Why Data Quality Matters in Enhancing your Customer Experience

Dr Mukesh K Mohania
Mohan N Dani
IBM India Software Labs

InformationOnDemand**India2011**

The Premier Conference for Information Management
Manage. Analyze. Govern.

February 2, 2011

Hyatt Regency I Mumbai, India

Data Quality -- Realities

- In developing economies, like India, many customers require Data Quality as they adopt BI and leverage their information assets

  - *Banking Companies* wants to create a single consistent view of their customer
  - *Telecom companies* wants to offer consistent service to their high worth customers irrespective of the line of business to which they belong

- Lack of standards and format specification results in poorly captured data and standardizing this data is a huge challenge

  - *Semantic Variation : Often semantic representing locality is interchangeably with area and area is a district etc*
  - *Semantic Exclusions: Many time information is dropped, e.g. missing road or floor number*
  - *Additional Information: Presence of multiple landmarks and non standard abbreviations*
  - *Syntactic Format variability: No particularly ordering of semantic elements*
  - *Multi-linguality and regional usage variations, Spelling errors*

- State of the art techniques used in developed word to improve Data Quality do not work hence we need techniques that can solve Data Quality problems of customers in developing economies
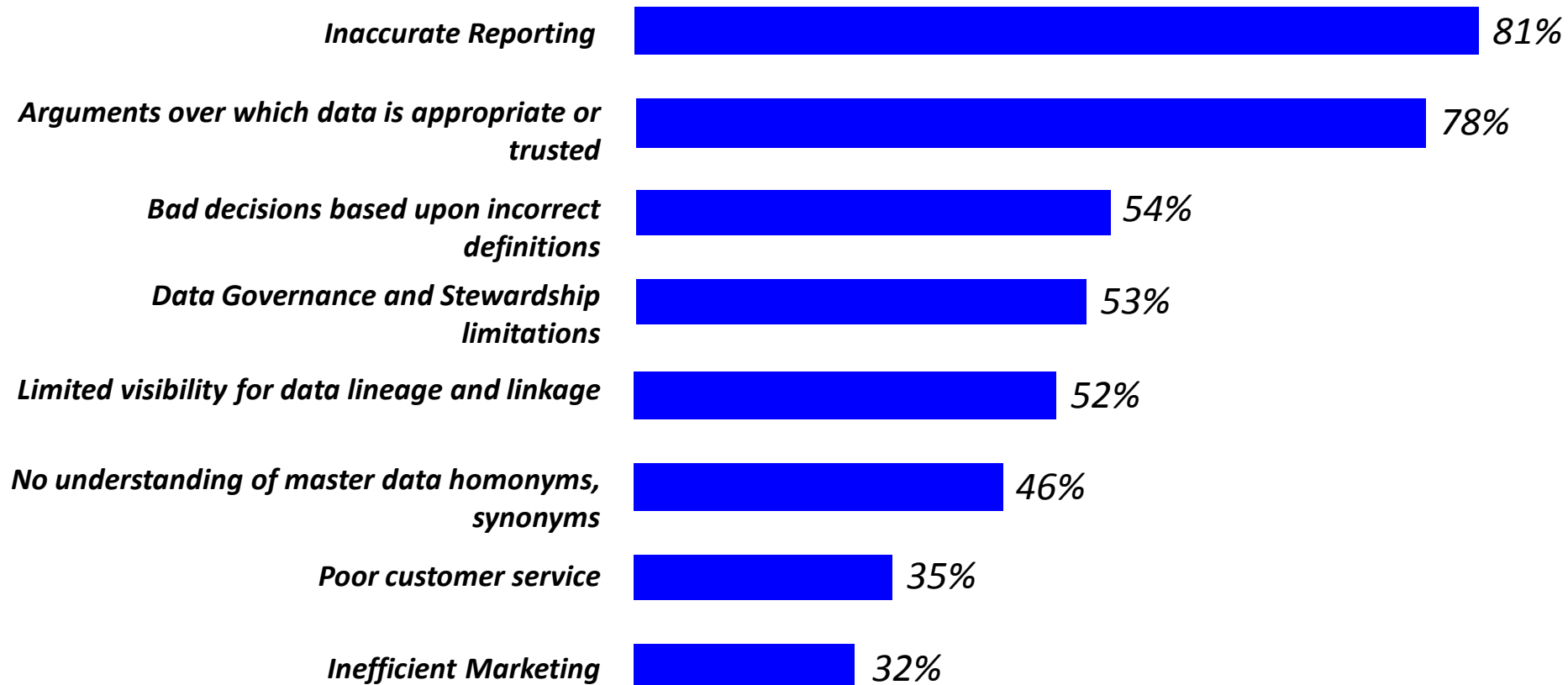
Impact of Poor Data Quality on Business

- Wrong, incomplete, non-standard addresses impact the business in the following ways -

  – Increased mailing costs

  – Expensive efforts in processing returned mail

  – Reduced customer satisfaction

  – Creates delays as well as missed opportunities

  – Increase in organisational risks

  – Challenges in data cleansing and validation efforts

  – Transliteration challenges

  – Not able to reach customers pro-actively

  – Not able to do Targeted marketing

  – Regulatory Compliance fines

How Organizations suffer from Bad Data?

What Kind of Problems?

