Data Quality Considerations for Long Term Data Retention

ABSTRACT------------------------------

IT shops are faced with new requirements to retain specific data for long periods of time; often decades. This is overloading operational databases and causing multiple problems. Data Quality is a factor that is often overlooked when crafting solutions. Yet it presents unique and challenging problems for the data management expert.

This presentation explains the concept of long term data retention of database data, outlining several problems concerning the quality of data. It covers problems created by changing data structures, loss of expert people, loss of access to the originating applications, changing reference data, and the need to work with data from multiple sources for the same application. The presentation shows how data with high quality can decay over time.

The data quality dimensions of completeness, clarity, and understanding are more applicable than the dimension of data accuracy. The consequences of ignoring data quality issues are also addressed.

BIOGRAPHY------------------------------

Jack Olson

Jack Olson has worked in the commercial software development business for 40 years. His career has mostly consisted of architecting solutions to IT problems in the area of database systems and tools. He spent 17 years in IBM development labs working on such notable products as CICS, IMS, DB2, and AIX. He worked at BMC software as Corporate Architect, as Vice President of Development at Peregrine Systems, and as Chief Technology Officer for Evoke Software and NEON Enterprise Software. He has worked with several other startup companies in recent years as a consultant, advisor, or board member. He is currently an independent consultant. Jack has published two books: “Data Quality: the Accuracy Dimension”, 2003 and “Database Archiving: How to Keep Lots of Data for a Long Time”, 2008. Jack has a BS degree in Mathematics from the Illinois Institute of Technology and an MBA from Northwestern University.

SvalTech

Data Quality Considerations for Long Term Data Retention

"Database Archiving: How to Keep Lots of Data for a Long Time"
Jack E. Olson, Morgan Kaufmann, 2008

Jack E. Olson
jack.olson@SvalTech.com
512-565-9584

1

SvalTech

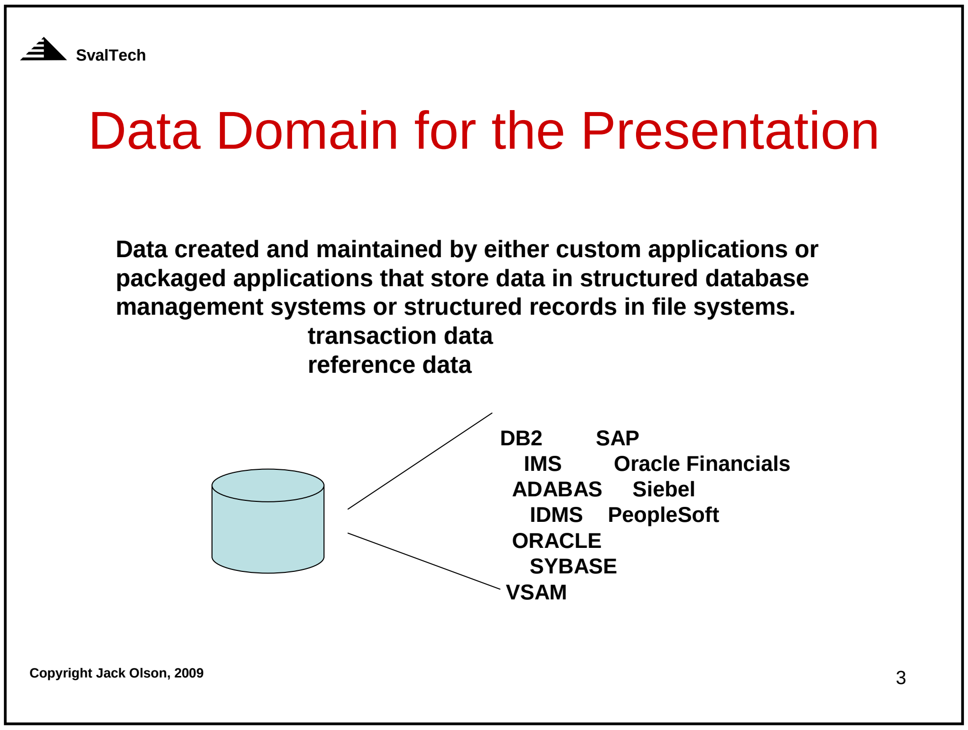
Presentation Roadmap

Define Data Domain
Define Long Term Data Retention
Define Data Time Lines

Problems With Using a Single Database Approach
- **Loss of clarity of understanding**
- **Metadata change corruption**
- **Reference data change corruption**

Problems with Using a Segmented DB Approach
- **Database Archiving**
- **Failure to associate Metadata to segments**
- **Failure to protect from data loss**
- **Failure to advise user of Metadata differences**

2



SvalTech

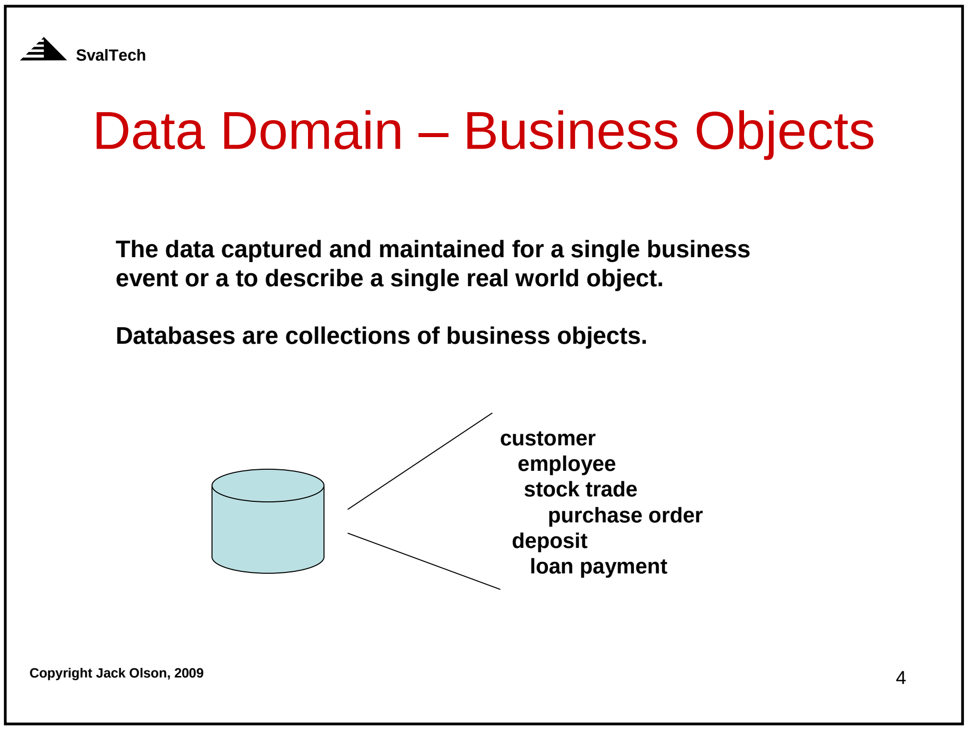
Data Domain for the Presentation

Data created and maintained by either custom applications or packaged applications that store data in structured database management systems or structured records in file systems.
transaction data
reference data

