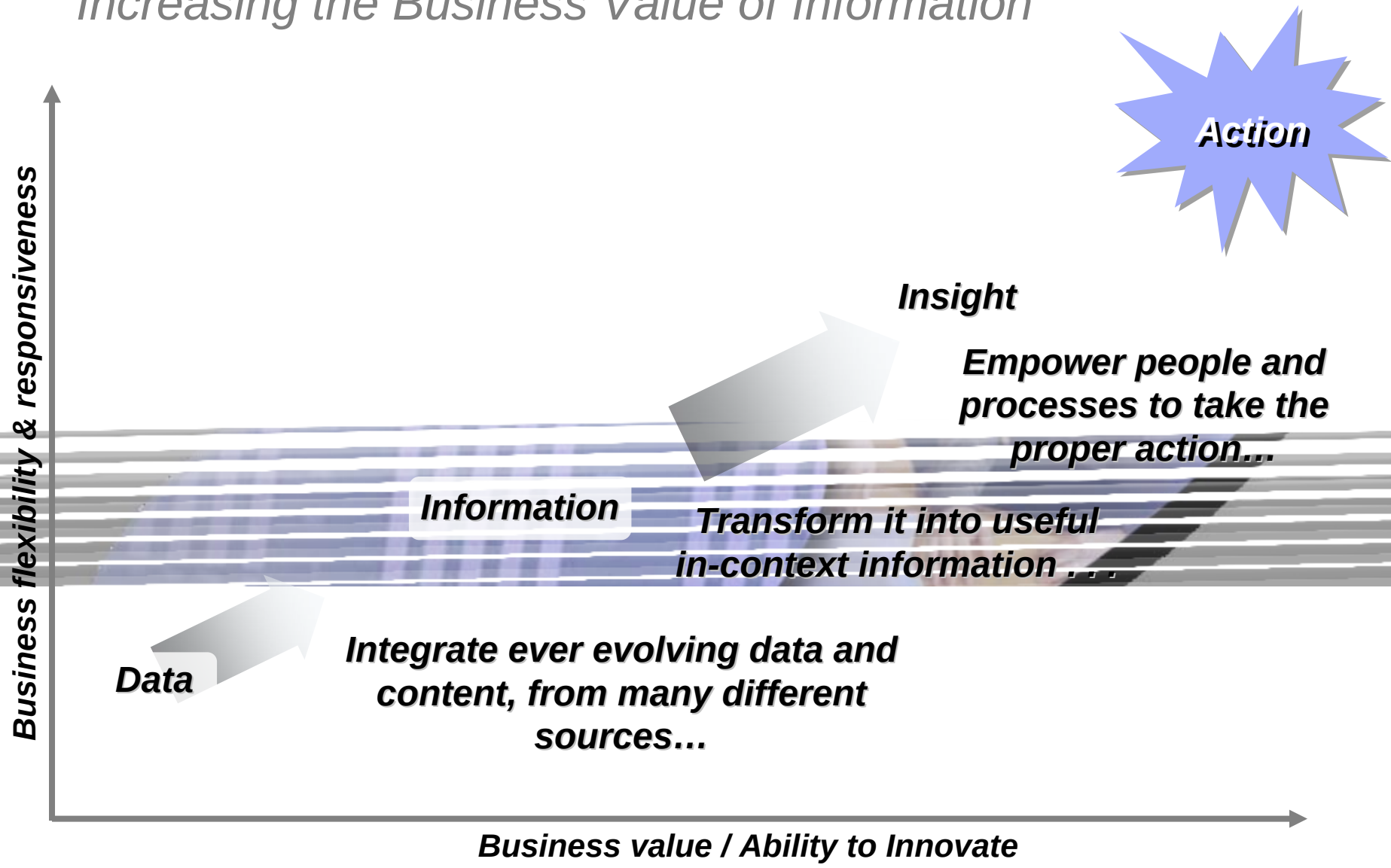
- **Multiple versions of the truth**
  - Problems managing customer, product and partner interactions
  - Regulatory compliance inhibited by poor transparency

- **Lack of trusted information**
  - Incomplete, out-of-date, inaccurate, misinterpreted data
  - Difficult to understand or control how information is used

- **Lack of agility**
  - Inability to take advantage of opportunities for innovation
  - Escalating costs due to inflexible systems and changing needs

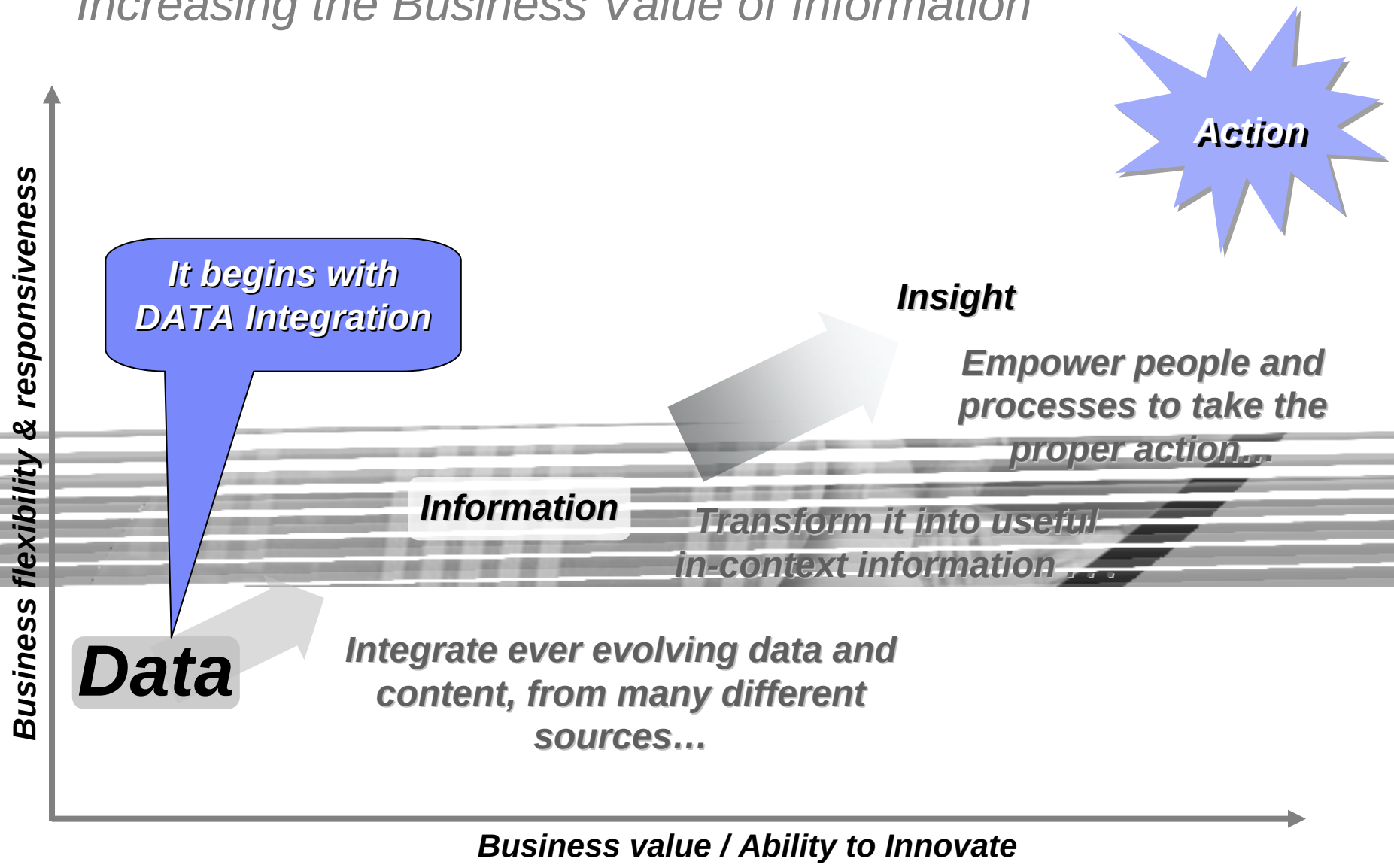
Information on Demand

Increasing the Business Value of Information



Information on Demand

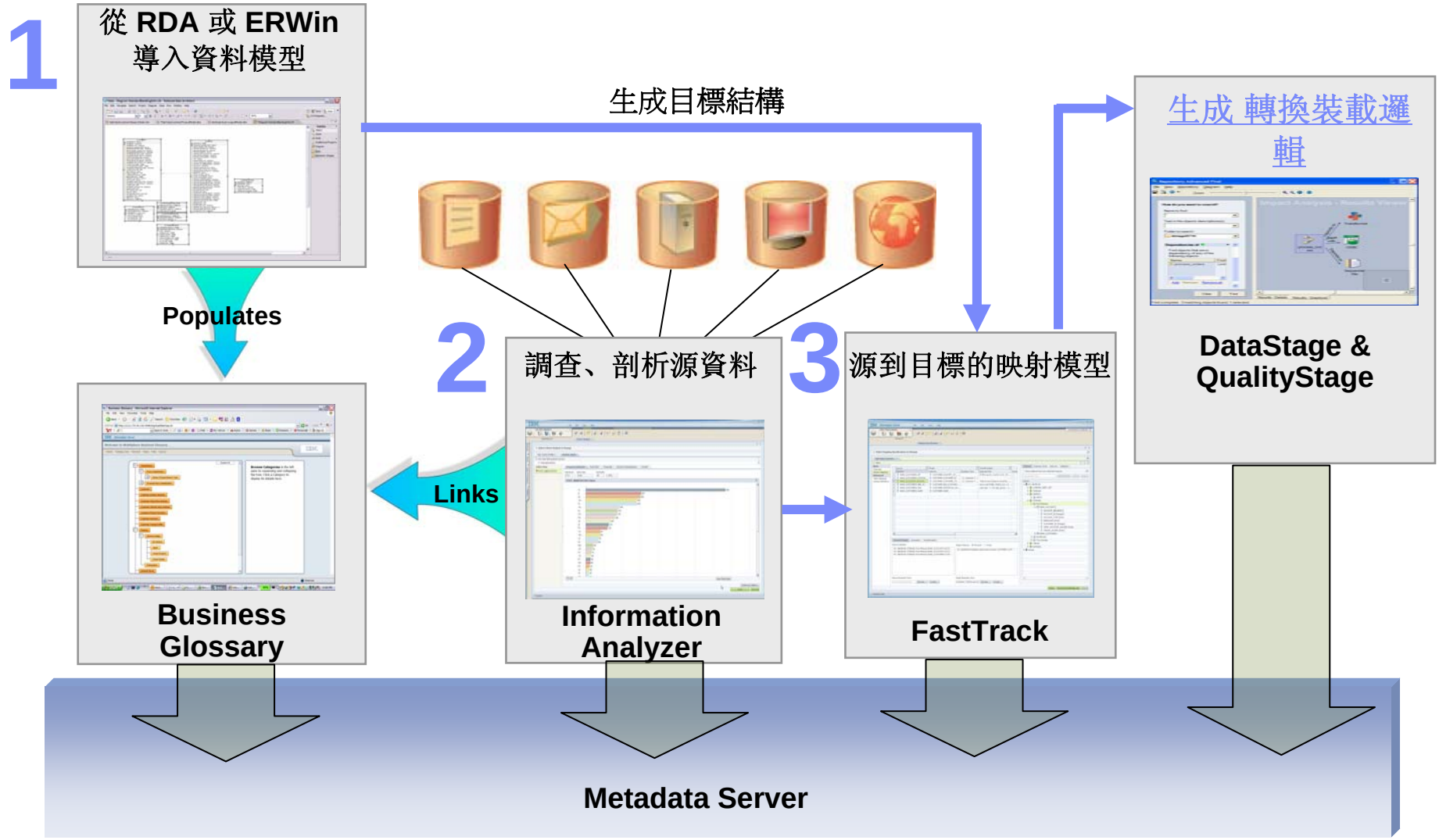
Increasing the Business Value of Information



Integration – Some Definitions

- Process-Centric
  - EAI (Enterprise Application Integration)
    - Process-centric, automate workflows and process dependencies
    - Leverage process assets across applications, to create new applications
- Data-Centric
  - Federation
    - Data and content centric "pull"
    - Virtualizes access to information sources: structured, unstructured, content
  - ETL (Extract, Transform, Load)
    - Data placement and transformation solutions
    - Typically scheduled, but moving to more real-time models
  - Data Event Publishing
    - Data centric "push"
    - Driver behind other data movement and integration models
  - Replication
    - Creating and synchronizing like-to-like copies of data
    - Focus on performance, multi-directional capabilities

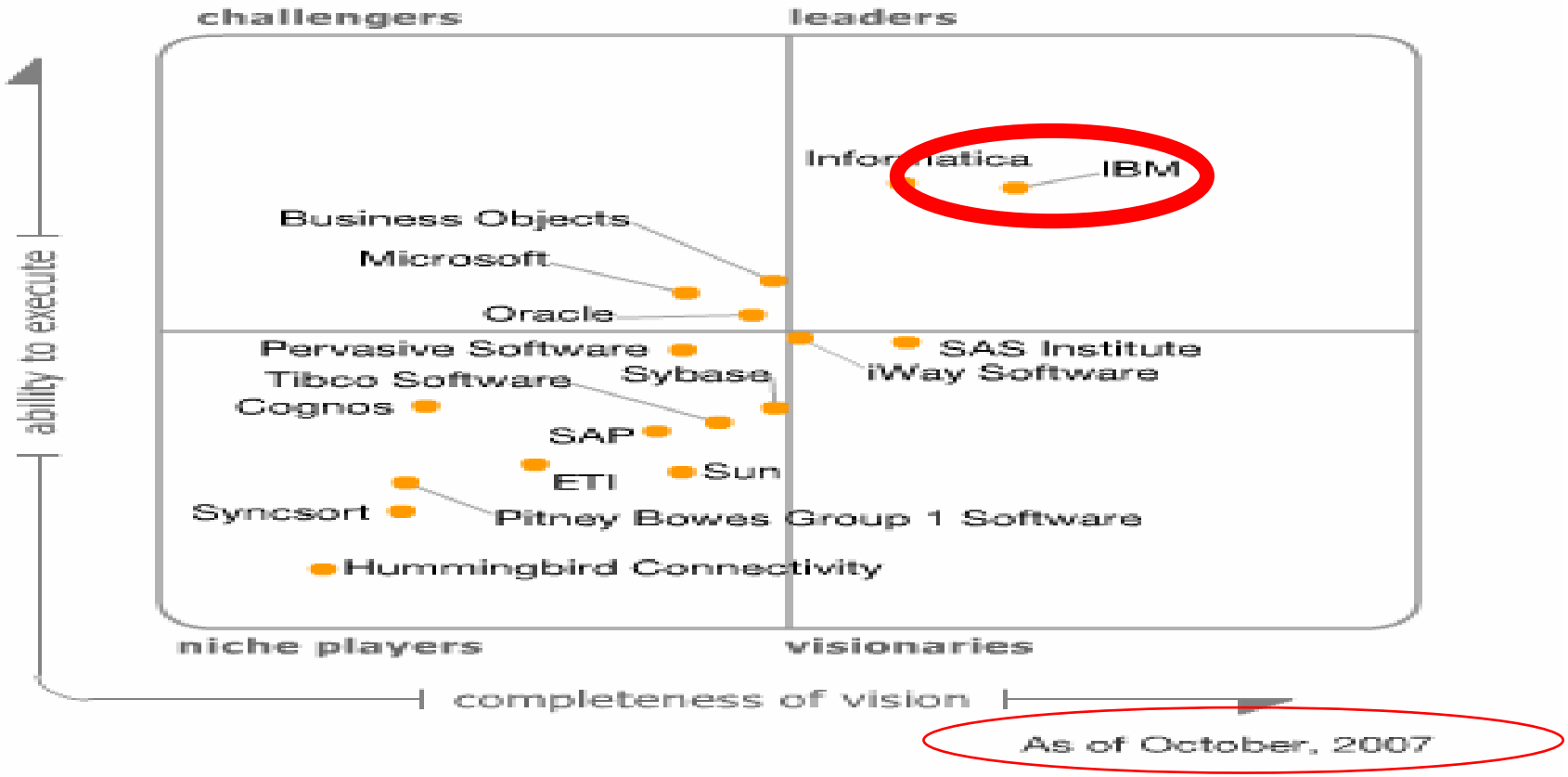
資料整合流程



簡化 & 滿意：降低項目時間、風險和成本!

從 Gartner Magic Quadrant 報告尋找最佳解決方案

Figure 1. Magic Quadrant for Data Integration Tools, 2007



Source: Gartner

IBM 於十六項評比中勇奪九項評比第一名 !!

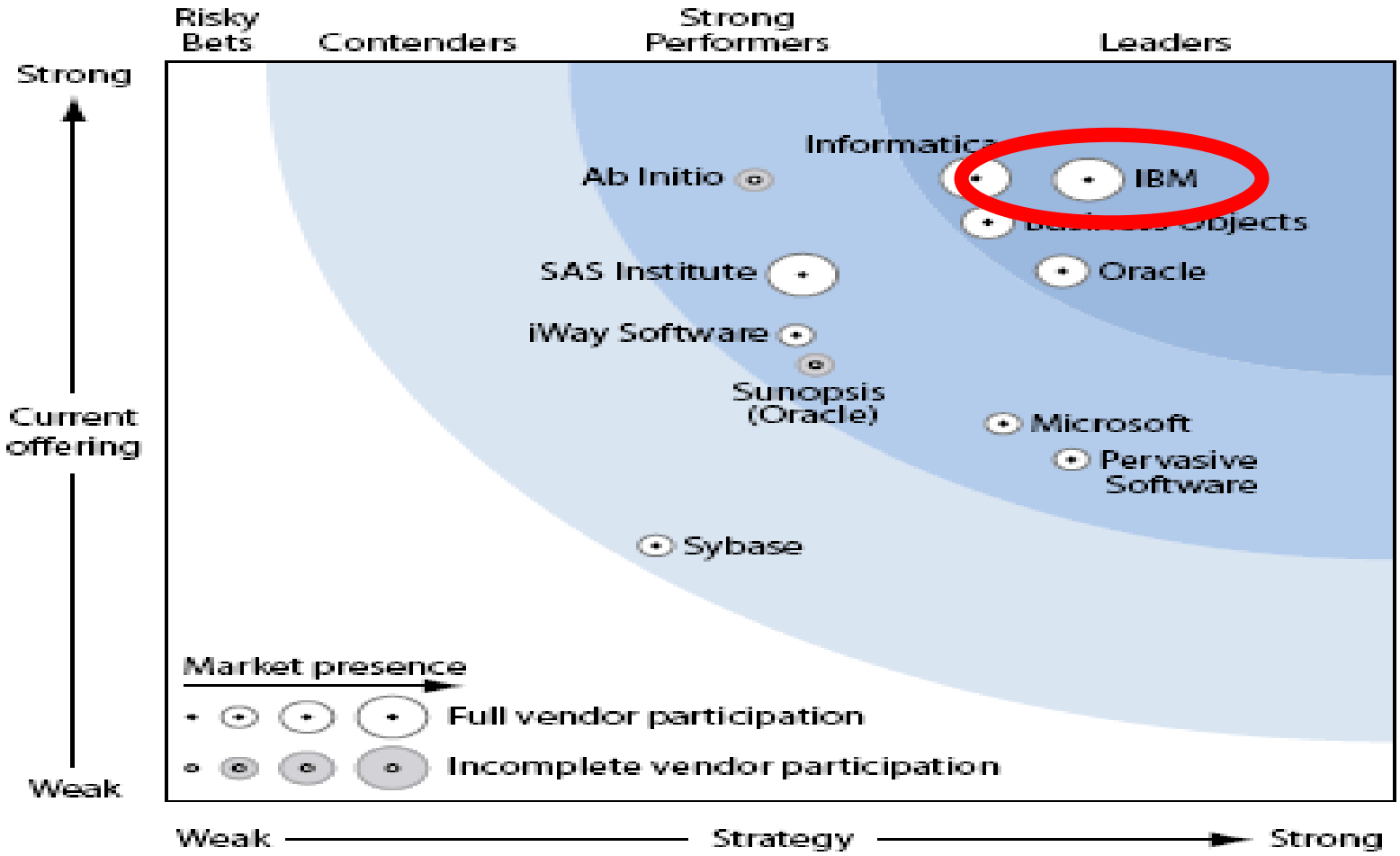
Figure 3 Forrester Wave™: Enterprise ETL, Q2 '07 (Cont.)

	Forrester's Weighting	Business Objects	IBM	Informatica	iWay Software	Microsoft	Oracle	Pervasive Software	SAS Institute	Sybase
CURRENT OFFERING	50%	3.91	4.20	4.21 1	3.15	2.56	3.58	2.32	3.57	1.73
Server capability	20%	3.80	4.15	4.45 1	3.70	2.85	3.55	2.00	3.85	2.15
Integration options	20%	4.45	4.75 1	4.45	4.15	2.30	3.40	3.20	4.45	2.20
Tool environments	20%	4.15	4.15 1	4.15	2.60	2.90	4.00	1.90	3.60	1.70
Support and training	5%	3.95	4.50	4.05	3.55	4.25	4.90 1	2.95	4.65	2.45
Additional data integration techniques	10%	3.40	3.60	4.00 1	4.20	2.00	2.60	1.60	0.80	1.20
Information management	25%	3.55	4.00 1	4.00	1.85	2.15	3.55	2.35	3.50	1.10
STRATEGY	50%	3.34	3.75 1	3.29	2.56	3.40	3.64	3.68	2.59	1.99
Product strategy	35%	4.00	4.50 1	4.00	2.50	3.50	3.50	3.50	3.00	2.00
Corporate strategy	15%	3.00	3.65	3.70 1	3.00	3.00	3.00	3.00	3.00	3.00
Cost	25%	3.10	2.20	2.20	3.10	4.00	4.00 1	4.20	2.50	3.00
Partnerships extending reach	25%	2.85	4.30 1	3.15	1.85	2.90	3.85	3.80	1.85	0.35
MARKET PRESENCE		3.44	4.55 1	4.35	2.18	2.67	3.82	2.95	4.25	2.53
Company financials	30%	4.00	4.50	4.75 1	2.25	2.75	2.75	2.75	5.00	2.25
Installed base	50%	3.20	4.40 1	4.00	1.80	2.00	4.30	3.20	3.50	2.50
Employees	20%	3.20	5.00 1	4.60	3.00	4.20	4.20	2.60	5.00	3.00

All scores are based on a scale of 0 (weak) to 5 (strong).

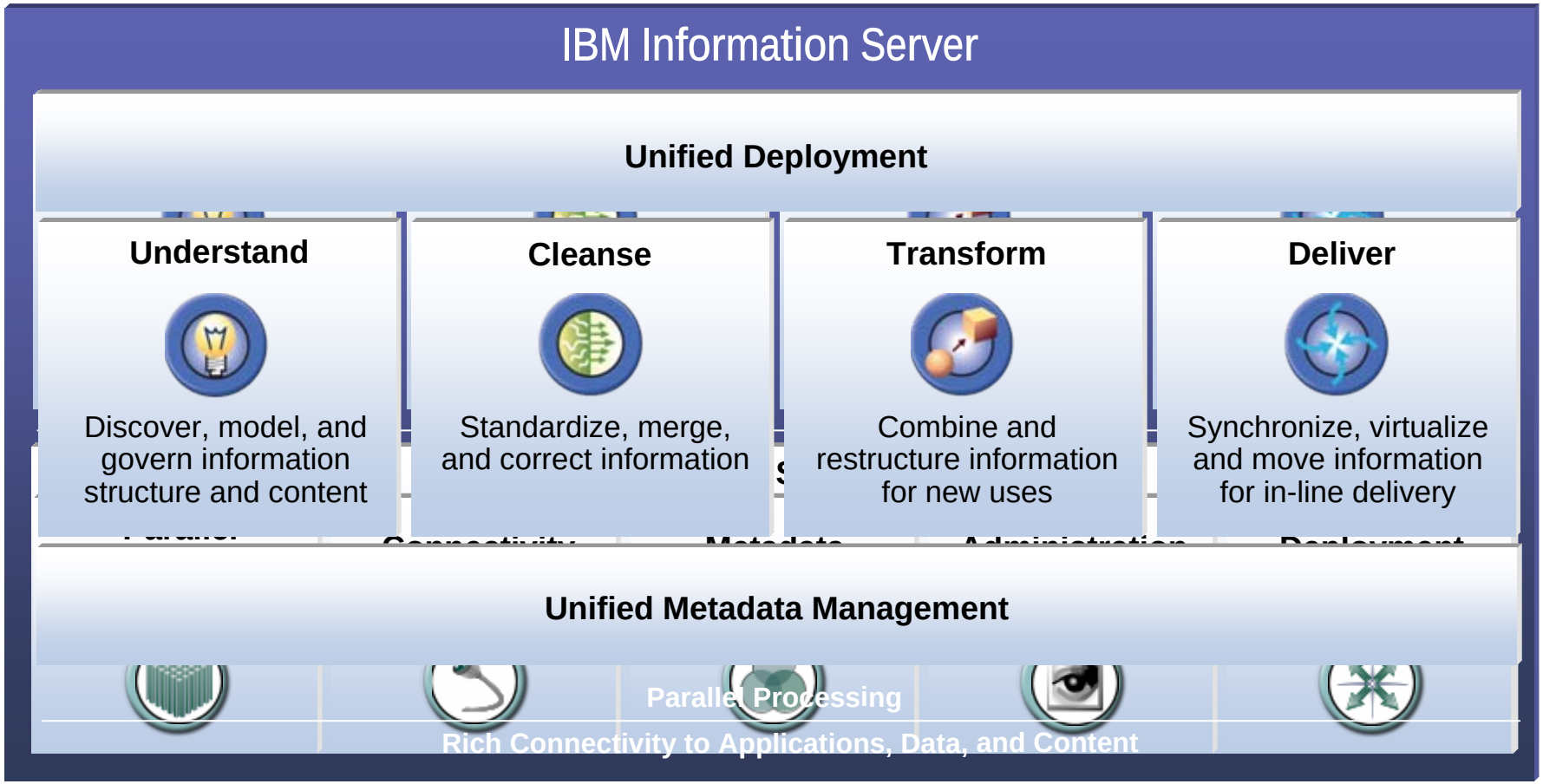
Source: Forrester Research, Inc.

Forrester 告訴你：選 IBM 就對啦!



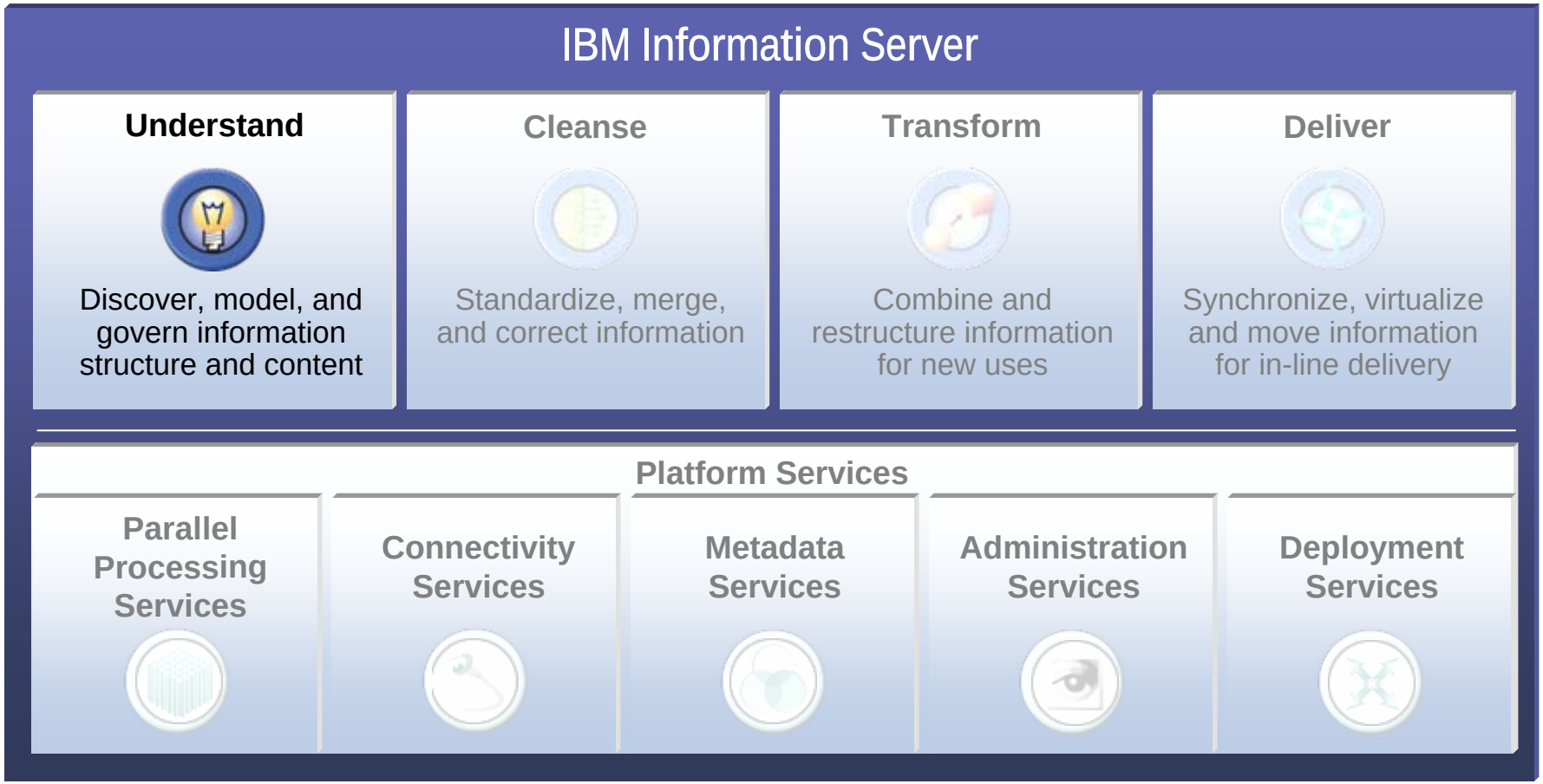
The IBM Solution: IBM Information Server

Delivering information you can trust



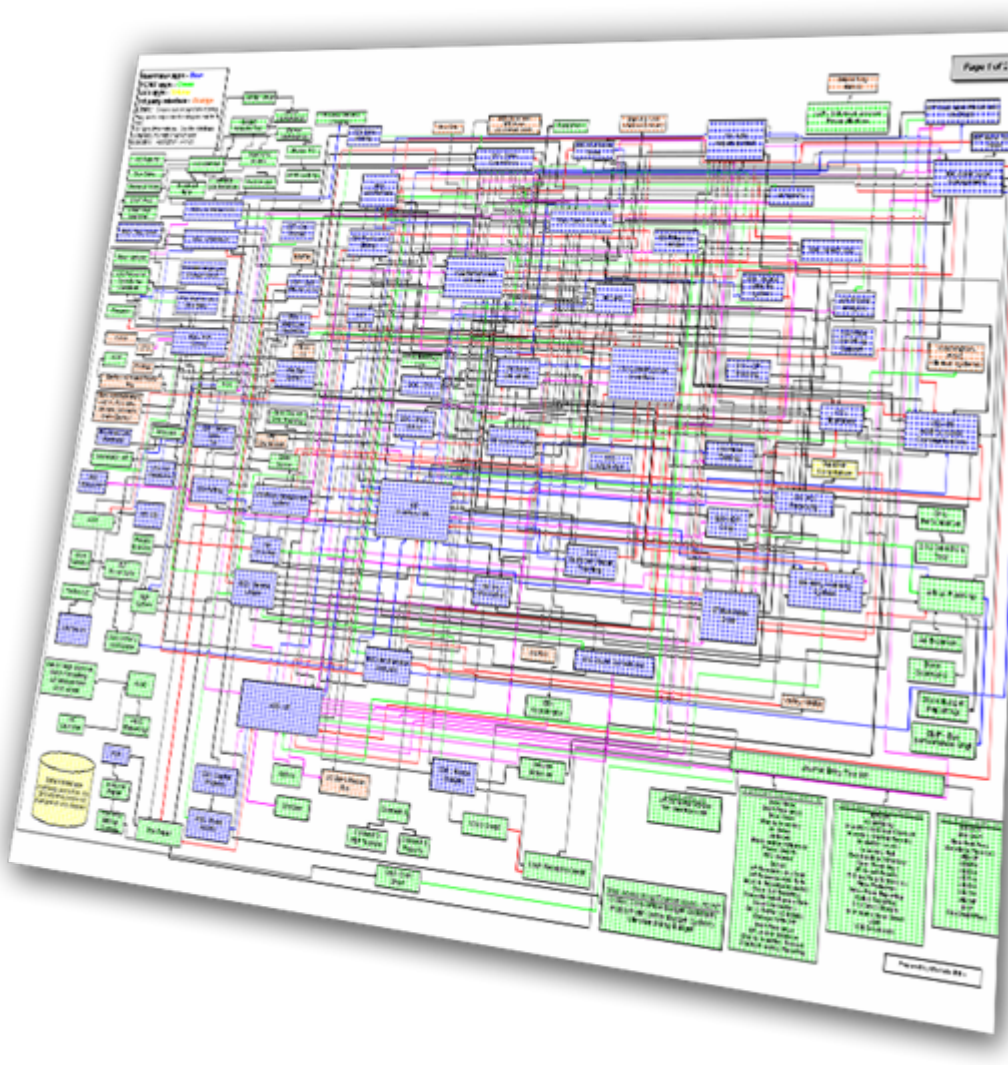
The IBM Solution: IBM Information Server

Delivering information you can trust

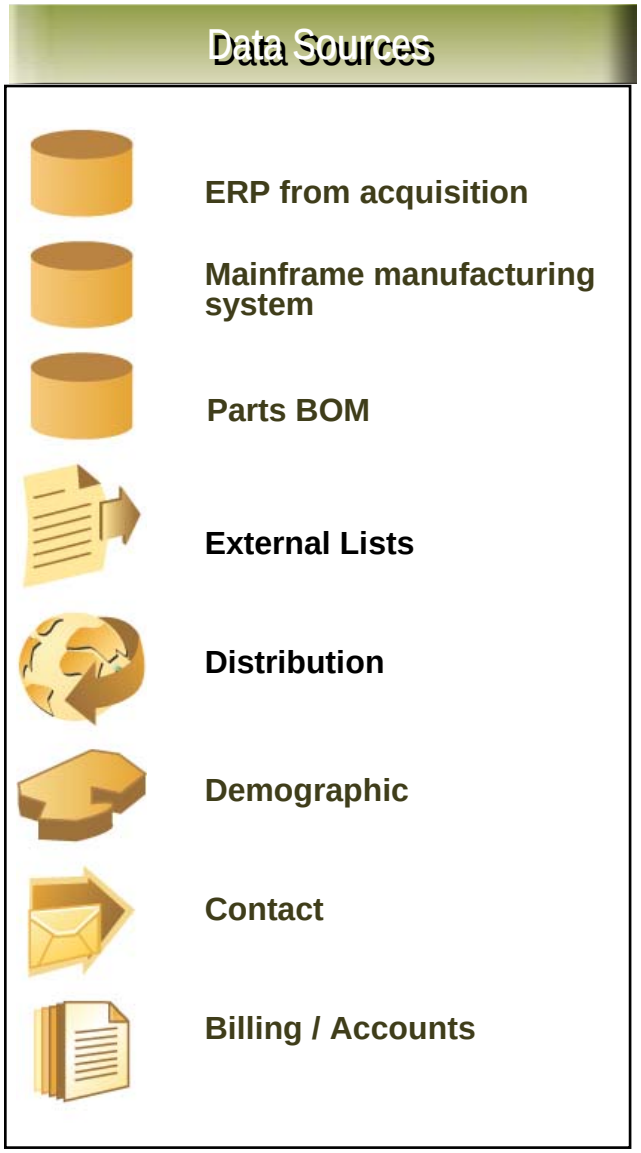


The Information Challenge

- What data sources are out there?
- How are they related to each other?
- What exactly is in the source data?
- How is it organized?
- What's the quality of the data?
- Is any data missing?
- Is any data duplicated?
- Is it fit for it's intended purpose?
- How do we monitor sources for changes in quality over time?



Data Profiling



Critical Problems:

- You don't know what data is really in your legacy systems
- Sources have changed or are new and unknown

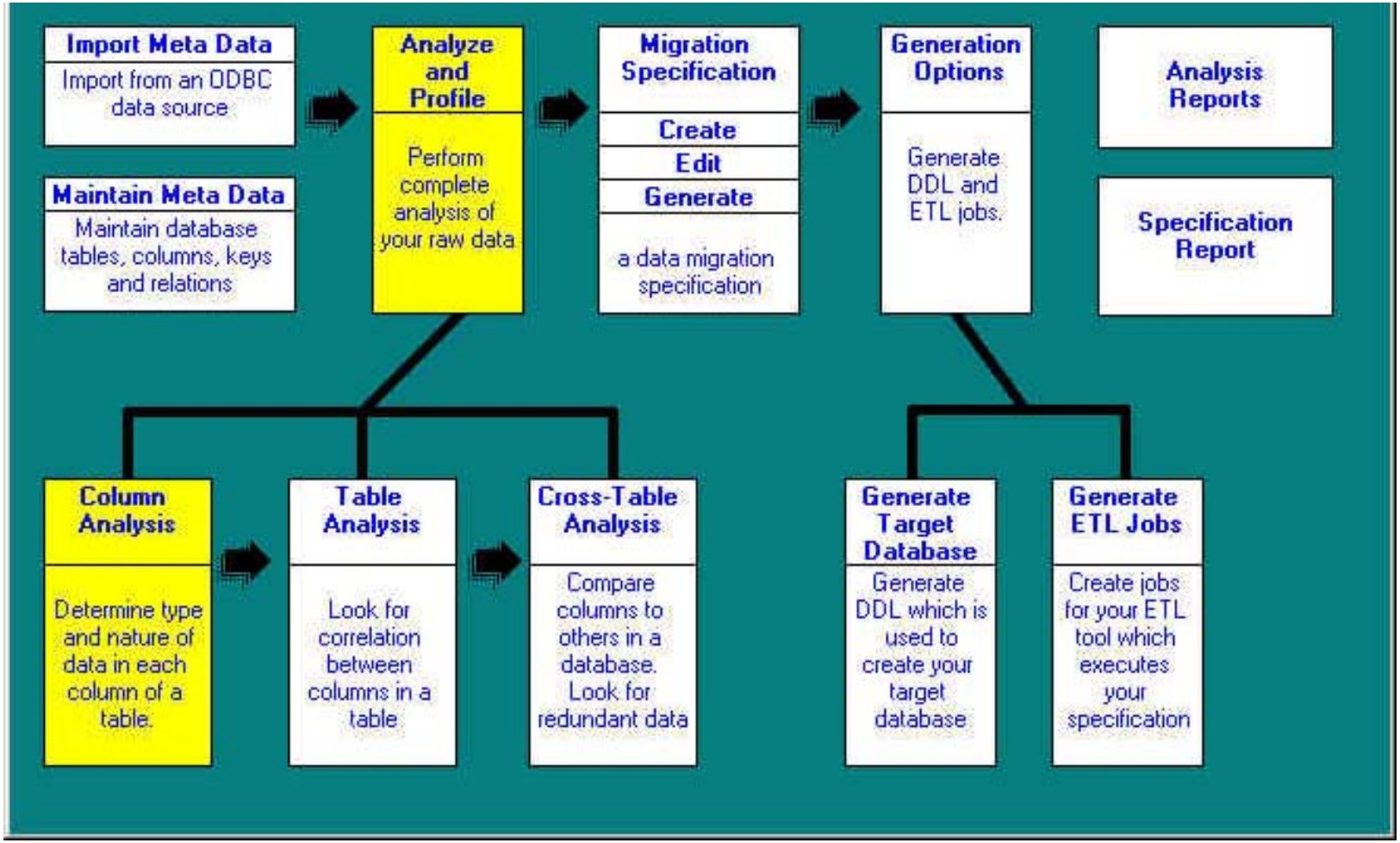
Why?

- Data values and relationships are inconsistent and divergent from documented rules
- Incomplete and missing documentation
- Data sources are never static and frequently change without warning

Alternative Approach

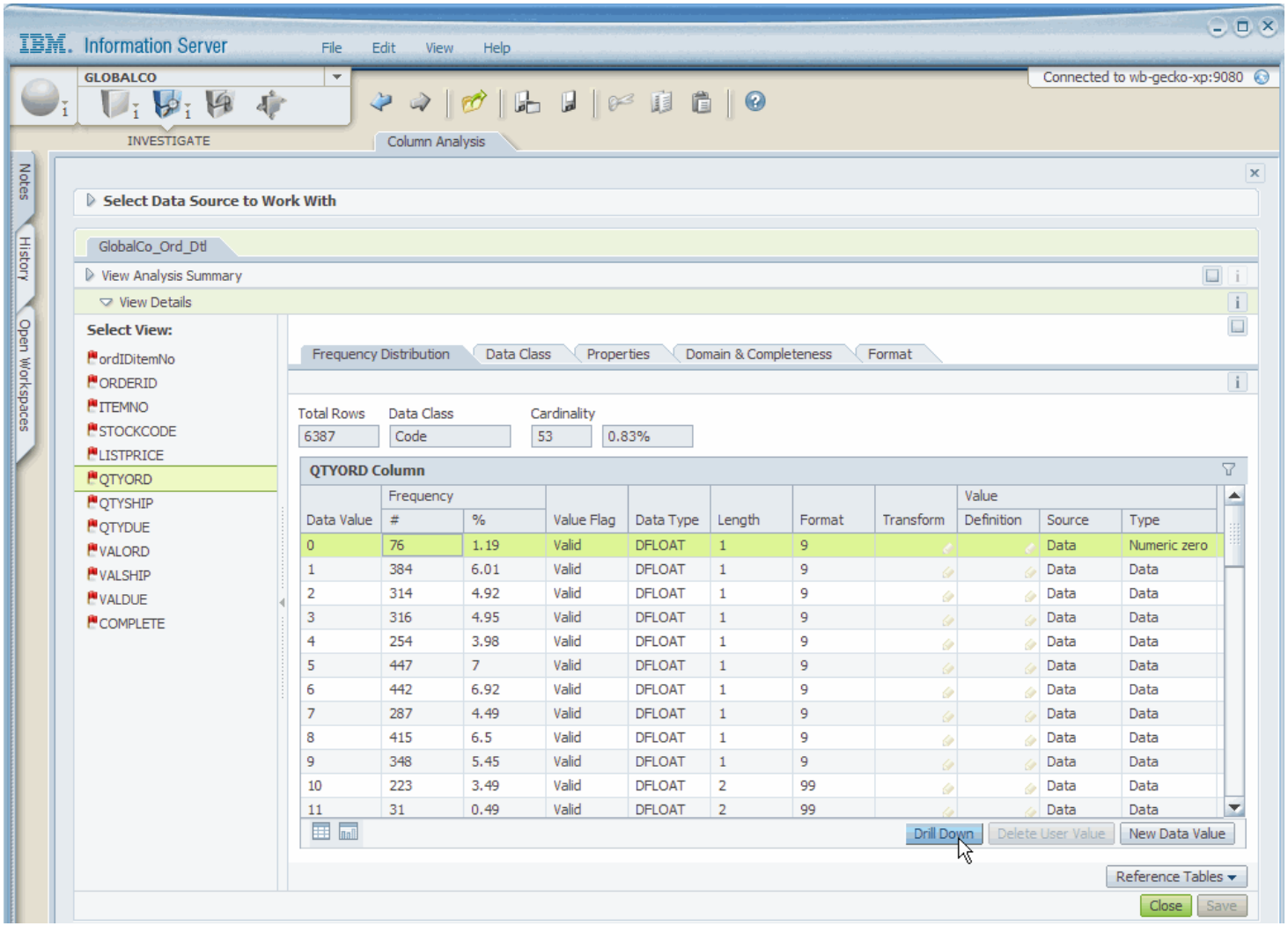
- Labor intensive, resource devouring process
- Never review 100% of data elements
- No infrastructure to support maintenance
- No standardized approach across projects
- 1st generation tools document but don't address the problem resolution

XX人壽 Data Profiling 步驟



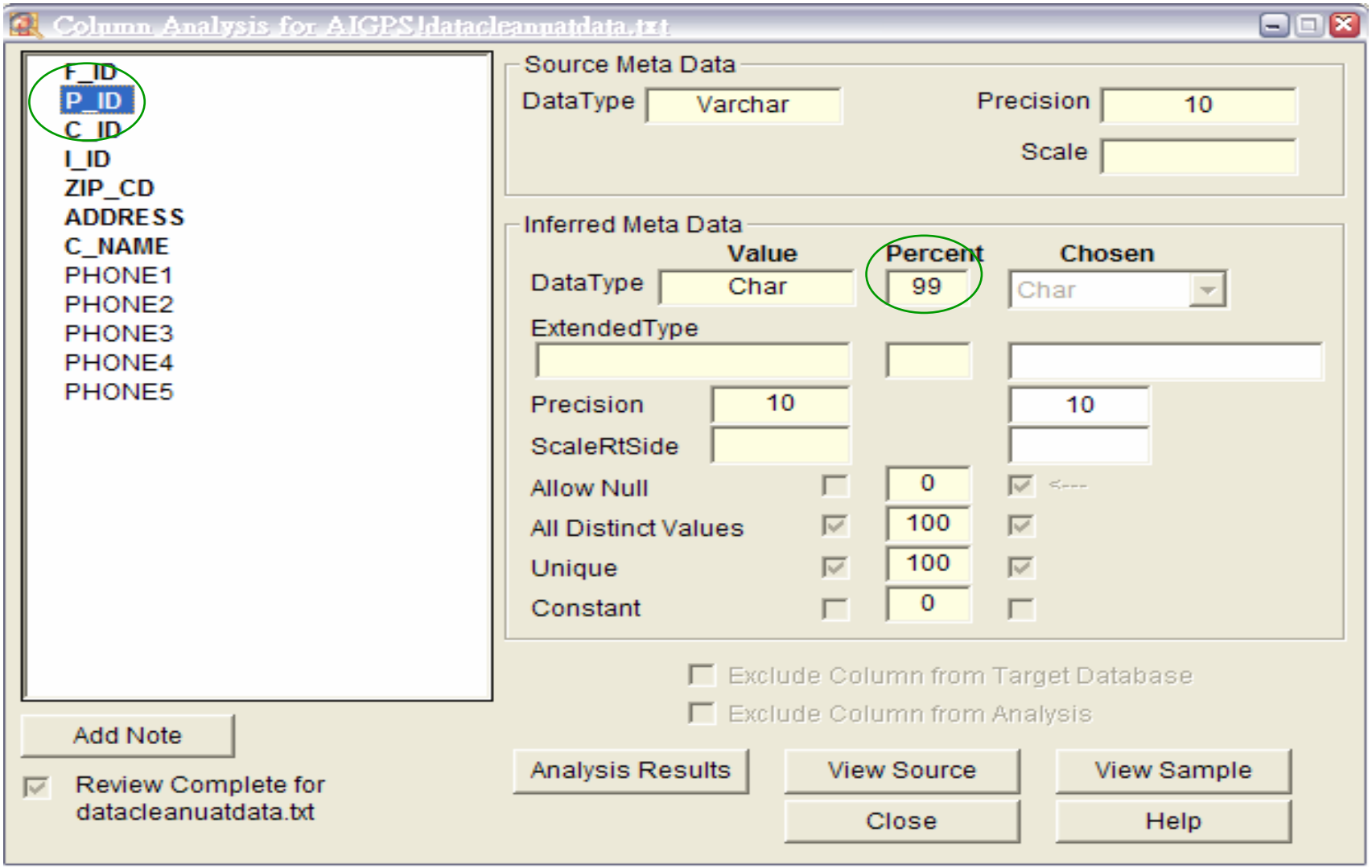
Data Profiling: Column Analysis

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

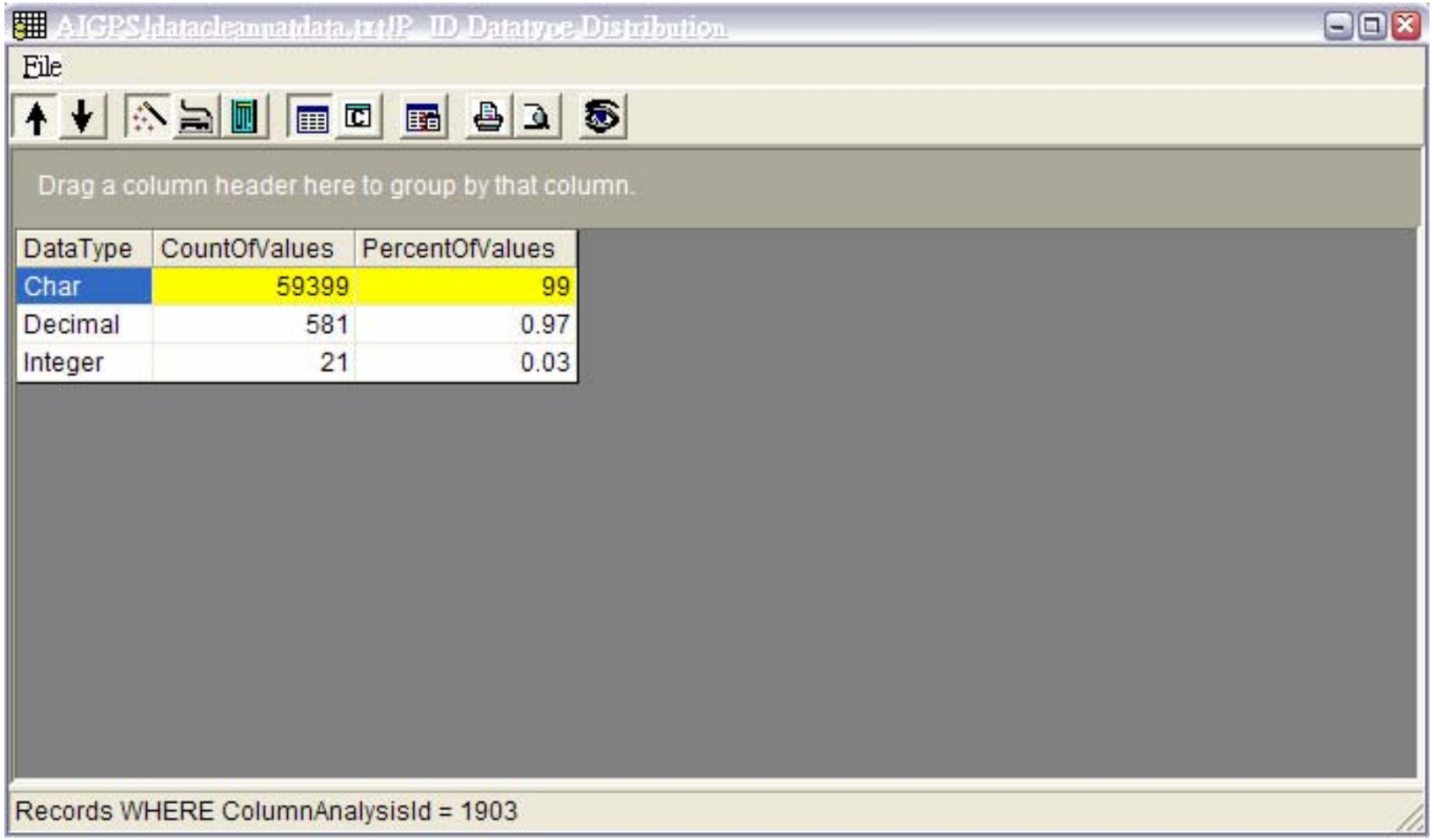


Data Value	Frequency #	Frequency %	Value Flag	Data Type	Length	Format	Transform	Value Definition	Value Source	Value Type
0	76	1.19	Valid	DFLOAT	1	9			Data	Numeric zero
1	384	6.01	Valid	DFLOAT	1	9			Data	Data
2	314	4.92	Valid	DFLOAT	1	9			Data	Data
3	316	4.95	Valid	DFLOAT	1	9			Data	Data
4	254	3.98	Valid	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