Problem	Percentage
Inaccurate Reporting	81%
Arguments over which data is appropriate or trusted	78%
Bad decisions based upon incorrect definitions	54%
Data Governance and Stewardship limitations	53%
Limited visibility for data lineage and linkage	52%
No understanding of master data homonyms, synonyms	46%
Poor customer service	35%
Inefficient Marketing	32%

TDWI, October 2006

Common Data Problems

- Lack of information standards
  - Different formats & structures across different systems

```
Ashok Kumar    416 Anand Niketan, New Delhi, India 110021
A Kumar Four sixteen Street 8, Anand Niketan, Delhi 110021
      Mr. Ashok Kr  #416 Anand Niketan, N Delhi, 21
```

- Data Surprises in Individual Fields
  - Data misplaced in the database
  - Special characters in the data

```
      Email          Tax ID        Telephone
    91,,,,             228-02-1975   6173380300
  ranivrgeoi@yahoo.co.in   025-37-1888   415-392-2000
 ,CYRUS_DASTUR@HOTMAIL.COM 34-2671434   3380321
      HP 15 State St.    508-466-1200  Orlando
```

- The redundancy nightmare
  - Duplicate records with a lack of standards

```
  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
```
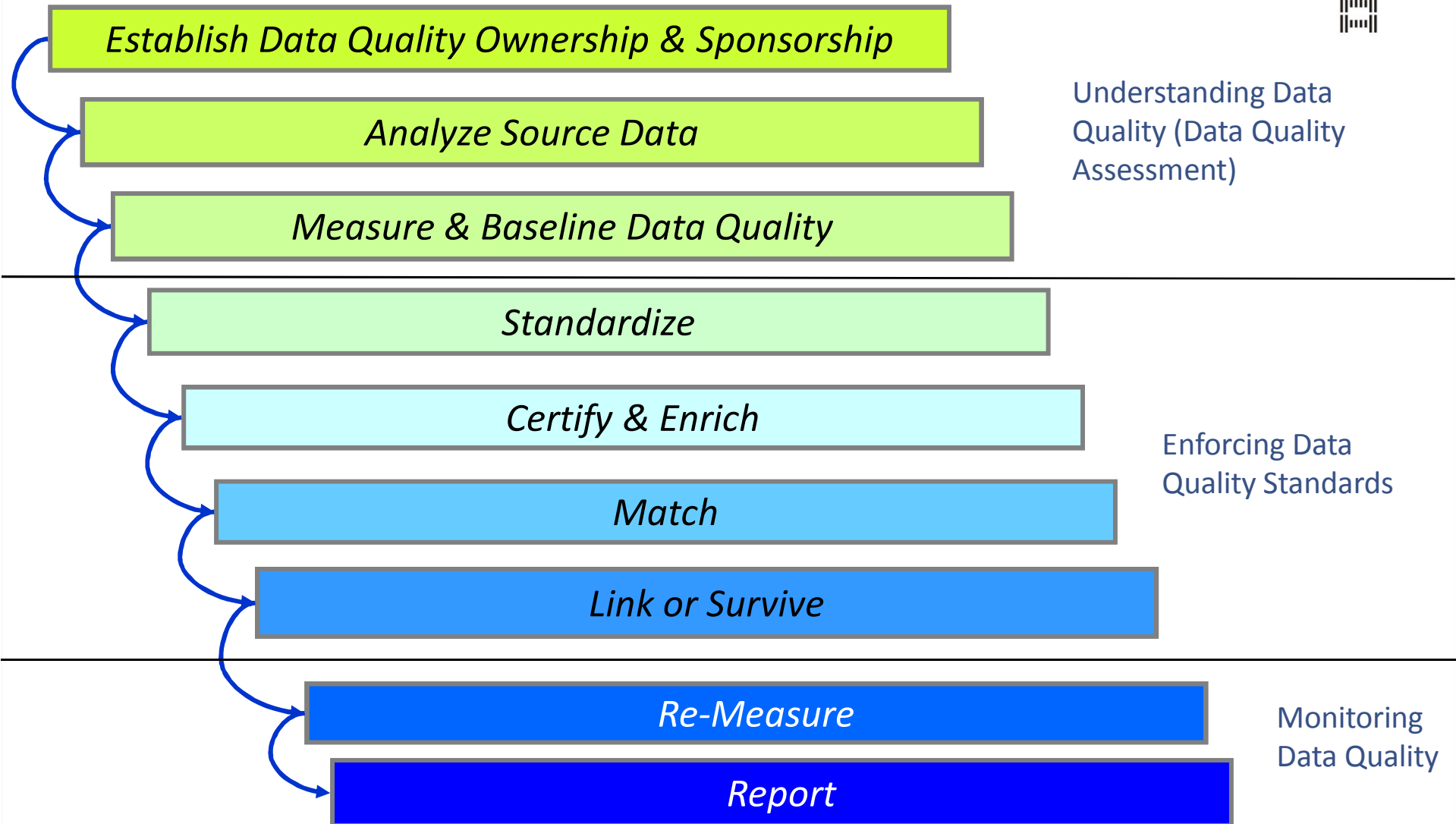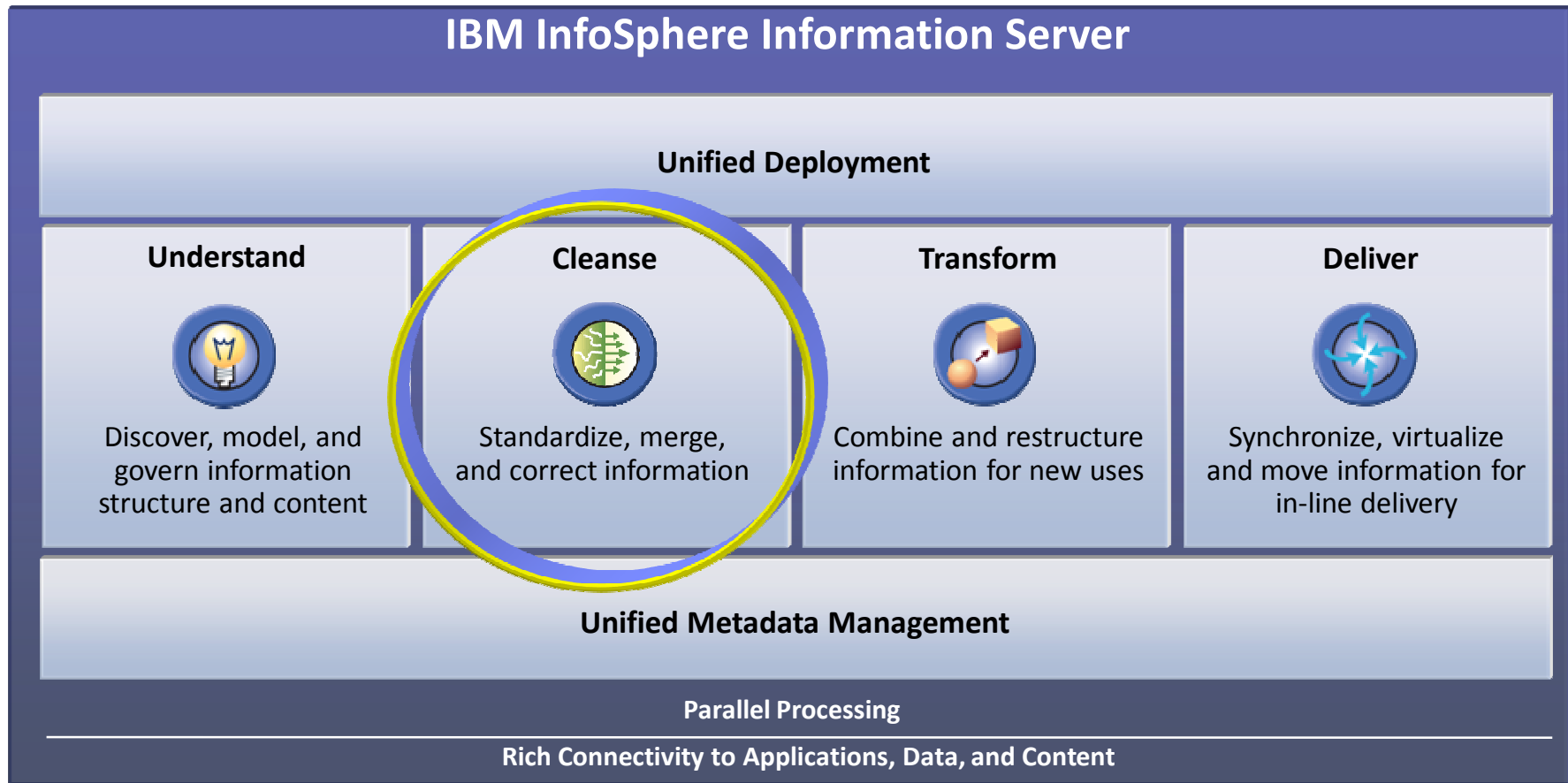
Solution & Approach

