The MIT Information Quality Industry Symposium, 2007

Applying search technology to data matching, consolidation, and cleansing

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

:::fast

The MIT Information Quality Industry Symposium, 2007
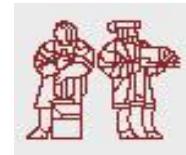
All Data is Dirty

|	
Forrt Lauderdale	**Fort Lauderdaleq**
Fort ALuderdale	**Fort Lauderrdale**
Fort Laderdale	**Fort Laudewrdale**
Fort Laudedale	**FORT LUADERDALE**
Fort Lauderale	**Fort. Lauderdale**
Fort Lauderdal	**Fort.Lauderdale**
Fort Lauderdale	**Ft Lauderdale**
FORT LAUDERDALE FLA.	**Ft. Lauderdale**
Fort LAuderdale,	**FT. LAUDERDALE**
FtLauderdale	**FT.LAUDERDALE**

:::*fast*

The MIT Information Quality Industry Symposium, 2007
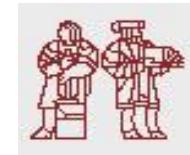
Search technology emphasizes connecting to the User

**FAST Connects Users To Information, Products and Communities**

USERS

FrontOffice: Search IS the Portal

Search Connects: Service Delivery Framework

BackOffice: Adaptive Information Warehouse

| **Information** | **Products** | **Communities** |

XML

::::fast

3
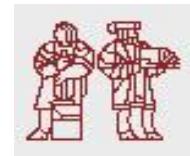
The Adaptive Information Warehouse

The first business intelligence solution built on search. Redefining how organizations access, interact with and exploit information.

The Adaptive Information Warehouse (AIW)

Advanced data integration and cleansing

Dynamic, real-time structured and unstructured data

**Integrated intuitive business intelligence portal**

| Data Cleansing | FAST ESP | Radar |
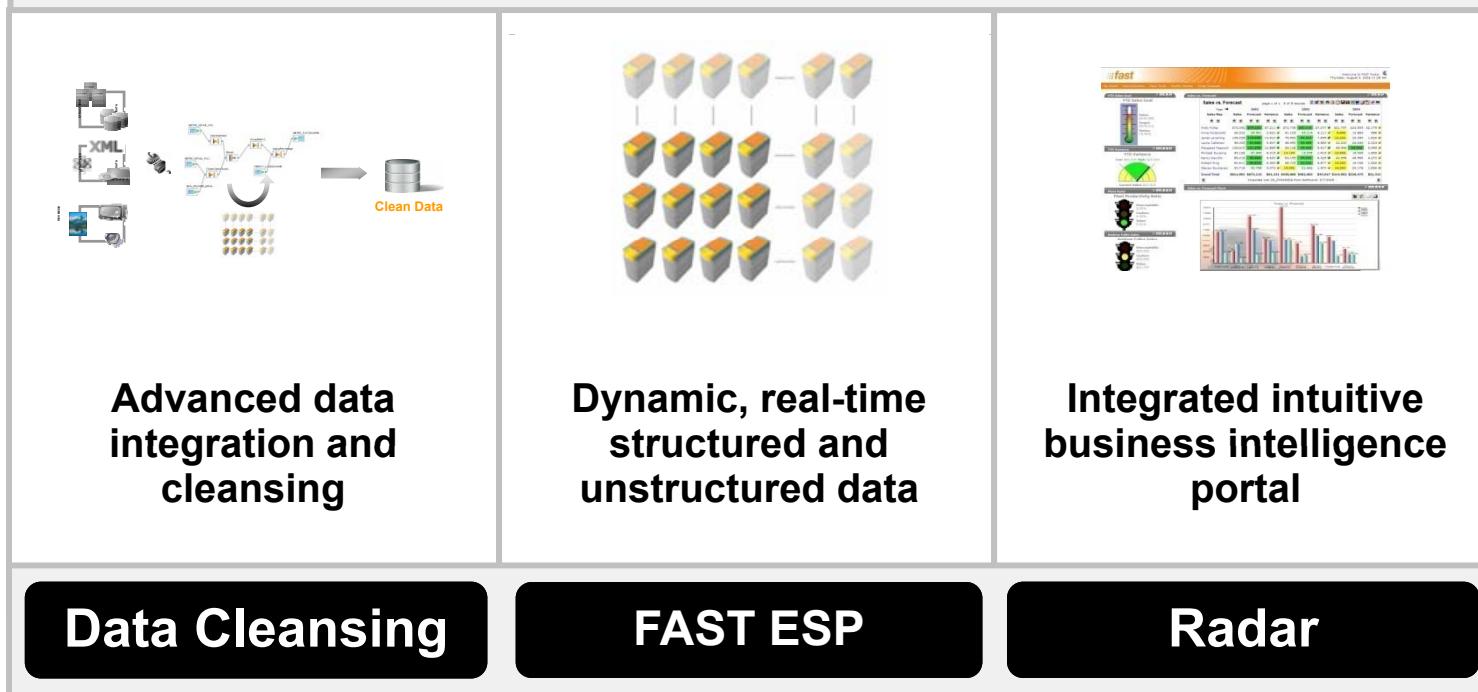
The MIT Information Quality Industry Symposium, 2007
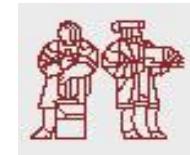
Customer ID	Last Name	First Name	Address	City	State	Zip
12093-28937	Williams	Joshua	41 Smiley Ave	Haverhill	MA	01830
23801-12938	Simpson	Mark	194 No 8 Rd	New Wilmington	PA	16142
28934-83112	Karel	Adrienne	2435 Park Ave. Southeast	Milpitas	CA	95035
38472-20937	Meyers	John	10932 Wellington St.	Albuquerque	NM	87106
42348-00198	Welch	Jack	1338 Ruswood Dr	Abilene	TX	79601
72831-21230	Thatcher	Margaret	9331 E Cactus Ln S	Sun Lakes	AZ	85248
92837-43224	Johnson	Alice	872 N. Campbell Ave.	Springfield	MO	65802

