



Original Software

# Importance of Managing Test Data

## **“Saint Valentines Data Massacre”**

John Miner  
Original Software  
Director of Operations



ChoicePoint



POLO.COM  
RALPH LAUREN

NORTHWESTERN  
UNIVERSITY

ERNST & YOUNG  
Quality In Everything We Do

EQUIFAX



LexisNexis

HOUSE COOPERS

usbank  
Five Star Service Guaranteed



# Data Security Breaches



cardsystems  
the power of the right solution

Bank of America



Deloitte



citi financial

Honeywell

Internal Revenue Service  
United States Department of the Treasury

LINENS-N-THINGS  
dream big, pay little

WELLS FARGO

OHIO  
UNIVERSITY



SENTRY  
INSURANCE



KAISER PERMANENTE

UNITED STATES DEPARTMENT OF  
DEFENSE



MARINES  
THE FEW. THE PROUD.



H&R BLOCK



Original Software

## Top 10 US Customer Data Loss Incidents

Company / Organization	Number of affected customers	Date of initial disclosure
CardSystems	40 million	June 17
Citigroup	3.9 million	June 6
DSW Shoe Warehouse	1.4 million	March 8
Bank of America	1.2 million	February 25
Time Warner	600,000	May 2
LexisNexis	310,000	March 9
Ameritrade	200,000	April 19
Polo Ralph Lauren	180,000	April 14
ChoicePoint	145,000	February 15
Boston College	120,000	March 17
Source: InformationWeek; public disclosures by companies		

Incidents Reported between 02/05 and 08/06

**256 Data Breaches**  
**90,722,452 Customers Impacted**



Original Software

## Surveys Conducted (Ponemon Institute)



23 million U.S. adults

**"We regret to inform you that your data was compromised or lost"**

- 20% terminated their accounts
- Another 40% are thinking about it.



**\$6.3 million  
average  
per-incident  
cost**



**40 % blamed on  
outsourcers,  
contractors,  
consultants,  
business partners**



Original Software

# Surveys Conducted (Ponemon Institute)

Per incident...

**\$5 million** (direct costs)

- legal services
- notification letters
- telecommunications costs
- product discounts
- increased staffing

**+** **\$7.5 million** (opportunity costs)

- customer loss
- increased difficulty recruiting new customers

**+** **\$1.3 million** (indirect costs)

- Lost employee productivity



**=** On average  
**\$140** (per lost record)

From "Lost Customer Information: What Does a Data Breach Cost Companies?"



Original Software

# Penalties are Significant



If you use a credit card or have applied for credit, your personal information is in the ChoicePoint system!

- “The Federal Trade Commission hit data broker ChoicePoint Inc. with the [largest civil penalty](#) in the agency’s history for allowing sensitive customer information to get into the hands of con artists.
- The commission levied a [\\$10-million penalty on top of \\$5 million in restitution](#) – a total that amounts to more than 10% of the company’s 2005 profit.”

*Los Angeles Times, January 27, 2006 – ChoicePoint is Fined for Data Breach*



# The world is watching!

- “Privacy Rights Clearinghouse”
  - 236,808,763 records containing sensitive personal information involved in breaches since January 2005
  - 233 incidents in 2008 (*as of 8/25/08*)
  - <http://www.privacyrights.org/ar/ChronDataBreaches.htm>
- Risk of exposure can cause significant damage
  - Lawsuits, Reputation, Loss of Customers, etc.





# Reaction

- Forced Internal data procedural audits
- Regulatory Compliance
  - Sarbanes-Oxley Act
  - PCI
  - United States Gramm-Leach-Bliley Act
  - Health Insurance Portability and Accountability Act (HIPAA)
  - Australia Privacy Amendment Act of 2000
  - European Union Personal Data Protection Directive
  - Hong Kong Personal Data (Privacy) Ordinance of 1995
  - New Zealand Privacy Act of 1993
  - ETC ...
- Over 50% of companies now have a Data Privacy Officer





# Most Common Data Security Loss



**Ponemon Institute**  
“Data Security Tracking Study”

“Gartner research found that 70 percent of security incidents incurring actual losses are inside jobs.”