A Process For Data Quality

Establish Data Quality Ownership & Sponsorship

Analyze Source Data

Measure & Baseline Data Quality

Understanding Data Quality (Data Quality Assessment)

Standardize

Certify & Enrich

Match

Link or Survive

Enforcing Data Quality Standards

Re-Measure

Report

Monitoring Data Quality

InformationOnDemandIndia2011
Manage. Analyze. Govern.

IBM InfoSphere Information Server
A Platform Enabling Enterprise Data Quality

Data Cleansing: InfoSphere QualityStage

- Provides specialized data quality processing
  - Ensures clean, standardized, de-duplicated information
  - Enables a single version of the truth
  - Supports global postal verification

- Provides visual tools for designing quality rules and matching logic
  - Seamlessly integrated with DataStage (one engine, one metamodel, one UI)
  - Precisely calibrates matching rules

- Allows quality logic to be deployed seamlessly within ETL, or as shared services

Subject Matter Experts

Data Analysts

Cleanse

InfoSphere QualityStage™

Standardize and correct source data fields, and match records together across sources to create a single view
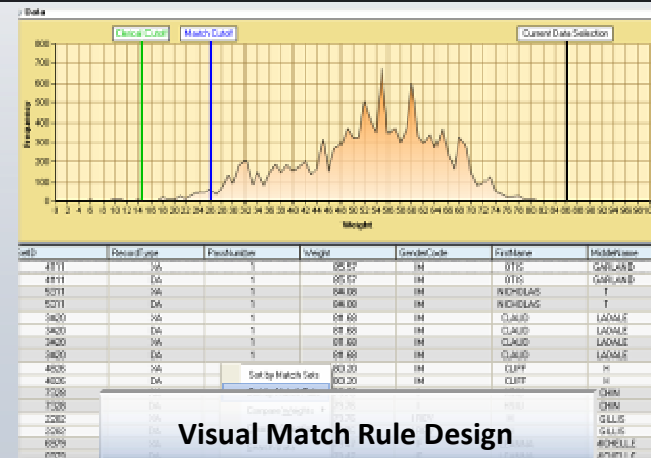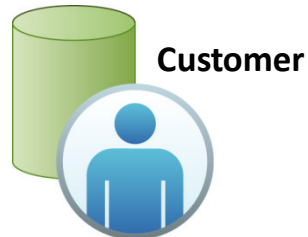
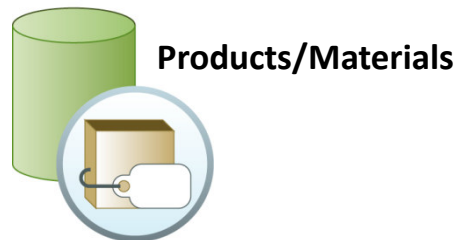**Visual Match Rule Design**

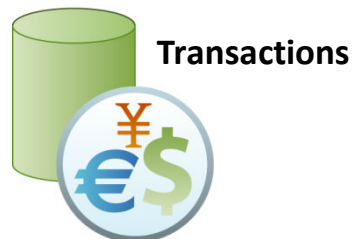# How will you get an accurate, consolidated view of your business?
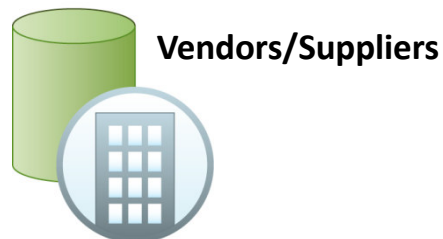
InfoSphere QualityStage Process
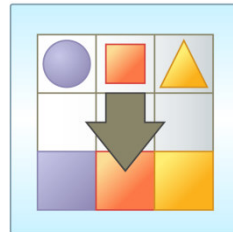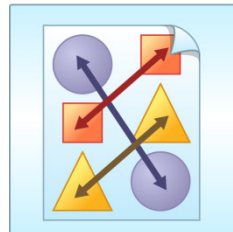
Customer
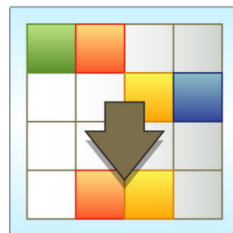
Products/Materials

Transactions

Vendors/Suppliers

1. Free Form Investigation

2. Data Standardization

3. Data Matching

4. Data Survivorship

Target Database with Consolidate Views

Indian Address Standardization

- IBM's QualityStage Address Standardization for India:

  - Standardizes pan India addresses (urban, rural, military etc) and provide a great degree of consistently in token definitions for producing high standardization outputs

  - Has the ability to be invoked real time by other applications in the enterprise so that data standardization can be handled real time

Typical Indian Address

- **Rachit Arora, opp Leela Garden, Vile Parle, Pkt**

  3,12, Block A, Subhash RD, Sector 17, MH, 400057

FirstName	Rachit	LastName	Arora

| DoorNo	Unit	SubUnit	StreetName	StreetType	LandmarkPosition
12	17 SECTOR	3 POCKET A BLOCK	SUBHASH	ROAD	OPPOSITE
Landmark	Area	City	District	State	Pincode
LEELA GARDEN	VILE PARLE EAST	MUMBAI	MUMBAI	MAHARASHTRA	400057

Rule Set Development Process

- **Quickly Adapt to customer data**

  - Quickly discover semantic and syntactic variants using context based clustering