Customer Records from Corporate CRM Database

FAST Search Index
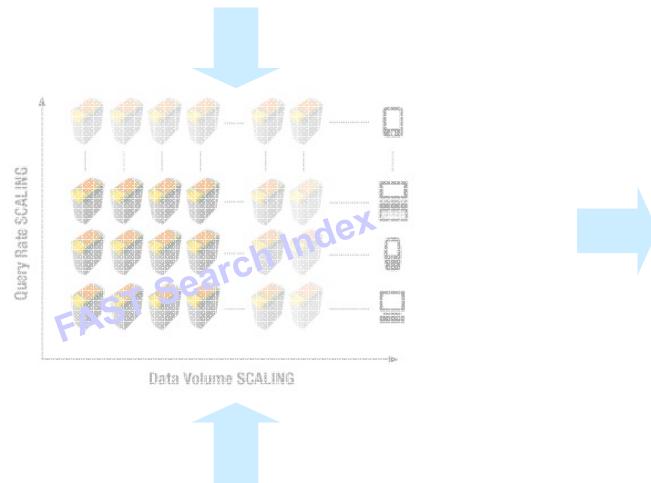
Match, merge, cleanse.

Last Name	First Name	Address	City	State	Zip	Phone No.
Carroll	Adrian	2435 Park Ave. S East	Milpitas	CA	95035	408-130-2897
Davidson	Henry	2997 Shore Dr	Merrick	NY	11566	212-287-8821
Johnson	Alice	872 N. Campbell Ave.	Springfield	MO	65802	417-862-3465
Meiers	John	10932 Willinton	Albaquerqee	NM		505-792-0323
Phillips	Natalie	4916 N Hoyne Ave	Chicago	IL	60625	773-315-2872

Regional Sales Call Center /

Customer Service Database

:::fast

5

The MIT Information Quality Industry Symposium, 2007
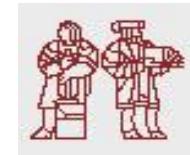
Data matching, cleansing, and consolidation