DB2
SAP
IMS
Oracle Financials
ADABAS
Siebel
IDMS
PeopleSoft
ORACLE
SYBASE
VSAM

Copyright Jack Olson, 2009

3



SvalTech

Data Domain – Business Objects

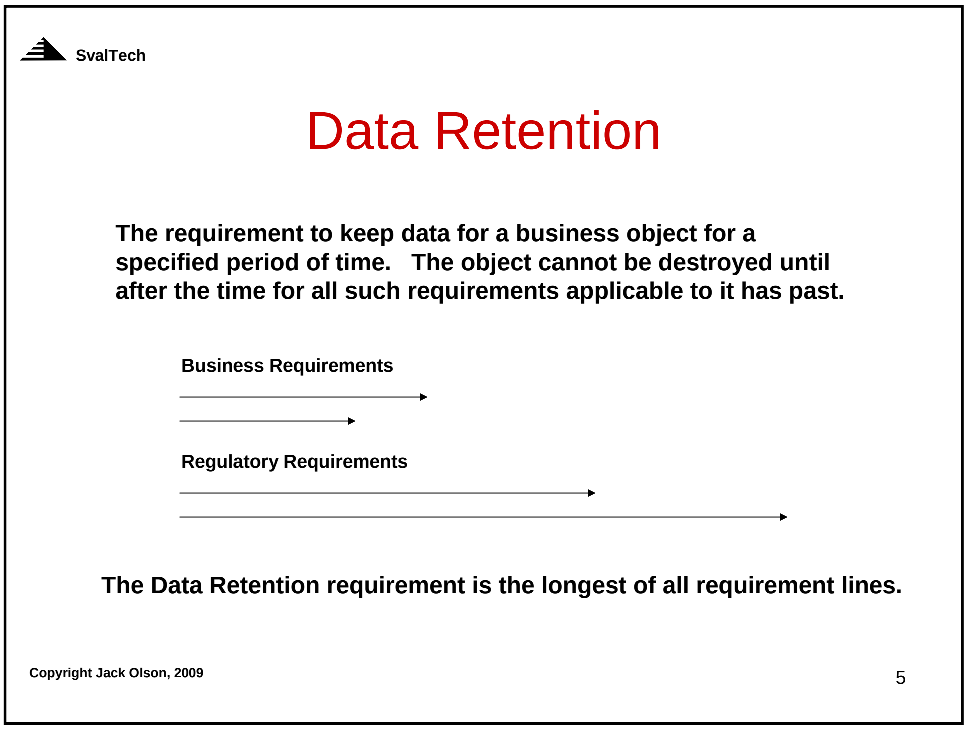
The data captured and maintained for a single business event or a to describe a single real world object.

Databases are collections of business objects.

customer
employee
stock trade
purchase order
deposit
loan payment

Copyright Jack Olson, 2009

4



SvalTech

Data Retention

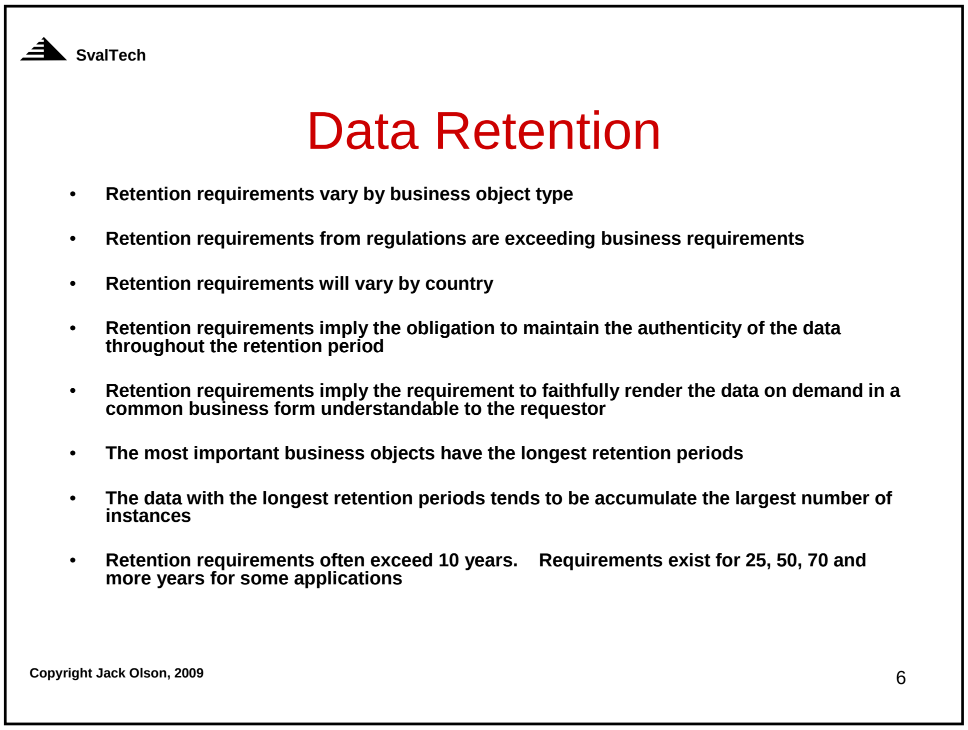
The requirement to keep data for a business object for a specified period of time. The object cannot be destroyed until after the time for all such requirements applicable to it has past.

Business Requirements

Regulatory Requirements

The Data Retention requirement is the longest of all requirement lines.

5



SvalTech

Data Retention

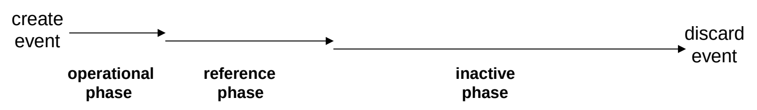
- **Retention requirements vary by business object type**
- **Retention requirements from regulations are exceeding business requirements**
- **Retention requirements will vary by country**
- **Retention requirements imply the obligation to maintain the authenticity of the data throughout the retention period**
- **Retention requirements imply the requirement to faithfully render the data on demand in a common business form understandable to the requestor**
- **The most important business objects have the longest retention periods**
- **The data with the longest retention periods tends to be accumulate the largest number of instances**
- **Retention requirements often exceed 10 years. Requirements exist for 25, 50, 70 and more years for some applications**

6

SvalTech

Data Time Lines

for a single instance of a data object



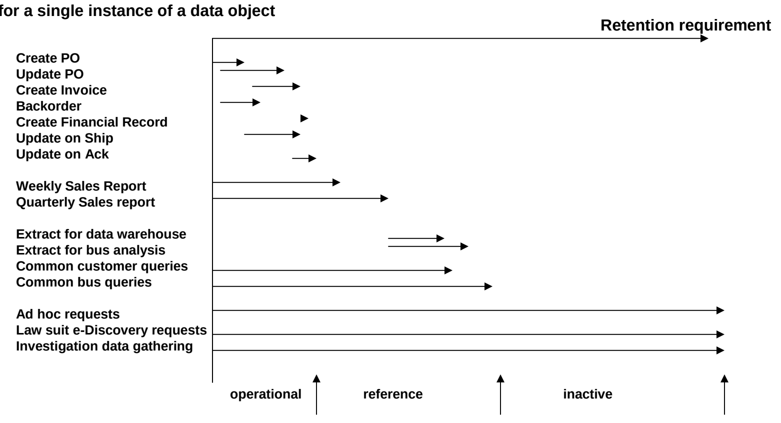
operational phase	can be updated, can be deleted, may participate in processes that create or update other data
reference phase	used for business reporting, extracted into business intelligence or analytic databases, anticipated queries
inactive phase	no expectation of being used again, no known business value, being retained solely for the purpose of satisfying retention requirements. Must be available on request in the rare event a need arises.

7

SvalTech

Data Process Time Lines

for a single instance of a data object



8

SvalTech

Some Observations

- **Some objects exit the operational phase almost immediately (financial records)**
- **Some objects never exit the operational phase (customer name and address)**
- **Most transaction data has an operational phase of less than 10% of the retention requirement and a reference phase of less than 20% of the retention requirement**
- **Inactive data generally does not require access to application programs: only access to ad hoc search and extract tools**

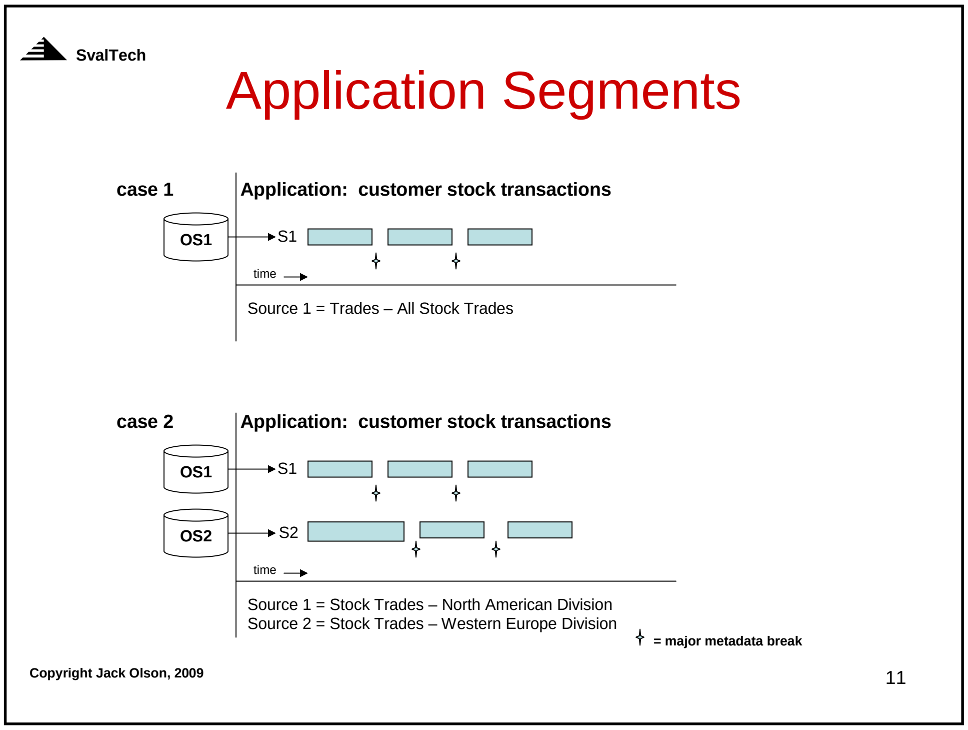
SvalTech

Application Segments

An application segment is a set of business objects generated from a single version of an application where all objects in the segment have data consistent with a single metadata definition.

