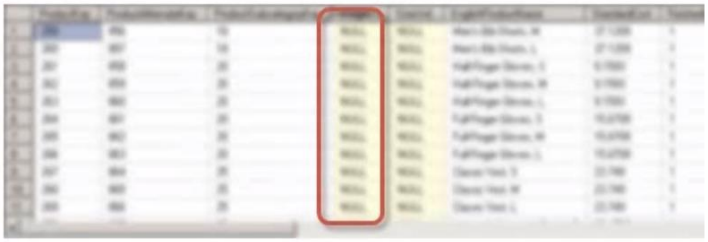
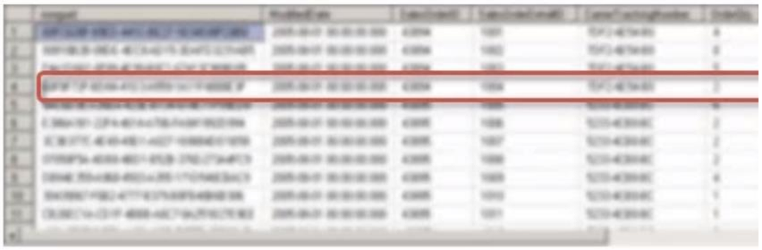
Types of Dirty Data

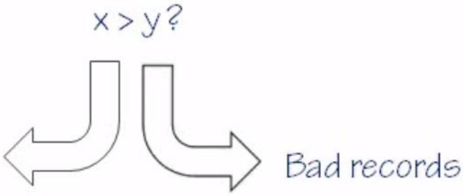
- Column Problems



- Record Problems

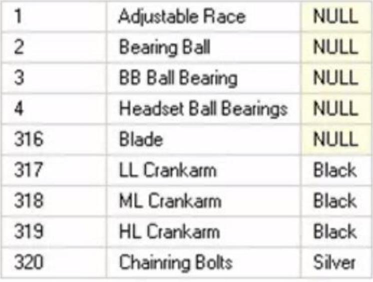


- Business Rule Problems



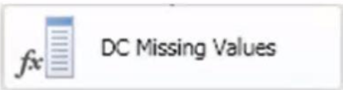
Column Problems: Missing Data Default

- Problem: Missing Data

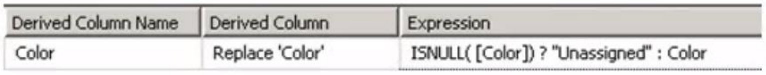


1	Adjustable Race	NULL
2	Bearing Ball	NULL
3	BB Ball Bearing	NULL
4	Headset Ball Bearings	NULL
316	Blade	NULL
317	LL Crankarm	Black
318	ML Crankarm	Black
319	HL Crankarm	Black
320	Chainring Bolts	Silver

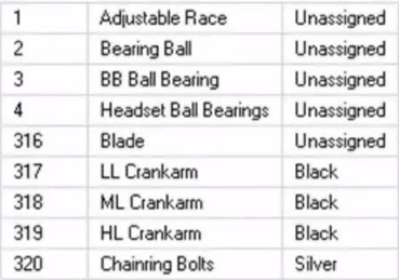
- Solution: Add Default



DC Missing Values



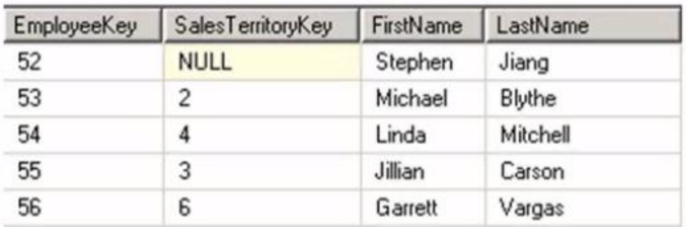
Derived Column Name	Derived Column	Expression
Color	Replace 'Color'	ISNULL([Color]) ? "Unassigned" : Color



1	Adjustable Race	Unassigned
2	Bearing Ball	Unassigned
3	BB Ball Bearing	Unassigned
4	Headset Ball Bearings	Unassigned
316	Blade	Unassigned
317	LL Crankarm	Black
318	ML Crankarm	Black
319	HL Crankarm	Black
320	Chainring Bolts	Silver

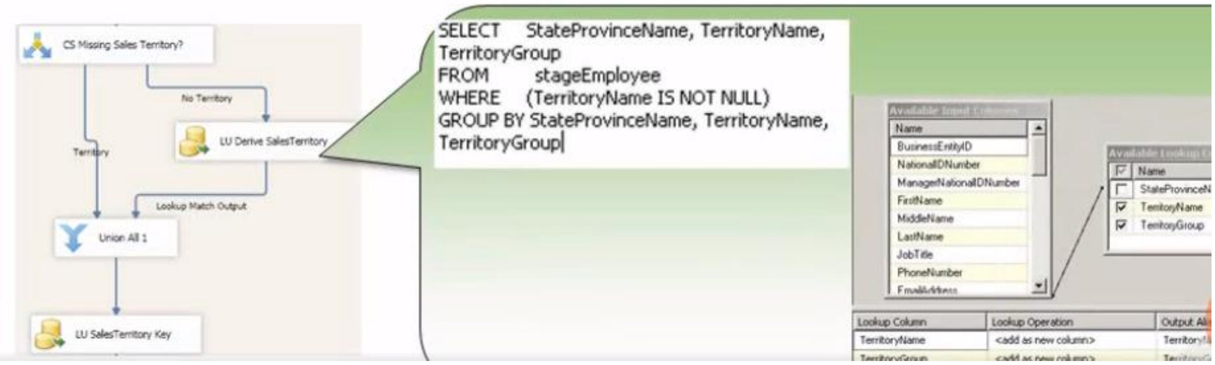
Column Problems: Derive Missing Data

- Problem: Missing Data



EmployeeKey	SalesTerritoryKey	FirstName	LastName
52	NULL	Stephen	Jiang
53	2	Michael	Blythe
54	4	Linda	Mitchell
55	3	Jillian	Carson
56	6	Garrett	Vargas