This speaks to a growing industry-wide understanding that outside hackers are no longer the main threat to data security. “

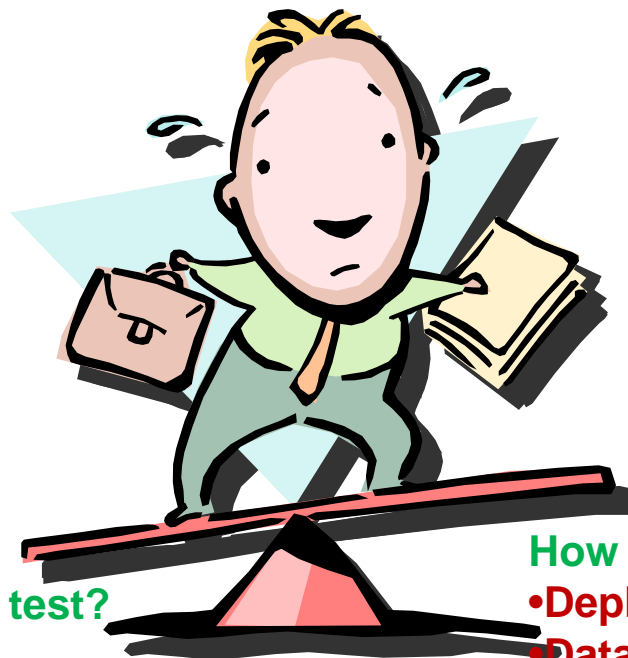
**LabRat Magazine, May 2005**  
“Preventing Confidential Internal Data Leaks”



# The Quality Conundrum

- Resources, risk and deadlines are often in conflict

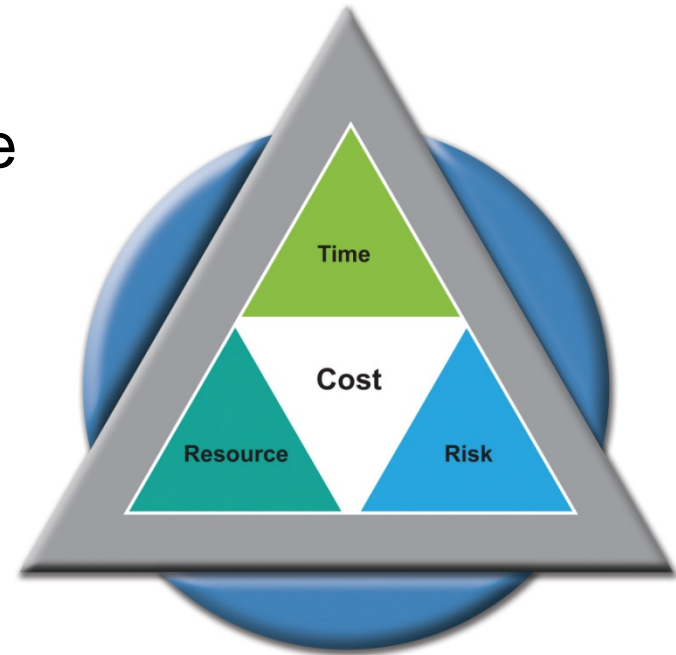
How long do I have?



How much do I test?

How much risk am I willing to take?

- Deploying untested applications
- Data privacy as related to test data





# Secure Approaches

- Don't test, just put it in production.
- Test lightly, carefully in production.
- Encrypt data at the source.
  - Challenging to implement, support and test
- Create a test/staging environment with copies of production data
  - Locked down, clean-room approach
  - Extremely limiting and inefficient for testing efforts
  - Does not address need to expose systems externally
- Create a test/staging environment with de-identified test data
  - Best option for obscuring personal information
  - Addresses need to expose systems externally

*Secure but dangerous!*  
*Let's be serious please.*  
*What does that say?*



# Why is test data hard to secure?

Access beyond normal application interfaces is needed for testing and support purposes

- Need to look “under the covers”
- Need to directly manipulate test data
- Need to share data with developers – or application vendor