- Biggest problem for all data warehouses
    - Merge of multiple sources of data
    - No common key

- FAST solution: unique combination of traditional DWH logic (ETL) and fuzzy matching
    - Full ETL tool for the traditional ETL logic
    - Use search engine core for fuzzy match
   - Apply linguistic techniques to data quality issues

:::*fast*

6

The MIT Information Quality Industry Symposium, 2007
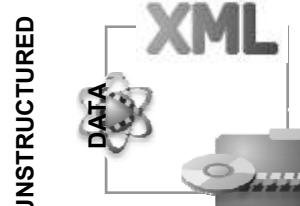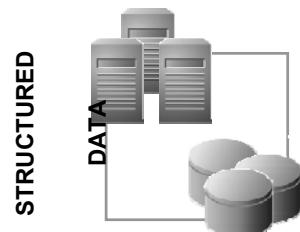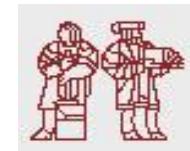
Cleansing Example:

Schibsted built Internet Yellow Pages and White Pages from scratch in 8 weeks outperforming existing providers on data quality
98 % Cleansing Precision

STRUCTURED DATA

UNSTRUCTURED DATA

RICH MEDIA

METRIC_DETAIL_FAC...

OnlyThisWeek

GroupMetrics

METRIC_FACT(DASHB...

Merge

AdjustPercentage

METRIC_DETAIL_FAC...

ClusterOpenClosed...

WEEKLY_LOAD(DASHB...

BUG_FEATURE_DETAI...

Clean Data

THE SEMANTIC INDEX

THE SEMANTIC INDEX

▶ **Content Fusion**
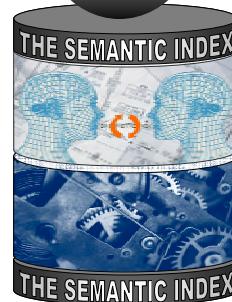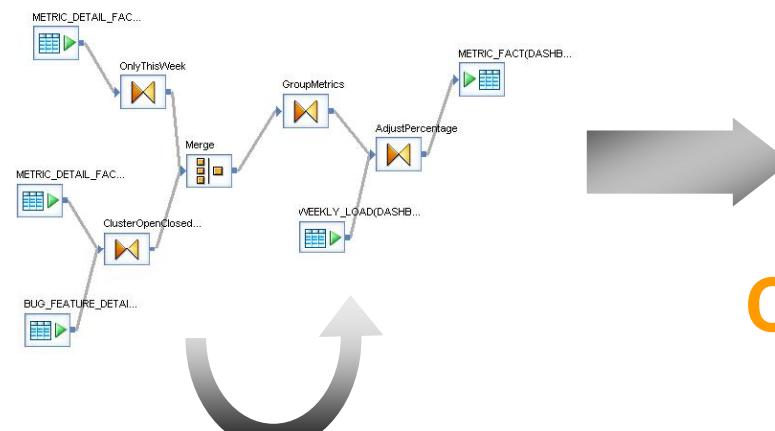
▶ **Semantic and Phonetic Cleansing Analysis**

:::fast

7
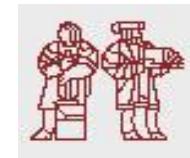
Example system configuration

Ling. matching

XML
XML
XML

Cleansing

'ambigous' database

Master database

Editorial Inspection GUI

"Back office"

"Front office"