A metadata break is a point in the life of the operational database where a change in metadata is implemented that changes the structure of the data or the manner in which data is encoded.

- An application will have many segments over time
- Minor changes in metadata can sometimes be implemented without forcing a segment change
- Major metadata changes will always generate a segment change where data created in the previous segment cannot be recast to the new metadata definition without some compromise in the data
- Application segments can be generated in parallel with one operational implementation using one version of the application at the same time that another operational instance is using a different version of the application



SvalTech

Application Segments

case 1
Application: customer stock transactions
OS1
S1
time

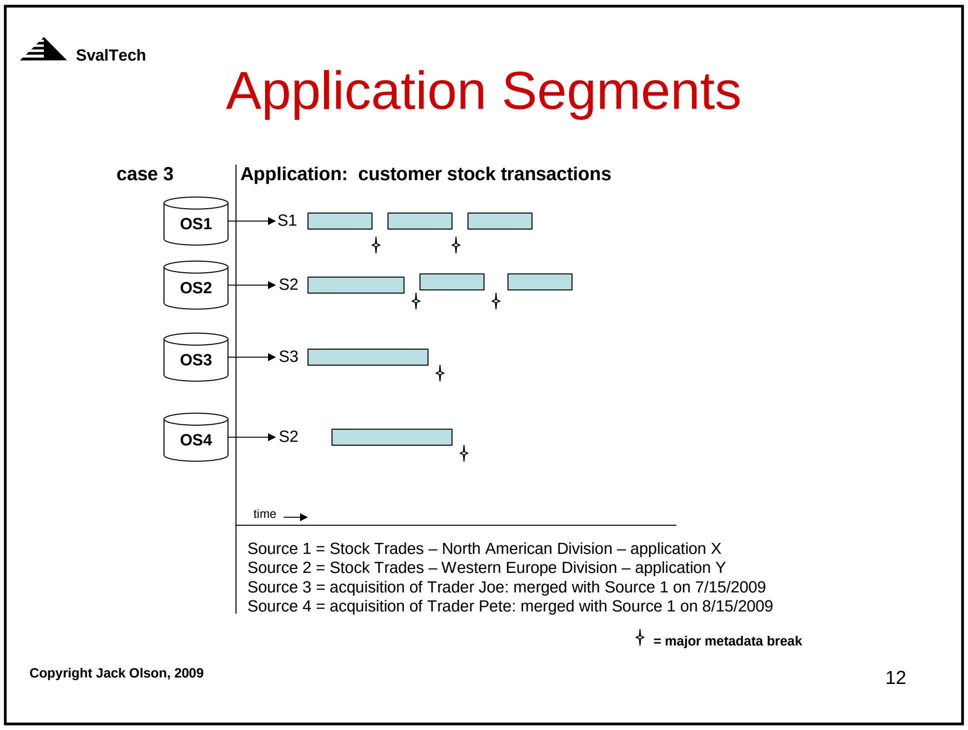
Source 1 = Trades – All Stock Trades

case 2
Application: customer stock transactions
OS1
S1
OS2
S2
time

Source 1 = Stock Trades – North American Division
Source 2 = Stock Trades – Western Europe Division

⟡ = major metadata break

11



SvalTech

Application Segments

Source 1 = Stock Trades – North American Division – application X
Source 2 = Stock Trades – Western Europe Division – application Y
Source 3 = acquisition of Trader Joe: merged with Source 1 on 7/15/2009
Source 4 = acquisition of Trader Pete: merged with Source 1 on 8/15/2009

⟡ = major metadata break

12

SvalTech

Problems with Using a Single Operational Database Approach

Single Operational Database Approach:

Keeping business object data in the single active operational database until the retention period expires and then deleting it.

Operational database contains business objects in all 3 phases

Single objects may exist in the database for decades

This is the most common method for handling long term data retention requirements

13

SvalTech

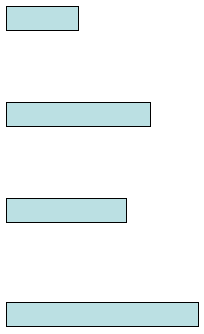
Problem 1: Loss of Clarity of Understanding

Database Structure
- **DDL**
- **DBD/PSB**

External Metadata
- **formal metadata repository**
- **auxiliary metadata repository**
- **copybooks,**

Application externalizations
- **display windows**
- **reports**

Knowledge in Employee Heads
- **IT workers**
- **business unit workers**



How much do you depend on each of these areas for interpreting data that you see?

How accurate are each of these?

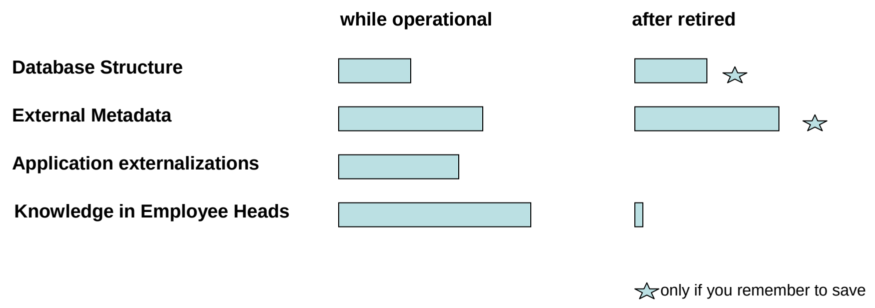
How complete are each of these?

14

SvalTech

Problem 1: Loss of Clarity of Understanding

While still operational, clarity is maintained.
When application is retired, clarity begins to erode: rapidly



Copyright Jack Olson, 2009

15

SvalTech

Problem 1: Loss of Clarity of Understanding Avoidance

- **Capture Metadata**
  - Formal metadata repositories
  - Original database architecture design
  - Application program screens/ reports
  - Extract product repositories
- **Validate Metadata**
  - Match to real data
  - Have verification review meetings
- **Enhance Metadata**
  - Logical information
  - Change information
- **Build access routines outside of application**
  - Reports
  - Displays
- **Validate accessibility on annual basis**
  - Select a probable user
  - Have them access retired data through access routines

Copyright Jack Olson, 2009

16

SvalTech

Problem 2: Metadata Change Corruption

The problem with metadata changes is that

the DBMS only supports one version of data definition

which means that old data must be manipulated to conform to the new definition

which often results in data elements being missing or inconsistent

a future user of the data does not know which instances are good and which are not.

When the scope of data in a DBMS covers a short time period the corruption may be acceptable.

The cumulative effect of change corruption over many years can render old data instances highly inaccurate and misleading.

17

SvalTech

Problem 2: Metadata Change Corruption

Example 1:

Add a column to an existing table. All old rows have value "NULL" inserted for this column. (or worse yet, a single default value that is NOT NULL).

