Data Quality

Loading the fact tables and dimension tables

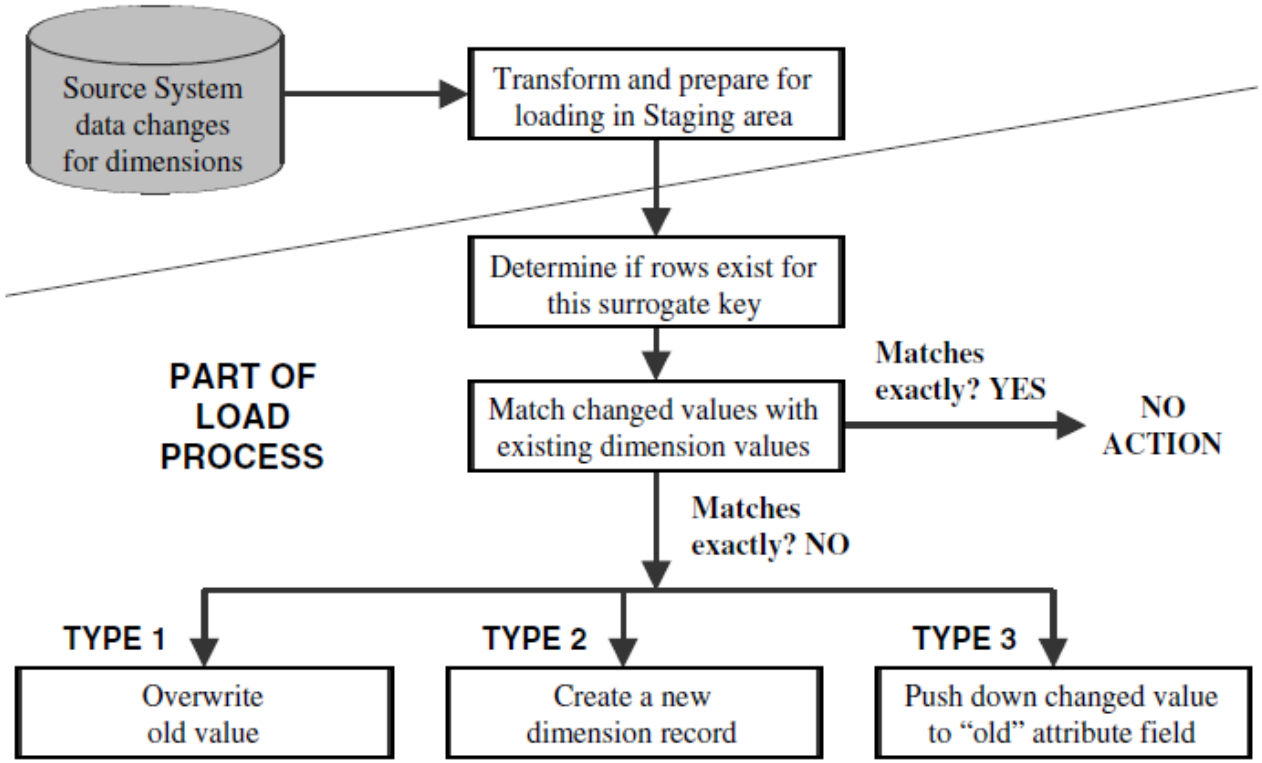


Figure 12-13 Loading changes to dimension tables.

- The procedure for maintaining the dimension tables includes two functions:
  - The initial loading of the tables
  - Applying the changes on an ongoing basis. Let us consider two issues.
- Incremental extracts for fact tables
  - Consist of new transactions
  - Consist of update transactions
  - Use database transaction logs for data capture
- Incremental loads for fact tables
  - Load as frequently as feasible
  - Use indexes
  - Apply parallel processing techniques

Data Quality

- What is Data: An abstraction/representation/ description of something in reality
- What is Data Quality: Accuracy + Data must serve its purpose/user expectations

Indicators of quality of data

- **Accuracy**: Correct information, e.g., address of the customer is correct
- **Domain Integrity**: Allowable values, e.g., male/female
- **Consistency**: The content and its form is same across all source system, e.g., product code of a product ABC in one system is 1234 then in other system it must be 1234 for that particular product

Indicators of quality of data (Cont.)

- **No Redundancy**: Data is not redundant, if for some reason for example efficiency the data is redundant then it must be identified accordingly
- **Completeness**: There are no missing values in any field
- **Conformance to Business rules**: Values are according to the business constraints, e.g., loan issued cannot be negative
- **Well defined structure**: Whenever the data item can be divided in components it must be stored in terms of components/well structure, e.g., Muhammad Ahmed Khan can be structured as first name, middle name, last name. Similar is the case with addresses

Indicators of quality of data (Cont.)

