



The MIT Information Quality Industry Symposium, 2007



# Demonstrations of linguistic data matching, consolidation, and cleansing

**Jeff Fried**  
**VP Advanced Solutions**  
**[Jeff.Fried@fastsearch.com](mailto:Jeff.Fried@fastsearch.com)**



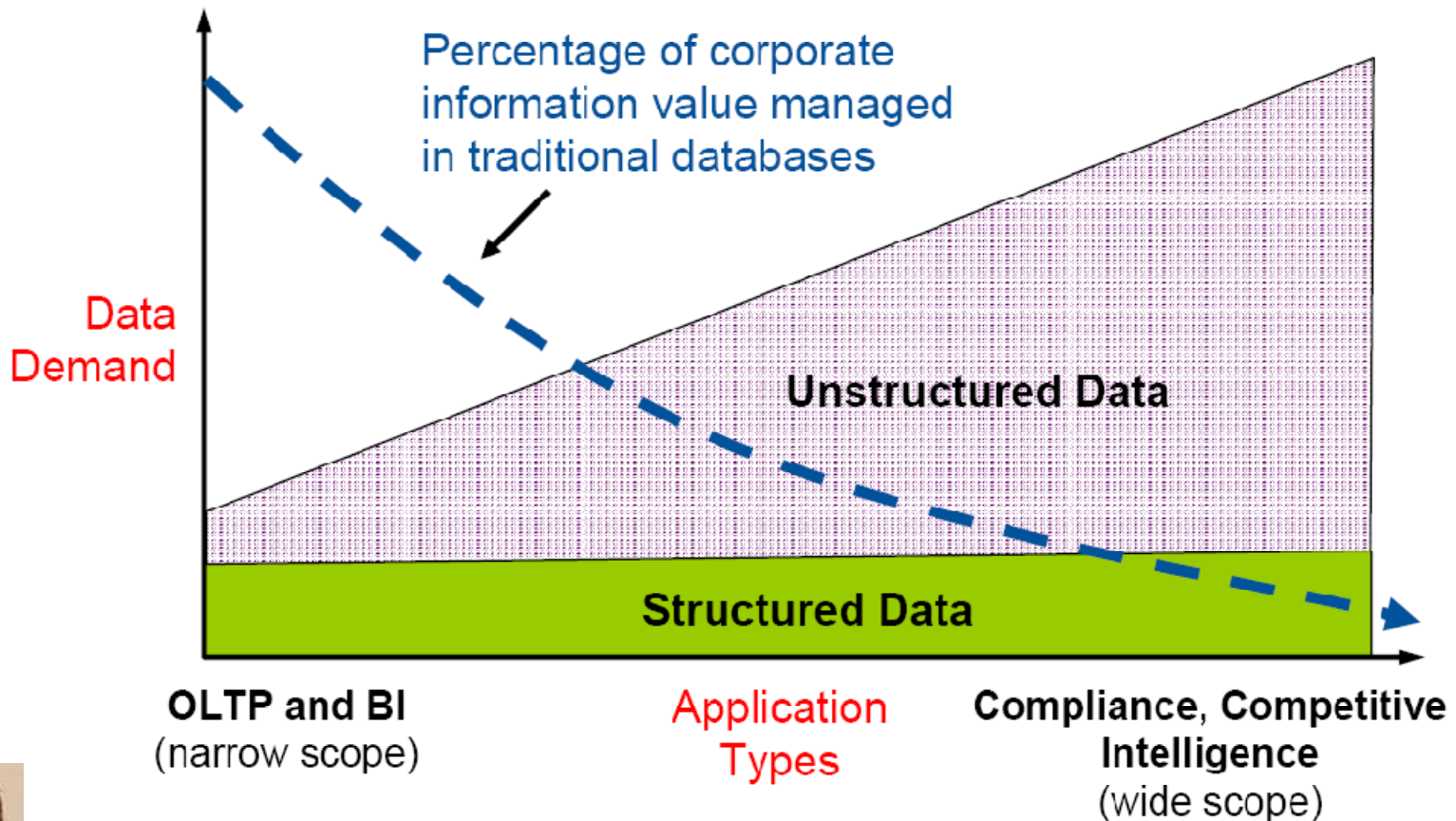


## Data Quality

- Still people entering the data
  - We do typos
  - We duplicate
  - We mess up



# Unstructured Information in the Warehouse



Mark Beyer



# Data Consolidation and ETL Studio



Direct access to RDBMs for info from some Telco's