ALTER TABLE PERSONNEL ADD COLUMN MILITARY_SERVICE CHARACTER 10

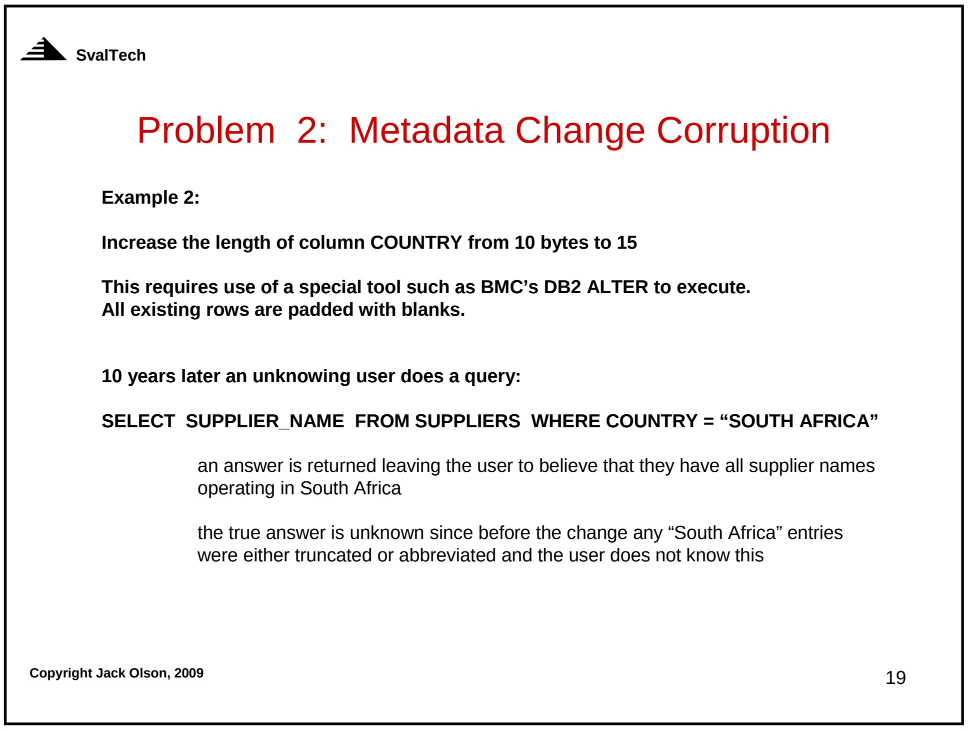
10 years later an unknowing user does a query:

SELECT NAME FROM PERSONNEL WHERE MILITARY_SERVICE = "NAVY"

an answer is returned leaving the user to believe that they have everyone who served in the navy.

the true answer is unknown

18



SvalTech

Problem 2: Metadata Change Corruption

Example 2:

Increase the length of column COUNTRY from 10 bytes to 15

This requires use of a special tool such as BMC's DB2 ALTER to execute.
All existing rows are padded with blanks.

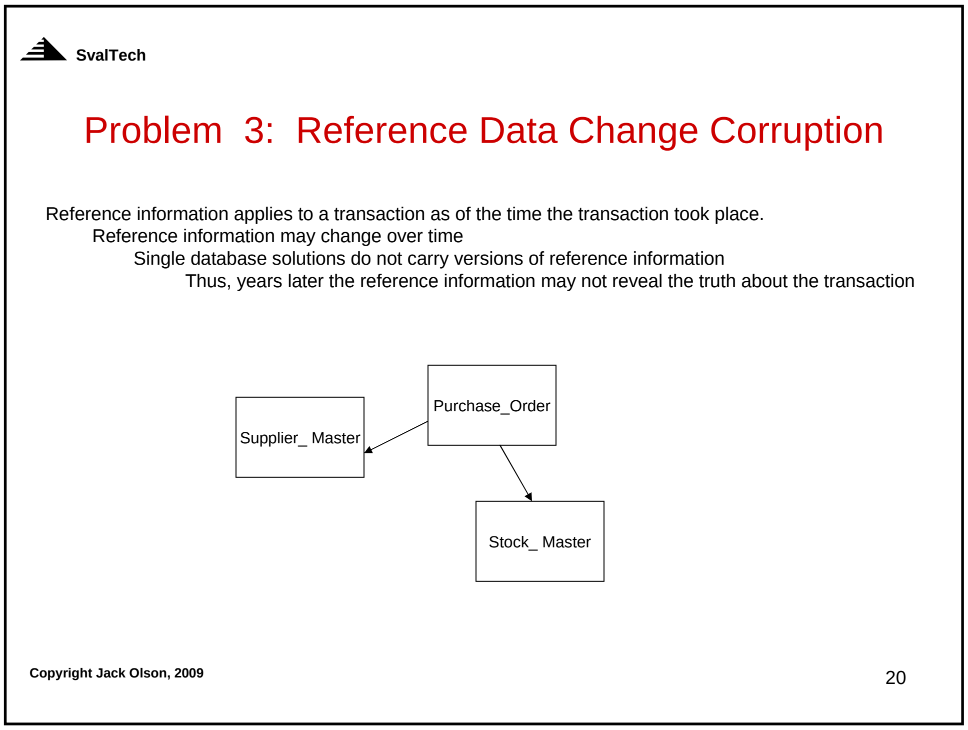
10 years later an unknowing user does a query:

SELECT SUPPLIER_NAME FROM SUPPLIERS WHERE COUNTRY = "SOUTH AFRICA"

an answer is returned leaving the user to believe that they have all supplier names operating in South Africa

the true answer is unknown since before the change any "South Africa" entries were either truncated or abbreviated and the user does not know this

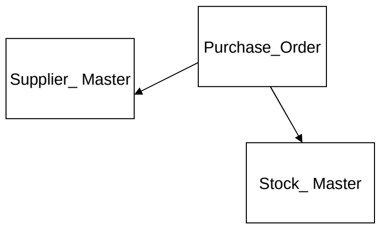
19



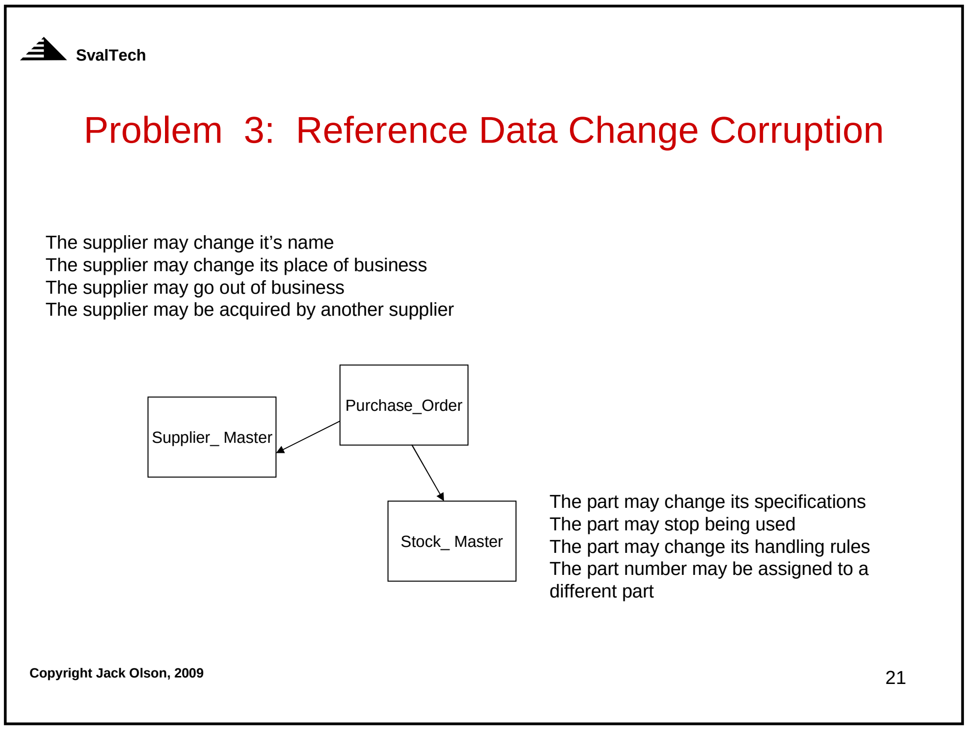
SvalTech

Problem 3: Reference Data Change Corruption

Reference information applies to a transaction as of the time the transaction took place.
Reference information may change over time
Single database solutions do not carry versions of reference information
Thus, years later the reference information may not reveal the truth about the transaction



