# Data Quality and Database Archiving: The Intersection of Two Important Data Management Functions

## ABSTRACT

This presentation shows that when database archiving technology is employed for large database applications that have long data retention periods, the data quality is preserved. It includes a short tutorial on the basics of database archiving. It shows how keeping data in operational systems for long periods of time creates many opportunities for the data quality to erode. It concludes with a detailed explanation of why a robust database archiving implementation prevents erosion from occurring and thus preserves the original quality for all time.

## BIOGRAPHY

**Jack Olson**
Chief Executive Officer
SvalTech, Inc.

Jack Olson has spent 40 years developing of systems software with a specialty in DBMS and Database tool technologies. 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. Jack is currently CEO of SvalTech, Inc., a company dedicated to the technology of Database Archiving. Jack has published two books: "Data Quality: the Accuracy Dimension", 2003 and "Database Archiving: How to Keep Lots of Data for a Very Long Time", 2009. Jack has a BS degree in Mathematics from the Illinois Institute of Technology and an MBA from Northwestern University.

2011 MIT Information Quality Industry Symposium

# Data Quality and Database Archiving

**The intersection of two important Data Management Functions**

**"Database Archiving: How to Keep Lots of Data for a Long Time"**
**Jack E. Olson, Elsevier, 2009**

Jack E. Olson
jack.olson@SvalTech.com



1

SvalTech

## Presentation Roadmap

**Database Archiving Basics**

**Data Quality Problems With Single, Operational Database Approach**
- **Long term loss of clarity of understanding**
- **Metadata change corruption**
- **Reference data changes**
- **Database Consolidation (mergers and acquisitions)**

**Using Database Archiving for Improved Data Quality**
- **Education and Awareness**
- **Early Business Records Capture**
- **Managing Data and Metadata within Application Segments**
- **Capture Extended Metadata (become application independent)**
- **Freeze Reference Data**
- **Metadata Change Sensitive Data Access**

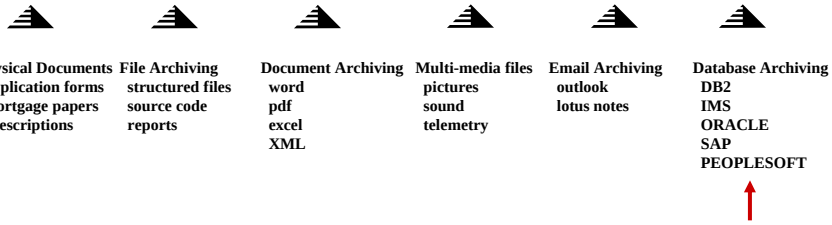


2

SvalTech

# Database Archiving

The process of removing selected data records 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.





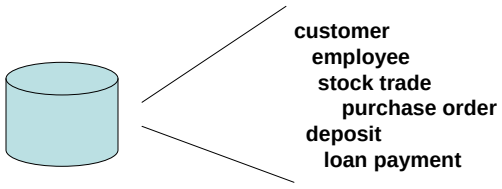



SvalTech

# Business Records

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

**Databases are collections of business records.**

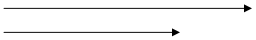


SvalTech

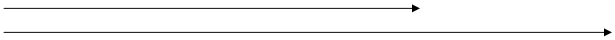
# Data Retention

**The requirement to keep data for a business record for a specified period of time. The record 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 generally exceed business requirements**
- **Retention requirements 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 tend to have the longest retention periods**
- **The data with the longest retention periods tend to have the largest number of instances**
- **Retention requirements often exceed 10 years. Requirements exist for 25, 50, 70 and more years for some applications**