  - Each record is represented by a Signature for better performance

  - Addition of discovered variants to dictionaries and classification tables
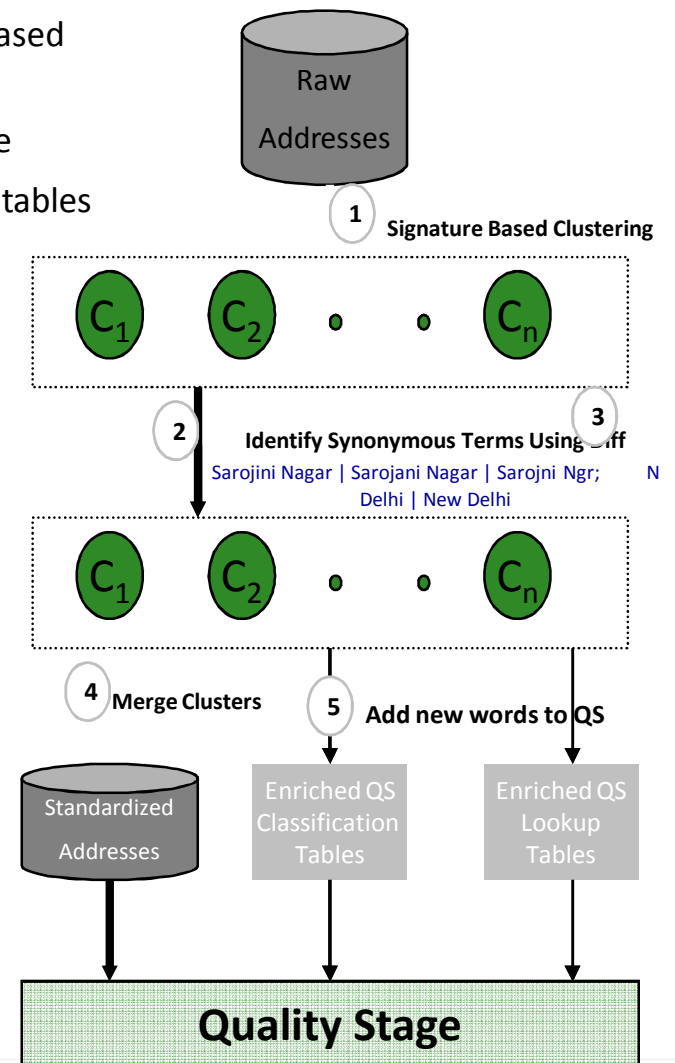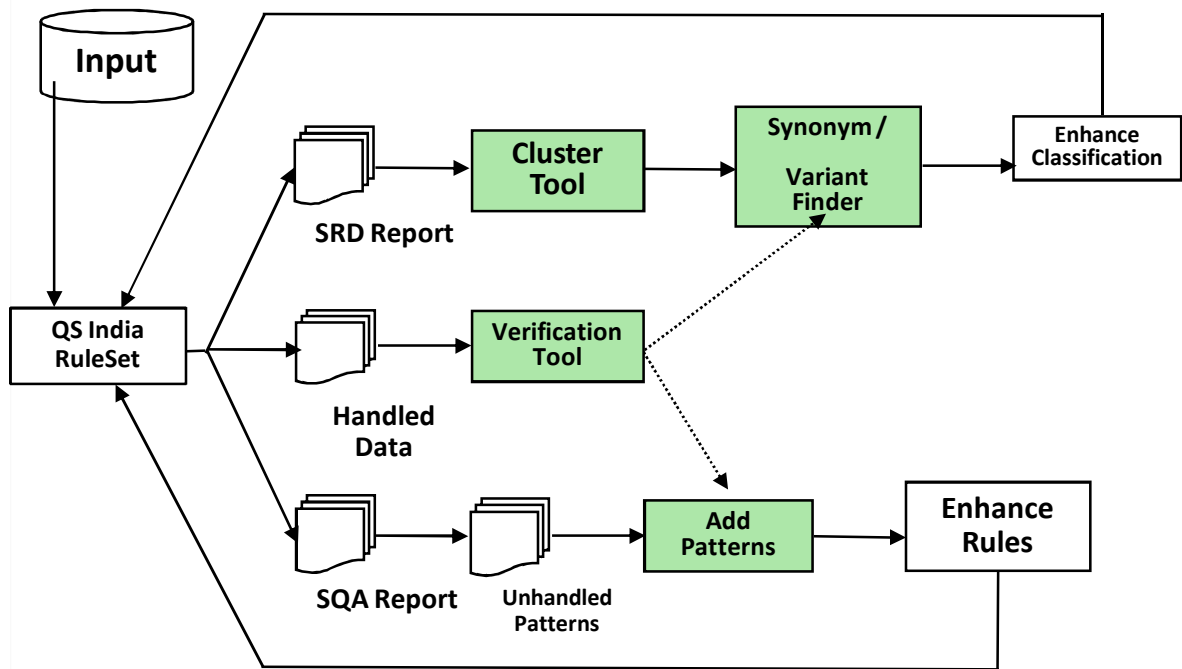
RDR and Taxonomy Construction