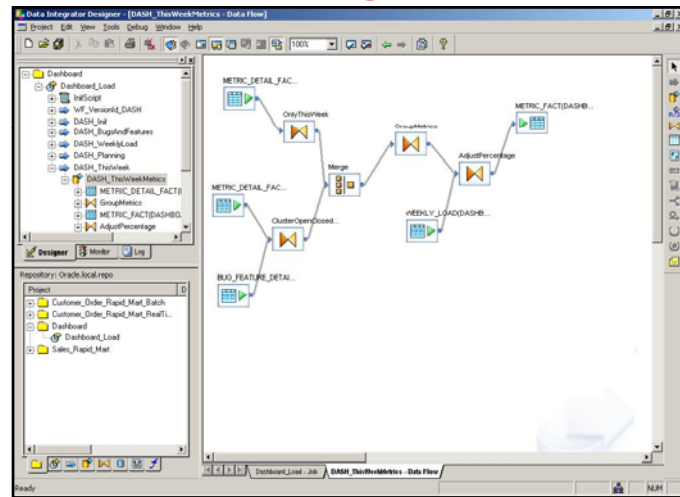
XML feed from other Telco's



Flat files (CSV or fixed) from the 'laggards'



Logic for matching and cleansing



Logic for ETL

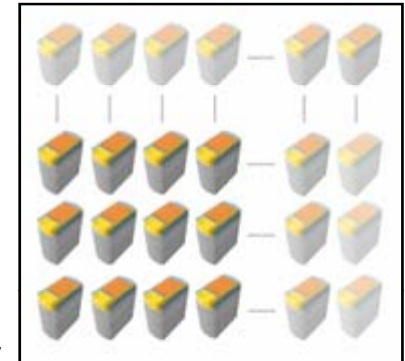
clean data

Master database for persistent storage



Lookup to ESP for Matching

Ordered hits (by quality)

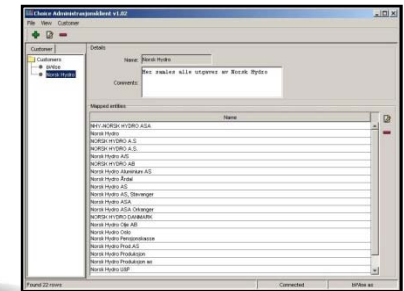


Cleansed data Back to ESP for next round Cleansing or for 'search'

Ambiguous data (close hits or unidentified)

