ORACLE®

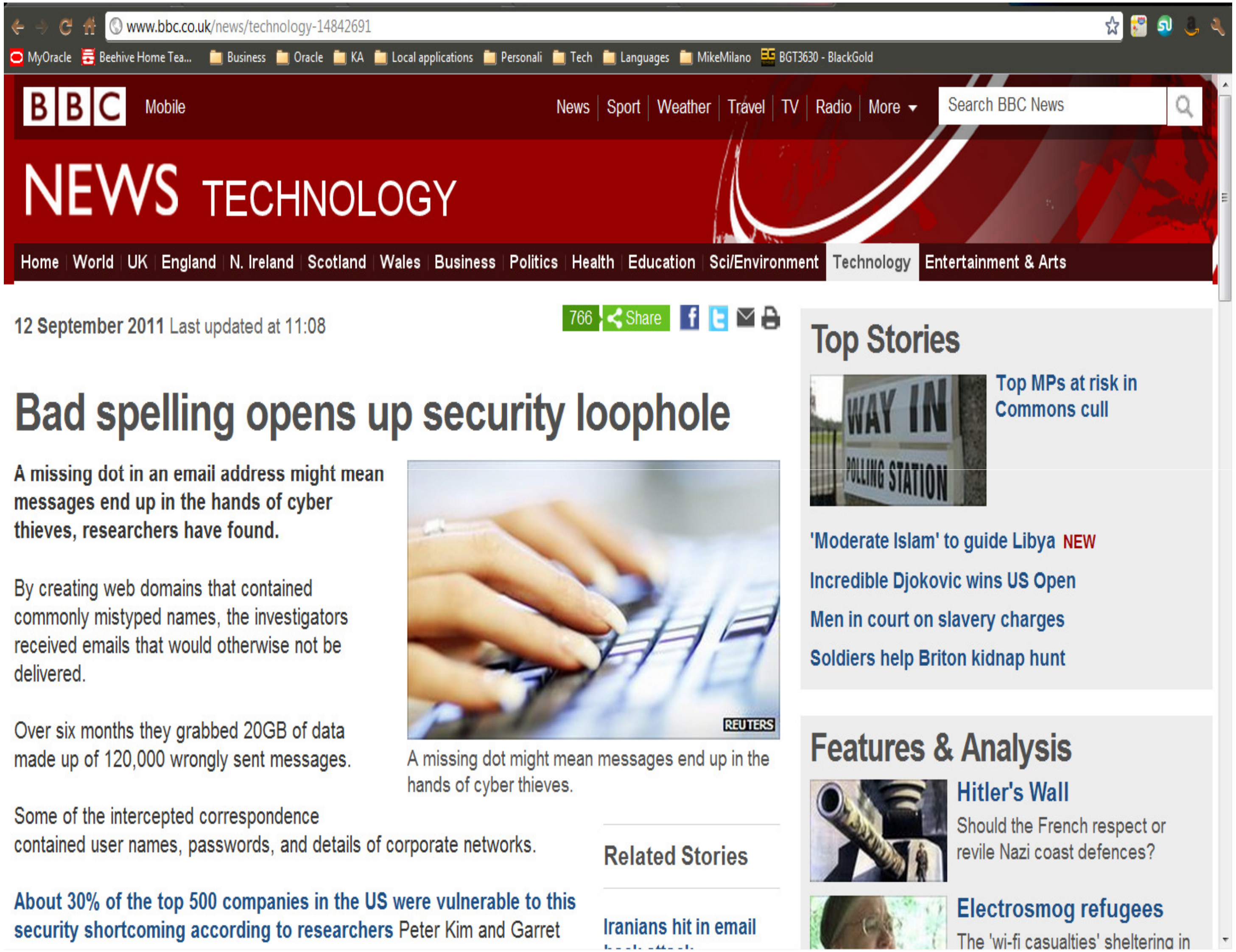
Get more value from your DWH

Ugo Pollio – Sales Consulting and Business Development
Oracle EMEA Data Integration Solutions

Program Agenda

- The challenge of Quality of Data
- Oracle Data Integration Solution
- Get More Value from your DWH
- Q&A

Data Quality issues are on your daily business



12 September 2011 Last updated at 11:08

Bad spelling opens up security loophole

A missing dot in an email address might mean messages end up in the hands of cyber thieves, researchers have found.

By creating web domains that contained commonly mistyped names, the investigators received emails that would otherwise not be delivered.

Over six months they grabbed 20GB of data made up of 120,000 wrongly sent messages.

Some of the intercepted correspondence contained user names, passwords, and details of corporate networks.

About 30% of the top 500 companies in the US were vulnerable to this security shortcoming according to researchers Peter Kim and Garret

A missing dot might mean messages end up in the hands of cyber thieves.

Related Stories

Iranians hit in email

Top Stories

Top MPs at risk in Commons cull

'Moderate Islam' to guide Libya NEW

Incredible Djokovic wins US Open

Men in court on slavery charges

Soldiers help Briton kidnap hunt

Features & Analysis

Hitler's Wall

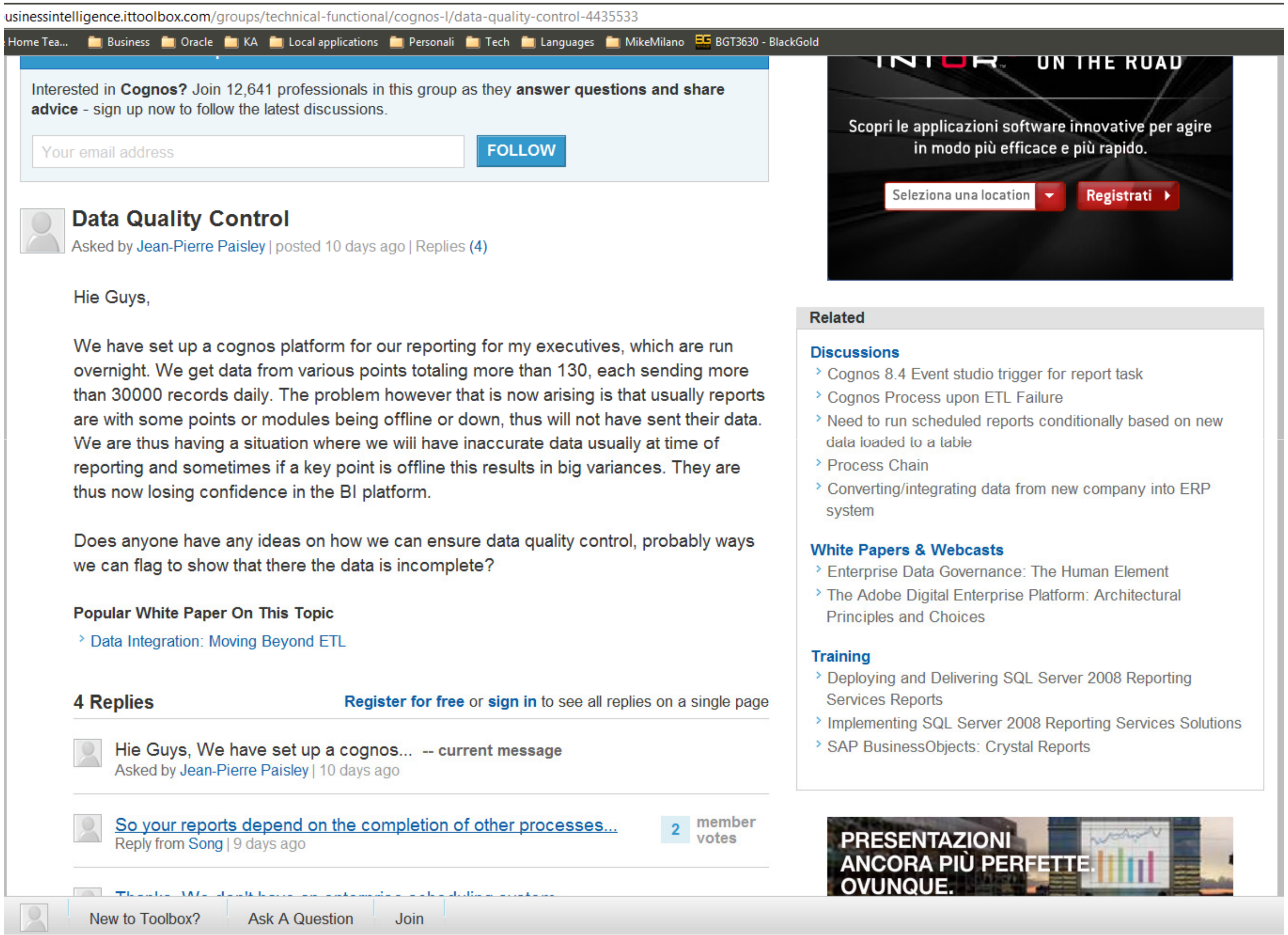
Should the French respect or revile Nazi coast defences?

Electrosmog refugees

The 'wi-fi casualties' sheltering in

http://www.bbc.co.uk/news/technology-14842691

An example in DWHs…



usinessintelligence.ittoolbox.com/groups/technical-functional/cognos-l/data-quality-control-4435533

Home Tea... Business Oracle KA Local applications Personali Tech Languages MikeMilano BGT3630 - BlackGold

Interested in **Cognos?** Join 12,641 professionals in this group as they **answer questions and share advice** - sign up now to follow the latest discussions.

Your email address
FOLLOW

Data Quality Control

Asked by Jean-Pierre Paisley | posted 10 days ago | Replies (4)

Hie Guys,

We have set up a cognos platform for our reporting for my executives, which are run overnight. We get data from various points totaling more than 130, each sending more than 30000 records daily. The problem however that is now arising is that usually reports are with some points or modules being offline or down, thus will not have sent their data. We are thus having a situation where we will have inaccurate data usually at time of reporting and sometimes if a key point is offline this results in big variances. They are thus now losing confidence in the BI platform.

Does anyone have any ideas on how we can ensure data quality control, probably ways we can flag to show that there the data is incomplete?

Popular White Paper On This Topic

- Data Integration: Moving Beyond ETL

4 Replies — **Register for free** or **sign in** to see all replies on a single page

Hie Guys, We have set up a cognos... -- **current message**
Asked by Jean-Pierre Paisley | 10 days ago

So your reports depend on the completion of other processes... — 2 member votes
Reply from Song | 9 days ago

Thanks. We don't have an enterprise scheduling system...

Related

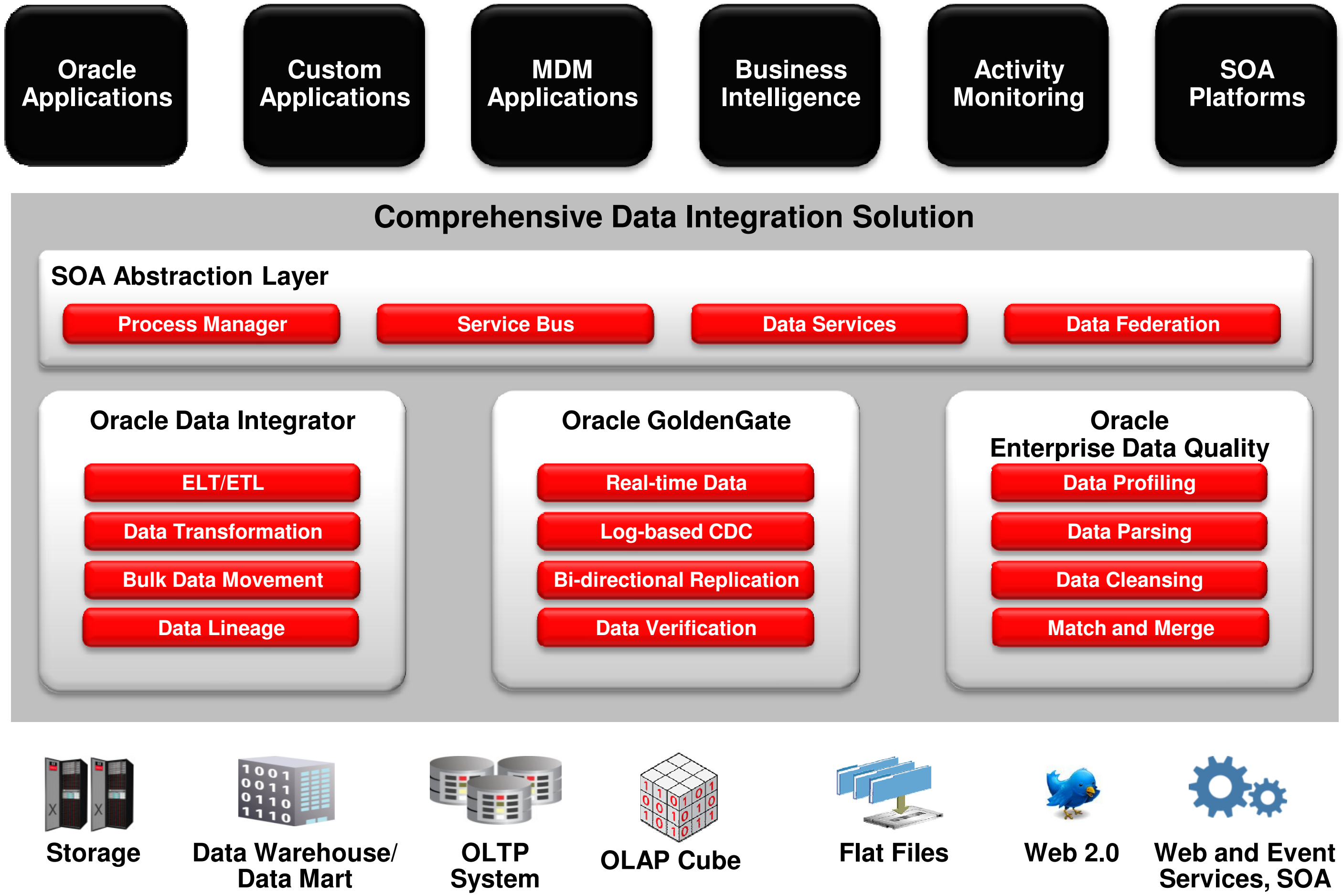
Discussions
- Cognos 8.4 Event studio trigger for report task
- Cognos Process upon ETL Failure
- Need to run scheduled reports conditionally based on new data loaded to a table
- Process Chain
- Converting/integrating data from new company into ERP system

White Papers & Webcasts
- Enterprise Data Governance: The Human Element
- The Adobe Digital Enterprise Platform: Architectural Principles and Choices

Training
- Deploying and Delivering SQL Server 2008 Reporting Services Reports
- Implementing SQL Server 2008 Reporting Services Solutions
- SAP BusinessObjects: Crystal Reports

Oracle Data Integration Solution

Best-in-class Heterogeneous Platform for Data Integration



Oracle Applications

Custom Applications

MDM Applications

Business Intelligence

Activity Monitoring

SOA Platforms

Comprehensive Data Integration Solution

SOA Abstraction Layer

- Process Manager
- Service Bus
- Data Services
- Data Federation

Oracle Data Integrator

- ELT/ETL
- Data Transformation
- Bulk Data Movement
- Data Lineage

Oracle GoldenGate

- Real-time Data
- Log-based CDC
- Bi-directional Replication
- Data Verification

Oracle Enterprise Data Quality

- Data Profiling
- Data Parsing
- Data Cleansing
- Match and Merge

Storage

Data Warehouse/ Data Mart

OLTP System

OLAP Cube

Flat Files

Web 2.0

Web and Event Services, SOA

ODI Saves Money

E-LT Runs on Existing Servers with Shared Administration

Typical: Separate ETL Server

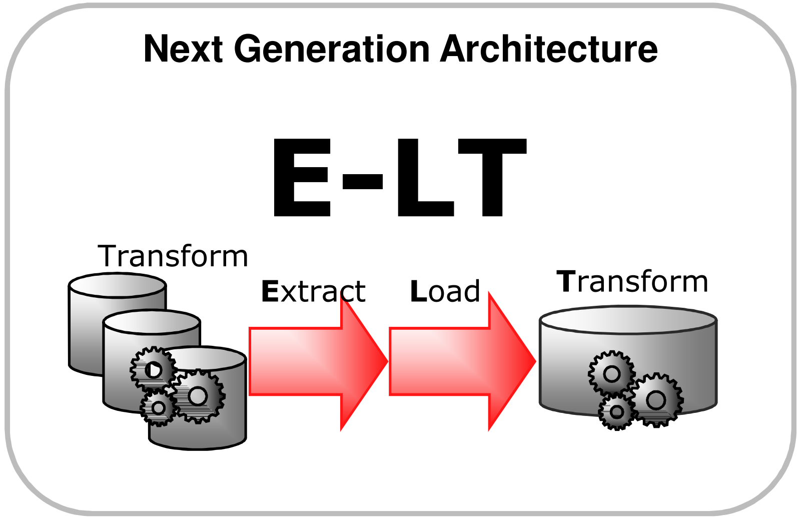
- Proprietary ETL Engine
- Expensive Manual Parallel Tuning
- High Costs for Standalone Server