- Improve Base accuracy of standard ruleset and decrease customization time
  - Base accuracy with the QS standard India Ruleset is 60-65% improve that to 80%
  - Reduce rule customization time anywhere between 4-12 weeks to customize the rules to reach 90-95% accuracy
- Organize Rules to make them manageable
  - Currently 2000+ rules which are very difficult to manage
  - No standard reference taxonomy exist for data cleansing
- Used the Ripple Down Rule framework for rule organization and rule writing
  - Version 1 of RDR based India-rule set developed. Plan to put these rules into production.
  - Improved performance by 5% while reducing number of rules
- Built an address taxonomy to be used by rule engine
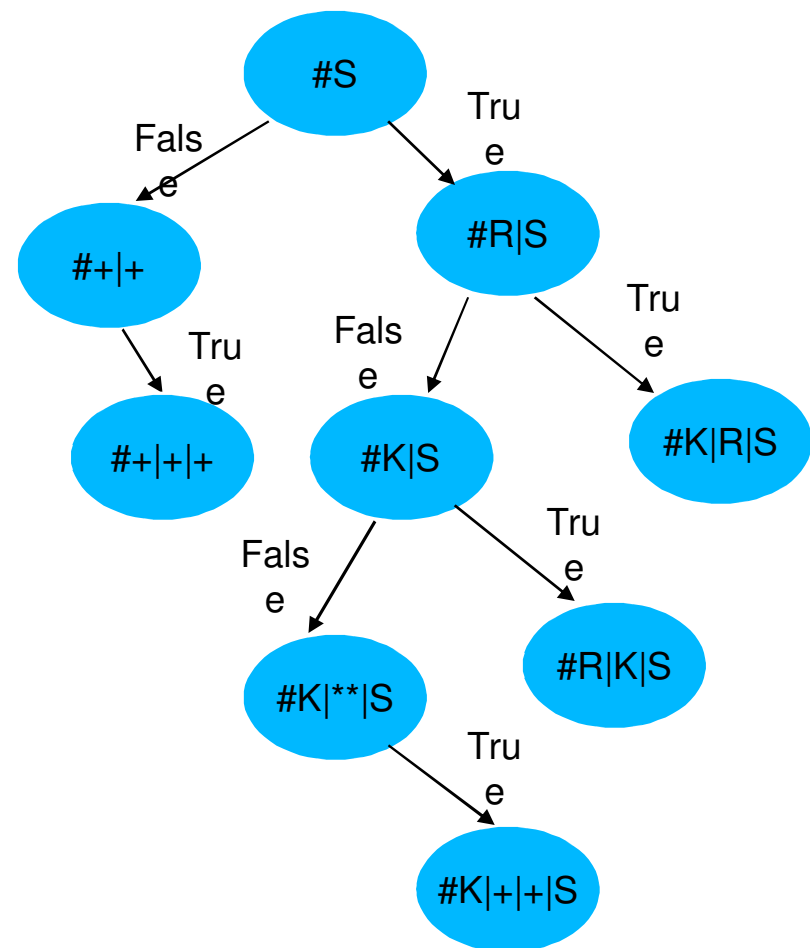
Key matching techniques with QualityStage

ASHOK KR   St JOHN'S ROAD   560034 12/8/62   ASHOKE KUMAR ST JOHN RD   560034 12/8/62

Deterministic Decisions Tables:
- Fields are compared
- Letter grade assigned
- Combined letter grades are compared to a vendor delivered file
- Result: Match; Fail; Suspect.
- This method is used when the organisation has control over every aspect of their data and can influence it to make decision with their inherent expertise and lineage and result is predictable.

Probabilistic Record Linkage:
- Fields are evaluated for degree-of-match
- Weight assigned: represents the **"information content"** by value
- Weights are summed to derived a total score
- Result: Statistical probability of a match
- This Method is best suited when the data lineage and aspects are unknown leading to best results.

Classic Transformation: Account to Customer

Source	Legacy Key	Name	Address	Phone	Birth Date
Life	70328574	Neema Uttam Mhatre	Sh Sai Smruti Malwani Church Malad W. MUMBAI 400095 India	781-259-9945	2/05/1940
Home	80328575	Ms Neema Mhatre	Sai Smruti Mawani Church, Malad (W), Mumbai 95	617-259-9000	
Auto	90238495	N. Mhatre	Sh Sai Smruti Malwani Ch, Malad W, Mumbai	781-295-9945	2/05/1941

Link Related Records to create cross-reference ids

Source	Legacy Key	Name	Address	Phone	Birth Date	Cust-ID
Life	70328574	Neema Uttam Mhatre	Sh Sai Smruti Malwani Church Malad W. MUMBAI 400095 India	781-259-9945	02/05/1940	0001
Home	80328575	Ms Neema Mhatre	Sai Smruti Mawani Church, Malad (W), Mumbai 95	617-259-9000		0001
Auto	90238495	N. Mhatre	Sh Sai Smruti Malwani Ch, Malad W, Mumbai	781-295-9945	2/05/1941	0001

