

## Enforcing Data Quality via Metadata Lineage

Wilshire Metadata Conference

David Plotkin

Data Quality Manager

Wells Fargo Consumer Credit Group

## Introduction

- Understanding data lineage
- Mapping the information chain
- The data lineage metamodel
- Identifying the data quality rules with data profiling
- Segregating data that fails the rules
- Managing transformation metadata
- The Benefits of quality data lineage

## Understanding data lineage

WELLS  
FARGO

- Except for the initial data input point, all data comes from someplace:
  - ✿ System of record
  - ✿ Extract
  - ✿ Result of a transformation (many different kinds)
  - ✿ Data warehouse
  - ✿ Outside information provider
- To figure out what rules hold, you must know where the data came from.
- To understand the information flow, you must map the information chain

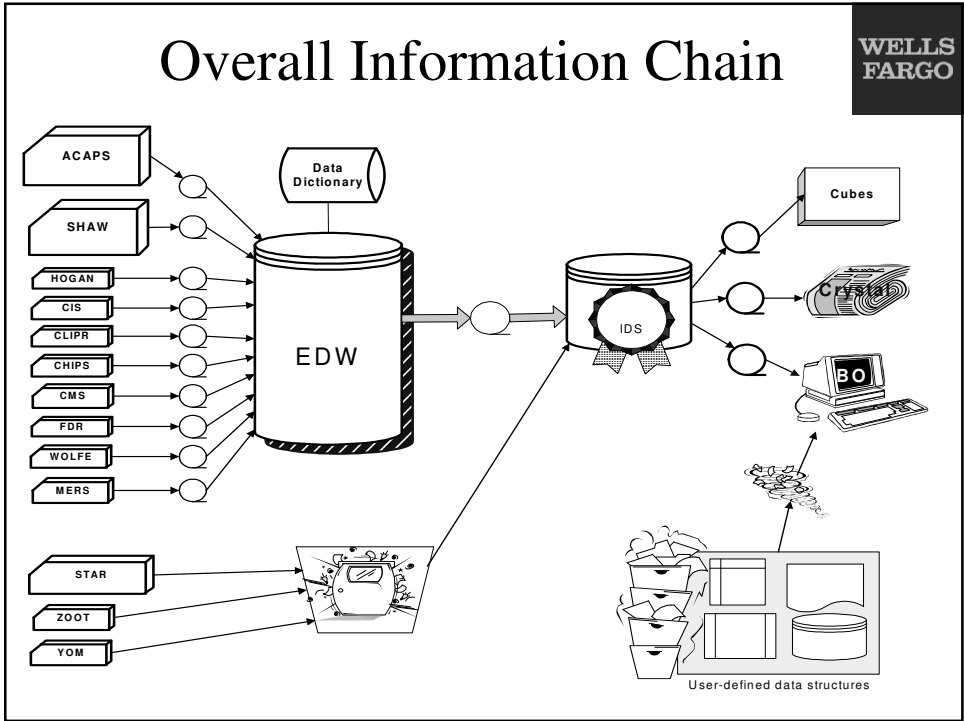
## Mapping the Information Chain

WELLS  
FARGO

- The information chain maps the flow of data, including:
  - ✿ Data Producers (including systems of record)
  - ✿ Data consumers (including reports)
  - ✿ Transformations
  - ✿ Data Stores (with metadata)
- Provides a “map” for where data rules need to be applied:
  - ✿ What data quality rule?
  - ✿ Applied at what point?