- Solution: Use Lookup Query to Derive Data

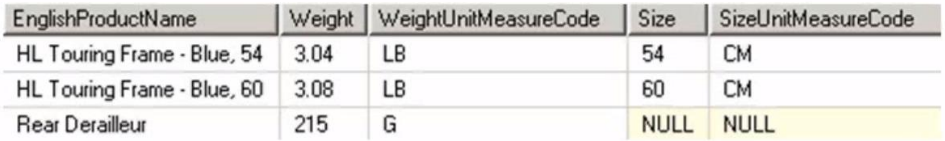


```
SELECT    StateProvinceName, TerritoryName,
TerritoryGroup
FROM      stageEmployee
WHERE     (TerritoryName IS NOT NULL)
GROUP BY StateProvinceName, TerritoryName,
TerritoryGroup
```

Lookup Column	Lookup Operation	Output Al
TerritoryName	<add as new column>	Territory
TerritoryGroup	<add as new column>	TerritoryG

Column Problems: Translation

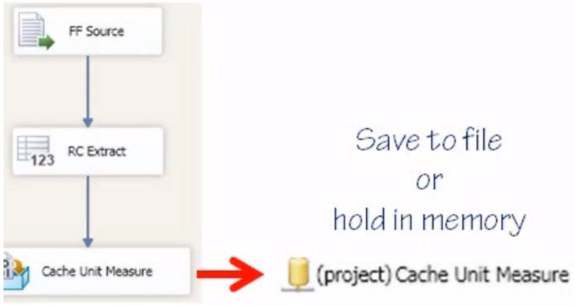
- Problem: Code to Translate



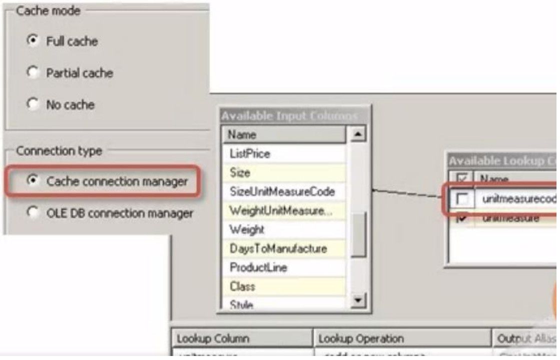
EnglishProductName	Weight	WeightUnitMeasureCode	Size	SizeUnitMeasureCode
HL Touring Frame - Blue, 54	3.04	LB	54	CM
HL Touring Frame - Blue, 60	3.08	LB	60	CM
Rear Derailleur	215	G	NULL	NULL

- Solution: Lookup (optionally with Cache Transform)

1. Extract Step



2. Lookup Step

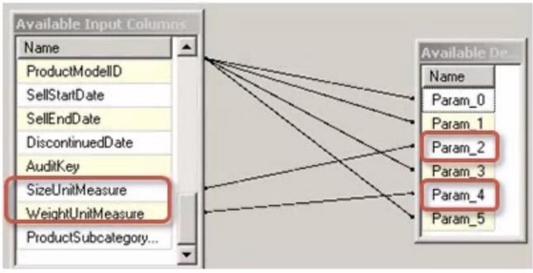


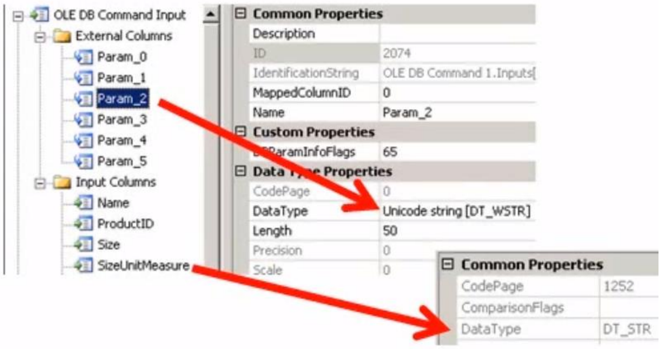
Cache mode
● Full cache
○ Partial cache
○ No cache

Connection type
● Cache connection manager
○ OLE DB connection manager

Column Problems: Data Type

- Problem: Incompatible Data Types

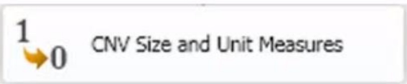


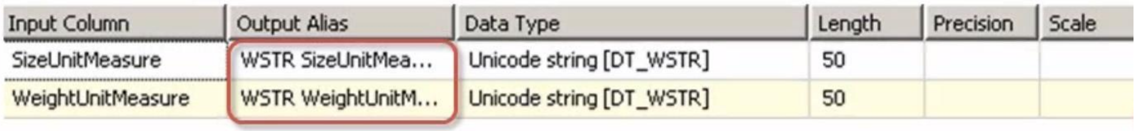


Common Properties	
Description	
ID	2074
IdentificationString	OLE DB Command 1.Inputs[
MappedColumnID	0
Name	Param_2
Custom Properties	
DBParamInfoFlags	65
Data Type Properties	
CodePage	0
DataType	Unicode string [DT_WSTR]
Length	50
Precision	0
Scale	0