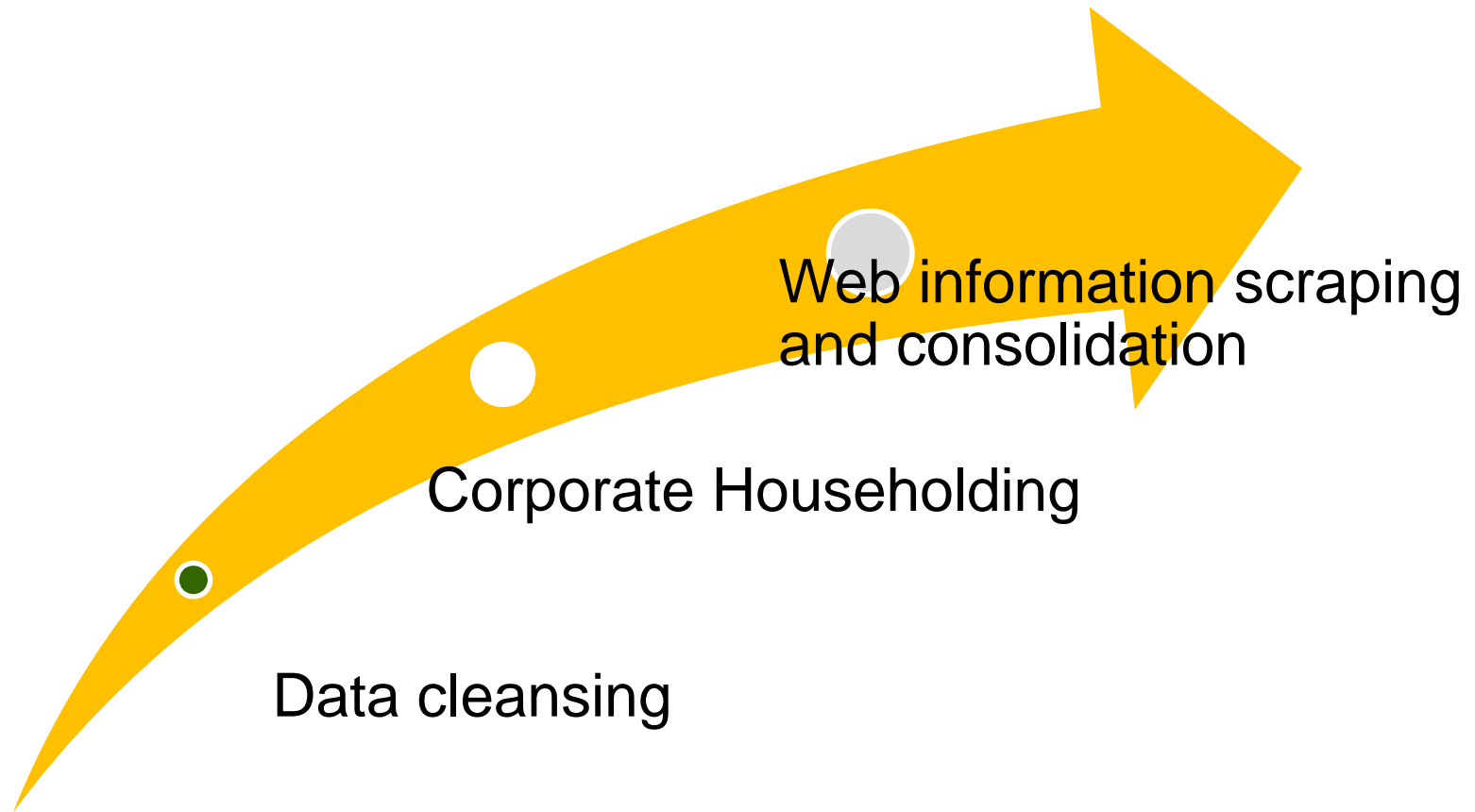
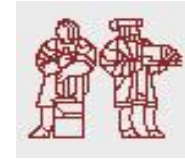


'Error' database for manual inspection, correction, storage/learning





The MIT Information Quality Industry Symposium, 2007



**fast**



# The MIT Information Quality Industry Symposium, 2007



## Viewing data, we notice issues

The screenshot shows a Microsoft Internet Explorer browser window displaying a corporate dashboard. The browser's address bar shows the URL `http://localhost/Radar/Maintenance/UserPages.aspx?PluginPage=101`. The dashboard features the 'fast' logo and a navigation menu with items like 'Home Page', 'Sales YTD', 'My Portal', 'Administration', 'Data Tools', and 'Maintenance'. Two main data sections are visible: 'Brazil sales' and 'Gauge 102 Northwind Sales'. Each section includes a traffic light icon and a list of thresholds: 'Unacceptable', 'Caution', and 'Value'. The 'Brazil sales' section shows a yellow light and values of 95,000 (Unacceptable), 110,000 (Caution), and 107,658 (Value). The 'Gauge 102 Northwind Sales' section shows a green light and values of 1,000,000 (Unacceptable), 1,250,000 (Caution), and 1,354,459 (Value). A search box with a 'Search' button is also present in the 'MySearch' section.

Section	Unacceptable	Caution	Value
Brazil sales	95,000	110,000	107,658
Gauge 102 Northwind Sales	1,000,000	1,250,000	1,354,459