20



SvalTech

Problem 3: Reference Data Change Corruption

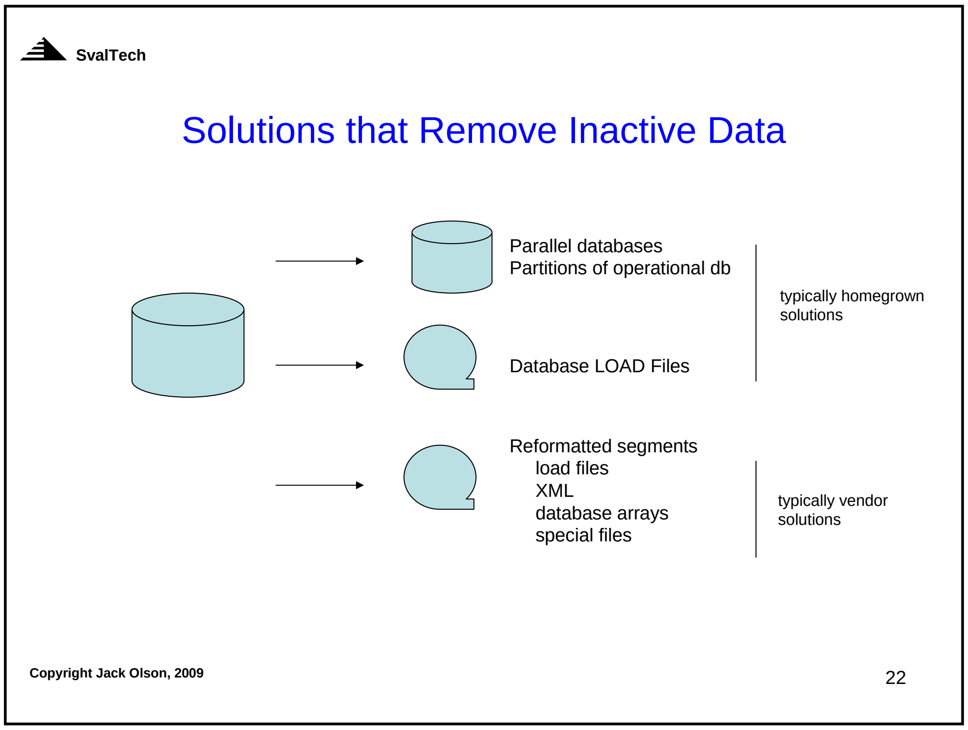
The supplier may change it's name
The supplier may change its place of business
The supplier may go out of business
The supplier may be acquired by another supplier

Purchase_Order
Supplier_ Master
Stock_ Master

The part may change its specifications
The part may stop being used
The part may change its handling rules
The part number may be assigned to a different part

Copyright Jack Olson, 2009

21



SvalTech

Solutions that Remove Inactive Data

Parallel databases
Partitions of operational db

Database LOAD Files

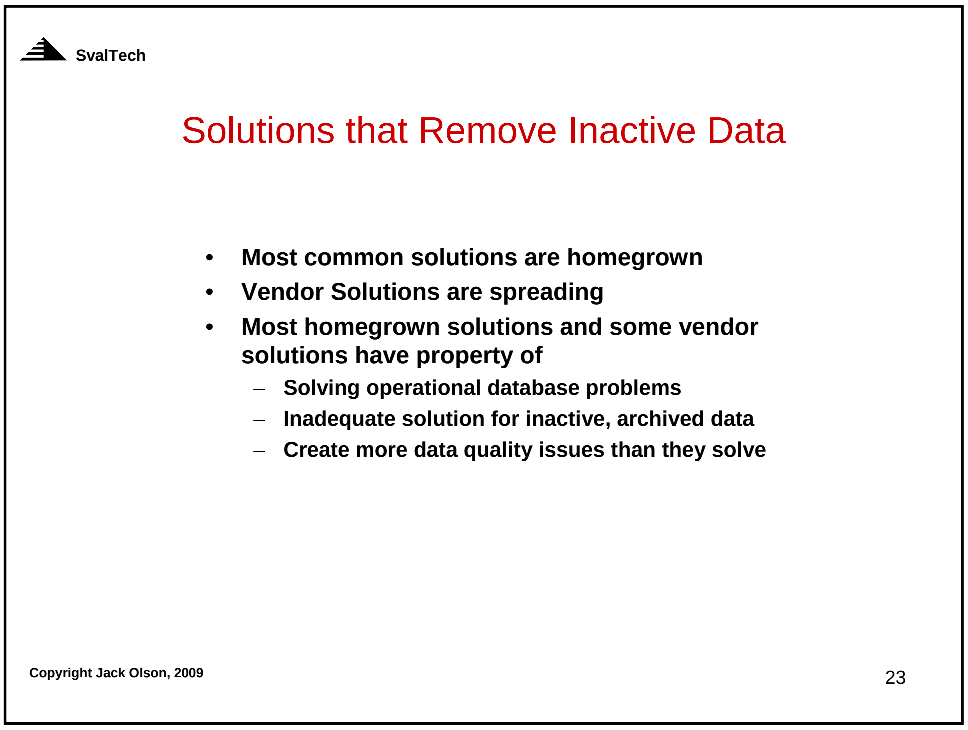
typically homegrown solutions

Reformatted segments
- load files
- XML
- database arrays
- special files

typically vendor solutions

Copyright Jack Olson, 2009

22



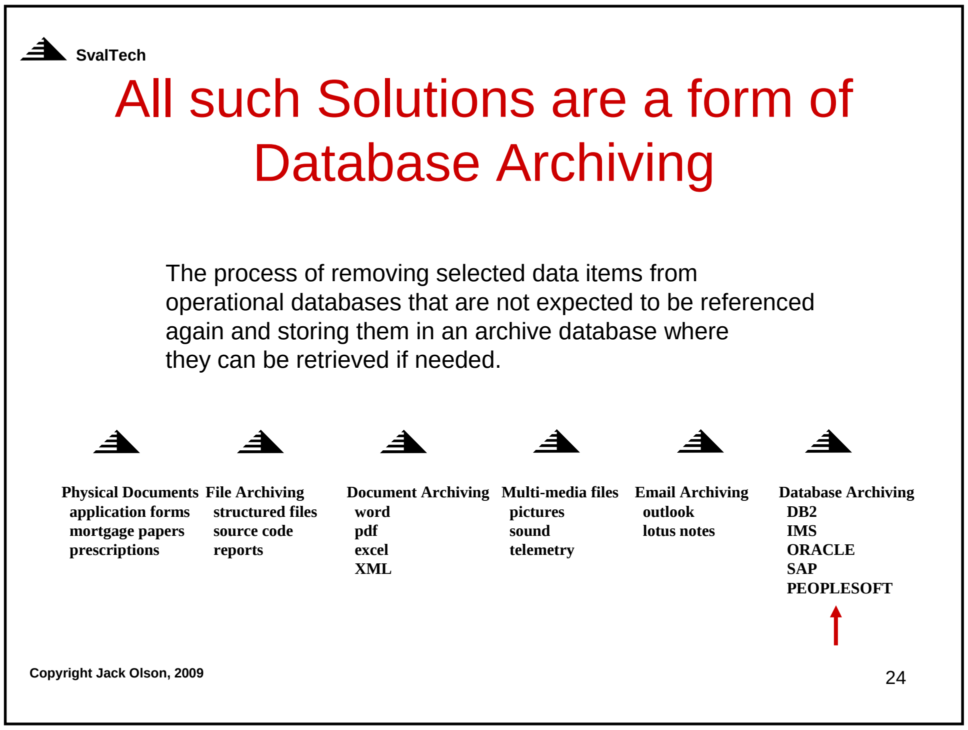
SvalTech

Solutions that Remove Inactive Data

- **Most common solutions are homegrown**
- **Vendor Solutions are spreading**
- **Most homegrown solutions and some vendor solutions have property of**
  - **Solving operational database problems**
  - **Inadequate solution for inactive, archived data**
  - **Create more data quality issues than they solve**

Copyright Jack Olson, 2009

23



SvalTech

All such Solutions are a form of Database Archiving

The process of removing selected data items from operational databases that are not expected to be referenced again and storing them in an archive database where they can be retrieved if needed.