Common Properties	
CodePage	1252
ComparisonFlags	
DataType	DT_STR

- Solution: Data Conversion

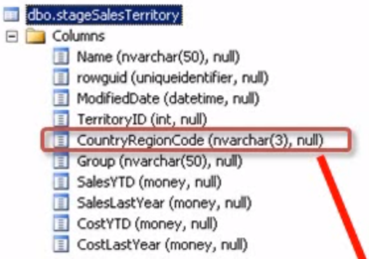




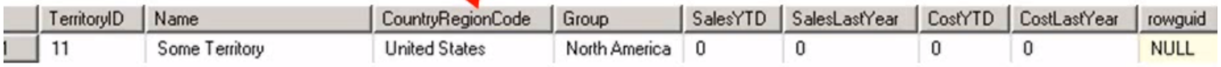
Input Column	Output Alias	Data Type	Length	Precision	Scale
SizeUnitMeasure	WSTR SizeUnitMea...	Unicode string [DT_WSTR]	50		
WeightUnitMeasure	WSTR WeightUnitM...	Unicode string [DT_WSTR]	50		

Column Problems: Truncation

- Problem: Input Column Size > Destination Column Size

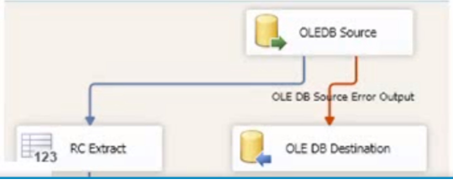


```
dbo.stageSalesTerritory
  Columns
    Name (nvarchar(50), null)
    rowguid (uniqueidentifier, null)
    ModifiedDate (datetime, null)
    TerritoryID (int, null)
    CountryRegionCode (nvarchar(3), null)
    Group (nvarchar(50), null)
    SalesYTD (money, null)
    SalesLastYear (money, null)
    CostYTD (money, null)
    CostLastYear (money, null)
```

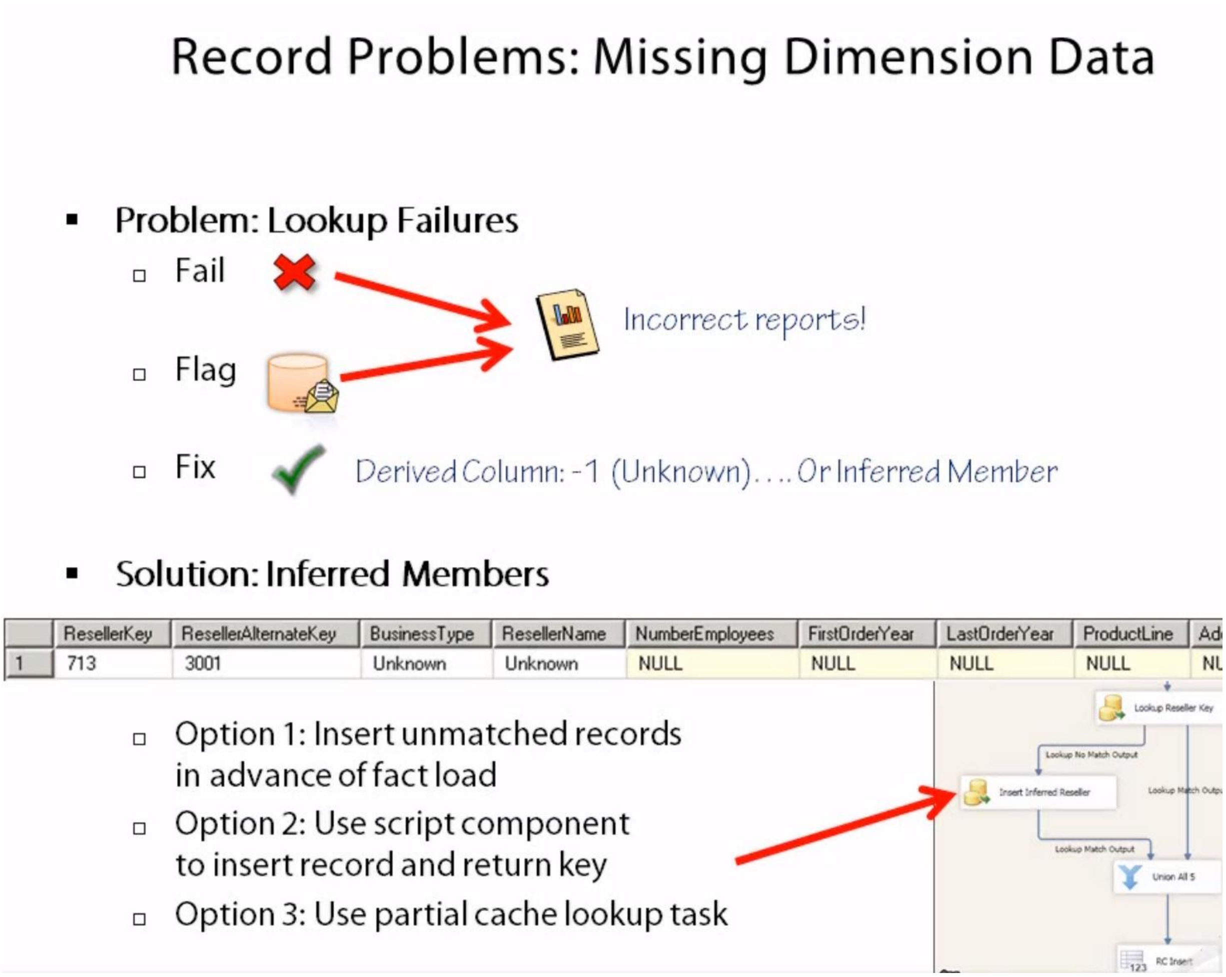


	TerritoryID	Name	CountryRegionCode	Group	SalesYTD	SalesLastYear	CostYTD	CostLastYear	rowguid
1	11	Some Territory	United States	North America	0	0	0	0	NULL

- Solution 1: Derived Column with Left() or Substring()
  - But…Consider the impact on downstream operations
- Solution 2: Flag the record for manual intervention



Implement process to manage these errors!



