The Root Cause of Unstructured Data Problems is Not What You Think

Bruce Thompson, CEO

Action Information Systems

www.expeditefile.com

SNIA™

What is this presentation all about?

- Today's storage model has no idea what it is storing.

- **Expedite implements a new storage model that knows what business information is being stored.**

So What? Who's Suffering?

- Every aspect of unstructured data management is limited by the old storage model:
  - Backup Industry
  - File Servers
  - HSM, ILM, Tier 2, NearLine, LTFS, etc.
    - Hit wall first (>10 years ago)
- Storage model is not wrong, just very limited.
- It's time for a new model….

Why Know What Is Stored?

- Solves many storage problems
- Enables currently impossible features
- Storage management is MUCH easier
- Opens new markets
- New revenue opportunities
- Fun…

WARNING!

- Changes the way we think about storage
- Can make storage people “uncomfortable”
- Users may know more about this than we do
- Change can make investments less “relevant”
- Any use of the word “process” is Kryptonite!
- Storage people find it easy to dismiss…

Ways to Dismiss the New Model

- That's Impossible!
- It's not my job, strategy, scope, etc.
- We don't / can't / won't do processes
- Every process is completely unique anyway
- Requires massive OS and storage changes
- Demands massive professional services
- No one is asking for it
- Customers won't understand it

The Opposites Are True!

All Are False!

What is Unstructured Data?

- Pile-of-Files?
  - Ugly bags of mostly bits?
- Who gets to define it?
  - The ones trying to run their companies with unstructured data
- New community of uses
  - Information Owners – Responsible for data
  - Information Stewards – Manage the data
- How do THEY define Unstructured Data?

When Asked, what do they say?

- Policies
- Contracts
- Easements
- Patents
- Professional Agreements
- Tax Records
- Employee Agreements
- Incident Reports
- Change Requests
- Time Sheets
- Expense Reports
- Purchase Orders
- Engineering Reports
- BoD Minutes
- Flight Records
- Explosive permits

Information Asset

- What business people work with, communicate with, get evaluated against, and paid to control
  - Atomic unit of business information
- Defined as a set of files, tracking data, processes, states, rules, people, etc., that collectively, are meaningful to their organization.
  - All have lifecycles
  - Can include thousands of files