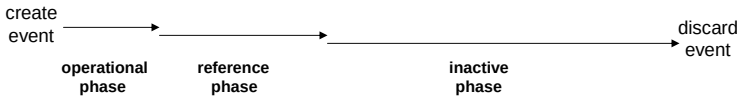


6

SvalTech

# Data Time Lines

**for a single instance of a business record**




| | |
|---|---|
| **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.** |

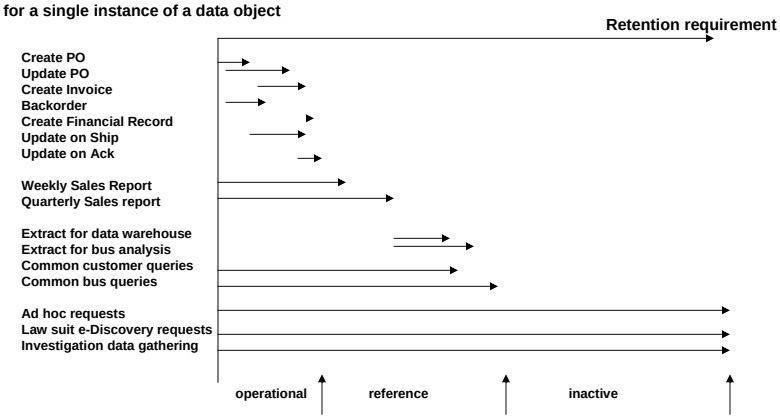


7

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

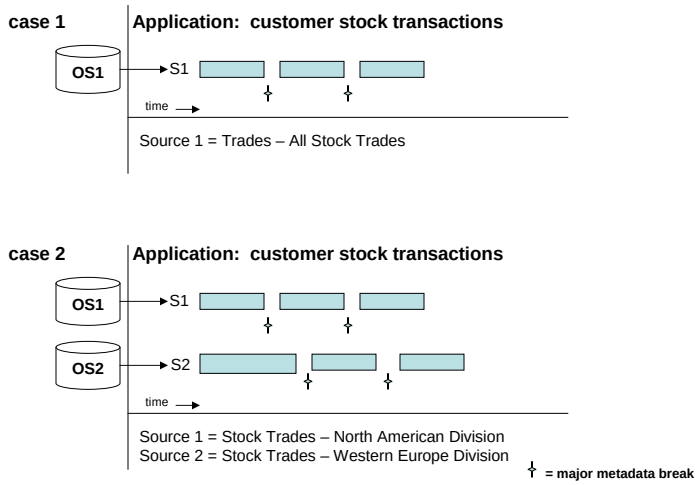
**An application segment is a set of business records generated from a single version of an application where all records in the segment have data consistent with a single metadata definition.**

**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

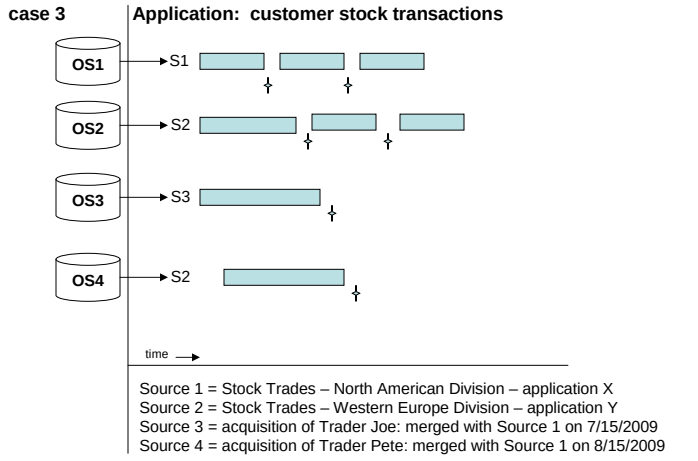





11

SvalTech

# Application Segments





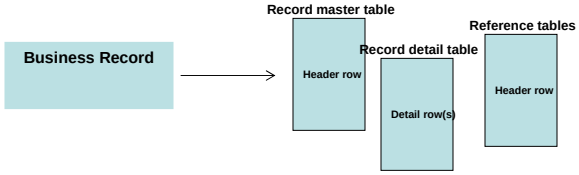

12

SvalTech

# Business Record Dependencies

**An electronic business record (stored in a formal database) has many objects it depends on for finding, viewing, and interpreting the data stored for the record.**

**These are called Business Record Dependencies.**




**Business Record Dependencies:**
- **Hardware**
- **Operating System**
- **DBMS (type and version)**
- **application data model**
- **application metadata**
- **application programs**
- **reference tables (master data)**
- **subject matter experts**