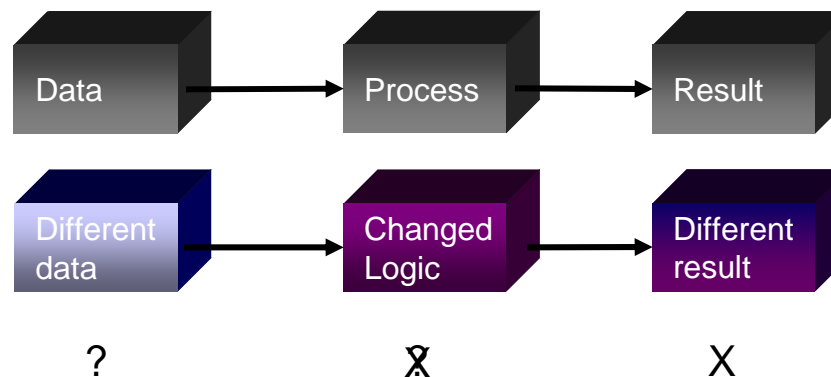
# Test Data Strategy

Address data privacy concerns

- De-Identify your test data

Remove the data variable from regression testing:

- Known starting point(s) for test data
- Protect test data for easy reuse
- Manage test data as part of the test pack





# Test Data

Often the bane of good, secure regression testing

## How

- ... much time is spent creating/maintaining it?
- ... often is bad data the core cause of testing issues?
- ... is the data reset to known states?
- ... does delays in refreshing impact cycle time of tests?
- ... is data management tied to testing automation?
- ... is it secured, scrambled?

## Who

- ... creates/manages/shares the data?
- ... has access? QA? Dev? DBA's?
  - *Full data detail in all its glory?*
  - *Common source of breaches*



How long do I have?

How much do I test?

How much risk am I  
willing to take?



Original Software

# How to get my Test Data







# Test in Production

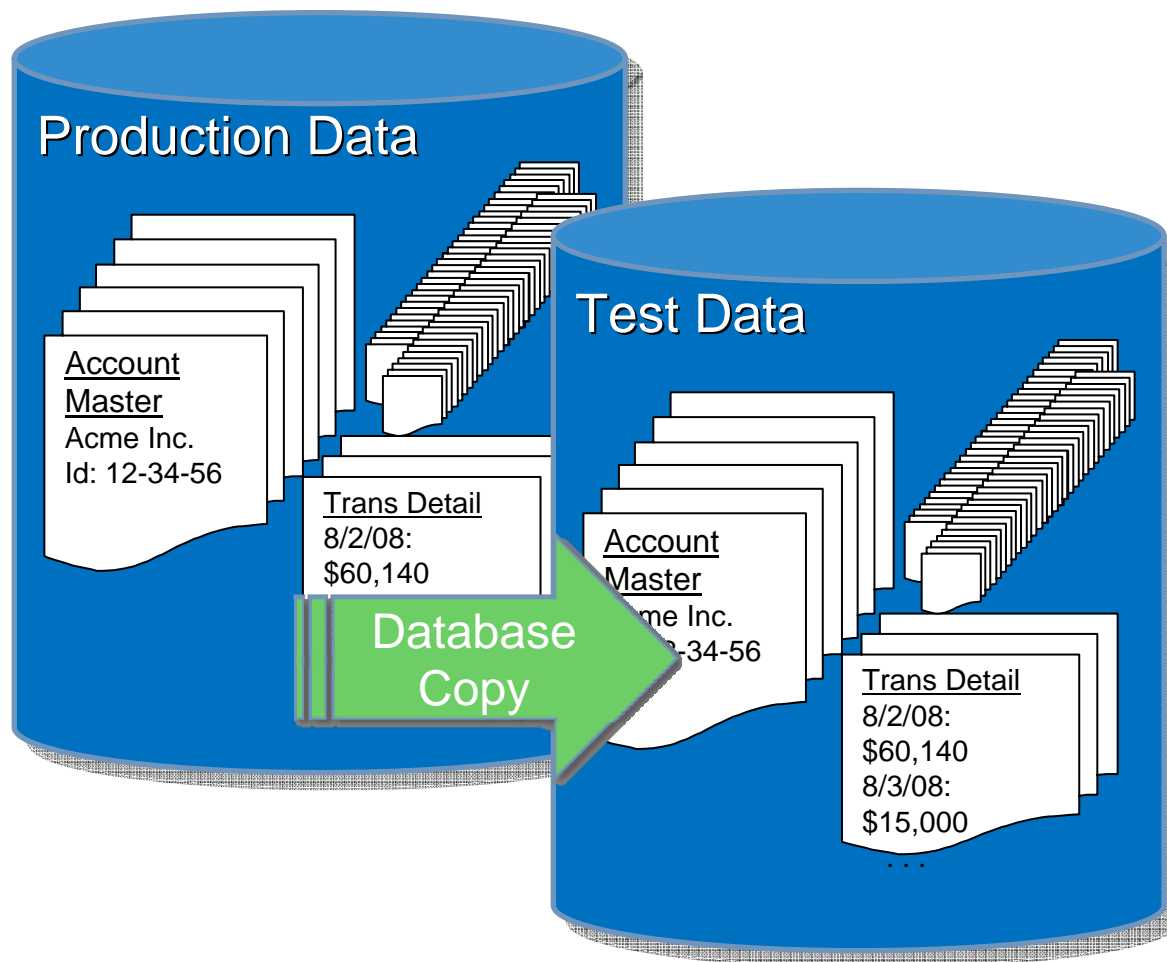
- To heck with Test Data and Test Environments, just put it in production and hope for the best.