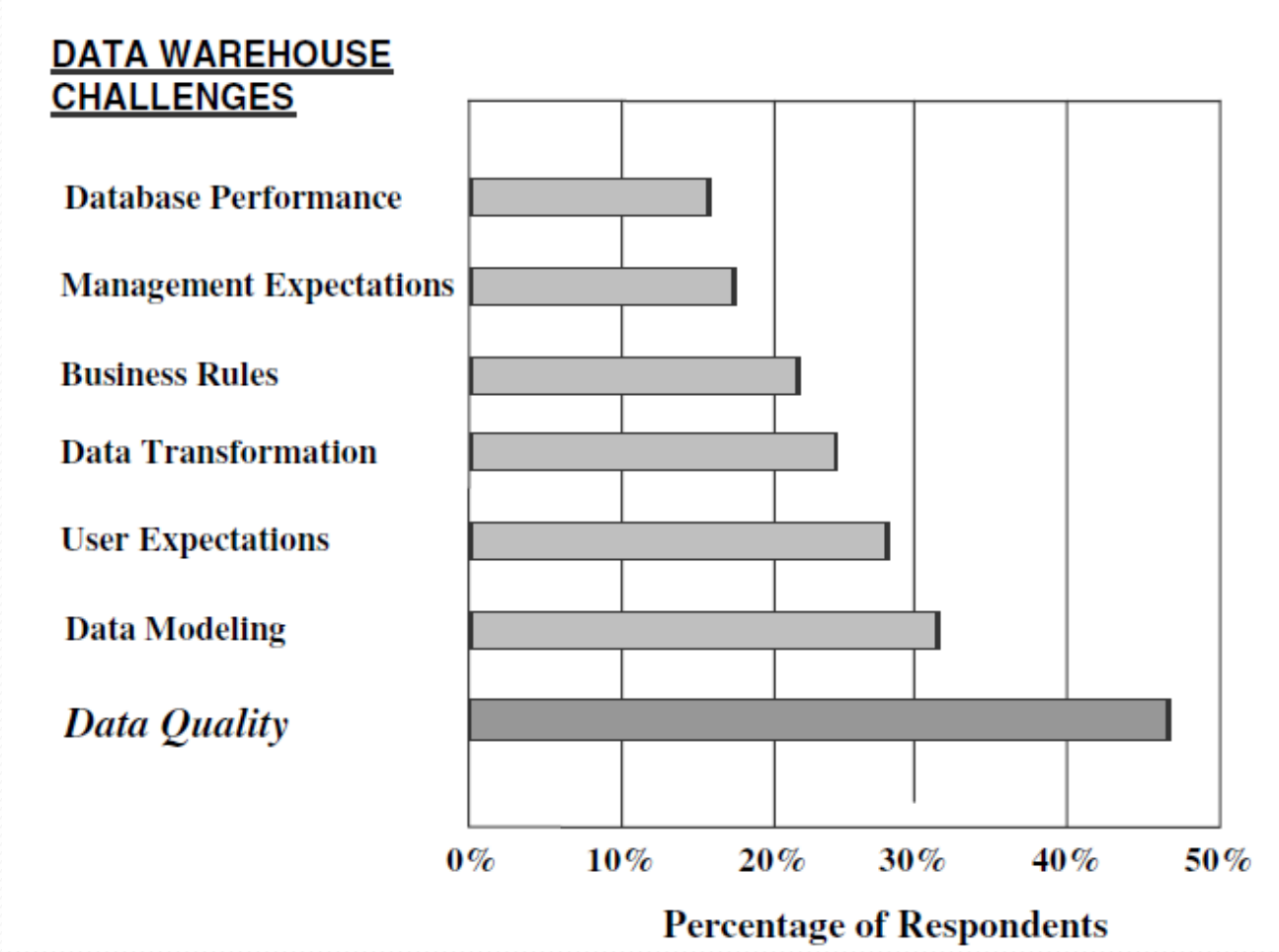
- **Data Anomaly**: Fields must contain that value for which it was created, e.g., State field cannot take the city name
- **Proper Naming convention**
- **Timely**: timely data updates as required by user
- **Usefulness**: The data elements in data warehouse must be useful and fulfill the requirements of the users otherwise data warehouse is not of any value

Indicators of quality of data (Cont.)