Data Load

For Search

Search Portal
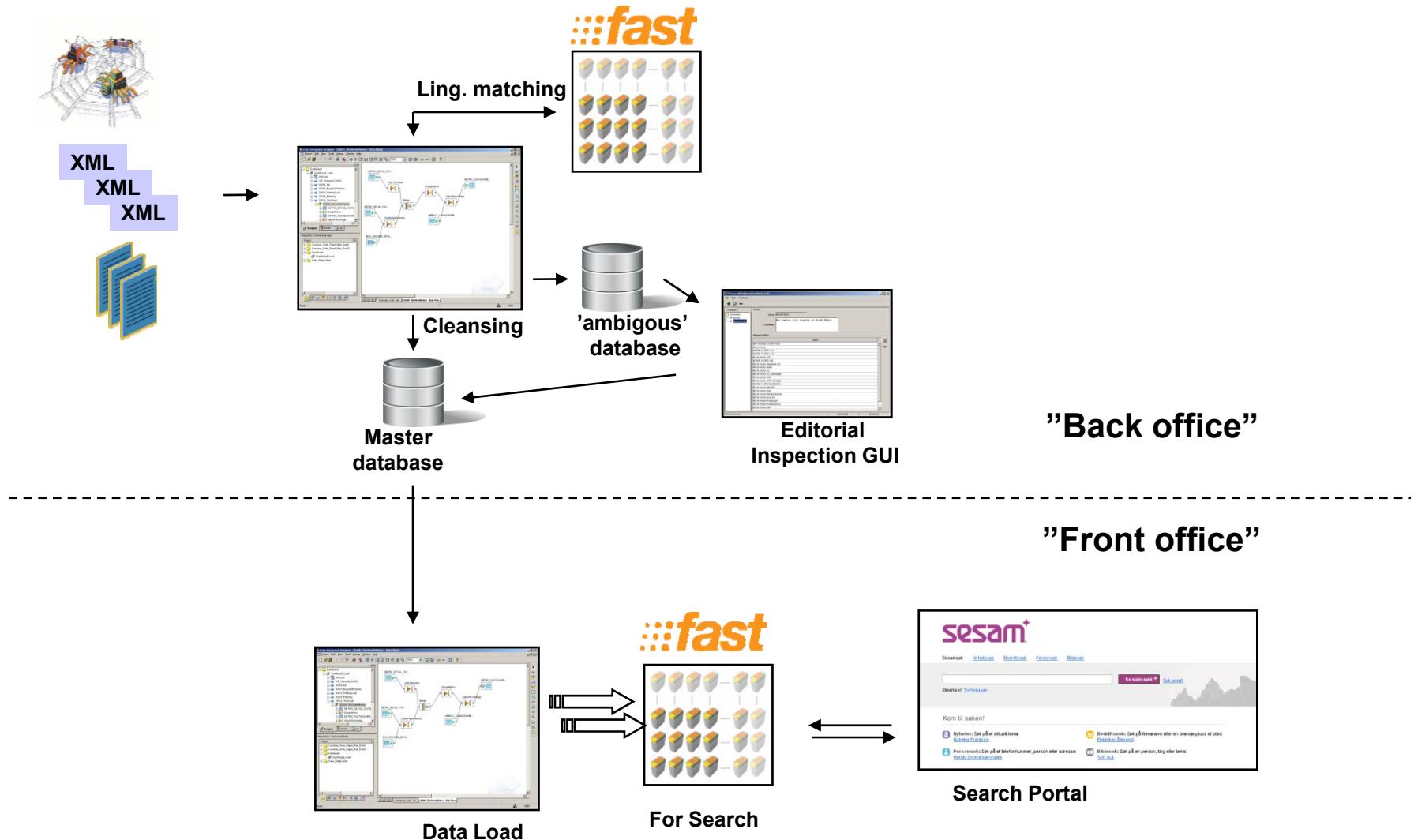
Data Warehouse of the Year Award 2006

Data Cleansing Scenario One: "Master" DB exists
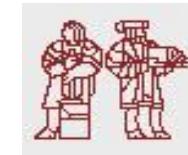
One ' Master' exists, other datasources are cleansed towards it

Step one:

- The Customer Master is indexed by FAST ESP

Step two:

A datasource is cleansed according to the Master

Result: a corrected version of that datasource

10

Data Cleansing Scenario Two:
# No "Master" DB
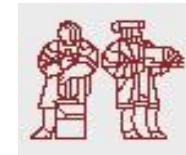
Two or more databases exist, and a merge is needed