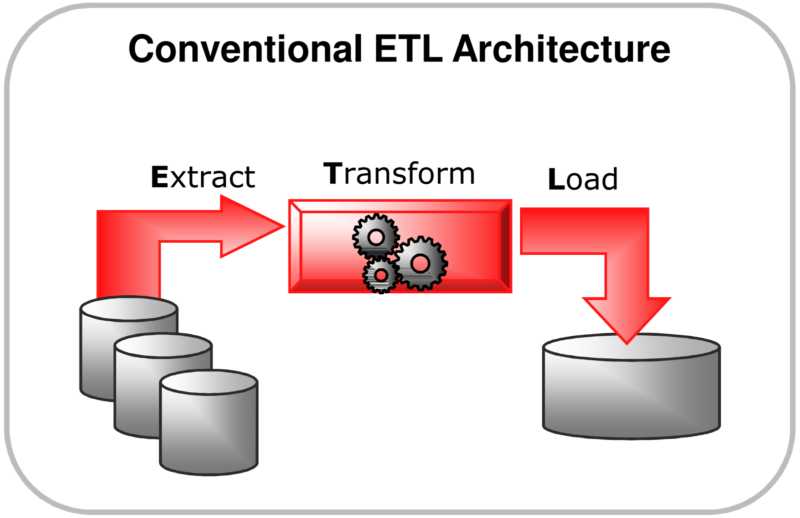
ODI: No New Servers

- **Lower Cost:** Leverage Compute Resources & Partition Workload efficiently
- **Efficient:** Exploits Database Optimizer
- **Fast:** Exploits Native Bulk Load & Other Database Interfaces
- **Scalable:** Scales as you add Processors to Source or Target
- **Manageability:** unified Enterprise Manager

Benefits

- Better Hardware Leverage
- Easier to Manage & Lower Cost
- Simple Tuning & Linear Scalability

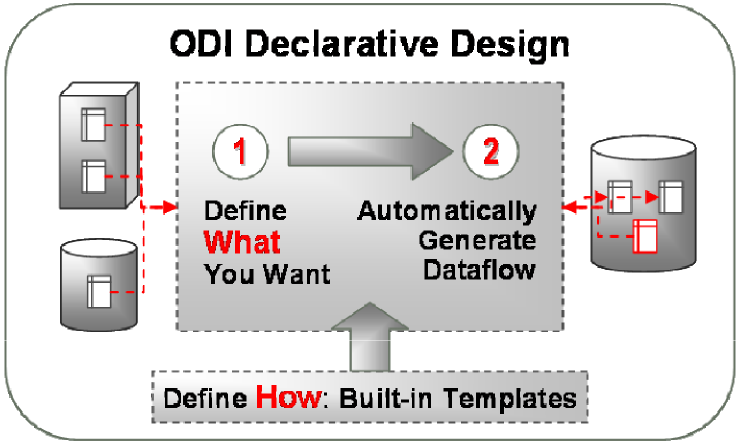


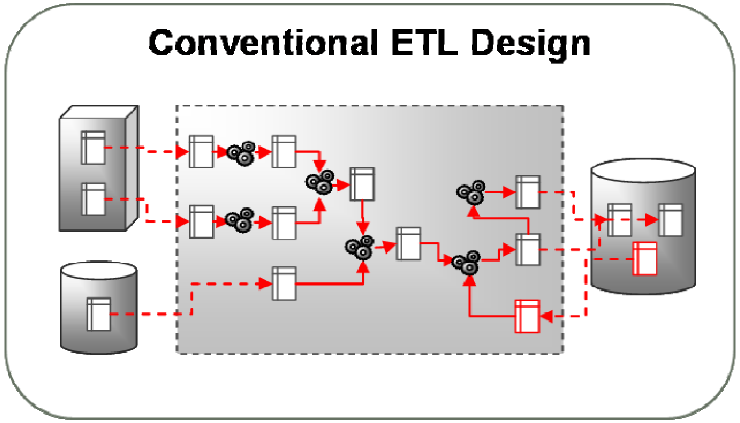


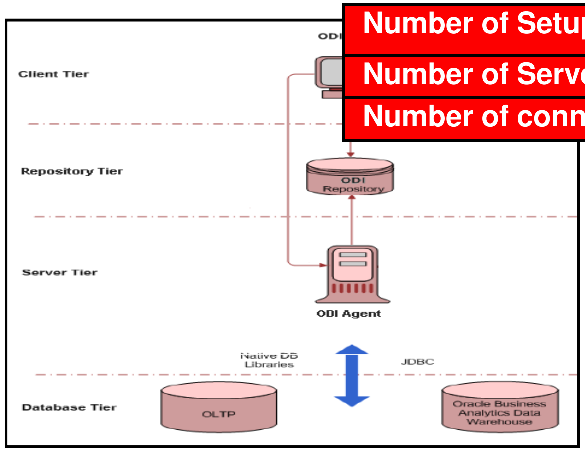
ODI is Simpler

Speed Project Delivery and Time to Market with ODI

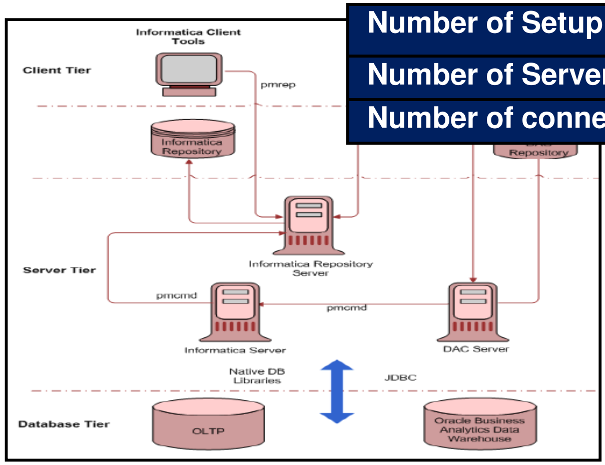
- Development Productivity
  - 40% Efficiency Gains
- Environment Setup (ex: BI Apps)
  - 33-50% Less Complex


ODI Declarative Design
1
2
Define What You Want
Automatically Generate Dataflow
Define How: Built-in Templates





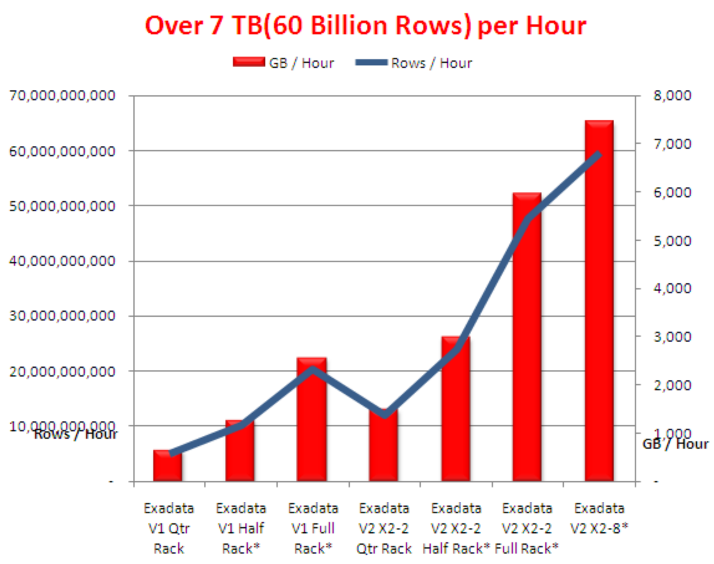
Number of Setup Steps	7
Number of Servers	1
Number of connections	3

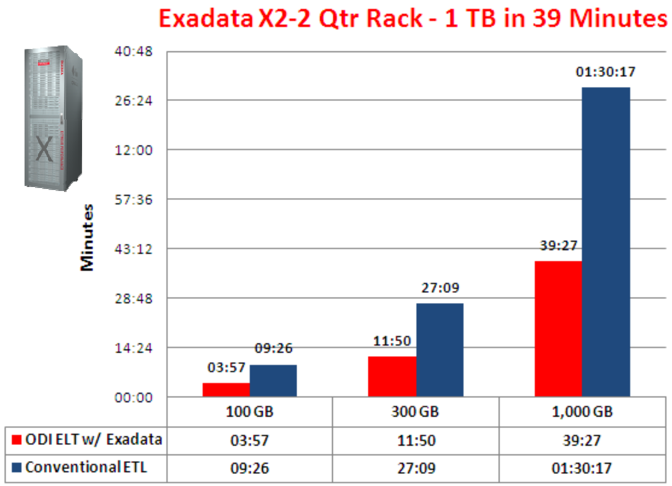


Number of Setup Steps	10
Number of Servers	3
Number of connections	7

ODI is Faster

Up to 7TB per hour of real world data loading and complex transformations





	100 GB	300 GB	1,000 GB
🟥 ODI ELT w/ Exadata	03:57	11:50	39:27
🟦 Conventional ETL	09:26	27:09	01:30:17

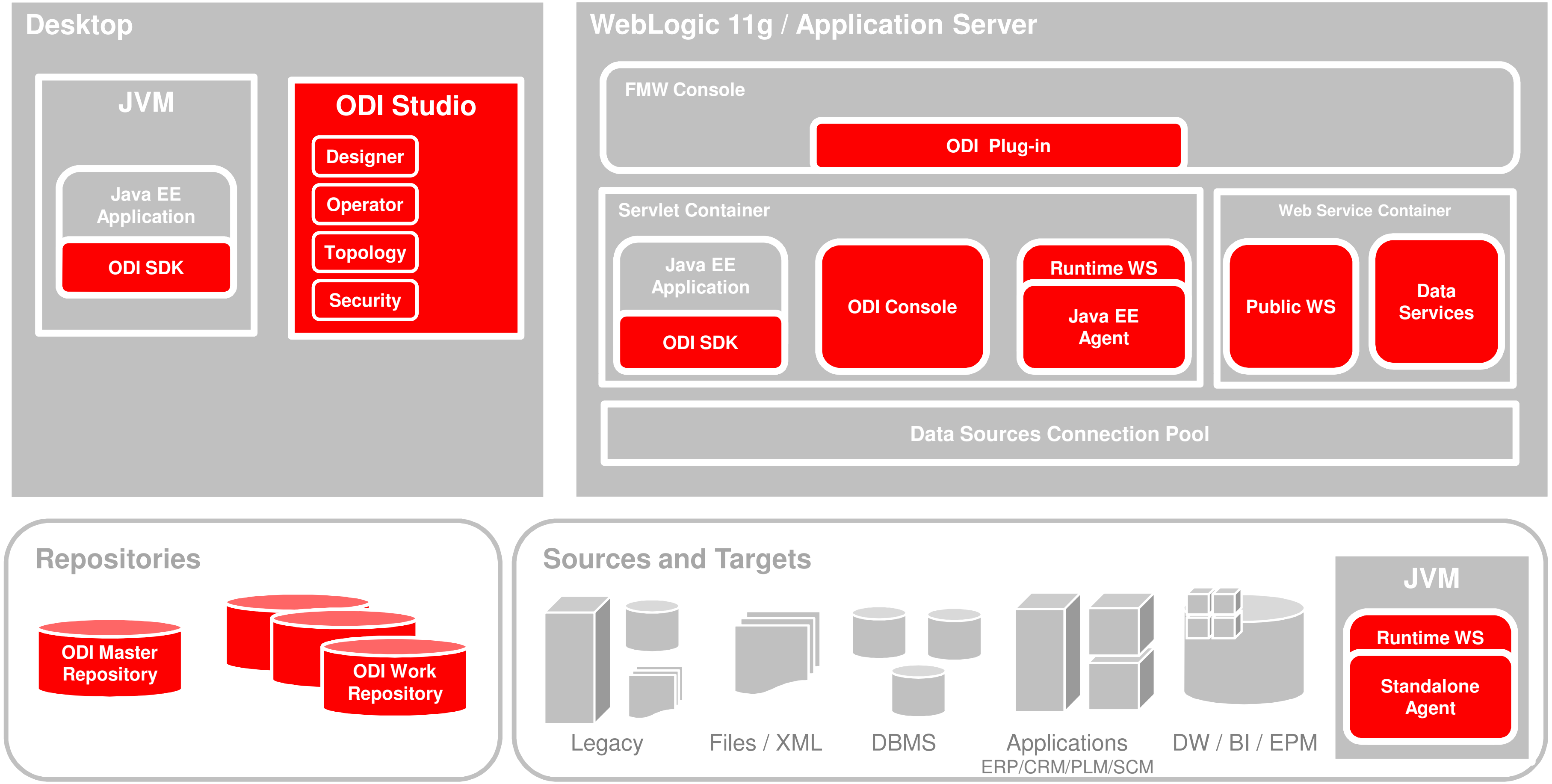
ODI ELT (on Exadata/any DW)

- ODI scales with Exadata
  - Loads increase linearly as Exadata scales
- ODI runs on Exadata – no ETL hardware required
  - No new hardware required as data sets grow
  - ODI processes used only during integration runs
  - Exadata continually available for OLTP, BI, DW, etc
- Common administration, monitoring and management
- All the benefits of rapid tools-based ETL development

Conventional ETL

- As data sets grow, more hardware ($$) needed to scale
- ETL parallel optimization and design ($$$) is heavily dependent on resources available to the ETL environment
  - Sources, integrations, targets must be designed to match processing power of ETL environment
  - Source flat files split to match # of ETL engine CPU's
  - Integration grid setup appropriately to match # of ETL engine CPU's
  - Target partitions, table spaces to match # of ETL engine CPU's
- ETL engine hardware resources only used for ETL
  - Cannot be utilized for OLTP, BI, DW, etc.
- Hardware not co located, multiple vendors
- Different management, monitoring and administration from database and BI infrastructure ($$)

Product Architecture
for Enterprise Scale Deployment



Unique Qualities of Enterprise Data Quality

Integrated DQ Solution

- Seamless integration of all core DQ capabilities
- Engineered for business users
- Integrated team collaboration and management

Modern Architecture

- Easy to configure and integrate ‘DQ Services’
- Modern, open architecture (Java, SOA, etc.)
- Collaborative, multi-user project support

Designed as a Platform

- Enables innovation and reuse
- Enables delivery of complete business solutions
- Allows partners opportunity to showcase their expertise and specialization

Oracle EDQ for Customer Data

Fastest Time To Value

- Easy to use - Single screen combines all key DQ functions
- Simple graphical UI, no coding required
- Collaborative environment with integrated issue management
- Designed for business user

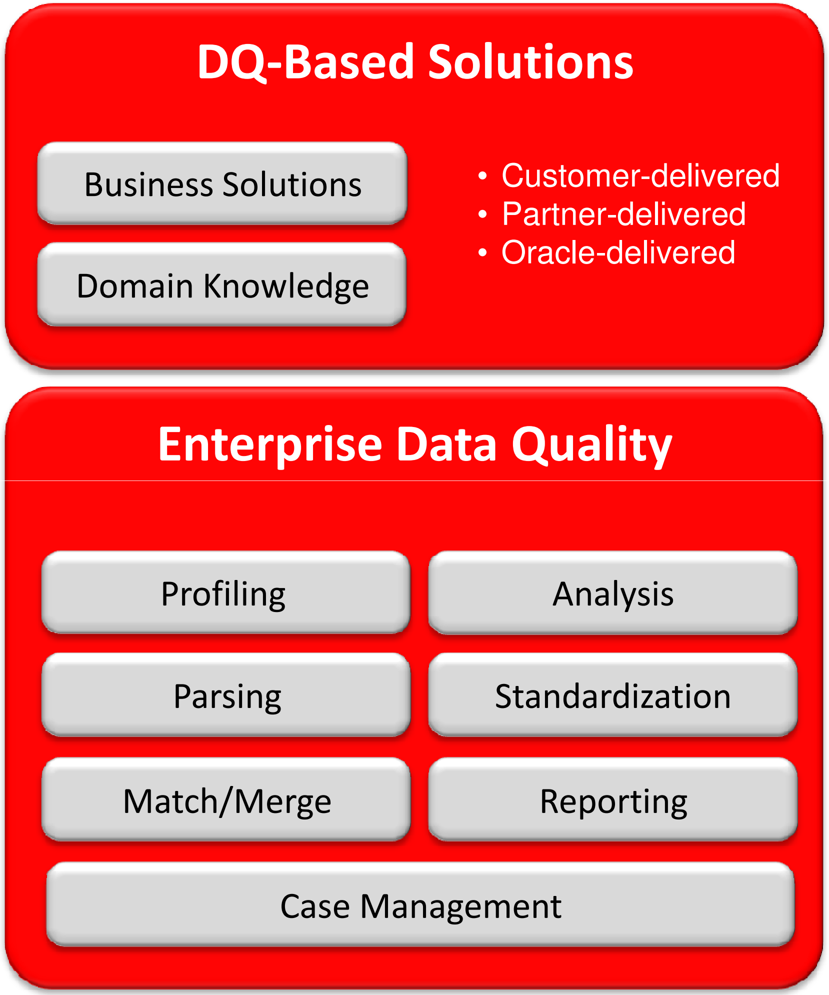
Most Tunable Solution

- Fully tunable rules
- Closed-loop rule building with instant feedback
- No 'black boxes'
- Easy to extend and re-use

Lowest Total Cost of Ownership

- Integrated solution
- Rapid integration and deployment
- Low maintenance overheads
- Scalable performance