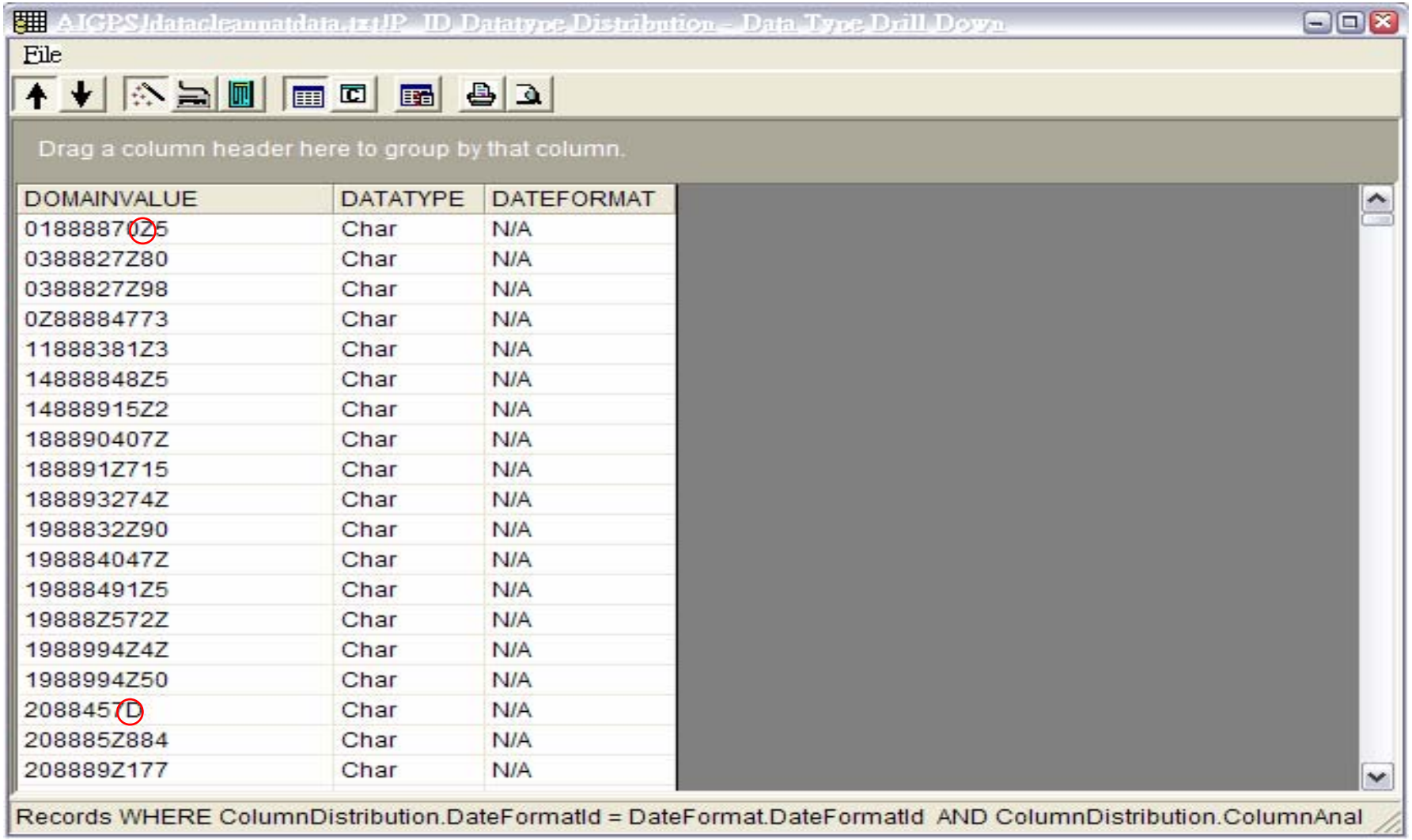
P_ID Column 屬性分析



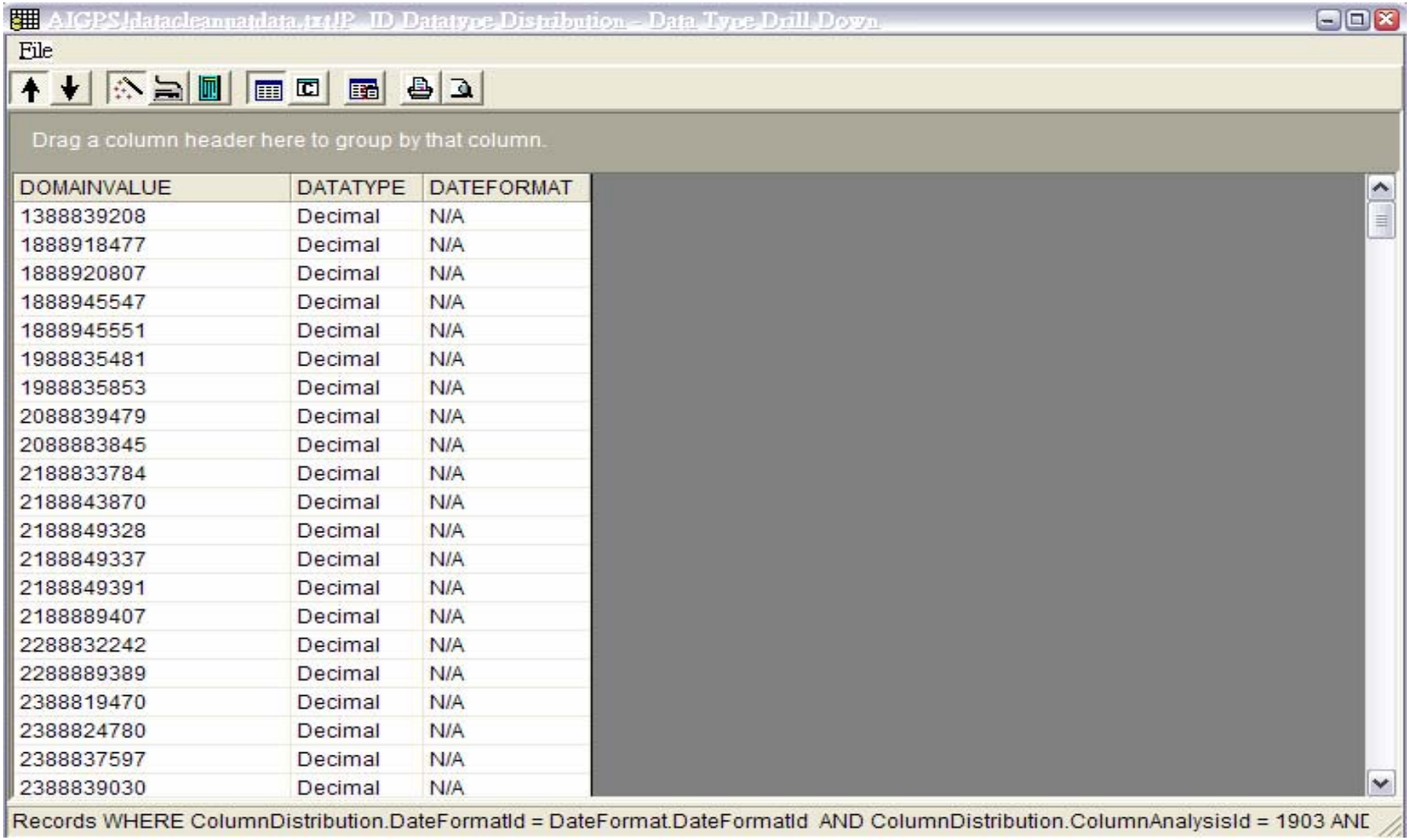
Column Analysis – P_ID Column 屬性，筆數，與百分比



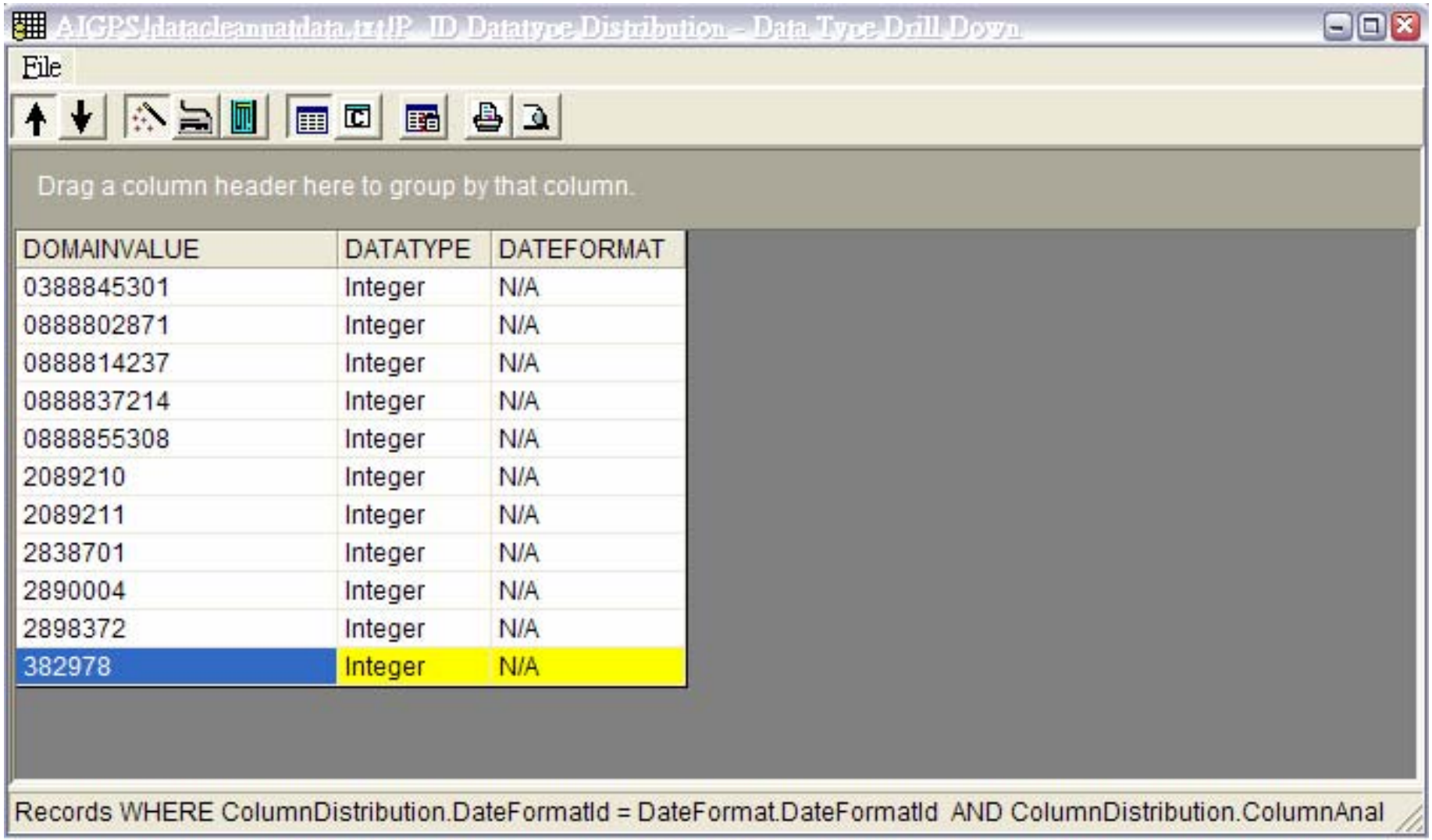
Column Analysis – P_ID Column資料值 – 屬於 Char 資料



Column Analysis – P_ID Column資料值 – 屬於 Decimal 資料



Column Analysis – P_ID Column資料值 – 屬於 Integer 資料



所有Column Analysis Report(Partial)

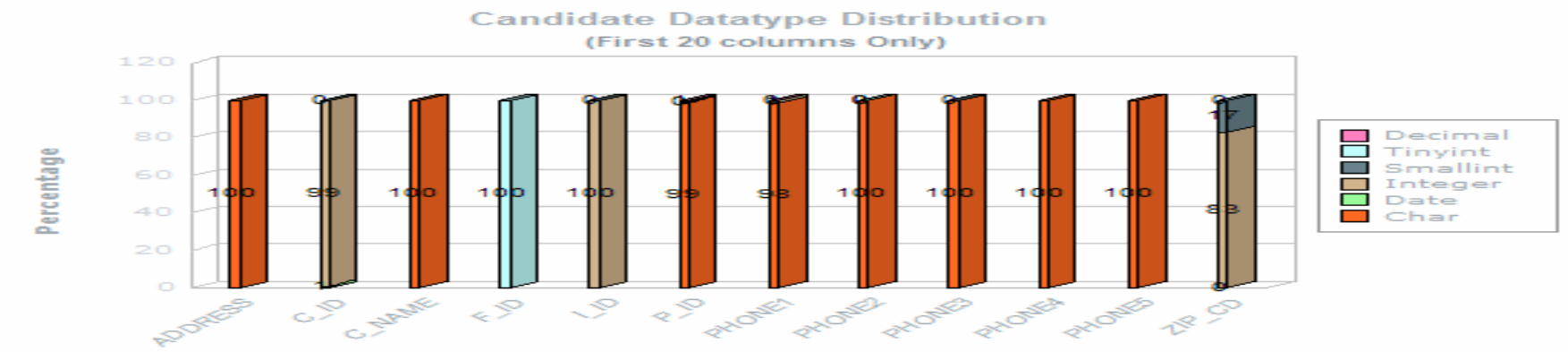
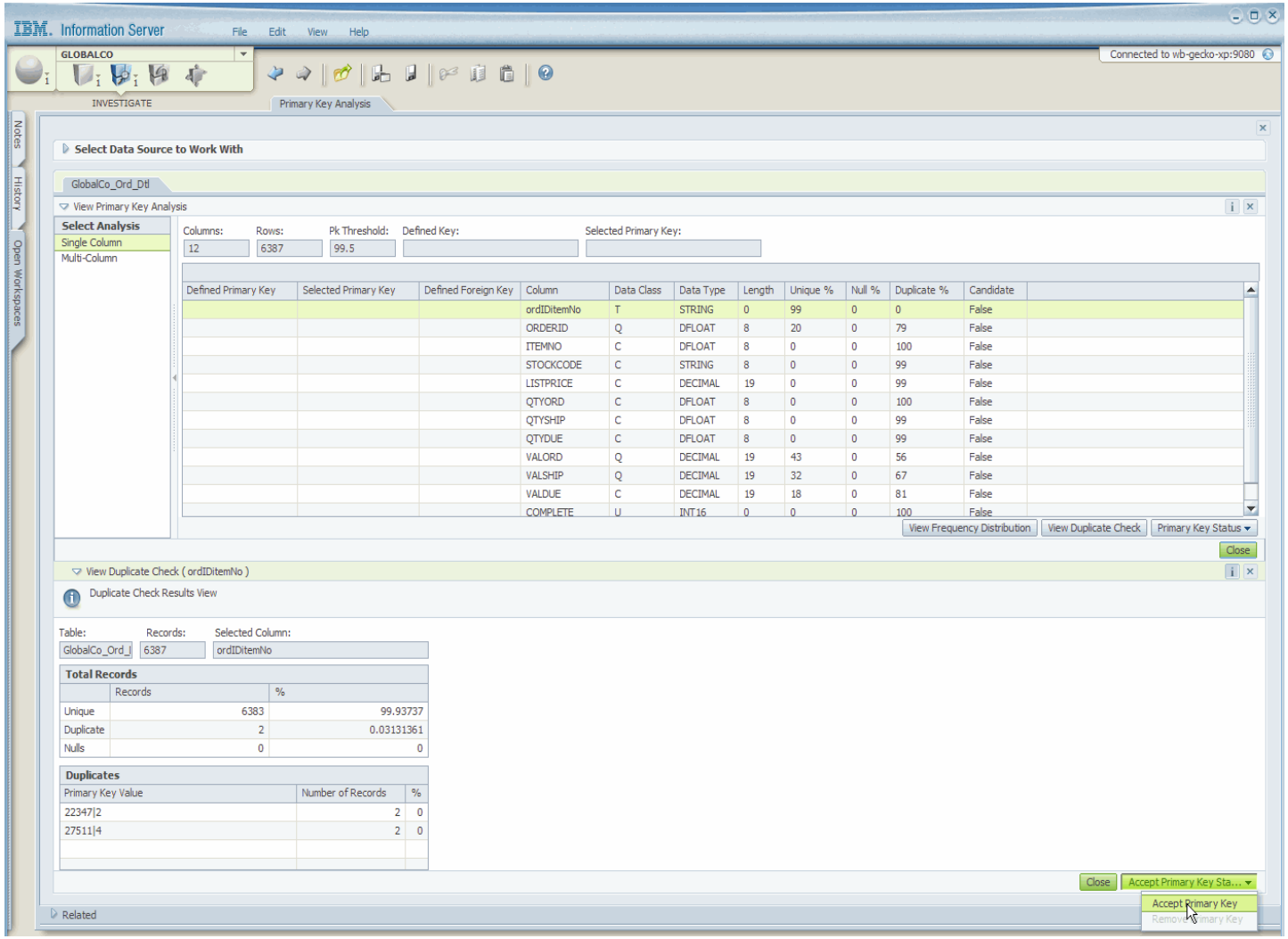


Table Name	Column Name	Data Type Date Format	Summary of Types Percent	 Count
datacleanuatdata.txt				
	ADDRESS	Char	100.00%	65,421
	C_ID	Smallint	0.30%	197
		Integer	98.66%	63,909
		Date MMDDYY	0.23%	146
		Date YYMMDD	0.81%	522
	C_NAME	Char	100.00%	65,419
	F_ID	Tinyint	100.00%	65,421
	I_ID	Smallint	0.19%	126

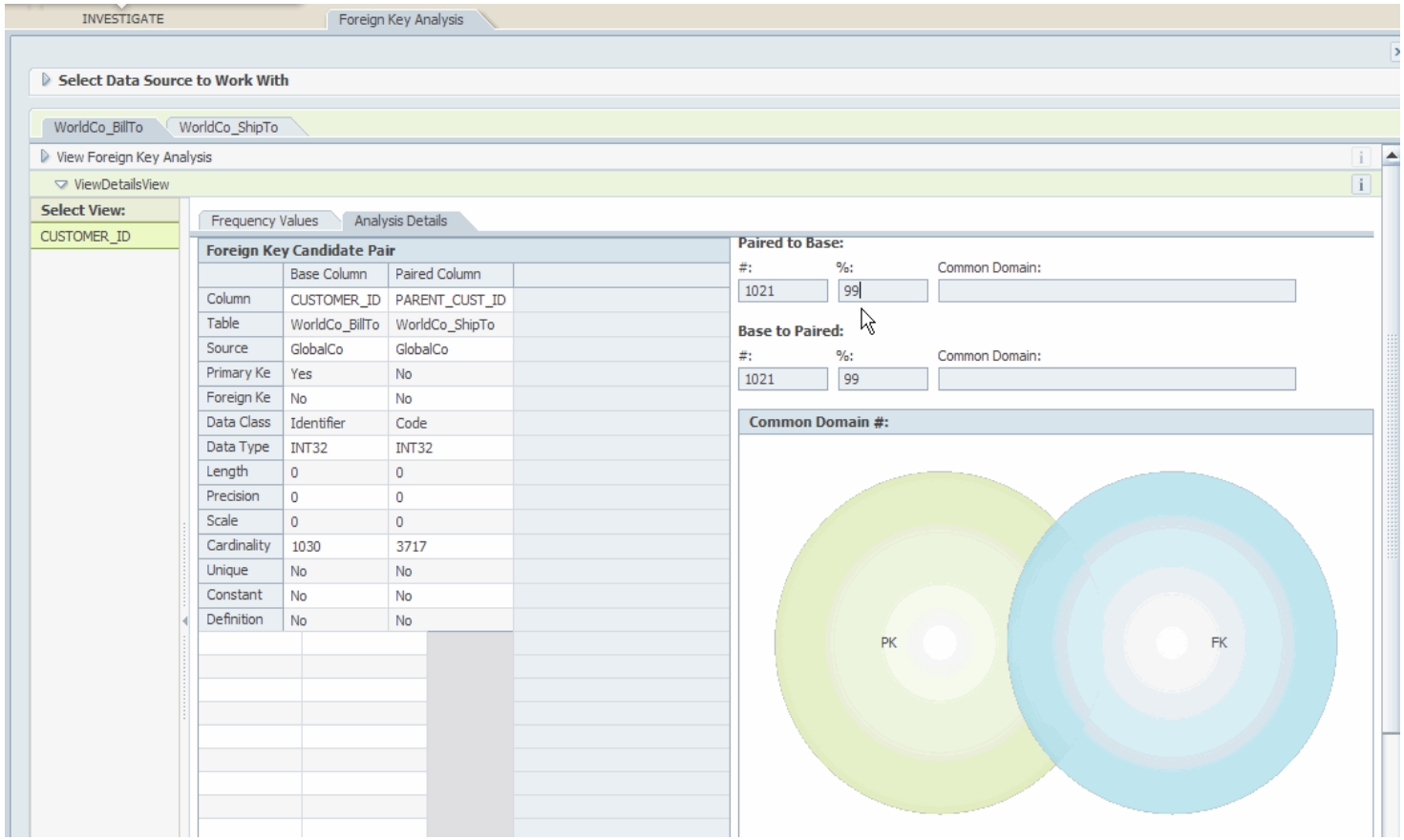
Data Profiling: Table Analysis

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



Data Profiling: Cross Table Analysis

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



Data Profiling: Baseline Analysis

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

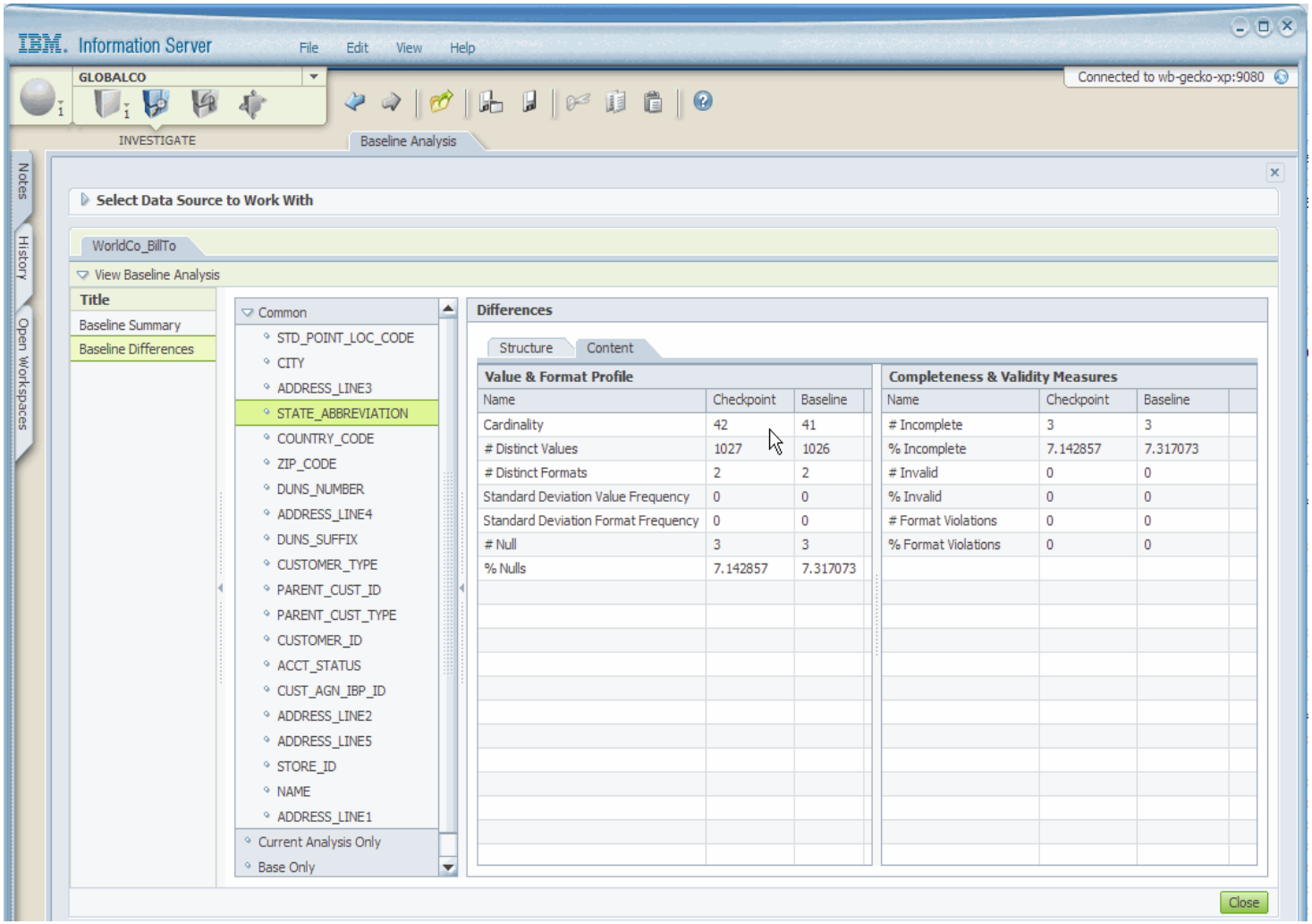
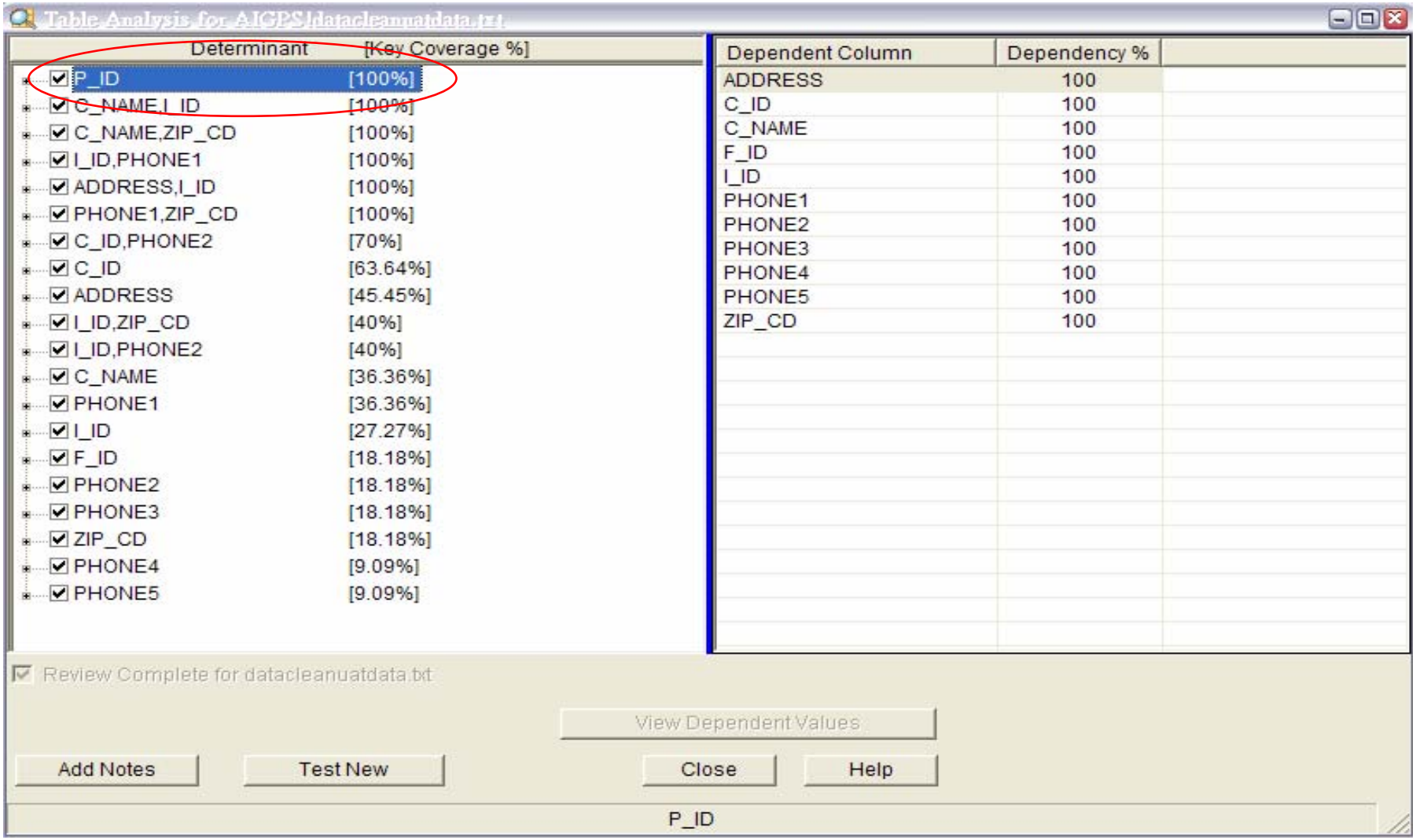
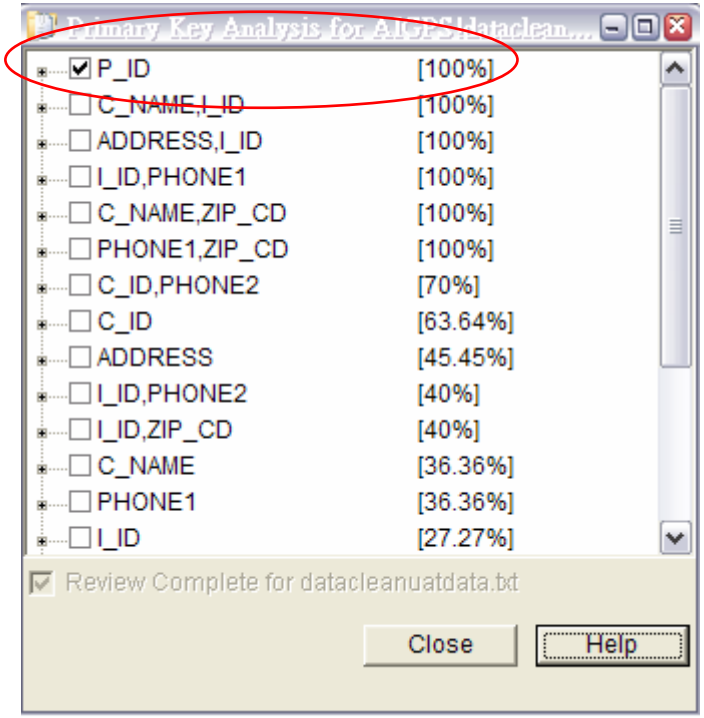


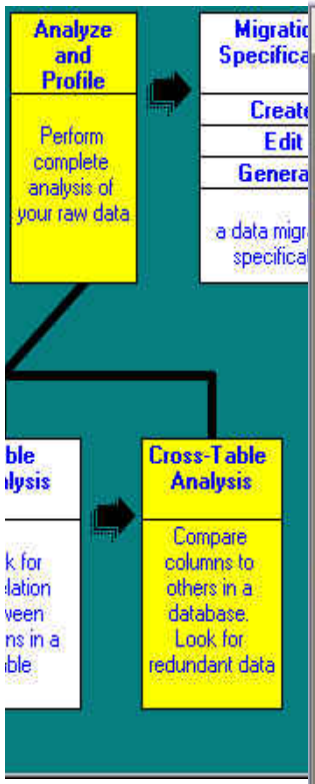
Table Analysis – Check Dependency



Primary Key Analysis – Check Dependency



Cross-Table Analysis – 確認兩個tables的從屬關係(Example)



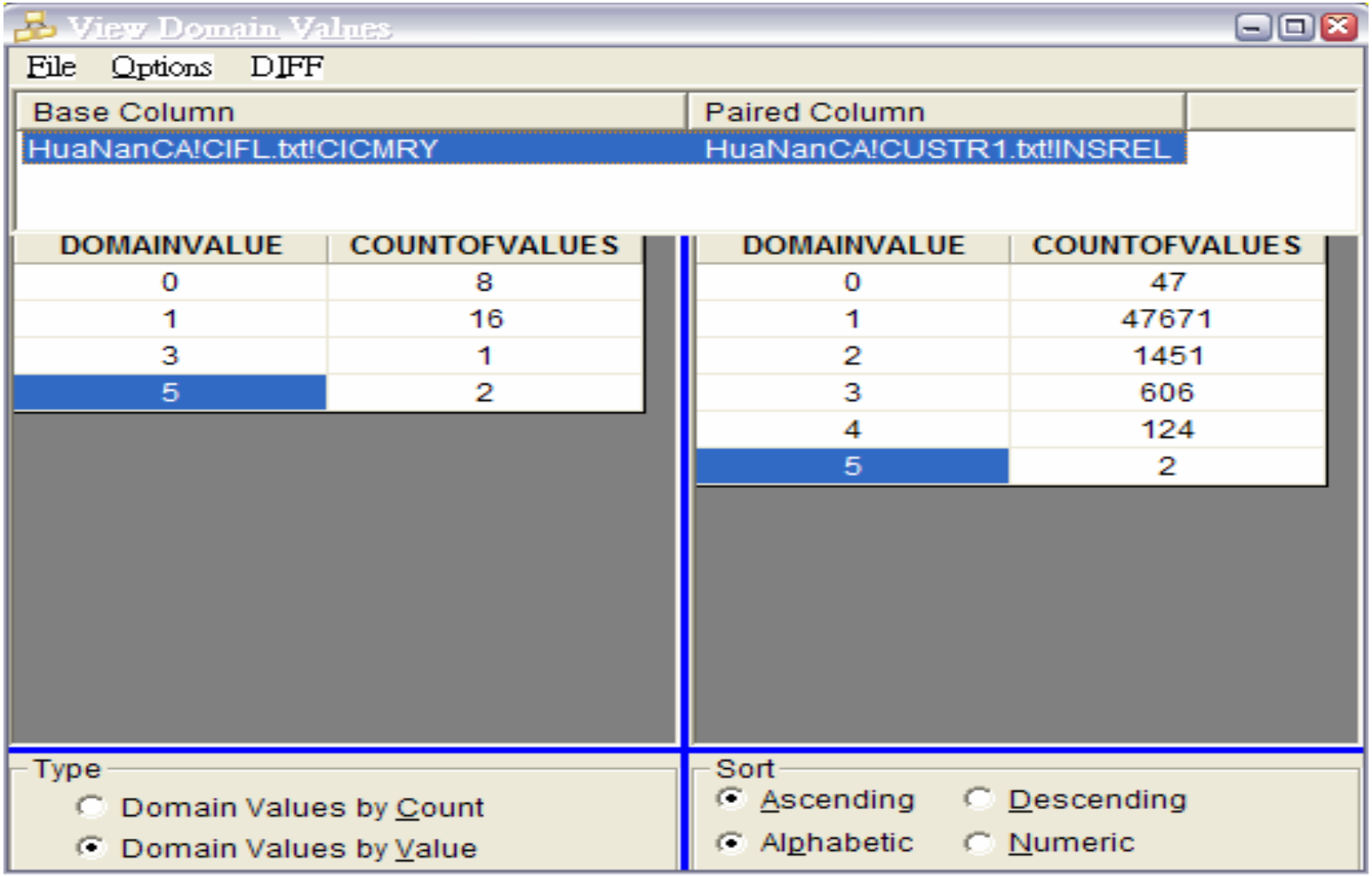
CIFL.txt Cross-Table Analysis Review

File

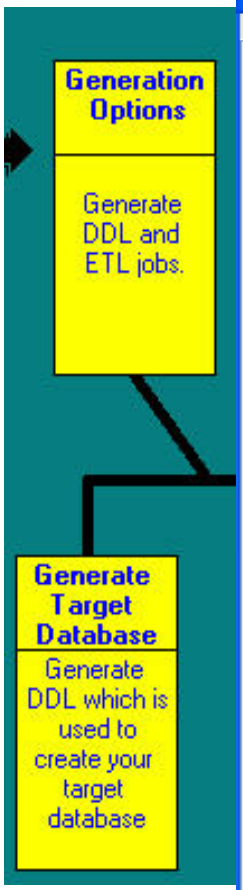
Drag a column header here to group by that column.

BASEDBNAME	BASETABLENAME	PAIREDDBNAME	PAIREDTABLENAME	BASECOLUMNNAME	PAIREDCOLUMNNAME	PERCENTAGE	EXCLUDEDFLAG
HuaNanCA	CIFL.txt	HuaNanCA	CUSTR1.txt	CICMRY	INSREL	100	N
		HuaNanCA	CUSTR1.txt	CICMRY	DEPENDENTS	100	N
		HuaNanCA	CUSTR1.txt	CICMRY	YR_IN_COMP	100	N
		HuaNanCA	CUSTR1.txt	CICMRY	INCOME_ANN	100	N
		HuaNanCA	CUSTR1.txt	CICMRY	YR_THERE	100	N
		HuaNanCA	CUSTR1.txt	CICWIL5	INSREL	100	N
		HuaNanCA	CUSTR1.txt	CICWIL8	INSREL	100	N
		HuaNanCA	CUSTR1.txt	CICWIL5	DEPENDENTS	100	N
		HuaNanCA	CUSTR1.txt	CICWIL8	DEPENDENTS	100	N
		HuaNanCA	CUSTR1.txt	CICWIL5	YR_IN_COMP	100	N
		HuaNanCA	CUSTR1.txt	CICWIL5	INCOME_ANN	100	N
		HuaNanCA	CUSTR1.txt	CICWIL5	YR_THERE	100	N
		HuaNanCA	CUSTR1.txt	CICWIL8	YR_IN_COMP	100	N
		HuaNanCA	CUSTR1.txt	CICWIL8	INCOME_ANN	100	N
		HuaNanCA	CUSTR1.txt	CICWIL8	YR_THERE	100	N
		HuaNanCA	CUSTR1.txt	CICCHI	YR_IN_COMP	100	N
		HuaNanCA	CUSTR1.txt	CICCHI	INCOME_ANN	100	N
		HuaNanCA	DI270.txt	CICWIL8	GCBUSTYP	100	N
	DI270.txt	HuaNanCA	CIFL.txt	GCBUSTYP	CICWIL8	100	N

確定從屬關係(Example)



產生目的資料庫的 table definition(DDL) (Example)



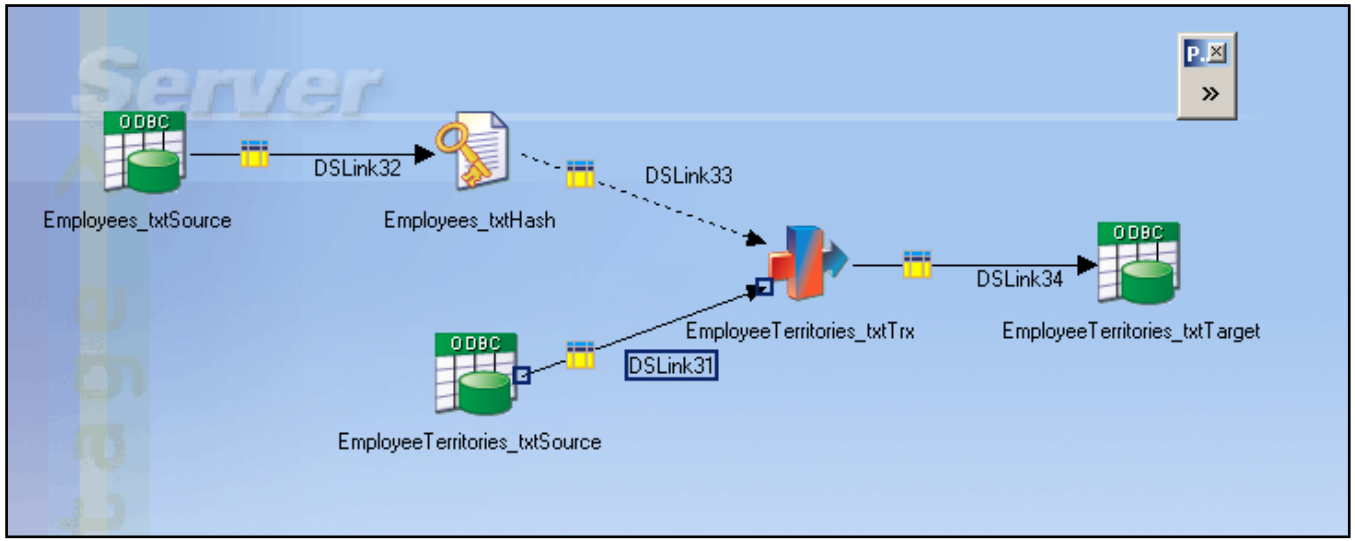
NewDB.sql - Notepad

File Edit Format View Help

```
--   SQL script file: C:\00_LANDPAD_00\NewDB.sql
--   DDL commands generated on Thu Aug 26 20:46:31 2004.
--
--   Database Name:      NewDB
--   Business Name:      New DB
--
--   Generation Options:
--       DropTable:      Yes
--       Primary Keys:   Yes
--       Foreign Keys:   Yes
--       Script File:    C:\00_LANDPAD_00\NewDB.sql
--
--   Destination Database Type: SQL Server -- 7.x

set QUOTED_IDENTIFIER on


-- Drop all referential constraints.

ALTER TABLE EmployeeTerritories
      DROP CONSTRAINT MGX_EmployeeTerritories_FK2
go

ALTER TABLE EmployeeTerritories
      DROP CONSTRAINT MGX_EmployeeTerritories_FK4
go

--   Table     Name: Employees
--   Business Name: Employees
DROP TABLE Employees
go

CREATE TABLE Employees (
      EmployeeID              smallint            NOT NULL,
      SSN                     char(11)            NOT NULL,
      LastName                varchar(14)         NOT NULL,
      FirstName               varchar(14)         NOT NULL,
      Title                   varchar(30)         NOT NULL,
      TitleofCourtesy         char(4)             NOT NULL,
      BirthDate               datetime            NOT NULL,
      HireDate                datetime            NOT NULL,
      Address                 varchar(32)         NOT NULL,
      City                    varchar(16)         NOT NULL,
      Region                  char(4)             NULL,
      PostalCode              char(7)             NOT NULL,
      Country                 char(3)             NOT NULL,
      HomePhone               char(14)            NOT NULL,
      Extension               smallint            NOT NULL,
      Notes                   varchar(34)         NULL,
      DivisionID              smallint            NOT NULL
)
go

ALTER TABLE Employees
      ADD  PRIMARY KEY (EmployeeID)
go
```

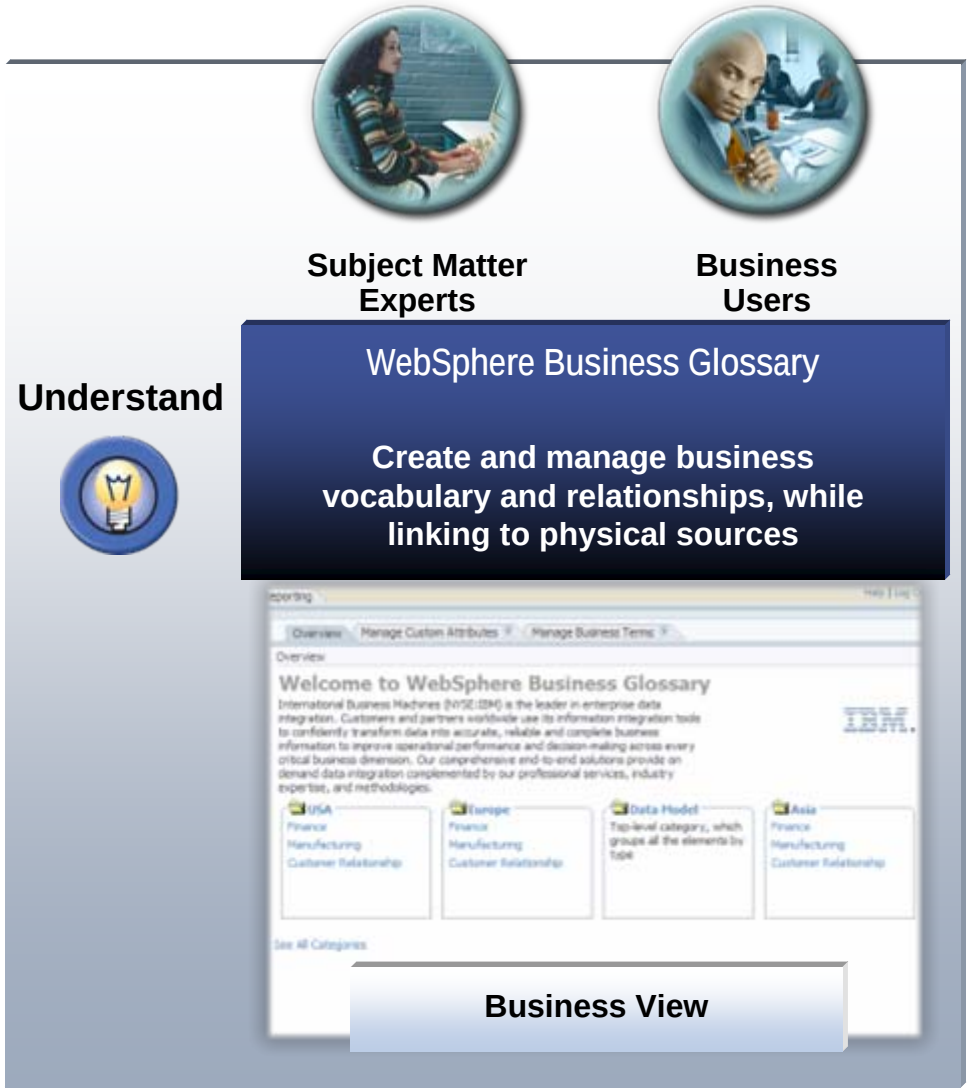
產生整合來源資料的ETL job -- 完美的結合！ (Example)

- Generate jobs to migrate source data to your target database (Oracle, SQL Server, etc.)
- Incorporates DataStage functions into mappings for use in the generated DataStage job definitions
- Automatically generates DataStage jobs (.dsx files) that can be imported to DataStage



Business Understanding: WebSphere Business Glossary

- Allows business users to record their view of the business
  - Aligns business & IT for better results
- Provides business context to information technology assets
  - Reduces project risk & shortens time to value
- Establishes responsibility and accountability
  - Establishes data governance and control



Database = DB2

Schema = NAACCT

Table = DLYTRANS

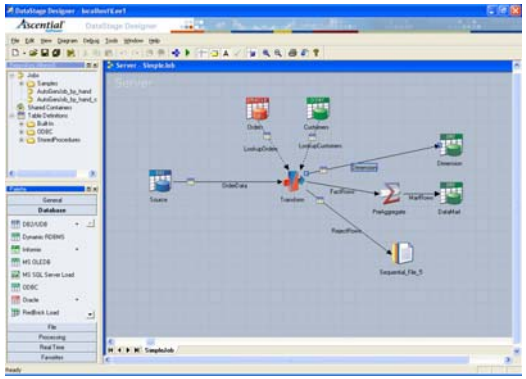
Column = ACCT_NO

data type = char(11)

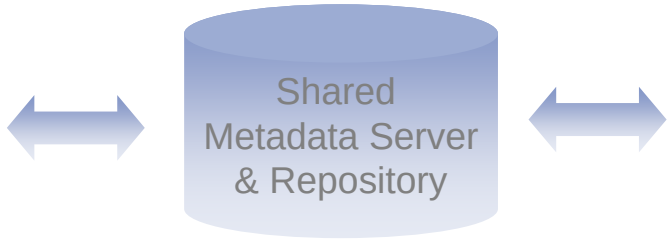
GL Account Number

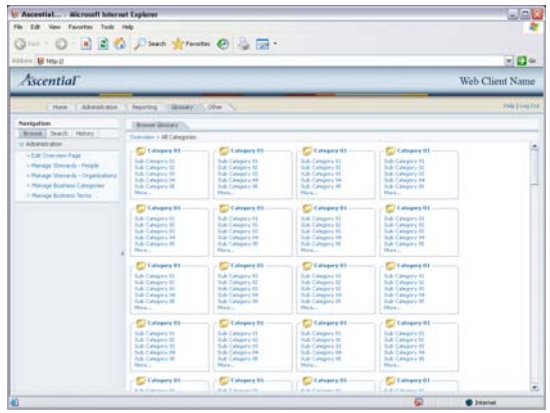
The ten digit account number. Sometimes referred to as the account ID. This value is of the form L-FIIIIVVVV.

Benefit: Create a common vocabulary between business & technical users



WebSphere DataStage





WebSphere Business Glossary

Architectural Understanding: Rational Data Architect

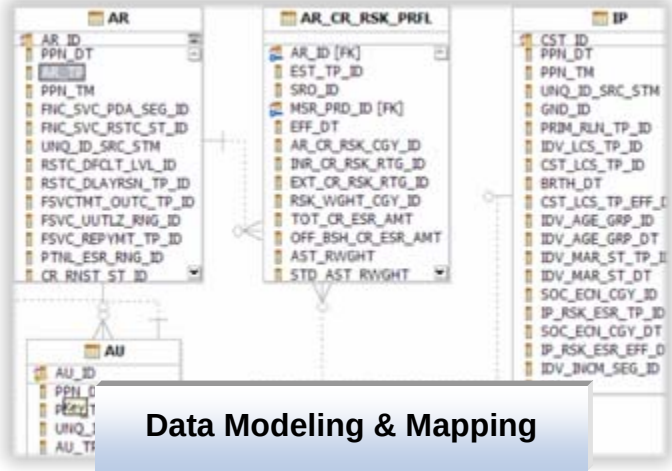
- Allows data models to be linked to sources and business terms
  - Facilitates alignment of business & IT for better results
- Discovers & maps relationships across models and sources
  - Reduces time to value and speeds project implementations
- Works with IBM Industry Models
  - Accelerates projects and provides a proven industry foundation

Subject Matter Experts

Architects

Rational Data Architect

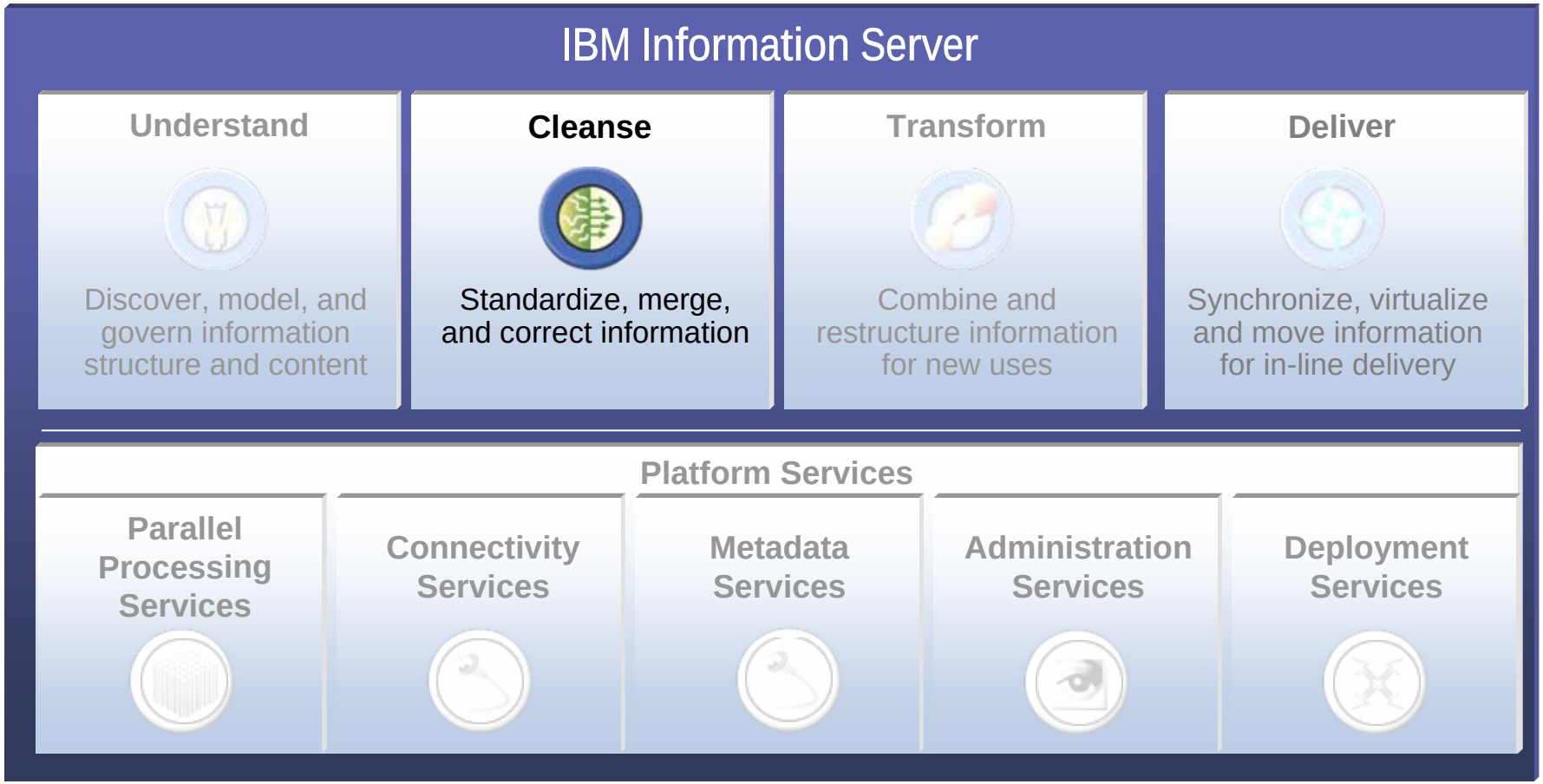
Create and manage business vocabulary and relationships, while linking to physical sources



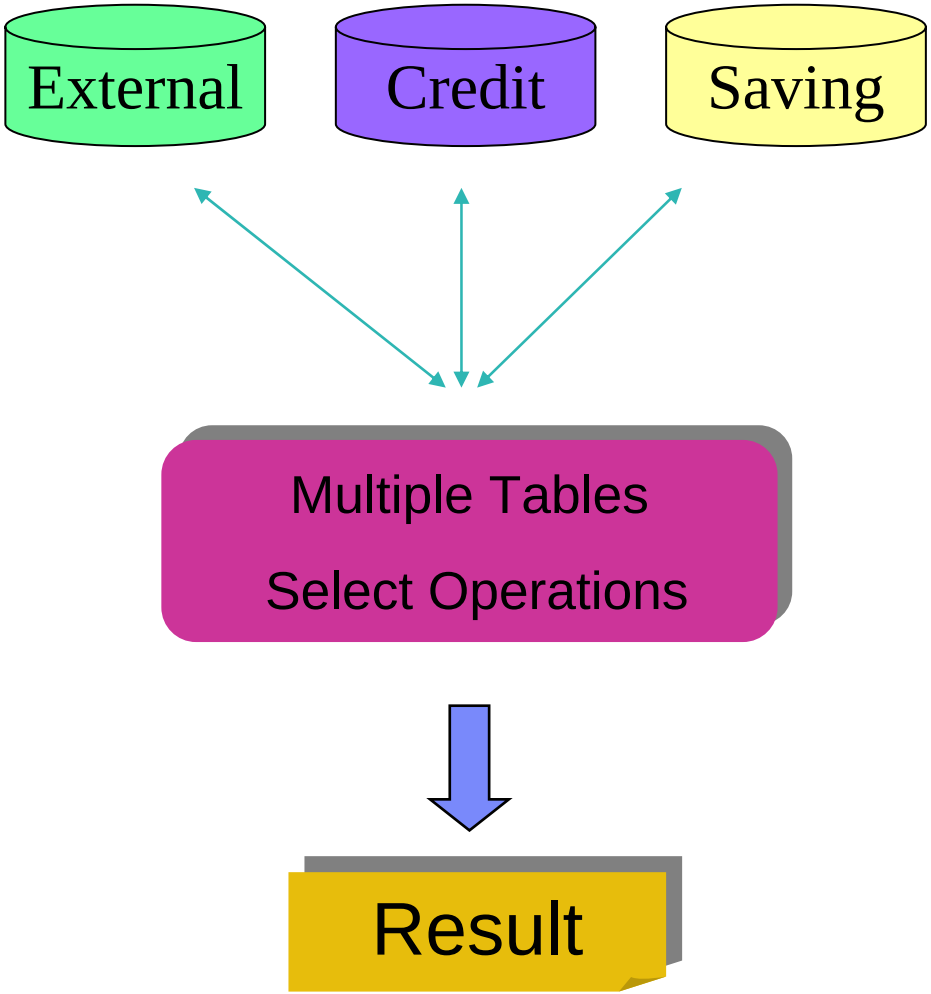
Data Modeling & Mapping

The IBM Solution: IBM Information Server

Delivering information you can trust

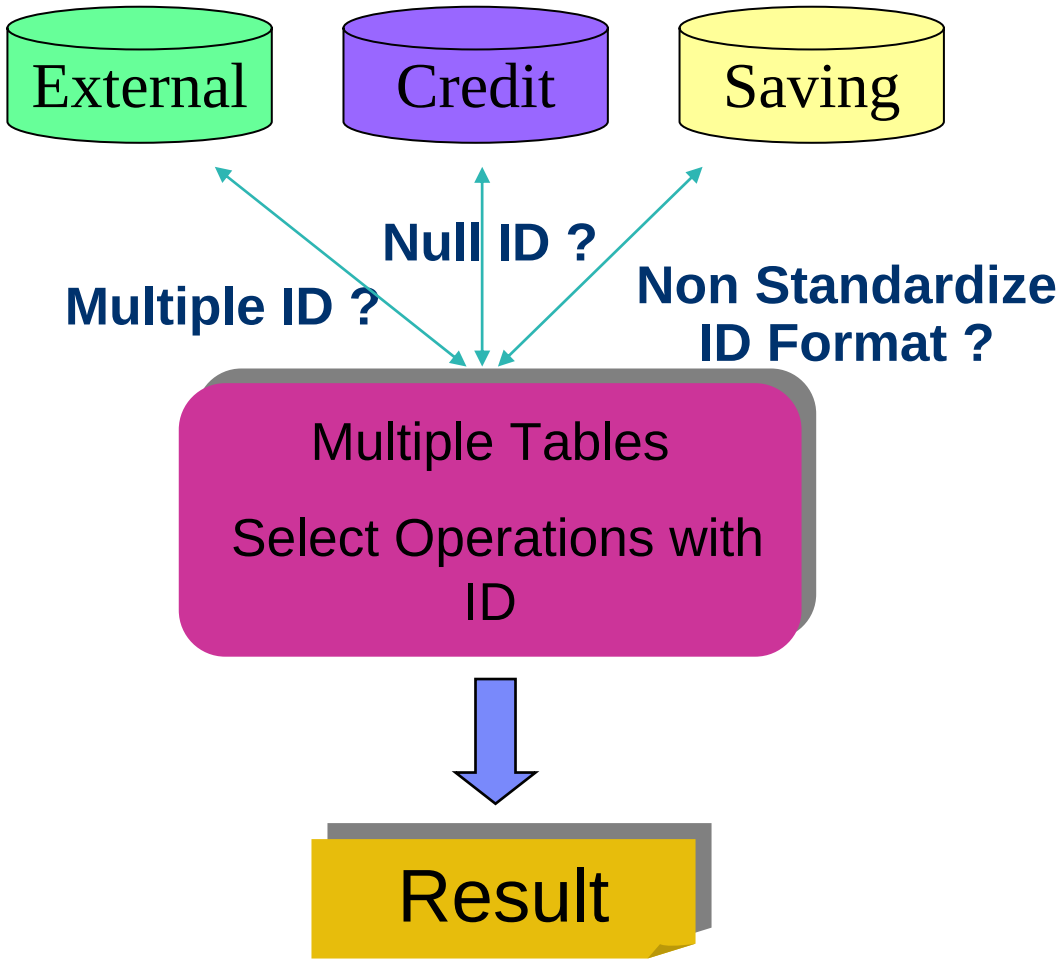


How To Do? By ID Number?