Record Problems: Missing Dimension Data

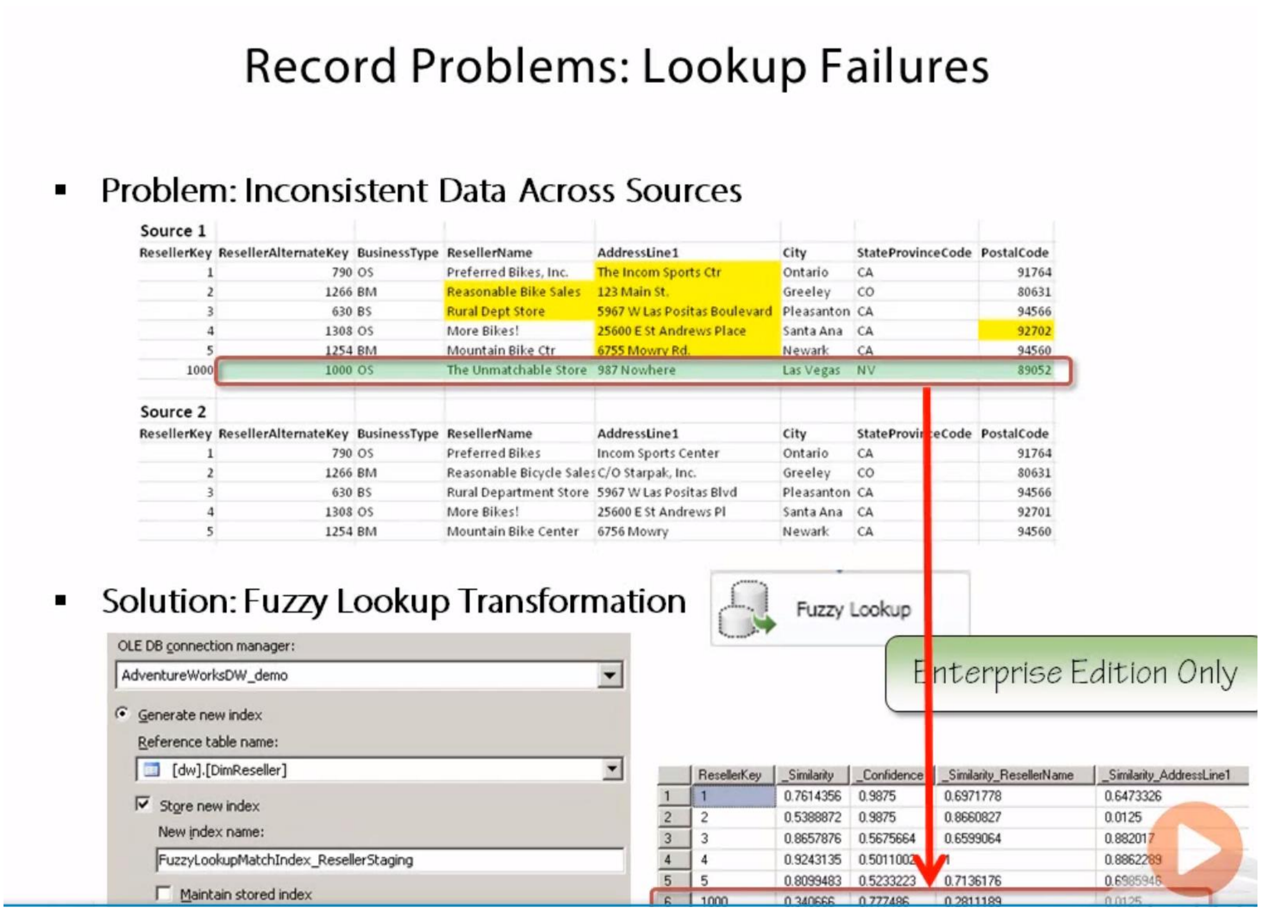
- Problem: Lookup Failures
  - □ Fail ✖
  - □ Flag
  - □ Fix ✓ Derived Column: -1 (Unknown)….Or Inferred Member

Incorrect reports!