- **Entity and Referential Integrity**: Entity integrity means every table must have a primary key and it must be unique and not null. Referential integrity enforces parent child relationship between tables, you can not insert a record in child table unless you have a corresponding record in parent table

Problems due to quality of data

- Businesses ranked data quality as the biggest problem during data warehouse designing and usage



Types of Data Quality Problems

- Dummy values: For example, to pass a check on postal code, entering dummy or not precise information such as 4444 (dummy) or 54000 for all regions
- Absence of data values: For example not a complete address
- Unofficial use of field: For example writing comments in the contact field of the customer

Types of Data Quality Problems

- **Cryptic Information**: At one time operation system was using 'R' for "remove" then later for "reduced" and some other time point for "recovery"
- **Contradicting values**: compatible fields must not contradict, e.g., two fields ZIP code and City can have values 54000 and Lahore but not some other city name for ZIP code 54000

Types of Data Quality Problems

- Violation of business rule: Issued loan is negative
- Reused primary keys: For example, a business has 2 digit primary key. It can have maximum 100 customers. When a 101th customer comes the business might archive the old customers and assign the new customer a primary key from the start. It might not be a problem for the operation system but you need to resolve such issues because DW keeps historical data.

Types of Data Quality Problems

- Non-unique identifiers: For example different product codes in different departments
- Inconsistent values: one system is using male/female to represent gender while other system is using 1/0
- Incorrect values: Product Code: 466, Product Name: “Crystal vas”, Height:”500 inch”. It means that product and height values are not compatible. Either product name or height is wrong or maybe the product code as well

Types of Data Quality Problems

- Erroneous integration: A person might be a buyer or seller to your business. Your customer table might be storing such person with ID 222 while in seller table it might be 500. In data warehouse you might need to integrate this information. The persons with IDs 222 in both tables might not be same

Data Cleansing Decisions/ Issues

- ***Which Data to Cleanse***

  The cost of cleaning up all data in the data warehouse is enormous.

  By ignoring the cleansing of unimportant data as long as all the important data is cleaned up.

- ***Where to Cleanse***
  - staging area
    - Cleansing the data in the staging area is comparatively easy.
    - But Data pollution will keep flowing into the staging area from the source systems.
    - The source systems will continue to suffer from the consequences of the data corruption.
  - Source System
    - If you attempt to cleanse the data in the source systems, you are taking on a complex, expensive, and difficult task.

Data Cleansing Decisions/ Issues

- ***How to Cleanse***
  - Vendor tools
  - In-house programming

- ***How to Discover the Extent of Data Pollution***

  Make a list that reflects the sources of pollution found in the environment, then determine the extent of the data pollution with regard to each source of pollution

- ***Setting Up a Data Quality Framework***

  Most companies serious about data quality pull all these factors together and establish a data quality framework. Essentially, the framework provides a basis for launching data quality initiatives. It embodies a systematic plan for action. The framework identifies the players, their roles, and responsibilities.

- Institute a data quality framework.
- Assign roles and responsibilities.
- Select tools to assist in the data purification process.
- Prepare in-house programs as needed.