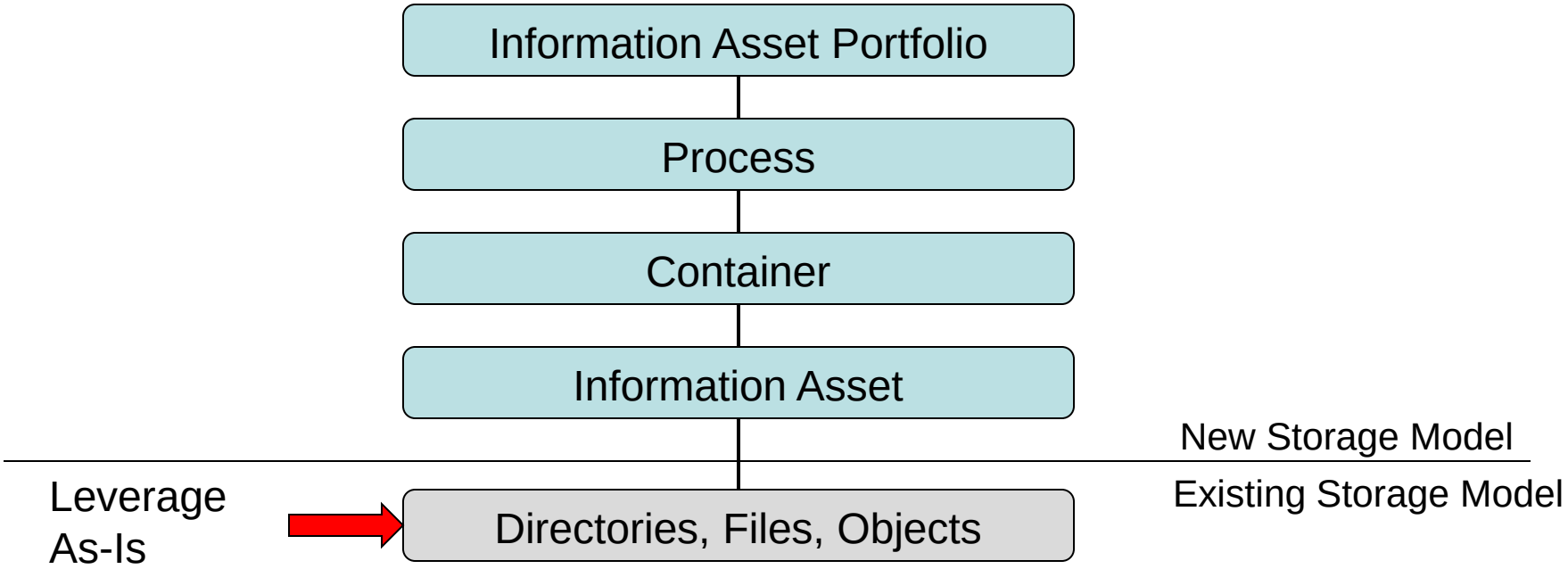
Information Asset Example: Contract

Possible Contract Components	
The Word document	The PDF version suitable to send to the client
The client name	Client contact information
The type of contract	The value of the contract
Template that was used to create the contract	The audit log of everything that happened to it
Previous versions of the contract	Rejected versions
List of people who need to approve it	The sales rep who created it
The costing spreadsheet used to create sections	Scanned signature page
List of people who have not yet approved it	List of people to be notified when approved
Proof that people actually approved it	Proof that the customer received the copy
A cryptographic signature to detect tampering	A mirrored copy for protection
A copy to put up on the website	Login of users who can access it via the website
Copies of important emails from the customer	Any related photos, sketches, or drawings
References to previous contracts	Customer account number
Validation script to run against the contract	How long people are given to approve or reject it

How do they control them?

- When asked, users very quickly start to discuss collections of information assets, all with the same state.
- We call these a Container.
- Where we can automatically apply storage management functions!

The New "Business Stack"



Requirements Difficult to Get?

- Talk to business people in THEIR terms, not storage terms.
- I haven't seen a process that couldn't be defined by an information owner in less than 60 seconds. (usually much quicker)

- The challenge is to be able to do something with their requirements.

WHAT CAN STORAGE DO WITH THIS NEW MODEL?

Here is a simple quiz…

- Want to implement a storage function on files…

A: Requires a way to set a “bit” on the file.

B: Requires the bit to be stored.

C: Implement the storage function.

Q: Which one’s harder?

Problem with Storage

- Setting the bit is orders of magnitude harder!

- Don't know which files to set the bit.
- Don't know when to set the bit.
- Don't know when NOT to set the bit.

- This new model of unstructured data provides this missing intelligence to answer these questions.

Users View of “Storage Policies”

Example of Bit Setting Difficulty

- Ability to set a file to read only vs. read/write.
  - Been there since the introduction file systems
- How are file servers ACTUALLY configured?
  - Full control!
- Anyone can do anything on any file at any time anonymously.

…Tragically Trivial To Trash!

Upset Customer

Can Backup Help?

- What can backup do when a file is deleted or corrupted?

- It can't do anything. Must assume its valid.
- Not provided enough information to know the difference between a change and a corruption.
- Forced to leave the decision to a human to notice the problem and manually run restore.
- Roll the media? Lose the only good copy?

What the Users Want

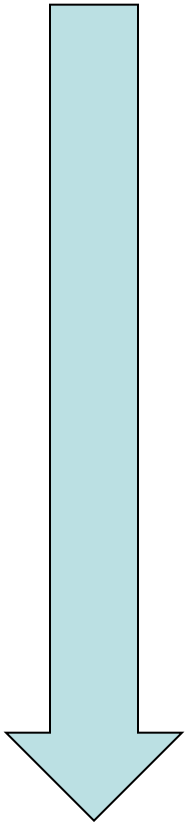
- When the asset is created or changed, protect it and back it up immediately. It is important.
- If someone deletes it by mistake or corrupts it, restore it automatically and immediately.
- But, allow changes to it…
  - Done via process controls
- What happens if you deliver this?
  - Users trust the data
  - Don’t make copies – Most Efficient Dedupe!

New Requirements

- This user community will have a new set of requirements for the storage industry.
  - White paper outlining >100 differences.
- Revolve around their definition of unstructured data.
- Our software, Expedite, provides the bridge between their requirements (market) and today's storage implementations.