# The MIT Information Quality Industry Symposium, 2007



## We click on the gauge

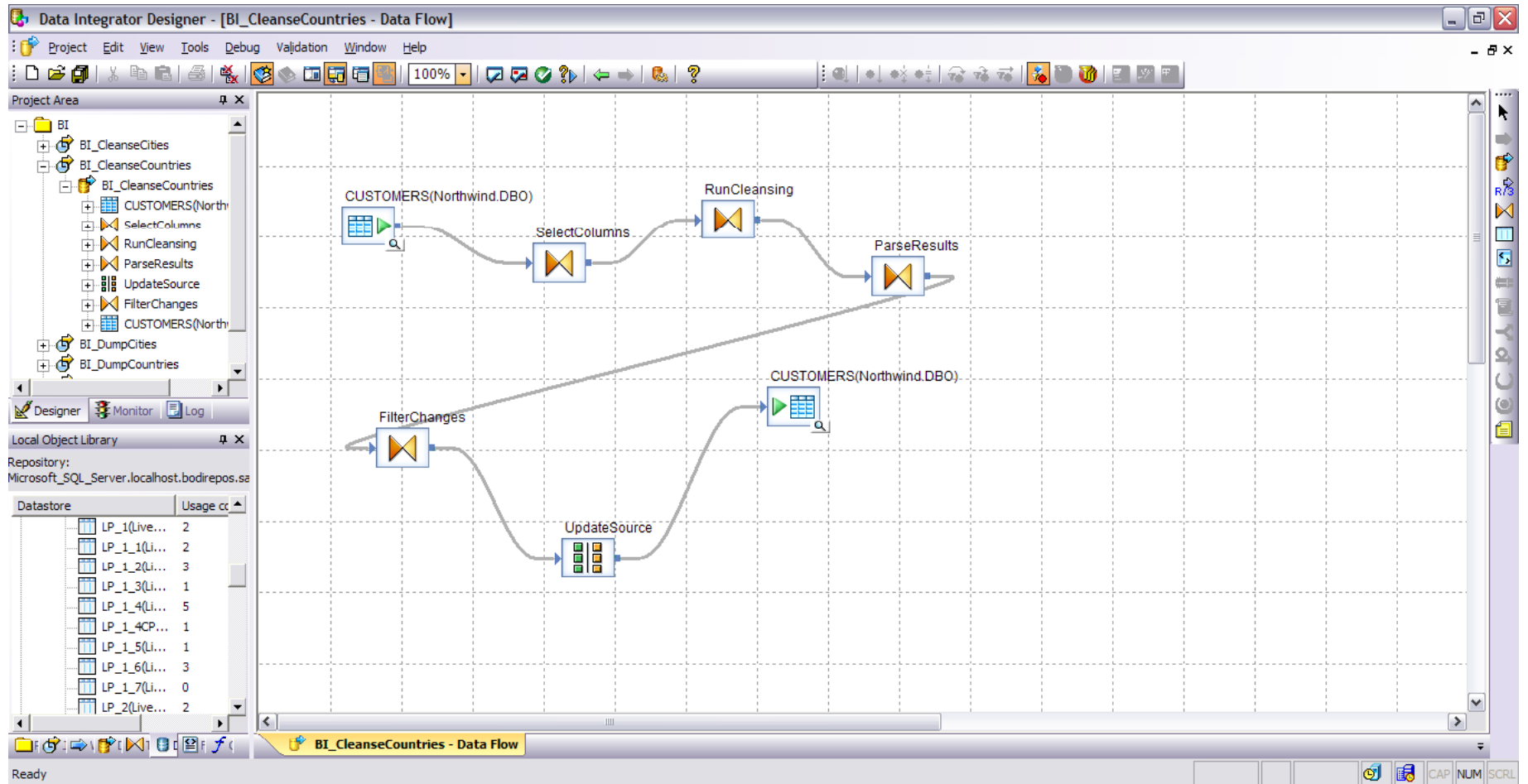
The screenshot shows a Microsoft Internet Explorer browser window displaying a web application. The browser's address bar shows the URL: `http://localhost/Radar/DataTools/DrillDown2.aspx?AggID=102&stage=1.1`. The page header includes the 'fast' logo and a welcome message: 'Welcome, Per Anders Friday, June 1, 2007 12:26 PM'. The main content area features a table titled 'Sales by country' with 24 records. The table has two columns: 'Country' and 'Sales'. The 'Argentina' row is circled in red. The table data is as follows:

Country	Sales
Argentina	8,119
Austria	139,497
Belgium	35,135
Brazil	7,311
Brazil	107,658
Canada	55,334
Denmark	34,782
Finland	19,778
France	85,499
Germany	244,641
Ireland	57,317
Italiano	1,546
Italy	15,159
Mex.	1,403
Mexico	22,671
Norway	5,735
Poland	3,532
Portugal	12,469





# We cleanse the data







# The MIT Information Quality Industry Symposium, 2007



## Now look at the result

Corporate Radar - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites RSS Print Mail New Tab

Google Start Bokmerker 38 blokkert ABC Controller Send til Innstillinger

Address http://localhost/Radar/Maintenance/UserPages.aspx?PluginPage=101 Go

**fast**

Welcome, Per Anders  
Friday, June 1, 2007 12:28 PM

Home Page Sales YTD

My Portal Administration Data Tools Maintenance

Brazil sales MySearch

**Brazil sales**

Unacceptable: 95,000  
Caution: 110,000  
Value: 114,968

**Global sales**

**Gauge 102 Northwind Sales**

Unacceptable: 1,000,000  
Caution: 1,250,000  
Value: 1,354,459

Done Trusted sites





# The MIT Information Quality Industry Symposium, 2007



## Details

Corporate Radar - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites RSS Mail Print Writable Folder Help