Oracle EDQ for Customer Data



DQ-Based Solutions

Business Solutions

Domain Knowledge

- Customer-delivered
- Partner-delivered
- Oracle-delivered

Enterprise Data Quality

Profiling

Analysis

Parsing

Standardization

Match/Merge

Reporting

Case Management

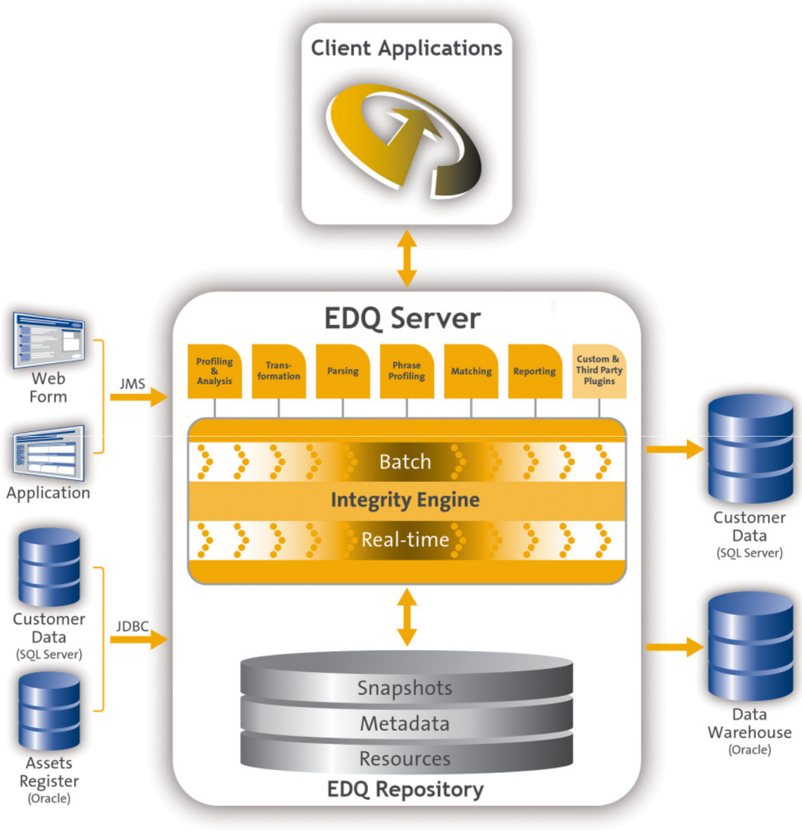
Solutions

- Pre-configured business solutions and domain knowledge
- Can be delivered by
  - Customers
  - Partners
  - Oracle
- Solutions can be components, processes, domain knowledge, up to complete applications

Data Quality Platform

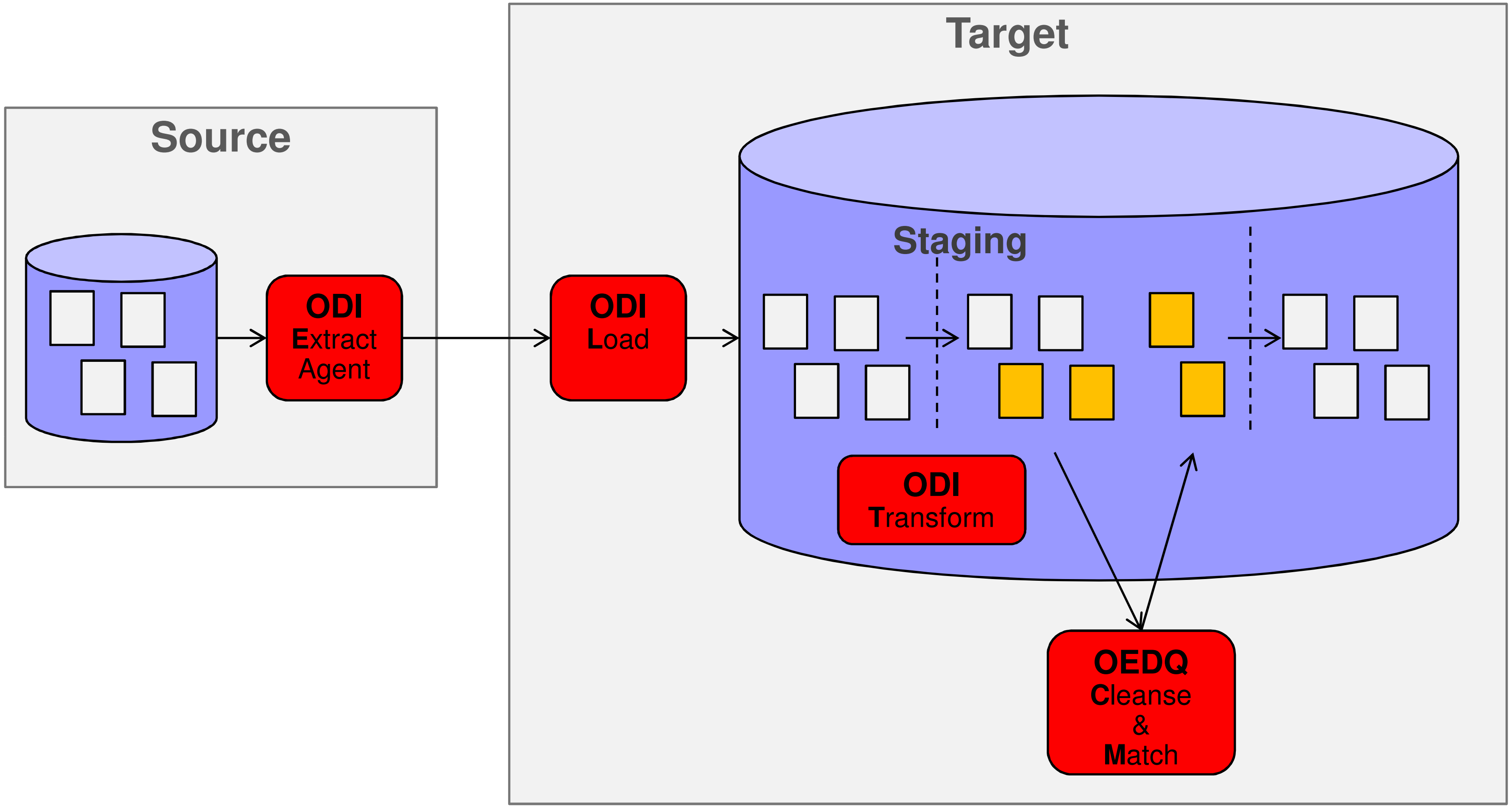
- End-to-end enterprise data quality solution spanning both customer and product data quality
- Highly configurable to match specific business needs
- Case management tools for tracking and web-based KPI reporting for increased productivity

Product Architecture



- All Java Server (Stateless)
- Java Webstart Client Applications
- Fully integrated with a single repository and UI
- Batch and Real-time Execution
- Connects to virtually any source/target of data
- Platform Independent

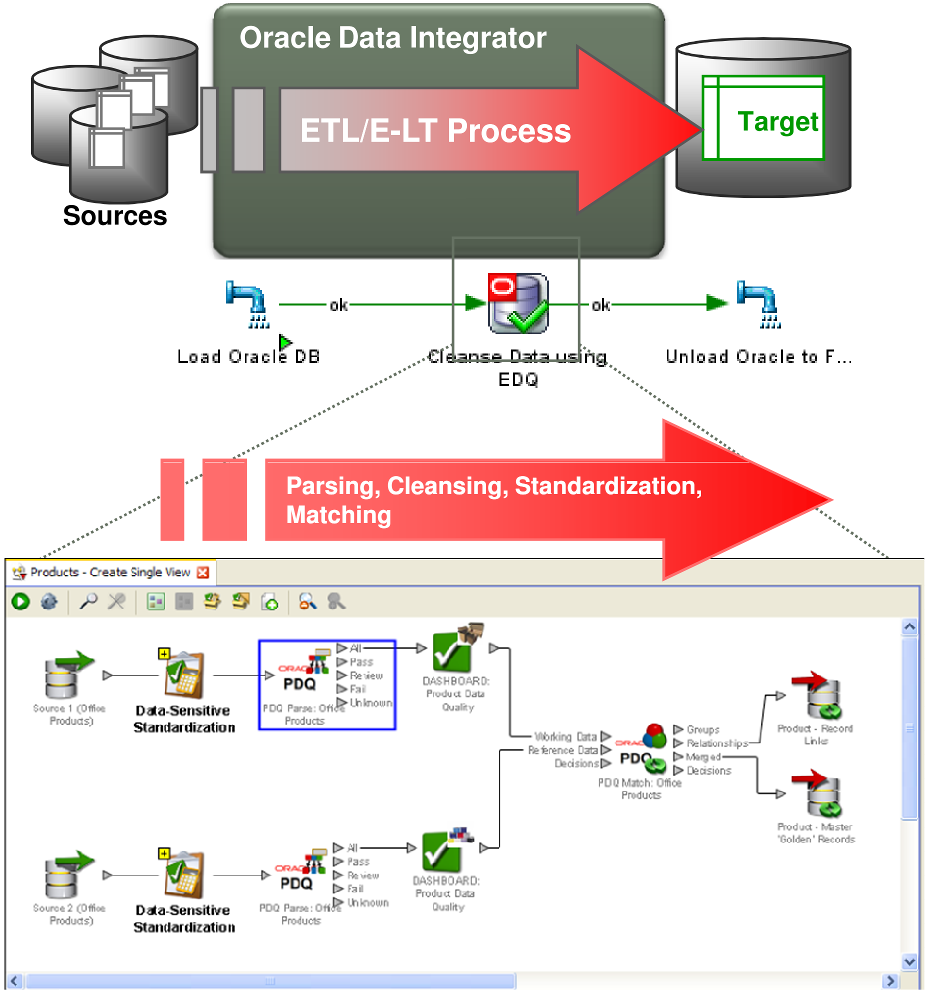
ODI & OEDQ – ELT+CM



Extract > **L**oad > **T**ransform > **C**leanse > **M**atch

Integrated Data Quality with ODI

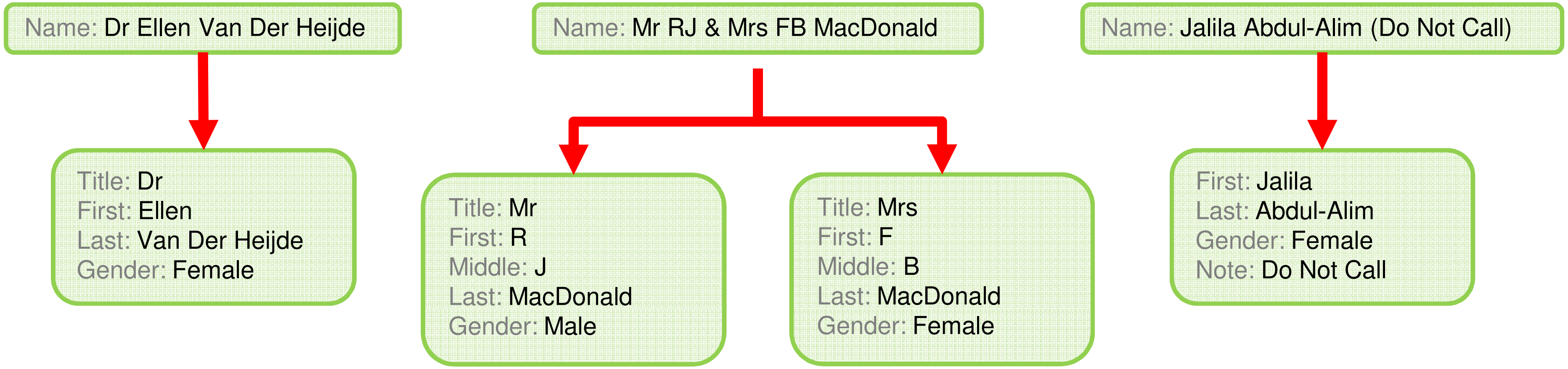
Oracle Enterprise Data Quality Runtime with Data Integrator



Oracle Data Integrator
ETL/E-LT Process
Sources
Target
ok
ok
Load Oracle DB
Cleanse Data using EDQ
Unload Oracle to F...
Parsing, Cleansing, Standardization, Matching
Products - Create Single View
Source 1 (Office Products)
Data-Sensitive Standardization
PDQ
PDQ Parse: Office Products
All
Pass
Review
Fail
Unknown
DASHBOARD: Product Data Quality
Working Data
Reference Data
Decisions
PDQ Match: Office Products
Groups
Relationships
Merged
Decisions
Product - Record Links
Product - Master 'Golden' Records
Source 2 (Office Products)
Data-Sensitive Standardization
PDQ
PDQ Parse: Office Products
All
Pass
Review
Fail
Unknown
DASHBOARD: Product Data Quality

- Best of breed Quality
  - Proven, scalable DQ engine
  - Rich capabilities for cleansing, standardization, validation, match and merge
  - Extensible by customers
- Out-of-box integration
  - ODI integrates with Quality functions via pre-built ODI OpenTool
  - Drag and drop graphical icon for inserting DQ flows into ODI

Oracle Enterprise Data Quality – Standardization



Name: Dr Ellen Van Der Heijde

Title: Dr
First: Ellen
Last: Van Der Heijde
Gender: Female

Name: Mr RJ & Mrs FB MacDonald

Title: Mr
First: R
Middle: J
Last: MacDonald
Gender: Male

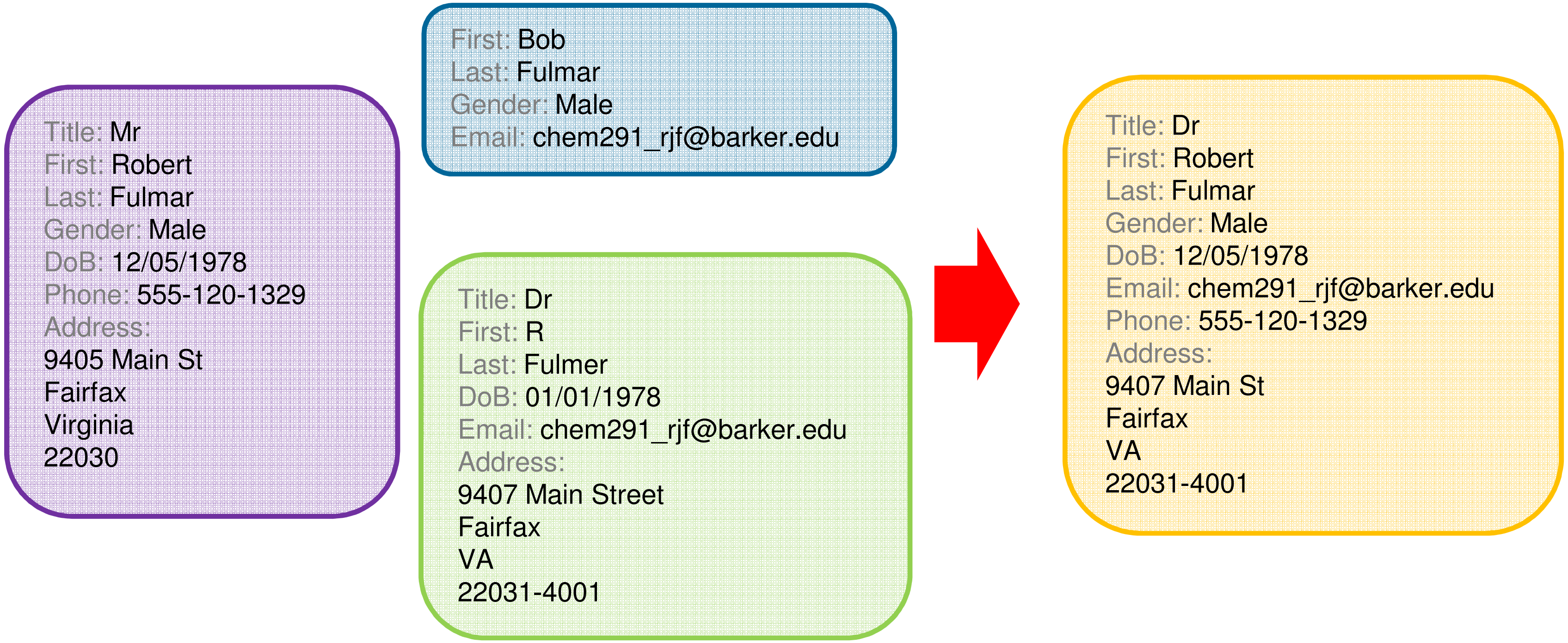
Title: Mrs
First: F
Middle: B
Last: MacDonald
Gender: Female

Name: Jalila Abdul-Alim (Do Not Call)

First: Jalila
Last: Abdul-Alim
Gender: Female
Note: Do Not Call

- Standardize, Transform and Parse
- Split names and name elements
- Identify individuals and businesses
- Derive additional attributes

Oracle Enterprise Data Quality - Matching



Title: Mr
First: Robert
Last: Fulmar
Gender: Male
DoB: 12/05/1978
Phone: 555-120-1329
Address:
9405 Main St
Fairfax
Virginia
22030