Create a Customer Profile with the best information from all sources
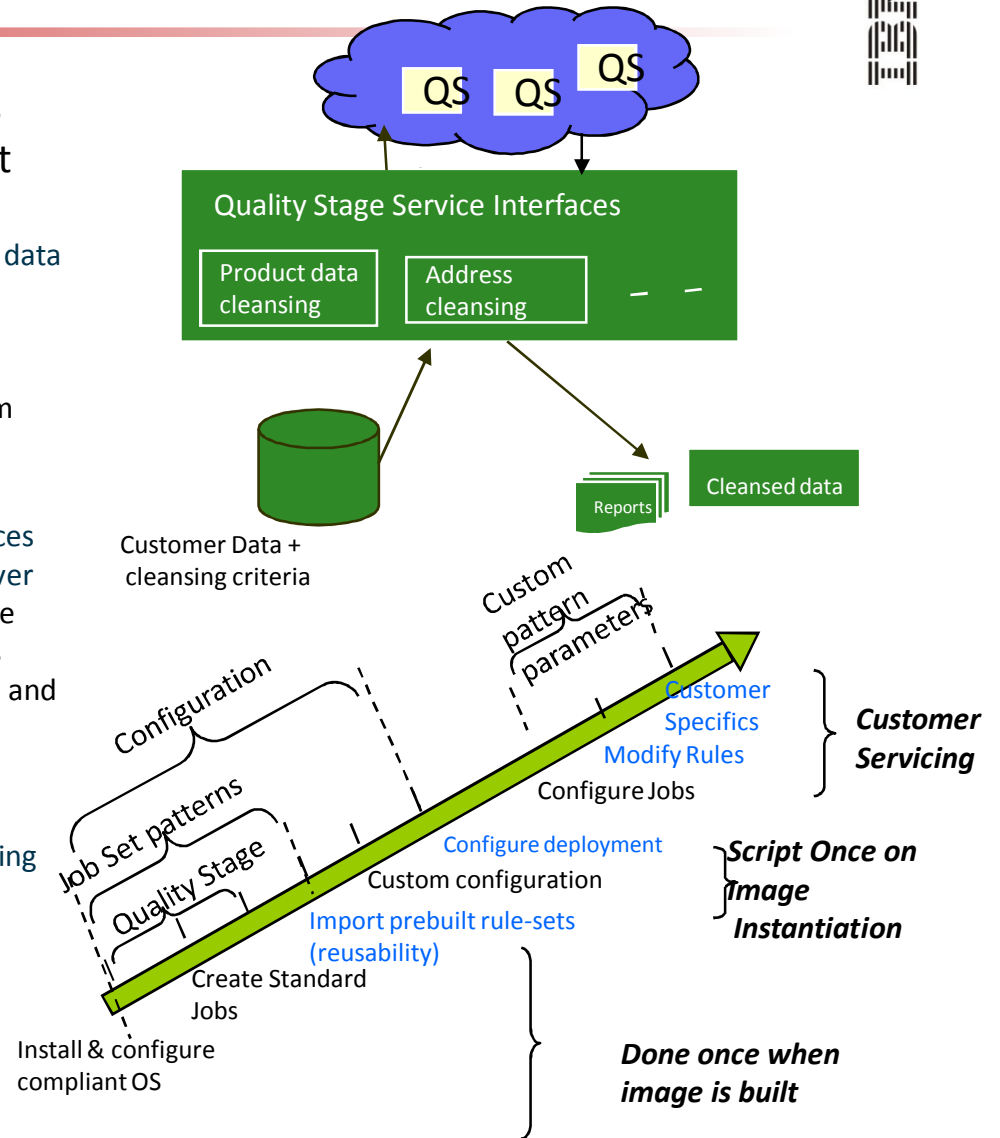
Source	Name	Address	Phone	Birth Date	Cust-ID
CP 0001	Ms. Neema Uttam Mhatre	Sh Sai Smruti Malwani Church, Malad (W), Mumbai 400095, India	617-259-9000	02/05/1940	

Cleansing as a Transient Service

- There is an increasing demand from customers who want data cleansing services as a transient offering
  - Customer often seek one time cleansing of their critical data (Telecom New Zealand)
    - Provides the business the agility to address their immediate data quality issues while they plan to adopt a long term enterprise data quality program

- Reduce Cost and time to deliver
  - To offer "bundled capability" of IBM products and services on a cloud environment to reduce cost and time to deliver
    - IBM QS image management and reusability of rule development and configuration across customers
    - IBM cleansing solutions offered on the IBM cloud and the Amazon cloud

- Competitive play
  - Competition (e.g. Informatica) is already offering cleansing solutions on the cloud

QS QS QS

Quality Stage Service Interfaces

Product data cleansing

Address cleansing

Reports

Cleansed data

Customer Data + cleansing criteria

Custom pattern parameters

Configuration

Job Set patterns

Quality Stage

Customer Specifics
Modify Rules

Customer Servicing

Configure Jobs

Configure deployment

Script Once on Image Instantiation

Custom configuration

Import prebuilt rule-sets (reusability)

Create Standard Jobs

Install & configure compliant OS

Done once when image is built

InformationOnDemand**India2011**
Manage. Analyze. Govern.

Data Quality needs in your IT projects

- ## Scenario 1: Master Data Integration

  - Your organization has decided to embark on an MDM (or CDI) initiative. This can include any and all of the previous scenarios in terms of data quality challenges – to gain a single version of the truth, you need clean, concise data.

  - Doing MDM without a comprehensive approach to Master Data Integration (MDI) including data quality will never deliver consistent results.

  - One or more application environments for CRM, ERP, or home grown applications. Need a manner for prevention – to proactively empower users and the applications they use to ensure the quality of data.

- ## Scenario 2 : Business Intelligence Initiatives

  - Doing some campaign management and need analytics of their customer and sales data. The data warehouse needs to be the source of truth. Leverage DW for an integrated view of the enterprise information assets and to drive strategic decisions (customer spend analysis, customer and product profitability, etc.)

  - Need to apply data quality in concert with the integration (ETL) process in order to clean up customer and product data.

- ## Scenario 3: Application Migration

  - Facing a migration from 1 or more legacy applications to a single application structure. May be characterized or include ERP/CRM instance consolidation

  - Great opportunity to clean up data – both in content and structure as well as addressing the duplicates that may exist.

  - Without doing this, you run the risk of creating a single, inconsistent version of the truth!

- ## Scenario 4: Database centric operations

  - Company wants to do a marketing campaign – data lives in a number of different files/databases. Quality is suspect, duplicates within and across databases.