Google G Start [Icons] Bokmerker 38 blokkert ABC Kontroller Send til Innstillinger

Address http://localhost/Radar/DataTools/DrillDown2.aspx?AggID=102&stage=1.1 Go

**fast**

Welcome, Per Anders  
Friday, June 1, 2007 12:28 PM

Home Page Sales YTD

My Portal Administration Data Tools Maintenance

### Sales by country

page 1 of 1 21 of 21 records

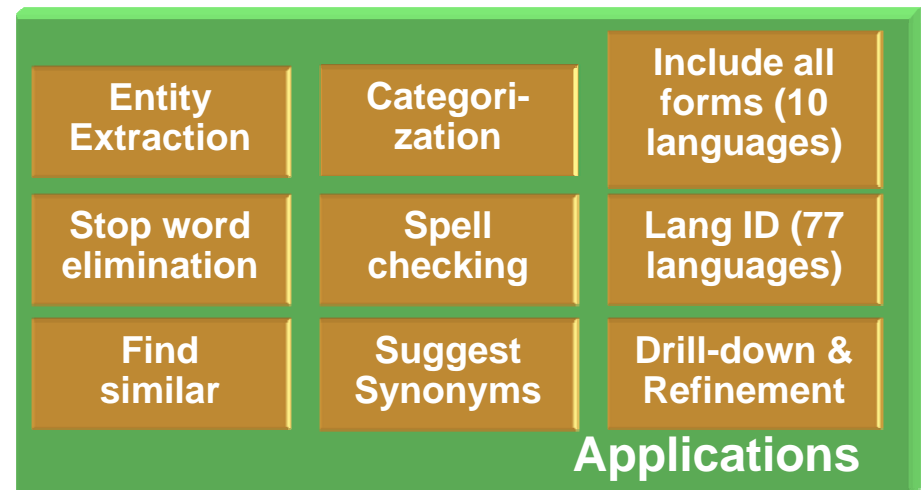
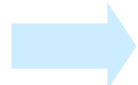
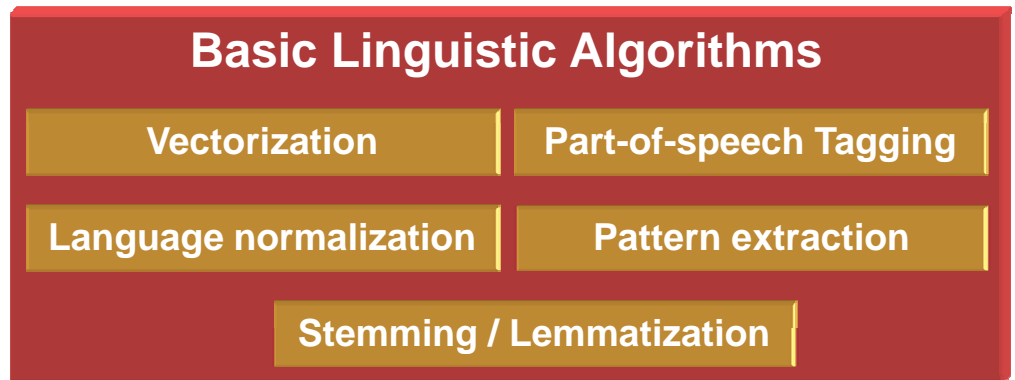
Country	Sales
Argentina	8,119
Austria	139,497
Belgium	35,135
Brazil	114,968
Canada	55,334
Denmark	34,782
Finland	19,778
France	85,499
Germany	244,641
Ireland	57,317
Italy	16,705
Mexico	24,073
Norway	5,735
Poland	3,532
Portugal	12,469
Spain	19,432
Sweden	59,524
Switzerland	32,920

Done Trusted sites





# Linguistic Fundamentals





The MIT Information Quality Industry Symposium, 2007



# Q&A

**fast**



The MIT Information Quality Industry Symposium, 2007



Thank you!

**fast**  
*find the real value of search*

**Jeff Fried**  
**VP Advanced Solutions**  
**Jeff.Fried@fastsearch.com**

**fast**