Levels of Asset Control

1. Files and Directories
2. First Stage
3. Identification
4. Categorized
5. Process
6. Tailored Process
7. Custom Process
8. Purpose built application



Increasing levels of Process controls

EXPEDITE

Why Is Expedite Different? B.A.S.I.C.

B. Business Users. (Owners and Stewards)

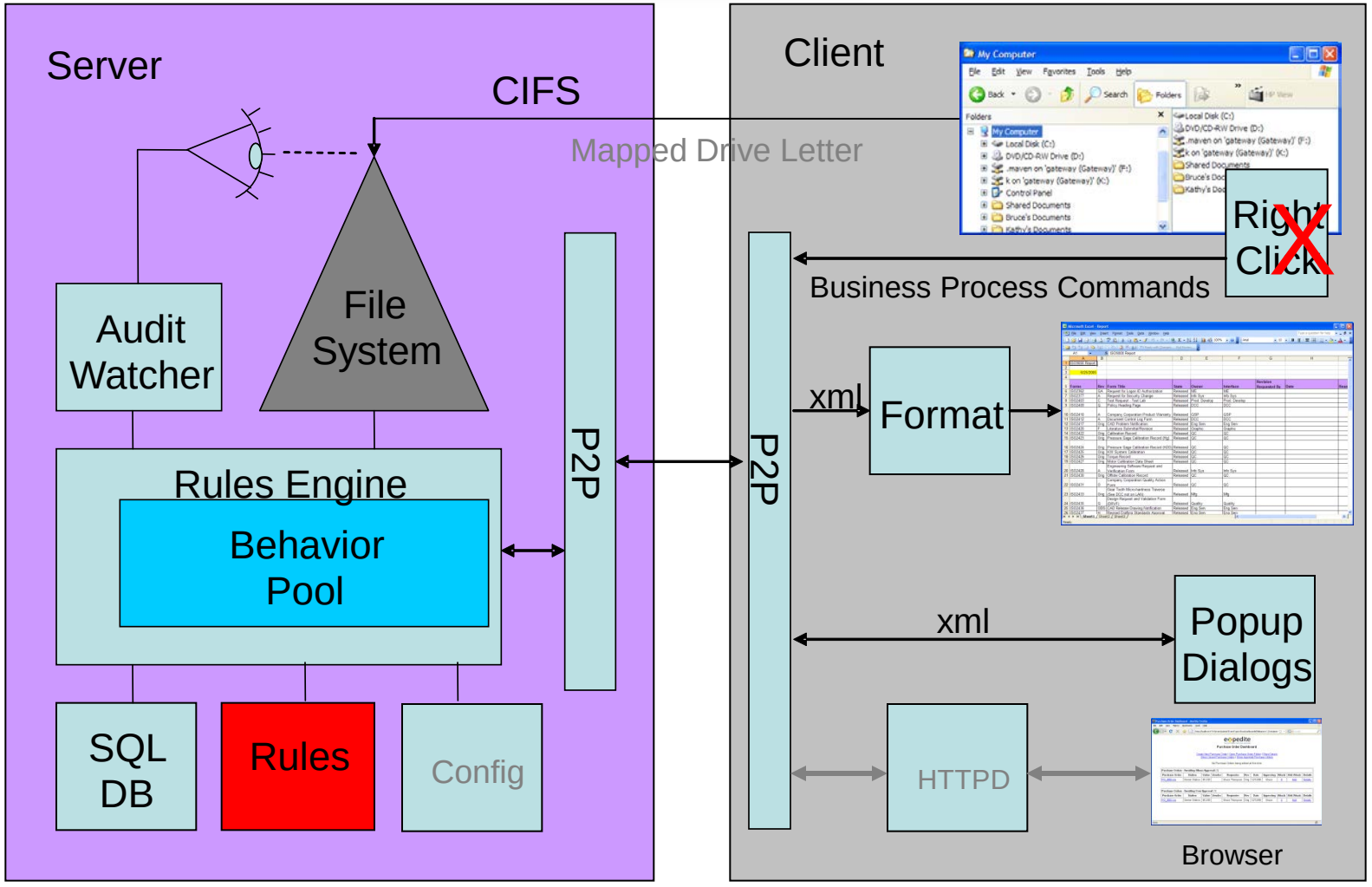
A. Assets – Unit of information is the asset.