13

SvalTech

# Database Archiving Goals

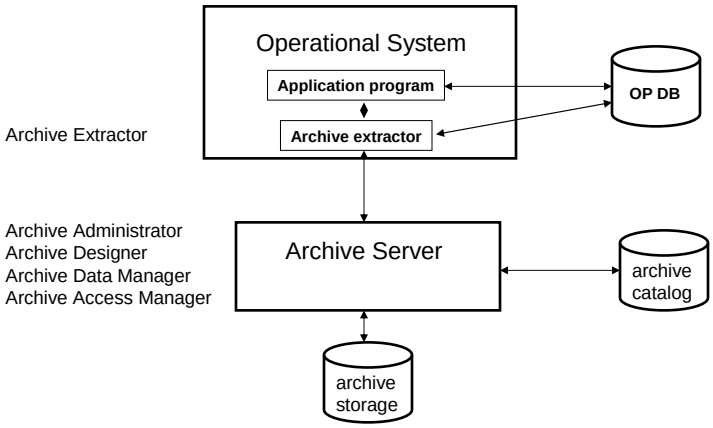
- **Eliminate as many business record dependencies as possible**
- **Preserve data as it appeared when first created**
- **Convert data to a storage form that is more suitable for long term data retention**
- **Prevent ANY updates to the data once archived**
- **Restrict access to archived data to minimum number of people as possible**
- **Maintain audit records on all access to archived data**



14

SvalTech

# Architecture of Database Archiving






15

SvalTech

# Database Archive Logical Structure

Application

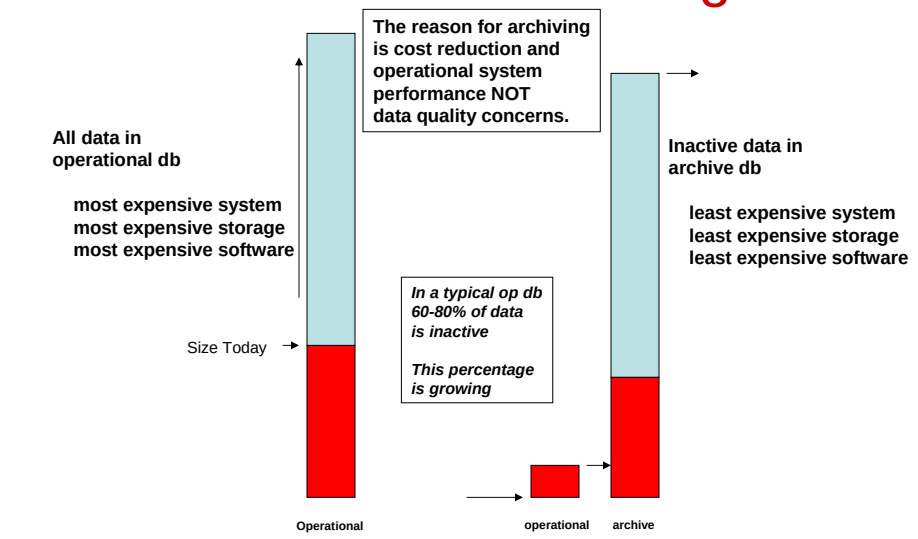
- **Application Segment 1**
  - **Source data description**
    **Archive metadata**
    **Access audit**
  - **Archive event 1**
    - **Business Record Data**
      **Reference Data**
      **Indexes audit**
  - **Archive event 2**
- **Application Segment 2**
- **Application Segment 3**



16

SvalTech

# Reason for Archiving






17

SvalTech

# Data Quality Issues

**The longer you keep business records in the operational database the greater the risk of degradation of the quality of the data.**

**This is due to changes in the Business Record Dependencies.**

**If none of the Business Record Dependencies change for the life of the business record then there will be no degradation.**

**What are the chances?????**



18

SvalTech

# 1: Platform Changes

Hardware
Operating Systems
DBMS

- **Usually involve other changes at the same time**
  - Application programs
  - Data Structure Changes
- **Platform issues for data usually involve transformations**
  - Data encoding pages
  - Limits on field sizes, numbers
  - Encoding of date/time
- **Can also include latent data quality issues covered up by older systems**
  - Lack of enforcement of unique keys
  - Lack of data type enforcement
  - Lack of NULL indicator support



19

SvalTech

# 1: Platform Changes

- **Single Operational Database Requirement**
  - Convert data to replacement infrastructure
    - Convert structure definitions as required
    - Perform data transformations as best as can
    - Resolve exposed quality issues as best as can
- **Use of Database Archiving**
  - Produce application archive segments for data on older systems
  - Transform structure and data minimally to archive platform
  - Audit all data errors found in archiving