- Solution: Inferred Members

	ResellerKey	ResellerAlternateKey	BusinessType	ResellerName	NumberEmployees	FirstOrderYear	LastOrderYear	ProductLine	Ad
1	713	3001	Unknown	Unknown	NULL	NULL	NULL	NULL	NU

  - □ Option 1: Insert unmatched records in advance of fact load
  - □ Option 2: Use script component to insert record and return key
  - □ Option 3: Use partial cache lookup task



Record Problems: Lookup Failures

- Problem: Inconsistent Data Across Sources

Source 1

ResellerKey	ResellerAlternateKey	BusinessType	ResellerName	AddressLine1	City	StateProvinceCode	PostalCode
1	790	OS	Preferred Bikes, Inc.	The Incom Sports Ctr	Ontario	CA	91764
2	1266	BM	Reasonable Bike Sales	123 Main St.	Greeley	CO	80631
3	630	BS	Rural Dept Store	5967 W Las Positas Boulevard	Pleasanton	CA	94566
4	1308	OS	More Bikes!	25600 E St Andrews Place	Santa Ana	CA	92702
5	1254	BM	Mountain Bike Ctr	6755 Mowry Rd.	Newark	CA	94560
1000	1000	OS	The Unmatchable Store	987 Nowhere	Las Vegas	NV	89052

Source 2

ResellerKey	ResellerAlternateKey	BusinessType	ResellerName	AddressLine1	City	StateProvinceCode	PostalCode
1	790	OS	Preferred Bikes	Incom Sports Center	Ontario	CA	91764
2	1266	BM	Reasonable Bicycle Sales	C/O Starpak, Inc.	Greeley	CO	80631
3	630	BS	Rural Department Store	5967 W Las Positas Blvd	Pleasanton	CA	94566
4	1308	OS	More Bikes!	25600 E St Andrews Pl	Santa Ana	CA	92701
5	1254	BM	Mountain Bike Center	6756 Mowry	Newark	CA	94560

- Solution: Fuzzy Lookup Transformation

Fuzzy Lookup

Enterprise Edition Only

OLE DB connection manager: AdventureWorksDW_demo

● Generate new index

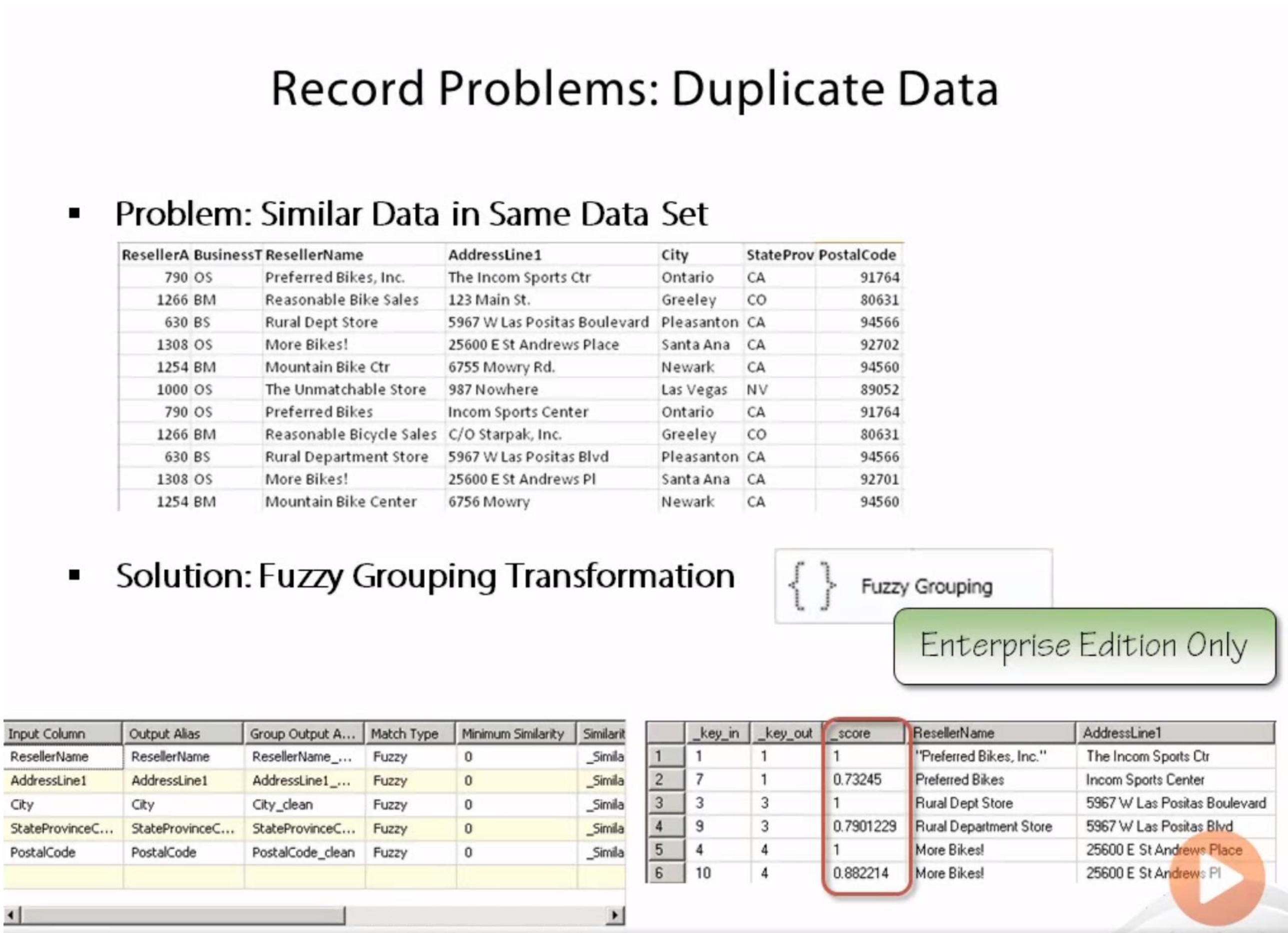
Reference table name: [dw].[DimReseller]