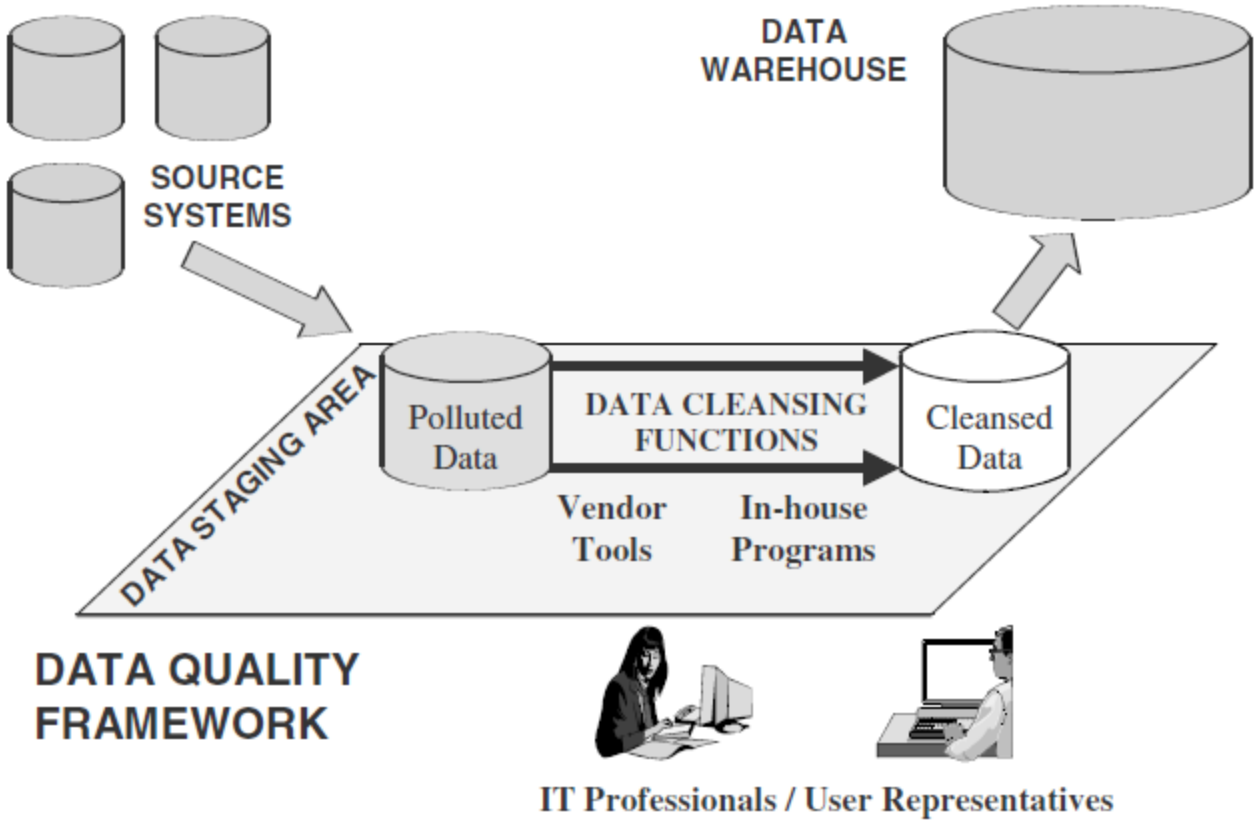
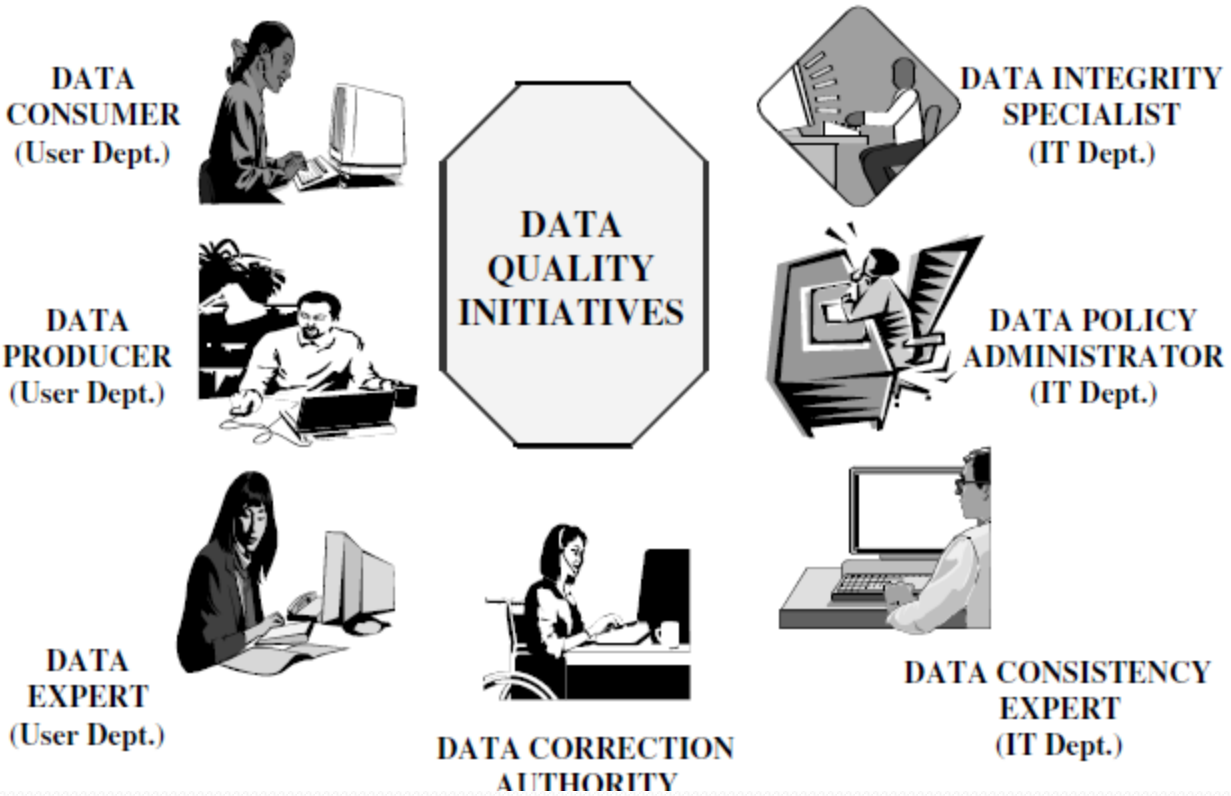


Figure 13-7 Overall data purification process.


DATA CONSUMER (User Dept.)
DATA PRODUCER (User Dept.)
DATA EXPERT (User Dept.)
DATA QUALITY INITIATIVES
DATA CORRECTION AUTHORITY
DATA INTEGRITY SPECIALIST (IT Dept.)
DATA POLICY ADMINISTRATOR (IT Dept.)
DATA CONSISTENCY EXPERT (IT Dept.)