20

SvalTech

## 2: Data Structure & Metadata Changes

**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.**





SvalTech

## 2: Metadata Changes

**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

SvalTech

## 2: Metadata Changes

**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





SvalTech

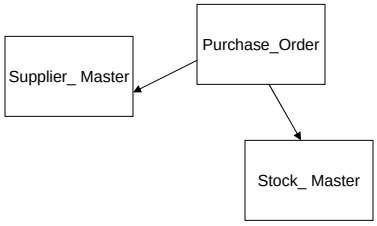
## 2: Metadata Changes

- **Single Operational Database Requirement**
  - **Convert data to match new definitions**
    - **Make up values for new columns**
    - **Truncate columns or extend with blanks**
    - **Use NULL on columns inappropriately**
- **Use of Database Archiving**
  - **Create a Metadata Break and start a new application archive segments**
  - **Data created on old definitions stored with metadata in old segment**
  - **New data created after the change stored with metadata in new segment**

SvalTech

## 3: Reference Data Changes

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



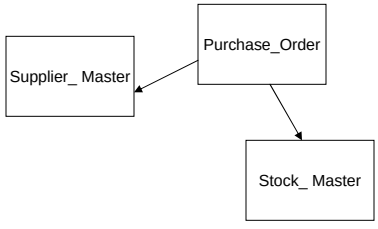



25

SvalTech

## 3: Reference Data Changes

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




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



26

SvalTech

## 3: Reference Data Changes

- **Single Operational Database Requirement**
  - **Old Business Records always connect to current definition of reference data instead of the reference data that existed at the time of create**
  - **The original reference data is not maintained and hence not retrievable**
- **Use of Database Archiving**
  - **Each archive event encapsulates the reference data into the event package**
  - **The business record is viewed with reference data as it existed at the time of archive**



27

SvalTech

## 4: Application Program Changes

incremental improvements
application renovation
application replacement

**Much of a user's interpretation of data is achieved through the application program forms, reports, screen displays, and screen prompts**

**When an application program changes, old data may not yield accurate results due to data structure changes or changes in the way a column is encoded**

**This is generally handled through converting old data to be structurally compatible with the new data definitions.**
**columns are populated with default values**
**column values are changed**

**Sometimes the old data can not be changed to be accurate when used with the new application**

**When an application is discontinued (retired) in favor of a newer application, the old data still retains a dependency on the old application, which retains a dependency on the old system. Management is anxious to get rid of the old costs and thus often does so.**



28

SvalTech

## 4: Application Program Changes

- **Single Operational Database Requirement**
  - Data is often converted to form of new data structures with attendant problems
  - Data is kept separate with old versions of application also retained
  - Data is kept separate but access is restricted through direct SQL only
- **Use of Database Archiving**
  - The switch t the new application version or new application is treated as a metadata break.
  - Each archive segment contains its own metadata and data in original form



29

SvalTech

## 5: Subject Matter Expert Changes

Subject Matter Experts tend to exist only when the application is current. They are knowledgeable about the data structures and rules as they currently exist as opposed to what they may have meant in the past.

When an application is retired, the SME's will disappear quickly. This removes their knowledge from supporting the retired data.



30

SvalTech

## 5: Subject Matter Expert Changes

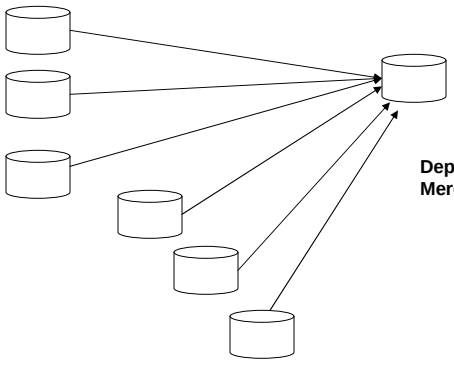
- **Single Operational Database Requirement**
  - **Maintain SME on all current and past versions of the application**
  - **Maintain SME for all retired applications**
- **Use of Database Archiving**
  - **Having complete metadata with original data separated into application archive segments reduces dependencies on SMEs.**



31

SvalTech

## 6: Getting it All at One Time



**Departmental System Consolidation**
**Merger and Acquisition Consolidation**



32

SvalTech

## Summary of Database Archiving Benefits for Maintaining Data Quality

- **Captures business record and reference data at time it becomes inactive**
  - **Data Never Changes when in Archive**
  - **All Access return same values no matter when in life-cycle**
- **Places data in an environment where it is independent of original application environment**
- **Avoids taking actions that will corrupt the quality of data**





SvalTech

## Final Thoughts

**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.**