# Enforcing Data Quality Through Data Lineage Metadata

David Plotkin, Wells Fargo Consumer Credit Group



## Importance of the Information Chain

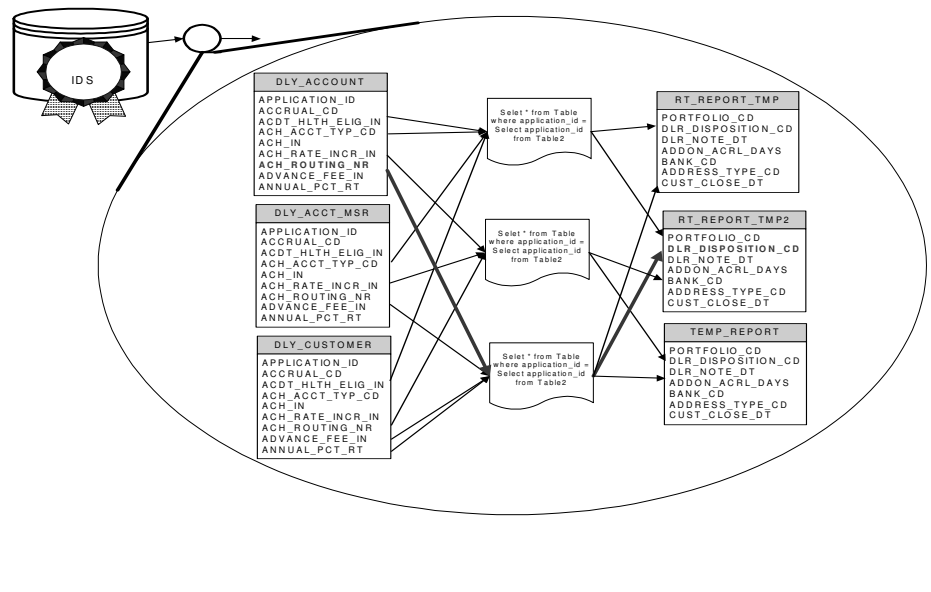
- Addresses the need to understand data flows: capture, movement, and transformation of data.
- Provides the context for lineage – the overall “map” that governs the relationship between data elements.

# What is Lineage?



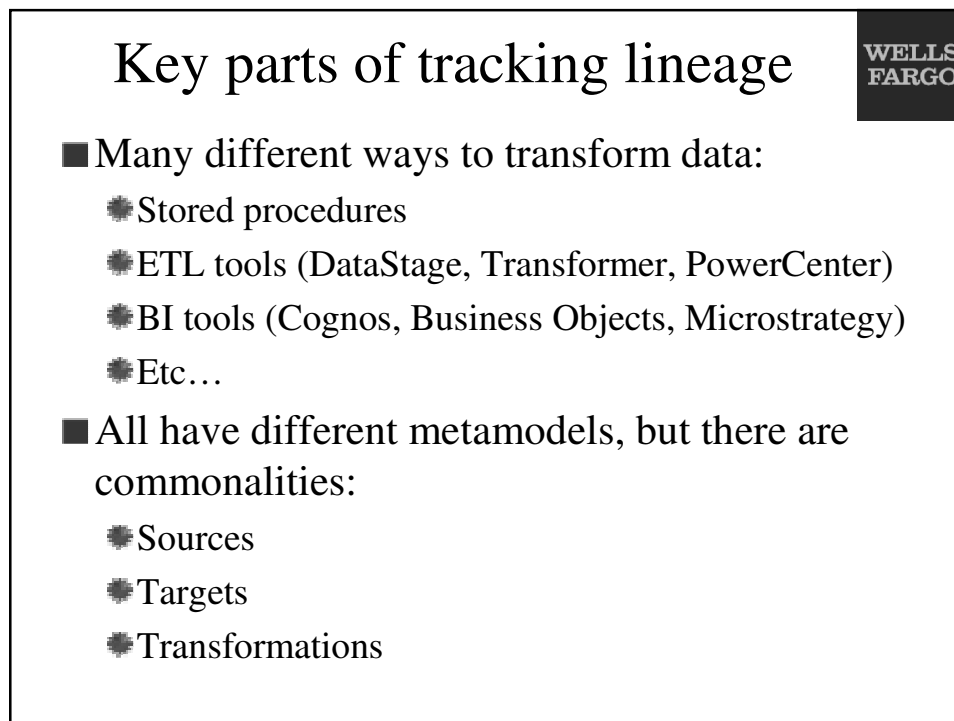
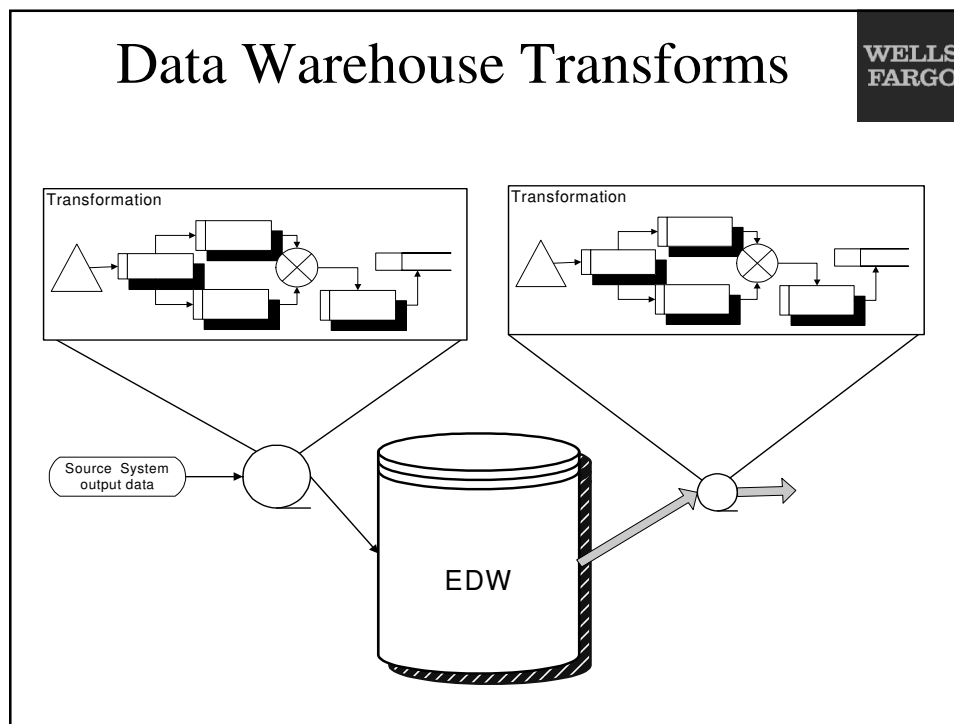
- Shows how data elements are related to each other:
  - Identical elements. Only need to define these elements once, cutting down on work. Also ties report element definitions to their source.
  - Transformed elements. Helps define elements because you know where they came from and how they were operated on.
- Simplifies Stewardship:
  - All “identical” elements have the same steward.
  - Stewardship propagated across the chain.