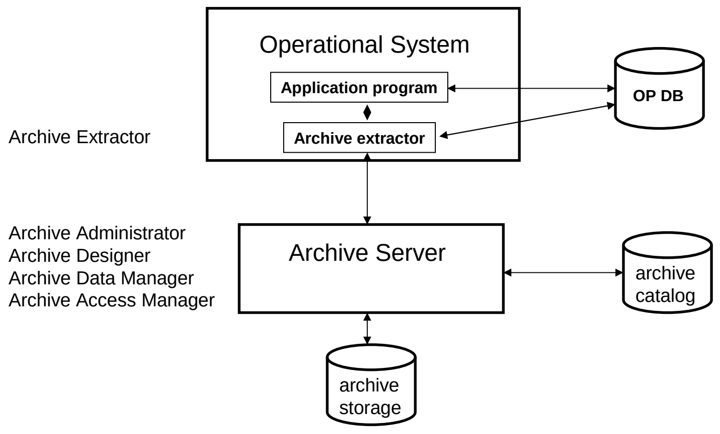
Physical Documents	**File Archiving**	**Document Archiving**	**Multi-media files**	**Email Archiving**	**Database Archiving**
application forms	**structured files**	**word**	**pictures**	**outlook**	**DB2**
mortgage papers	**source code**	**pdf**	**sound**	**lotus notes**	**IMS**
prescriptions	**reports**	**excel**	**telemetry**		**ORACLE**
		XML			**SAP**
					PEOPLESOFT

Copyright Jack Olson, 2009

24

SvalTech

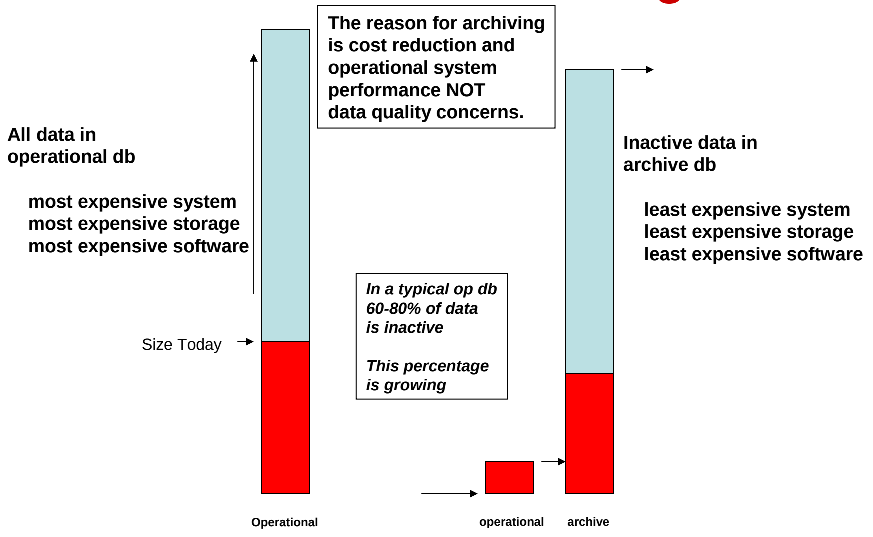
Architecture of Database Archiving



25

SvalTech

Reason for Archiving

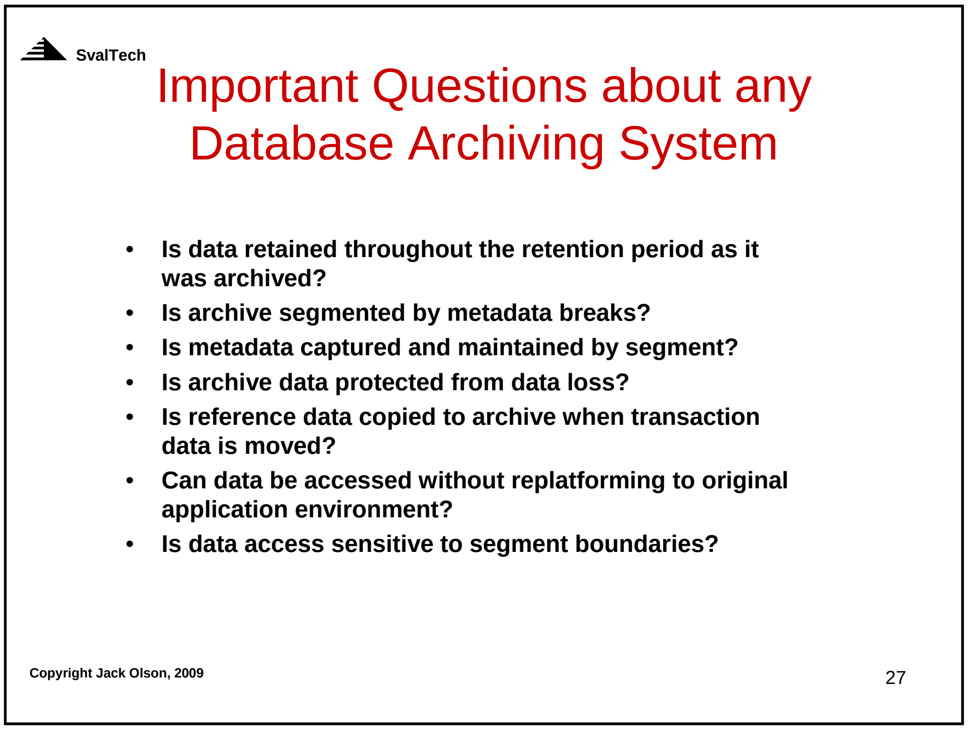


The reason for archiving is cost reduction and operational system performance NOT data quality concerns.

In a typical op db 60-80% of data is inactive

This percentage is growing

26

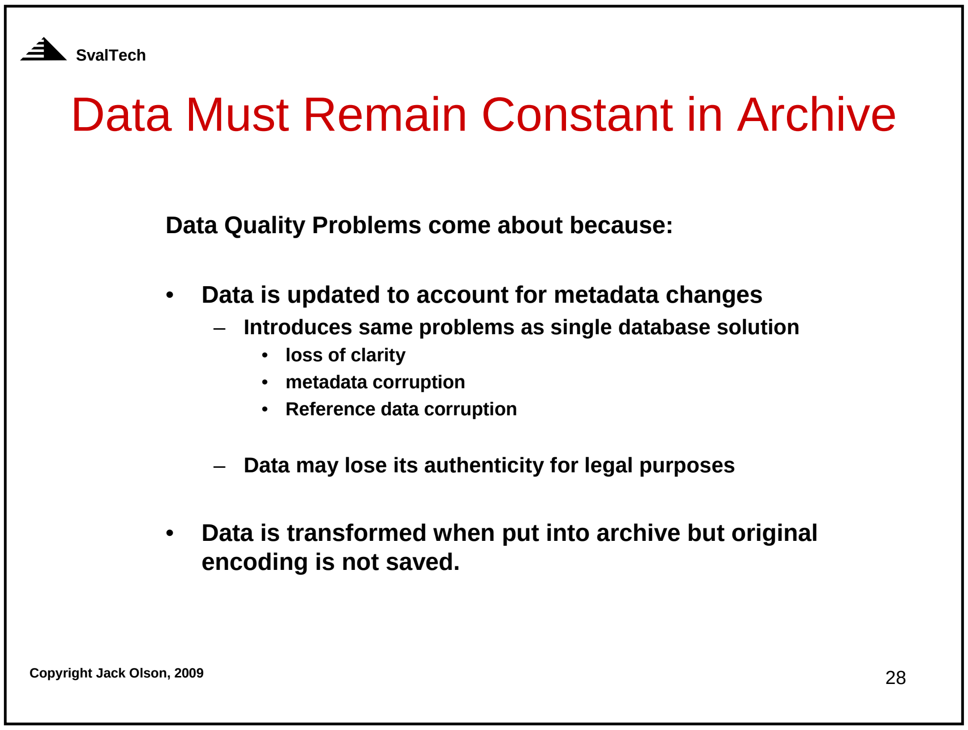


SvalTech

Important Questions about any Database Archiving System

- **Is data retained throughout the retention period as it was archived?**
- **Is archive segmented by metadata breaks?**
- **Is metadata captured and maintained by segment?**
- **Is archive data protected from data loss?**
- **Is reference data copied to archive when transaction data is moved?**
- **Can data be accessed without replatforming to original application environment?**
- **Is data access sensitive to segment boundaries?**