Original Software

# Copy Production Data



Can be ...

- Too much
- No data privacy
- Difficult to “reset”
- Can hinder test coverage
- Most common

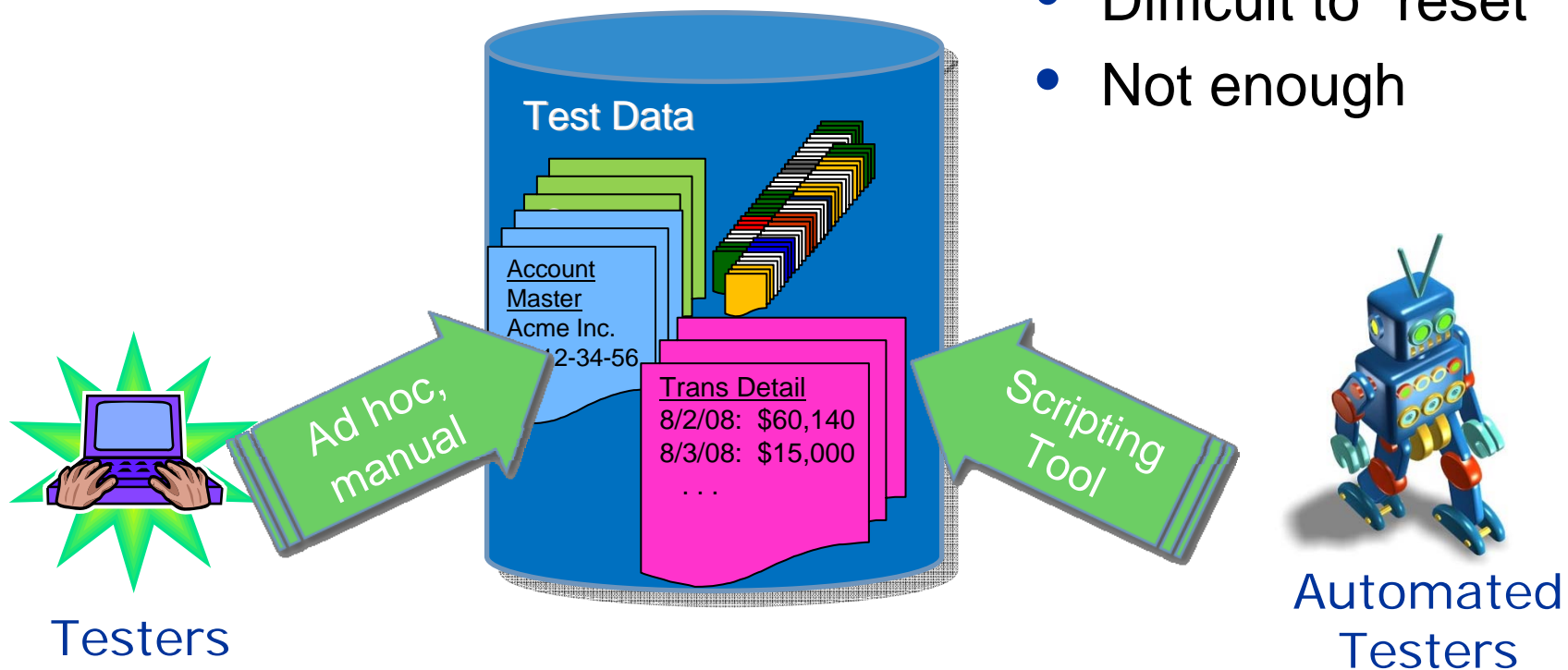


Original Software

# Generate Test Data

Can be ...