- **Use ID Number to find out the customers in Saving Account System but not neither in External Database nor Credit Card System**

Is it that simple?????



Actual Situation

External Database

Lack of Standard

Unique Key	Identification Number	Title	Name	Age	DOB	Gender	MS	Credit Card Acct	Address
BACD-0001	z000000(1)	先生	郭靖	45	24/6/1959	M	N	5886-8466-8295-7605	四川省阿坝藏族羌族自治州
BACD-0002	z0000002	先生	张三丰	99	1/11/1905	M	N	9912210629331030	南宁市太平洋世纪广场汇春路
BACD-0003	s000000(3)	先生	司马懿	72	8/8/1932	M	N	8827-6156-5365-7720	重庆市高新区科园二路 7 号
BACD-0004	z000000(4)	先生	张之洞	91	6/7/1913	M	N	3813592345491130	西安市南二环永松路西何家村
BACD-0005		先生	陈平	42	28/11/1962	M	Y	7335910007512850	福州市鼓楼区西洪路 181 号
BACD-0006	z000000-6	女仕	苏小小	105	2/2/1900		N	2939787387971440	合肥市双河三村 11 栋 406
BACD-0007	000000(7)	女仕	黄蓉	28	13/1/197		Y	60	羊西线蜀汉路428号老房子酒
BACD-0008	z000-000(8)	女仕	张菁	26	9/11/197	M	Y	16	成都市羊西线产渐醒区围城大
BACD-0009	z000000/(9)	先生	张良	77	21/9/1927	M	N	2635-7651-8358-0435	羊西线御都花园别墅旁，金都
BACD-0010	z000001(0)	女仕	傅红雪	59	21/9/1945	M	Y	4716323283136880	北京市 延 庆 县 延 庆 镇
BACD-0011	z000001(1)	先生	韩信	18	17/4/1986	M	Y	8978667939305010	北京市朝阳区建国门外大街建
BACD-0012	z000001(2)	先生	任我行	19	22/2/1986	M	Y	2070056786285840	陕西 省 大 荔 县 冯 翊 路
BACD-0013	z000001(3)	先生	林平之	77	22/7/1927	M	N	6263-9030-1596-6404	中 央 路 3 5 巷

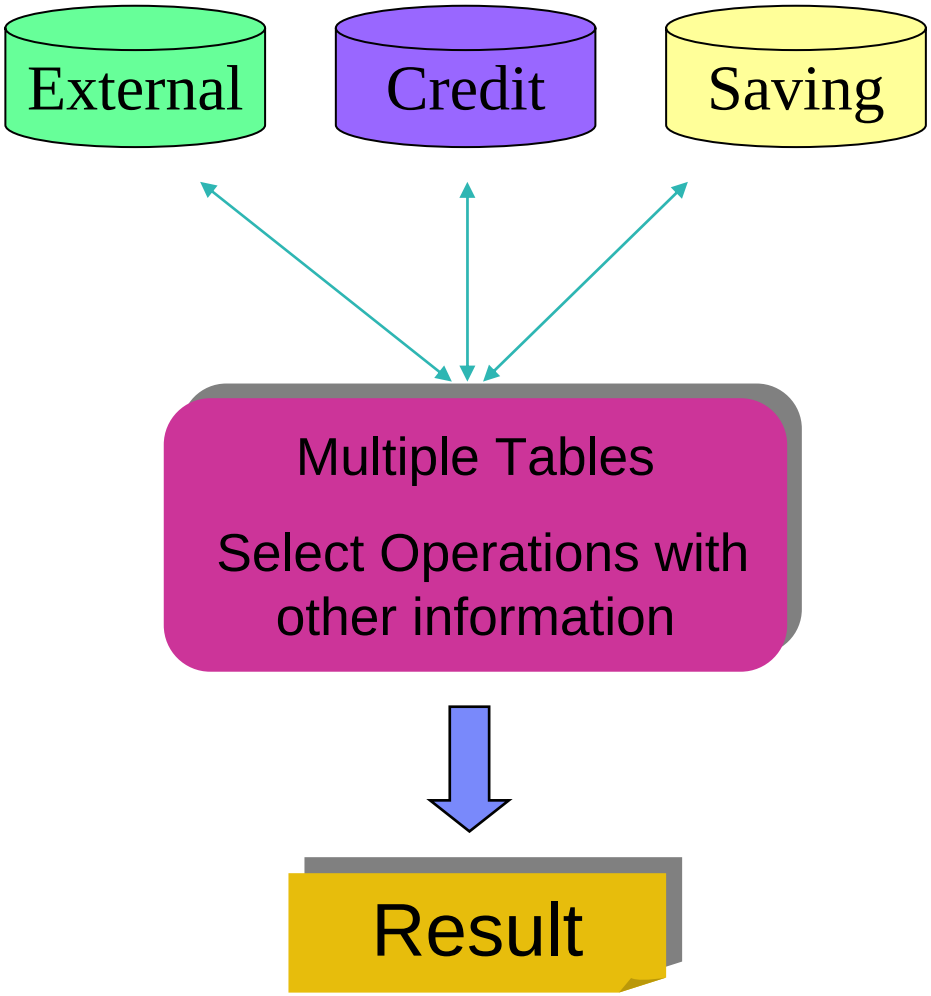
Multiple IDs?

Null Value ?

HSDB Saving

Unique Key	Last Name	First Name	Title	ID	Age	DOB	Martial Status	Gender	Saving Card Number	City
BBS-0001	汤	汉斯	先生	z0000016	42	30/6/1962	Y	M	7210768211	青铜
BBS-0002	黄	蓉	女仕	z000000(7)	28	13/1/1977	Y	F	7212130654	应城
BBS-0003	郭	靖	先生	z000000(1)	45	24/6/1959		M	7217255356	四川
BBS-0004	陈方	安生	女仕	W854253(6)	30	12/1/1975	Y	F	721-8-742464	孝感
BBS-0005	方	安生	女仕	WA8542536	30	12/1/1975	Y	F	7210082353	贵州
BBS-0006	苏	小小	女仕	z000000-6	105	14/2/1895	Y	F	7216811477	
BBS-0007	花	木兰	女仕	z000001(5)	54	30/6/1950		F	7214380818	
BBS-0008	姚	明	先生		82	16/1/1923	N	M	7217853124	吴忠
BBS-0009	克	林顿	先生	z000001(7)	36	10/9/1968		F	721-1-861172	基隆
BBS-0010	萧	十一郎	先生	z10000004	77	21/9/1927	Y	M	7217177272	
BBS-1011	萧	十一郎	先生	z100000(4)	77	21/9/1927	Y	M	721-0-100770	北京
BBS-0012	苏	轼	先生	z9876542	87	22/8/1917	Y	M	7213585546	福州

How about other information?



- We cannot simply use ID number, then how about use other information?
- Such as Name, phone number, address…etc
- Can it really solve the problem?

Different System Format

Unique Key	Identification Number	Title	Name	Age	DOB	Gender	MS	Credit Card Acct	Address	Phone
BACD-0001	z000000(1)	先生	郭靖	45	24/6/1959	M	N	5886-8466-8295-7605	四川省阿坝藏族羌族自治州九寨沟县漳扎镇九安宾馆三层零二室	(86)-27312878

Different Structure

Unique Key	Last Name	First Name	Title	ID	Age	DOB	Martial Status	Gender	Saving Card Number	City	Street Address	Building Name	Block Number	Floor Number	Room Number
BBS-0003	郭	靖	先生	z000000(1)	45	24/6/1959		M	7217255356	四川省阿坝藏族羌族自治州九寨沟县漳扎镇		九安宾馆			三零二室

Record from one system

Unique Key	ID/Passport Number	Family Name	Given Name	Salutation	Age	MS	Gender	Credit Card Number	Province	City	District	Building Name	Floor Number	Room Number	Phone Number
BBC-0001	X237470(2)	郭	靖	先生	45	Y	M	7791742551504780	四川		漳扎镇	九安宾馆		302室	(86)-87543237
BBC-1121	X237470-2	郭	靖	先生	45	Y		4874-8493-1921-1492	四川省	阿坝藏族羌族自治州					(+86)-028-87543237
BBC-2113	WK3788624	郭	靖				M	6323409777617890		阿坝藏族羌族自治州 漳扎镇		九安宾馆		三零二室	(86)-27312878
BBC-2534	WM3802367	郭	靖	先生	45			3443-5345-1086-2270	四川		九寨沟县				028-87543237
BBC-3121	W380236(7)	郭	靖		45	Y	M	4833-1227-5409-4733				九安宾馆	3层	02室	13987654321

Null Value

Wrong Fields

Records from different system

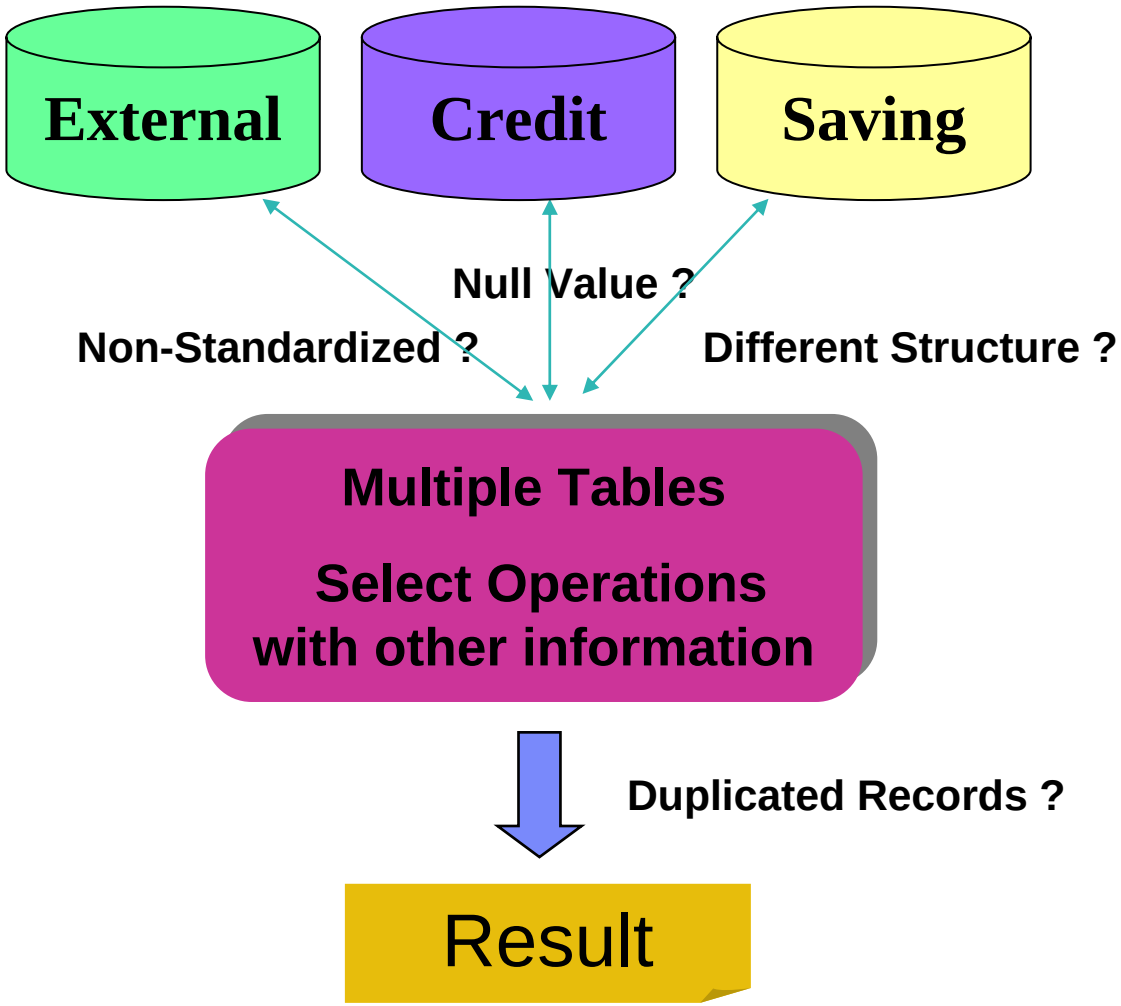
Unique Key	Identification Number	Title	Name	Age	DOB	Gender	MS	Credit Card Acct	Address	Phone
BACD-0007	000000(7)	女仕	黄蓉	28	13/1/1977		Y	6056314333726720	北京大兴区双河南里天兴公寓四单元五楼八室	(86)-021-2789-4133

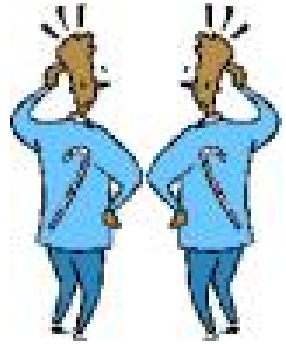
Unique Key	**ID/Passport Number**	Family Name	Given Name	Salutation	Age	MS	Gender	Credit Card Number	Province	City	District	Street Address	Building Name	Block Number	Floor Number	Room Number	Phone Number
BBC-0021	BK357061-3	黄	蓉	女仕	28	Y	F	3086-3594-9679-9788			大兴区	双河南里	天兴公寓		5楼		021-2789-4133

Standard ?

Unique Key	Last Name	First Name	Title	ID	Age	DOB	Martial Status	Gender	Saving Card Number	City	Street Address	Building Name	Block Number	Floor Number	Room Number
BBS-0002	黄	蓉	女仕	z000000(7)	28	13/1/1977	Y	F	7212130654			天兴公寓 4单元		5楼	8室

A solution????


External
Credit
Saving
Null Value ?
Non-Standardized ?
Different Structure ?
Multiple Tables
Select Operations
with other information
Duplicated Records ?
Result



The Data Quality Challenge

- **Lack of information standards**
  - Different formats & structures across different systems
- **Data surprises in individual fields**
  - Data misplaced in the database
- **Information buried in free-form fields**
- **Data myopia**
  - Lack of consistent identifiers inhibit a single view
- **The redundancy nightmare**
  - Duplicate records with a lack of standards

```
Kate A. Roberts   416 Columbus Ave #2, Boston, Mass 02116
Catherine Roberts Four sixteen Columbus APT2, Boston, MA 02116
Mrs. K. Roberts   416 Columbus Suite #2, Suffolk County 02116
```

Name	Tax ID	Telephone
J Smith DBA Lime Cons.	228-02-1975	6173380300
Williams & Co. C/O Bill	025-37-1888	415-392-2000
1st Natl Provident	34-2671434	3380321
HP 15 State St.	508-466-1200	Orlando

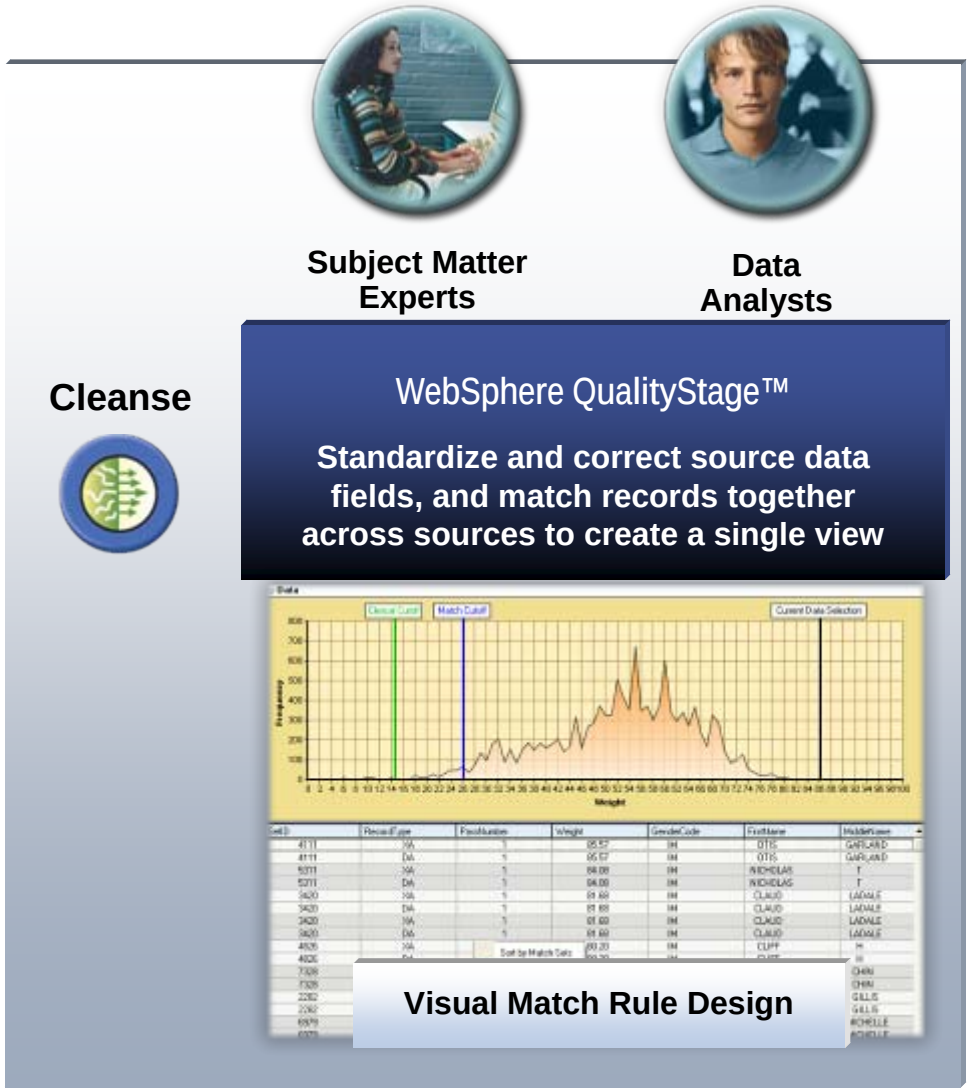
```
WING ASSY DRILL 4 HOLE USE 5J868A HEXBOLT 1/4 INCH
WING ASSEMBY, USE 5J868-A HEX BOLT .25" - DRILL FOUR HOLES
USE 4 5J868A BOLTS (HEX .25) - DRILL HOLES FOR EA ON WING ASSEM
RUDER, TAP 6 WHOLES, SECURE W/KL2301 RIVETS (10 CM)
```

```
19-84-103    RS232 Cable 6' M-F CandS
CS-89641     6 ft. Cable Male-F, RS232 #87951
C&SUCH6      Male/Female 25 PIN 6 Foot Cable
```

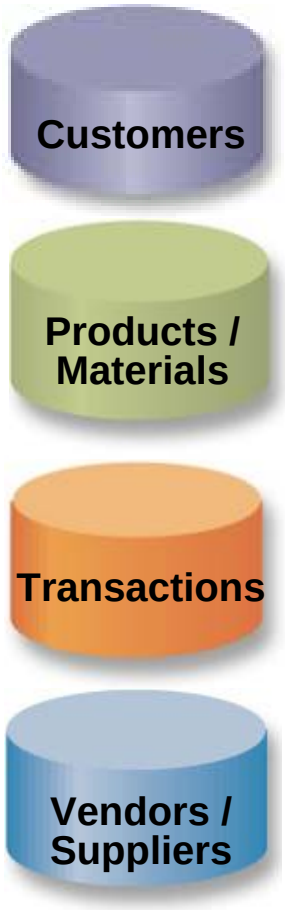
```
90328574    IBM                  187 N.Pk. Str. Salem NH 01456
90328575    I.B.M. Inc.          187 N.Pk. St. Salem NH 01456
90238495    Int. Bus. Machines   187 No. Park St Salem NH 04156
90233479    International Bus. M. 187  Park Ave Salem NH 04156
90233489    Inter-Nation Consults 15 Main Street Andover MA 02341
90345672    I.B. Manufacturing   Park Blvd. Bostno MA  04106
```

Data Cleansing: WebSphere QualityStage

- Ensures clean, standardized, de-duplicated information
  - Reduces project risk and supports better business results
- Matches together records across systems
  - Enables a single version of the truth
- Supports global postal verification
  - Cleanses international data to support requirements across geographies



How Does WebSphere QualityStage Work?



Cleansing Process

1. **Data Investigation**
2. **Data Standardization**
3. **Data Matching**
4. **Data Survivorship**

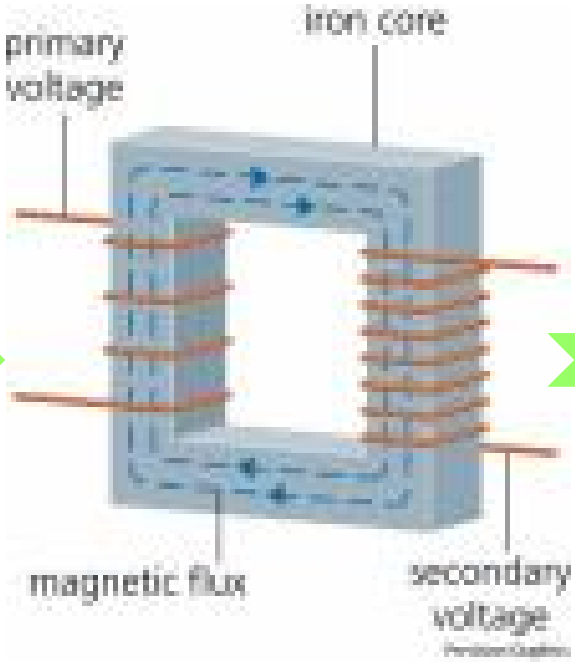
Accurate, cleansed data that drives critical decisions

Merge Sources Files

External

Credit

Saving



Merged File

File Format

External Database	HSBD Credit	HSBD Saving
Unique Key	Unique Key	Unique Key
Identification Number	ID/Passport	Last Name
Title	Family Name	First Name
Name	Given Name	Title
Age	Salutation	ID
DOB	Age	Age
Gender	MS	DOB
MS	Gender	Martial Status
Credit. Acct	Credit Card Number	Gender
Address	Province	Saving Card Number
Phone	City	City
	District	Street Address
	Street Address	Building Name
	Building Name	Block Number
	Block Number	Floor Number
	Floor Number	Room Number
	Room Number	Country Code
	Phone Number	Phone Number
		Extension

Unique Record Key
Source File Description
Sequential Record Number
Original Unique Key
ID/Passport Number
Title
Name Gender Martial Status
Name
Gender
Age
Date Of Birth
Martial Status
Address
Phone Number
Saving Account
Credit Card Number

Sources Files Format

External DataBase

Unique Key	Identification Number	Title	Name	Age	DOB	Gender	MS	Credit Card Acct	Address
BACD-0001	z000000(1)	先生	郭靖	45	24/6/1959	M	N	5886-8466-8295-7605	四川省阿坝藏族羌族自治州九
BACD-0002	z0000002	先生	张三丰	99	1/11/1905	M	N	9912210629331030	南宁市太平洋世纪广场汇春路
BACD-0003	s000000(3)	先生	司马懿	72	8/8/1932	M	N	8827-6156-5365-7720	重庆市高新区科园二路 7 号
BACD-0004	z000000(4)	先生	张之洞	91	6/7/1913	M	N	3813592345491130	西安市南二环永松路西何家村
BACD-0005		先生	陈平	42	28/11/1962	M	Y	7335910007512850	福州市鼓楼区西洪路 181 号
BACD-0006	z000000-6	女仕	苏小小	105	2/2/1900		N	2939787387971440	合肥市双河三村 11 栋 406
BACD-0007	000000(7)								
BACD-0008	z000-000(8)								
BACD-0009	z000000/(9)								
BACD-0010	z000001(0)								
BACD-0011	z000001(1)								
BACD-0012	z000001(2)								
BACD-0013	z000001(3)								

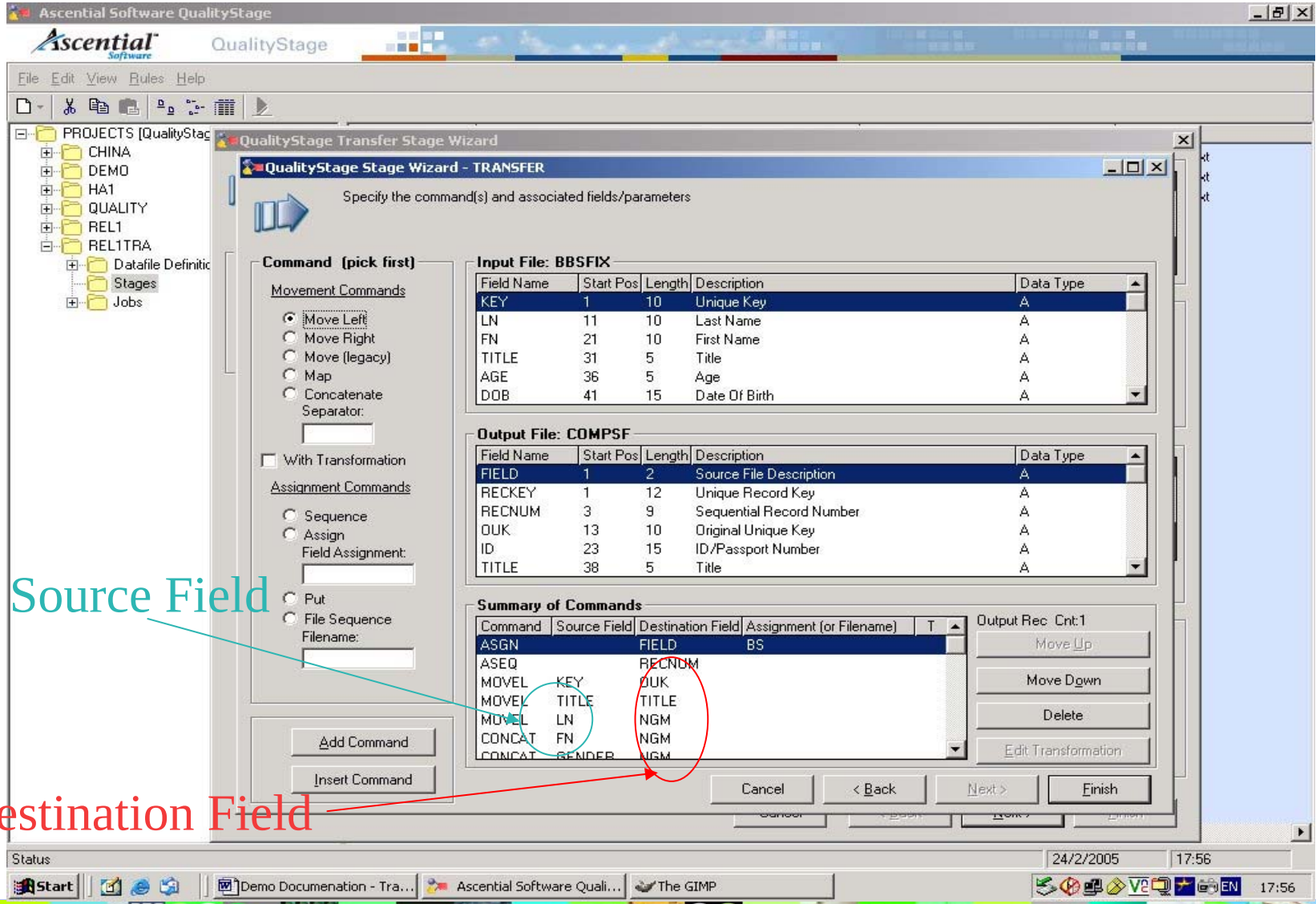
HSBD Credit

Unique Key	ID/Passport Number	Family Name	Given Name	Salutation	Age	MS	Gender	Credit Card Number	Province	City
BBC-0001	X237470(2)	郭	靖	先生	45	Y	M	7791742551504780	四川	
BBC-1121	X237470-2	郭	靖	先生	45	Y		4874-8493-1921-1492	四川省	阿坝藏族羌族自治
BBC-2113	WK3788624	郭	靖				M	6323409777617890		阿坝藏族羌族自治
BBC-2534	WM3802367	郭	靖	先生	45			3443-5345-1086-2270		阿坝藏族羌族自治
BBC-3121	W380236(7)	郭	靖		45	Y	M	4833-1227-5409-4733	四川	
BBC-1851	W854253(6)	陈	方安生	女仕	30	Y	F	3103291894201560		
BBC-0007	WA8542536	方	安生	女仕	30	Y		7357-4523-1038-0431		
BBC-9675	AT207									
BBC-1889	AL207									
BBC-3117	CX276									
BBC-8991	CX276									

HSDB Saving

Unique Key	Last Name	First Name	Title	ID	Age	DOB	Martial Status	Gender	Saving Card Number	City
BBS-0001	汤	汉斯	先生	z0000016	42	30/6/1962	Y	M	7210768211	青铜
BBS-0002	黄	蓉	女仕	z000000(7)	28	13/1/1977	Y	F	7212130654	应城
BBS-0003	郭	靖	先生	z000000(1)	45	24/6/1959		M	7217255356	四川
BBS-0004	陈方	安生	女仕	W854253(6)	30	12/1/1975	Y	F	721-8-742464	孝感
BBS-0005	方	安生	女仕	WA8542536	30	12/1/1975	Y	F	7210082353	贵州
BBS-0006	苏	小小	女仕	z000000-6	105	14/2/1895	Y	F	7216811477	
BBS-0007	花	木兰	女仕	z000001(5)	54	30/6/1950		F	7214380818	
BBS-0008	姚	明	先生		82	16/1/1923	N	M	7217853124	吴忠
BBS-0009	克	林顿	先生	z000001(7)	36	10/9/1968		F	721-1-861172	基隆
BBS-0010	萧	十一郎	先生	z10000004	77	21/9/1927	Y	M	7217177272	
BBS-1011	萧	十一郎	先生	z100000(4)	77	21/9/1927	Y	M	721-0-100770	北京
BBS-0012	苏	轼	先生	z9876542	87	22/8/1917	Y	M	7213585546	福州

Transfer from Source Files to Target File

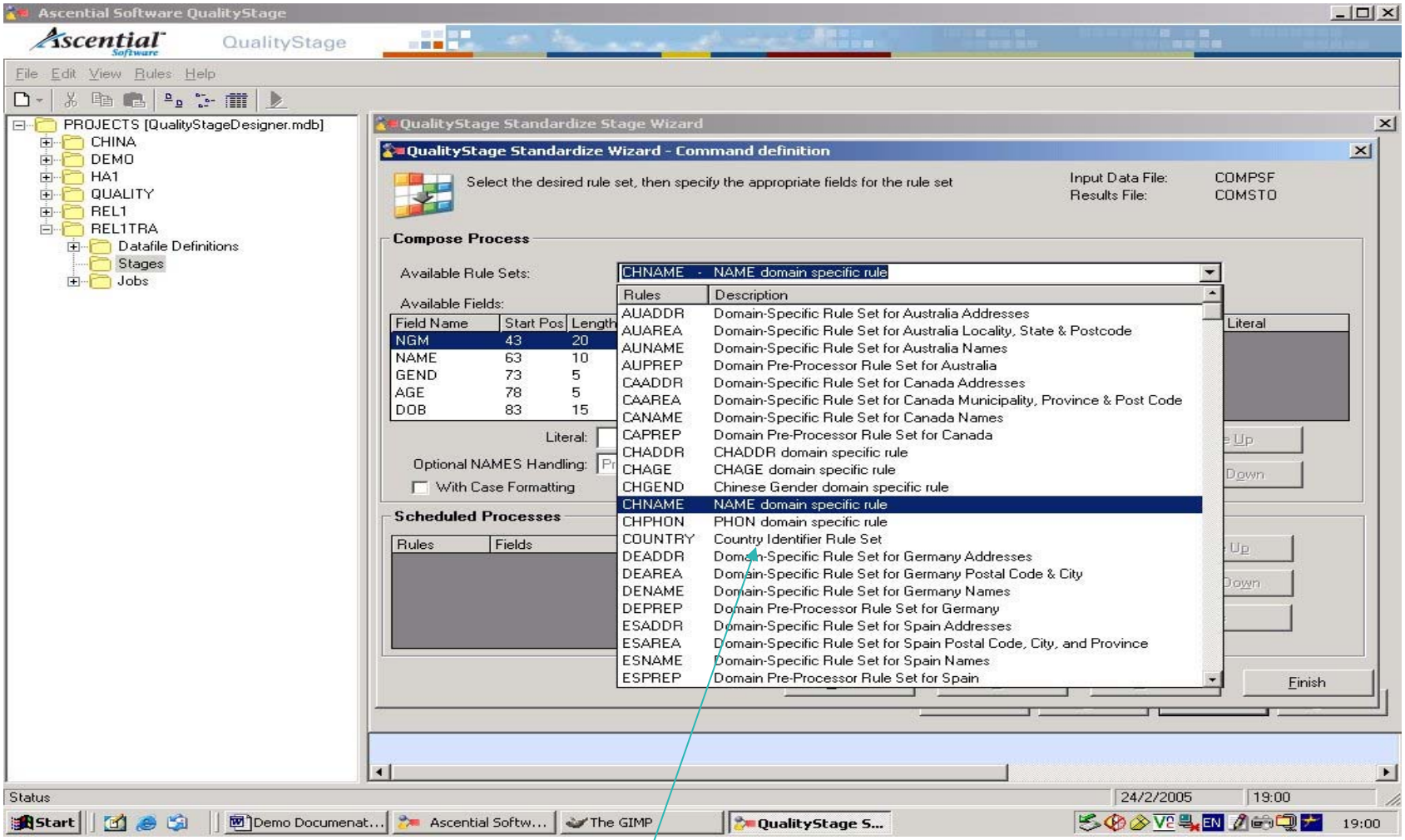


Target File

- Without coding a line of program
- Simple fields selection and define

RECKEY		OUK	ID	TITLE	NGM	NAME	GEND	AGE	DOB
BS000000001	1	Unique Key	ID	Title	Last Name First Name	Last Name	Gende	Age	DOB
BS000000002	2	BBS-0001	z0000016	先生	汤 汉斯_M_Y	汤 汉斯	M	42	30/6/1962
BS000000003	3	BBS-0002	z000000(7)	女仕	黄 蓉_F_Y	黄 蓉	F	28	13/1/1977
BS000000004	4	BBS-0003	z000000(1)	先生	郭 靖_M_	郭 靖	M	45	24/6/1959
BS000000005	5	BBS-0004	W854253(6)	女仕	陈方 安生_F_Y	陈方 安生	F	30	12/1/1975
BS000000006	6	BBS-0005	WA8542536	女仕	方 安生_F_Y	方 安生	F	30	12/1/1975
BS000000007	7	BBS-0006	z000000-6	女仕	苏 小小_F_Y	苏 小小	F	105	14/2/1895
BS000000008	8	BBS-0007	z000001(5)	女仕	花 木兰_F_	花 木兰	F	54	30/6/1950
BS000000009	9	BBS-0008		先生	姚 明_M_N	姚 明	M	82	16/1/1923
BS000000010	10	BBS-0009	z000001(7)	先生	克 林顿_F_	克 林顿	F	36	10/9/1968
BS000000011	11	BBS-0010	z10000004	先生	萧 十一郎_M_Y	萧 十一郎	M	77	21/9/1927
BS000000012	12	BBS-1011	z100000(4)	先生	萧 十一郎_M_Y	萧 十一郎	M	77	21/9/1927
BS000000013	13	BBS-0012	z9876542	先生	苏 轼_M_Y	苏 轼	M	87	22/8/1917
BS000000014	14	BBS-0013	z9876541	女仕	郭 襄_F_Y	郭 襄	F	44	22/3/1960
BS000000015	15	BBS-0014	z9876540	女仕	黄 忠_M_Y	黄 忠	M	1	12/1/2004
BS000000016	16	BBS-0015	z9876539	先生	汤 告鲁斯_M_Y	汤 告鲁斯	M	81	9/8/1923

Standardization – By simple rules selection



Select Appropriate Rules

Standardize Results - Name

Input Name	Symbol	Family Surname	Family Name	Given Name	Gender	Martial Status
闻人	C	闻	闻	人	NA	NA
王田	FF	王	王	田	NA	NA
张良	F+	张	张	良	NA	NA
人马	+F	人	人	马	NA	NA
顺治	++	顺	顺	治	NA	NA
夏侯惇	C+	夏侯	夏侯	惇	M	Y
項虞姬	FFF	项	项虞	姬	F	Y
朱元璋	FF+	朱	朱	元璋	M	N
趙李錢孫	FFFF	赵	赵 李	钱孙	F	Y
雷公羊和	FCF	雷	雷公羊	和	F	Y
上官雷和	CFF	上官	上官雷	和	F	Y
上官雷和	CFF	上官	上官	雷和	M	Y
趙李錢孫	FFFF	赵	赵	李钱孙	M	Y
上官雷和	CFF	上官	上官	雷和	M	Y
雷公羊和	CC	雷	雷公羊	和	M	Y
上官雷和	CFF	上官	上官	雷和		Y

Standardize Results - Phone

Input Number	Symbol	Country Code	Dialing Code	Area Code	Line Number	Mobile Number	Extension
(+852)-27312878	(+^)^	852			27312878		
(+886)-02-23779435	(+^)^^	886	0	2	23779435		
(+886)-27143828	(+^)^	886			27143828		
(86)-021-2789-4133	(^)^^^	86	0	21	27894133		
(86)-028-91327983	(^)^^	86	0	28		91327983	
(86)-91327983	(^)^	86				91327983	
(+86)-010-13987654321 E 12	(+^)^^E^	86	0	10		13987654321	12
+852-96781923(8)	+^^(^)	852				96781923	8
2178-4321 Ext.5	^^E^				21784321		5
2178-4321(5)	^^(^)				21784321		5
6432 7980	^^				64327980		

Standardize Results

Input

Unique Key	Address
BS000000004	四川省阿坝藏族羌族自治州九寨沟县漳扎镇九安宾馆三零二室
BA000000002	四川省阿坝藏族羌族自治州九寨沟县漳扎镇九安宾馆 302室
BC000000002	四川漳扎镇九安宾馆302室
BC000000003	四川省阿坝藏族羌族自治州
BC000000004	阿坝藏族羌族自治州漳扎镇九安宾馆三零二室
BC000000005	四川九寨沟县
BC000000006	九安宾馆3层02室

Standardized

Unique Key	Province	Province Type	Autonomous City Name	Autonomous City Type	District Name	District Type	Town Name	Town Type	Building Name	Floor Number	Floor Type	Unit Number	Unit Type
BS000000004	四川	省	阿坝藏族羌族	自治州	九寨沟	县	漳扎	镇	九安宾馆			302	室
BA000000002	四川	省	阿坝藏族羌族	自治州	九寨沟	县	漳扎	镇	九安宾馆			302	室
BC000000002	四川						漳扎	镇	九安宾馆			302	室
BC000000003	四川	省	阿坝藏族羌族	自治州									
BC000000004			阿坝藏族羌族	自治州			漳扎	镇	九安宾馆			302	室
BC000000005	四川				九寨沟	县							
BC000000006									九安宾馆	3	层	02	室

Standardize Results

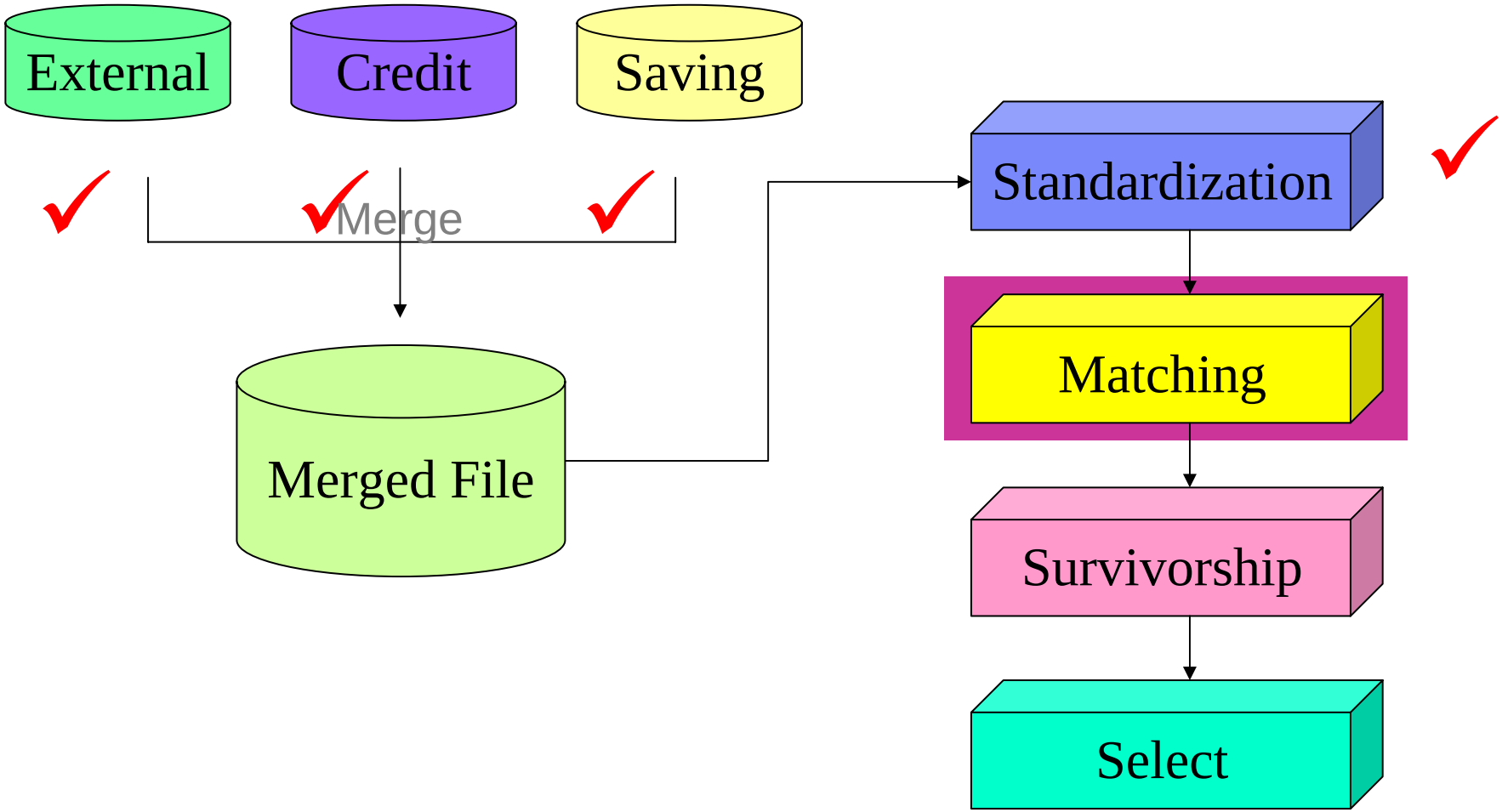
Input

Unique Key	Address
BS000000003	天兴公寓4单元 5楼 8室
BA000000008	北京大兴区双河南里天兴公寓四单元五楼八室
BC000000022	大兴区双河南里天兴公寓5楼

Standardized

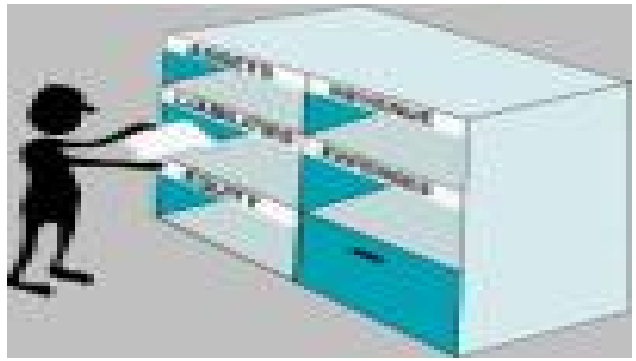
Unique Key	Province	District	District Type	Street Name	Building Name	Block Value	Block Type	Floor Number	Floor Type	Unit Number	Unit Type
BS000000003					天兴公寓	4	单元	5	楼	8	室
BA000000008	北京	大兴	区	双河南里	天兴公寓	4	单元	5	楼	8	室
BC000000022		大兴	区	双河南里	天兴公寓			5	楼	8	室

Processes


External
Credit
Saving
Merge
Merged File
Standardization
Matching
Survivorship
Select

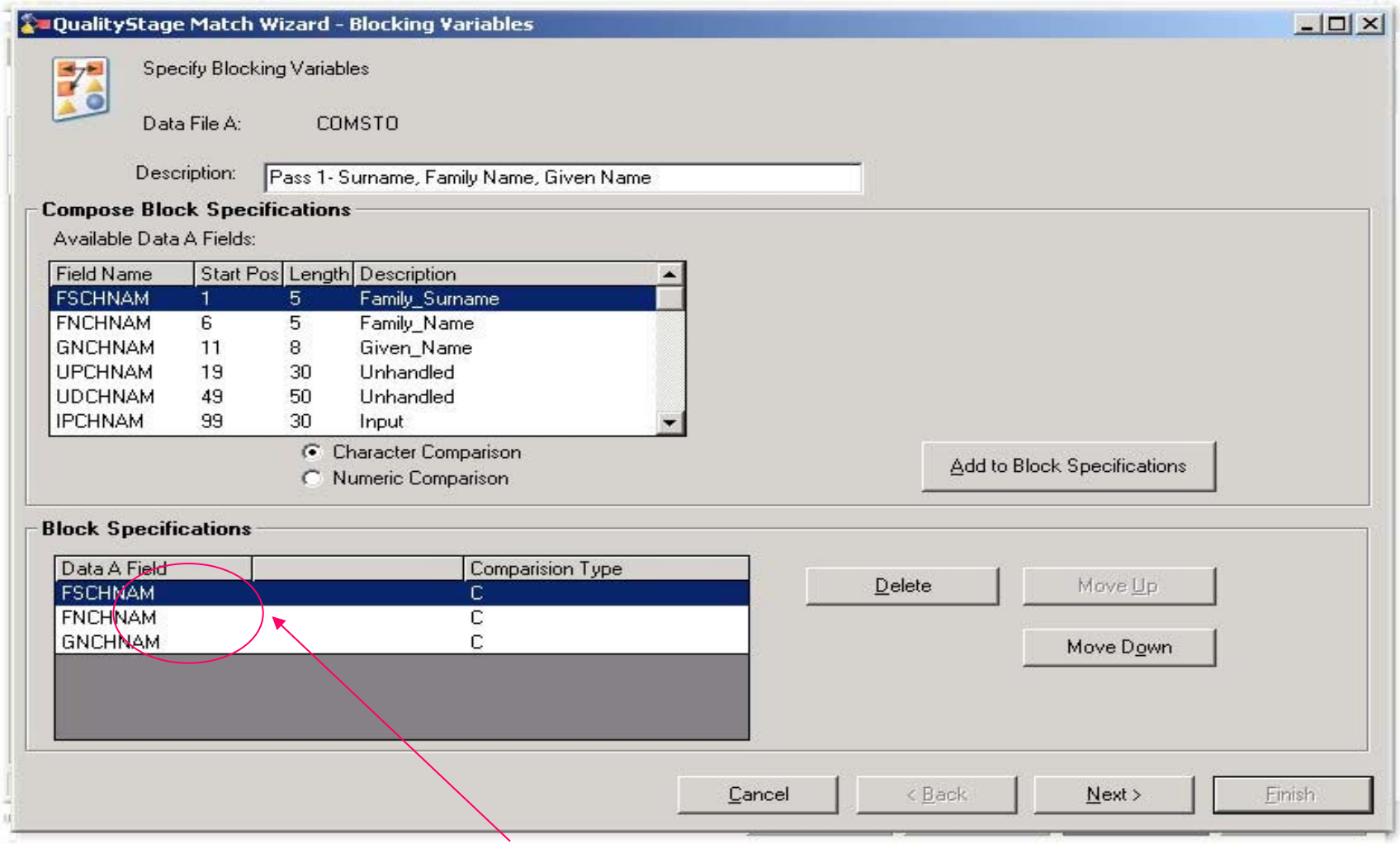
Matching – QS Approach

- Step 1: Groups records with same specific data value into Block
  - e.g. City, Family Name



- Steps 2: Compares the similarity of standardized data fields among Blocked records

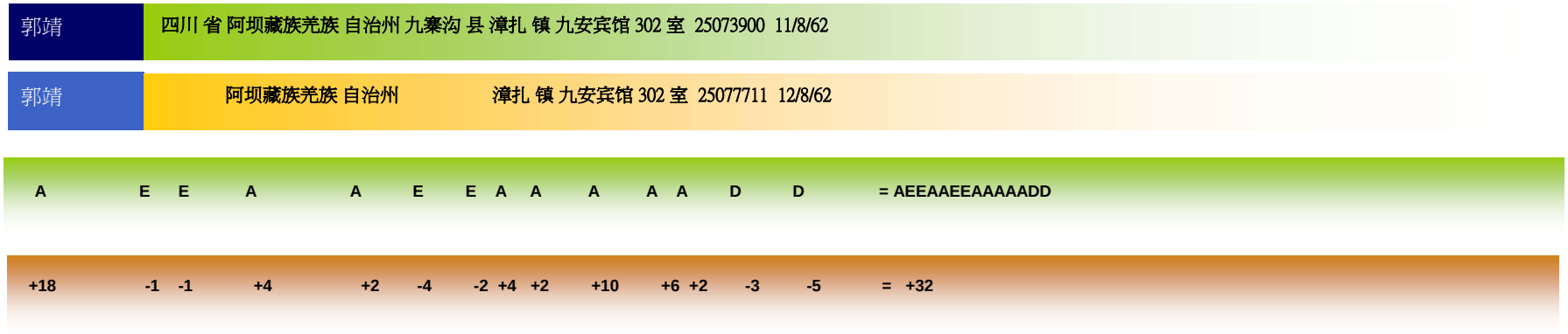
Define Blocks



Define the Grouping Criteria

Super Matching Capabilities

Are these two records a match?