First: Bob
Last: Fulmar
Gender: Male
Email: chem291_rjf@barker.edu

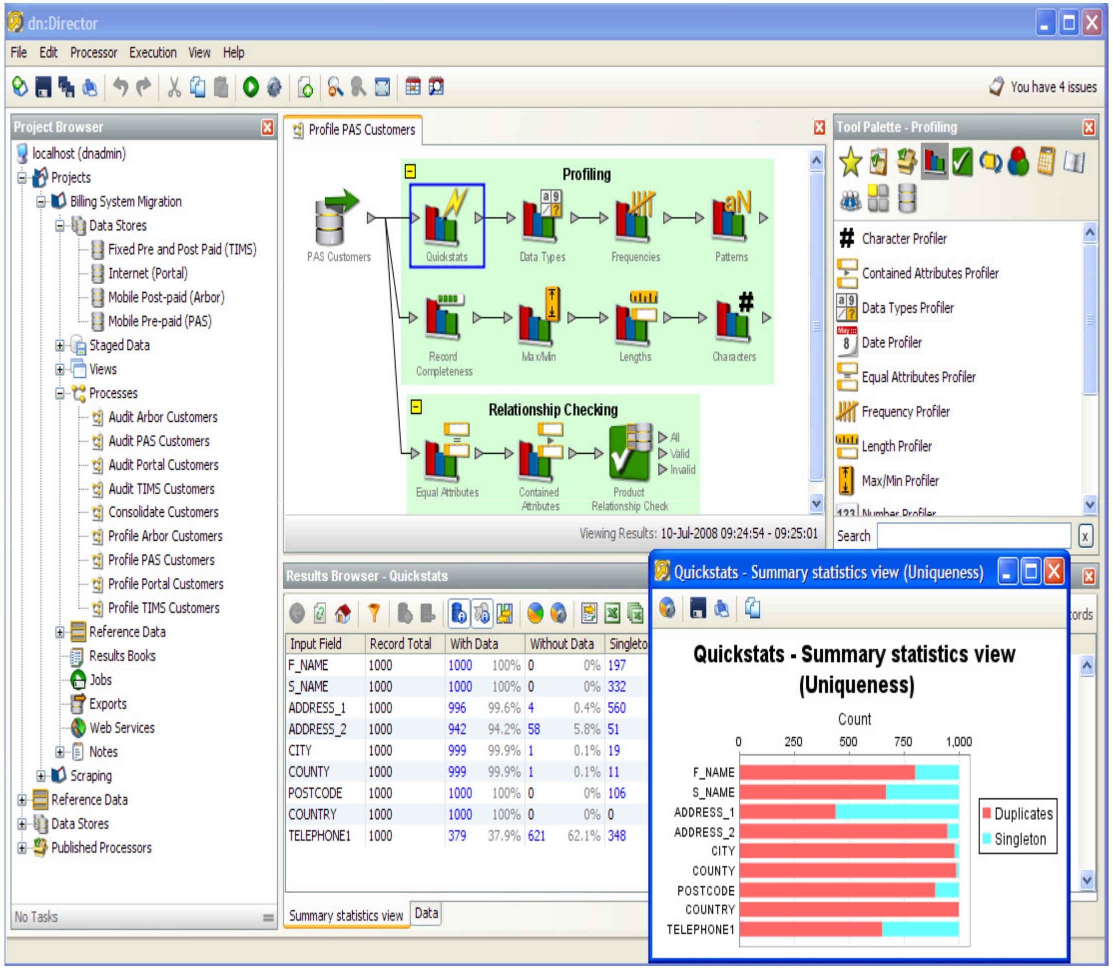
Title: Dr
First: R
Last: Fulmer
DoB: 01/01/1978
Email: chem291_rjf@barker.edu
Address:
9407 Main Street
Fairfax
VA
22031-4001

Title: Dr
First: Robert
Last: Fulmar
Gender: Male
DoB: 12/05/1978
Email: chem291_rjf@barker.edu
Phone: 555-120-1329
Address:
9407 Main St
Fairfax
VA
22031-4001

- Match & Merge data from disparate sources
- Create ‘best’ record based on survivorship rules

Built for the Business User

- Short learning curve & time-to-value
- Solution for owners of the business problem
- Integrated team collaboration



Input Field	Record Total	With Data		Without Data		Singleto
F_NAME	1000	1000	100%	0	0%	197
S_NAME	1000	1000	100%	0	0%	332
ADDRESS_1	1000	996	99.6%	4	0.4%	560
ADDRESS_2	1000	942	94.2%	58	5.8%	51
CITY	1000	999	99.9%	1	0.1%	19
COUNTY	1000	999	99.9%	1	0.1%	11
POSTCODE	1000	1000	100%	0	0%	106
COUNTRY	1000	1000	100%	0	0%	0
TELEPHONE1	1000	379	37.9%	621	62.1%	348

Oracle EDQ for Product Data

Built from the ground up for product data

- Handles the variability of product data – structure, standards, categories
- Use Customer's data to build references

Ability to govern a largely un-governed process

- Stewardship, oversight and remediation combined in a single interface
- Optimal combination of automation and remediation

The Product Data Problem – Unstructured & Non-Standard

What is this?

10hp motor 115V Yoke mount

MOT-10,115V, 48YZ,YOKE

mtr, ac(115) 10 horsepower 115volts

This 10hp yoke mounted motor is rated for 115V with a 5 year warranty

10 Caballos, Motor, 115 Voltios

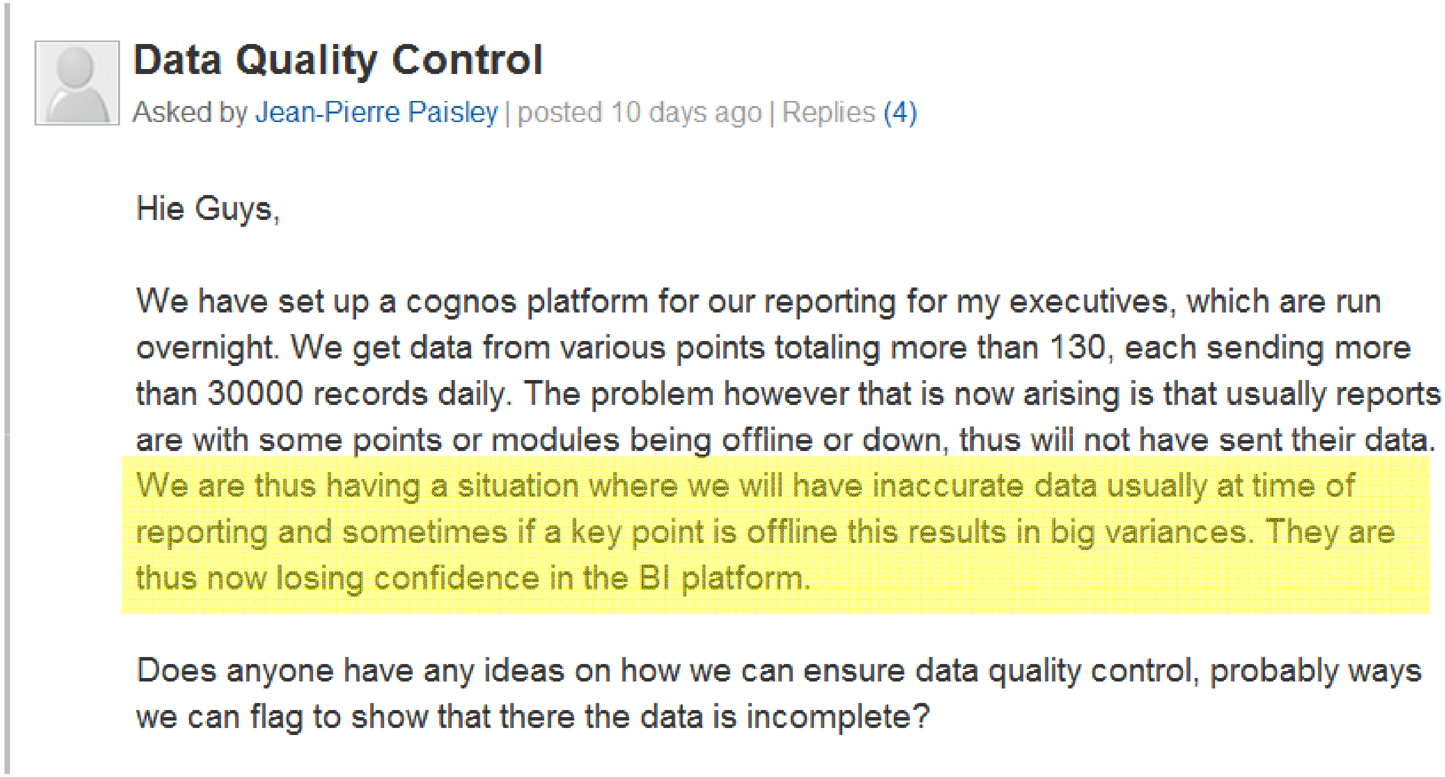
TEAO HP = 10.0 1725RPM 115V 48YZ YOKE MTR

Motor, TEAO, 1725 RPM, 48YZ, 15 Voltios, Montaje de Yugo, hp = 10

Item	Motor
Classification	26101600
Power	10 horsepower
Voltage	115
Mounting	Yoke

Product data is much more variable and unpredictable than other data types

An example in DWHs…



Data Quality Control

Asked by Jean-Pierre Paisley | posted 10 days ago | Replies (4)

Hie Guys,

We have set up a cognos platform for our reporting for my executives, which are run overnight. We get data from various points totaling more than 130, each sending more than 30000 records daily. The problem however that is now arising is that usually reports are with some points or modules being offline or down, thus will not have sent their data. We are thus having a situation where we will have inaccurate data usually at time of reporting and sometimes if a key point is offline this results in big variances. They are thus now losing confidence in the BI platform.

Does anyone have any ideas on how we can ensure data quality control, probably ways we can flag to show that there the data is incomplete?

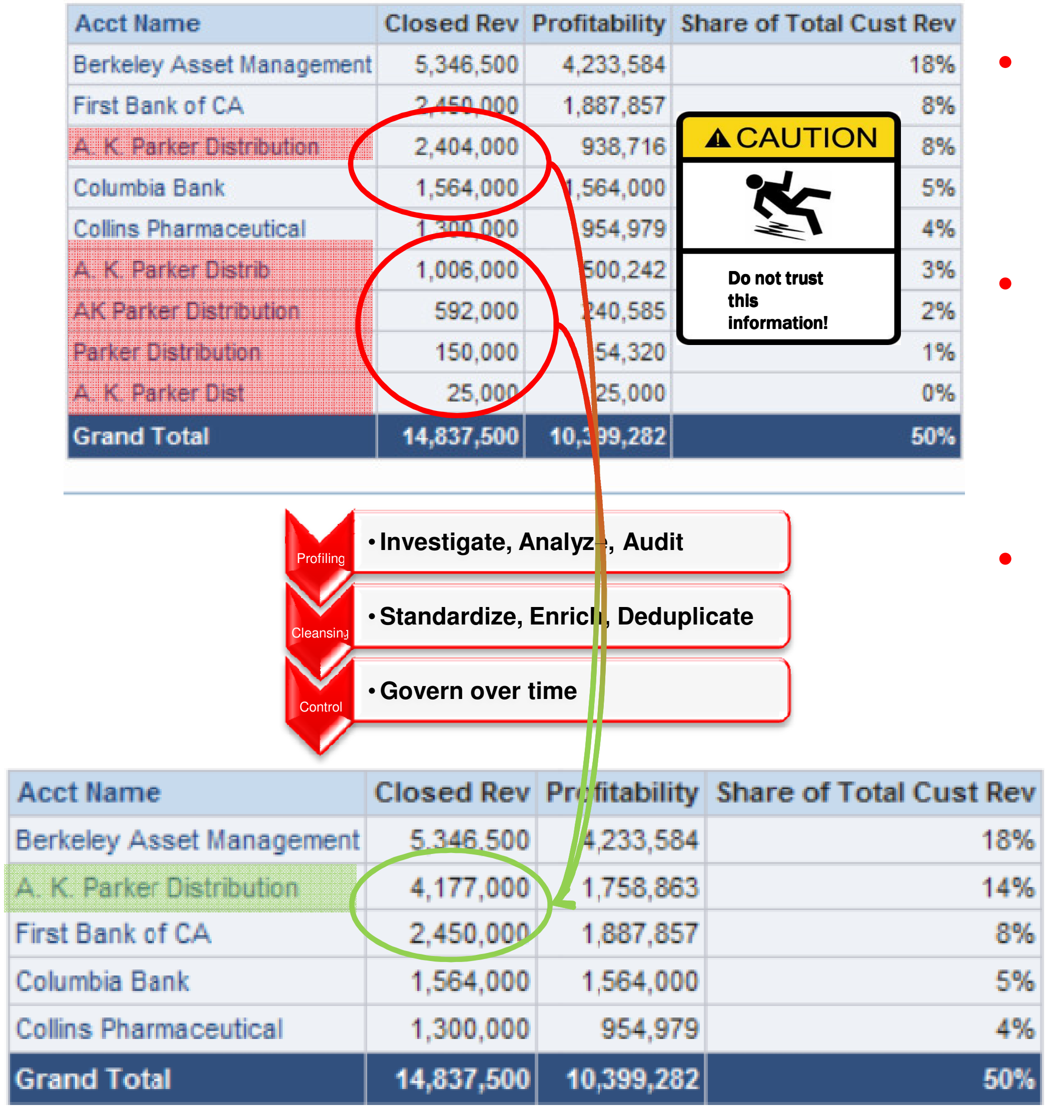
Build your Data Analysis strategy on trustable data

Data Quality for Business Intelligence

- Get Information from data is always a challenge:
  - Because of lack of standards, or different standards across different sources
  - Because of missing values and typos
  - Robust Matching features are the only that solve duplicate issues
- Business executors always need trustable data:
  - For taking the right decision
  - To get business insights and further develop the business
  - Discover potential gaps
- Customer believe in Oracle DQ strategies because:
  - Best of breed technology
  - Most flexible and customizable as per customer's needs

Next Gen Data Warehouse

Change data into valuable Information



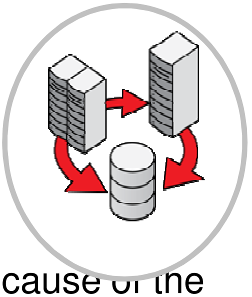
Acct Name	Closed Rev	Profitability	Share of Total Cust Rev
Berkeley Asset Management	5,346,500	4,233,584	18%
First Bank of CA	2,450,000	1,887,857	8%
A. K. Parker Distribution	2,404,000	938,716	8%
Columbia Bank	1,564,000	1,564,000	5%
Collins Pharmaceutical	1,300,000	954,979	4%
A. K. Parker Distrib	1,006,000	500,242	3%
AK Parker Distribution	592,000	240,585	2%
Parker Distribution	150,000	54,320	1%
A. K. Parker Dist	25,000	25,000	0%
Grand Total	14,837,500	10,399,282	50%

⚠ CAUTION

Do not trust this information!

- Investigate, Analyze, Audit
- Standardize, Enrich, Deduplicate
- Govern over time

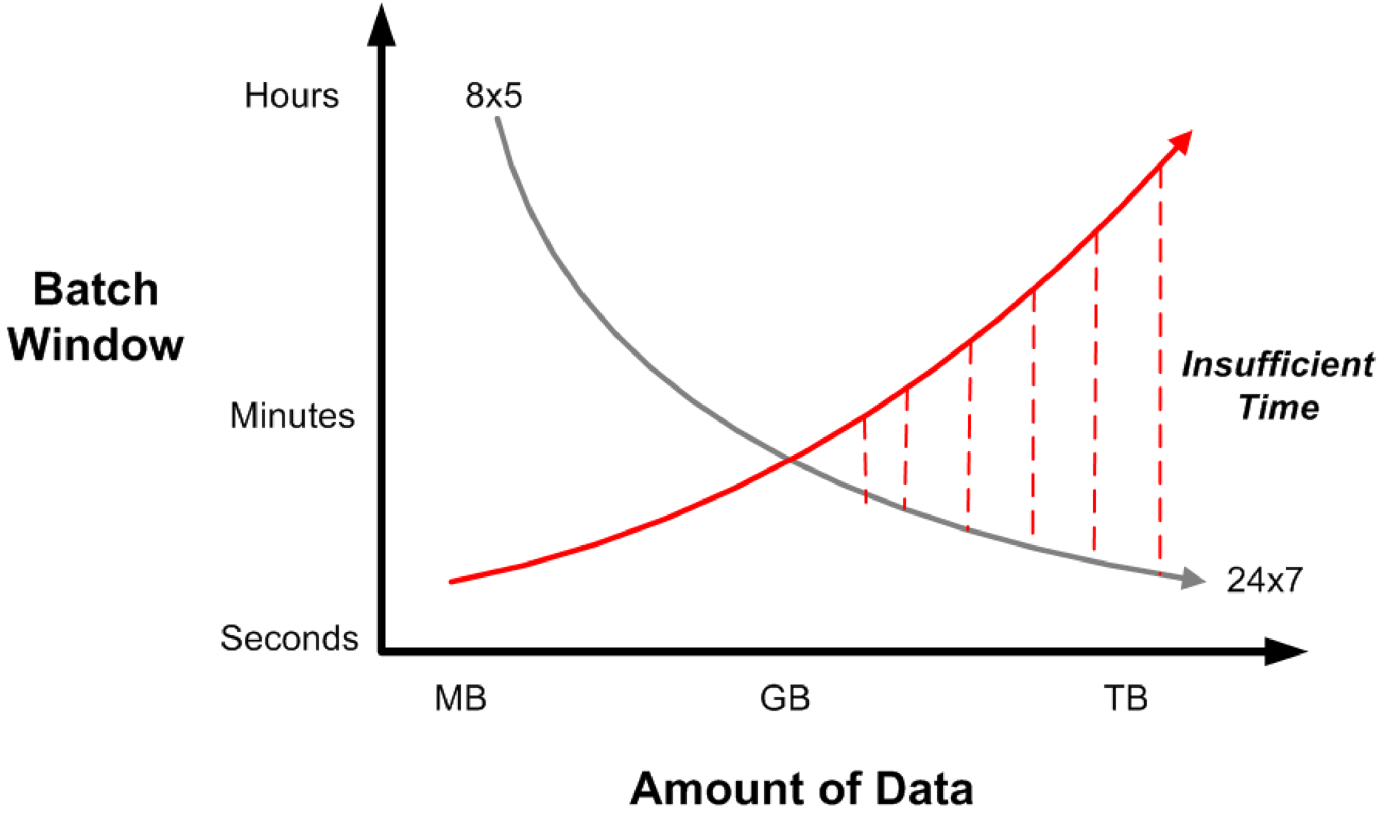
Acct Name	Closed Rev	Profitability	Share of Total Cust Rev
Berkeley Asset Management	5,346,500	4,233,584	18%
A. K. Parker Distribution	4,177,000	1,758,863	14%
First Bank of CA	2,450,000	1,887,857	8%
Columbia Bank	1,564,000	1,564,000	5%
Collins Pharmaceutical	1,300,000	954,979	4%
Grand Total	14,837,500	10,399,282	50%