- Incomplete
- Not real world
- Difficult to “reset”
- Not enough

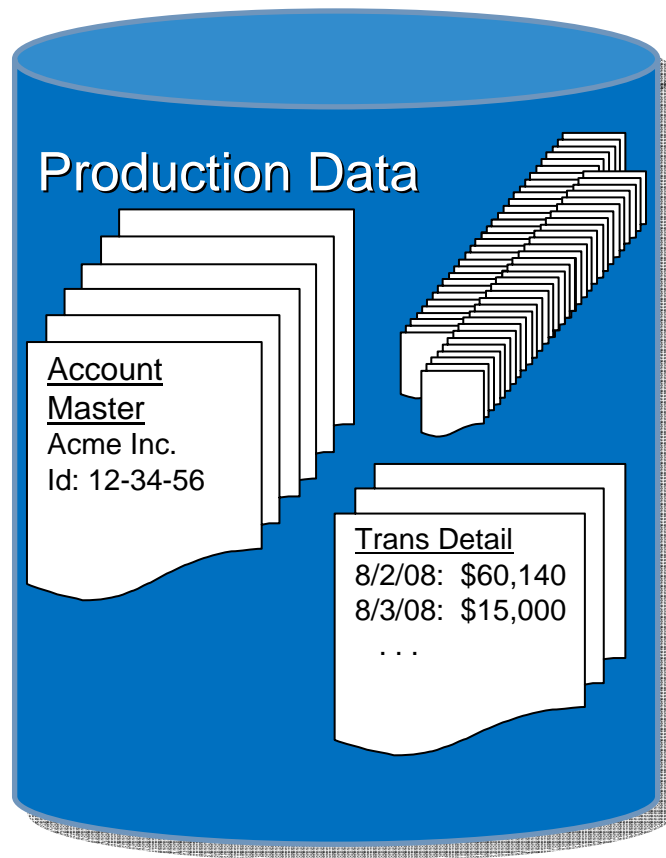




Original Software

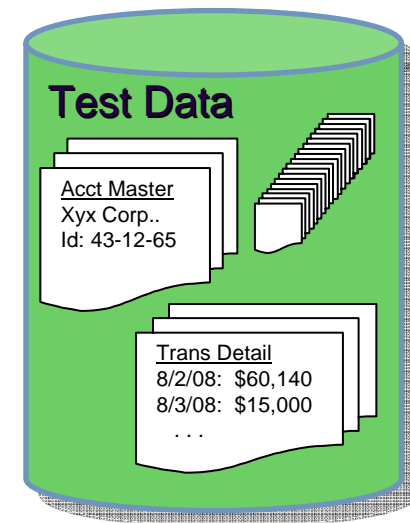
# Extract from Production

- Build and maintain SQL scripts
- Use test data automation tools
- Reuse extraction criteria against changed production data
- Apply scrambling to **de-identify**



SQL Scripts

Test Data  
Automation tools



Smaller, more manageable  
data environment



# Extraction from Production

If Production data available, the best option

Smaller, more manageable

- *Multiple copies – QA, Dev, Support*
- *Retain versions, save with releases*

Extract variations

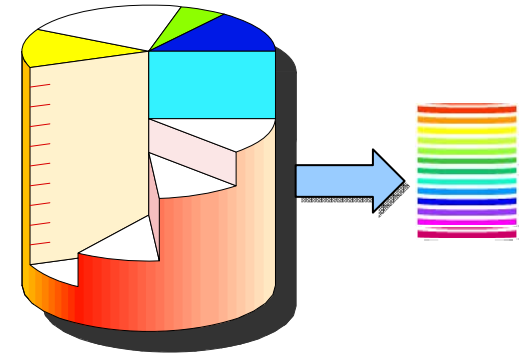
- *Can be re-extracted to include new data*
- *Different extractions to support different tests*

Seed your test data

- *Max records, fraction or percentage of file*
- *Sample important field combinations*