# Stored Procedure Transforms



# Enforcing Data Quality Through Data Lineage Metadata

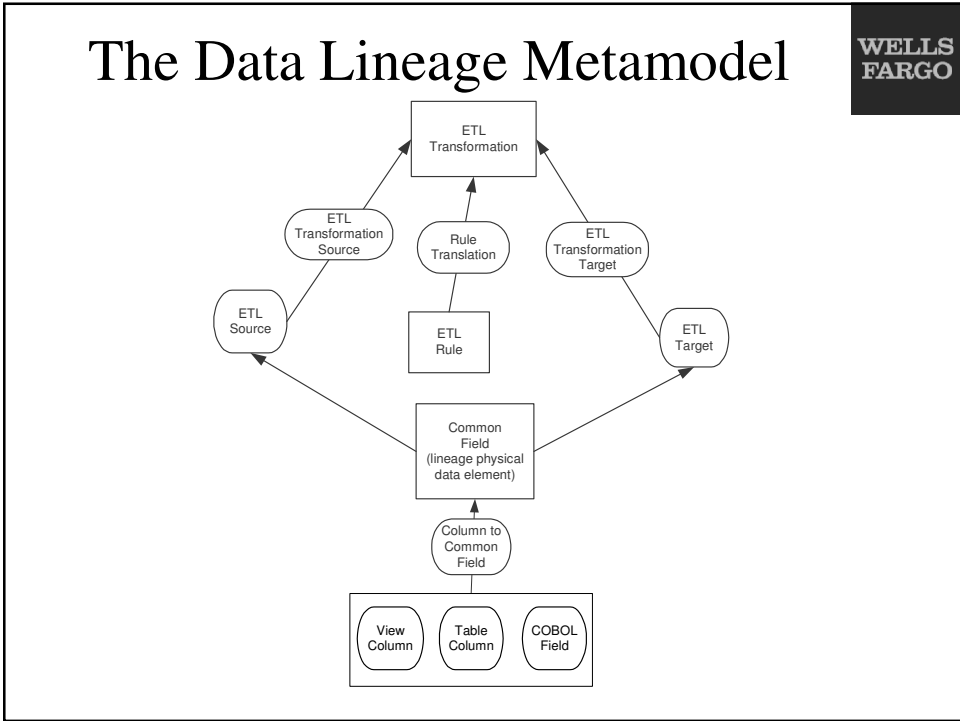
David Plotkin, Wells Fargo Consumer Credit Group



The DAMA International Symposium & WILSHIRE Meta-Data Conference  
Denver, Colorado ● April 23-27, 2006

# Enforcing Data Quality Through Data Lineage Metadata

David Plotkin, Wells Fargo Consumer Credit Group



## Discerning the Rules: Data Profiling

### ■ Data Profiling:

- Enables you to find and document the data quality rules
- Test rules you think are true
- These rules can be attached at the point in the Information Chain where they should be enforced.
- Data Profiling may discover issues with your data quality which you can use projects to fix.