- **The Business Issue**
  - BI Reports are not trustable, because of the state of source data
- **Reduce risks**
  - Improve data quality by integrating cleansing as part of the process
  - Eliminate data redundancies
- **Improve Business Insights**
  - Improved business insight with improved data quality
  - Better profiling of data to eliminate gaps in insight

Mission-Critical Systems and Batch Processing

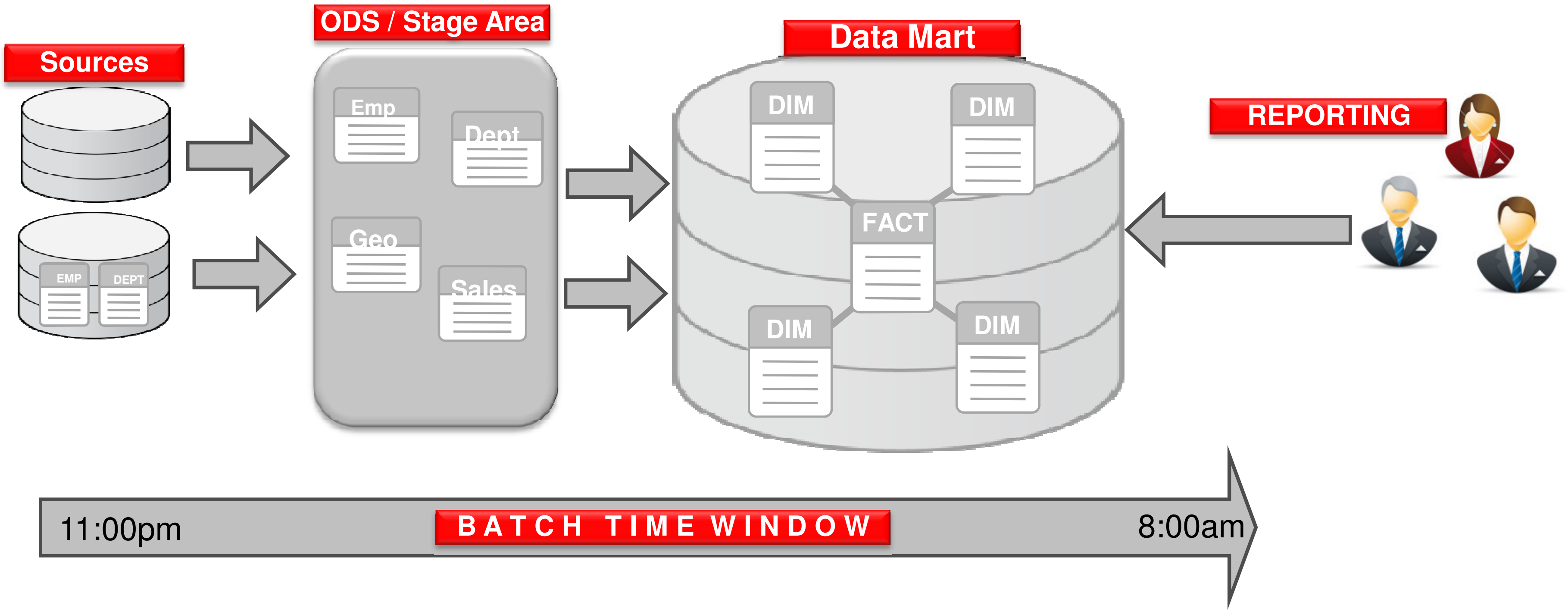
Too Much Data, Not Enough Time



What time of your day is your business *NOT* at it's peak?

E-LTQ In-line Predictive Quality

What if you go longer than the batch window?

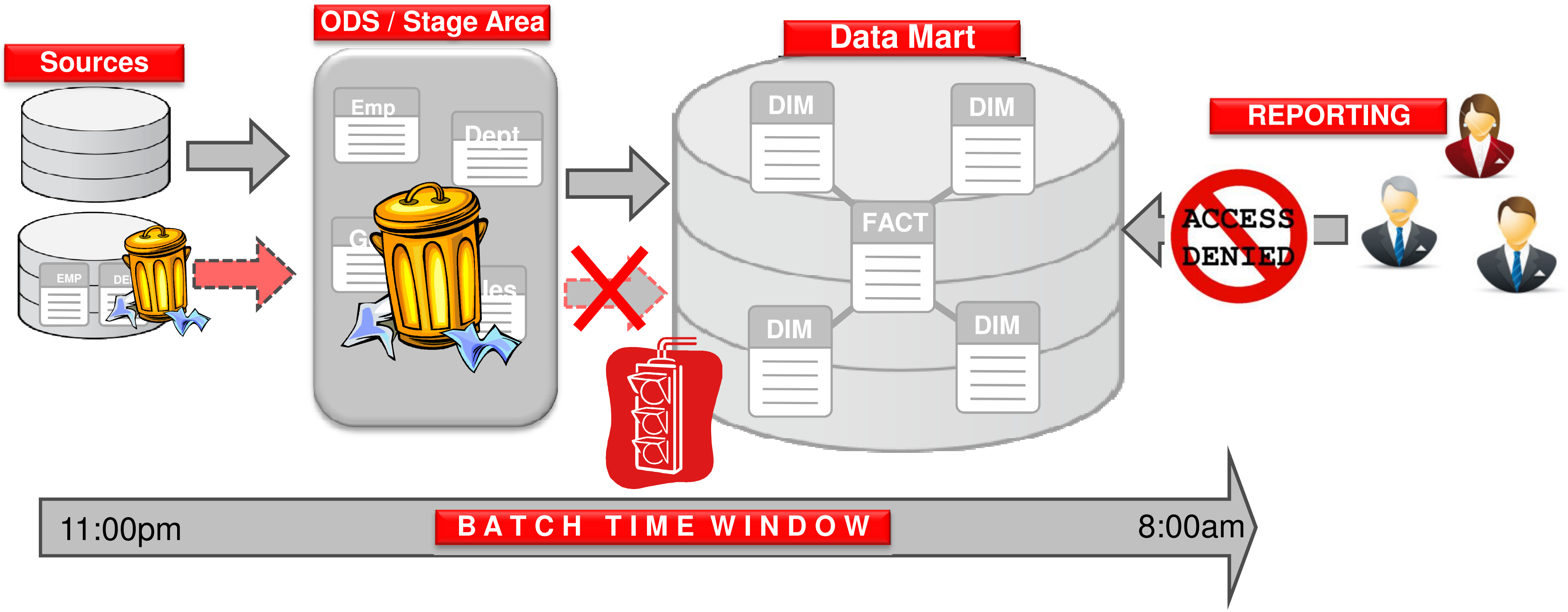


Data are loaded, transformed during a batch time window, before users and applications get access.
This windows easily becomes a challenge because of:

- data volume increasing
- only static controls can be applied on the flow, eventually discarding bad data. Discarded data potentially generate inconsistencies on the final target.

E-LTQ In-line Predictive Quality

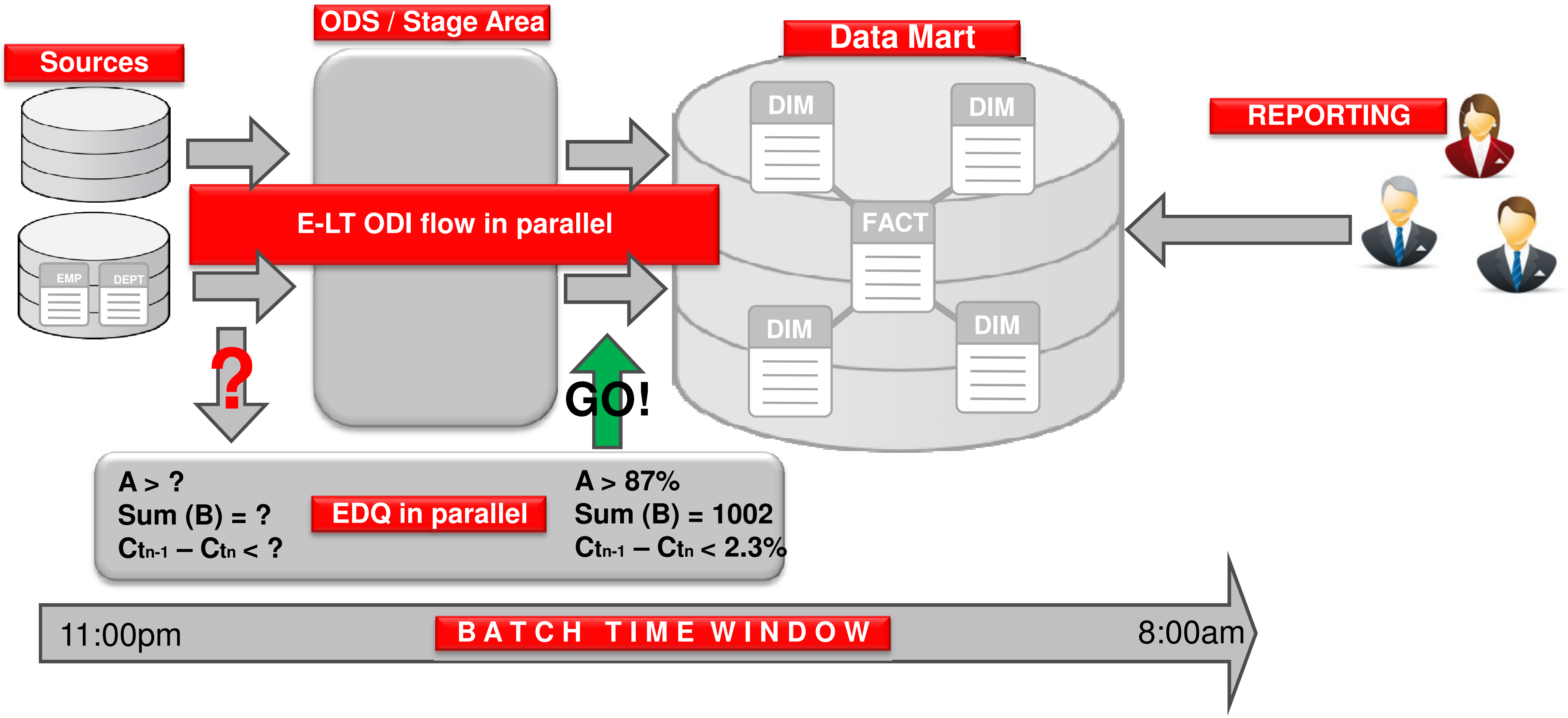
What if you discard or, worse, you load, bad data?



If bad data are processed without cleansing or discarded , the Data mart cannot be accessed until bad data are fixed. This usually is done by IT operations, with lot of efforts, Users & Manager pressing for restoring data ASAP.

E-LTQ In-line Predictive Quality

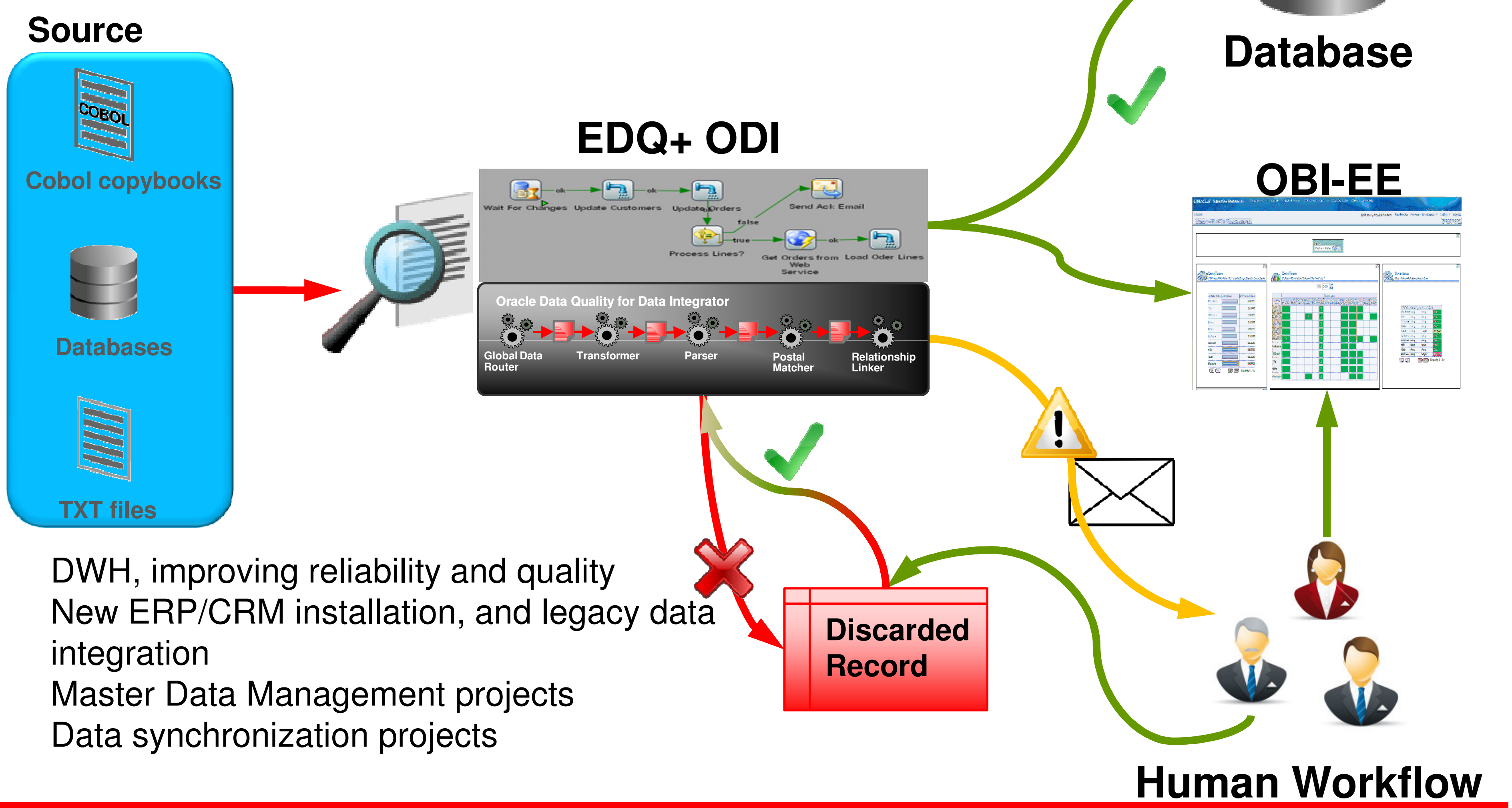
Check your data dynamically, ensure Quality



Check while you transform & Load your
Business Thresholds A > 80%, sum(B) = 1000, $C_{tn-1} - C_{tn} < 3\%$

Data Quality Firewall

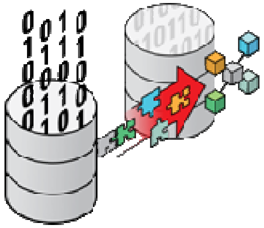
profile, repair, check, alert, report

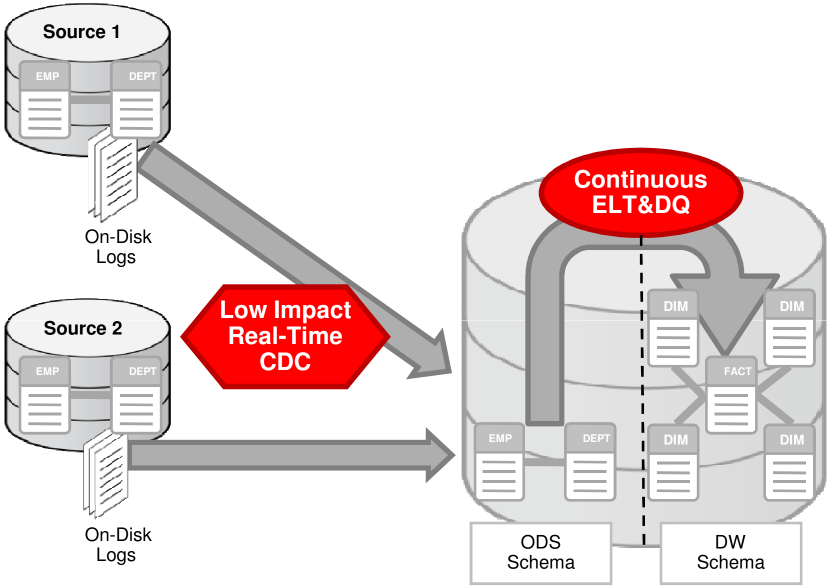


- DWH, improving reliability and quality
- New ERP/CRM installation, and legacy data integration
- Master Data Management projects
- Data synchronization projects

Business Intelligence

Real Time Data Warehouse



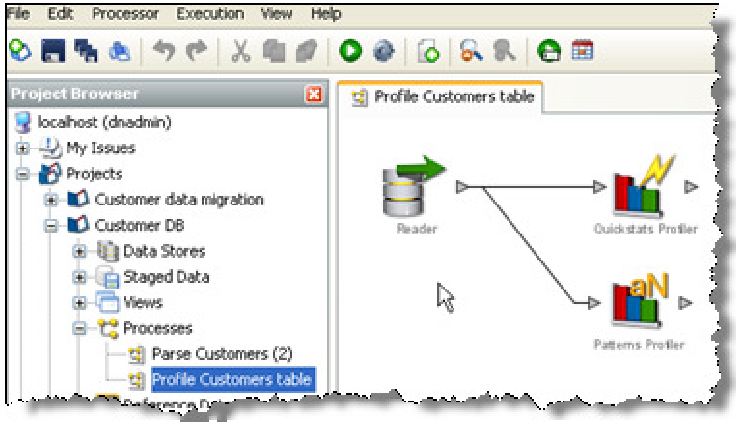


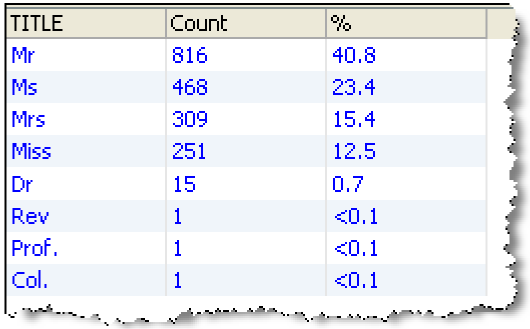
- **Solution**
  - Log-based capture of database transactions from source systems
  - Load to target with sub-second latency
  - Transformation performed on the database using E-LT in continuous mini-batches