Leverage for extract process for customer environments

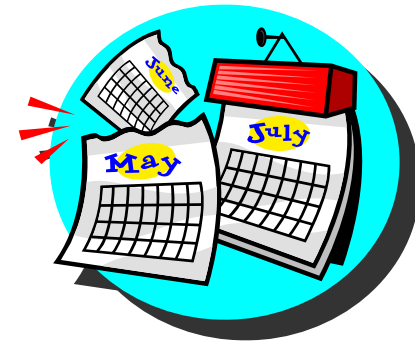
May require scrambling to secure privacy





# Data Warping & Scrambling

- Warp Date/Time in test data
  - Simulate key processing dates
    - *End of month*
    - *End of year*
  - Extend value of test data
- Scramble / De-identify sensitive data
  - Vertical and logical scrambling
  - Masking
  - Selective generation
- Maintain referential integrity
  - Ensure scrambling and warping are applied consistently



b1	t2	g3
p1	g2	b3
t1	b2	p3
g1	p2	t3

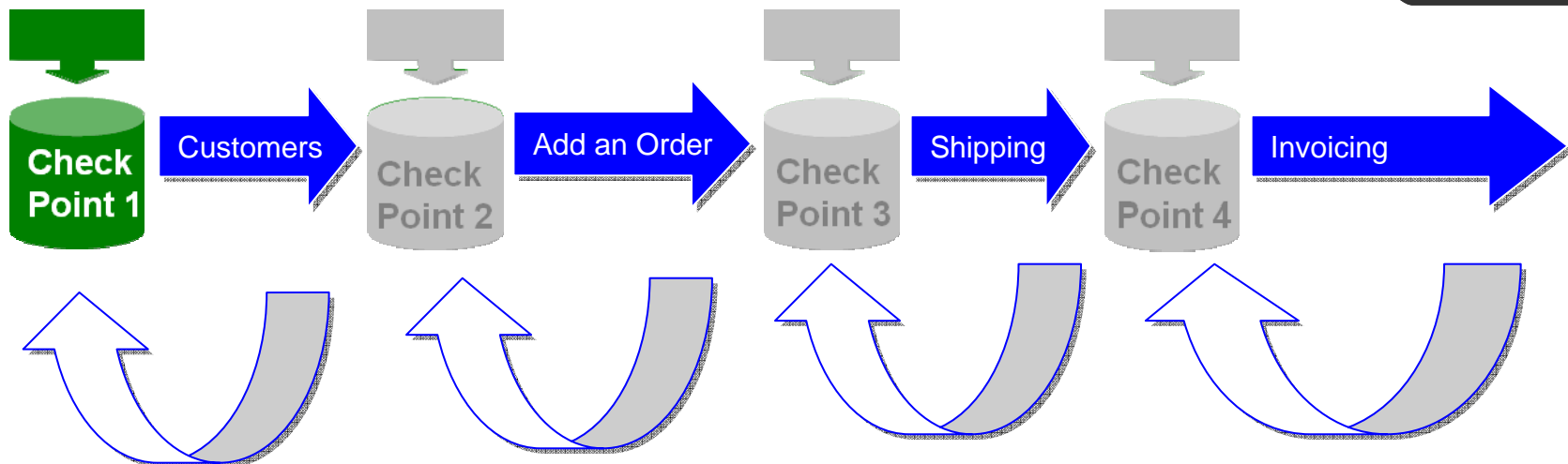


Original Software

# Protect Good Test Data

- Maintain pristine copy of data via checkpoints & rollbacks
  - Reduce need for new copies
  - Repeatability of tests using the exact same data
  - Reset to checkpoints within automated scripts