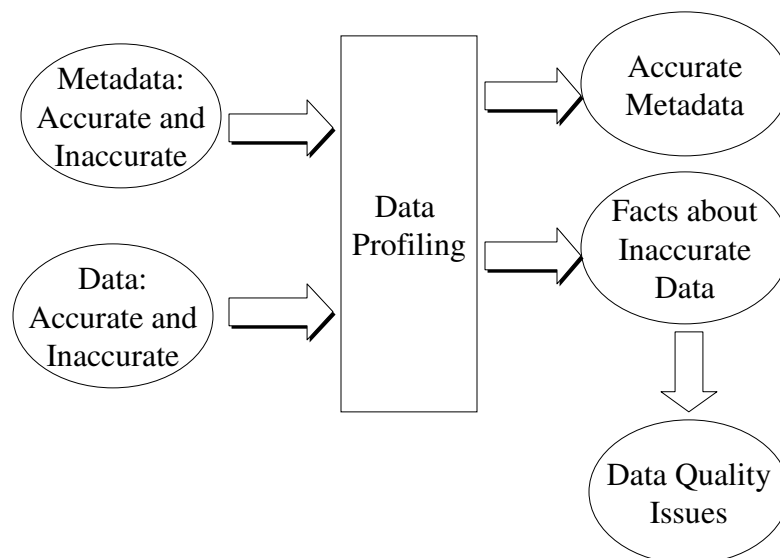
## The two parts of Data Profiling

WELLS  
FARGO

- *Discovery* uses sophisticated algorithms to detect possible patterns/rules within your data
- *Assertion testing* enables you to test rules Subject Matter Experts tell you are true and determine how well they fit the data
- In both cases, the “rules” must be reviewed by subject matter experts
- The results (rules tested, rule proposer, data that does not conform to the rules, subject matter experts consulted, etc) must be stored in a repository.
- The output is better metadata (rules) and better information about your data

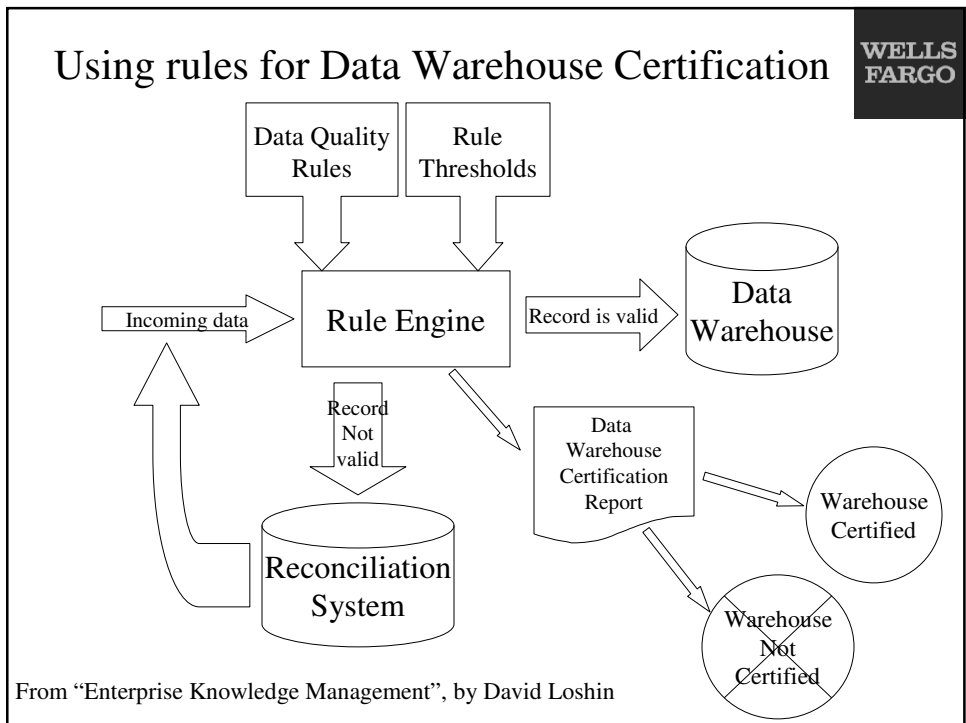
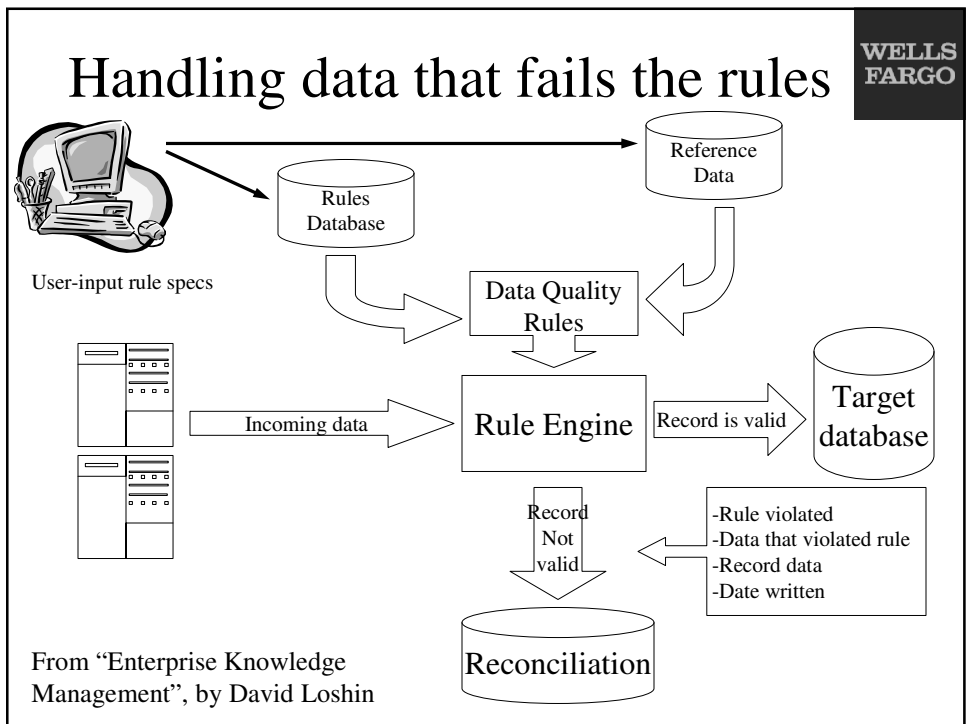
## Output from Data Profiling