27



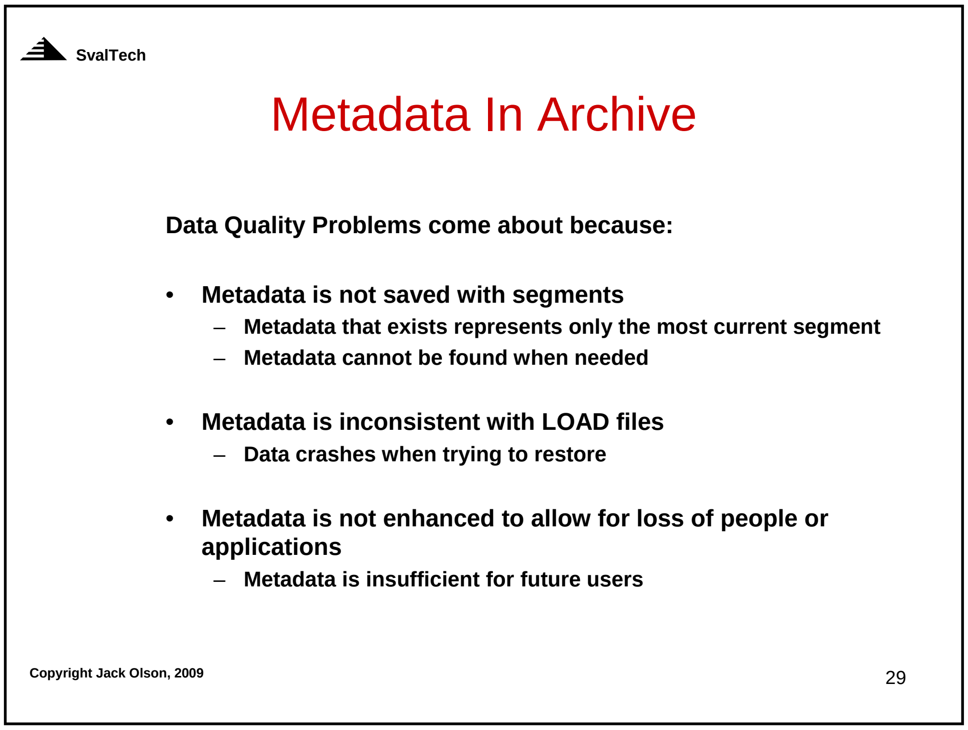
SvalTech

Data Must Remain Constant in Archive

Data Quality Problems come about because:

- **Data is updated to account for metadata changes**
  - **Introduces same problems as single database solution**
    - **loss of clarity**
    - **metadata corruption**
    - **Reference data corruption**
  - **Data may lose its authenticity for legal purposes**
- **Data is transformed when put into archive but original encoding is not saved.**

28



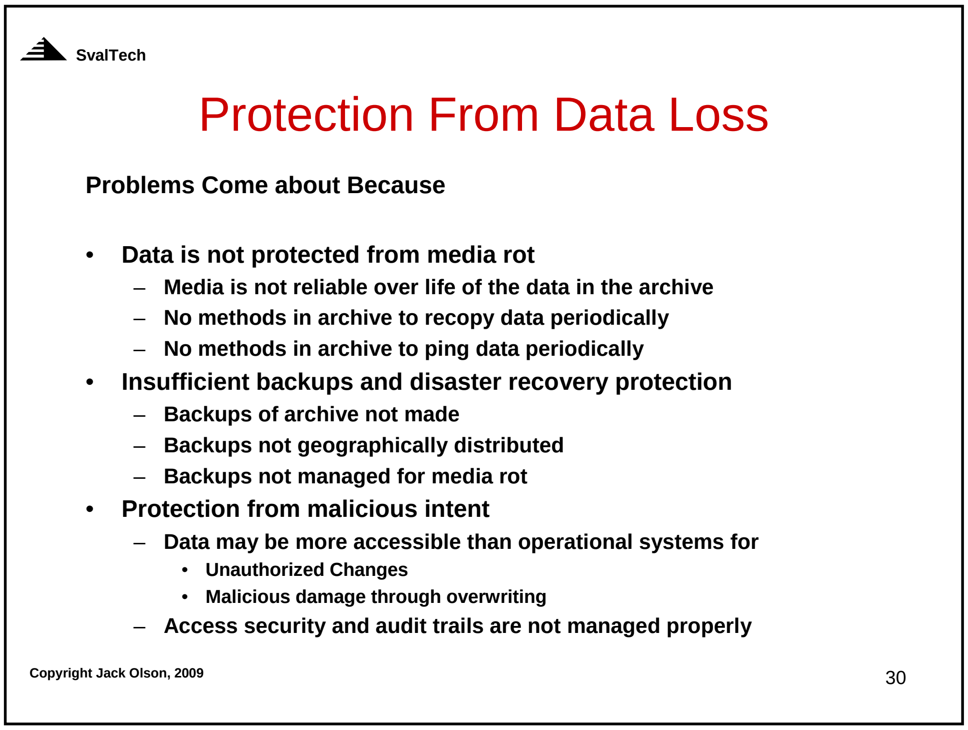
SvalTech

Metadata In Archive

Data Quality Problems come about because:

- **Metadata is not saved with segments**
  - **Metadata that exists represents only the most current segment**
  - **Metadata cannot be found when needed**
- **Metadata is inconsistent with LOAD files**
  - **Data crashes when trying to restore**
- **Metadata is not enhanced to allow for loss of people or applications**
  - **Metadata is insufficient for future users**

29



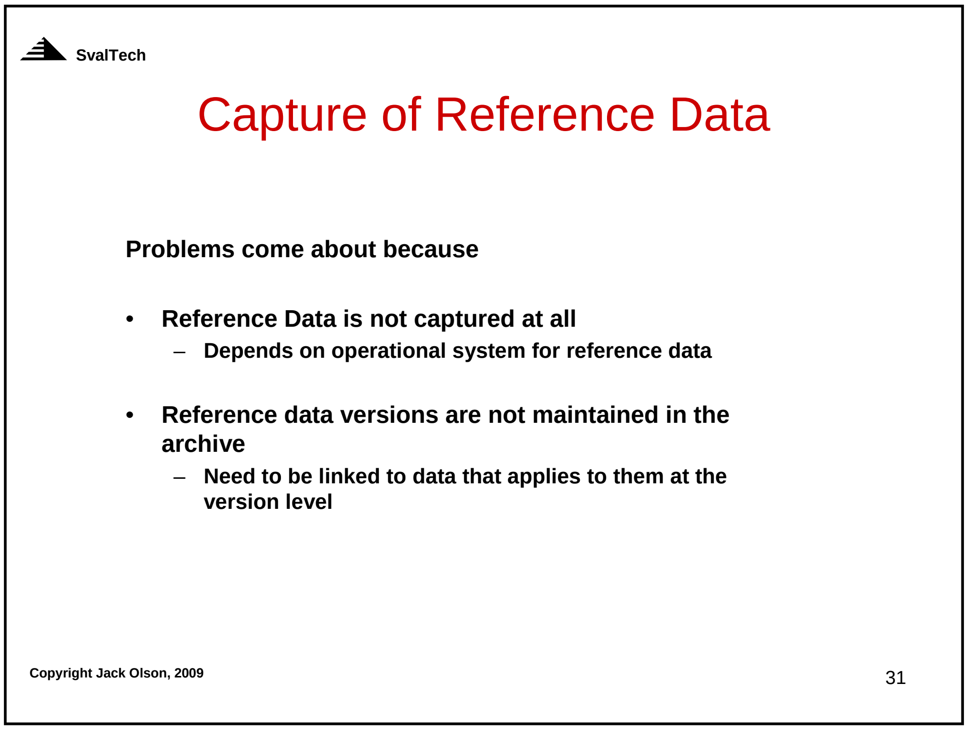
SvalTech

Protection From Data Loss

Problems Come about Because

- **Data is not protected from media rot**
  - **Media is not reliable over life of the data in the archive**
  - **No methods in archive to recopy data periodically**
  - **No methods in archive to ping data periodically**
- **Insufficient backups and disaster recovery protection**
  - **Backups of archive not made**
  - **Backups not geographically distributed**
  - **Backups not managed for media rot**
- **Protection from malicious intent**
  - **Data may be more accessible than operational systems for**
    - **Unauthorized Changes**
    - **Malicious damage through overwriting**
  - **Access security and audit trails are not managed properly**