S. Storage – Storage aware.

I. Integrated intelligence.

C. Cloud – Handles all the different topologies and access methods needed for today's business.

Basic Architecture



Basic Controls – First Stage Container

- How to setup a First Stage container?
- Answer 3 questions:
  - What to call it
  - Where it put it in the directory structure
  - Where to put it in the portfolio map
  - (optional) who has access to it

First Stage Features

- **Create** – From templates
- **Import** – Single files
- **Upload** – Dir Trees
- **Metadata** – Extendable
- **Attachments** – emails, multimedia, etc.
- **Permissions** – Protect from modification
- **Mirroring** – Auto recovery
- **Crypto-Signatures** – Corruption detection
- **Check-in/Checkout** – Modification process
- **Versioning** – History
- **Logging** – Forensics
- **Search** – Metadata, names, content
- **Archive** – Move to new location.
- **Obsolete** – Controlled deletion process

Setup Enough Containers…

- Result is the Information Asset Portfolio
  - Relationships among the Information Assets
- What users want to use to navigate information
  - Don’t want “enterprise search”
- Now can “set bits” at a very high level and keep them set as data is created and moved.

What can you do with the map?

- Navigate
  - “Go To” the data, not search for it.
- Setup automated storage management
- Documentation (processes, owners, stewards…)
- Security
- Access methods
- Audit
- Business Intelligence and analytics
- Can tell what is not important

Who else needs the portfolio?

- eDiscovery
- Classification
- Business Process Management
- Backup
- Active Archive Compliance
- Data Governance
- Master Data Management
- Information Assurance
- Virtualization
- Identity Management
- File Sharing
- Data Loss Prevention
- Cloud Storage
- Big Data
- Security
- Disaster Recovery

Challenge to SNIA

- Who is going to ultimately control the portfolio?
- Whoever controls the portfolio, will control the industries.
- Will the Storage Industry step up and drive this or will another industry take the lead?

Dealing with Dark Data

- No Silver Bullet
- Users want:
  - Determine what it is, what state it is in,…
  - Find what is important
  - Find what is valid
  - Really want these converted to information assets!
- Why is that so difficult?

Content to Context Conversion

- In the general case, it is not possible to recreate the context of a file through analysis of its content.
  - Plenty of evidence over the past several decades that this is true.
  - Asking for the creation of information that is simply not stored in the computer.
- Big Data's assertion that pouring all unstructured data into Hadoop may not yield useful results.

Dark Data – Attack Strategy

- Create the asset portfolio.
- Have places to put the existing files.
- Have some basic process controls moving forward.
- Move existing data up the process “value chain” as appropriate for the type of information.
  - Can require signification human interaction.
  - May not be worthwhile or completely possible.
    - Decide if business decisions are still pending.

ILM

- Promise of “slow, huge, cheap” devices?
- 60-80% of data not accessed for a year?
- Move the old to cheap storage!
- Significant potential for cost savings.
- Has been attempted many times. (6 for me!)
- All have been file system based (like LTFS).
- All have failed. (NOT a shared service for all)
- People have been doing this with paper forever!

ILM Challenges and Limitations

- Users and operating systems can't handle large delays.
- Something or someone must orchestrate device access to prevent massive performance problems.
- Ends up limited to a single application, sometimes a single person.
- What is needed is a common, integrated, shared, service available to all.

Required for ILM Support

- Must know what data can be moved to the library and when. **Can't guess**
- Can't "automatically" restore a file on access.
- Must provide an interface so users know before they are going to have to wait.
  - Let users decide how to proceed.
- Users have to know before hand that if they retrieve data, it is what they are looking for.
  - This is the hardest one of all.

What If We Did ILM Correctly?

- What if we were able to utilize huge, slow, cheap storage as a general purpose sharable service integrated into our processes?

- Convert **piling problem to a pipeline problem**.
- That is the only real winning strategy to attack the tsunami of unstructured data.
  - Can tape do this?

Conclusion

- Industry is limited by the storage model
- Every aspect of storage management can be improved, expanded, and integrated
  - Rare technology that can amplify the value of existing infrastructure investments
- The Storage Industry should control the portfolio
- A new set of users to empower
- Aligns business with IT
- Looking for progressive partners…

Contact Info

Bruce Thompson, CEO

Action Information Systems, Inc

www.expeditefile.com

- See “For Storage Professionals” for details.

brucet@expeditefile.com

303-912-3172