☑ Store new index

New index name: FuzzyLookupMatchIndex_ResellerStaging

□ Maintain stored index

	ResellerKey	_Similarity	_Confidence	_Similarity_ResellerName	_Similarity_AddressLine1
1	1	0.7614356	0.9875	0.6971778	0.6473326
2	2	0.5388872	0.9875	0.8660827	0.0125
3	3	0.8657876	0.5675664	0.6599064	0.882017
4	4	0.9243135	0.5011002	1	0.8862289
5	5	0.8099483	0.5233223	0.7136176	0.6985946
6	1000	0.340666	0.777486	0.2811189	0.0125



Record Problems: Duplicate Data

- Problem: Similar Data in Same Data Set

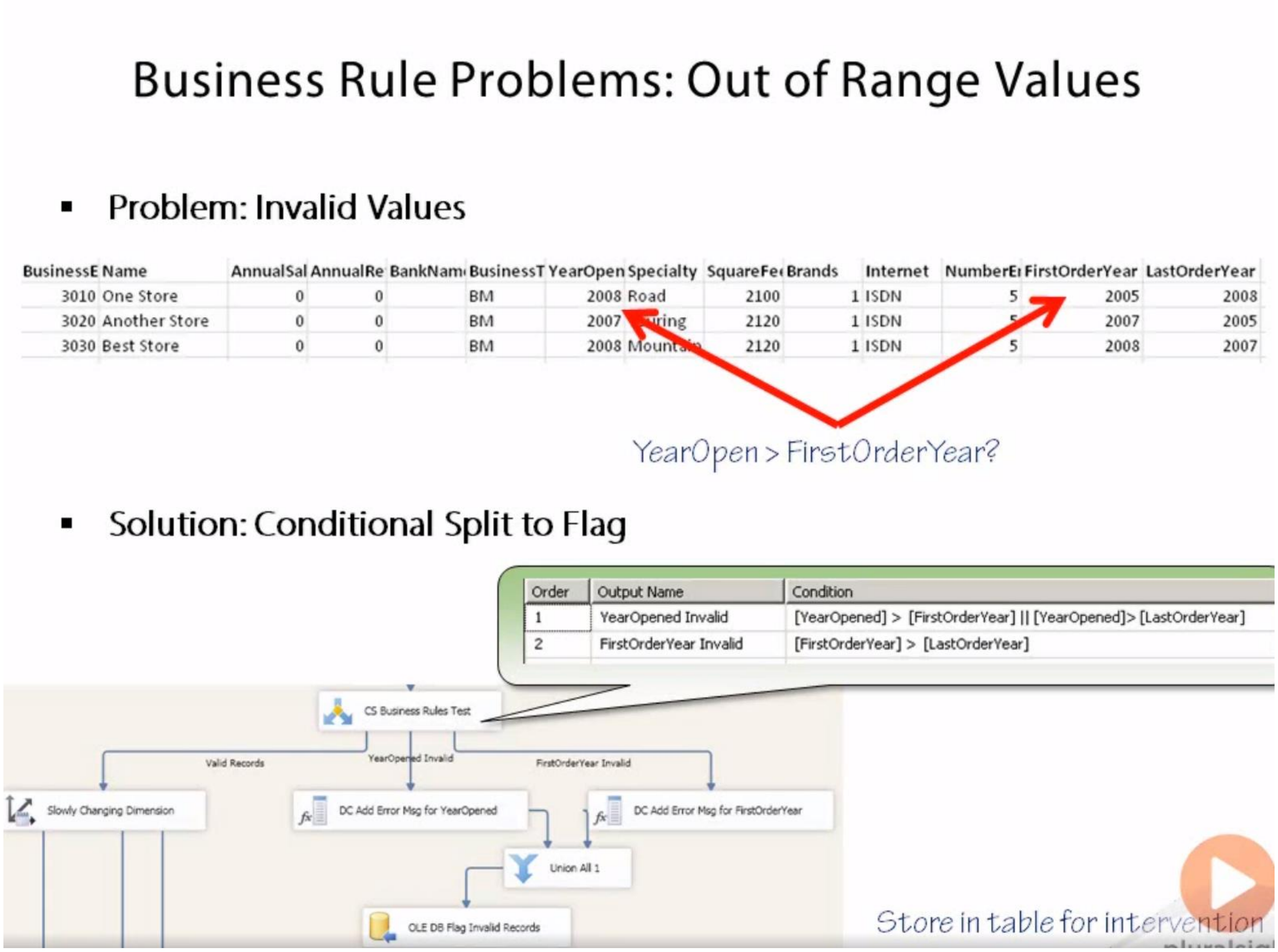
ResellerA	BusinessT	ResellerName	AddressLine1	City	StateProv	PostalCode
790	OS	Preferred Bikes, Inc.	The Incom Sports Ctr	Ontario	CA	91764
1266	BM	Reasonable Bike Sales	123 Main St.	Greeley	CO	80631
630	BS	Rural Dept Store	5967 W Las Positas Boulevard	Pleasanton	CA	94566
1308	OS	More Bikes!	25600 E St Andrews Place	Santa Ana	CA	92702
1254	BM	Mountain Bike Ctr	6755 Mowry Rd.	Newark	CA	94560
1000	OS	The Unmatchable Store	987 Nowhere	Las Vegas	NV	89052
790	OS	Preferred Bikes	Incom Sports Center	Ontario	CA	91764
1266	BM	Reasonable Bicycle Sales	C/O Starpak, Inc.	Greeley	CO	80631
630	BS	Rural Department Store	5967 W Las Positas Blvd	Pleasanton	CA	94566
1308	OS	More Bikes!	25600 E St Andrews Pl	Santa Ana	CA	92701
1254	BM	Mountain Bike Center	6756 Mowry	Newark	CA	94560