- **Benefits**
  - No resource / performance impact to OLTP
  - Fresh data available for better decision making
  - Get double-duty from database investment by using it for transformations
  - No batch windows necessary – key for global businesses

Key Capabilities

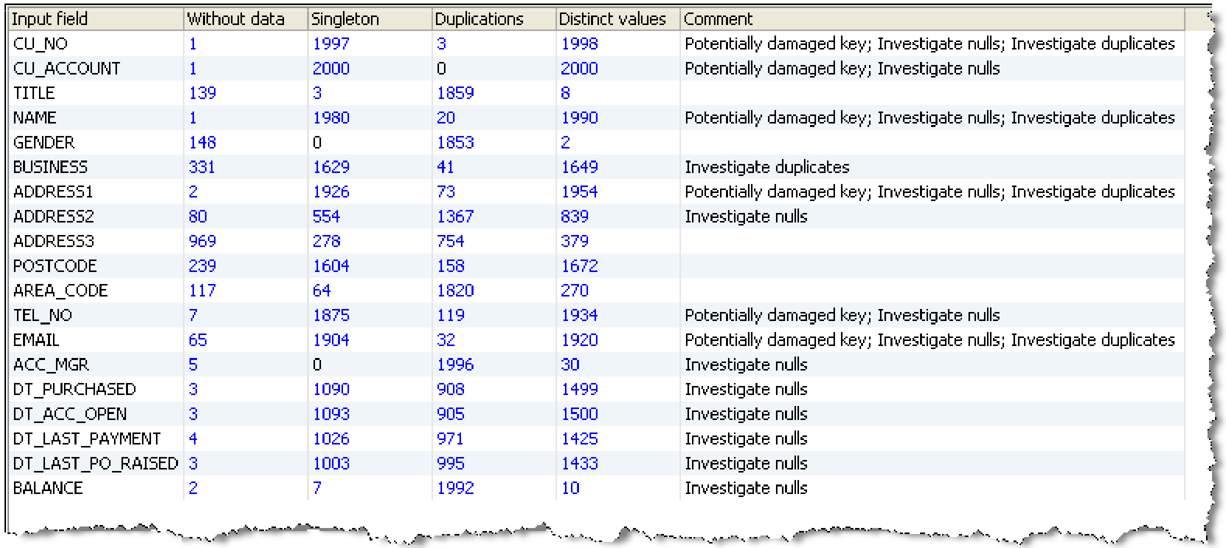
Profiling – Understand the data first





TITLE	Count	%
Mr	816	40.8
Ms	468	23.4
Mrs	309	15.4
Miss	251	12.5
Dr	15	0.7
Rev	1	<0.1
Prof.	1	<0.1
Col.	1	<0.1

Interactive exploration of data, identifying distribution and outlying values with drill-downs

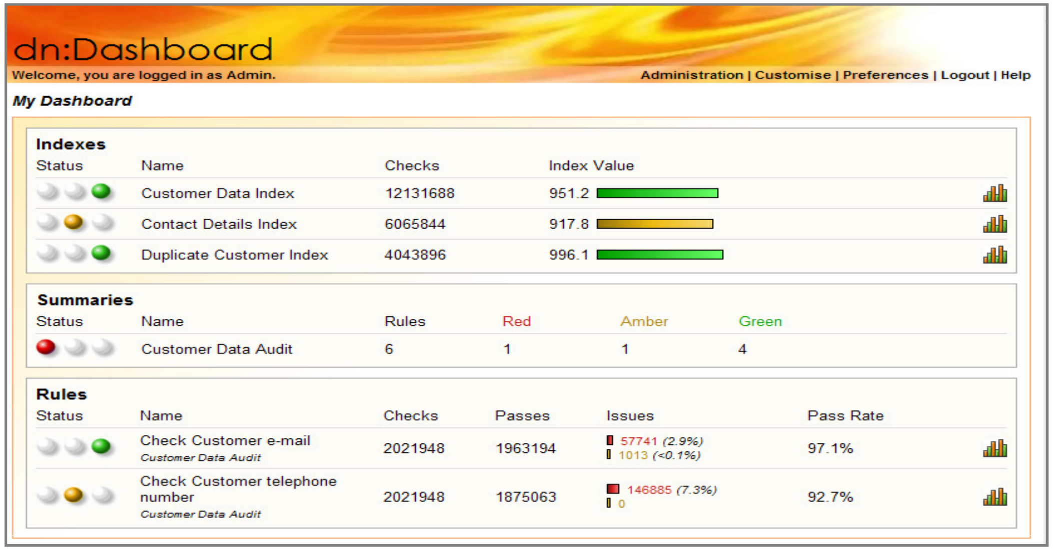


Input field	Without data	Singleton	Duplications	Distinct values	Comment
CU_NO	1	1997	3	1998	Potentially damaged key; Investigate nulls; Investigate duplicates
CU_ACCOUNT	1	2000	0	2000	Potentially damaged key; Investigate nulls
TITLE	139	3	1859	8	
NAME	1	1980	20	1990	Potentially damaged key; Investigate nulls; Investigate duplicates
GENDER	148	0	1853	2	
BUSINESS	331	1629	41	1649	Investigate duplicates
ADDRESS1	2	1926	73	1954	Potentially damaged key; Investigate nulls; Investigate duplicates
ADDRESS2	80	554	1367	839	Investigate nulls
ADDRESS3	969	278	754	379	
POSTCODE	239	1604	158	1672	
AREA_CODE	117	64	1820	270	
TEL_NO	7	1875	119	1934	Potentially damaged key; Investigate nulls
EMAIL	65	1904	32	1920	Potentially damaged key; Investigate nulls; Investigate duplicates
ACC_MGR	5	0	1996	30	Investigate nulls
DT_PURCHASED	3	1090	908	1499	Investigate nulls
DT_ACC_OPEN	3	1093	905	1500	Investigate nulls
DT_LAST_PAYMENT	4	1026	971	1425	Investigate nulls
DT_LAST_PO_RAISED	3	1003	995	1433	Investigate nulls
BALANCE	2	7	1992	10	Investigate nulls

Identify and quantify issues in data

Oracle Enterprise Data Quality – Audit

- Business Rules Check
- Cross Attribute Check
- Data Type Check
- Duplicate Check
- Email Check
- GBR Postcode Format Check
- Invalid Character Check
- Length Check
- List Check
- Logic Check
- Lookup Check
- No Data Check
- Pattern Check
- RegEx Check



dn:Dashboard

Welcome, you are logged in as Admin.

Administration | Customise | Preferences | Logout | Help

My Dashboard

Indexes

Status	Name	Checks	Index Value
○ ○ 🟢	Customer Data Index	12131688	951.2
○ 🟡 ○	Contact Details Index	6065844	917.8
○ ○ 🟢	Duplicate Customer Index	4043896	996.1

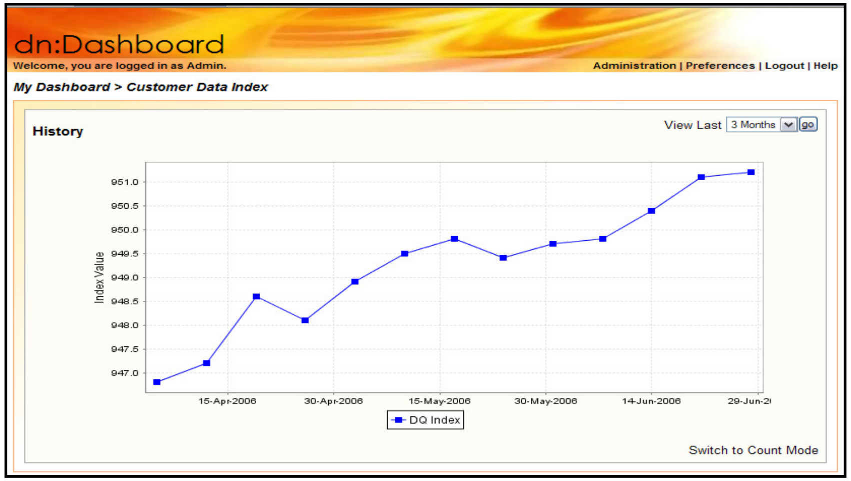
Summaries

Status	Name	Rules	Red	Amber	Green
🔴 ○ ○	Customer Data Audit	6	1	1	4

Rules

Status	Name	Checks	Passes	Issues	Pass Rate
○ ○ 🟢	Check Customer e-mail *Customer Data Audit*	2021948	1963194	57741 *(2.9%)* 1013 *(<0.1%)*	97.1%
○ 🟡 ○	Check Customer telephone number *Customer Data Audit*	2021948	1875063	146885 *(7.3%)* 0	92.7%

- Validate data against business rules
- Publish results to data quality dashboard



dn:Dashboard

Welcome, you are logged in as Admin.

Administration | Preferences | Logout | Help

My Dashboard > Customer Data Index

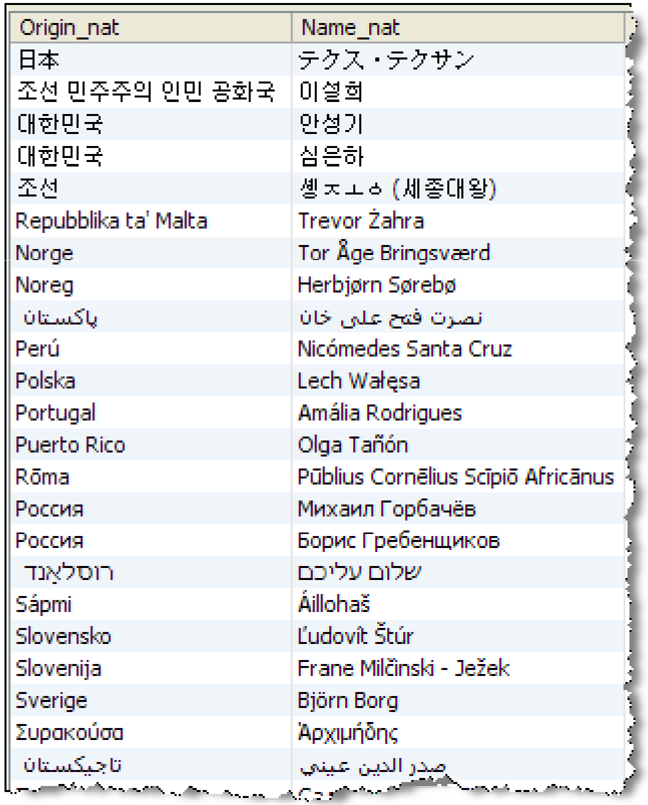
History

View Last 3 Months go

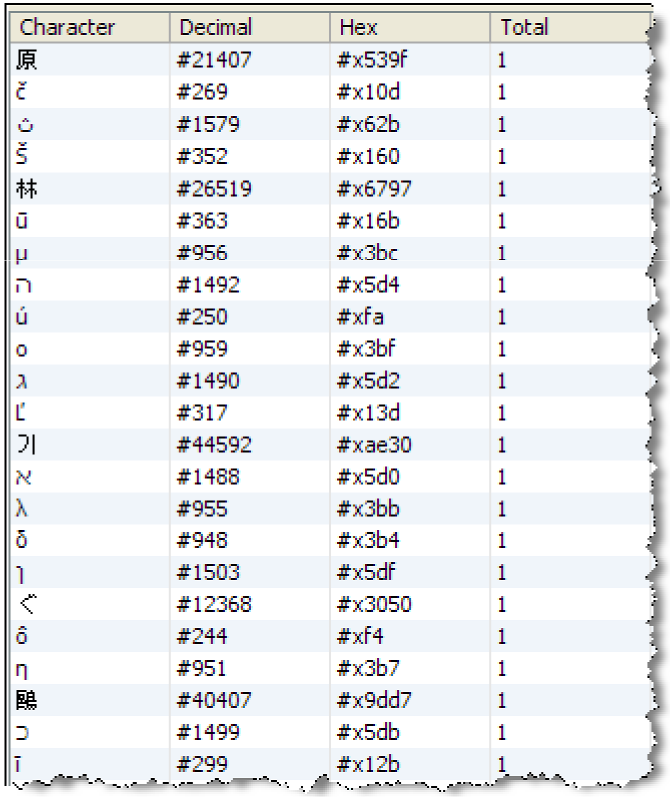
Switch to Count Mode

Key Features

- Fully Unicode Compliant



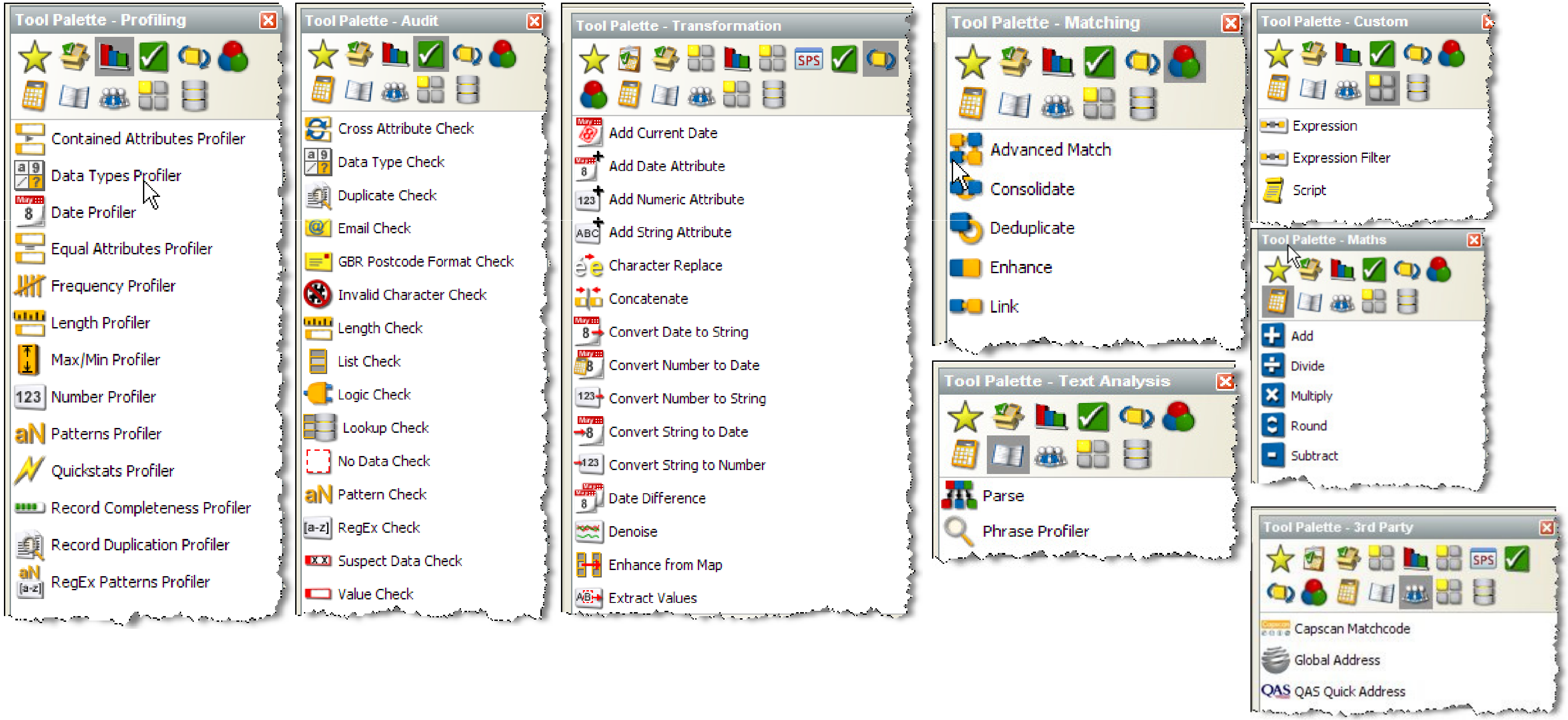
Origin_nat	Name_nat
日本	テクス・テクサン
조선 민주주의 인민 공화국	이설희
대한민국	안성기
대한민국	심은하
조선	솅ㅈㅗㆁ (세종대왕)
Repubblika ta' Malta	Trevor Żahra
Norge	Tor Åge Bringsværd
Noreg	Herbjørn Sørebø
پاکستان	نصرت فتح علی خان
Perú	Nicómedes Santa Cruz
Polska	Lech Wałęsa
Portugal	Amália Rodrigues
Puerto Rico	Olga Tañón
Rōma	Pūblius Cornēlius Scīpiō Africānus
Россия	Михаил Горбачёв
Россия	Борис Гребенщиков
רוסלאַנד	שלום עליכם
Sápmi	Áillohaš
Slovensko	Ľudovít Štúr
Slovenija	Frane Milčinski - Ježek
Sverige	Björn Borg
Συρακούσα	Ἀρχιμήδης
تاجیکستان	صدر الدین عینی