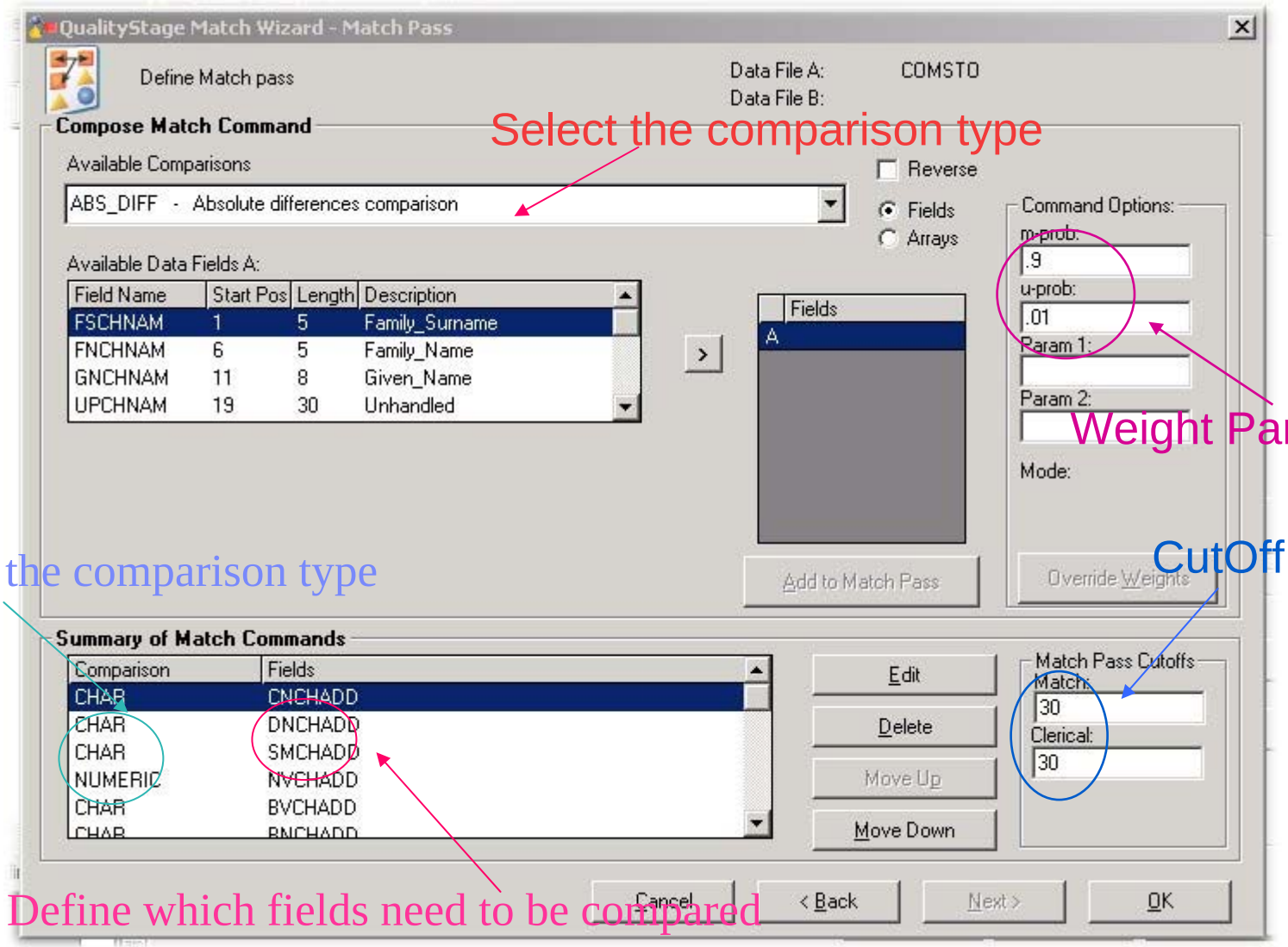


Deterministic Decisions Tables

Probabilistic Linkage

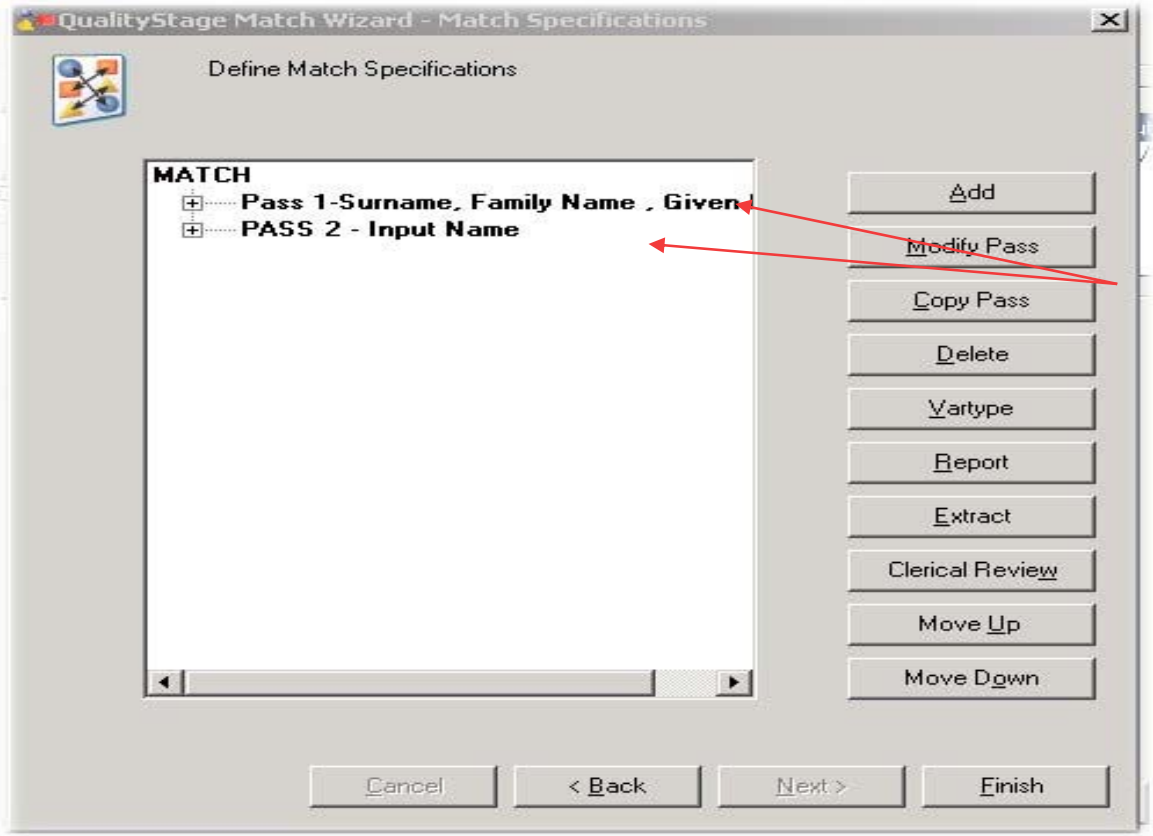
Frequency | Discriminating | Reliability

Define Blocks Records Comparison



Pass

- The settings are stored in a Pass



Multiple Passes

Examples : 3000 Records

3000 Records

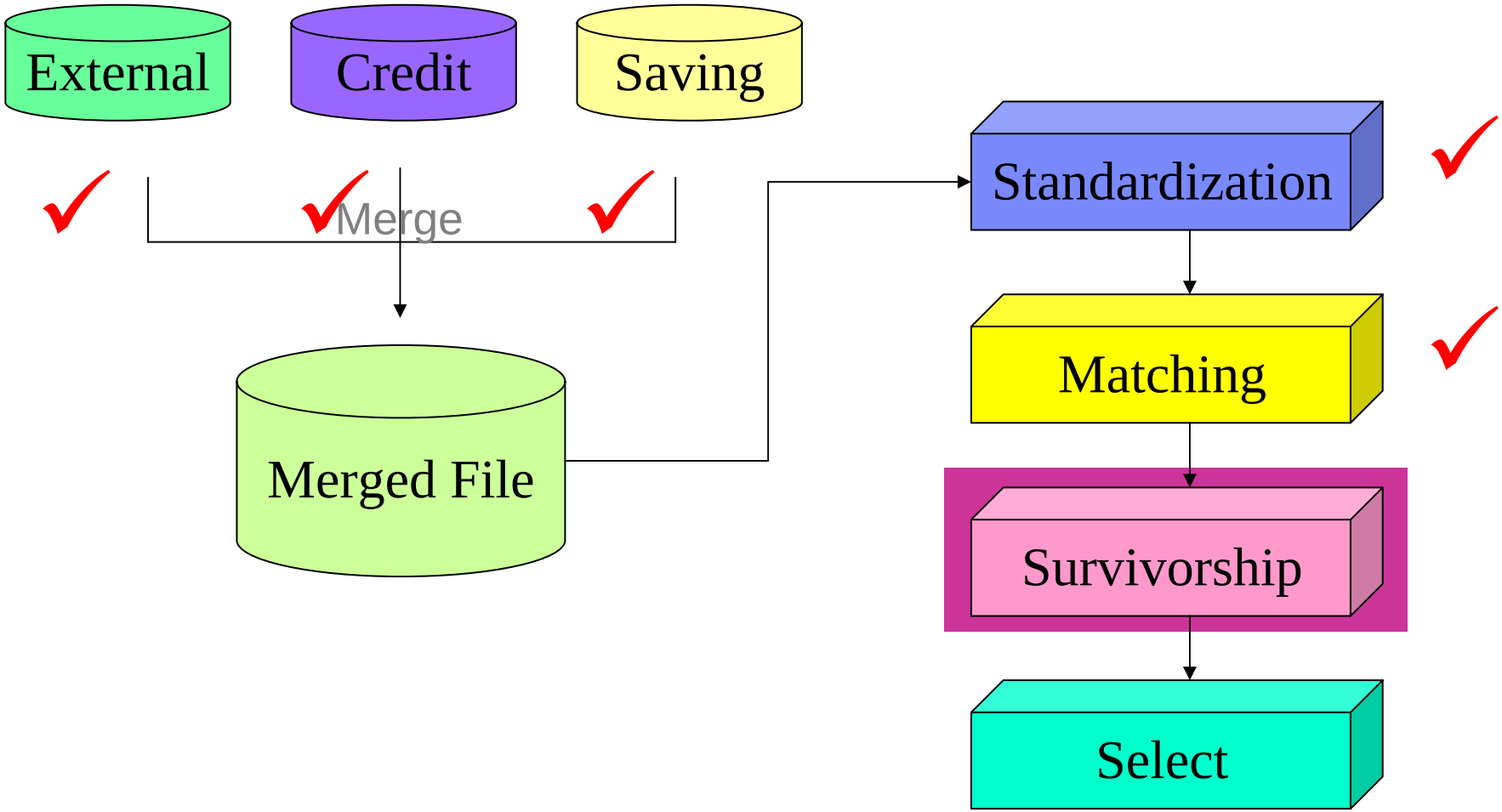
Pass 1 – Name , Address ,Phone (45)

2955 Records

Pass 2 – Phone (900)

2055 Records

Flow Diagram



Survivorship - Introduction

- Consolidate the records within a block to give a best records
- End user can set the selection criteria for each data field
- The most common criteria are
  - Longest (length), shortest(length) and Frequency

    Examples : Address (Longest)

    : Gender (Frequency)

Field Survivorship Examples

- Most Frequent

  Name , Address, Date of Birth, Gender, Martial Status and Age

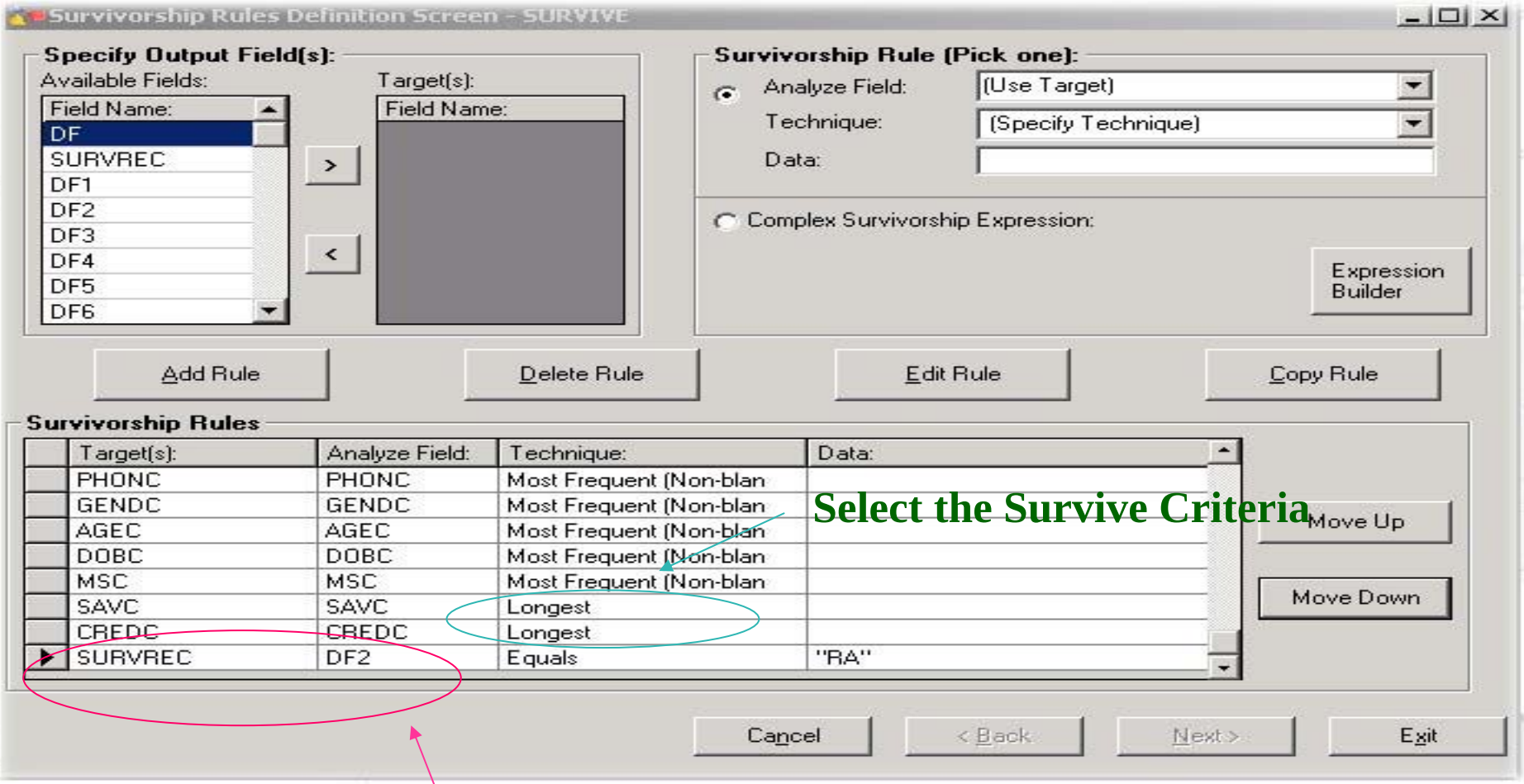
- Longest

  Saving Account and Credit Card Account

Block #	Mark	Family Name	Given Name	Gender	Age	DOB	Married
2	86.5	郭	靖	M	45	24/6/1959	N
2	66.92	郭	靖	M	45	24/6/1959	
2	73.02	郭	靖	M	45		Y
2	41.88	郭	靖		45		Y
2	59.57	郭	靖	M			
2	33.84	郭	靖		45		
2	32.41	郭	靖	M	45		Y

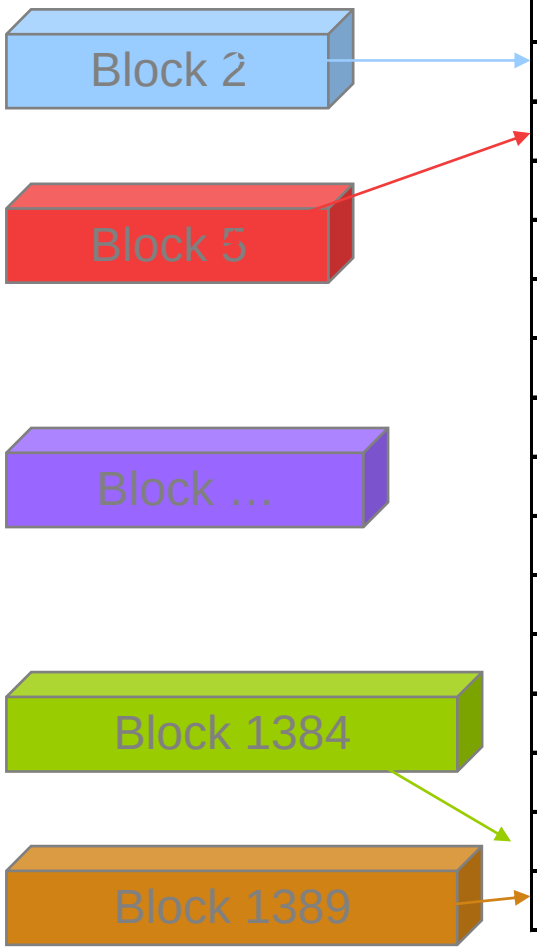
Block #	Mark	Family Name	Given Name	Gender	Age	DOB	Married
2	86.5	郭	靖	M	45	24/6/1959	Y

Select Survive Fields and Criteria



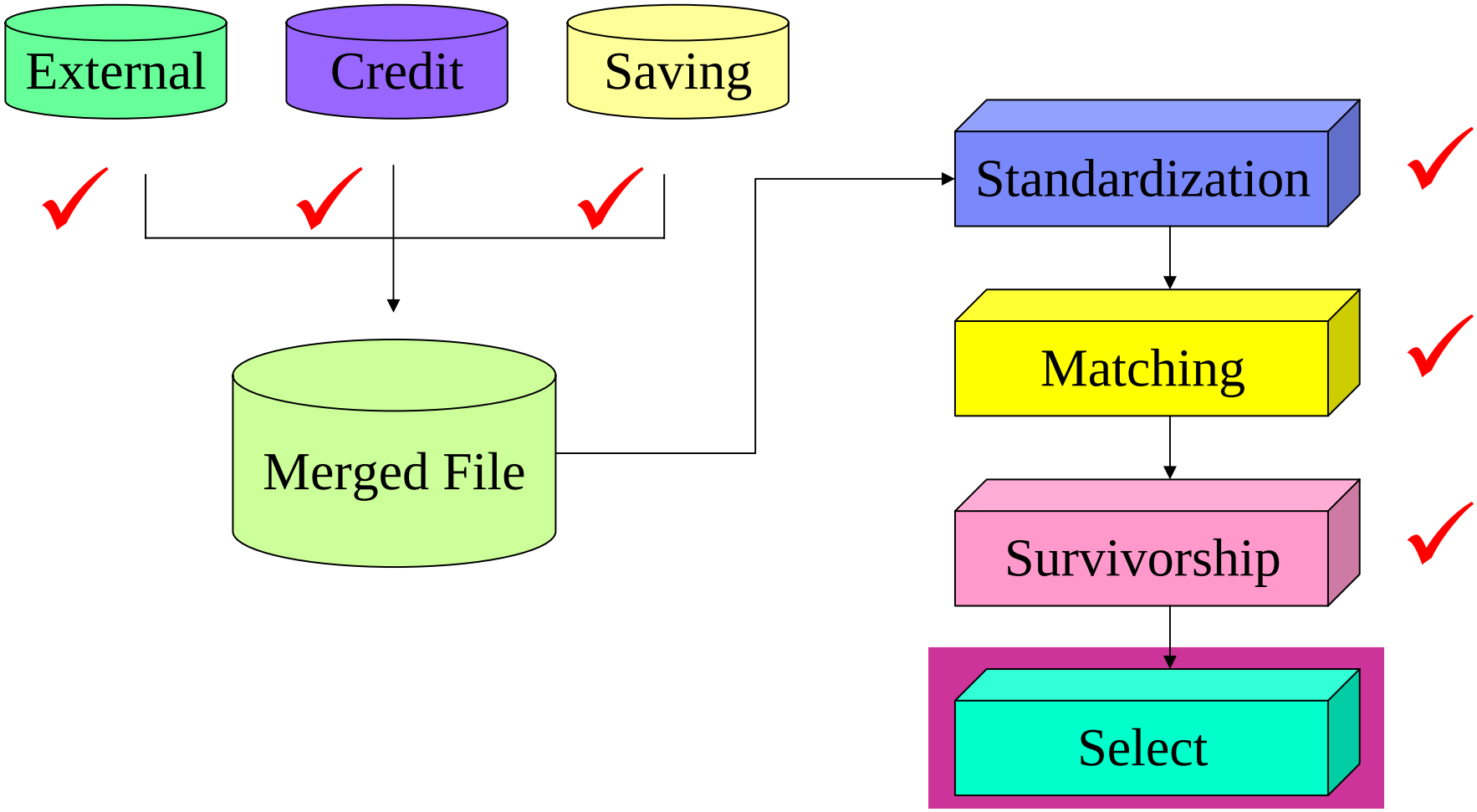
Credit Card and Saving Number is selected to Survive

Survivorship Results – Unique View



Block Number	Mark	Family Name	Given Name
2	86.5	郭	靖
5	68.22	张	之洞
8	86.1	黄	蓉
9	93.54	张	菁
14	95.9	林	平之
16	72.05	花	木兰
17	69.13	汤	汉斯
18	63.77	克	林顿
102	51.56	陈	健文
739	112.23	陈方	安生
740	95.5	方	安生
745	108.42	萧	十一郎
1379	72.21	司马	懿
1384	88.2	陈	平
1389	52.33	苏	小小

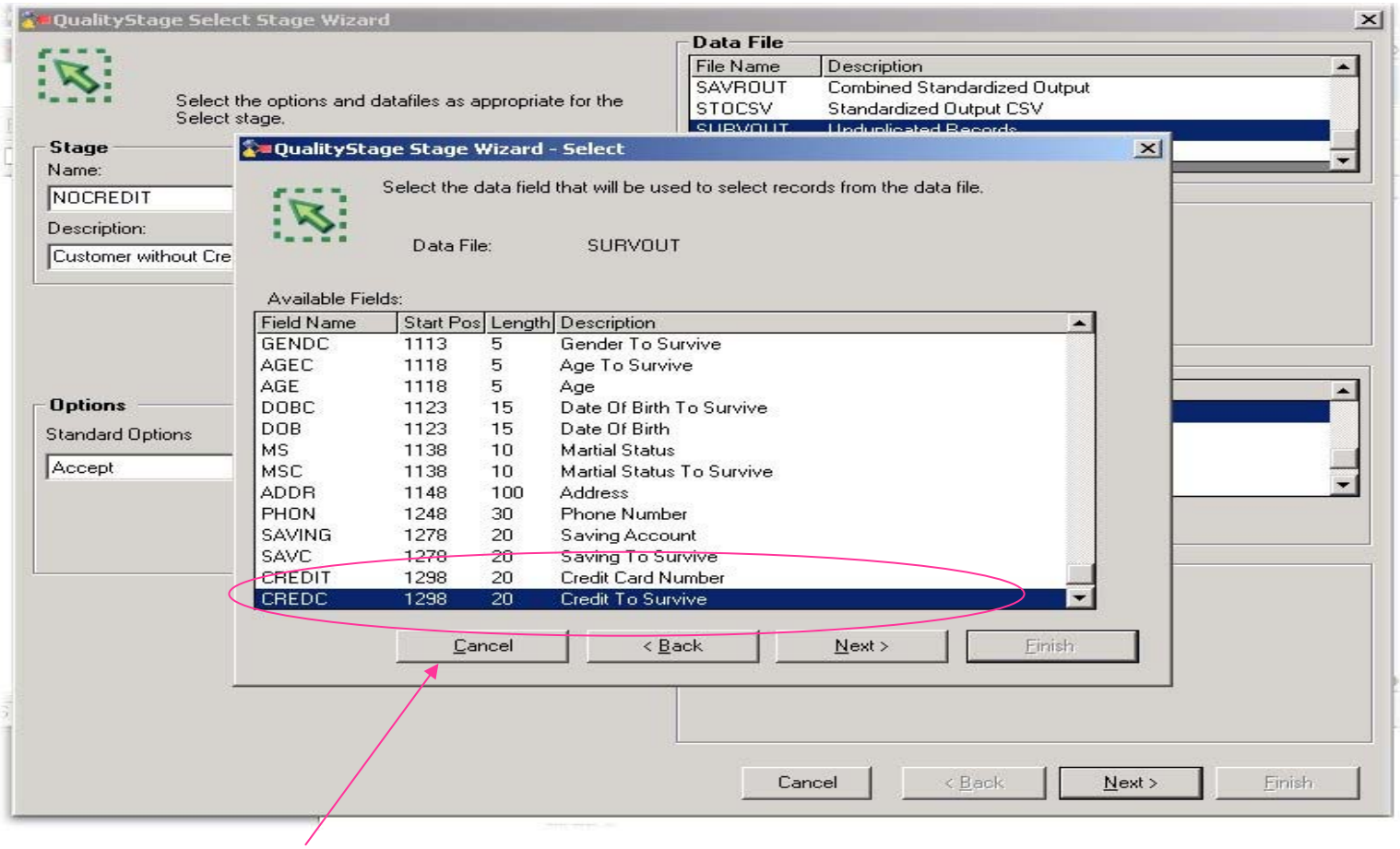
Flow Diagram



Selection - Introduction

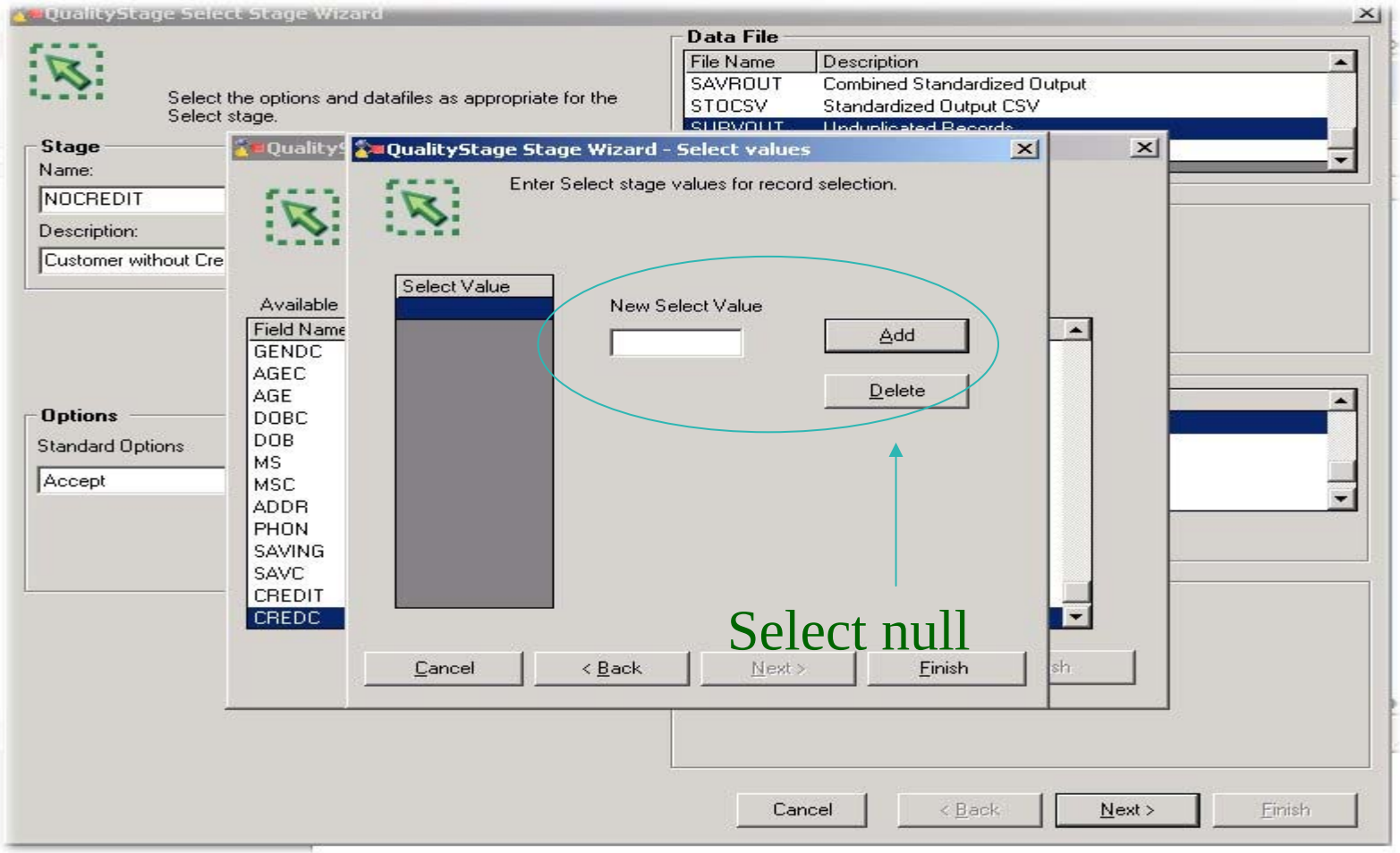
- We have a Customer database without duplicate records and best of breed information
- Then we can have a target customer list for customer without a credit card
- How ? By simple select those customer records without credit card information

Select Credit Card as Criteria



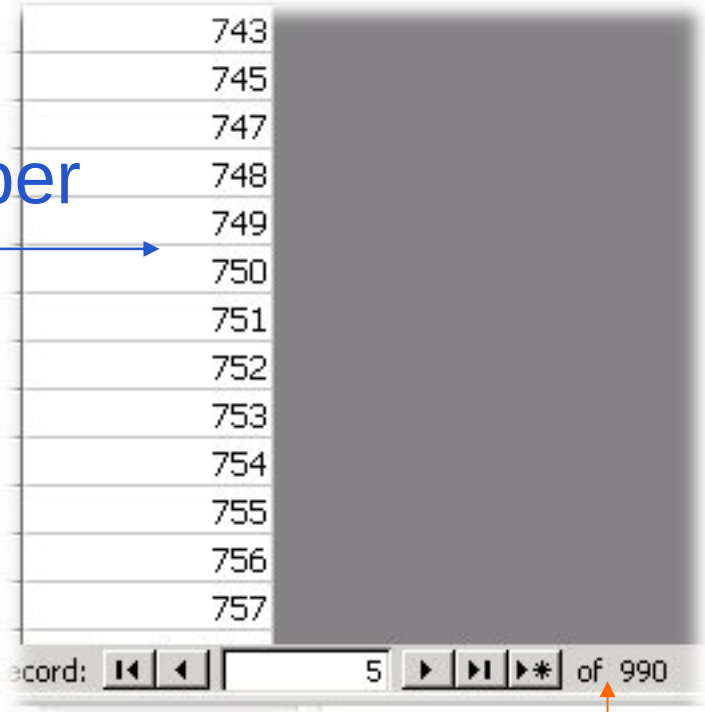
Select Credit Card as Select Criteria Field

Select Null Value for Credit Card



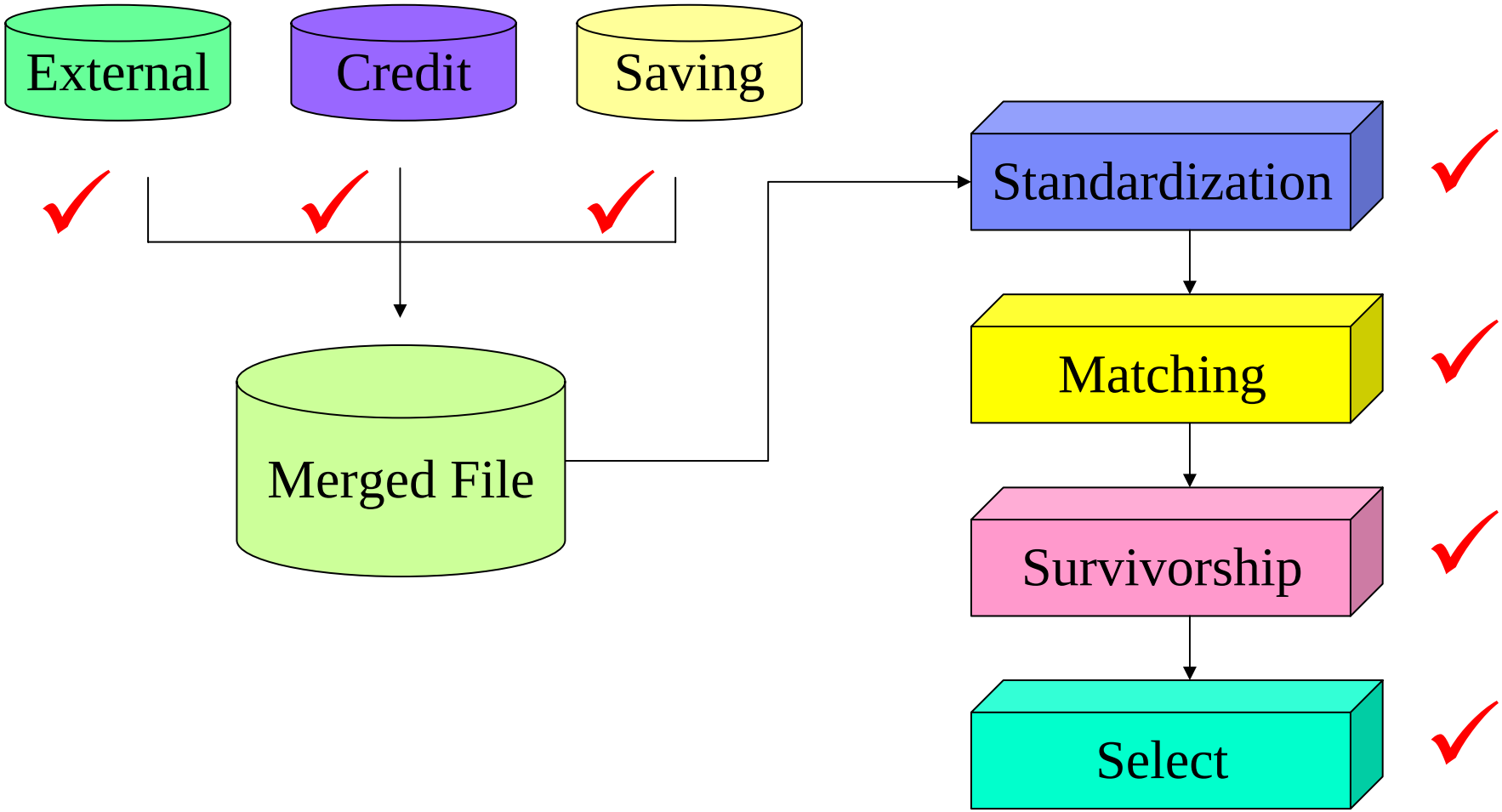
Results

Block Number



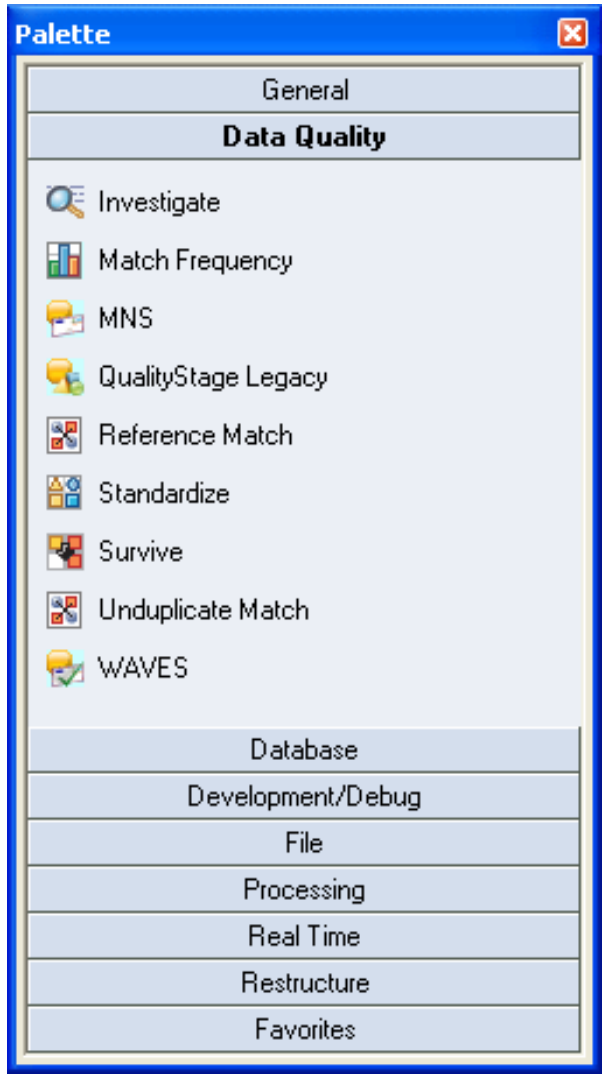
Only 990 HSBD Saving customers are found not using any Credit Card Service

Flow Diagram

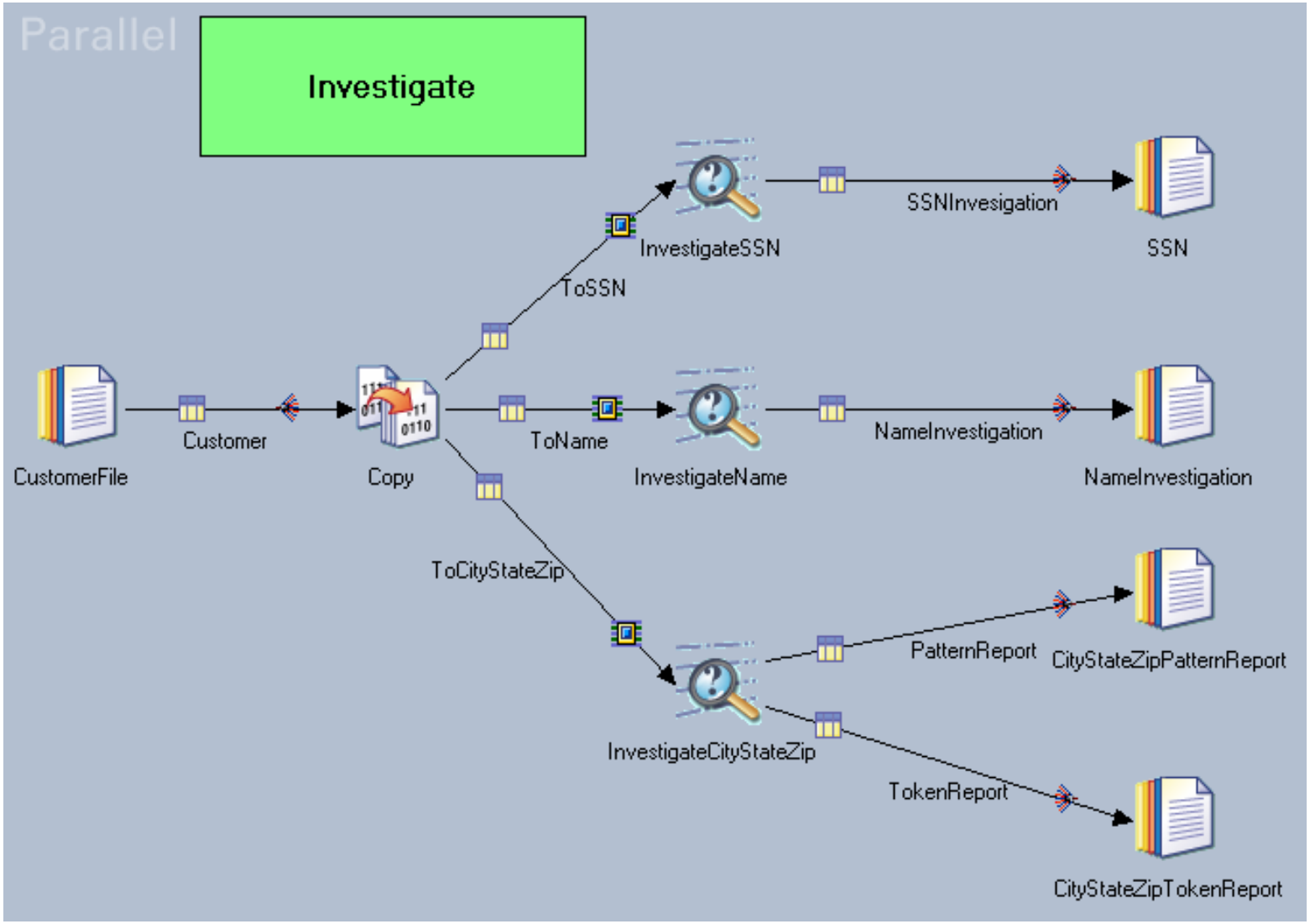


Single Design Environment

- All phases of data quality:
  - Investigate
  - Standardize
    - Domain and Multi-National
  - Match
    - Unduplicate
    - Reference
  - Survive
  - WAVES
  - Legacy (pre 8.0) support

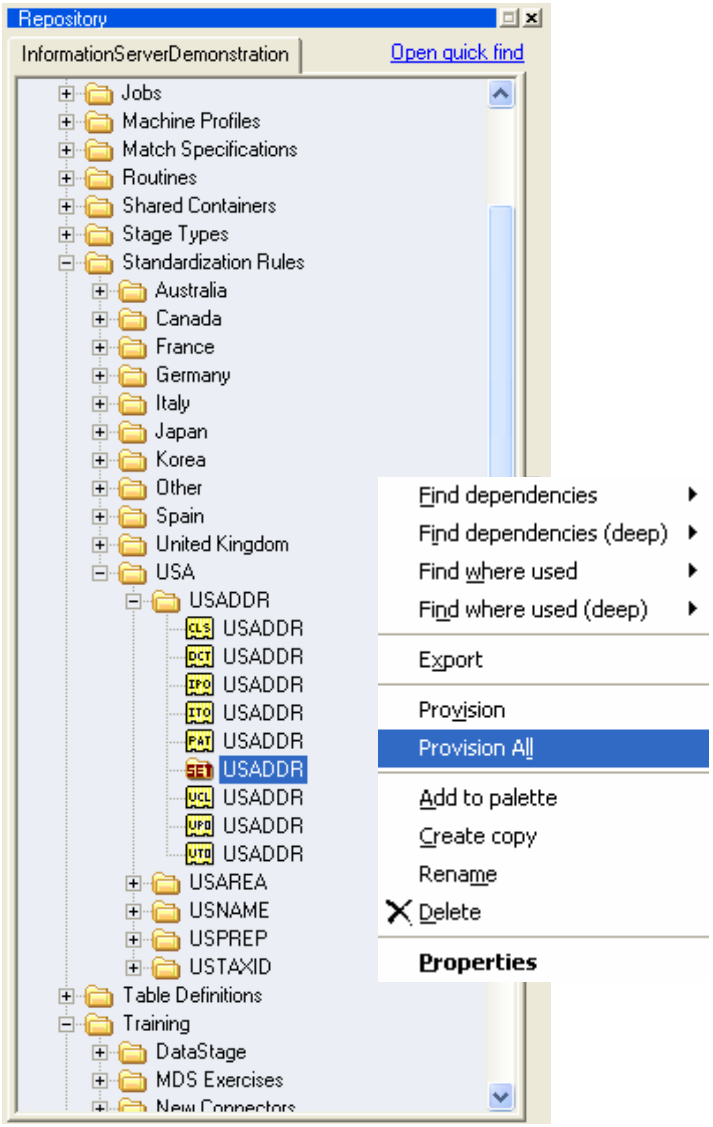


Investigations

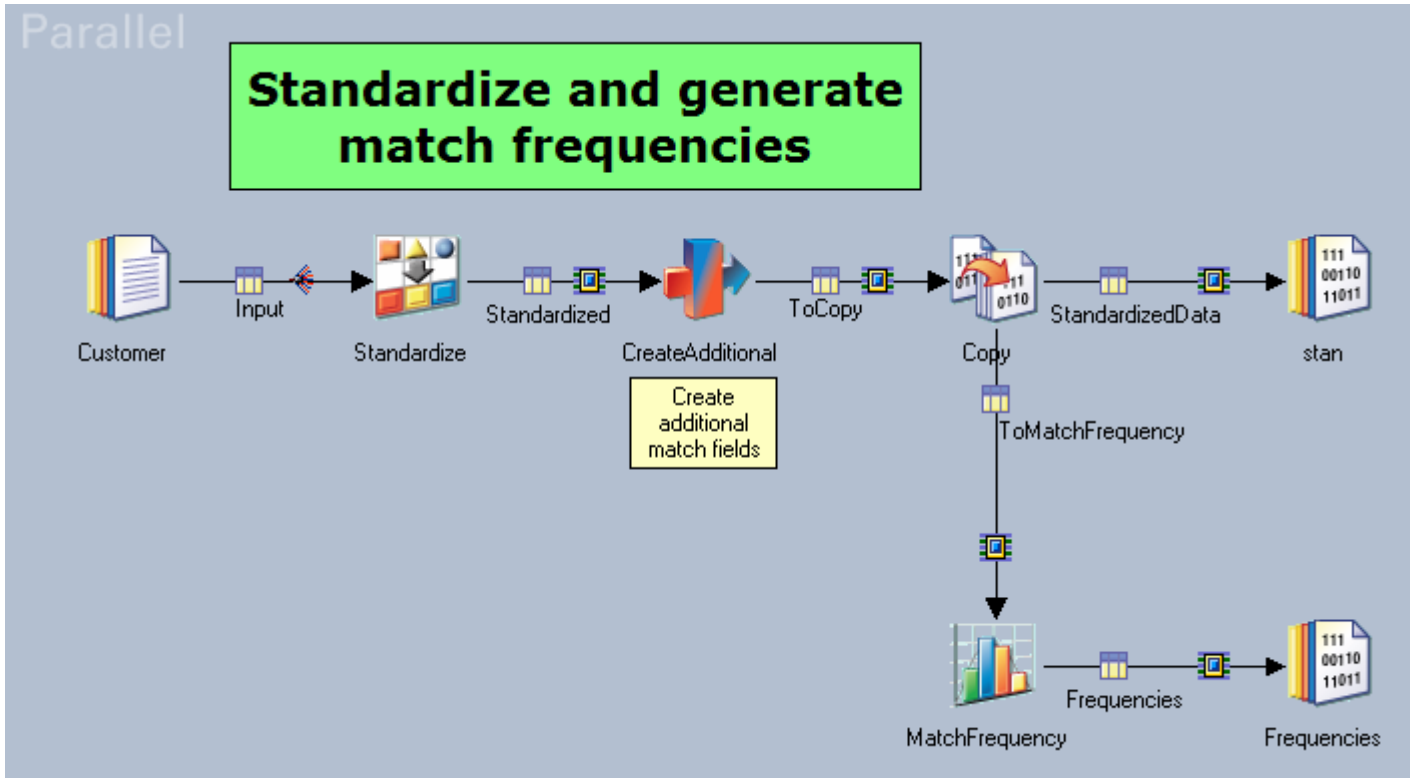


Provision Rules to be used

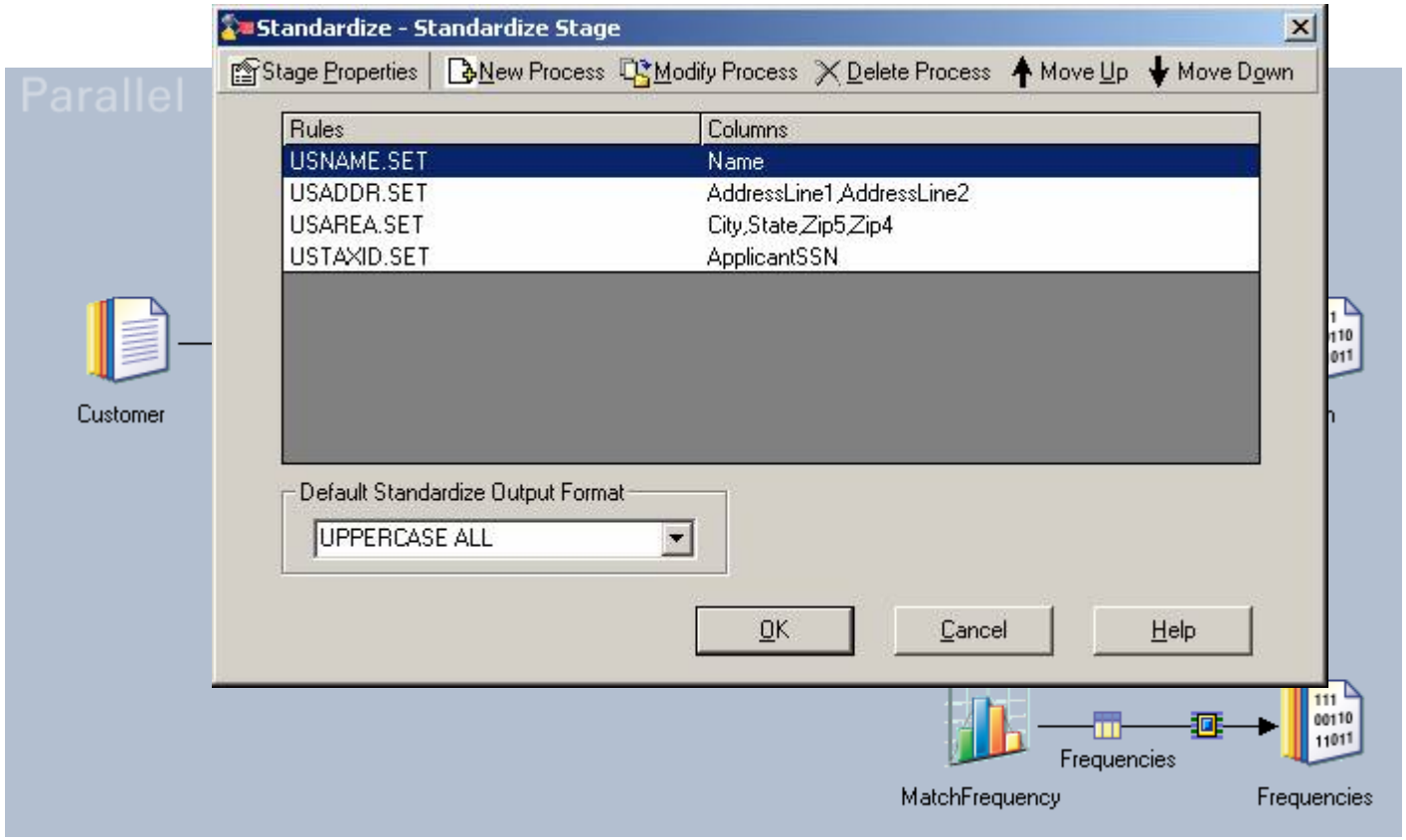
- Provisioning copies rules from repository to execution area
- Use ‘Provision all’



