Second Generation Archiving - Shining the Light on Dark Data

Acknowledgements and Disclaimers

Availability. References in this presentation to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates.

The workshops, sessions and materials have been prepared by IBM or the session speakers and reflect their own views. They are provided for informational purposes only, and are neither intended to, nor shall have the effect of being, legal or other guidance or advice to any participant. While efforts were made to verify the completeness and accuracy of the information contained in this presentation, it is provided AS-IS without warranty of any kind, express or implied. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this presentation or any other materials. Nothing contained in this presentation is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software.

All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer. Nothing contained in these materials is intended to, nor shall have the effect of, stating or implying that any activities undertaken by you will result in any specific sales, revenue growth or other results.

Abstract

Second Generation Archiving - Shining the Light on Dark Data

Second generation archiving solutions manage content based on value, removing unnecessary data to lower operating costs, storage costs, and improve infrastructure and backup performance.

Join this session to learn how to implement an archiving solution to meet storage cost reduction objectives, and why archiving should be a part of every IT department's core competency.

Agenda

The true cost of First Generation archiving

- A historical perspective
- Issues & pain points

o **Case study:** From First Generation email archiving to a lower cost, lower risk approach

o IBM capabilities to help IT manage information by value

Information volume is doubling every 2 years,
1.8 zettabytes in 2011.
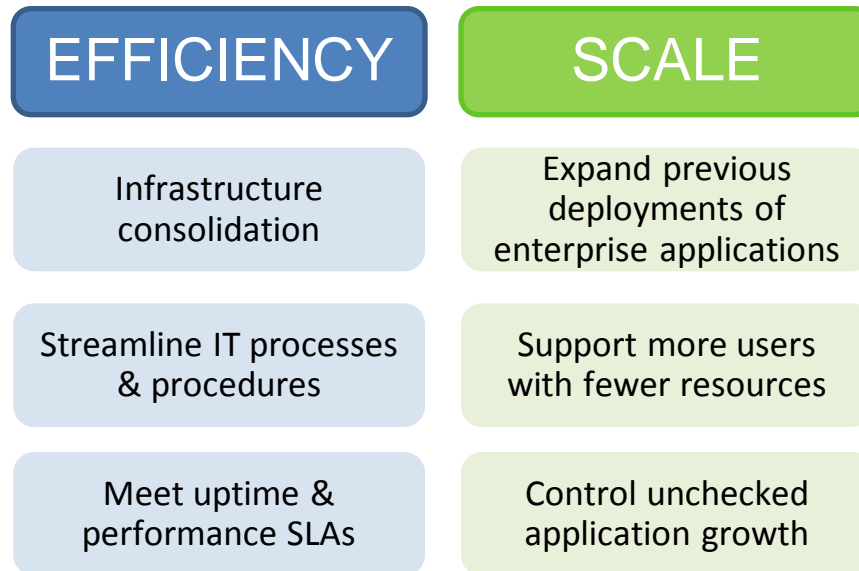
By 2020 it will be 50 times that amount.

The IT staff to manage it will grow less than 1.5 times.