Incremental Check Points



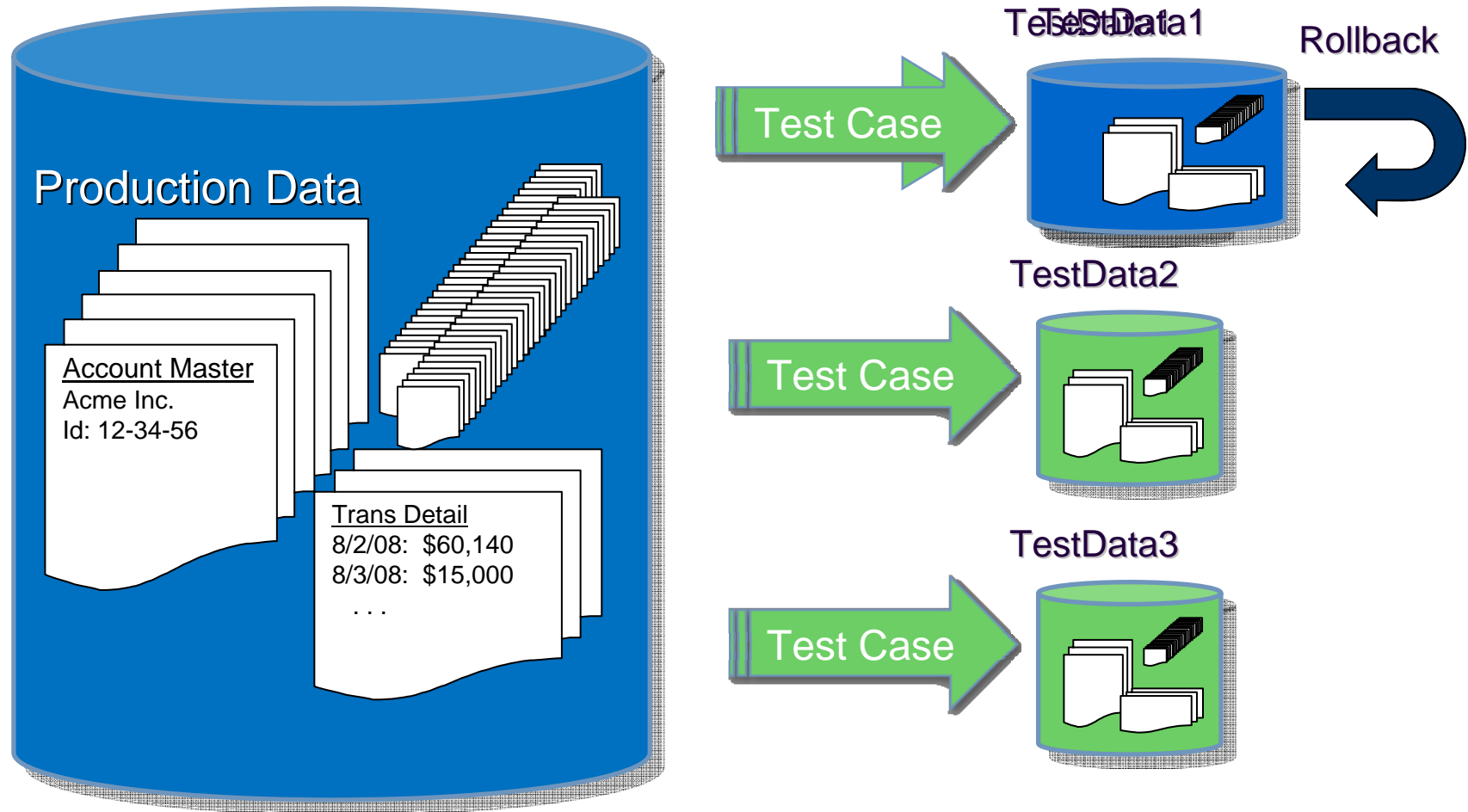
Incremental Roll-back





Original Software

# Testing Environment Management





# Importance of Test Data

- Test data is invaluable if it
  - Is available not inaccessible - expedite the process of testing
  - Is representative not repetitive - reduce the scope of testing
  - Is manageable not large - disk space can be a scarce
  - Is malleable not rigid - adjust dates to test past or future
  - Is easily refreshed not reliant on DBA's - be master of own timeline
- Securing test data is critical so that it
  - Has integrity not inconsistent - don't introduce defects
  - Is de-identified not recognizable - open testing to more people
  - Is valid not real - share test scenarios



Original Software

# Please visit [www.origsoft.com](http://www.origsoft.com)

**The ideal place for you to get up to speed with the key issues that surround the world of software testing.**

## Our Solutions

- Test Planning & Management
- Manual Functional Testing
- Automated Functional Testing
- Test Data Management
  - Creation, Extraction, Scrambling, Warping, Checkpoints, Rollbacks, Environment Management
- Database Transaction Visibility





Original Software



Thank You

John Miner  
jminer@origsoft.com  
630-321-0223  
[www.origsoft.com](http://www.origsoft.com)