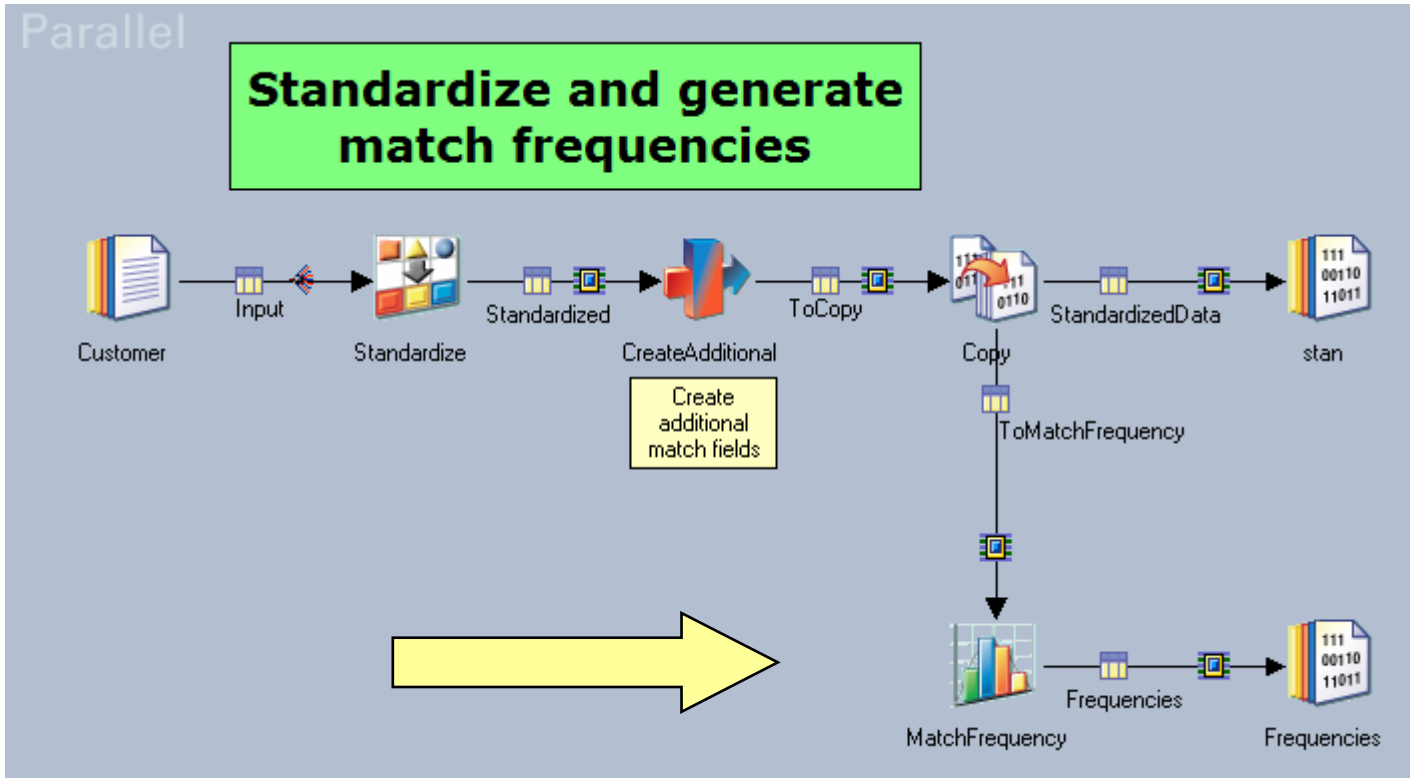
Standardization



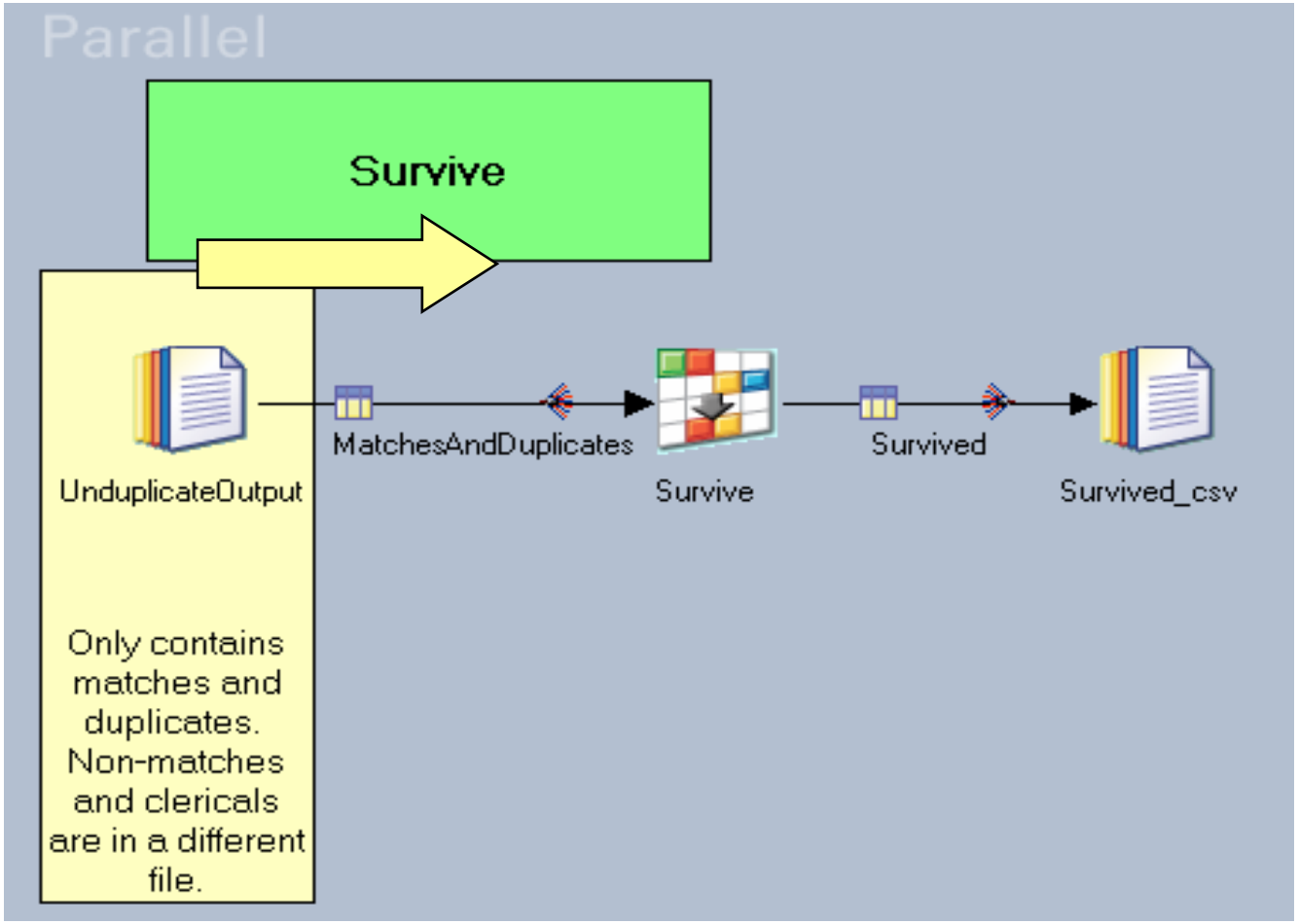
Standardization



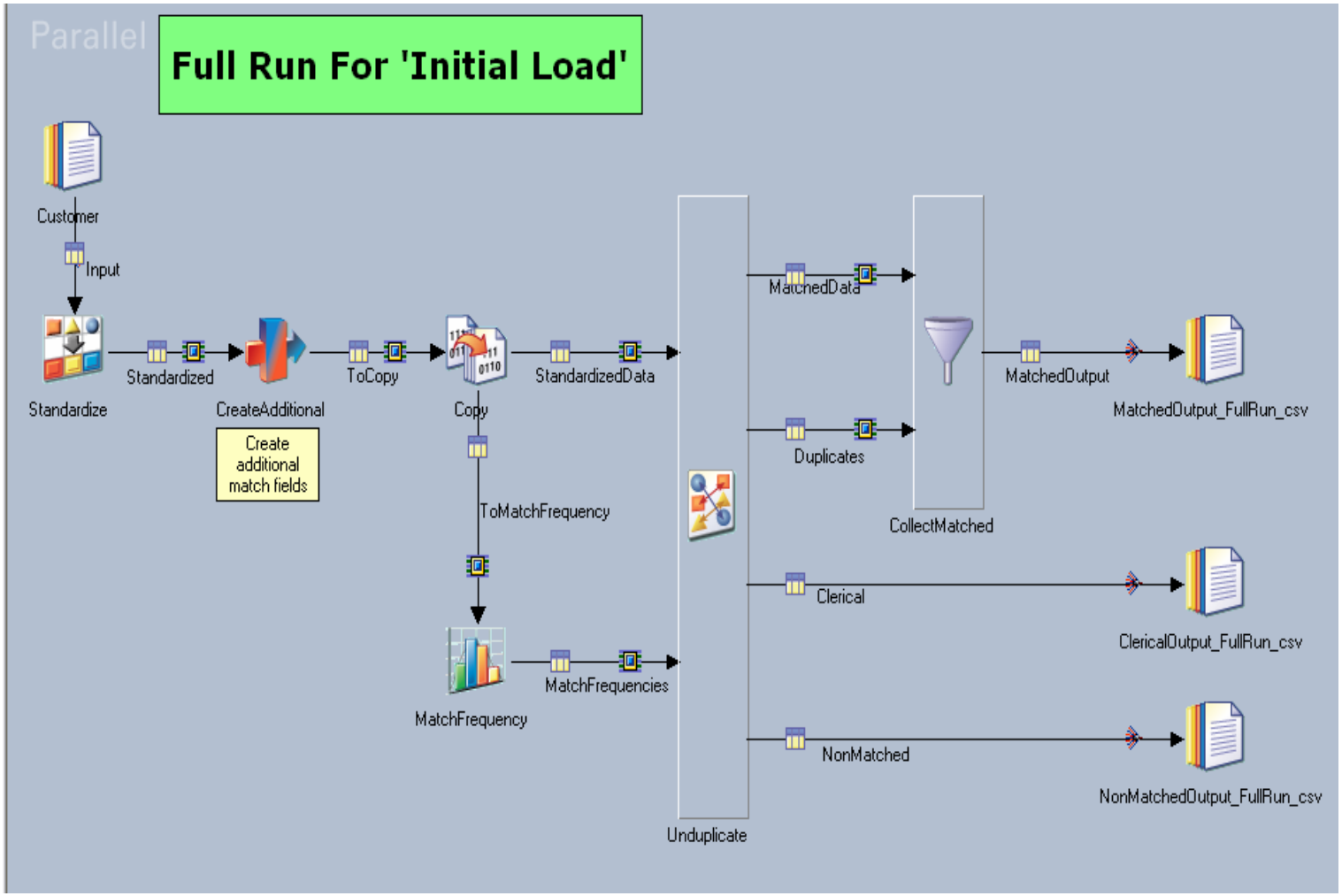
Match Frequency Generation



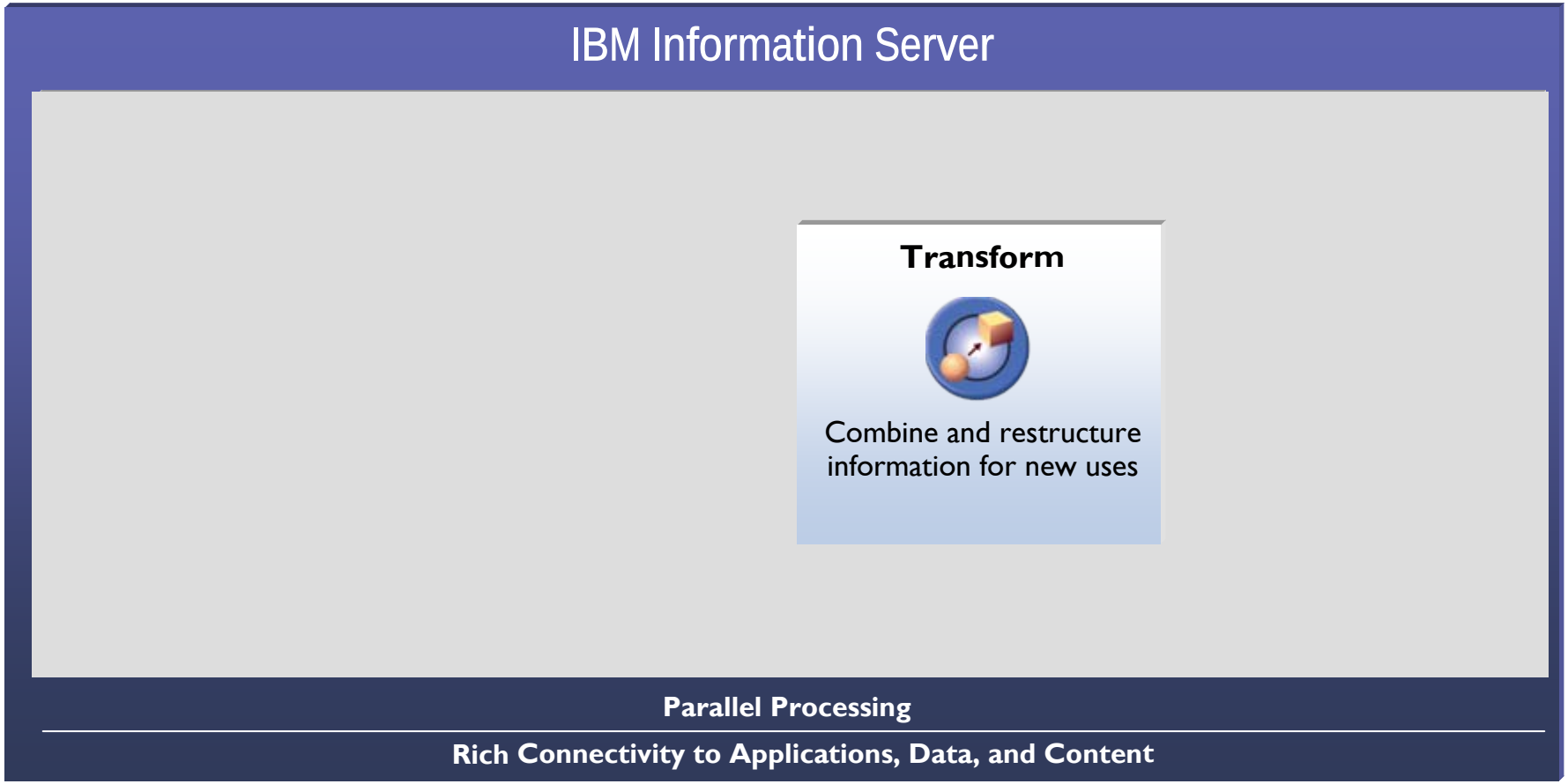
Survive


Parallel
Survive
UnduplicateOutput
MatchesAndDuplicates
Survive
Survived
Survived_csv
Only contains matches and duplicates. Non-matches and clericals are in a different file.

Full run in single design canvas

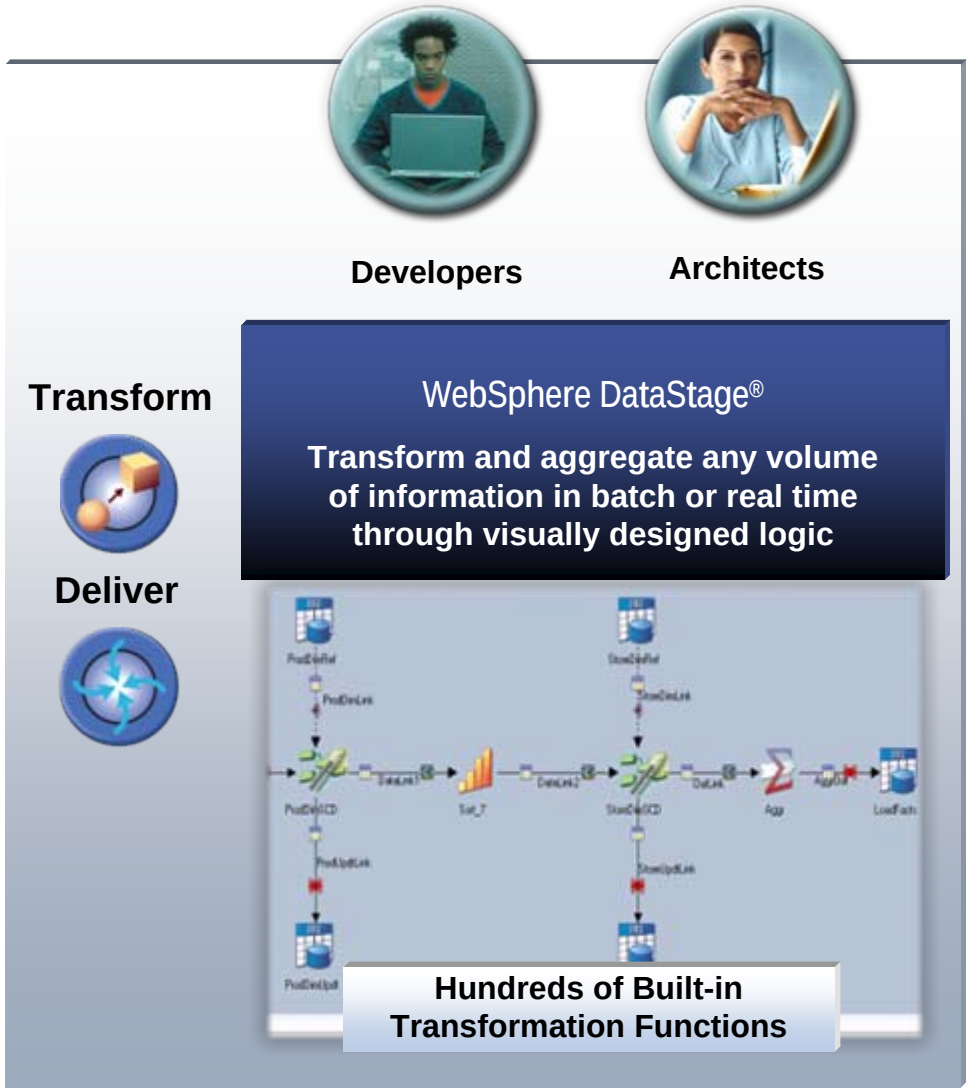


IBM Information Server
Delivering information you can trust

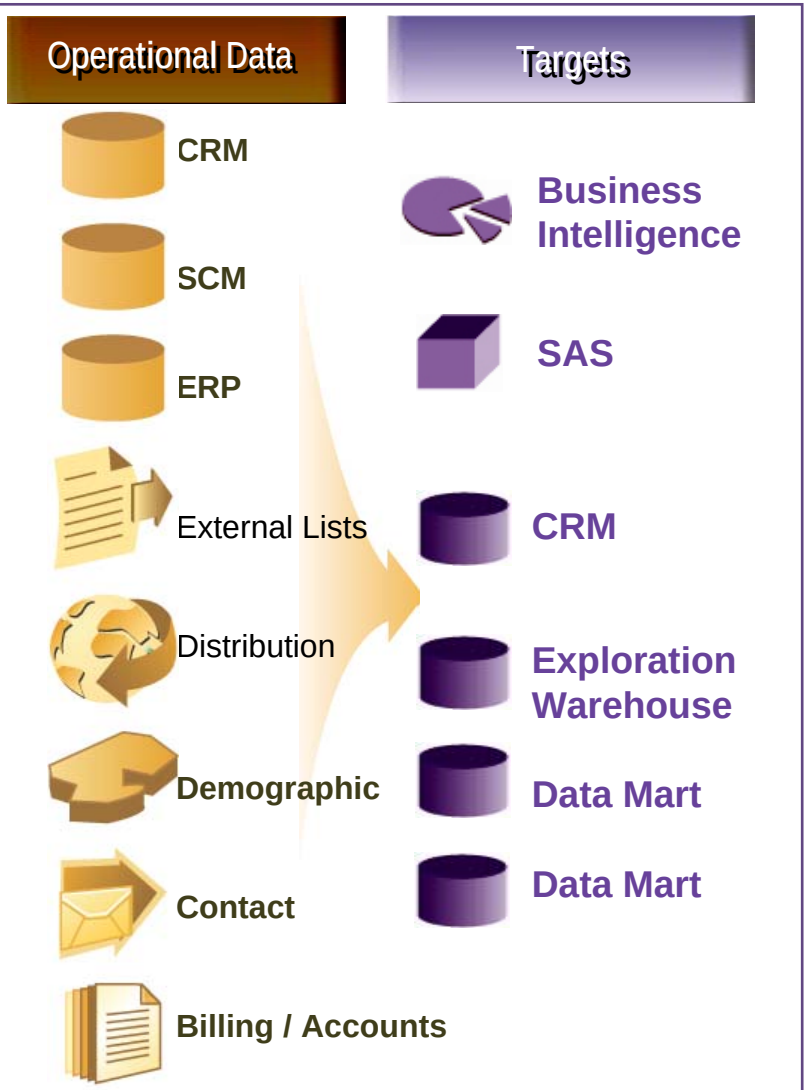


Data Transformation & Movement: WebSphere DataStage

- Provides codeless visual design of data flows with hundreds of built-in transformation functions
  - Speeds project delivery and reduces costs
- Complete ETL functionality with metadata-driven productivity
  - Deals with very large volumes of data
- Supports batch & real-time operations
  - Provides versatility to deal with many project requirements
- Provides integration from across the broadest range of sources

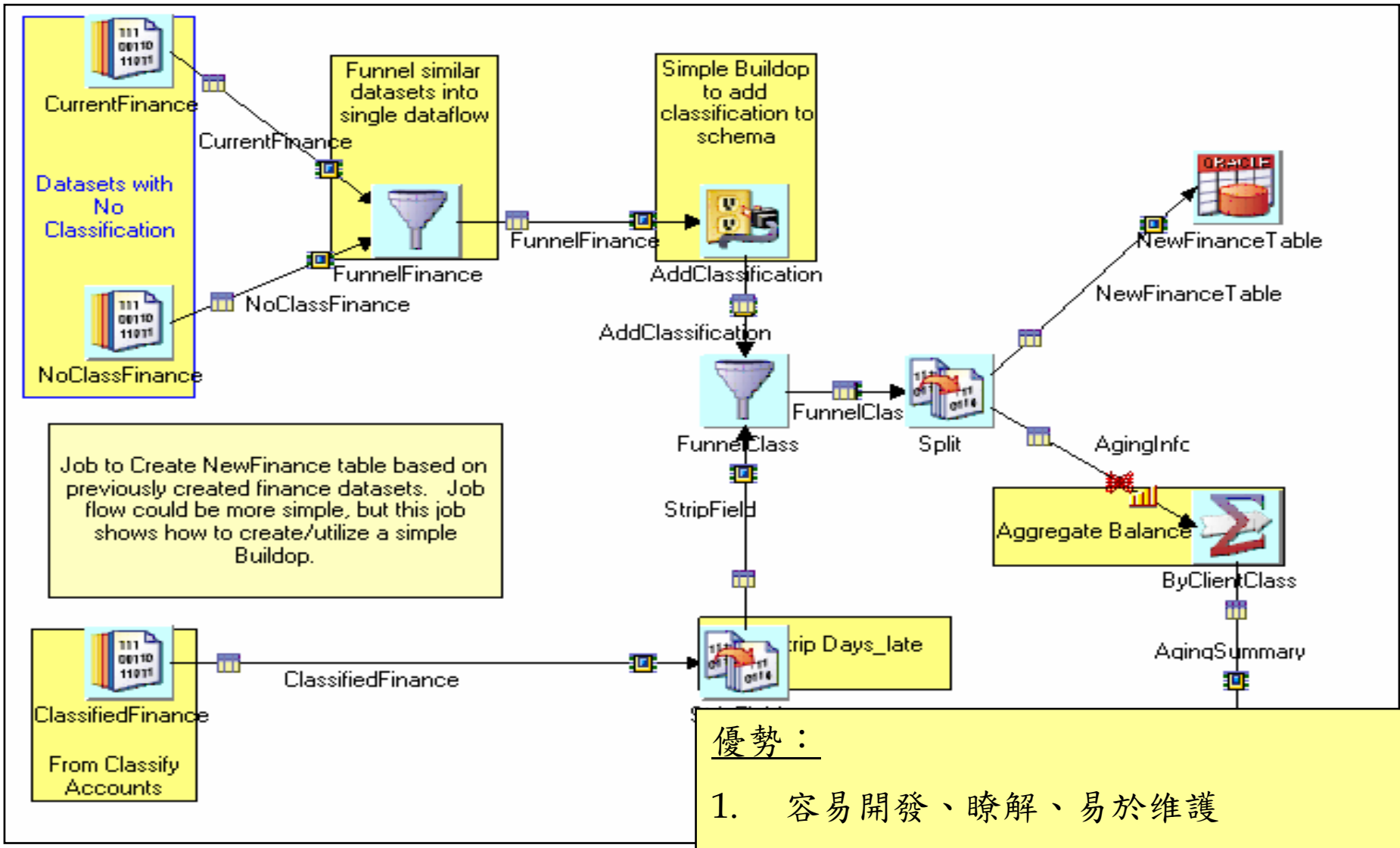


IBM WebSphere DataStage



- Processes and transforms large amounts of data in real-time or bulk mode
- Handles all transformations from simple to complex
- Manages multiple integration processes
- Integrates data from the widest range of enterprise and external data sources
- Provides direct connectivity to enterprise data applications as source or targets
- Leverages meta data for cross tool impact analysis and easy maintenance

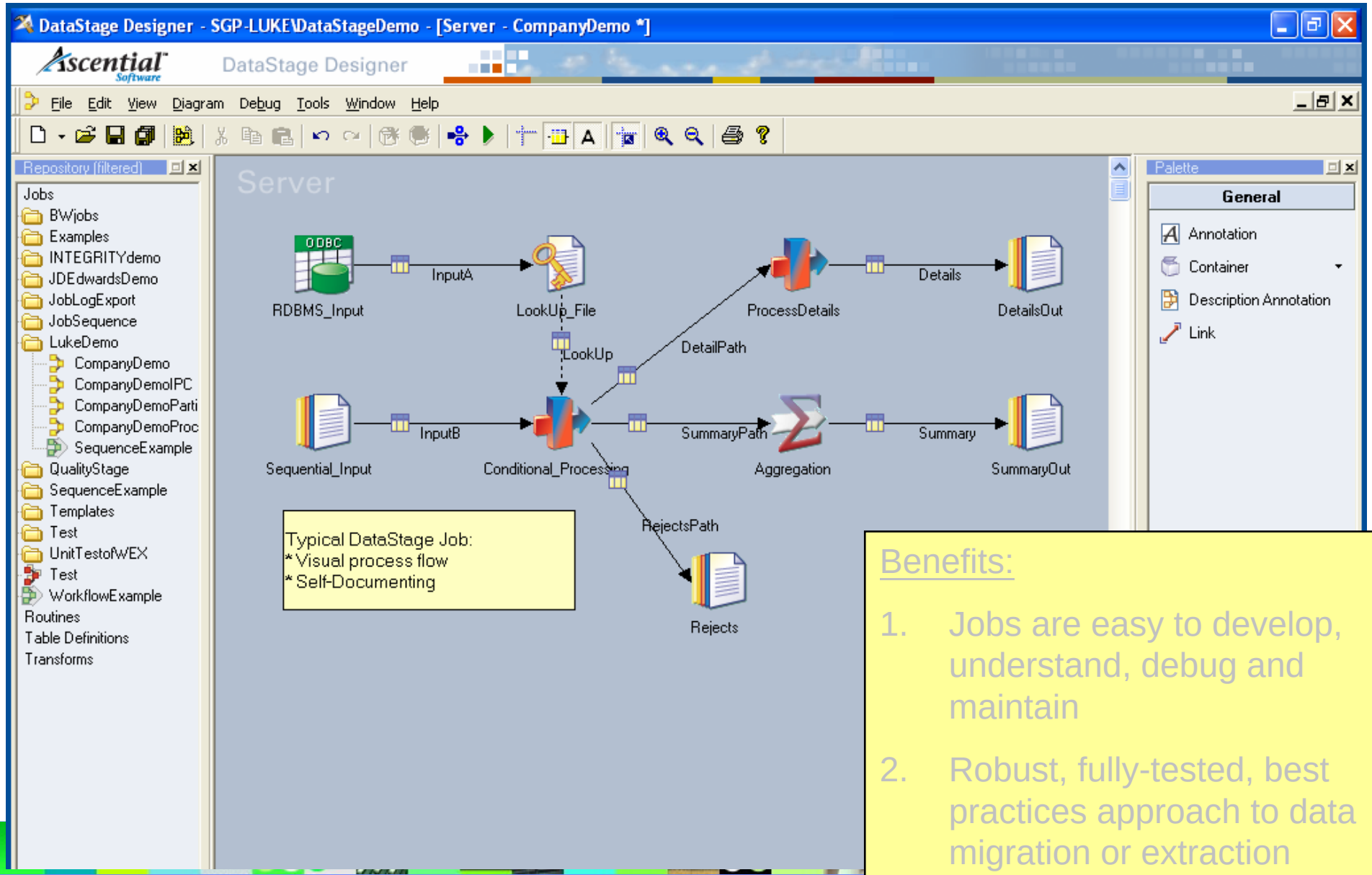
IBM DataStage提供全面的圖像設計及管理介面並內建眾多功能



優勢：

1. 容易開發、瞭解、易於维護
2. Gartner Research評選最好的ETL工具

Top-Down 圖形化的設計 – 清楚而且容易了解



Benefits:

1. Jobs are easy to develop, understand, debug and maintain
2. Robust, fully-tested, best practices approach to data migration or extraction

IBM DataStage 是一個強大的企業系統整合工具，具有如下的特點:

- 開放的、全面向的服務架構（Service Oriented Architecture, SOA）
- 正確的掌握資料特徵和資料品質
- 完整的內建資料轉換功能和流程安排
- 可重複使用的設計元件和轉換規則
- 執行效能可依軟硬體的成長而線性增長，無需更改設計
- 全面性的端到端元數據(Metadata)管理
- 完整的資料連接(Sequential, Hierarchical, Relational, Legacy, Email, Named Pipes, FTP, XML, Message Queues, Web Service, Java)
- 完全符合商業標準(XML, EDI, JMS, EJB, SOAP, JCA)
- 完整的企業整合解決方案

手工編寫系統的問題：

- 冗餘的業務規則和“多版本事實”
- 無法得到資料的統一視圖
- 不完善資訊流
- 不準確、不完整、不一致的資料
- 不能有效因應不斷快速增長的資料量
- 瑣碎、複雜、固定代碼架構
- 元數據的管理不易實現
- 很難分享已存在的程式碼，重覆設計相同功能的程式碼，既費時又費力

DataStage 企業版替代手工編寫：

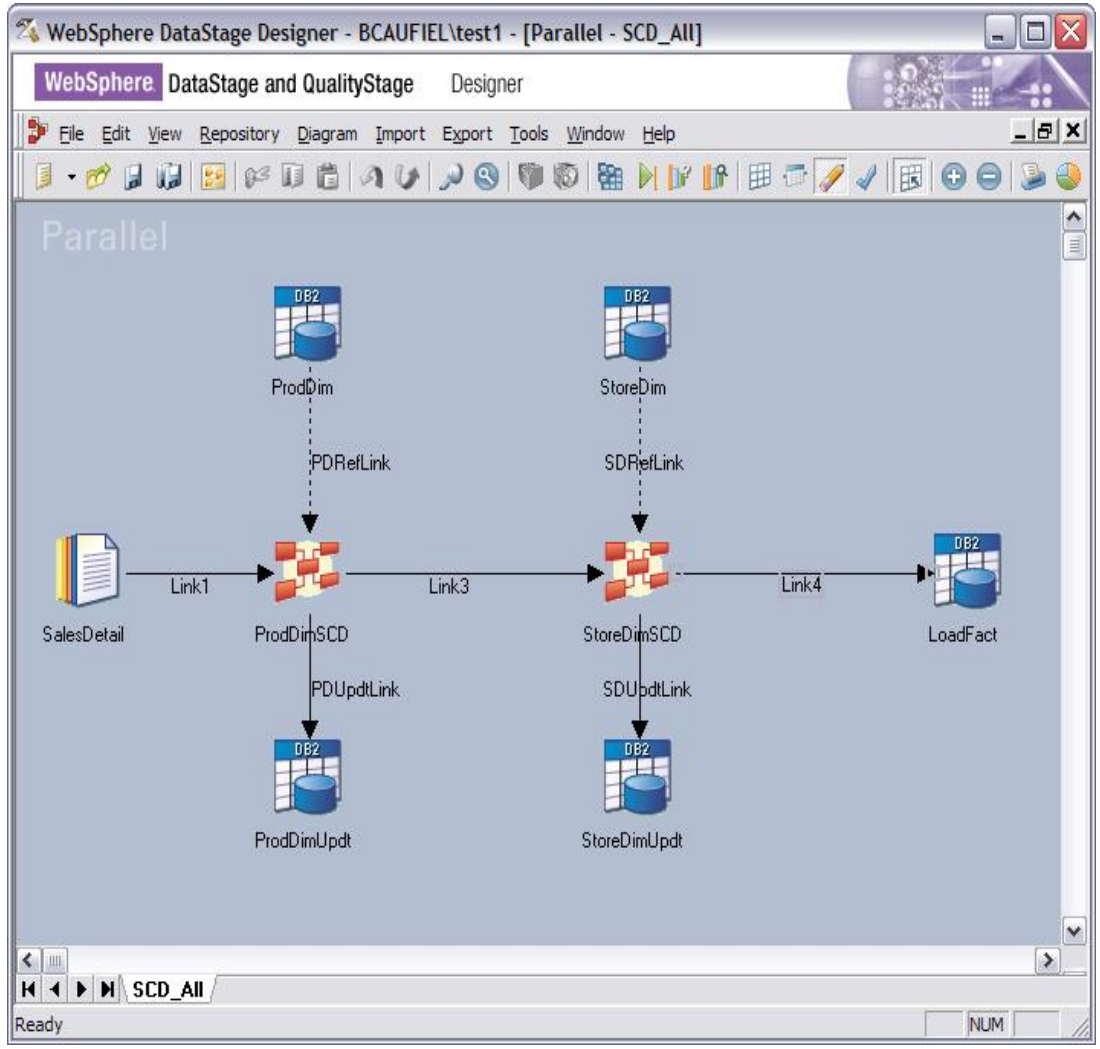
- 跨平臺，跨系統的運行能力
- 具有單一介面的設計和執行模式
- 統一的資料視圖
- 掌控完整的資料資訊流
- 呈現正確、完整、一致的資料
- 可及時反應資料容量的快速增長
- 簡單的元數據管理

使用DataStage 企業版的好處：

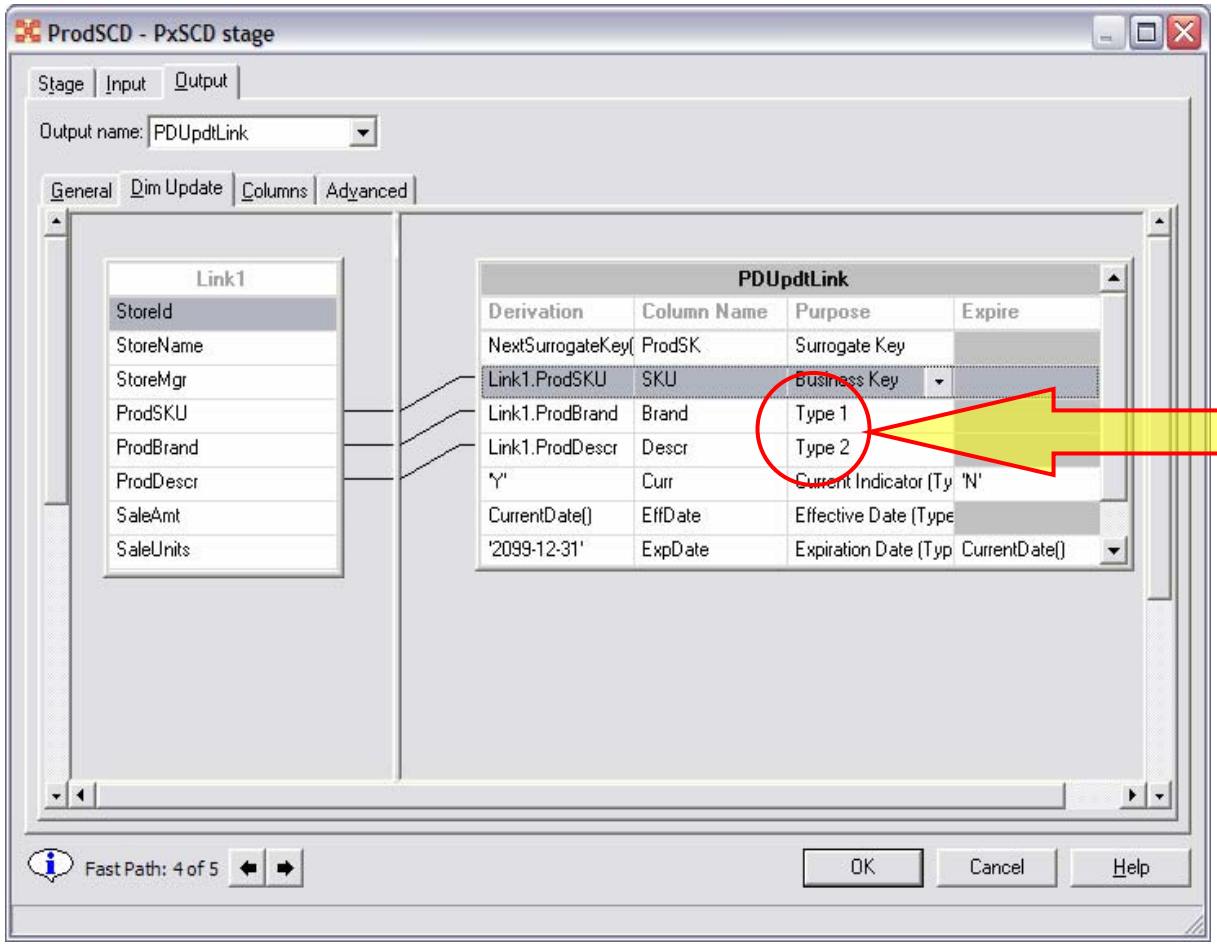
- 更快的設計、開發、上線，縮短專案時程
- 節省大量的開發和維護成本
- 具有最小的風險
- 可提高企業的生產力

Slowly Changed Dimension Enhancement

- New engine capabilities
  - Surrogate Key management
  - Updatable in-memory lookups
- New & enhanced stages
  - Surrogate Key Generator
  - Slowly Changing Dimension



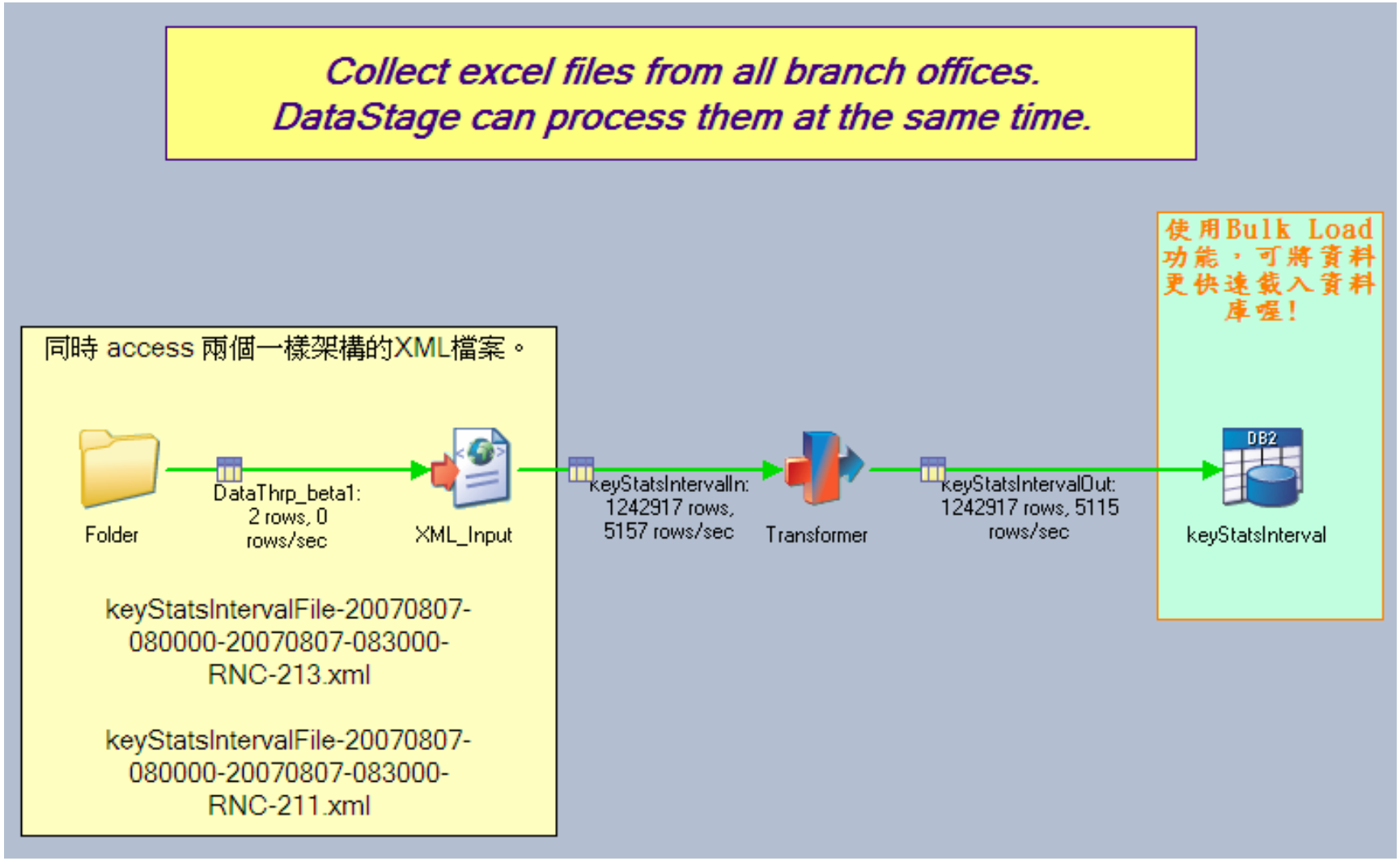
SCD: Updating the Dimension Table



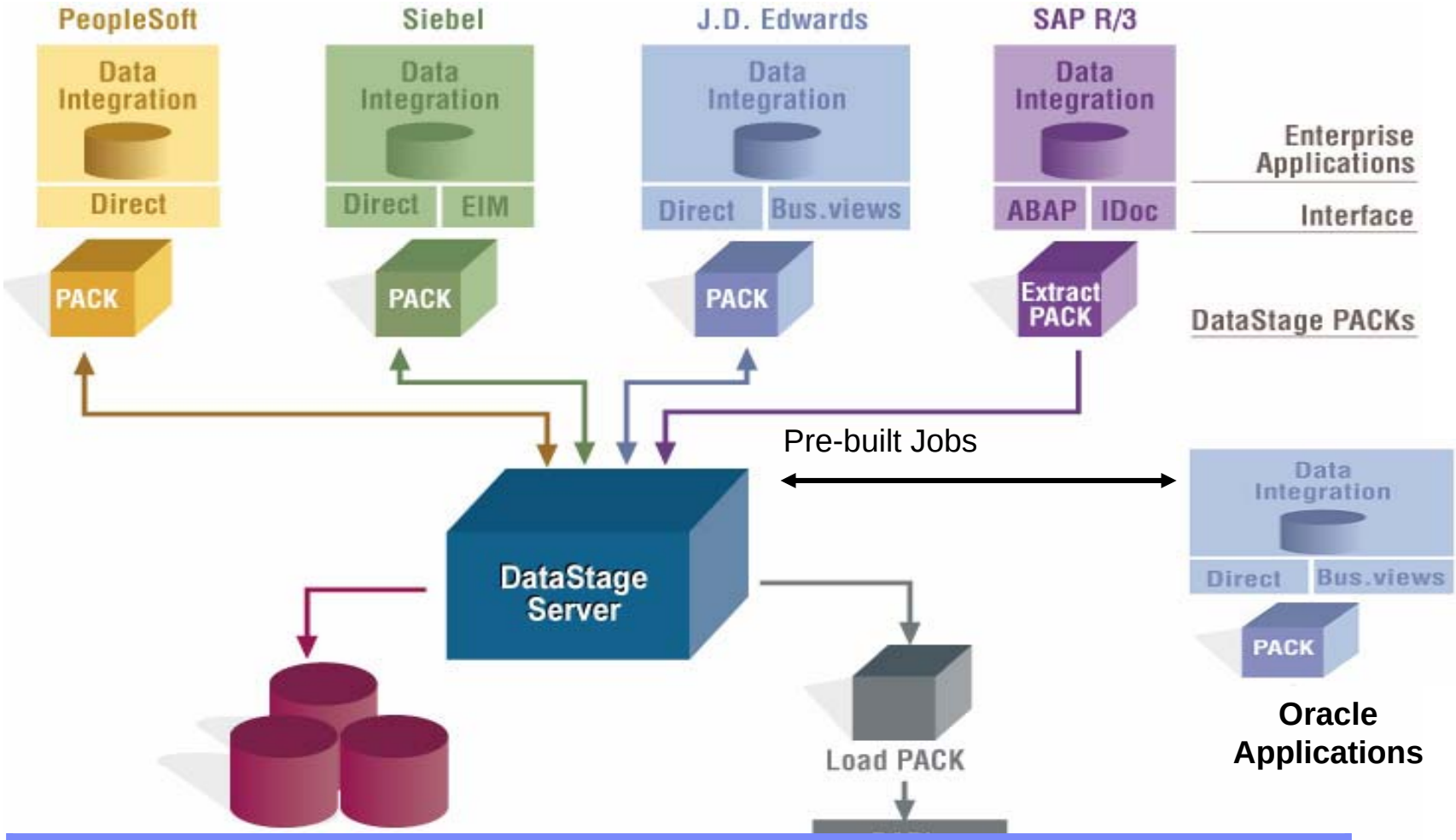
The **Derivation** column:

- Indicates how to detect that a **dimension row has changed**.
  - The action taken depends on the purpose code of the changed column. If it's a Type1, then the row will need to be updated. If it is a Type2 the row needs to be expired and a new one created.
  - Type2 changes are search for first.
- Indicates how **updated and new row values are to be computed**

Process your data from all branch offices at the same time



連接企業應用系統的能力



Benefits:

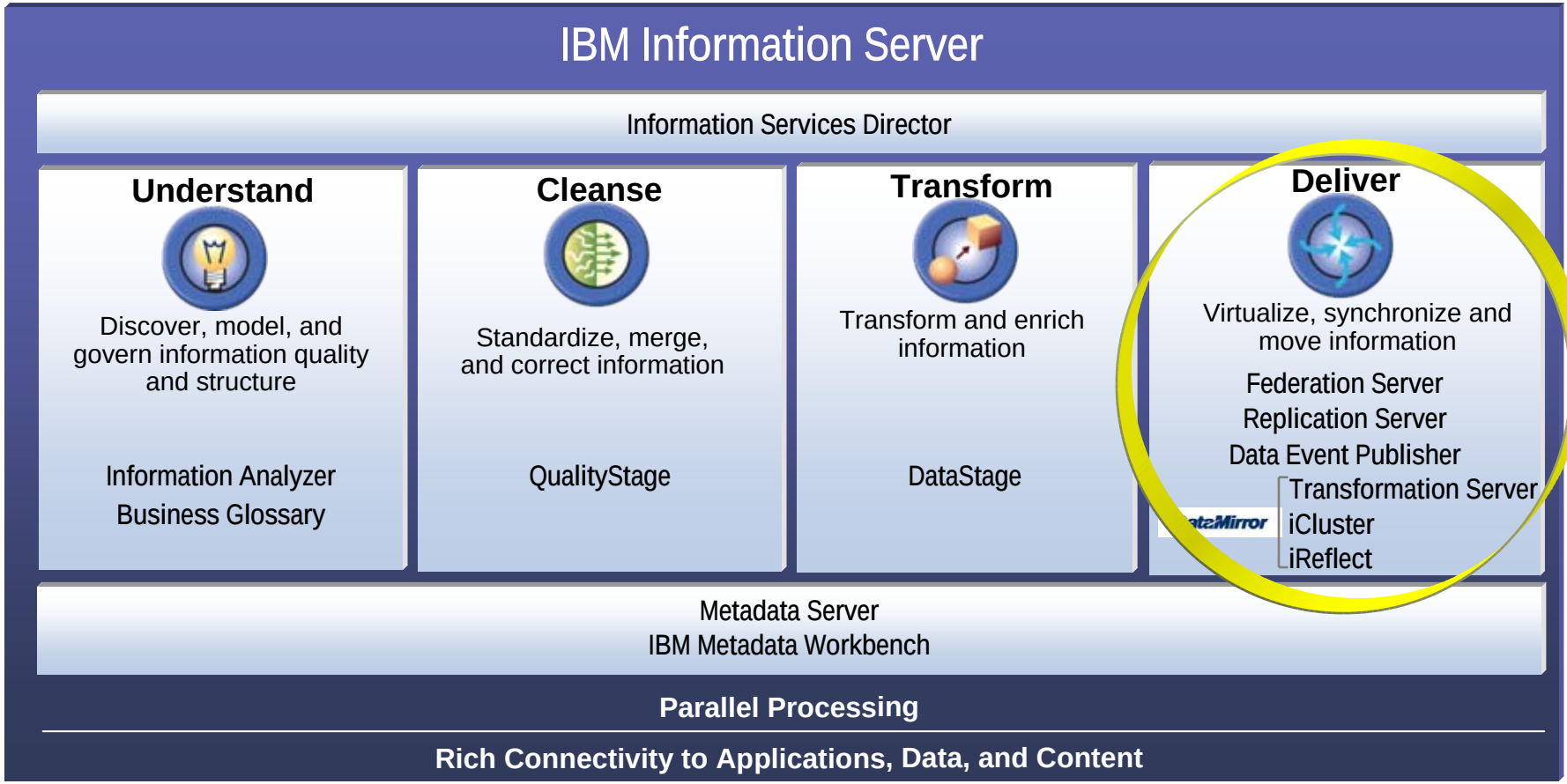
1. Uses the same visual paradigm of DataStage to work with enterprise apps
2. Removes the need to code at low-level API to work with enterprise apps

What Makes WebSphere DataStage Different?

Easy Design of Complex Data Processing	Benefits
Graphical, top-down design metaphor, with extensive library of pre-built functions & graphical sequencing	Faster time to market, Low cost to develop skills, Lower maintenance costs
Extensible, component-based architecture	Lower risk, Better capitalizes on existing investments
Strong reuse capabilities, including shared containers, routines, connection objects, and reusable services	Better consistency, faster time to market, stronger project leverage
Broad and deep connectivity, with bulk connectivity, changed data capture, and dynamic connectivity options	Better utility, better project flexibility, faster time to market
Rapid SOA deployment capability	Better utility, broader applicability
Massive Scalability	
Design serially, deploy in parallel	Able to deal with any data volume without logic changes, Greater utility
Metadata-driven Integration	
Unified metamodel across IBM Information Server	Speeds project delivery, Improves collaboration, Produces better results
Active metadata analysis, including diff, impact, and lineage	Better productivity, reduced risk

IBM Information Server Offerings

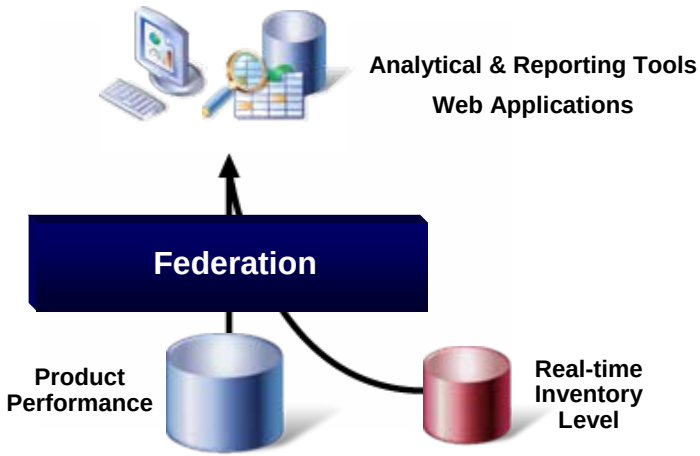
Delivering information you can trust



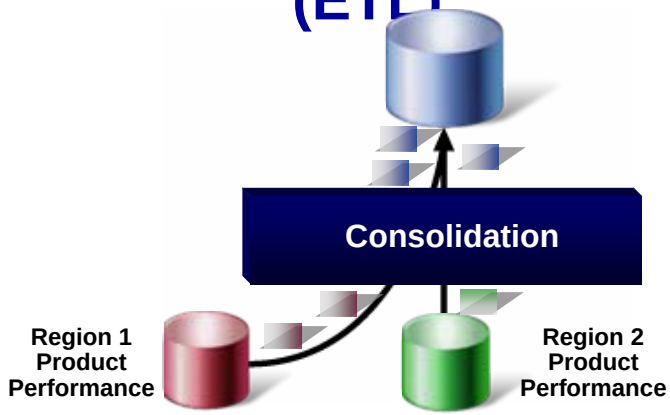
***Note: Transformation Server also implies Transformation/ES**

IBM Information Server for Different Styles of Integration

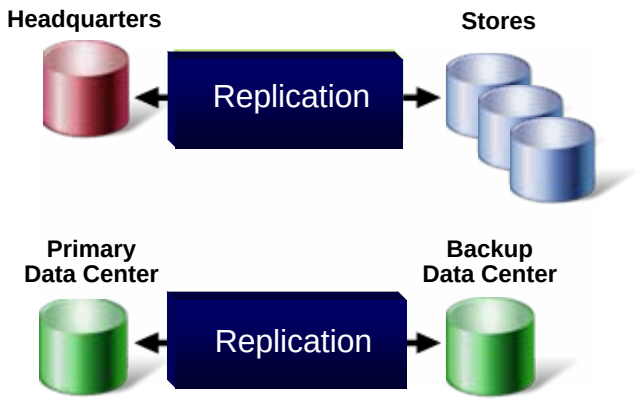
Federation



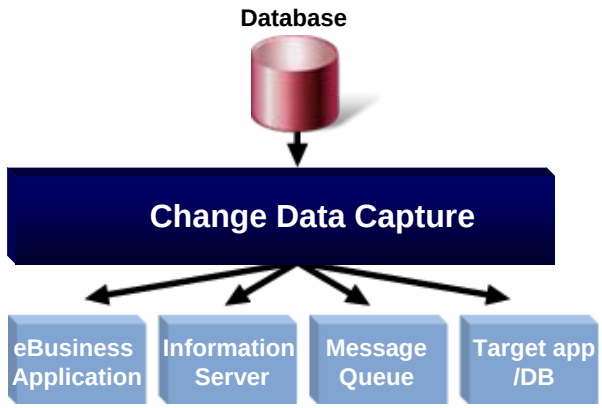
Consolidation (ETL)



Replication

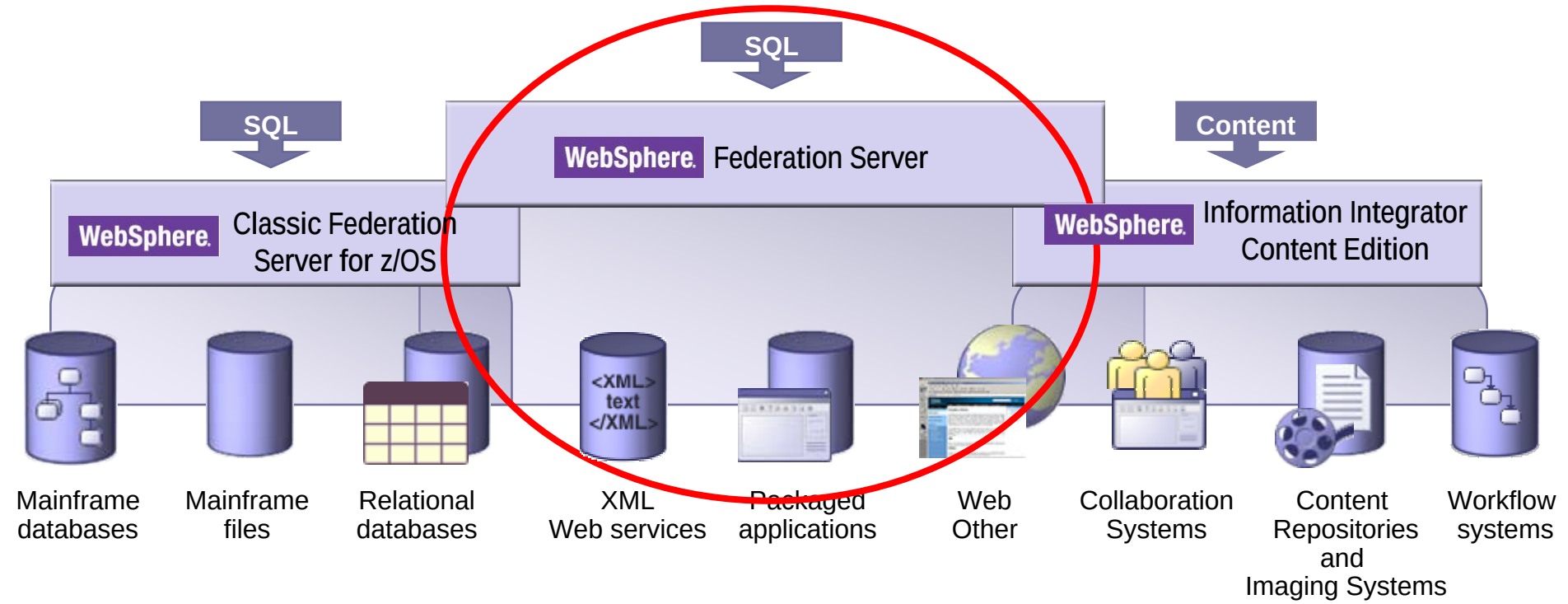


Change Data Capture



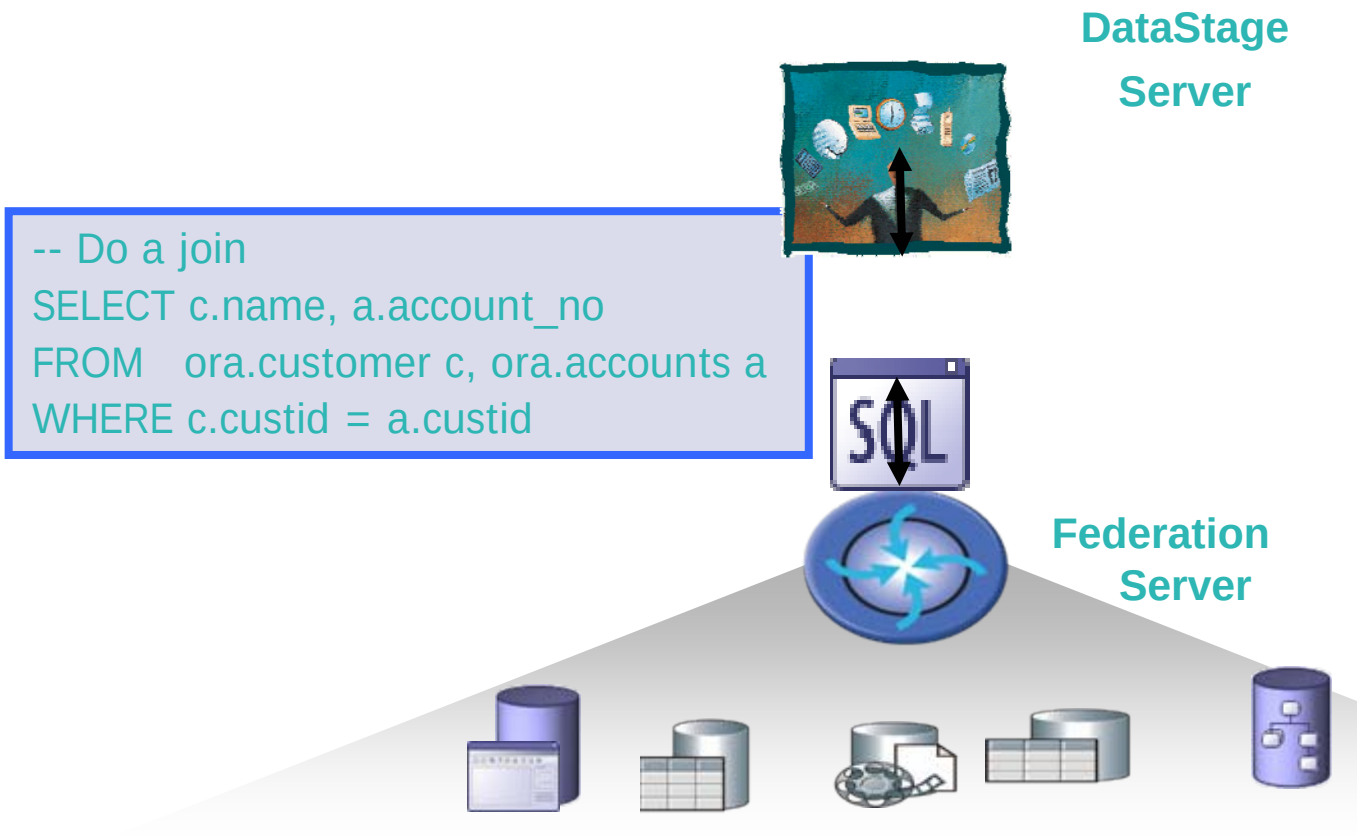
Combining Federation Servers

- WebSphere Federation Server can be combined with Classic Federation Server for z/OS and Information Integrator Content Edition to provide single query access to mainframe and unstructured sources

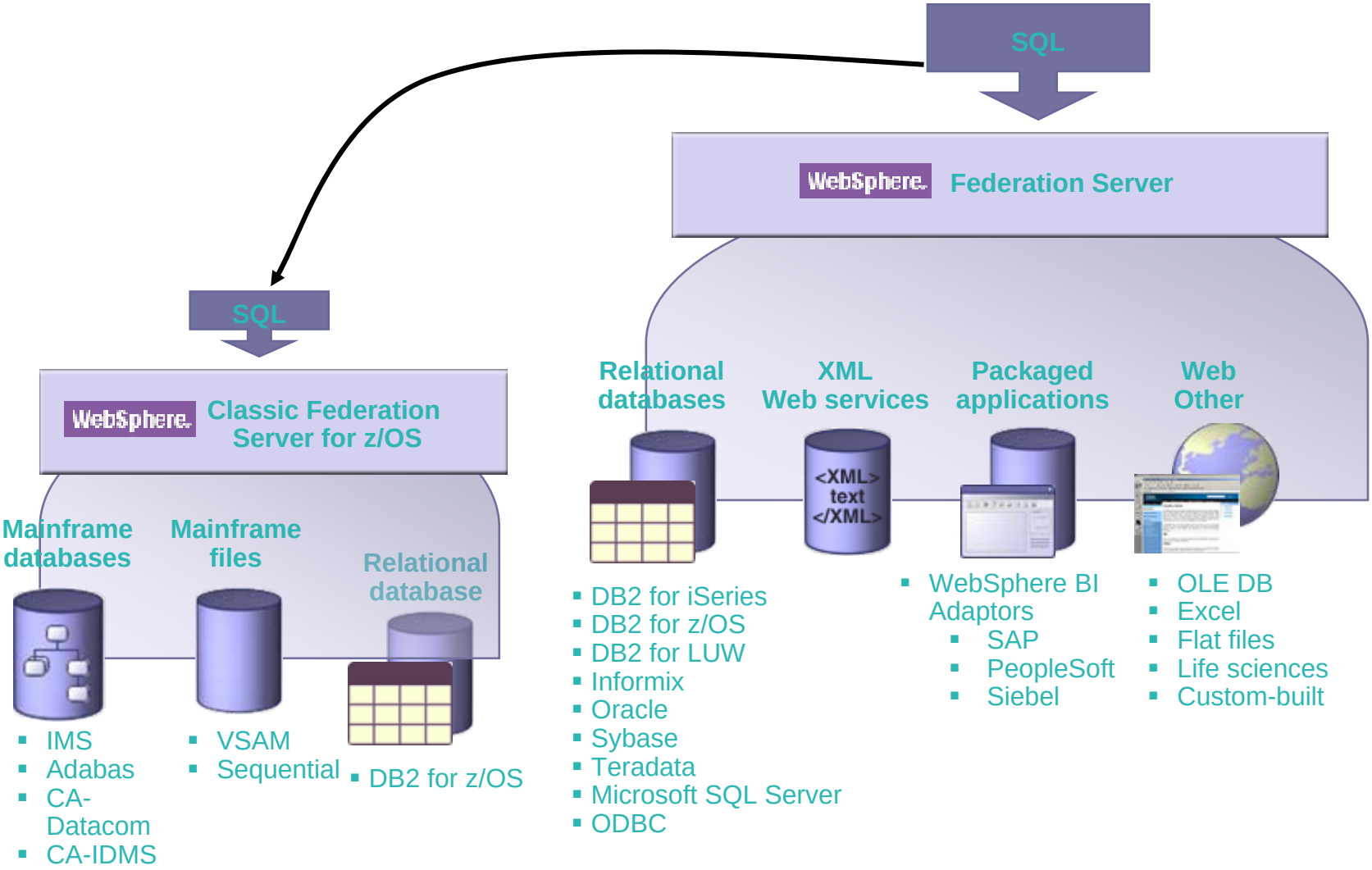


Data Federation w/ DataStage

Use SQL to access nicknames as if they were relational tables.
Works with wide variety of products and technologies.