- Solution: Fuzzy Grouping Transformation

Enterprise Edition Only

Input Column	Output Alias	Group Output A...	Match Type	Minimum Similarity	Similarit
ResellerName	ResellerName	ResellerName_...	Fuzzy	0	_Simila
AddressLine1	AddressLine1	AddressLine1_...	Fuzzy	0	_Simila
City	City	City_clean	Fuzzy	0	_Simila
StateProvinceC...	StateProvinceC...	StateProvinceC...	Fuzzy	0	_Simila
PostalCode	PostalCode	PostalCode_clean	Fuzzy	0	_Simila

	_key_in	_key_out	_score	ResellerName	AddressLine1
1	1	1	1	"Preferred Bikes, Inc."	The Incom Sports Ctr
2	7	1	0.73245	Preferred Bikes	Incom Sports Center
3	3	3	1	Rural Dept Store	5967 W Las Positas Boulevard
4	9	3	0.7901229	Rural Department Store	5967 W Las Positas Blvd
5	4	4	1	More Bikes!	25600 E St Andrews Place
6	10	4	0.882214	More Bikes!	25600 E St Andrews Pl



Business Rule Problems: Out of Range Values

- Problem: Invalid Values

BusinessE	Name	AnnualSal	AnnualRe	BankNam	BusinessT	YearOpen	Specialty	SquareFe	Brands	Internet	NumberE	FirstOrderYear	LastOrderYear
3010	One Store	0	0		BM	2008	Road	2100	1	ISDN	5	2005	2008
3020	Another Store	0	0		BM	2007	Touring	2120	1	ISDN	5	2007	2005
3030	Best Store	0	0		BM	2008	Mountain	2120	1	ISDN	5	2008	2007

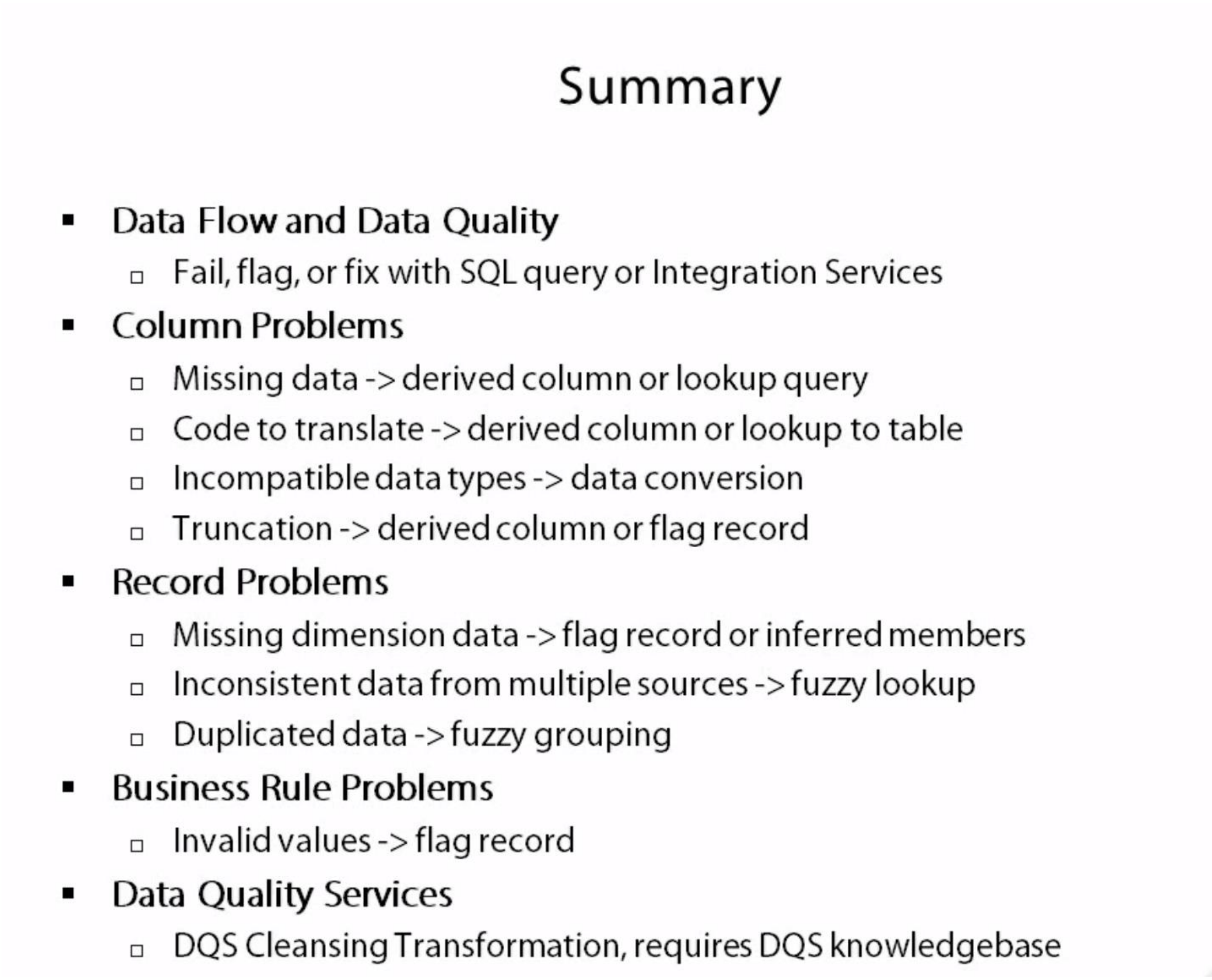
YearOpen > FirstOrderYear?