Incident records from Central Region

	id	lastname	firstname	address	city	state	zip
1	172-32-1176	White	Johnson	10932 Bigge Rd.	Menlo Park	CA	94025
2	213-46-8915	Green	Marjorie	309 63rd St. #411	Oakland	CA	94618
3	238-95-7766	Carson	Cheryl	589 Darwin Ln.	Berkeley	CA	94705
4	267-41-2394	O'Leary	Michael	22 Cleveland Av. #14	San Jose	CA	95128
5	274-80-9391	Straight	Dean	5420 College Av.	Oakland	CA	94609
6	341-22-1782	Smith	Meander	10 Mississippi Dr.	Lawrence	KS	66044
7	409-56-7008	Bennet	Abraham	6223 Bateman St.	Berkeley	CA	94705
8	427-17-2319	Dull	Ann	3410 Blonde St.	Palo Alto	CA	94301
9	472-27-2349	Gringlesby	Burt	PO Box 792	Covelo	CA	95428

Incidents from East Region

	id	lastname	firstname	address	city	state
1	807-91-6654	Panteley	Sylvia	1956 Arlington Pl.	Rockville	MD
2	527-72-3246	Greene	Morningstar	22 Graybar Hou...	Nashville	TN
3	722-51-5454	DeFrance	Michel	3 Balding Pl.	Gary	IN
4	712-45-1867	del Castillo	Innes	2286 Cram Pl. #..	Ann Arbor	MI
5	341-22-1782	Smith	Meander	10 Mississippi Dr.	Lawrence	KS
6	899-46-2035	Ringer	Anne	67 Seventh Av.	Salt Lake City	UT
7	998-72-3567	Ringer	Albert	67 Seventh Av.	Salt Lake City	UT
8	172-32-1176	Wite	Johnson	10932 Bigge Rd.	Menlo Park	CA
9	486-29-1786	Locksley	Charlene	18 Broadway Av.	San Francisco	CA
10	427-17-2319	Dull	Ann	3410 Blonde St.	Palo Alto	CA
11	846-92-7186	Hunter	Sheryl	3410 Blonde St.	Palo Alto	CA
12	672-71-3249	Yokomoto	Akiko	3 Silver Ct.	Walnut Creek	CA
13	724-08-9931	Stringer	Dirk	5420 Telegraph	Oakland	CA
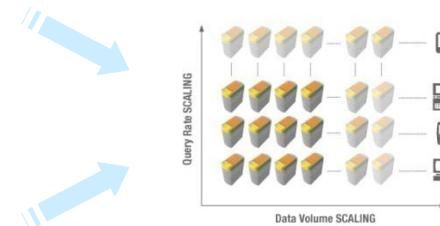
Cleansed Unioned Incident list

	id	lastname	firstname	address
1	172-32-1176	White	Johnson	10932 Bigge R
2	213-46-8915	Green	Marjorie	309 63rd St. #
3	238-95-7766	Carson	Cheryl	589 Darwin Ln
4	267-41-2394	O'Leary	Michael	22 Cleveland A
5	274-80-9391	Straight	Dean	5420 College A
6	341-22-1782	Smith	Meander	10 Mississippi
7	409-56-7008	Bennet	Abraham	6223 Bateman
8	427-17-2319	Dull	Ann	3410 Blonde S
9	472-27-2349	Gringlesby	Burt	PO Box 792
10	486-29-1786	Locksley	Charlene	18 Broadway A
11	527-72-3246	Greene	Morningstar	22 Graybar Ho
12	648-92-1872	Blotchet-Halls	Reginald	55 Hillsdale Bl
13	672-71-3249	Yokomoto	Akiko	3 Silver Ct.
14	712-45-1867	del Castillo	Innes	2286 Cram Pl.
15	722-51-5454	DeFrance	Michel	3 Balding Pl.
16	724-08-9931	Stringer	Dirk	5420 Telegrap
17	724-80-9391	MacFeather	Stearns	44 Upland Hts
18	756-30-7391	Karsen	Livia	5720 McAuley