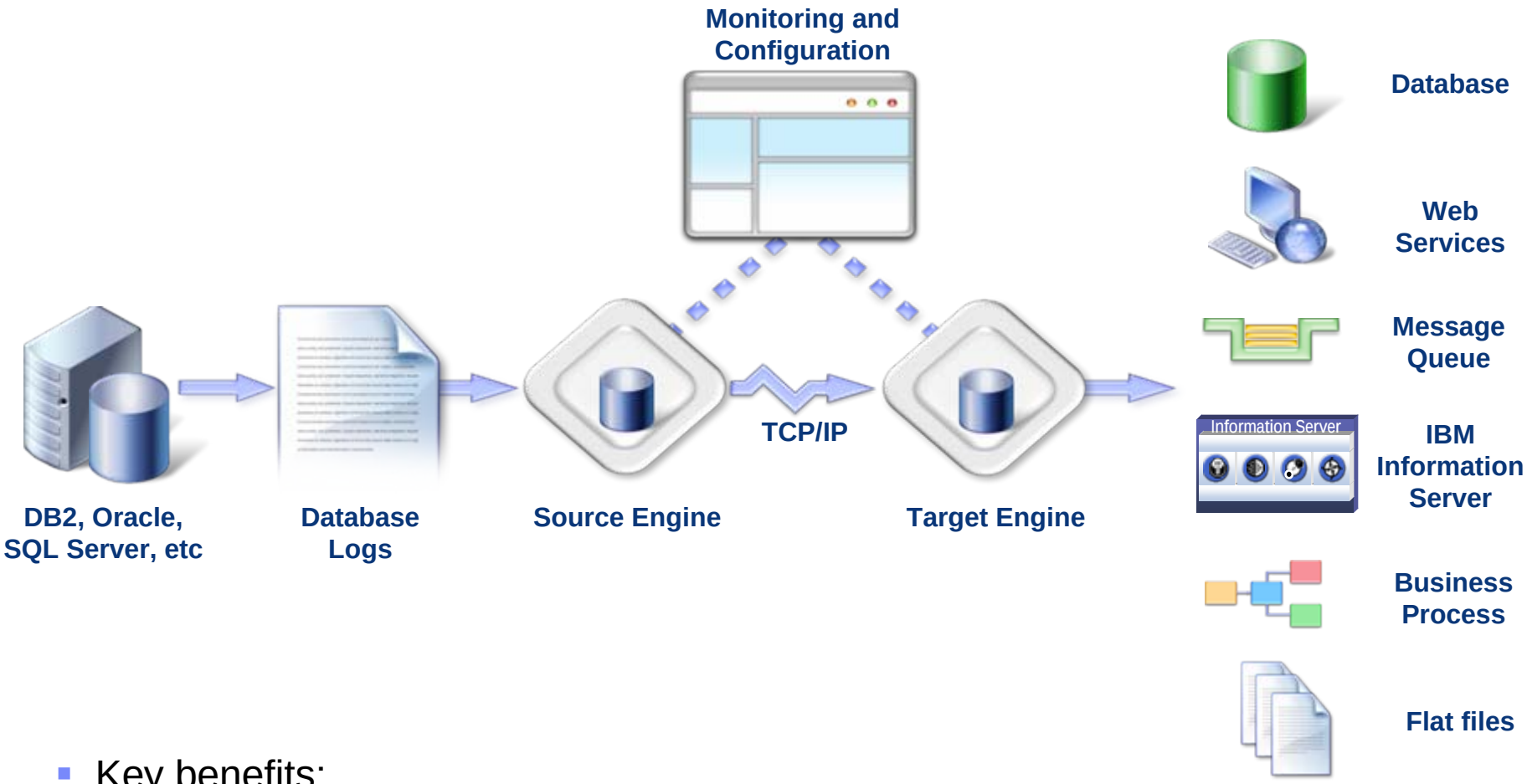


Open + Mainframe Data Sources supported !



Plus partner tools and custom-built connectors extend access to more sources

Log-Based Change Data Capture



- Key benefits:
  - Low impact
  - Flexible implementation
  - Heterogeneous platform support
  - Easy to use

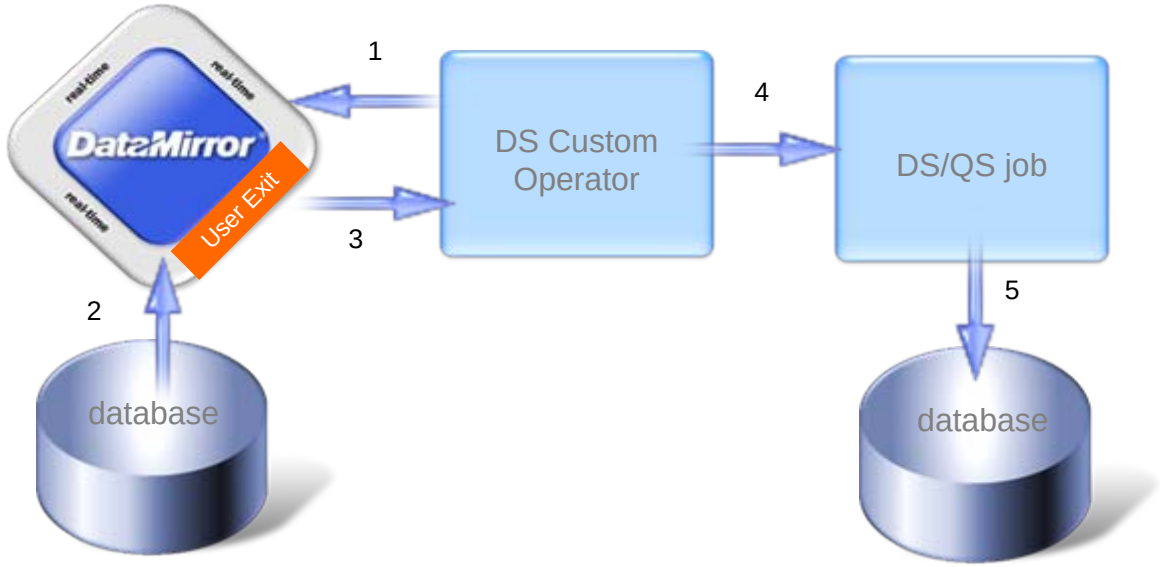
Low Impact

- Log-based CDC captures data without interacting with database
  - 0.05% system resources required to process 300+GB
- No changes or upgrades to applications and schemas required
- Peer-to-peer architecture does not require additional hardware
- Sending only changed data requires minimal network bandwidth

Direct Connect with DataStage

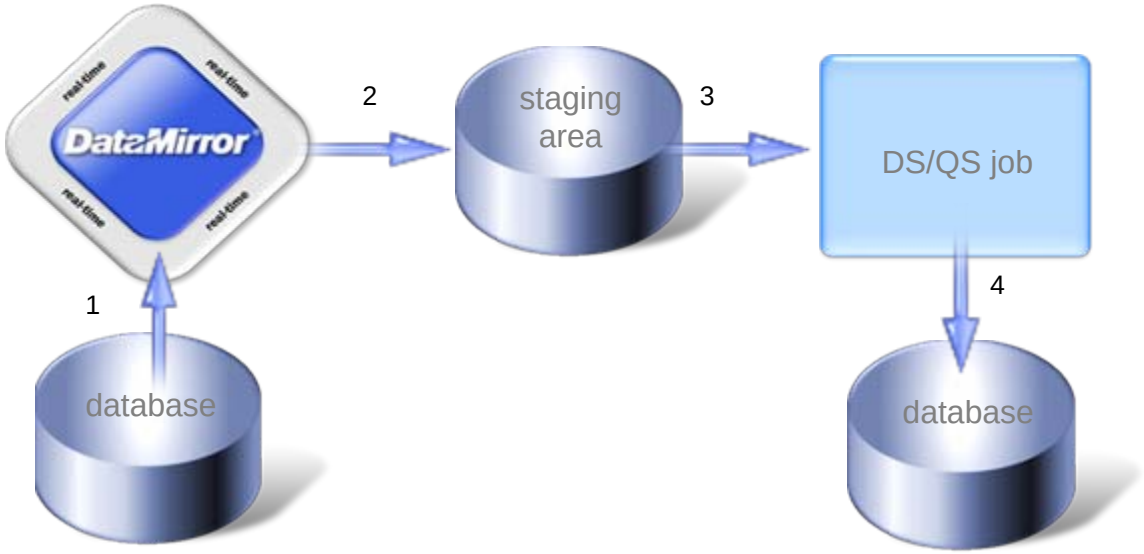
- Enabling real-time response to data changes and business events
  - Low impact log-based changed data capture
  - New palette stages on Information Server
  - Full bi-directional replication capabilities
  - Stream data changes into Information Server

Direct Connect – Available Today



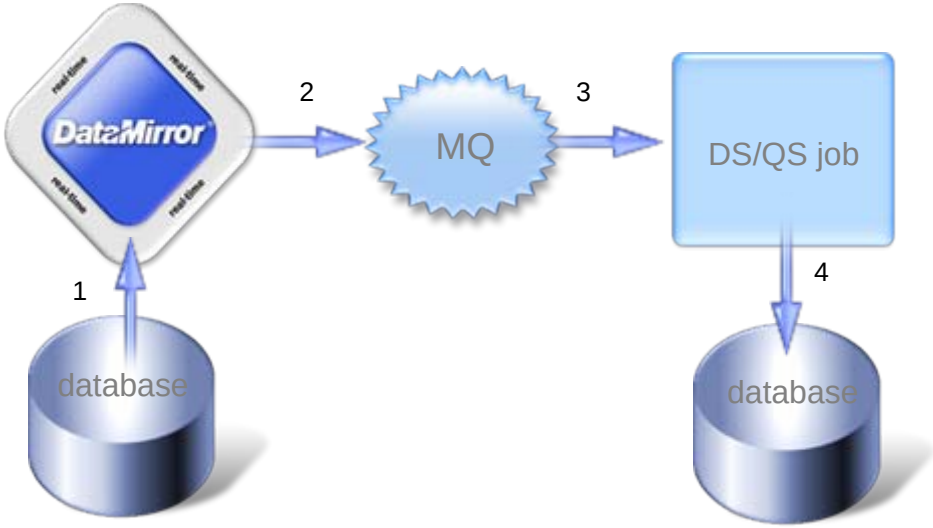
1. Custom operator, which runs on regular intervals, requests the changed data from DataMirror
2. DataMirror captures/collects changes made to remote database
3. Captured changes passed to user exit and writes to comm port
4. Custom operator passes data off to downstream stages
5. Update target database with changed data

Option 1 (out of the box) – Database Staging



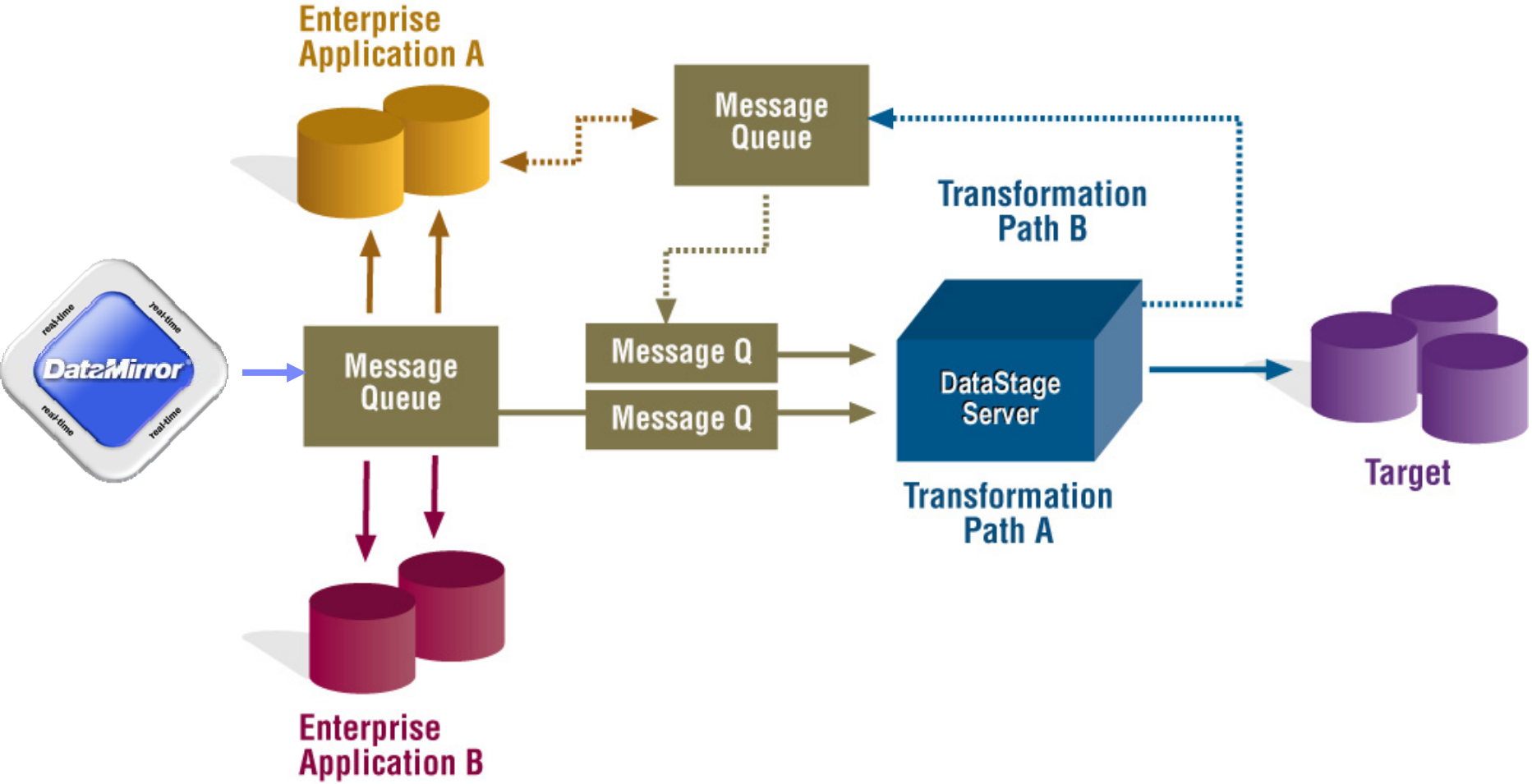
1. DataMirror captures change made to source database
2. DataMirror writes changes to a staging table.
3. DataStage reads the changes from the staging table, transforms and cleans the data as needed
4. Update target database with changes
5. Update internal tracking with last DataMirror bookmark processed

Option 2 (out of the box) – MQ based integration

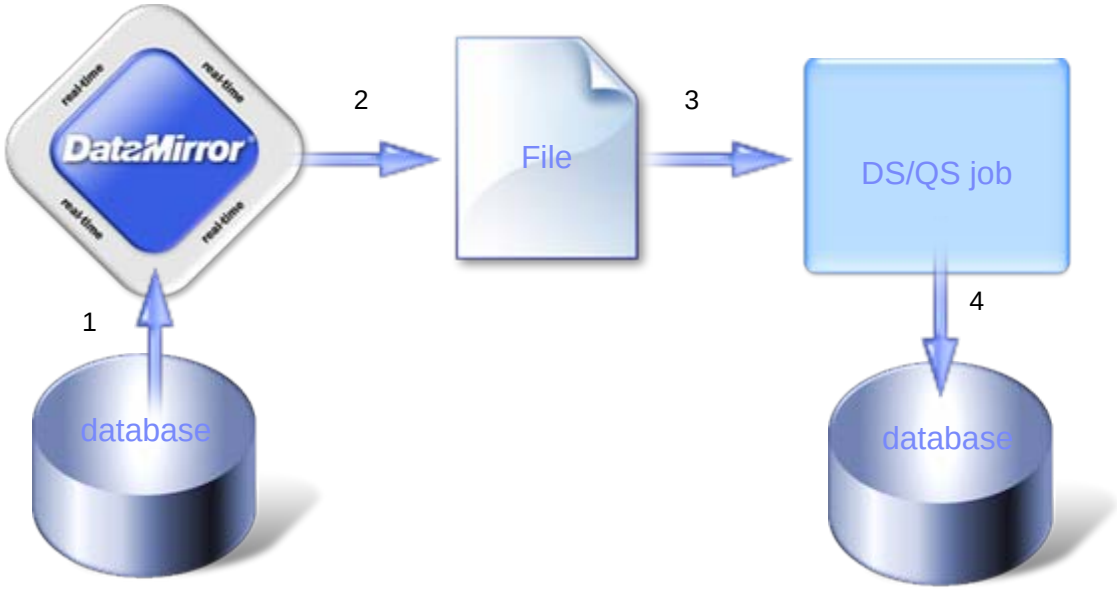


1. DataMirror captures/collects changes made to remote database
2. Captured changes written to MQ
3. DataStage (via MQ connector) processes message and passes data off to downstream stages
4. Updates written to target warehouse

結合IBM Message Queue展現及時資料整合

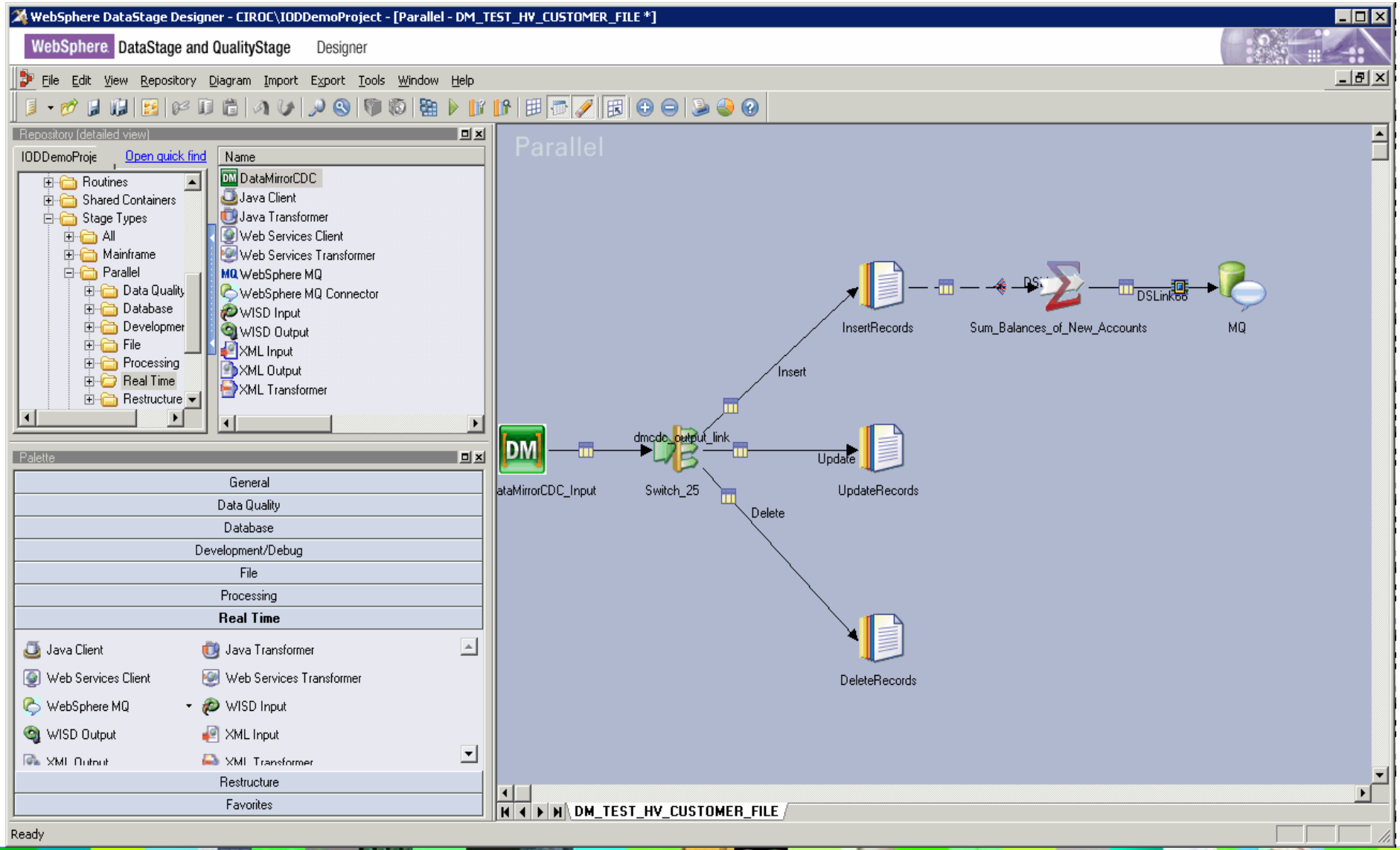


Option 3 – File Based



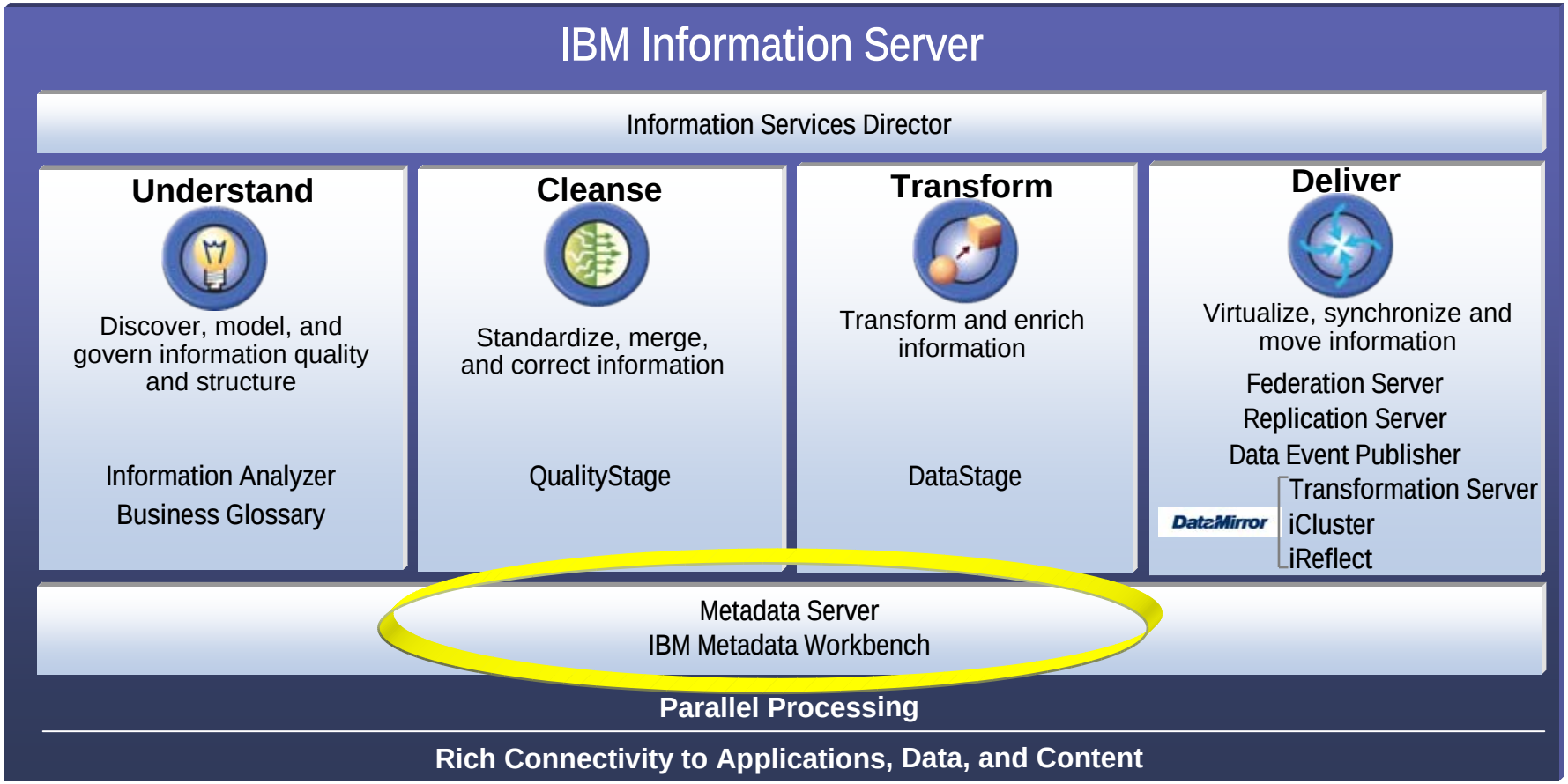
1. DataMirror captures changes made to source database
2. DataMirror writes each transaction to a file
3. DataStage reads the changes from the file
4. Update target database with changes

Get Changed Data from Source



IBM Information Server Offerings

Delivering information you can trust

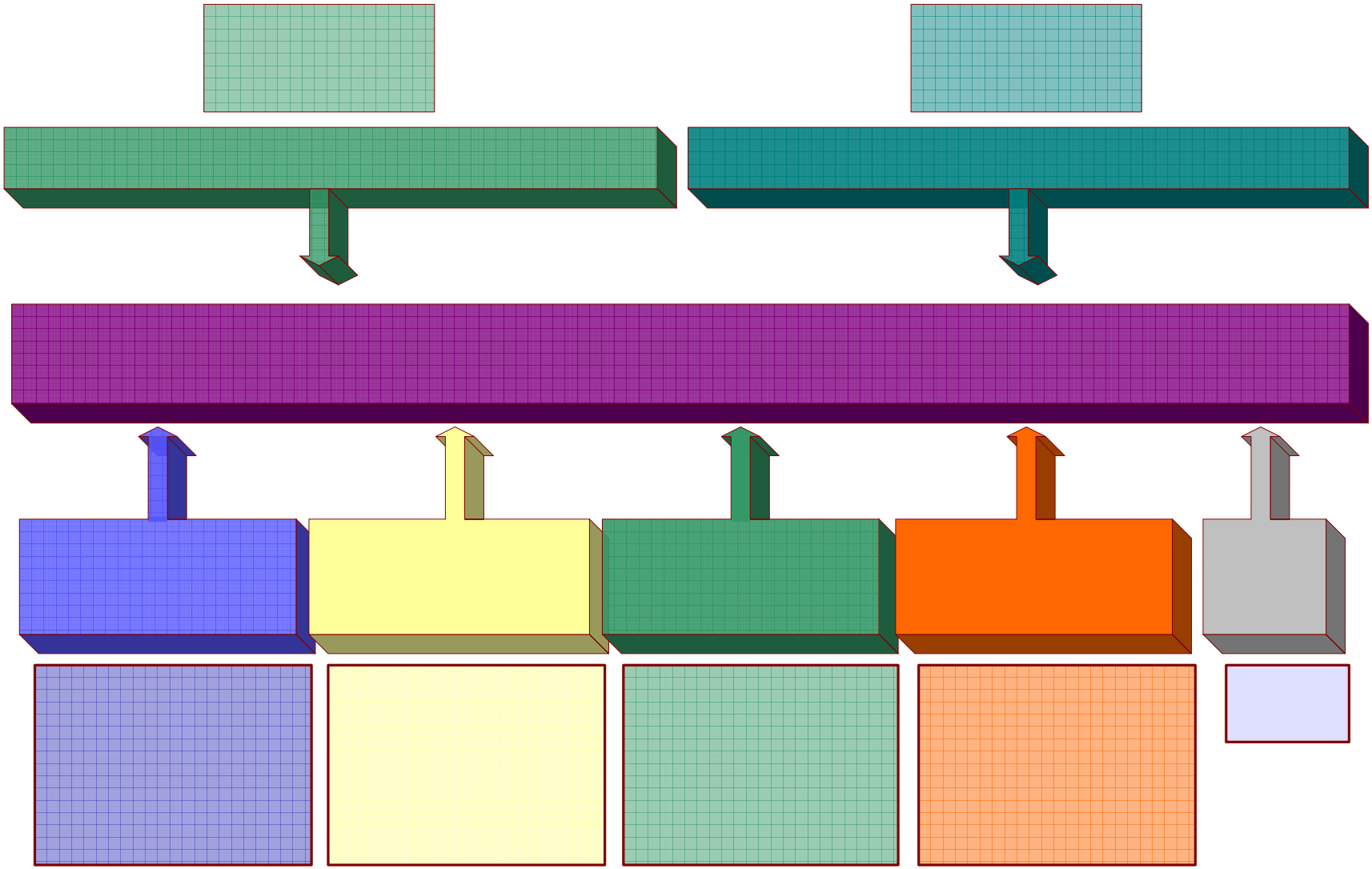


Note: Transformation Server also implies Transformation/ES

IBM Metadata Strategy – over several years

- IBM is uniquely positioned to meet customer demand
  - Widest portfolio of tooling
    - Significant challenge; greatest potential
  - Leading metadata support in existing offerings
    - WSRR, etc.
  - Cross divisional effort on consolidation (Eclipse, Eclipse Modeling Framework, etc.).

IBM's Metadata Vision



Industry Model Integration Points with Information Server


Architects


Data Analysts


Subject Matter Experts


Developers


DBAs

Rational Data Architect

Information Analyzer

Business Glossary

DataStage & QualityStage

Metadata Server

Import/Export Physical Models

Metadata Server

MetaBroker Import

CA ER/Win

Import/Export Logical Models

Data Models

Metadata Server

MetaBroker Import/Export

IBM Industry Models provide industry-proven acceleration by pre-populating IBM Information Server with definitions and designs

Role-Based Tools with Integrated Metadata


Business Users


Subject Matter Experts


Architects


Data Analysts


Developers


DBAs

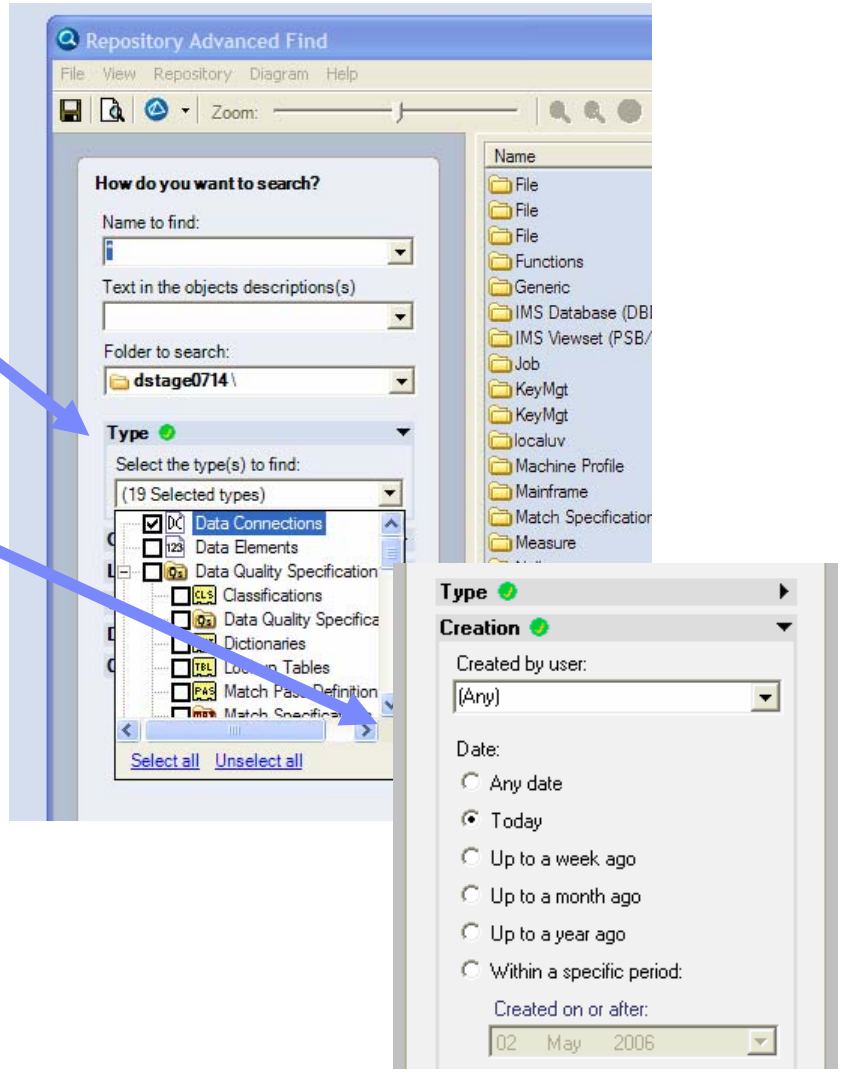
Unified Metadata Management

Design Operational

- Simplify Integration
- Facilitate change management & reuse
- Increase trust and confidence in information
- Increase compliance to standards

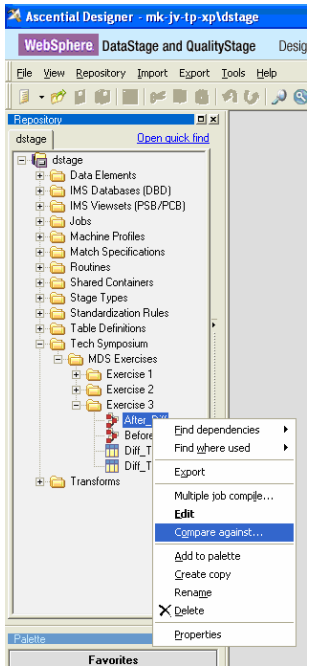
Find – Advanced Search Criteria

- Search on following criteria:
  - Object type
    - Job, Table Definition, Stage etc.
  - Creation
    - Date/Time
    - By User
  - Last Modification
    - Date/Time
    - By User
  - Where Used
    - What other objects use this object?
  - Dependencies of
    - What does this object use?
- Options
  - Case
  - Match on “name & description” or



Job, Table or Routine Difference

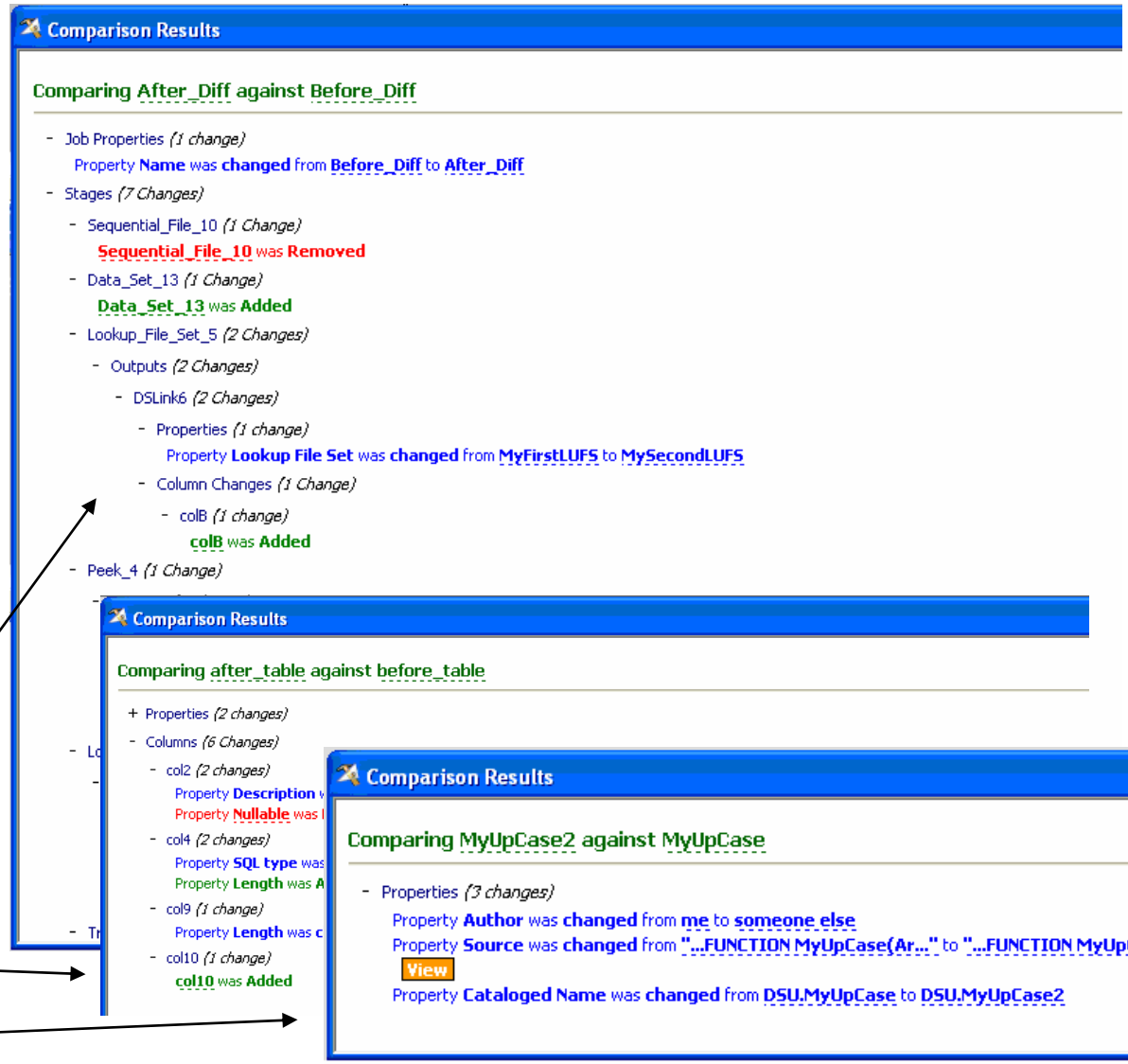
Available for Jobs, Tables & Routines



Textual report with hot links to the relevant editor in Designer.

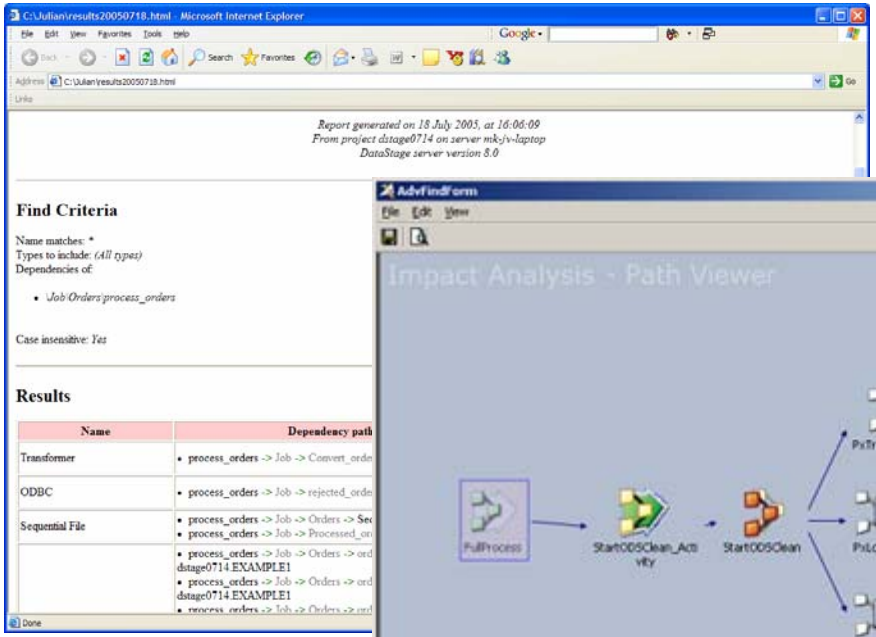
Tables

Routines

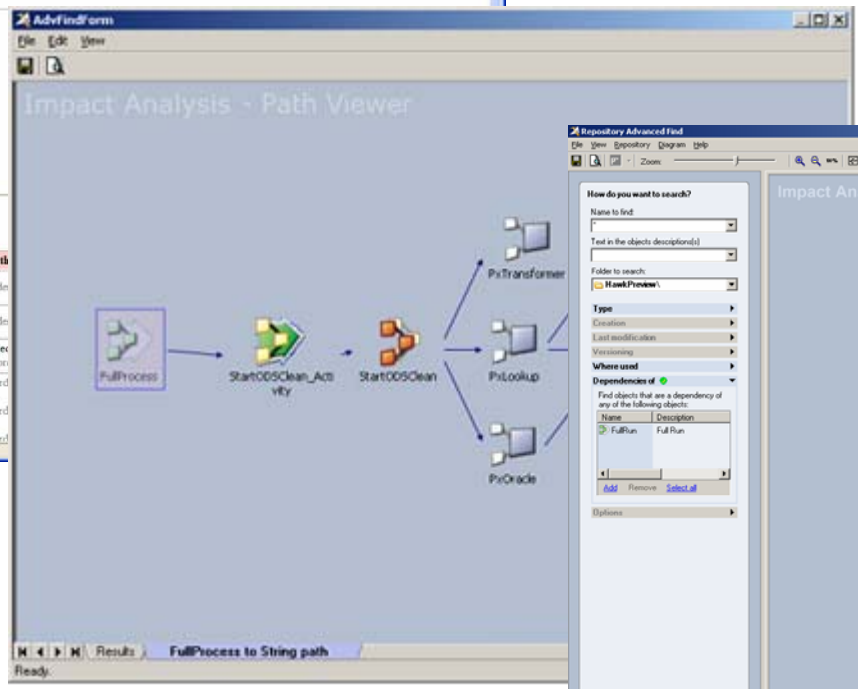


圖形化且易於解讀的特性，提供異動影響分析圖，可採用不同的角度切入，使開發者更易於估算，因業務需求的變動所需配合修改的幅度

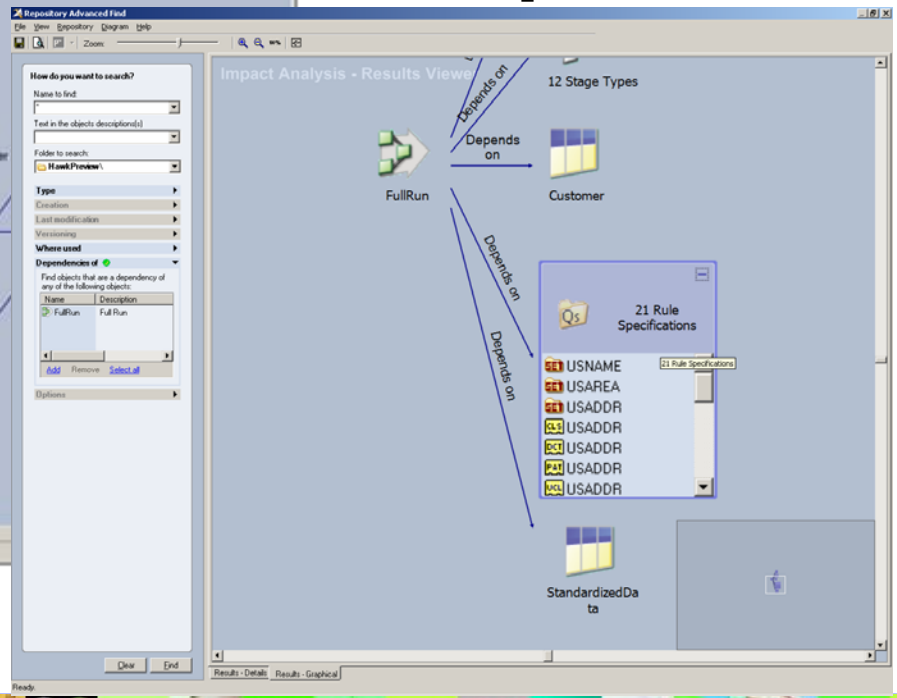
HTML View



Path View



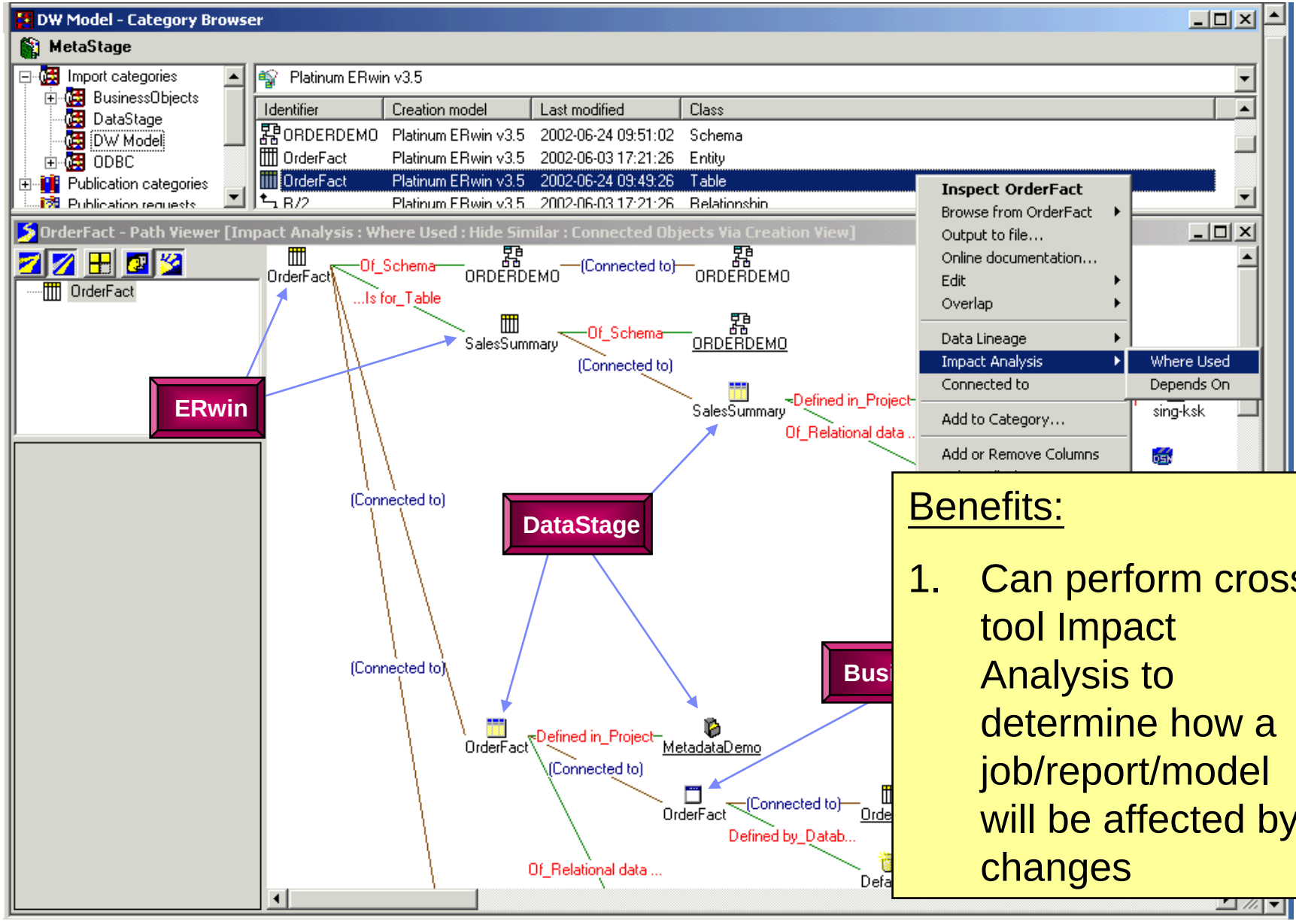
Graphical Tree View



MetaBrokers & Metadata Bridges

Metadata Bridges for Import - Information Server 8.0	
Adaptive Repository & Foundation (CWM XMI)	Meta Integration Repository (MIR) on Access
ArgoUML (UML XMI)	Meta Integration Repository (MIR) via ODBC
Borland Together (UML XMI)	Microsoft? Excel
BusinessObjects Crystal Reports (8.5 to 9.0)	Microsoft Repository 2.1b (XIF)
BusinessObjects Data Integrator (CWM XMI)	Microsoft Repository 3.x (MDC)
BusinessObjects Designer (File)	Microsoft SQL Server? 2005 Analysis Services (DSV)
BusinessObjects Designer (Repository)	Microsoft SQL Server 2005 Data Source View (DSV)
Categories and Terms	Microsoft SQL Server 2005 Integration Services (DSV)
CA AllFusion Component Modeler (4.x UML XMI)	Microsoft SQL Server 2005 Reporting Services (DSV)
CA AllFusion ERwin 4.x Data Modeler	Microsoft SQL Server 7.0 to 2005 Analysis Services (DSO)
CA AllFusion ERwin 7.x Data Modeler	Microsoft Visio Database (ERX)
CA AllFusion Gen 4.1a to 7.5	Microsoft Visio UML (UML XMI)
CA AllFusion Repository DS - ODBC	Microsoft Visual Studio/Modeler 2.0 (MDL)
CA COOL:Biz 5.1	MicroStrategy 7.0 to 8.0
CA COOL:BusinessTeam (GroundWorks 2.2.1)	NCR Teradata MDS 5.0 to 6.x
CA COOL:DBA (Terrain for DB2) 5.3.2	NoMagic MagicDraw (UML XMI)
CA COOL:Enterprise (ADW) 2.7	ODBC 3.0 MetaBroker
CA COOL:Xtras Mapper (TerrainMap for DB2)	OMG CWM 1.0 and 1.1 XMI 1.1
CA ERwin 3.x (ERX)	OMG CWM Pre-1.0 XMI 1.1
CA ParadigmPlus 3.52	OMG UML 1.1 to 1.4 XMI 1.x
Cobol Copybook Flat Files	Oracle Designer 1.3.2, 2.1.2, 6.0, 6i & 9i
Cognos 8 Framework Manager (File)	Oracle Warehouse Builder (CWM XMI)
Cognos Impromptu	Oracle Warehouse Builder 10.2
Cognos ReportNet Framework	Popkin System Architect 7.1.12 to 10.x
Manager (File)	ProActivity 3.x & 4.0
Cognos ReportNet ReportStudio (File)	SAS Data Integration Studio (MIR XMI)
Embarcadero ER/Studio 5.1 to 7.1	SAS ETL Studio (CWM XMI)
Gentleware Poseidon (UML XMI)	SAS Information Map Studio (MIR XMI)
Hummingbird ETL/Genio 5.04	SAS Management Console (MIR XMI)
Hyperion Application Builder (CWM XMI)	Select SE 7.0
Hyperion Essbase Integration Services 7.0	Silverrun-RDM 2.4.4 to 2.7.2
IBM DB2? Cube Views	Sybase PowerDesigner CDM 6.1.x
IBM DB2 OLAP Integration Server 8.1	Sybase PowerDesigner CDM 7.5 to 12.0
IBM DB2 Warehouse Manager (CWM XMI)	Sybase PowerDesigner OOM 9.0 (UML XMI)
IBM Rational Data Architect	Sybase PowerDesigner PDM 6.1.x
IBM Rational Rose? 2000e to 2003 (MDL)	Sybase PowerDesigner PDM 7.5 to 12.0
IBM Rational Rose 98(i) to 2000 (MDL)	Telelogic System Architect 7.1.12 to 10.x
IBM Rose XMI Toolkit 1.0 & 1.05 & 1.15 (UML XMI)	Telelogic Tau (UML XMI)
IBM VisualAge? for Java? 3.0 (UML XMI)	Unisys Rose UML utility 1.1 (UML XMI)
IBM WebSphere? Studio 3.0 (UML XMI)	Unisys Rose XMI Interchange (UML XMI)
Informatica PowerCenter (File)	User Information
Informatica PowerCenter (Repository)	Visible IE:Advantage 6.1
Merant App Master Designer 4.0	XML DTD 1.0 (W3C)
Meta Integration Metadata (MIM)	XML Schema 1.0 (W3C XSD)
Meta Integration Metadata (MIR XMI)	XML Schema (Microsoft XDR)