Character	Decimal	Hex	Total
原	#21407	#x539f	1
č	#269	#x10d	1
ث	#1579	#x62b	1
Š	#352	#x160	1
林	#26519	#x6797	1
ū	#363	#x16b	1
μ	#956	#x3bc	1
ה	#1492	#x5d4	1
ú	#250	#xfa	1
ο	#959	#x3bf	1
ג	#1490	#x5d2	1
Ľ	#317	#x13d	1
기	#44592	#xae30	1
א	#1488	#x5d0	1
λ	#955	#x3bb	1
δ	#948	#x3b4	1
ן	#1503	#x5df	1
ぐ	#12368	#x3050	1
ô	#244	#xf4	1
η	#951	#x3b7	1
鷗	#40407	#x9dd7	1
כ	#1499	#x5db	1
ī	#299	#x12b	1

Key Features

- Comprehensive DQ Functionality with a single UI and Repository



Tool Palette - Profiling

- Contained Attributes Profiler
- Data Types Profiler
- Date Profiler
- Equal Attributes Profiler
- Frequency Profiler
- Length Profiler
- Max/Min Profiler
- Number Profiler
- Patterns Profiler
- Quickstats Profiler
- Record Completeness Profiler
- Record Duplication Profiler
- RegEx Patterns Profiler

Tool Palette - Audit

- Cross Attribute Check
- Data Type Check
- Duplicate Check
- Email Check
- GBR Postcode Format Check
- Invalid Character Check
- Length Check
- List Check
- Logic Check
- Lookup Check
- No Data Check
- Pattern Check
- RegEx Check
- Suspect Data Check
- Value Check

Tool Palette - Transformation

- Add Current Date
- Add Date Attribute
- Add Numeric Attribute
- Add String Attribute
- Character Replace
- Concatenate
- Convert Date to String
- Convert Number to Date
- Convert Number to String
- Convert String to Date
- Convert String to Number
- Date Difference
- Denoise
- Enhance from Map
- Extract Values

Tool Palette - Matching

- Advanced Match
- Consolidate
- Deduplicate
- Enhance
- Link

Tool Palette - Text Analysis

- Parse
- Phrase Profiler

Tool Palette - Custom

- Expression
- Expression Filter
- Script

Tool Palette - Maths

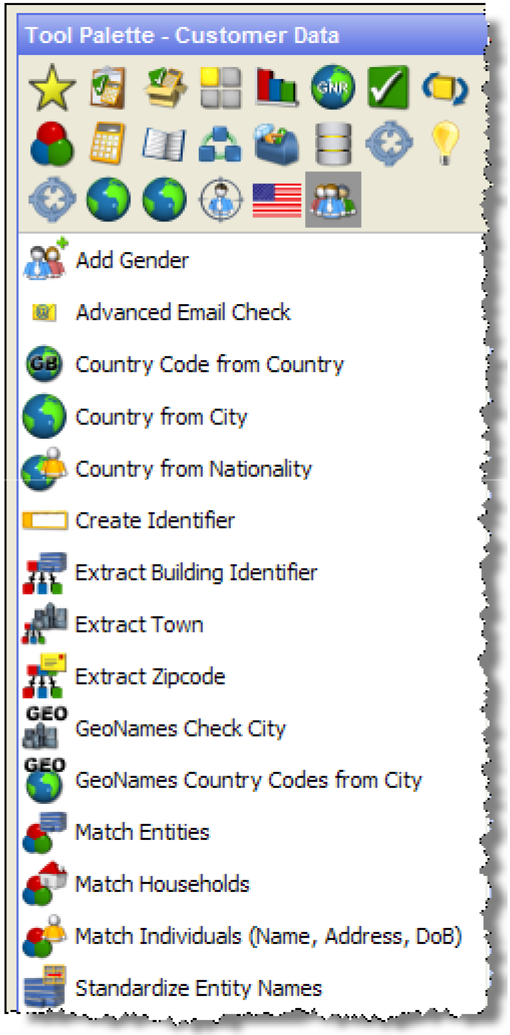
- Add
- Divide
- Multiply
- Round
- Subtract

Tool Palette - 3rd Party

- Capscan Matchcode
- Global Address
- QAS Quick Address

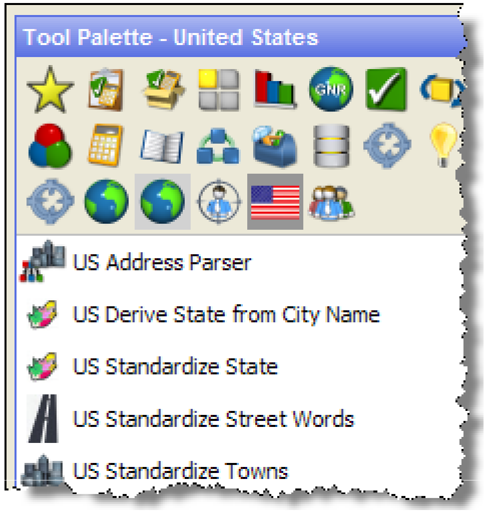
Key Features

- Provided Extensions for Customer Data and Locales
- Highly Extensible



Tool Palette - Customer Data

- Add Gender
- Advanced Email Check
- Country Code from Country
- Country from City
- Country from Nationality
- Create Identifier
- Extract Building Identifier
- Extract Town
- Extract Zipcode
- GeoNames Check City
- GeoNames Country Codes from City
- Match Entities
- Match Households
- Match Individuals (Name, Address, DoB)
- Standardize Entity Names

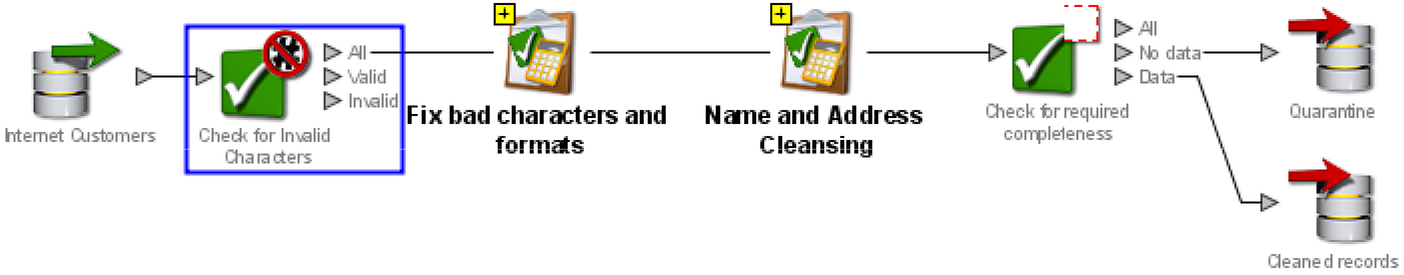


Tool Palette - United States

- US Address Parser
- US Derive State from City Name
- US Standardize State
- US Standardize Street Words
- US Standardize Towns

Transformation – Data Improvement

- Fully configurable data transformation rules
- Operates in both Batch and Real-Time
- Full control over data updates
- Original data always preserved (and all steps in between)
- Source data may either be staged and processed or 'streamed' through the process



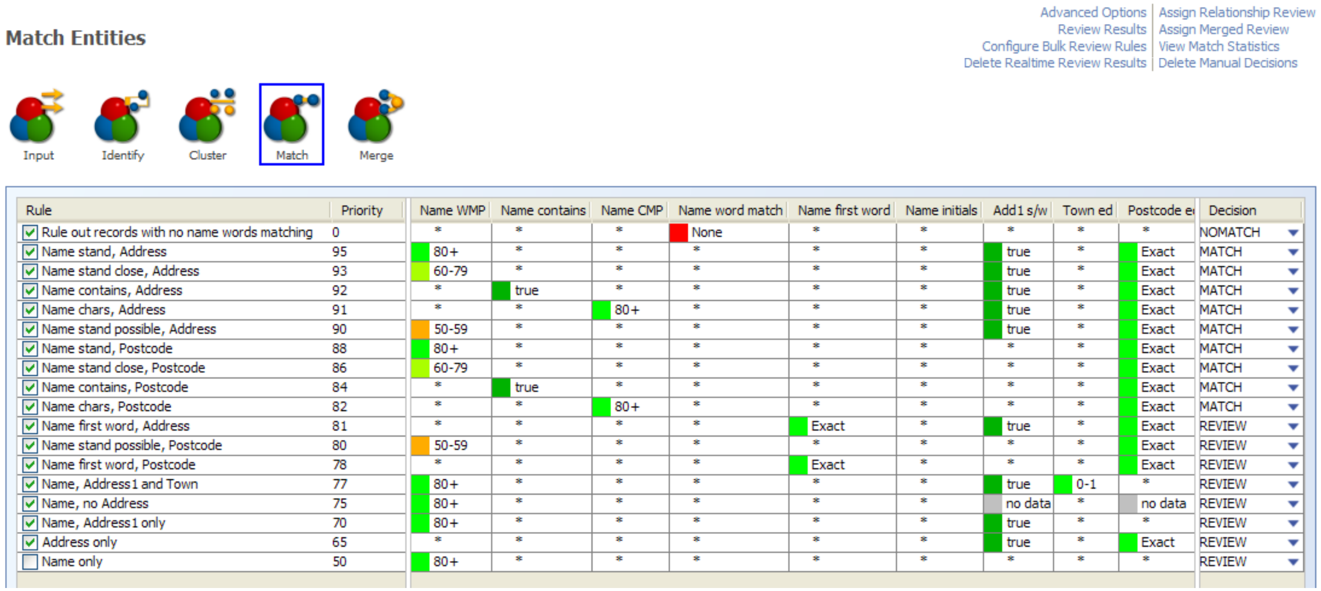
Internet Customers
All
Valid
Invalid
Check for Invalid Characters
Fix bad characters and formats
Name and Address Cleansing
All
No data
Data
Check for required completeness
Quarantine
Cleaned records

Use profiling results to create your own data improvement rules

Use provided processors for common tasks such as address standardization

Matching

- Designed for business users
- Flexible matching engine for any data with many comparison algorithms
- Provided template match processors for individual, entity and address matching
- Easy reuse of configured match processors
- Fully configurable outputs (Links, Groups, Master and Slaves, Best Record)
- Operates in both Batch and Real-Time
- See Match Essentials deck for more information on Matching



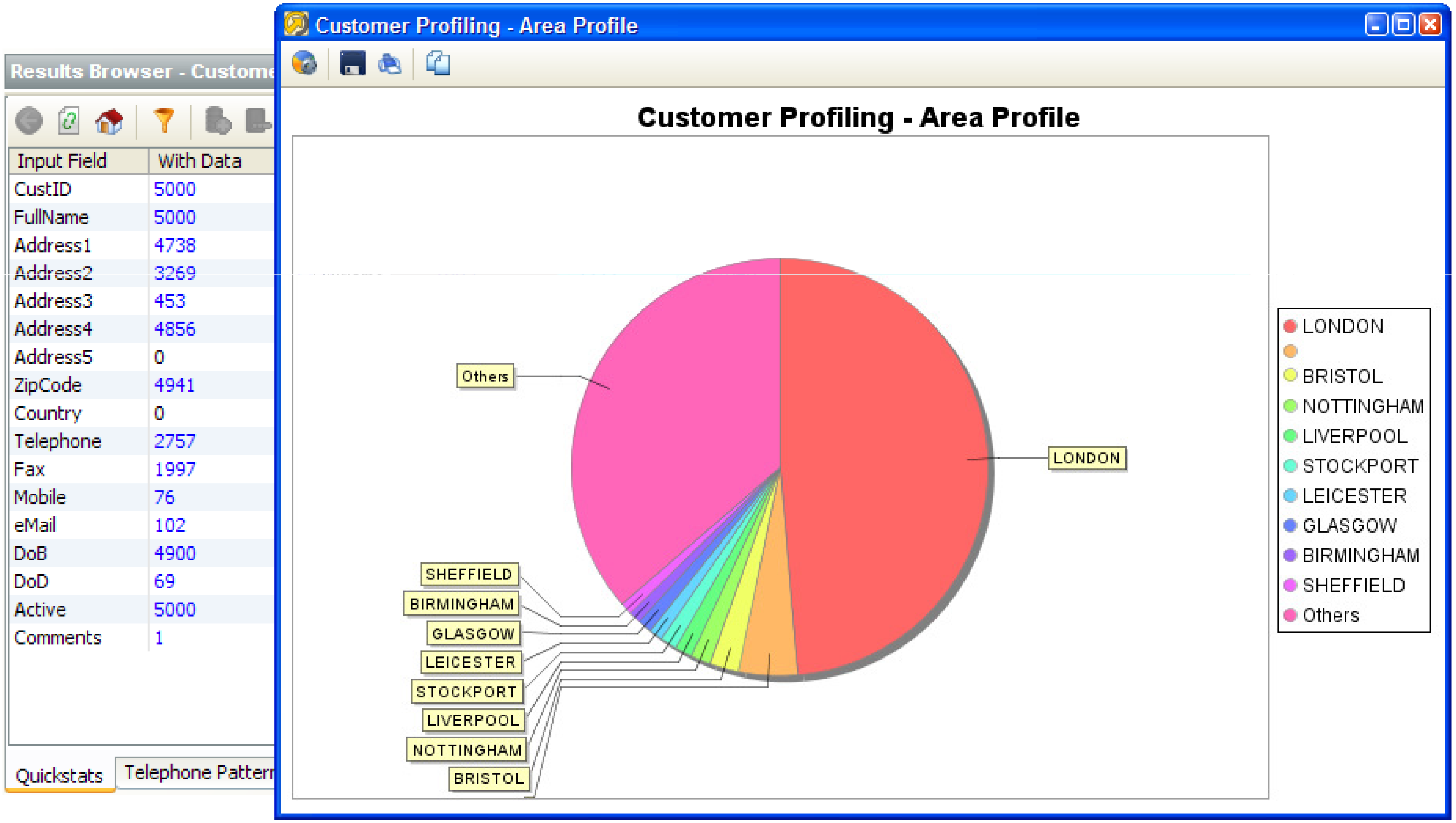
Rule	Priority	Name WMP	Name contains	Name CMP	Name word match	Name first word	Name initials	Add1 s/w	Town ed	Postcode e	Decision
☑ Rule out records with no name words matching	0	*	*	*	🟥 None	*	*	*	*	*	NOMATCH
☑ Name stand, Address	95	🟩 80+	*	*	*	*	*	🟩 true	*	🟩 Exact	MATCH
☑ Name stand close, Address	93	🟩 60-79	*	*	*	*	*	🟩 true	*	🟩 Exact	MATCH
☑ Name contains, Address	92	*	🟩 true	*	*	*	*	🟩 true	*	🟩 Exact	MATCH
☑ Name chars, Address	91	*	*	🟩 80+	*	*	*	🟩 true	*	🟩 Exact	MATCH
☑ Name stand possible, Address	90	🟧 50-59	*	*	*	*	*	🟩 true	*	🟩 Exact	MATCH
☑ Name stand, Postcode	88	🟩 80+	*	*	*	*	*	*	*	🟩 Exact	MATCH
☑ Name stand close, Postcode	86	🟩 60-79	*	*	*	*	*	*	*	🟩 Exact	MATCH
☑ Name contains, Postcode	84	*	🟩 true	*	*	*	*	*	*	🟩 Exact	MATCH
☑ Name chars, Postcode	82	*	*	🟩 80+	*	*	*	*	*	🟩 Exact	MATCH
☑ Name first word, Address	81	*	*	*	*	🟩 Exact	*	🟩 true	*	🟩 Exact	REVIEW
☑ Name stand possible, Postcode	80	🟧 50-59	*	*	*	*	*	*	*	🟩 Exact	REVIEW
☑ Name first word, Postcode	78	*	*	*	*	🟩 Exact	*	*	*	🟩 Exact	REVIEW
☑ Name, Address1 and Town	77	🟩 80+	*	*	*	*	*	🟩 true	🟩 0-1	*	REVIEW
☑ Name, no Address	75	🟩 80+	*	*	*	*	*	⬜ no data	*	⬜ no data	REVIEW
☑ Name, Address1 only	70	🟩 80+	*	*	*	*	*	🟩 true	*	*	REVIEW
☑ Address only	65	*	*	*	*	*	*	🟩 true	*	🟩 Exact	REVIEW
☐ Name only	50	🟩 80+	*	*	*	*	*	*	*	*	REVIEW

Reporting

- Highly flexible reporting interface
  - Export any Results views automatically to database/file
  - 1-click export of results to Excel from the Director client
- Dashboard reporting provides stakeholder view of Data Quality KPIs with trend analysis
- Example reports
  - Automatic Matches / Review Matches / Non-Matching Records
  - Match Group Size Report
  - Match Rule Report
  - Data Validity Report
  - Profiling Report
  - Etc.