WELLS  
FARGO



# Enforcing Data Quality Through Data Lineage Metadata

David Plotkin, Wells Fargo Consumer Credit Group



The DAMA International Symposium & WILSHIRE Meta-Data Conference  
Denver, Colorado • April 23-27, 2006

## Continuous Monitoring

WELLS  
FARGO

- Control (the “c” in Six Sigma’s DMAIC) is achieved by monitoring data quality via profiling.
- If you find a data quality “leaker”:
  - Can use lineage to trace it back to its source
  - Identify where the quality is being degraded
  - Correct the process to stop the leaking!



## Cultural Shift in Managing Lineage Metadata (1)

WELLS  
FARGO

- Impact analysis the old way (before tracking lineage):
  - Developers did it with little documentation
  - Testing was painful because rules were not well-understood
  - Done manually to figure out what transformations were touched by the change.
  - Impacts often missed, requiring rework with subsequent delays and poor customer experience
  - Every data quality effort started out with queries to find out:
    - Where data came from
    - What its rules were
    - How it was transformed

## Cultural Shift in Managing Lineage Metadata (2)

WELLS  
FARGO

- Impact analysis the new way (after tracking lineage):
  - Data Analysts do it: rule info and lineage is in repository
  - Testing uses the data quality requirements to test (50% saving)
  - Impact analysis is automated based on lineage in repository:
    - Lineage accurate because transformations are scanned into repository
    - Takes much less time (95% savings!)
    - Much higher accuracy, leads to 90% savings on rework
  - Data Quality efforts no longer need to talk to “old-timers” to figure out how things work.
  - Savings of >70% in figuring out where the data went wrong.
    - What rule got violated?
    - Where (in the Information Chain) is that rule applied?

## More Benefits of quality lineage

WELLS  
FARGO

- 15 Business Analysts and 20 Customer Analysts save about 30% of their time trouble-shooting unexpected results.
- 15 Developers save 25% of their time analyzing impacts to code needed for projects.
- Same 15 Developers save another 25% of their time helping analysts figure out where the problems are.  
*Developers now twice as productive!!*
- Huge impact on customer satisfaction – incidents where data is unavailable for multiple days (due to poor impact analysis) has been reduced by 90%.