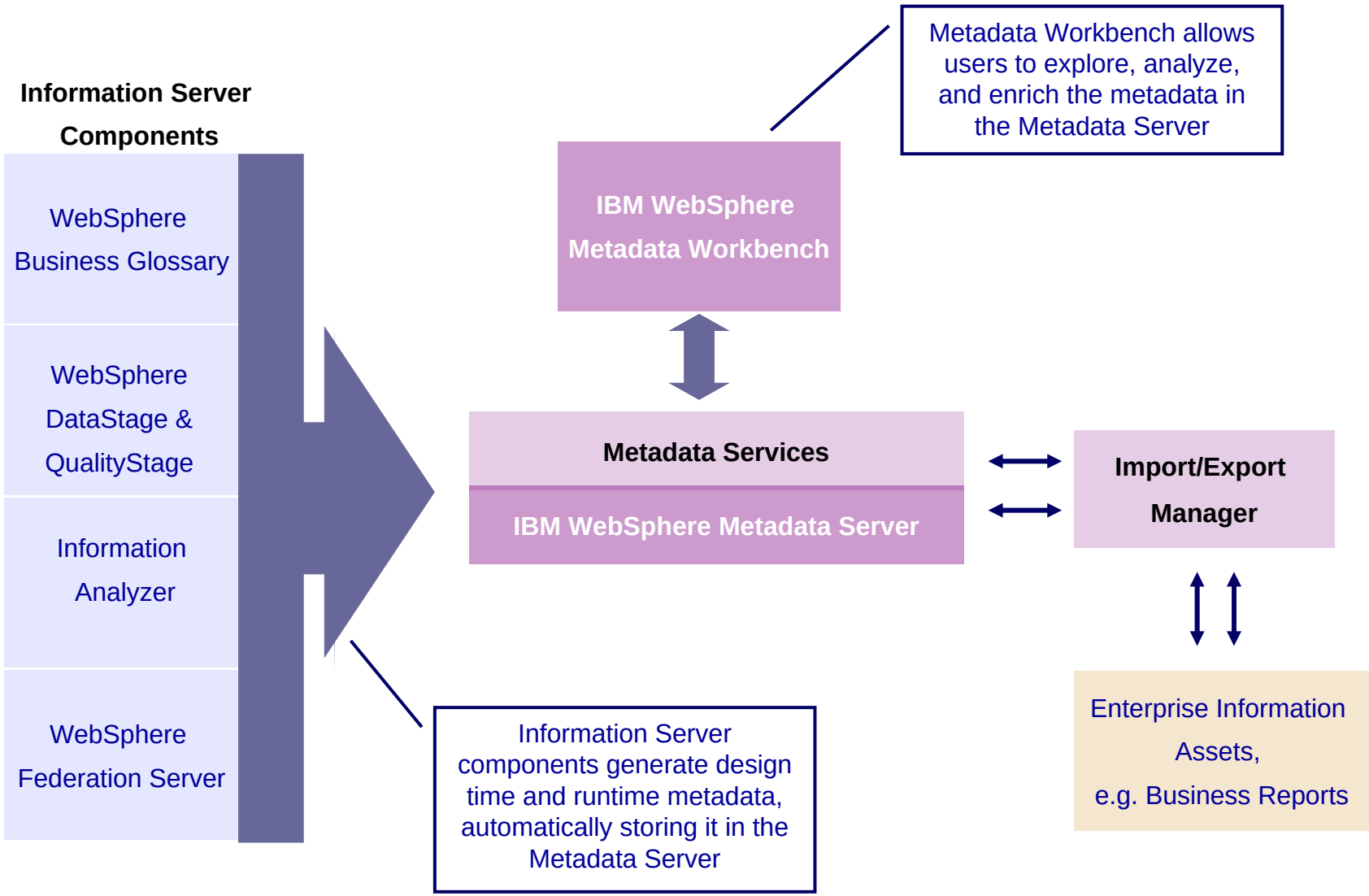
Benefit to Developers – Cross-tool Impact Analysis



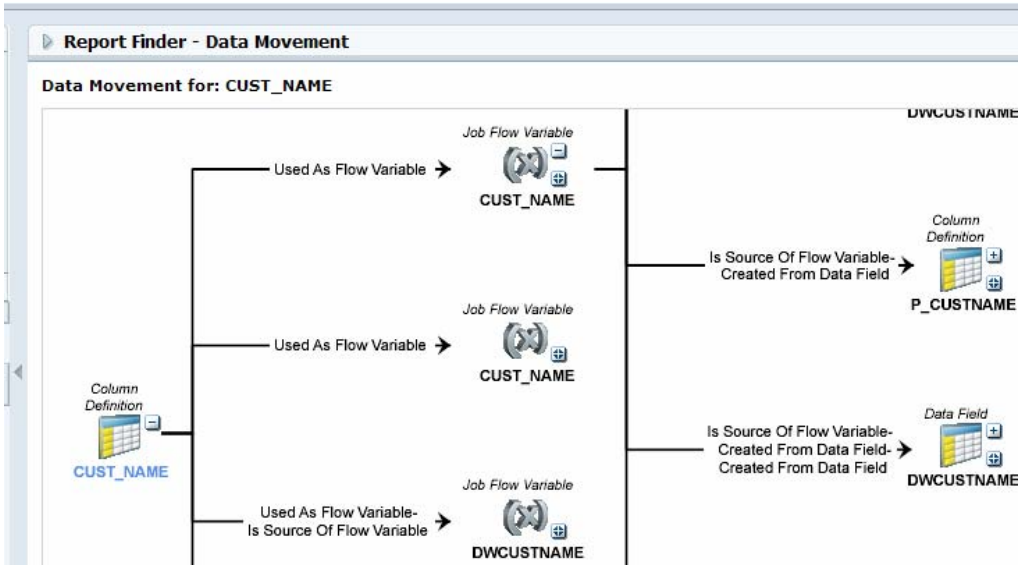
Benefits:

1. Can perform cross-tool Impact Analysis to determine how a job/report/model will be affected by changes

How does Metadata Workbench work?

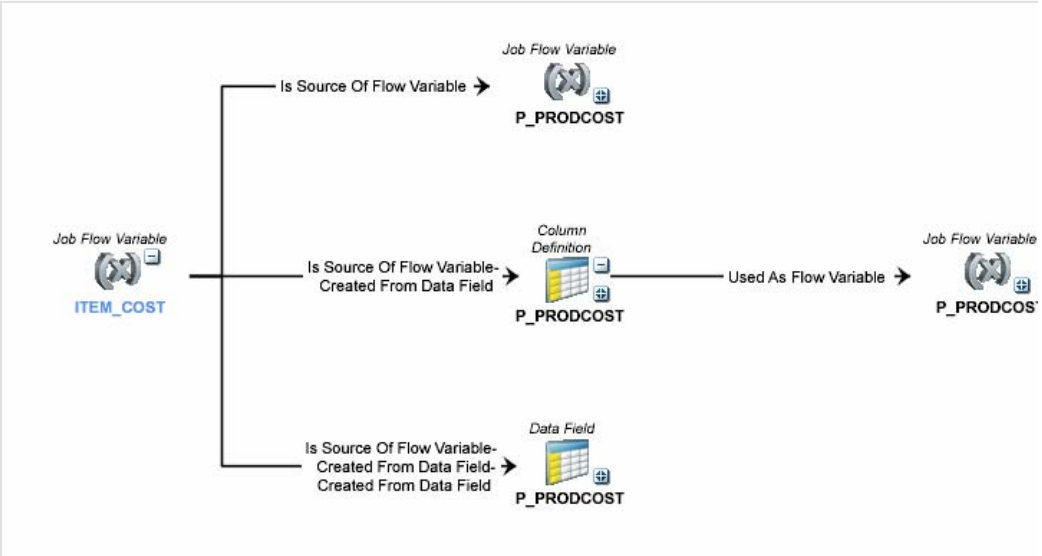


Impact Analysis based on a column



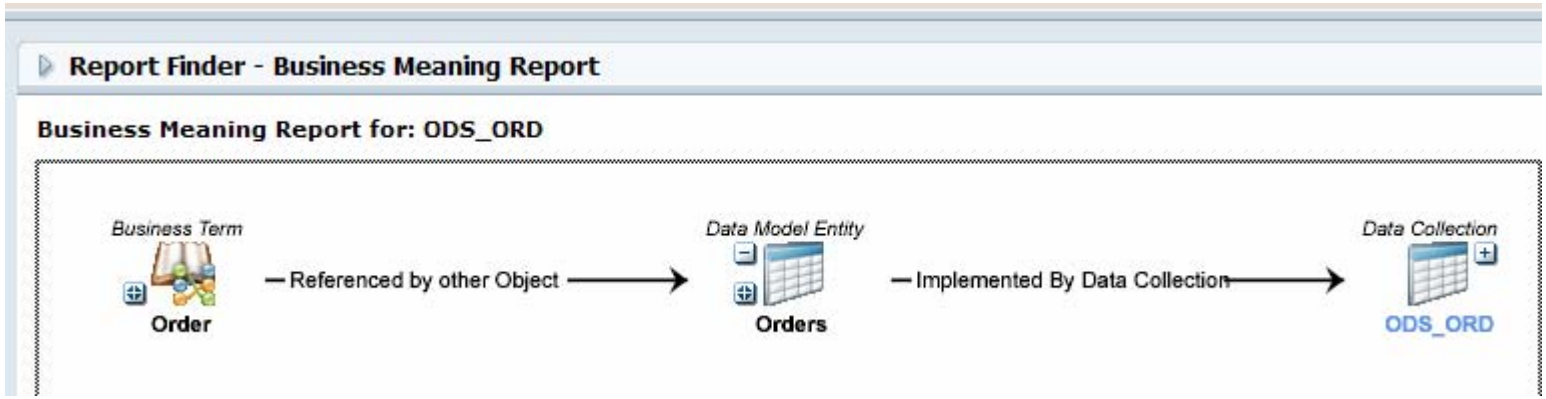
A DataStage ETL developer wants to make a change to the column CUST_NAME and needs to understand which jobs or reports will be impacted by this change.

Impact Analysis based on a DataStage Job



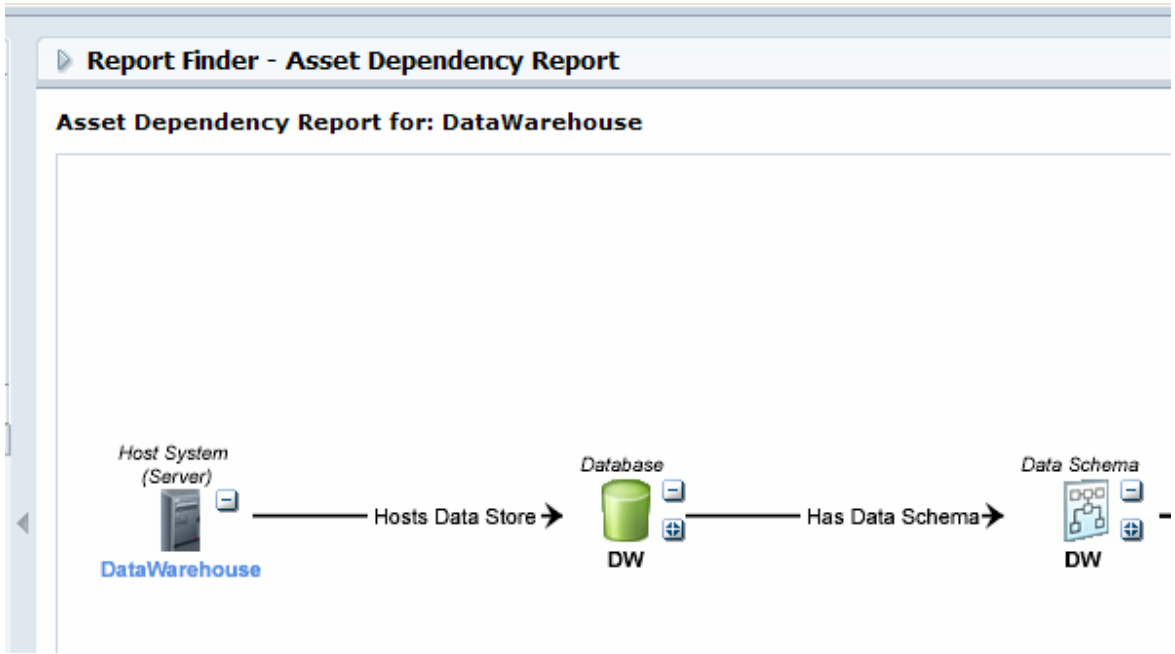
DataStage developer notices a problem in their job design that is causing an extra 0 to be added to the item cost. Is this distorting revenue reports?

Understand Business Meaning and Background



What kind of information is contained in the data collection ODS_ORD? *What else can I discover about this resource?*

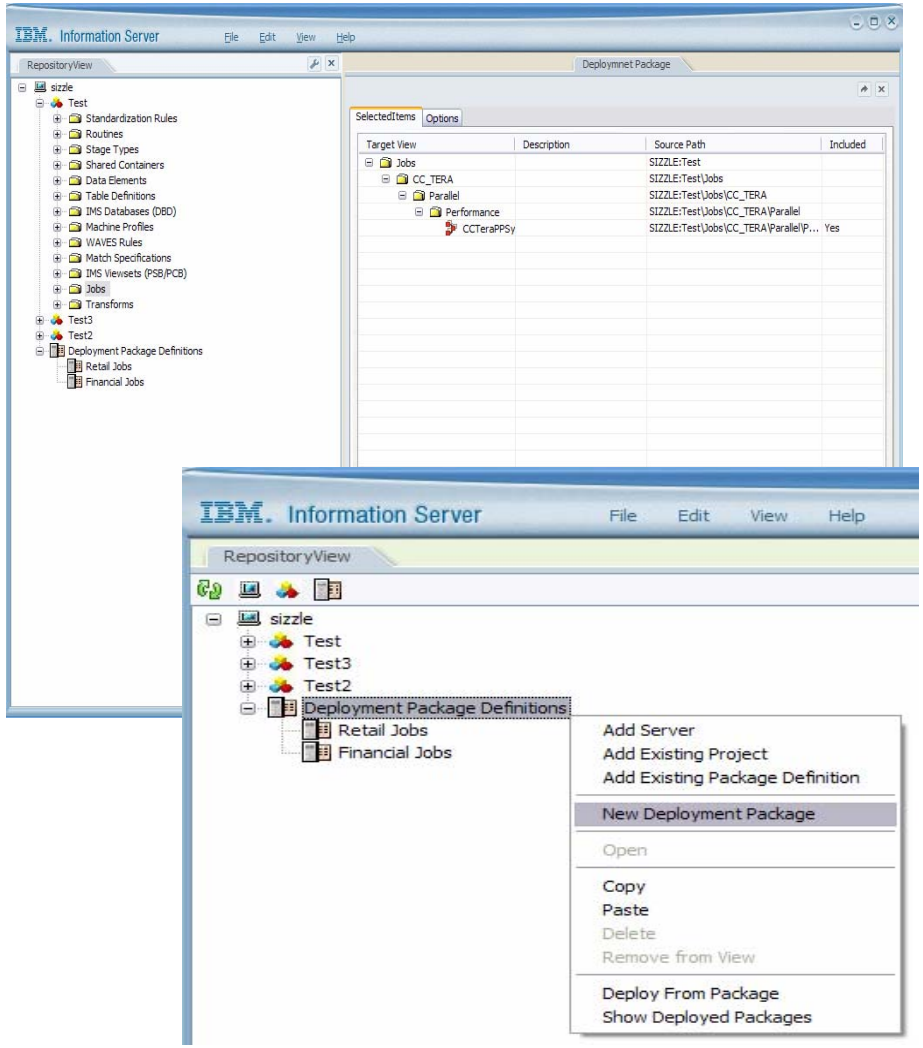
Analyze system dependencies



The data warehouse server needs to be upgraded. Who should I notify and coordinate downtime with?

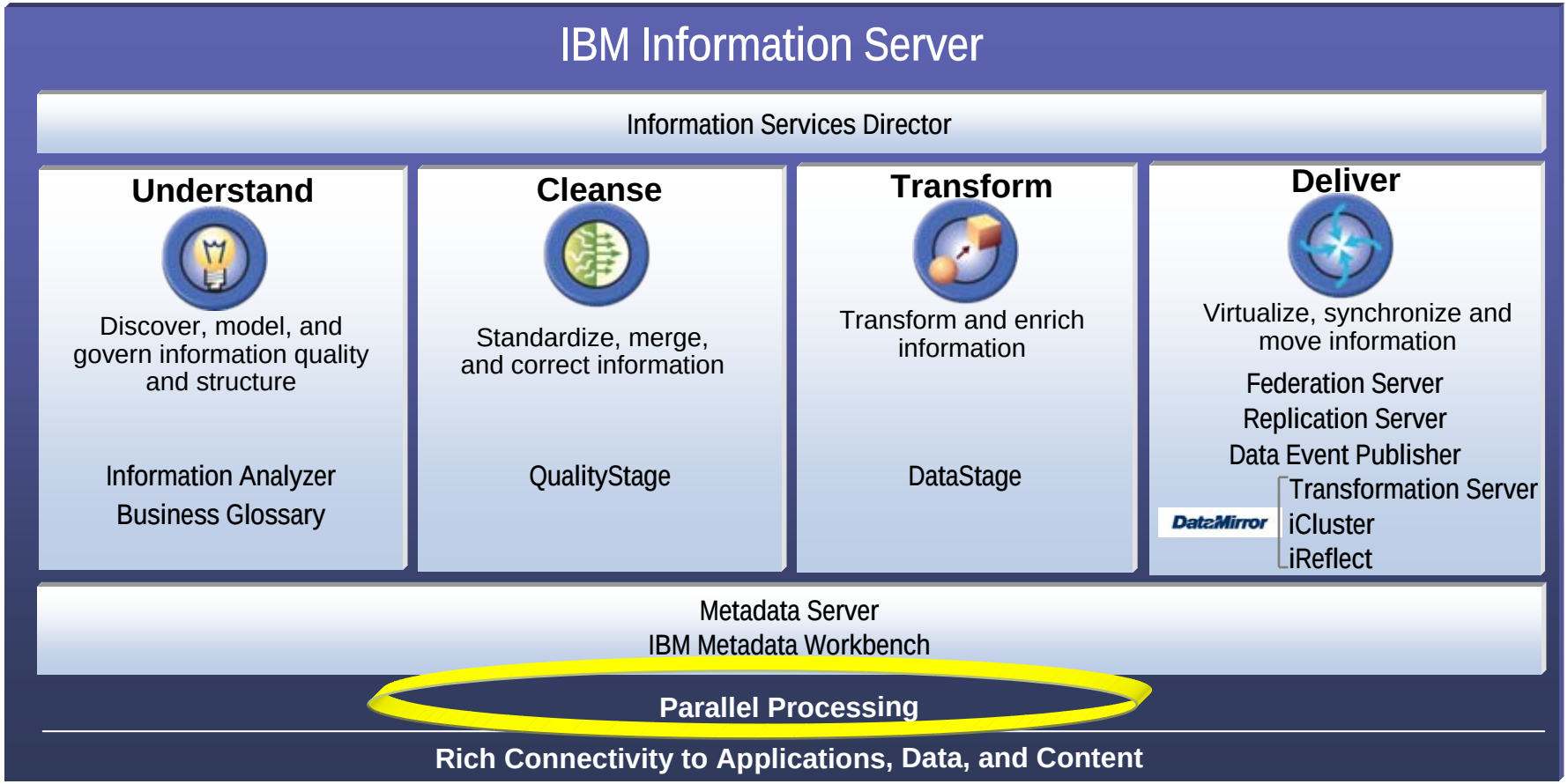
Version Control Targeted for 8.1

- Job Promotion - from Development to Test to Production
- Easy and integrated job movement from one environment to the other
- Deployment of the package to the target system
  - Deploy to the target system – either directly from the source Information Server, from the file system or source control system.
  - Deployment can be using the graphical interface or command line.
- Security checks in the flow as we go from one system to the other
- Support for external source control systems: early support for ClearCase, PVCS. Web Services support for general applications.
- In 8.0 supported by built-in import/export



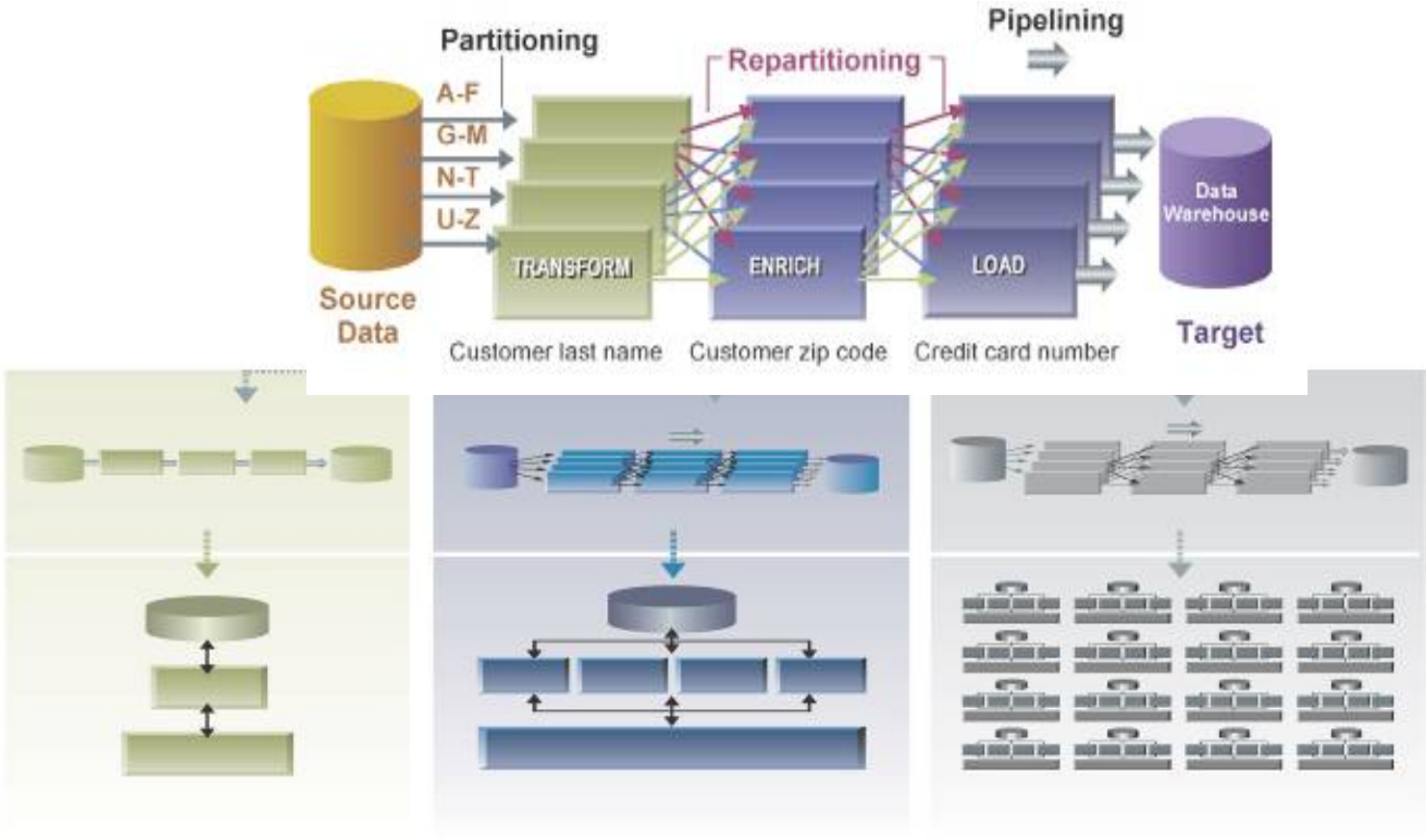
IBM Information Server Offerings

Delivering information you can trust



Note: Transformation Server also implies Transformation/ES

How Do Parallel Processing Services Work?



Uniprocessor

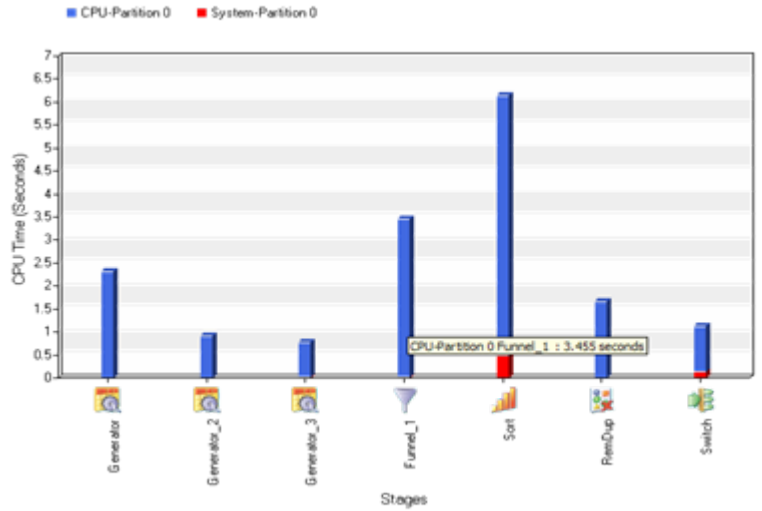
SMP System

MPP, GRID, and Clustered Systems

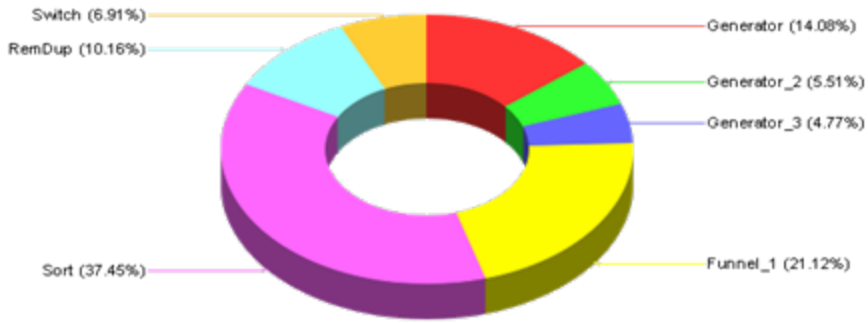
CPU Utilization

- Visualizes the time in CPU of each operator.
- Shows what operators were dominating the CPU at different points during the run.
- Percentage view shows what percentage of the CPU load of the job each stage on the canvas was responsible for.
- Inserted operators and Composite sub-operators automatically get bundled up in these results.
- Advanced users can see combination, which will change this chart to reflect each process and the stages contained within.

Total CPU and System Time

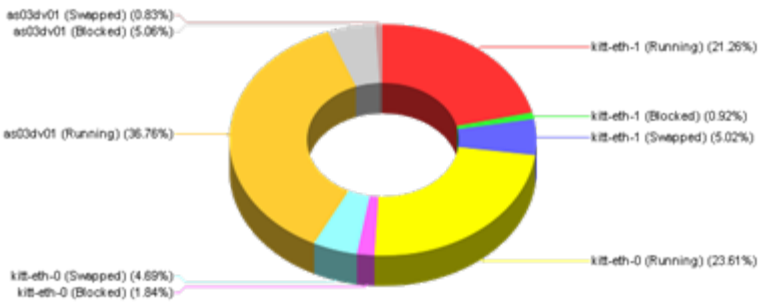


Percentage CPU Pie Chart

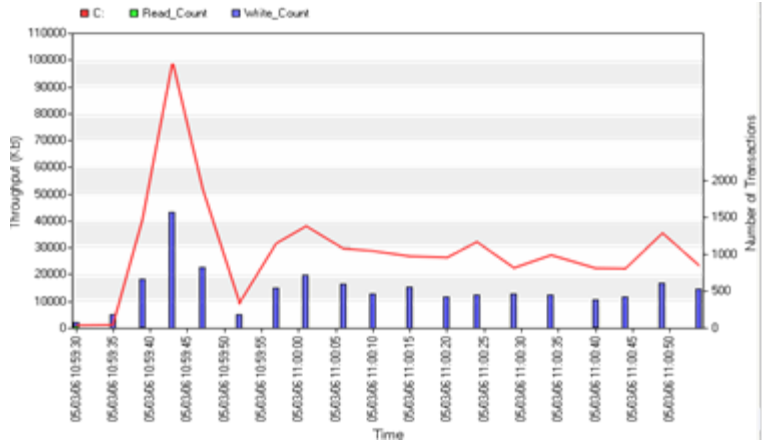


Physical Machine Utilization

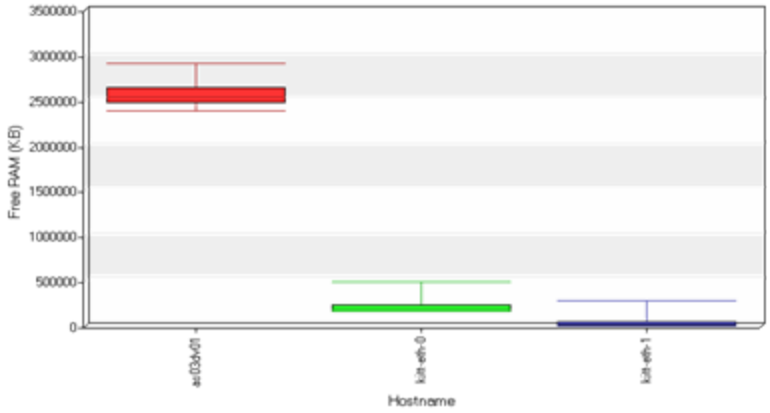
Average Process Distribution



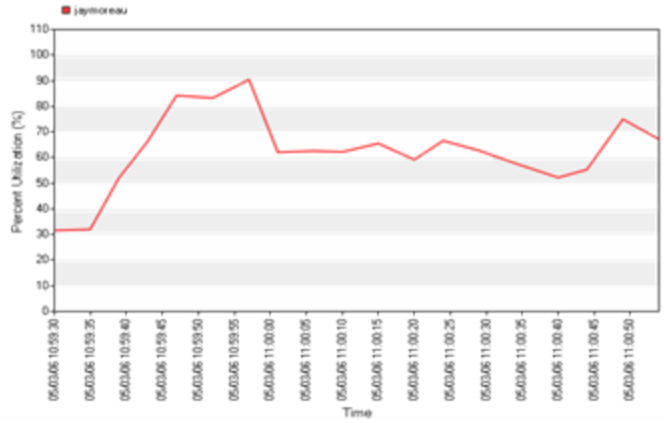
Disk Throughput



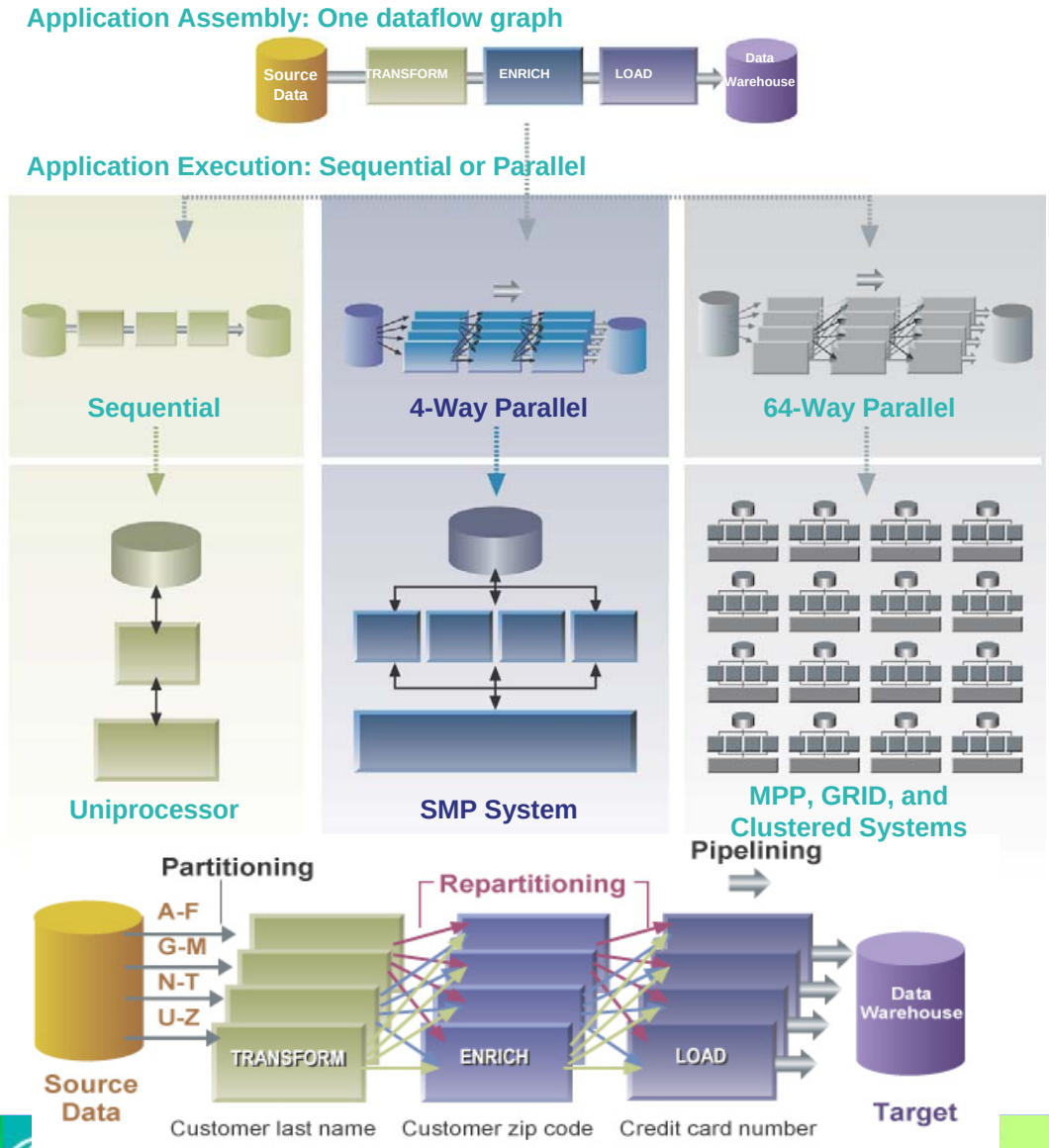
Free Memory Whisker Box



Percent CPU Utilization



馬力加速: 平行處理



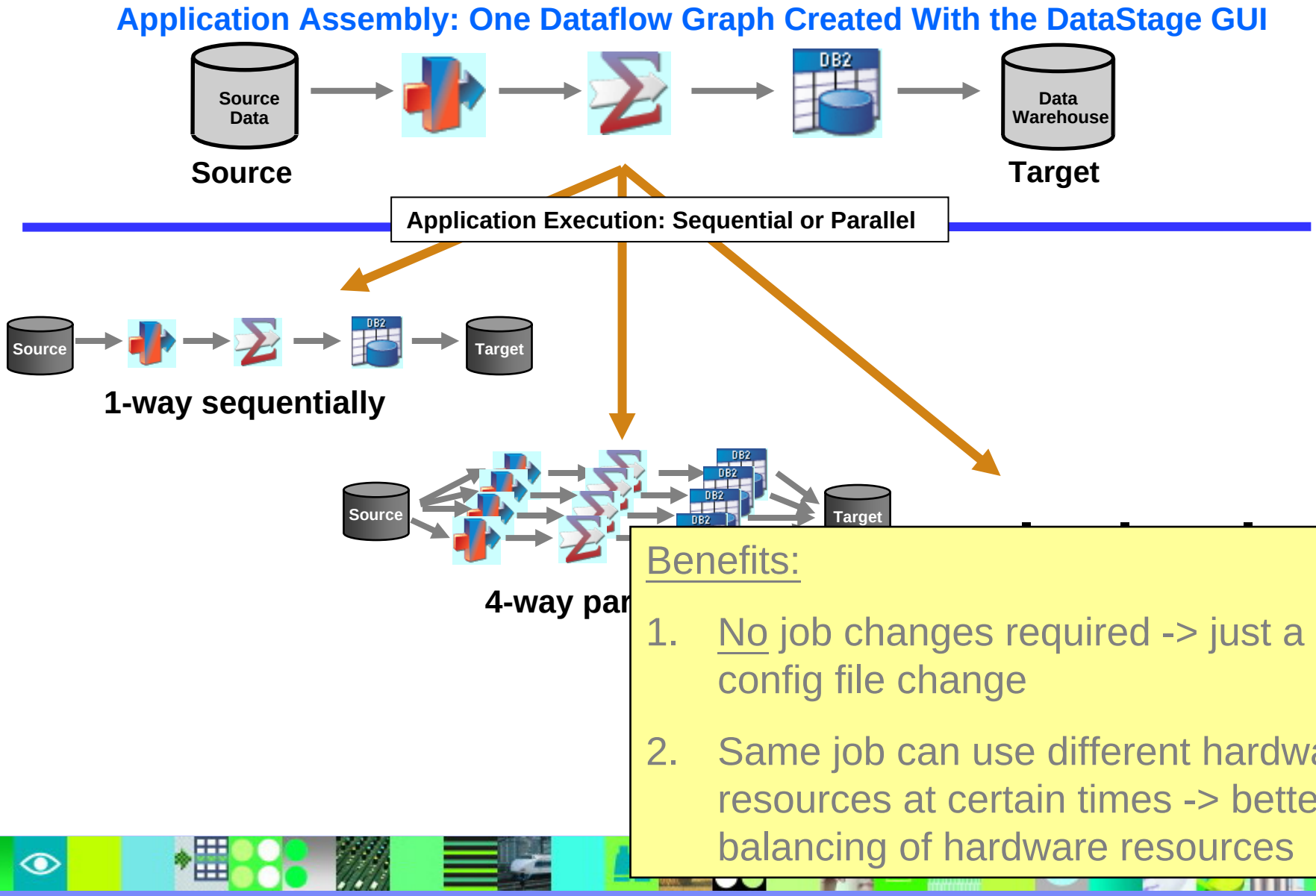
Why Enterprise Edition?

- Design sequentially, deploy in parallel
- Proven linear scalability
- Dynamic data partitioning and in-flight repartitioning of data
- Portable across SMP, Clustered, GRID, and MPP platforms
- Parallel RDBMS support, including IBM DB2, Oracle, Sybase, Informix, MS SQL server & Teradata
- Codeless parallelization
- Incorporate and parallelize existing applications into data integration process

Business Benefits

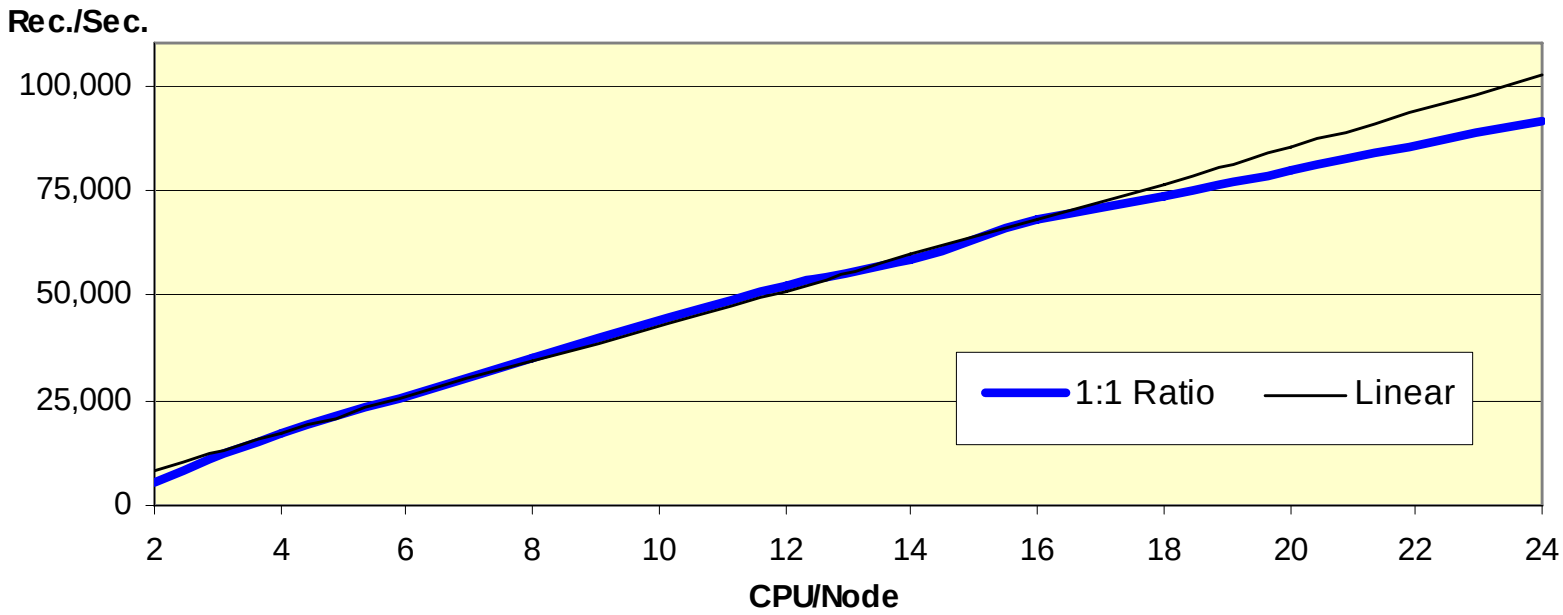
- Meet business commitments through higher productivity

有彈性的平行處理架構


Application Assembly: One Dataflow Graph Created With the DataStage GUI
Source Data
DB2
Data Warehouse
Source
Target
Application Execution: Sequential or Parallel
Source
DB2
Target
1-way sequentially
Source
DB2
DB2
DB2
DB2
Target
4-way par
Benefits:
1. No job changes required -> just a config file change
2. Same job can use different hardware resources at certain times -> better balancing of hardware resources

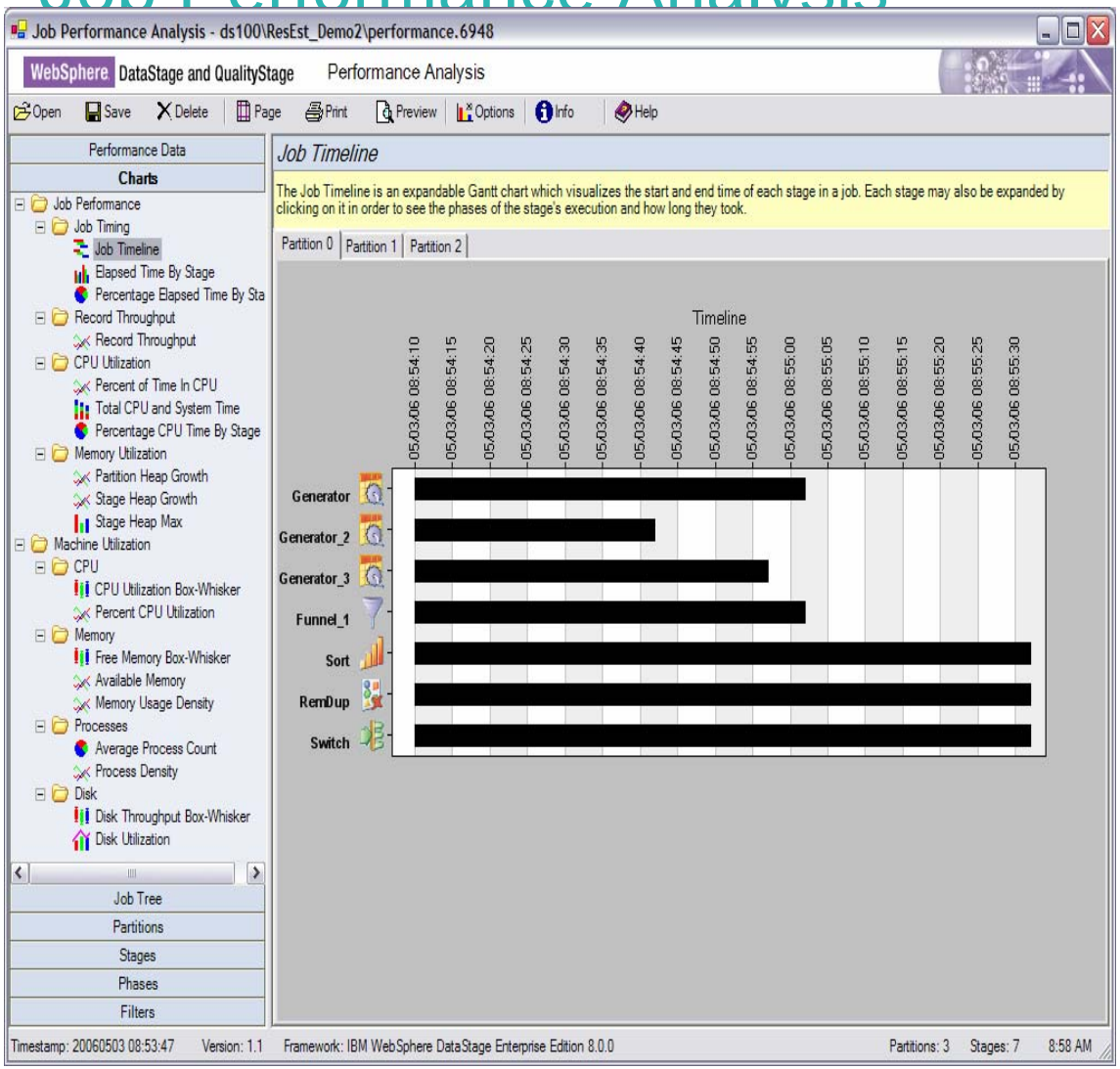
DataStage提供無限制的線性擴展能力，可隨業務成長的需求，提供可預測且倍數成長的執行效能，而不需更動ETL Job 設計

Benchmark: Scalable Data Integration Using Ascential DataStage Enterprise Edition



資料來源：**InfoSizing Performance BenchMark Report: DataStage XE Parallel Extender, Dec. 16, 2006**

Job Performance Analysis

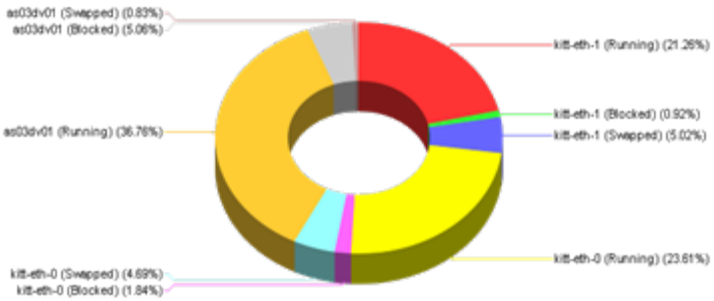


A visualization tool which:

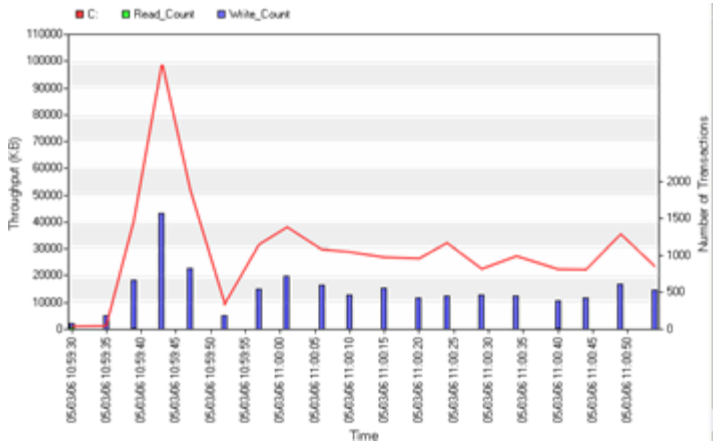
- Provides deeper insight into runtime job behavior.
- Offers several categories of visualizations, including:
  - Record Throughput
  - CPU Utilization
  - Job Timing
  - Job Memory Utilization
  - Physical Machine Utilization
- Hides runtime complexity by emphasizing the stages the customer placed on the designer canvas.