- Duplication identification and removal
- Approximate matching to help oversee minor errors
- Multi-field compare to ensure secure identification
- Tunable relative importance of each field
- More than 20 million names known by the plattform
- Spell checking and error correction included

- Private attributes from eiher source are merged to a super-set

::: fast

Search Is a Matching Service
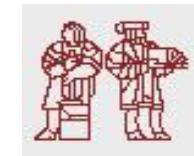
3-Gram

Score

ABUKABLAN, KHALID HUSSEIN OCEAN BLVD 3062 POMPANO BEACH

131 776

3-Gram

Search improves Data Quality through de-duplication

:::*fast*

12

Linguistic Capabilities Provide High Data Matching Accuracy
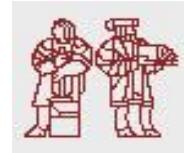
- **Score strings on similarity**
    - **Lai Nancy, 501 N. Clinton St. ≈ Nancy J. Lai, 501 North Clinton**

- **Levenshtein edit distance algorithms**
    - **Calculates the difference on two strings (names)**
    - **John Meyers ≈ John P. Meiers; Britney Spears ≈ Britney Spaers**

- **Phonetic match**
    - **Calculates an integer from the strings, and matches on 'sounds like'**
    - **Christin ≈ Kristin; Kierkegaard ≈ Kierkegård**

- **Scope Fields (new in FAST Instream 4.1 and FAST ESP 5.0)**
    - **Finds relevancy in a scope or a context – for example 'Person' vs. 'Address'.**

- **'Charset' normalization**
    - **For multinational chars and name combinations**
   - **André = Andre, Jürgen = Jurgen etc.**

:::*fast*

The MIT Information Quality Industry Symposium, 2007
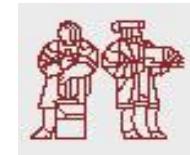
Linguistic Fundamentals

Lexicon Base

- Special terminology lexica
- Geographical and people's names
- Spellcheck dictionaries
- Subject-specific ontologies
- Synonymy Dictionaries
- Part-of-speech Dictionaries
- Inflection Dictionaries
- Language-specific Common Words

Basic Linguistic Algorithms

Vectorization	Part-of-speech Tagging
Language normalization	Pattern extraction

Stemming / Lemmatization

Applications

Entity Extraction	Categori-zation	Include all forms (10 languages)
Stop word elimination	Spell checking	Lang ID (77 languages)
Find similar	Suggest Synonyms	Drill-down & Refinement

:::fast

14

Where do Linguistics and other search technologies apply to data quality?
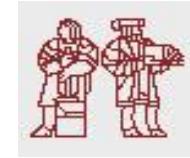
- On textual and/or multimedia data

- In combination with other rules-based techniques

- In hard-to-match situations
  - Match and cleanse with significantly fewer rules
  - Match and cleanse with much higher accuracy

- Where structuring "blobs" of text is useful
  - Entity Extraction
  - Relationship Extraction
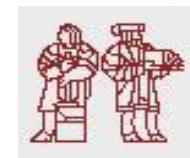
# Data Consolidation and ETL Studio

Direct access to RDBMs for info from some Telco's

XML feed from other Telco's

XML

Flat files (CSV or fixed) from the 'laggards'
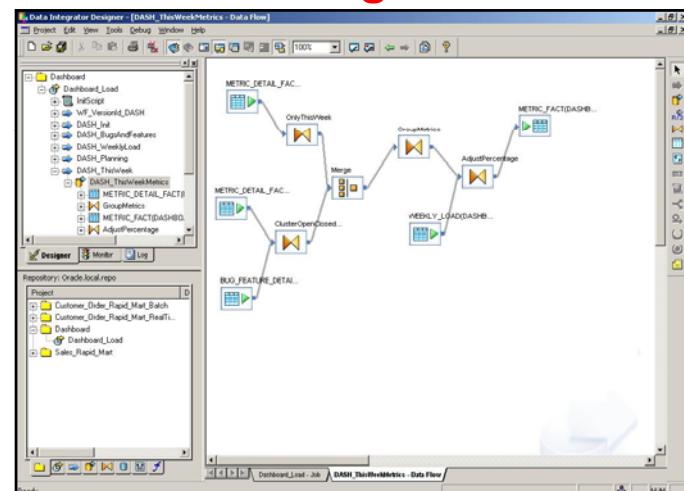
Logic for matching and cleansing