- Solution: Conditional Split to Flag

Order	Output Name	Condition
1	YearOpened Invalid	[YearOpened] > [FirstOrderYear] || [YearOpened]> [LastOrderYear]
2	FirstOrderYear Invalid	[FirstOrderYear] > [LastOrderYear]

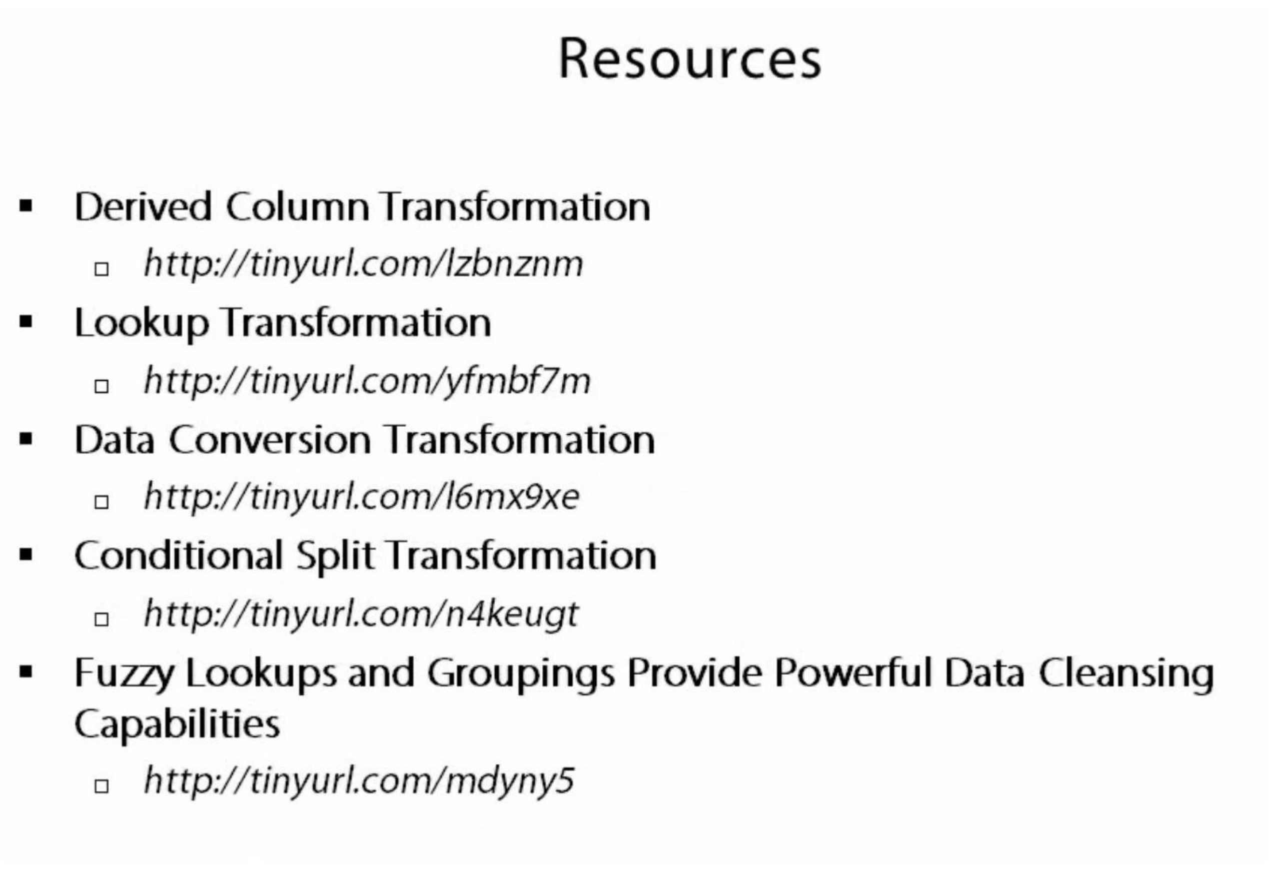
CS Business Rules Test
Valid Records
YearOpened Invalid
FirstOrderYear Invalid
Slowly Changing Dimension
DC Add Error Msg for YearOpened
DC Add Error Msg for FirstOrderYear
Union All 1
OLE DB Flag Invalid Records

Store in table for intervention



Summary

- Data Flow and Data Quality
  - Fail, flag, or fix with SQL query or Integration Services
- Column Problems
  - Missing data -> derived column or lookup query
  - Code to translate -> derived column or lookup to table
  - Incompatible data types -> data conversion
  - Truncation -> derived column or flag record
- Record Problems
  - Missing dimension data -> flag record or inferred members
  - Inconsistent data from multiple sources -> fuzzy lookup
  - Duplicated data -> fuzzy grouping
- Business Rule Problems
  - Invalid values -> flag record
- Data Quality Services
  - DQS Cleansing Transformation, requires DQS knowledgebase



Resources

- Derived Column Transformation
  - *http://tinyurl.com/lzbnznm*
- Lookup Transformation
  - *http://tinyurl.com/yfmbf7m*
- Data Conversion Transformation
  - *http://tinyurl.com/l6mx9xe*
- Conditional Split Transformation
  - *http://tinyurl.com/n4keugt*
- Fuzzy Lookups and Groupings Provide Powerful Data Cleansing Capabilities
  - *http://tinyurl.com/mdyny5*