Physical Machine Utilization

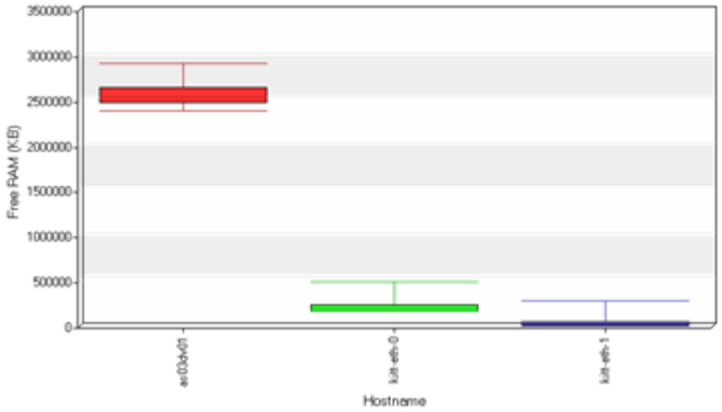
Average Process Distribution



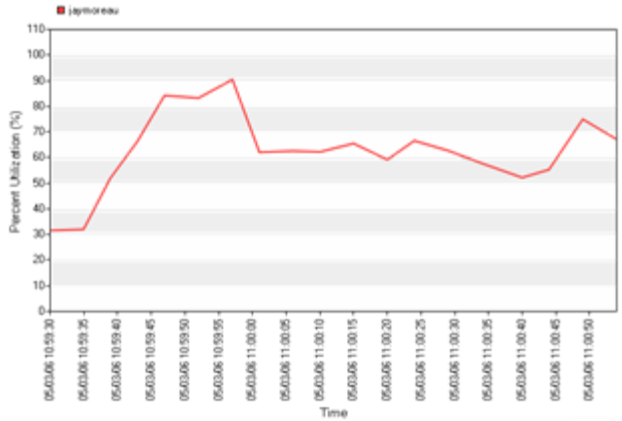
Disk Throughput



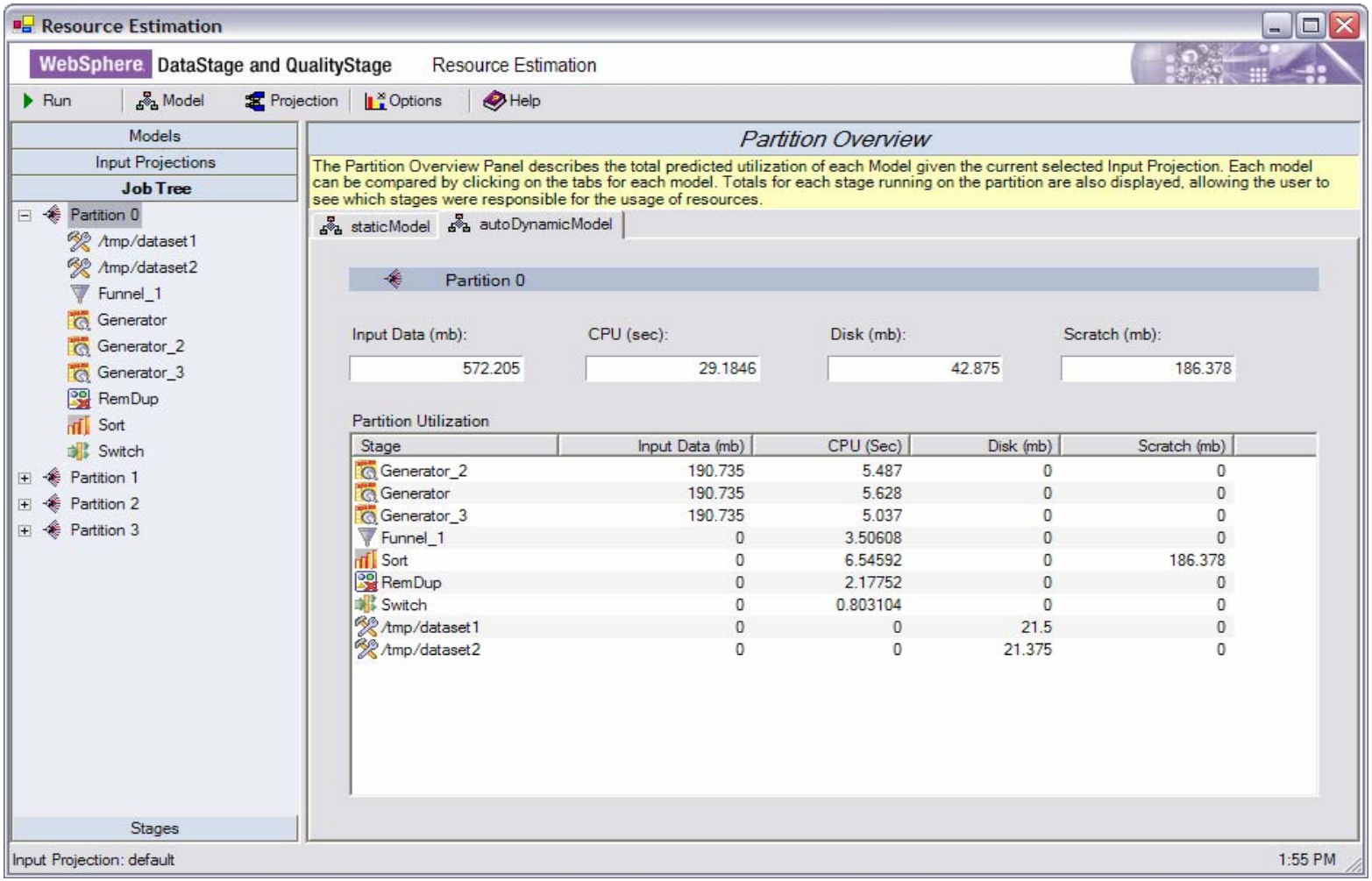
Free Memory Whisker Box



Percent CPU Utilization

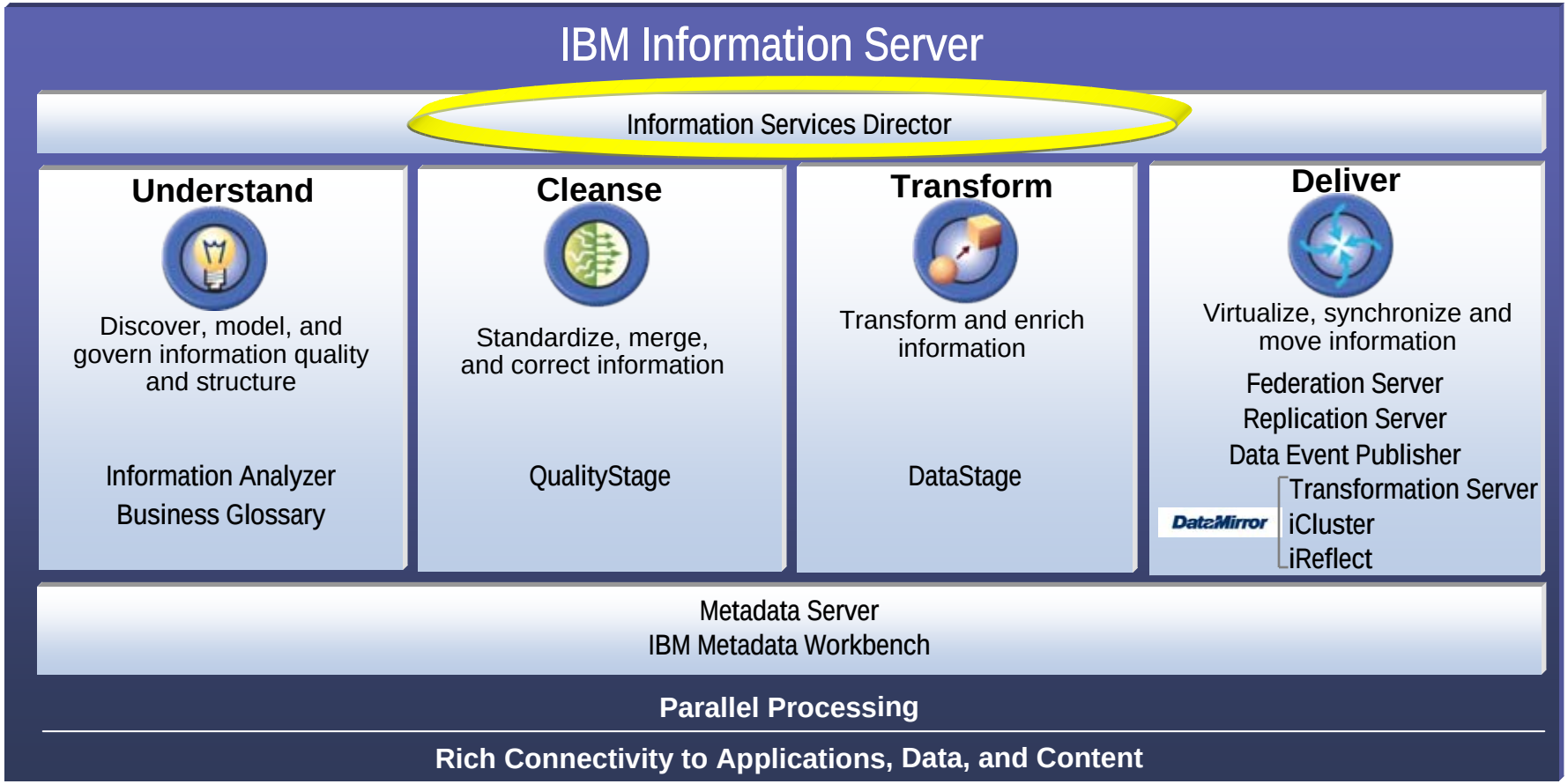


圖形化平行處理資源評估工具



IBM Information Server Offerings

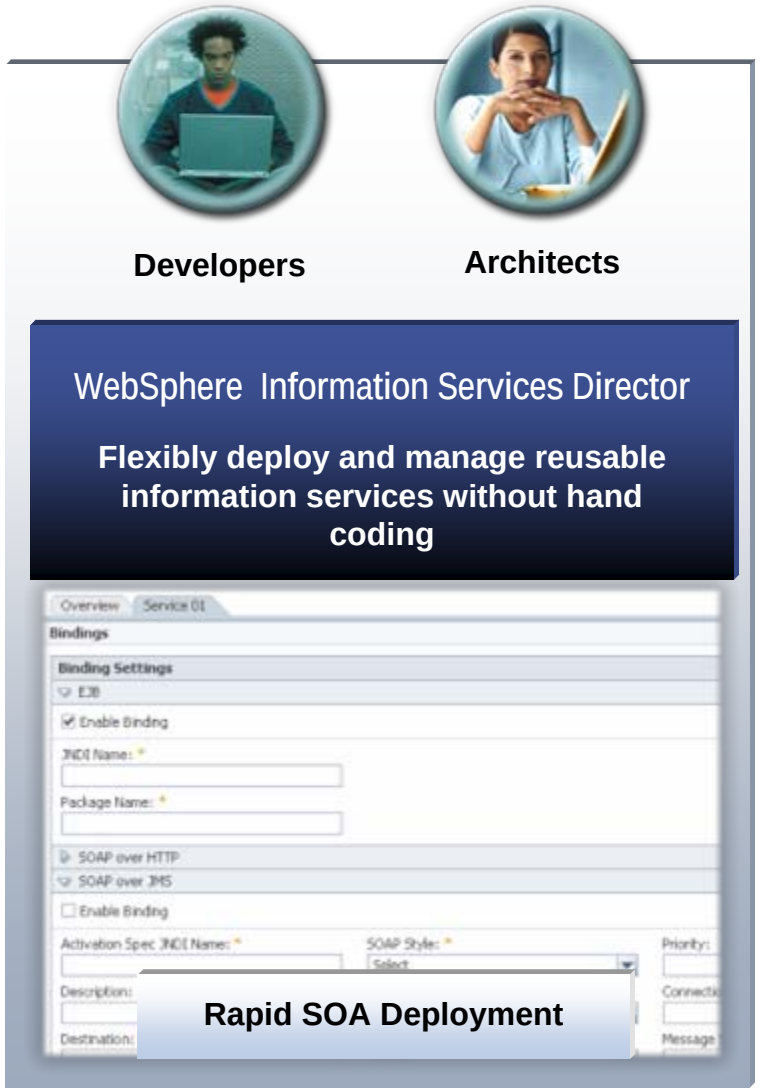
Delivering information you can trust



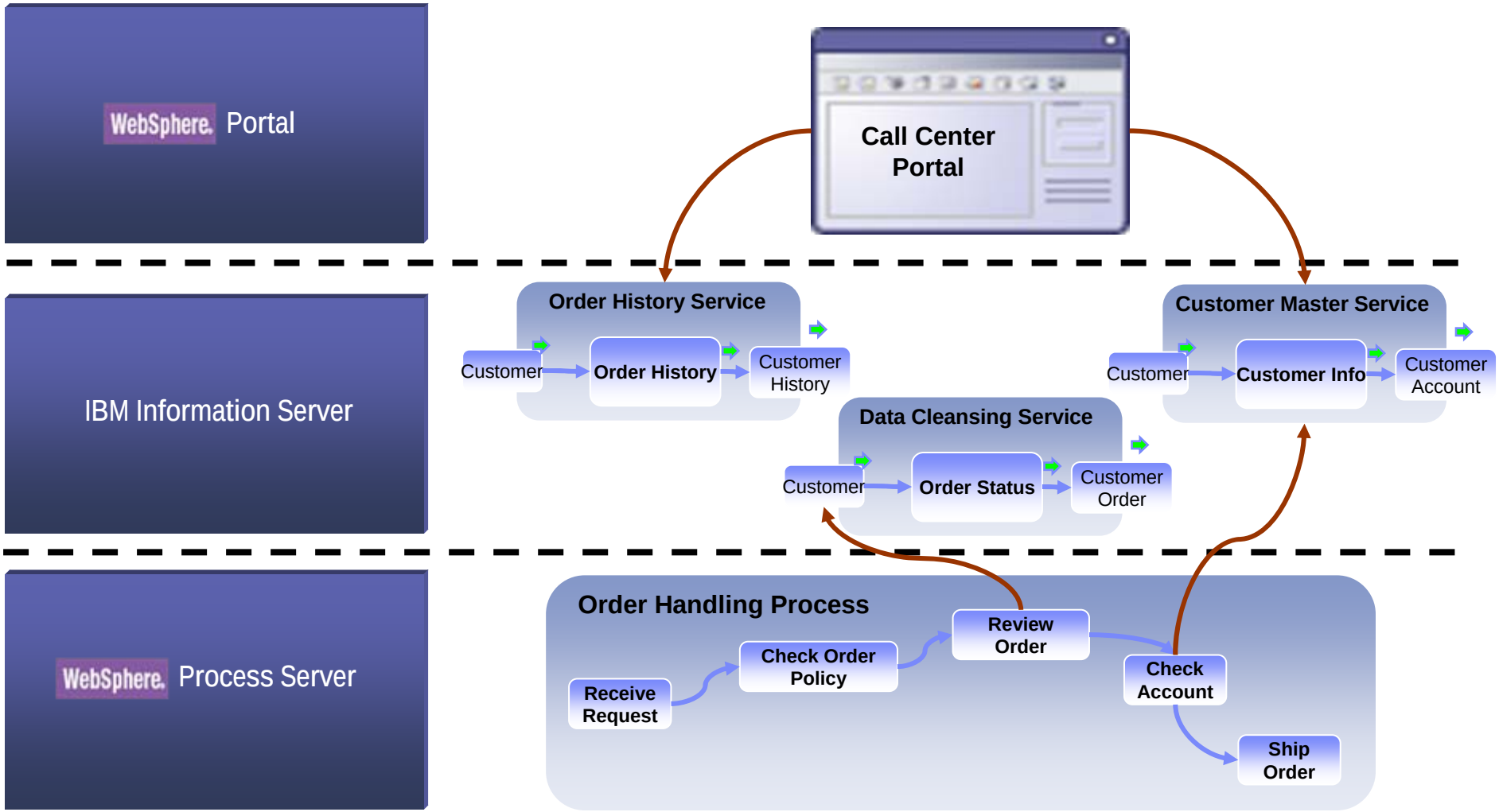
**Note: Transformation Server also implies Transformation/ES*

Rapid SOA Deployment: WebSphere Information Services Director

- Packages information integration logic as services that insulate developers from underlying sources
- Allows these services to be invoked as Enterprise Java Beans or Web services
- Provides load balancing & fault tolerance for requests across multiple Information Servers
- Provides foundation infrastructure for Information Services

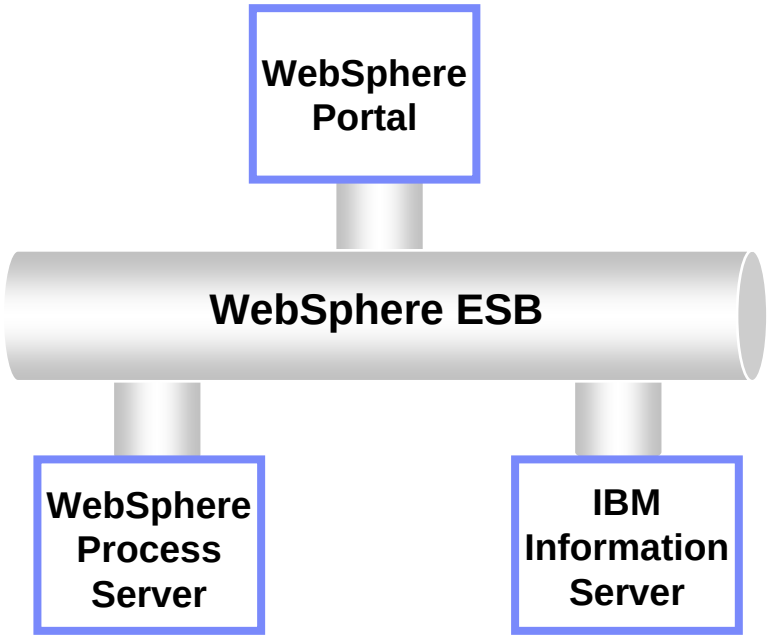


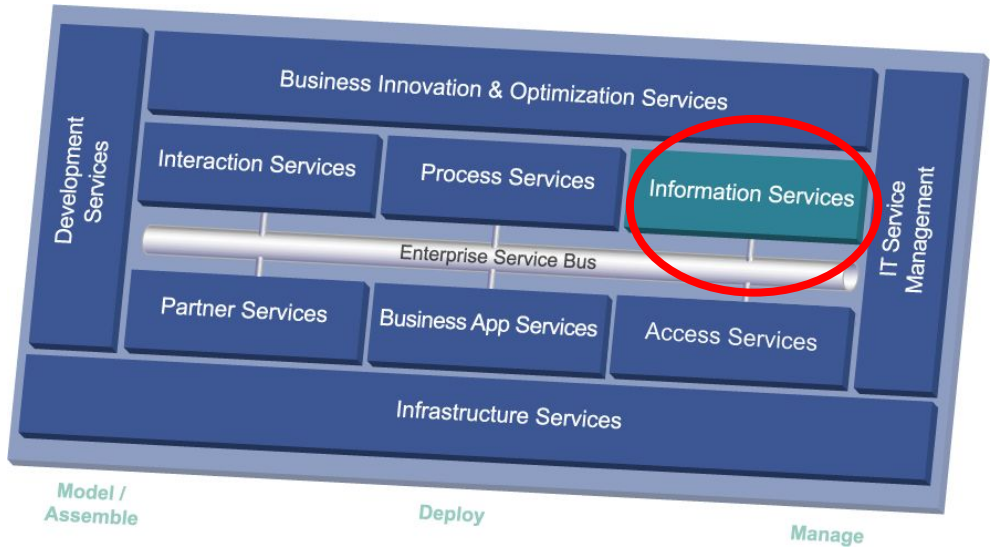
Actionable Information Services



Service Oriented Architecture
Information as a Service is Key

Getting the right data quickly and consistently for all applications continues to be a key challenge for many enterprises.
Forrester, January 2006

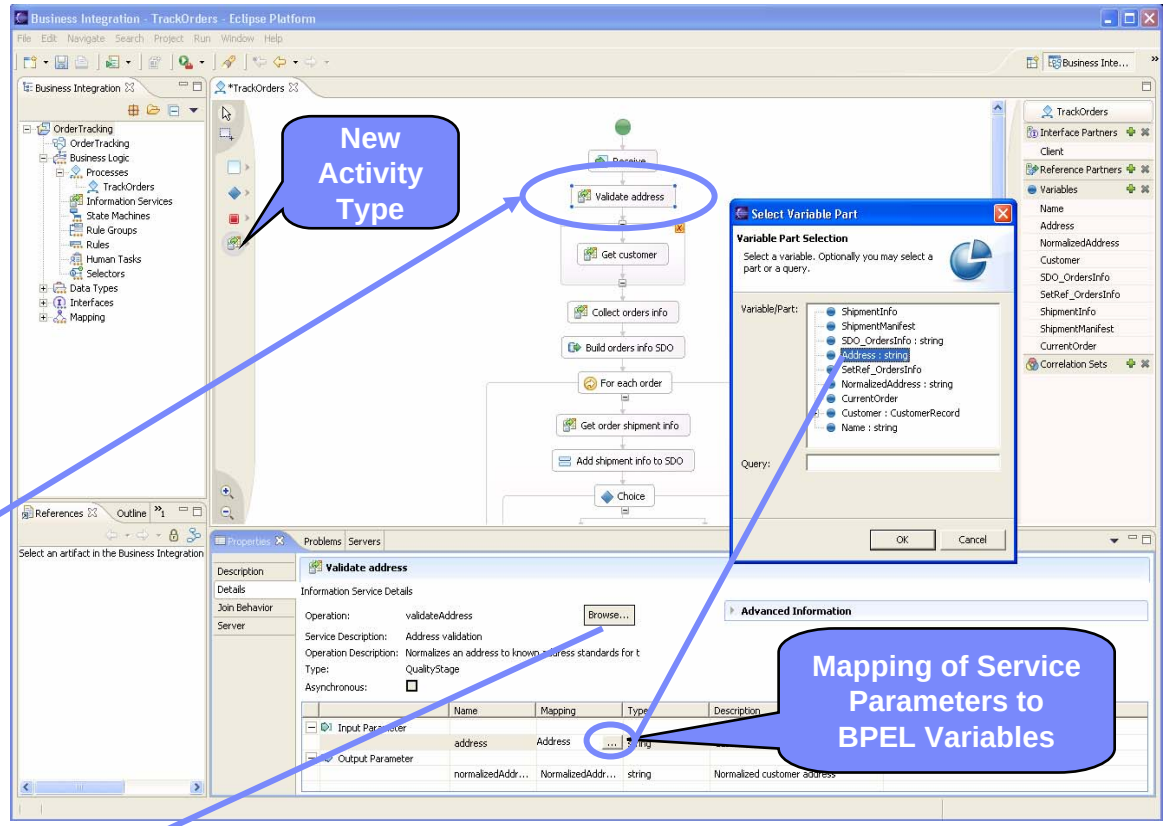




You will waste your investment in SOA unless you have enterprise information that SOA can exploit.
Gartner, March 2005

Using an Information Service in a Business Process

WebSphere Integration Developer



Compose information service as part of a broader business process

View the information service metadata from within the process development environment

Improves customer return on investment by facilitating reuse & ensures more consistent and controlled information across processes to improve results

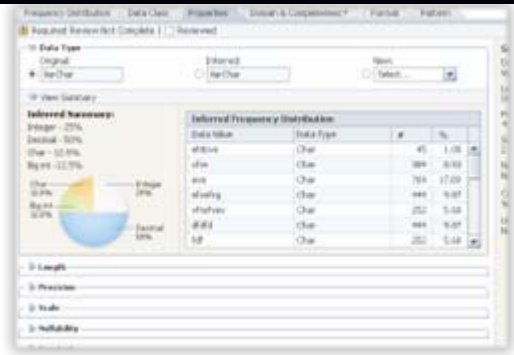
Applying Information Server to SOMA

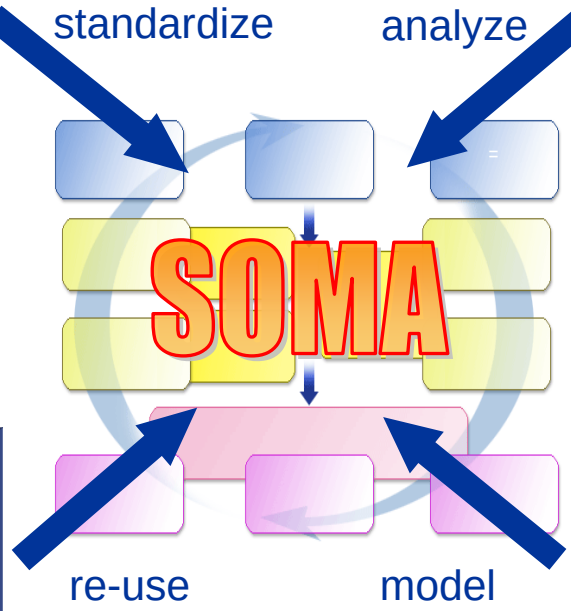
IBM WebSphere Business Glossary

Create and manage business vocabulary and relationships, while linking to physical sources

IBM WebSphere Information Analyzer

Analyze source data structures, and monitor adherence to integration and quality rules



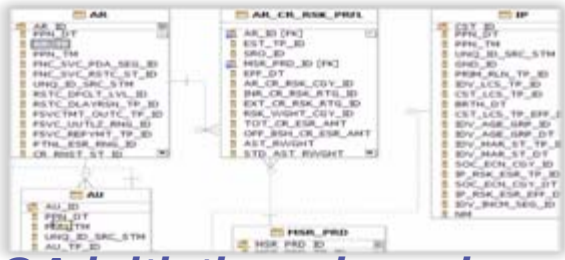


IBM Industry Models

Information models based on proven industry implementation experience.

IBM Rational Data Architect

Create conceptual, logical and physical data models



Reduces risk by incorporating information into key SOA initiatives, based on a proven methodology and best practices

全方位的SOA支援

1. Supported SOA Platform: WebSphere Application Server, BEA Weblogic Application Server, JBoss

2. Support SOAP (Simple Object Access Protocol), EJB (Enterprise Java Bean), JMS (Java Message Services) over HTTP, JMS over text

3. Both SOA services Provider and Consumer supported

4. Data Transformation and Cleansing supported

Service-Oriented Architecture (SOA)功能

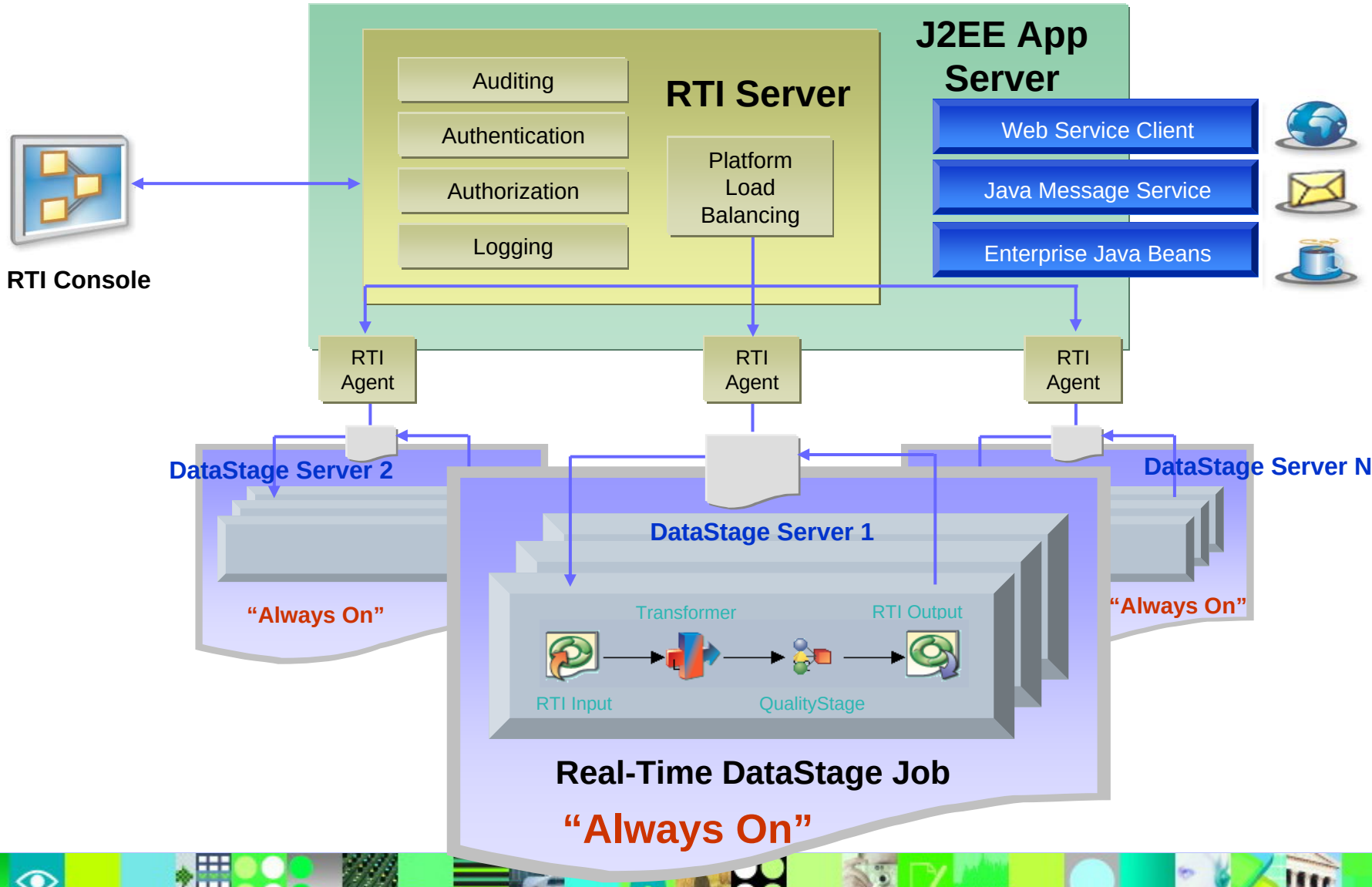


Illustration With Web Services (**Two Ways**)

Web Service Enters

Name	Value
Name ?	MR. JACK, EDWARD C.
Address1 ?	50 Washington Avenue
Address2 ?	Suite 320
City ?	Santa Clara
State ?	California

Data / Web Entry

Validated and Cleansed Result

```
Name:       MR EDWARD C JACK
Address:    50 WASHINGTON AVE STE 320
City:       SANTA CLARA
State:      CA
Postcode:   12345-1234
```

Calls Data Cleansing/ Scrubbing Web Services

Return Cleansing Data/ Scrubbing Web Services

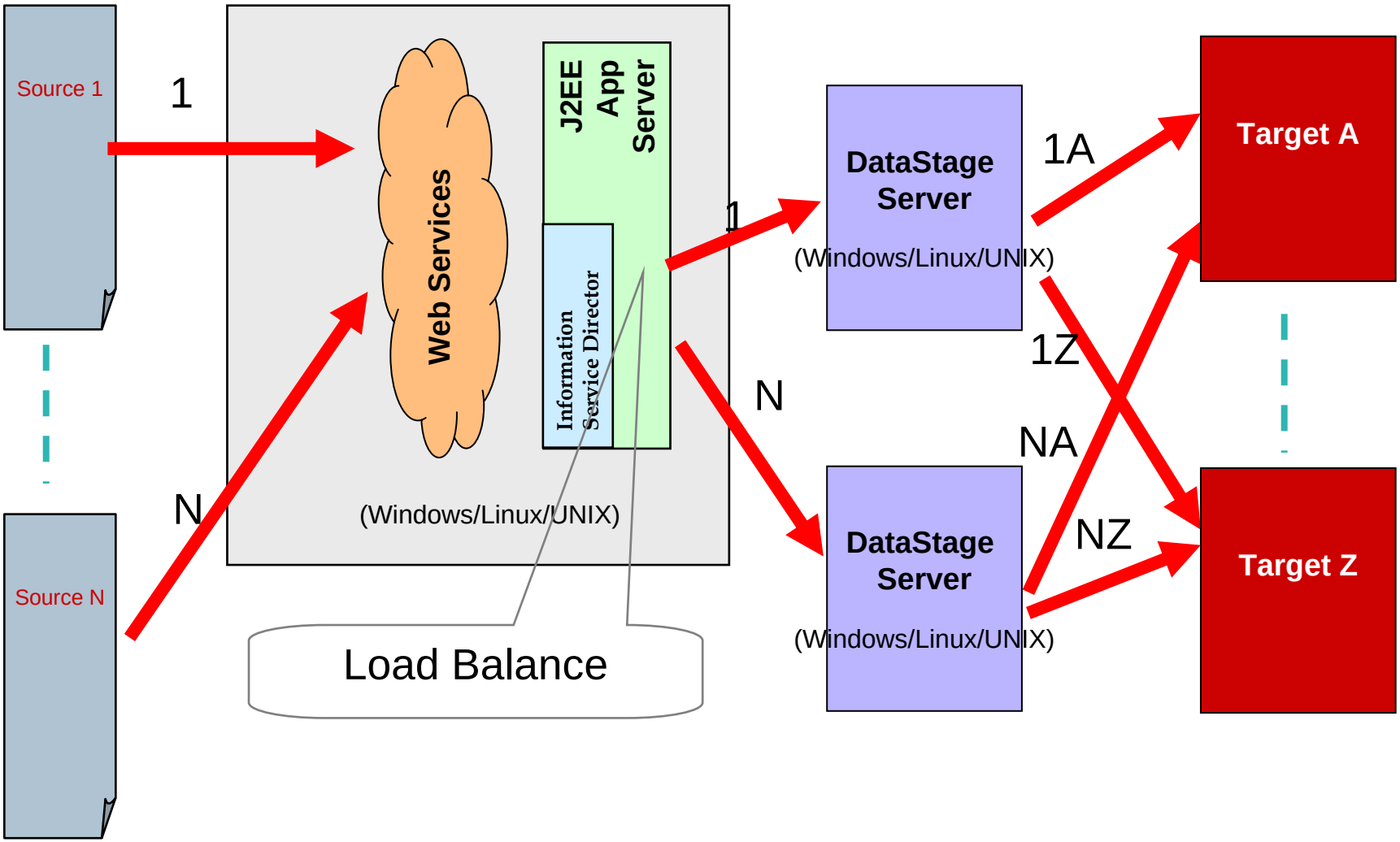
Web Services

Invokes RTI-web service As **Consumer**

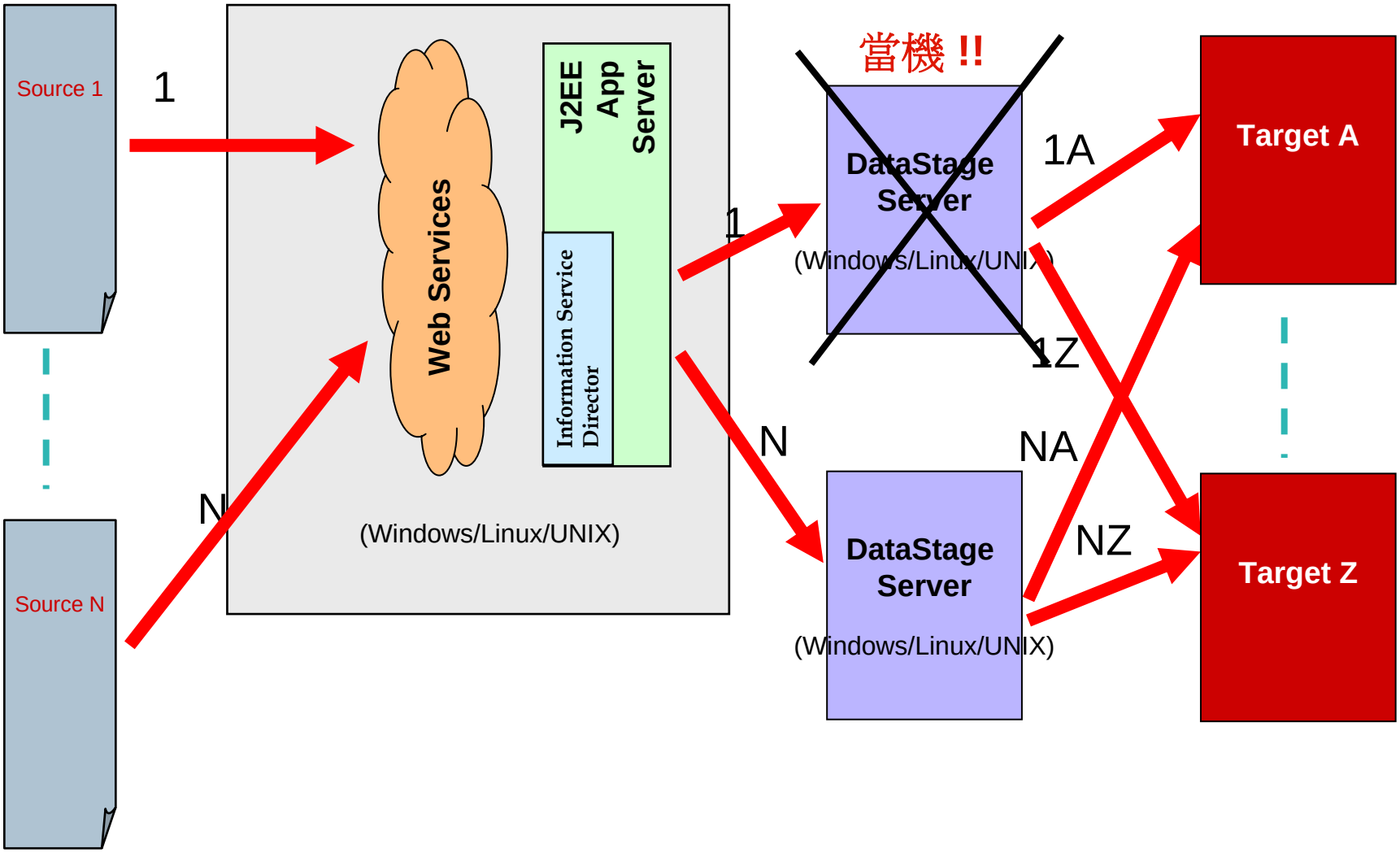
Real-Time ETL+ Data Validation / Cleansing

Invokes RTI-web service As **Provider**

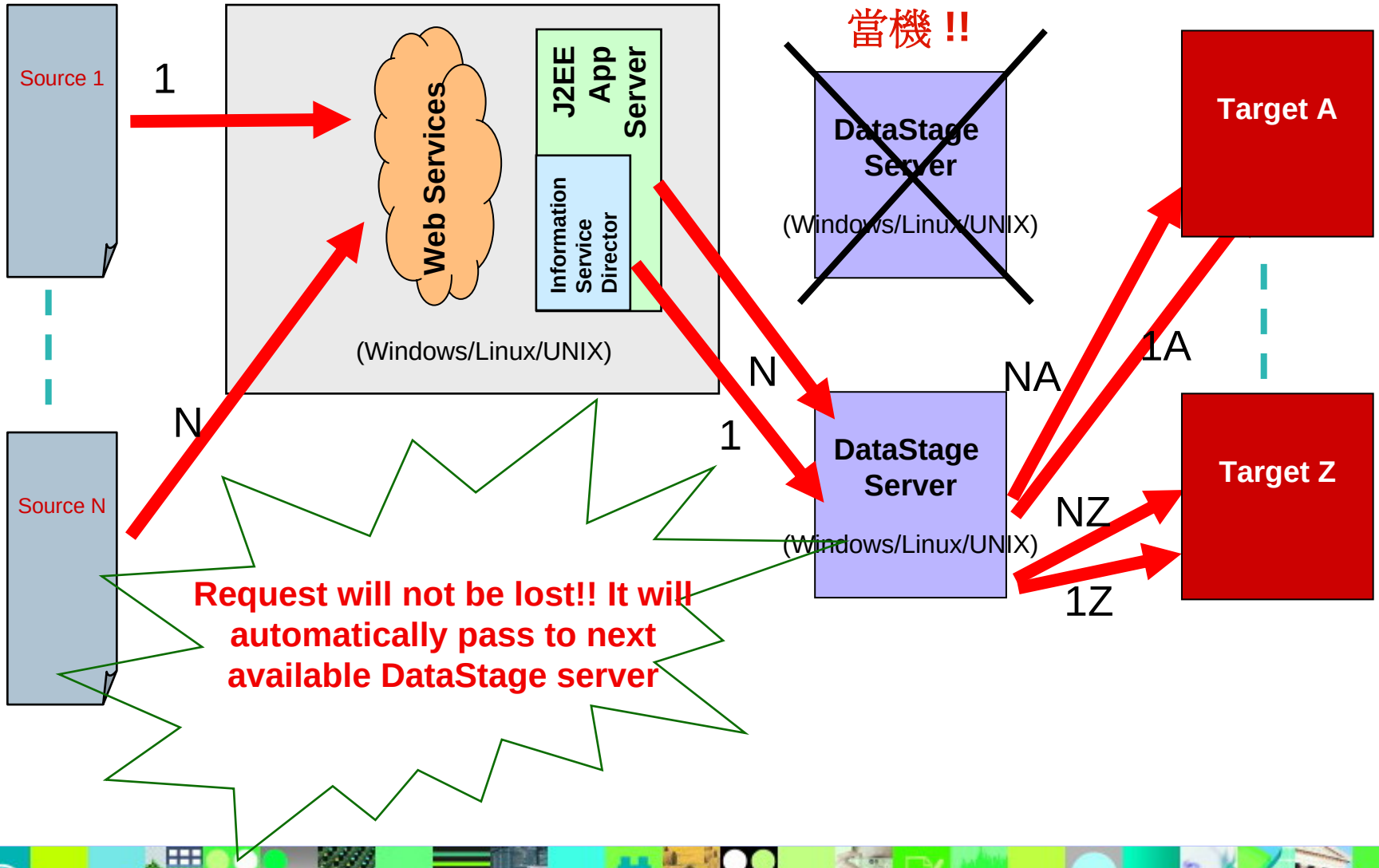
The Architecture of DataStage Server High Availability



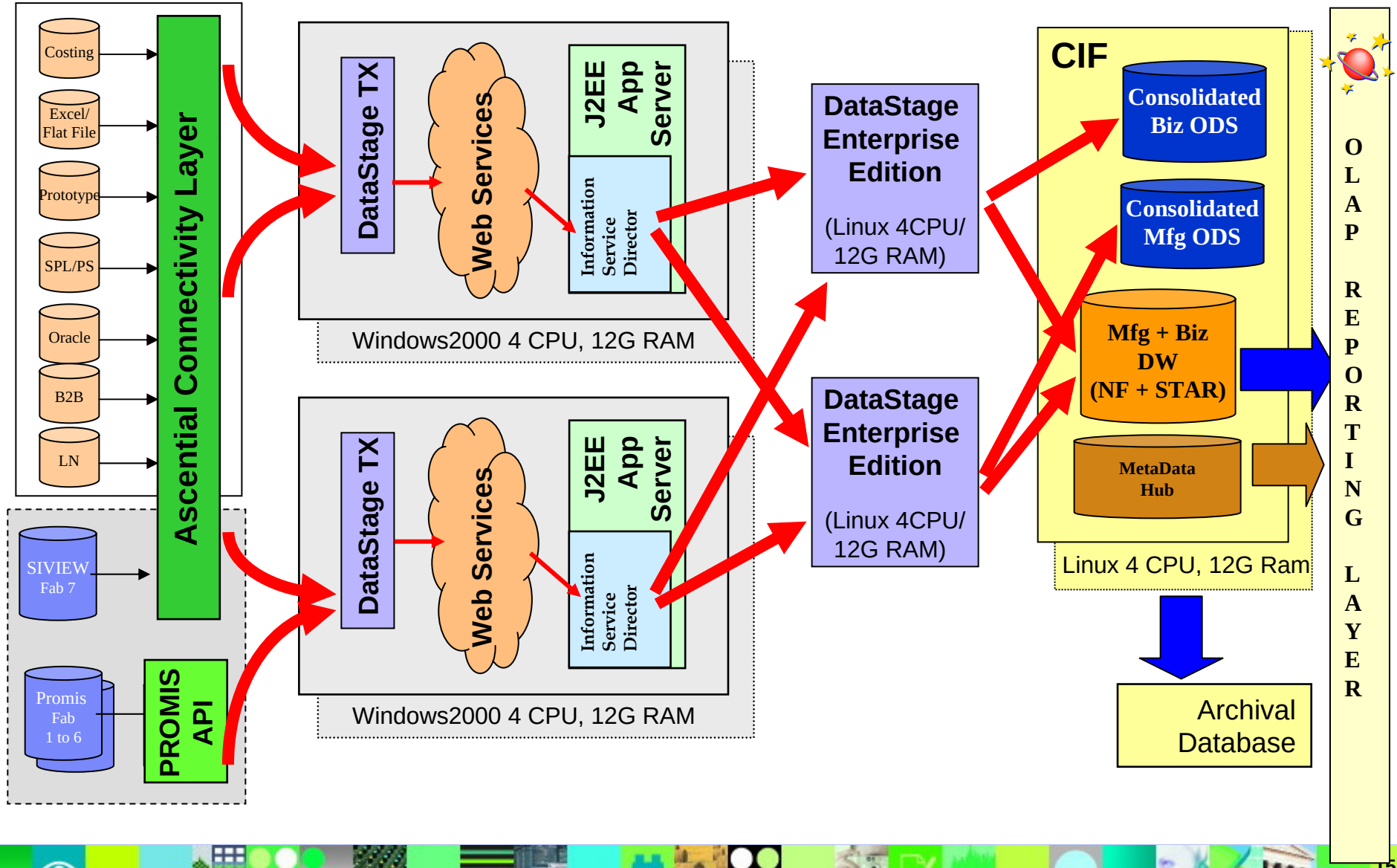
High Availability Process I



High Availability Process II

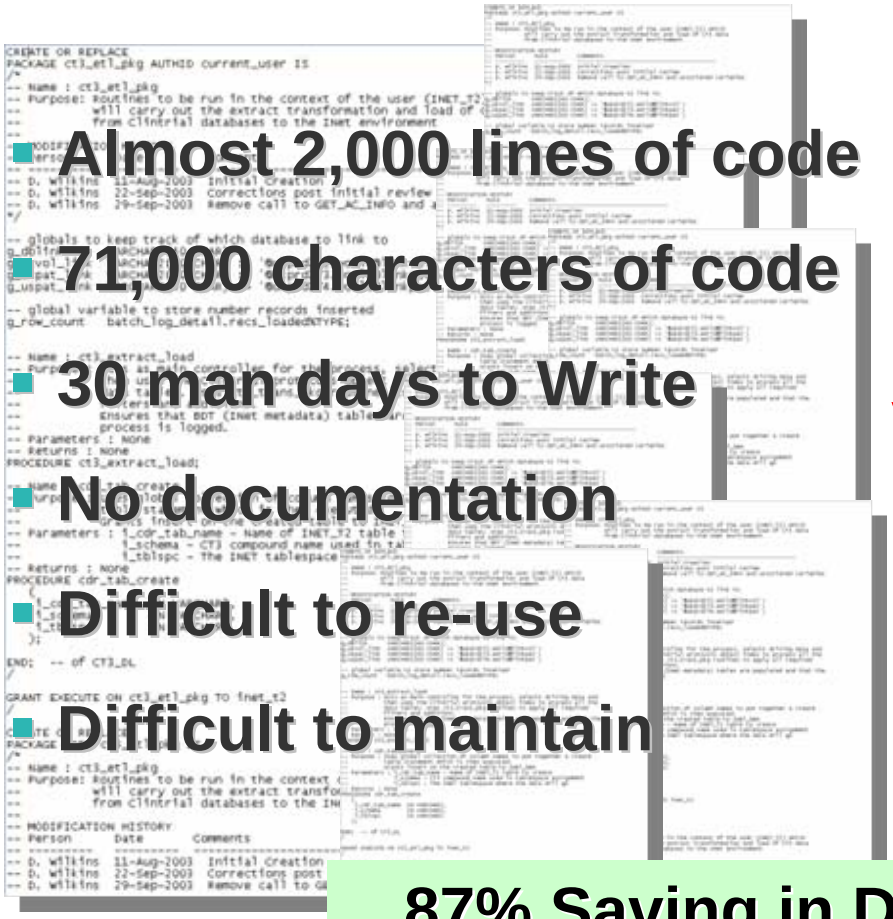


成功案例:全球第三大晶圓代工廠Chartered Semiconductor公司



DataStage 提升產能範例: Pharmaceutical data processing

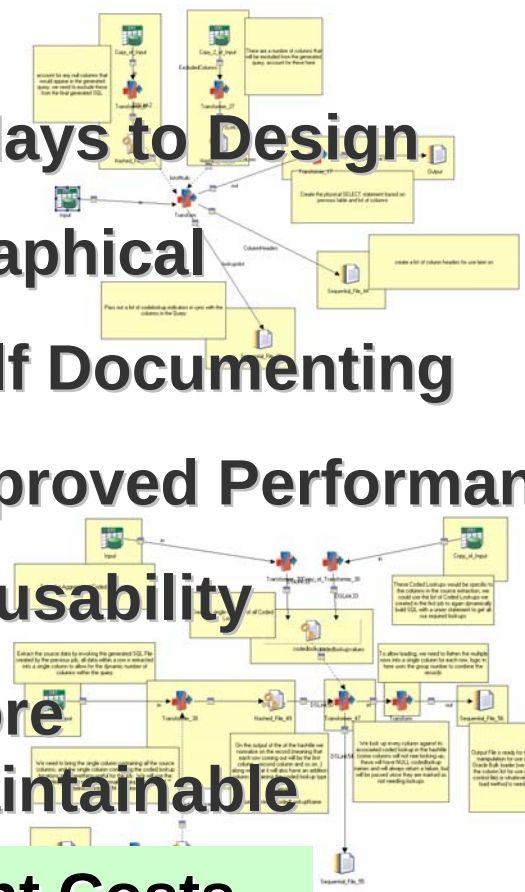
Legacy Development (Handcoding)



- Almost 2,000 lines of code
- 71,000 characters of code
- 30 man days to Write
- No documentation
- Difficult to re-use
- Difficult to maintain

versus

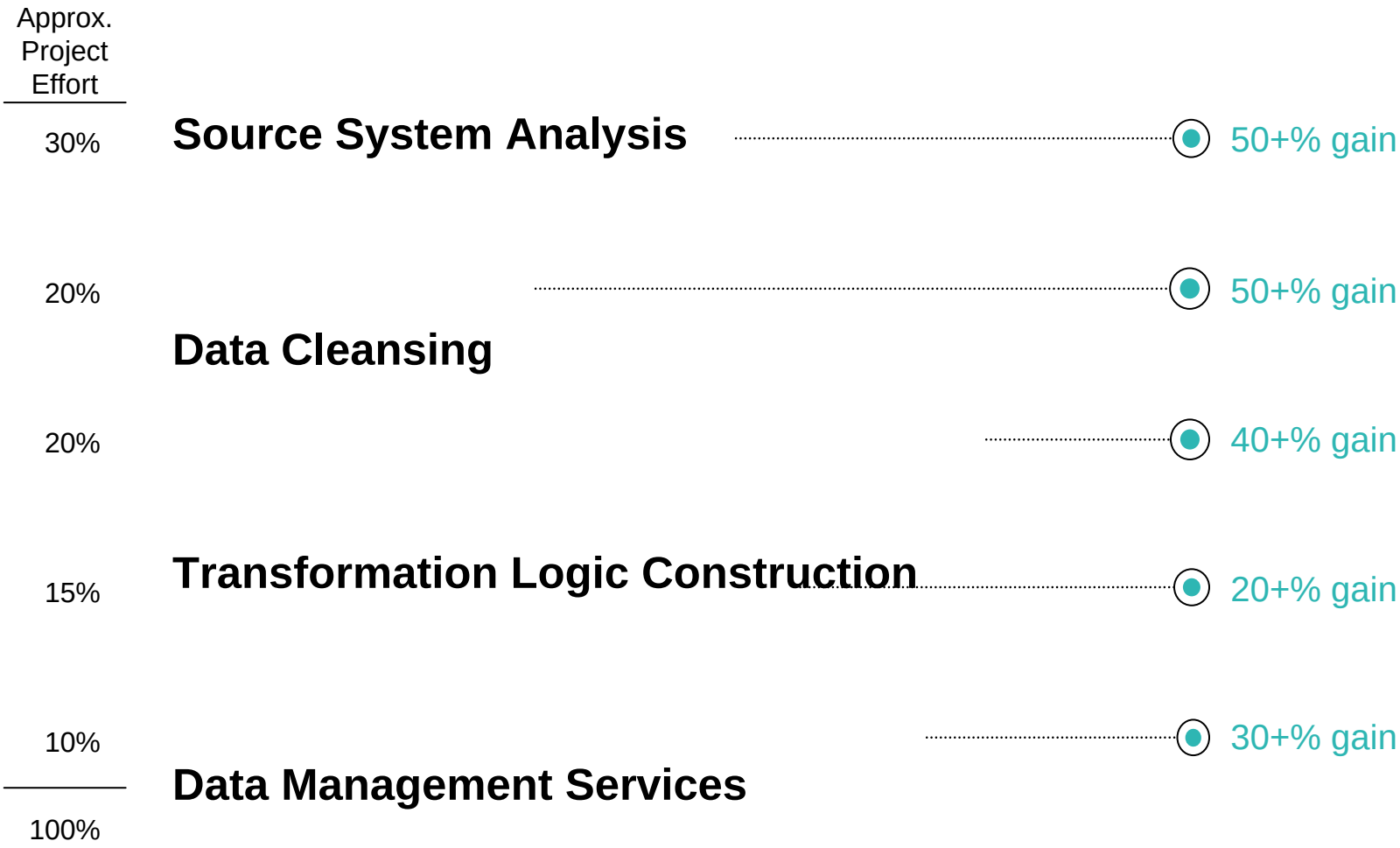
WebSphere DataStage



- 2 days to Design
- Graphical
- Self Documenting
- Improved Performance
- Reusability
- More Maintainable

87% Saving in Development Costs

IBM針對所服務過的資料整合專案所做的分析，發現相較於Hand Coding，採用DataStage可大幅提昇專案效益



資料來源：”Customers Achieve Significant Productivity Benefits--Example ETL Project”, IBM

WebSphere DataStage ROI: Retail customer

Company: A leading retail chain

Description	# Of Items	Hand Coding	Estimate	Actual Results
Technical Designs	122	867 days	15 days	
Interface Development				
Low Difficulty	35	237 days	13 days	
Medium Difficulty	28	661 days	42 days	
High Difficulty	49	936 days	147 days	
Total Effort		2,701 days	217 days	**203 Days!**

節省2498天

Global 2000 Profiting from Intelligent Information

IBM DataStage台灣建置實績(Partial)

- **金融業:** 大眾銀行、南山人壽、陽信銀行 、元大京華、台北富邦銀行、建華銀行、匯豐銀行、中華票券金融公司、荷蘭銀行、遠東國際商業銀行、摩根富林明證券投資顧問股份有限公司、聯合信用卡處理中心、中國信託、台新銀行....
- **製造業:** 台灣應材、統一企業、光寶科技、旺宏電子、中華映管(TFT 廠資料倉儲系統EDA專案等六廠) 、奇美電子、元太工業、億光電子、世界先進、華邦電子、南亞科技、友訊科技.....
- **電訊業:** 中華電信、新世紀資通、台灣大哥大、 泛亞電信、東森寬頻 ...
- **流通業:** 順發電腦、太古汽車.....
- **政府機構:** 健保局、疾病防治局、中科院、國防部國資中心、台中市政府、台北市政府自來水處、內政部戶役政、職訓局.....

Thank You

Question ?

Information Management software

Information Server Platforms

- Clients
  - Windows XP, Windows Vista
- Metadata Server (application server, repository)
  - Windows Server 2003
  - Linux
    - Red Hat Enterprise Linux AS (x86)
    - SUSE Enterprise Linux (x86, pSeries)
    - SUSE Enterprise Linux (zSeries)
  - Unix
    - AIX
    - Solaris
    - HP-UX (PA-RISC, Itanium)
- Engine
  - Windows Server 2003
  - Linux
    - Red Hat Enterprise Linux AS (x86)
    - SUSE Enterprise Linux (x86, pSeries)
    - SUSE Enterprise Linux (zSeries)
  - Unix
    - AIX
    - Solaris
    - HP-UX (PA-RISC, Itanium)
    - z/OS, Unix System Services (DataStage)