Lookup to ESP for Matching

Ordered hits (by quality)
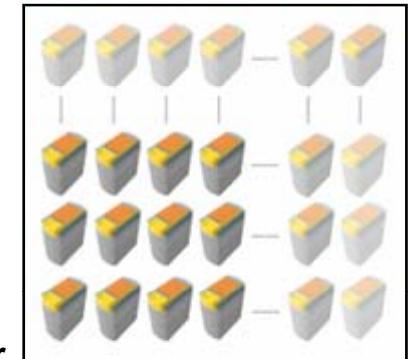
Cleansed data Back to ESP for next round Cleansing or for 'search'

Logic for ETL

clean data

Master database for persistant storage

Ambigous data (close hits or unidentified)
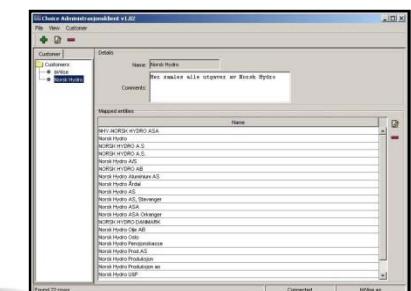
'Error' database for manual inspection, correction, storage/learning

fast

PG 701

The MIT Information Quality Industry Symposium, 2007
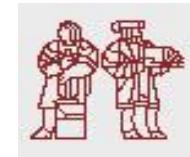
Key Cleansing Benefits

- Increased accuracy in search

- Improved entity recognition and relationship mapping

- Better analytics and intelligence

- Improvement and reuse of data for all applications

17

Cleansing: Impact on Discovery
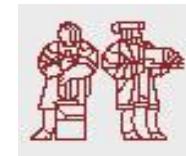
mozart | Search

Artist (100)

mozart (2810)
wolfgang amadeus MOZART (1980)
Mozart, Wolfgang amadeus (370)
W. A. MOZART (175)
Mozart (Complete Works) (152)
w.a. mozart (139)
w.a.mozart (47)
WOLFGANG AMADEUS MOZART (1756-1791) (39)
MOZART WOLFGANG AMADEUS (37)
Mozart, Wolfgang Amadeus (1756-1791) (34)
Mozart, W.A. (28)
Mozart, W. A. (27)
London Mozart Players (17)
wolfang amadeus mozart (17)
WA Mozart (16)
Wolfgang A. Mozart (16)
Wolgang Amadeus Mozart (12)
MOZART FESTIVAL ORCHESTRA (12)
Wolfgang Mozart (11)
W A Mozart (10)
80 more...

Artist (6)

Wolfgang Amadeus Mozart (1998)
Mozart Festival Orchestra (10)
Leopold Mozart (7)
Mozart & Haydn (4)
Mozart - Beethoven (4)
Beethoven Mozart (3)

::::fast

18

The Largest BI Limitations Today

2007
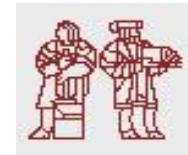
If We Build It,

<50% of Them Will Come

What else is wrong with BI today?

"Is everyone using BI who should be using BI?"

Region	Mean %
North America, n=284	32%
Europe, n=132	39%
Australia, n=44	36%
Japan, n=37	53%
China, n=44	40%

Mean %

Source: Gartner

Deployment Challenges
What are the barriers to adopting BI tools enterprise-wide?