  - Data surprises in individual fields (ex - business in personal field and email in the phone field)

  - Duplicates are likely to lead to incorrect campaign planning, execution and post campaign analysis

  - Manual cleansing – labour intensive

Putting it all Together: Information Architecture
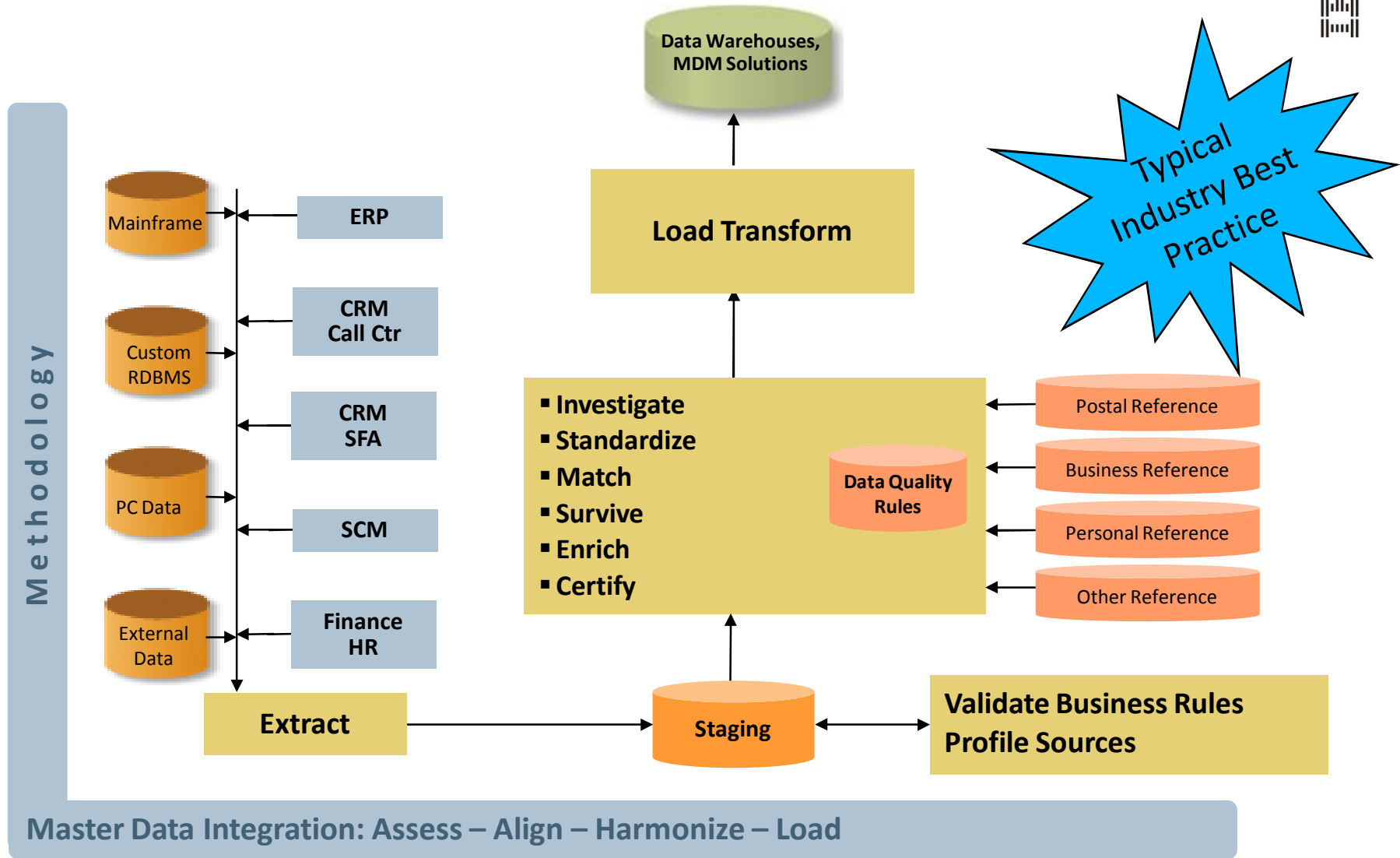
Enhance customer Experience with Data Quality - BFSI

- Standardized mail addresses significantly reduces the risk of mailers not reaching the customers

- Identify and match customers who have accounts in multiple branches and provide them with a consistent banking experience.

- Ability to define and identify household segments for profiling, cross-selling and up selling. Roll-out specialized services for such segments.

- Consolidated billing - Reduce mailing costs (monthly statements, notices etc). Reduce time and effort in processing returned mail

- When handling asset backed securities and mortgage backed properties in secondary markets, profile the properties based on the area to get demographic insights to the protection of the assets and predict the cash flow (negative/positive)

- Take advantage of pre-sorted mail discounts when sending group emails to NRI/foreign national customers

- Reduced cost from an IT perspective (data storage) as a result of data consolidation and removal of duplicates

Enhance customer Experience with Data Quality - Insurance

- Single, trusted view of customers across LOB

- Effectively share customer information with Internet, Agent/Brokers, Financial planners and Call centers

- Cross-Sell/Up-sell offers relevant & effective to individual customers

- Grow business with effective targeted marketing & Customer retention

- Customer retention

- Manage Claims: Enhance the quality, integration, and usage of real-time data to decrease loss ratios and improve customer service

- Fraud Claims : Recognize & reduce fraud claims

- Risk Management: Data Quality Accuracy reporting for Risk Management

- Analysis of Geographic Clustering of Risk

- Analysis of Agent Sales

Thank You

InformationOnDemand**India2011**
The Premier Conference for Information Management
Manage. Analyze. Govern.

February 2, 2011
Hyatt Regency I Mumbai, India