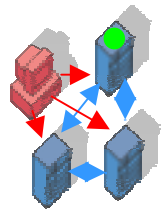
# Dirty Data

## The Big Unknown

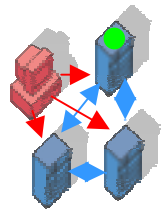


# Taurus Software: Dirty Data

**What costs the nation $600 billion per year?**

1. **The Defense Budget**
2. **Smoking**
3. **Dirty Data**

**Do we have dirty data?**



# Taurus Software: Dirty Data

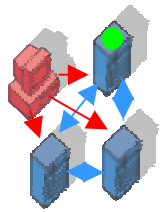
**Why is it a problem?**

**It’s always been a problem**

**Lehman Bros**

**National Steel**

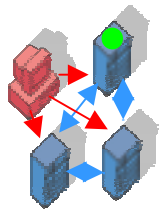
**Nike**



# Taurus Software: Dirty Data

**Your new RDBMS:**

- **Is strongly typed**
- **Has rules for data entry**
  - **List, range, format, rule**
- **Enforces referential integrity**



# Taurus Software: Dirty Data

**Your new RDBMS:**

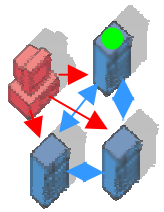
**Enforces R/I**

**Relationships and indexes**

**Adding an order**

**Bulk load**

**Transaction**



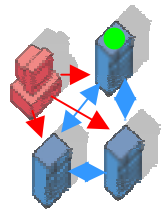
# Taurus Software: Dirty Data

**Two major problems**

- **Data items (inaccurate)**
- **Data entities (incomplete or illogical)**

**Caused by:**

- **Data entry and validation**
- **Poorly typed and designed DBMS**
- **Poor DELETE management**
- **Rogue programs**

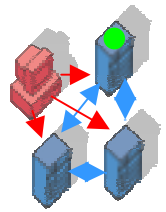


# Taurus Software: Dirty Data

## Data entity problems

**Orders without customers**
**Customer keys without customer detail**
**Multiple master records for the same customer**
**Customers who have never ordered**

**A hysterectomy claim for a male patient**
**Zip codes don’t exist in the city**
**A 400 lb shipment waiting for the mailman**



# Taurus Software: Dirty Data

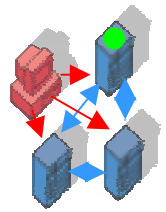
## Data item problems

### Illegal data:

- 23A 7  in a numeric
- ??/??/?? In a date
- characters < CHR(32) and > CHR(128)
- NULL/non-NULL, space and zero

### Format differs for same datatype

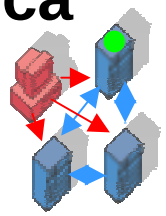
- 3/31/03, 31/03/2003, 3/2003
- (650) 482-2022, 650-482-2022



# Taurus Software: Dirty Data

**Data item problems**

- **Range errors**
  - **Patient age is 210**
  - **Order date is 1 Jan 2055**
  - **Toothbrush price is $599.95**
  - **19 fridges on one consumer order**
- **List errors**
  - **Mr., Mrs., Mx.**
  - **Main St, Chico, Main Str. chico, Chica**
  - **M, F, B, G**



## Solutions for the problem

**Analyze the scope**

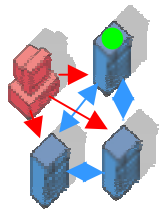
- **Read the data and and validate for a target**
- **Buy, build or outsource?**

**Define cleansing strategy**

- **Source, in-flight, target**

**Cleanse**

- **Today's data**
- **What about tomorrow's data?**



# Taurus Software: Dirty Data

**A demonstration**

**of IMAGE Dirty Data**

**And how**

**Oracle and SQL Server**

**React**