Reporting - Immediate drilldown reporting in Director

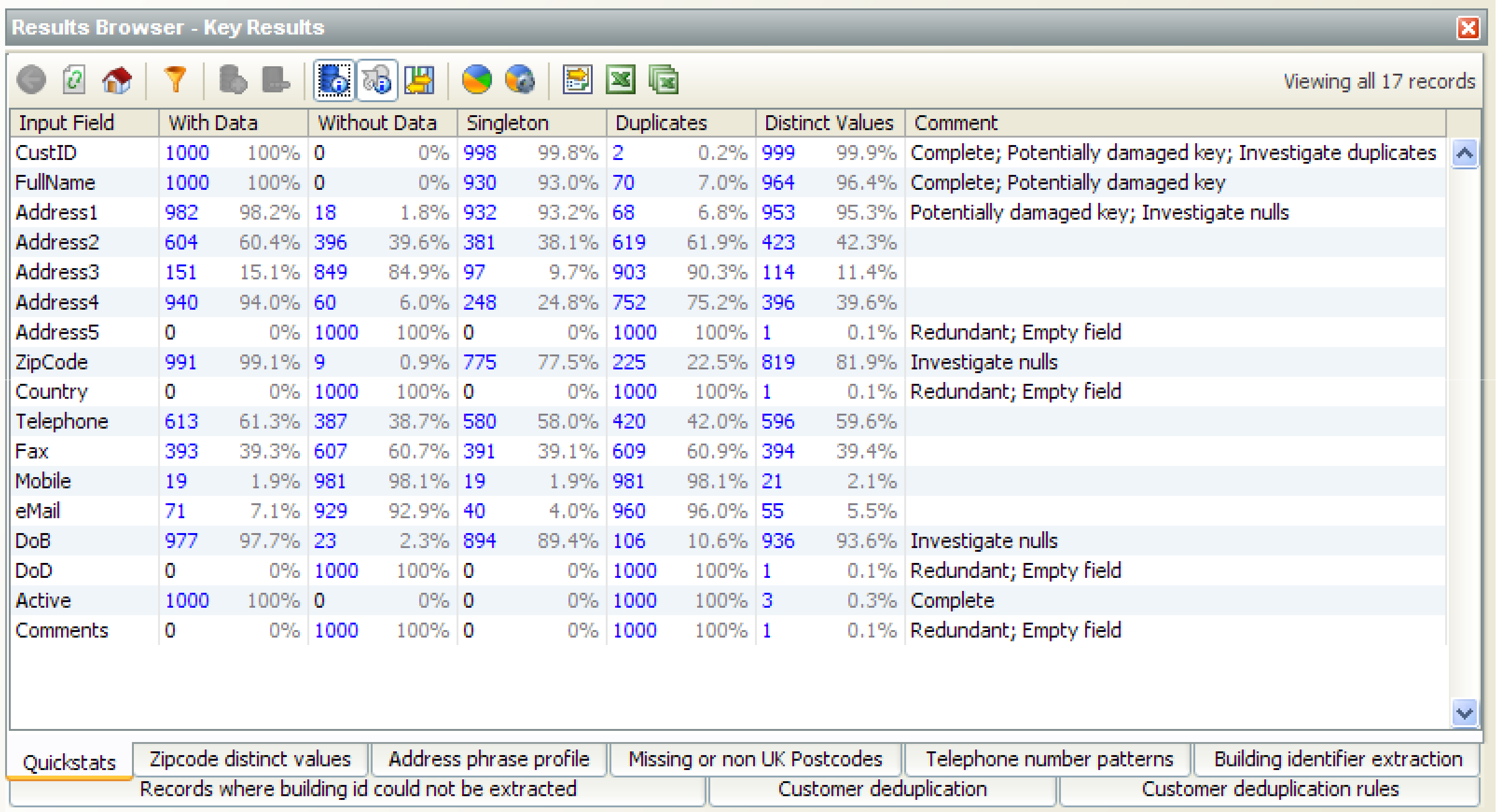
- Results Browser



Results Browser - Custome

Input Field	With Data
CustID	5000
FullName	5000
Address1	4738
Address2	3269
Address3	453
Address4	4856
Address5	0
ZipCode	4941
Country	0
Telephone	2757
Fax	1997
Mobile	76
eMail	102
DoB	4900
DoD	69
Active	5000
Comments	1

Reporting - Collect important results into Results Books



Results Browser - Key Results

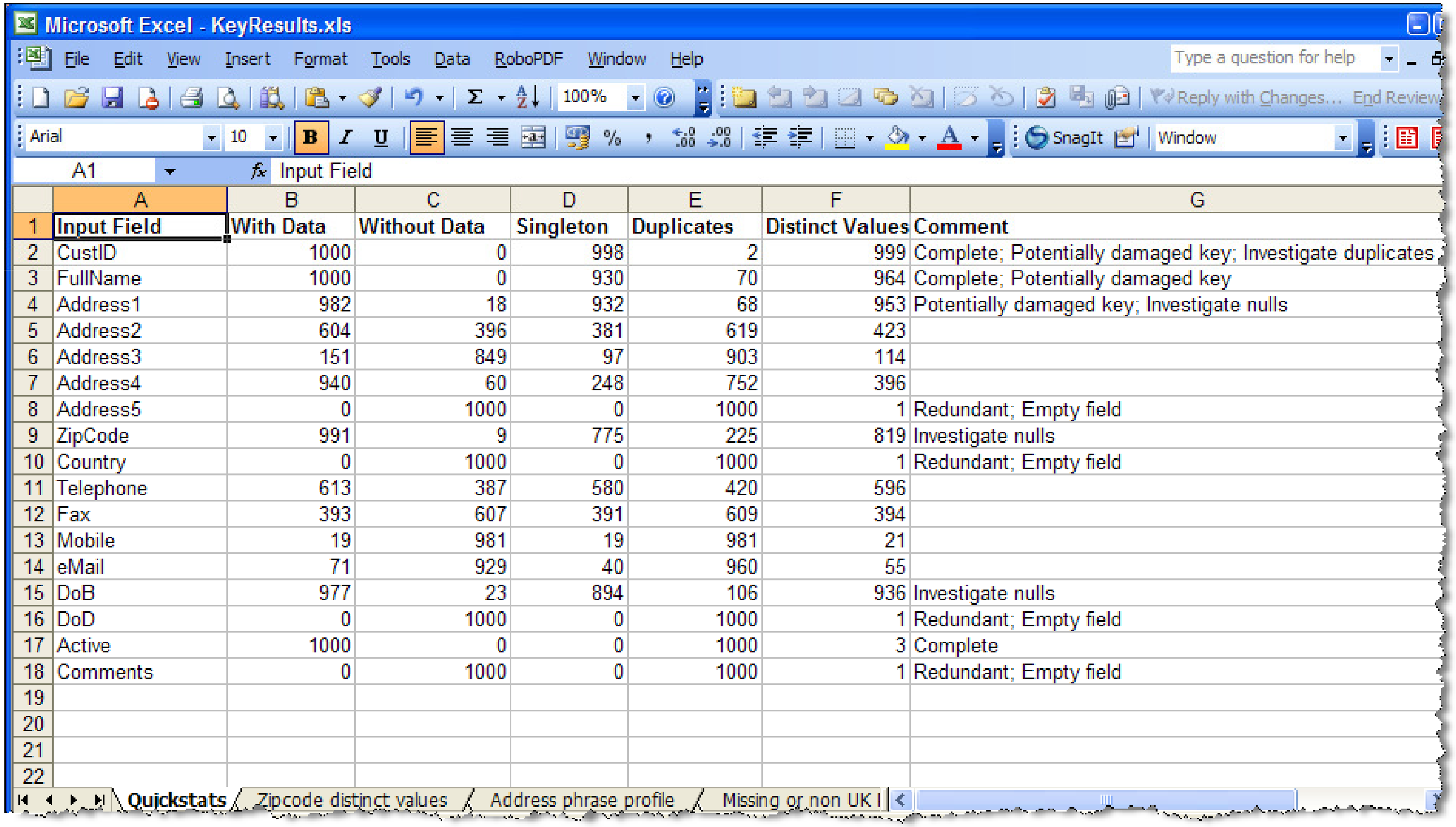
Viewing all 17 records

Input Field	With Data		Without Data		Singleton		Duplicates		Distinct Values		Comment
CustID	1000	100%	0	0%	998	99.8%	2	0.2%	999	99.9%	Complete; Potentially damaged key; Investigate duplicates
FullName	1000	100%	0	0%	930	93.0%	70	7.0%	964	96.4%	Complete; Potentially damaged key
Address1	982	98.2%	18	1.8%	932	93.2%	68	6.8%	953	95.3%	Potentially damaged key; Investigate nulls
Address2	604	60.4%	396	39.6%	381	38.1%	619	61.9%	423	42.3%	
Address3	151	15.1%	849	84.9%	97	9.7%	903	90.3%	114	11.4%	
Address4	940	94.0%	60	6.0%	248	24.8%	752	75.2%	396	39.6%	
Address5	0	0%	1000	100%	0	0%	1000	100%	1	0.1%	Redundant; Empty field
ZipCode	991	99.1%	9	0.9%	775	77.5%	225	22.5%	819	81.9%	Investigate nulls
Country	0	0%	1000	100%	0	0%	1000	100%	1	0.1%	Redundant; Empty field
Telephone	613	61.3%	387	38.7%	580	58.0%	420	42.0%	596	59.6%	
Fax	393	39.3%	607	60.7%	391	39.1%	609	60.9%	394	39.4%	
Mobile	19	1.9%	981	98.1%	19	1.9%	981	98.1%	21	2.1%	
eMail	71	7.1%	929	92.9%	40	4.0%	960	96.0%	55	5.5%	
DoB	977	97.7%	23	2.3%	894	89.4%	106	10.6%	936	93.6%	Investigate nulls
DoD	0	0%	1000	100%	0	0%	1000	100%	1	0.1%	Redundant; Empty field
Active	1000	100%	0	0%	0	0%	1000	100%	3	0.3%	Complete
Comments	0	0%	1000	100%	0	0%	1000	100%	1	0.1%	Redundant; Empty field

Quickstats
Zipcode distinct values
Address phrase profile
Missing or non UK Postcodes
Telephone number patterns
Building identifier extraction
Records where building id could not be extracted
Customer deduplication
Customer deduplication rules

Reporting - Output Results Books

Export Results Books as part of an automated job
1-click Export of a Results Book to Excel

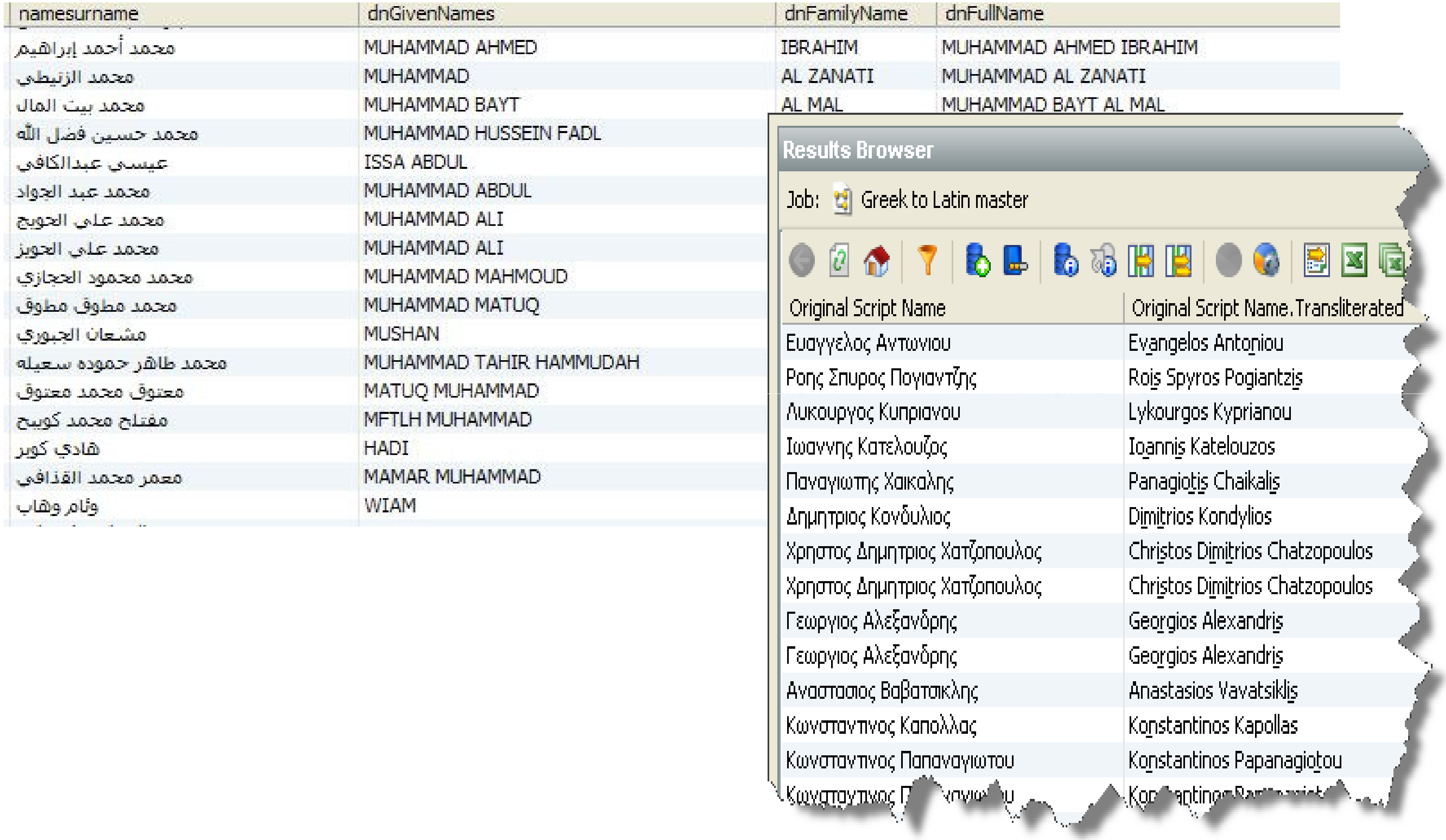


	A	B	C	D	E	F	G
1	Input Field	With Data	Without Data	Singleton	Duplicates	Distinct Values	Comment
2	CustID	1000	0	998	2	999	Complete; Potentially damaged key; Investigate duplicates
3	FullName	1000	0	930	70	964	Complete; Potentially damaged key
4	Address1	982	18	932	68	953	Potentially damaged key; Investigate nulls
5	Address2	604	396	381	619	423	
6	Address3	151	849	97	903	114	
7	Address4	940	60	248	752	396	
8	Address5	0	1000	0	1000	1	Redundant; Empty field
9	ZipCode	991	9	775	225	819	Investigate nulls
10	Country	0	1000	0	1000	1	Redundant; Empty field
11	Telephone	613	387	580	420	596	
12	Fax	393	607	391	609	394	
13	Mobile	19	981	19	981	21	
14	eMail	71	929	40	960	55	
15	DoB	977	23	894	106	936	Investigate nulls
16	DoD	0	1000	0	1000	1	Redundant; Empty field
17	Active	1000	0	0	1000	3	Complete
18	Comments	0	1000	0	1000	1	Redundant; Empty field
19							
20							
21							
22							

Arabic – Name Transcription Approach

- For 98+% of individual names, transcription occurs directly to the IC form of the name using a large database of Arabic names
- For the remaining names, a custom dictionary is used
- If a name is still unrecognized (<1%) it is transliterated using character-based transliteration and flagged as an exception
- Easy to add transcriptions for exceptions to the custom dictionary
- It is also possible to override the transcription for a specific names

Arabic, Non Latin – Name Transcription Example



namesurname	dnGivenNames	dnFamilyName	dnFullName
محمد أحمد إبراهيم	MUHAMMAD AHMED	IBRAHIM	MUHAMMAD AHMED IBRAHIM
محمد الزنيطى	MUHAMMAD	AL ZANATI	MUHAMMAD AL ZANATI
محمد بيت المال	MUHAMMAD BAYT	AL MAL	MUHAMMAD BAYT AL MAL
محمد حسين فضل الله	MUHAMMAD HUSSEIN FADL		
عيسى عبدالكافي	ISSA ABDUL		
محمد عبد الجواد	MUHAMMAD ABDUL		
محمد علي الحويج	MUHAMMAD ALI		
محمد علي الحويز	MUHAMMAD ALI		
محمد محمود الحجازي	MUHAMMAD MAHMOUD		
محمد مطوق مطوق	MUHAMMAD MATUQ		
مشعان الجبوري	MUSHAN		
محمد طاهر حموده سعيله	MUHAMMAD TAHIR HAMMUDAH		
معتوق محمد معتوق	MATUQ MUHAMMAD		
مفتاح محمد كويح	MFTLH MUHAMMAD		
هادي كوير	HADI		
معمر محمد القذافي	MAMAR MUHAMMAD		
وئام وهاب	WIAM		

Job: Greek to Latin master

Original Script Name	Original Script Name.Transliterated
Ευαγγελος Αντωνιου	Evangelos Antoniou
Ροης Σπυρος Πογιαντζης	Rois Spyros Pogiantzis
Λυκουργος Κυπριανου	Lykourgos Kyprianou
Ιωαννης Κατελουζος	Ioannis Katelouzos
Παναγιωτης Χαικαλης	Panagiotis Chaikalis
Δημητριος Κονδυλιος	Dimitrios Kondylios
Χρηστος Δημητριος Χατζοπουλος	Christos Dimitrios Chatzopoulos
Χρηστος Δημητριος Χατζοπουλος	Christos Dimitrios Chatzopoulos
Γεωργιος Αλεξανδρης	Georgios Alexandris
Γεωργιος Αλεξανδρης	Georgios Alexandris
Αναστασιος Βαβατσικλης	Anastasios Vavatsiklis
Κωνσταντινος Καπολλας	Konstantinos Kapollas
Κωνσταντινος Παπαναγιωτου	Konstantinos Papanagiotou
Κωνσταντινος Π[illegible]υ	Kon[illegible]antinos P[illegible]

Arabic – Name Matching Approach

- All known Latin variant representations of an Arabic name are recognized in matching using a dictionary of 5m variants
- The 5m may be filtered to the most frequent representations only if required
- High confidence matching even where transcription standards may be very different
- Can match both Arabic to Arabic and Arabic to English
- Can match Arabic to other languages via comprehensive transliteration capabilities for other languages
- Wide variety of additional matching algorithms and transformation capability, for example to cope with:
  - Missing names
  - Out of order names
  - Typos
  - Etc.
- Complete control over matching