- Ease-of-use issues with less technically-savvy employees
- Integration/compatibility issues with existing, multiple platforms
- Data quality problems
- Training internal staff too time-intensive and costly
- No clear ROI
- Issues with scalability across the entire organization
- No industry standard
- BI talent is too expensive
- Lower-than-expected analytic value
- Failure to address/integrate with our current or future BPM initiatives
- Other

% of Respondents

BI Built on Search Redefines How Organizations Access, Interact With and Exploit Business Intelligence
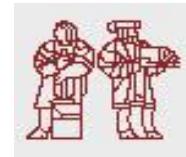
Drivers of Pervasive BI:

- Better business decisions with accurate, enriched data

- Increased productivity with ad-hoc, real-time performance

- Increased adoption with easy and intuitive access to intelligence

ease of use

democratize information

perform deeper analytics

currently just a handful of power users

high quality data

real-time information

break down silos

timely cross-enterprise information

integrate structured and unstructured data

speed-up decision making

suppress noise in data

::::fast

BI Built on Search Redefines How Organizations Access, Interact with and Exploit Information

1.
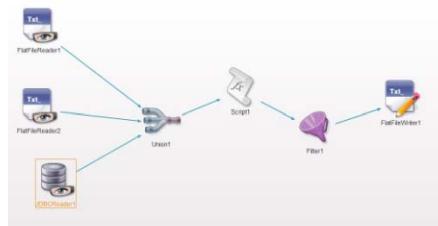
2.

3.

Better Business decisions with accurate data
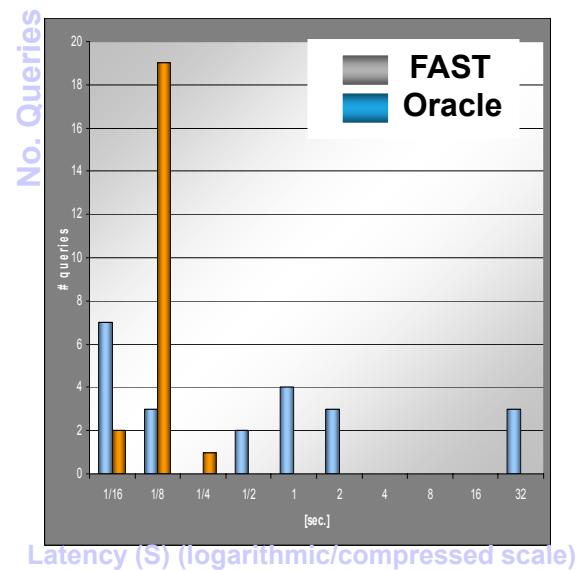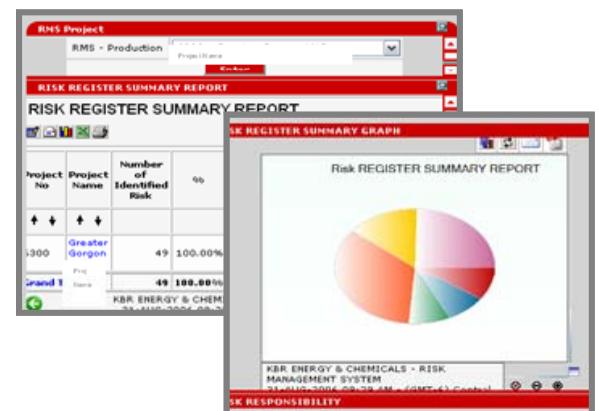
Latency (S) (logarithmic/compressed scale)

Increased productivity with ad-hoc, real-time performance

Increased Adoption with easy and intuitive access to intelligence

Analytical Portals built on Search

PG 707

# FAST Vision For Pervasive Business Intelligence

Ad-hoc, FAST access

$VVV^n LDB$
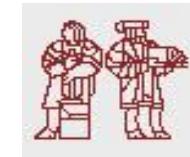
Exabyte scale information management
   Structured/Semi-structured/Unstructured

EASY analytics
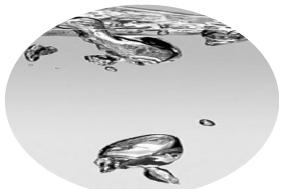
Data Improvement
   "Content Capture and Refinement"

Connections across information
   "Semantic Integration"

Real Time information

The MIT Information Quality Industry Symposium, 2007
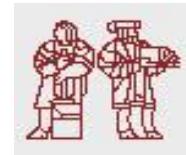
Thank you!

find the real value of search

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