30



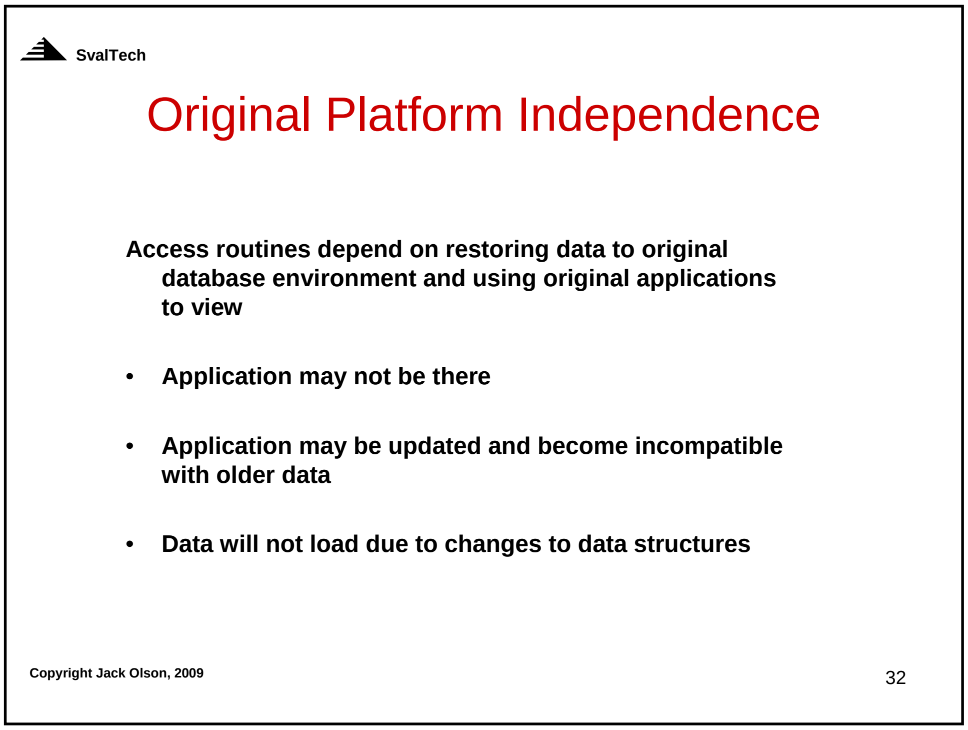
SvalTech

Capture of Reference Data

Problems come about because

- Reference Data is not captured at all
  - Depends on operational system for reference data
- Reference data versions are not maintained in the archive
  - Need to be linked to data that applies to them at the version level

31



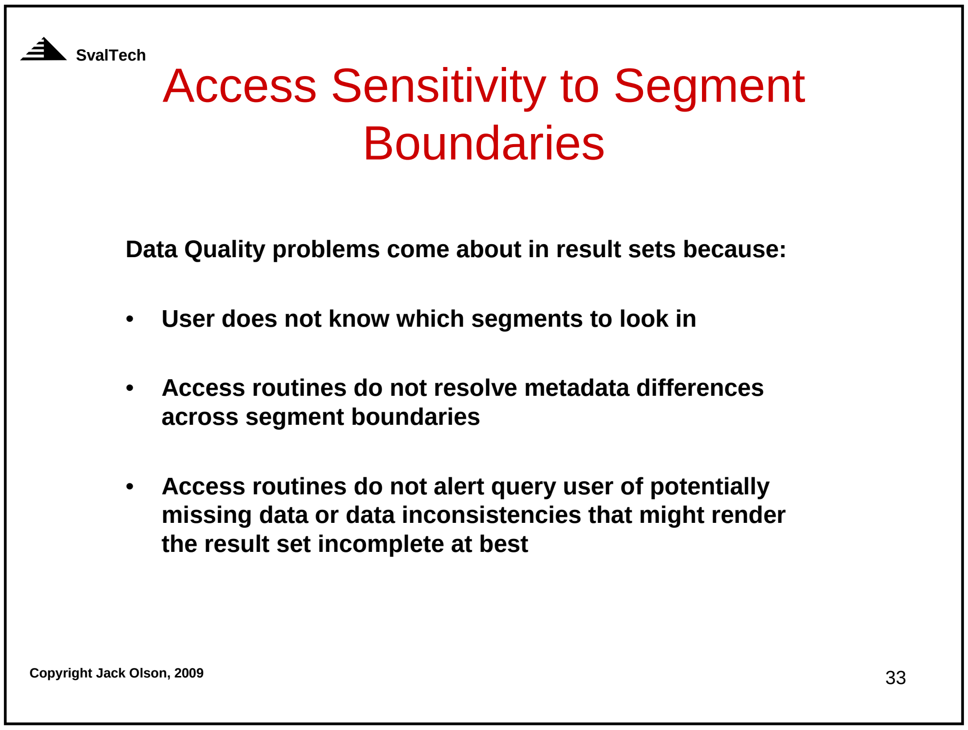
SvalTech

Original Platform Independence

Access routines depend on restoring data to original database environment and using original applications to view

- Application may not be there
- Application may be updated and become incompatible with older data
- Data will not load due to changes to data structures

32



SvalTech

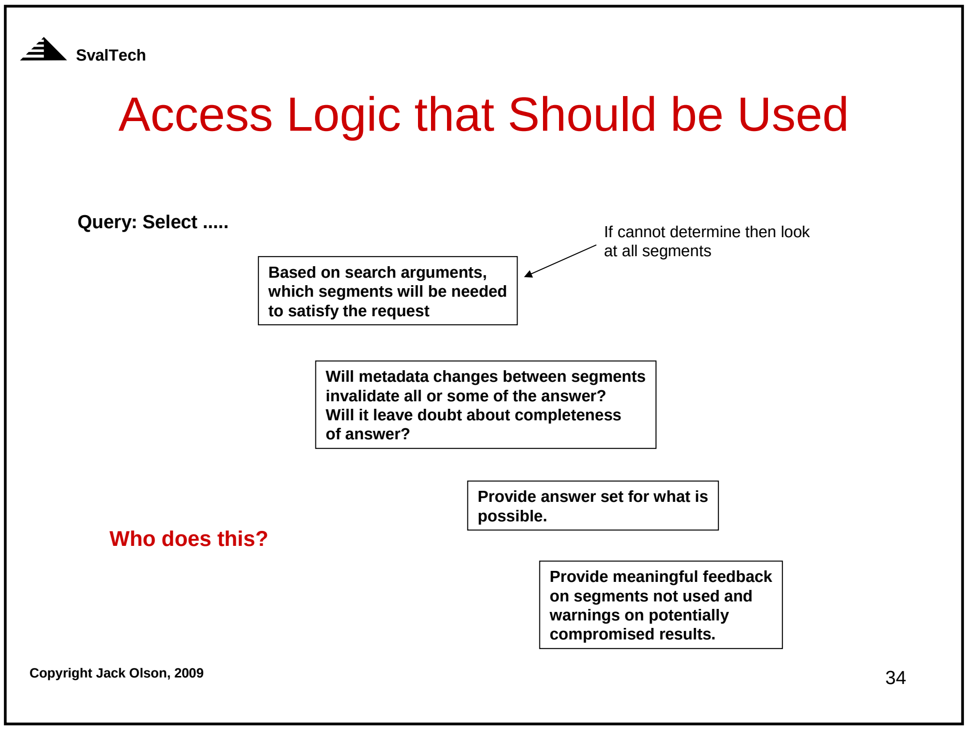
Access Sensitivity to Segment Boundaries

Data Quality problems come about in result sets because:

- **User does not know which segments to look in**
- **Access routines do not resolve metadata differences across segment boundaries**
- **Access routines do not alert query user of potentially missing data or data inconsistencies that might render the result set incomplete at best**

Copyright Jack Olson, 2009

33



SvalTech

Access Logic that Should be Used

Query: Select .....

Based on search arguments, which segments will be needed to satisfy the request

If cannot determine then look at all segments

Will metadata changes between segments invalidate all or some of the answer? Will it leave doubt about completeness of answer?

Provide answer set for what is possible.

Provide meaningful feedback on segments not used and warnings on potentially compromised results.

Who does this?

Copyright Jack Olson, 2009

34

SvalTech

Final Thought

Failure to address long term data quality erosion issues can lead to archived data being lost, rendered unusable, or meaningless.

A poorly designed strategy can appear to be working smoothly for years while data quality is eroding daily.

When the need for the data arises the consequences of bad design can be costly and an embarrassment to the corporation.

Good design needs to encapsulate application initial design, design change rules, data archiving processes, and on-going data management oversight.

35

SvalTech

Some Quotes

"When we go back to the archived data we cross our fingers.   So far we have been able to get what we want but it is always a major effort.   Someday it won't work."

"God help us if we have to use that data again."

In answer to the question of where do you store your archived data: "In black plastic garbage bags."

36