2020

2011
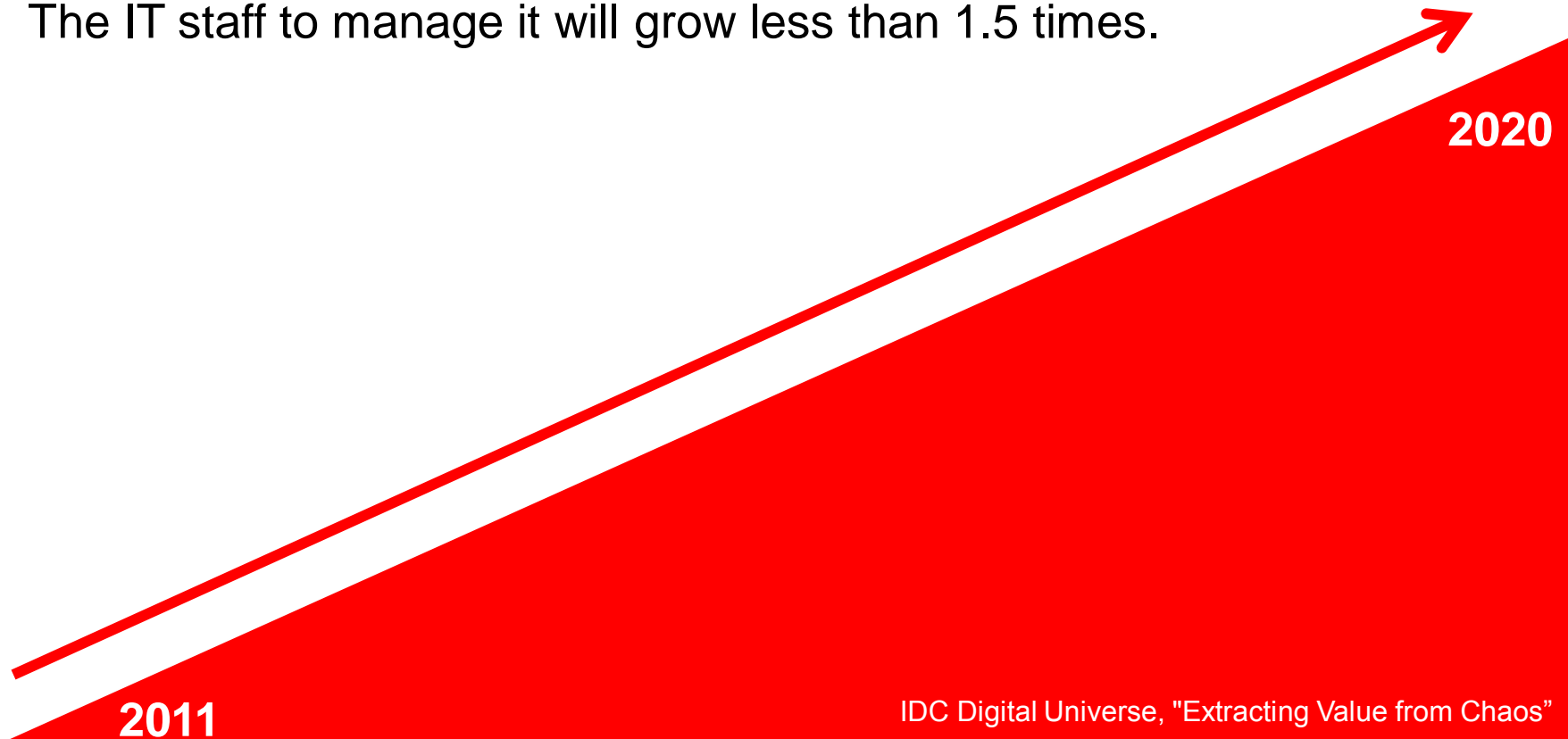
A Historical View

Archiving was introduced as means to keep active information produced in applications and enterprise systems *outside* of those systems in order to:

1) Lower infrastructure costs and improve operational efficiency which suffers as a result of large data volumes

 Faster backups on production systems

 Administration utilities ran more quickly

 Fewer failures and shorter time to recover from failures

2) Allow source systems to scale by removing old information
 Better performance from existing hardware

 Longer periods between hardware refresh & application upgrades

EFFICIENCY	SCALE
Infrastructure consolidation	Expand previous deployments of enterprise applications
Streamline IT processes & procedures	Support more users with fewer resources
Meet uptime & performance SLAs	Control unchecked application growth

The Third Driver: Compliance

Increasingly, ensuring compliance with regulations became a primary reason for implementing archiving of unstructured data with initial interest mainly from Financial Services.

Archiving provided a way to supervise employee communication with the public (i.e. NASD/FINRA for Broker-Dealers)

EFFICIENCY **SCALE**

Consolidate infrastructure / Retain information for duty or obligation / Expand previous deployments of enterprise applications

Streamline IT processes & procedures / Basis for transparency & internal governance / Support more users with fewer resources

Meet uptime & performance SLAs / Provide eDiscovery capability / Control unchecked application growth

Characteristics of The First Generation Archive

Archive Islands	Email centric; Difficult to expand to new data sources such as SharePoint, Social *(Impact: IT efficiency; Compliance holes)*
Keeps Everything	"Straight through Archiving" - All data is archived as if it is if equal importance *(Impact: eDiscovery Cost; Storage & Management; Legal Risk)*
Keeps Forever	Blanket retention period for all data, often not even turned on *(Impact: Over-retention creates large with additional storage costs, legal risk implications and eDiscovery cost)*
Limited eDiscovery	eDiscovery searches require IT (or a third party) to execute, adding delays, overhead, and service charges. Data often re-collected (duplication) for use in eDiscovery application. *(Impact: Time, Higher cost including penalties for not meeting deadlines, burdening IT staff, duplication of data*)
Impact User Productivity	Little or limited email client integration – separate web search interface required *(Impact: End user efficiency)*

Where and How IT can help remove cost & risk

4 Target Cost Compression Areas for Defensible Disposal and ILG Effectiveness

Where		How

Infrastructure storing data with no utility

1 **Data that can and should be disposed**
-- Application data
-- Files and end user docs
-- Email
-- Back up

DISPOSAL

**Analyze existing data in place before archiving;
Identify what is important;
Delete the rest.**

Storing data at a cost higher than its value

2 **Data that can and should be tiered and archived as its utility wanes or when its cost exceeds its value**
-- Aging application data, end user files, and email
-- Non-prod instances
-- Records and regulatory content with no business need

ARCHIVING

**Collect and archive only information with value
Compress, De-duplicate, migrate through storage tiers as value diminished**

Application costs without business value

3 **Applications, infrastructure and data that can be retired**
-- Moth balled applications with vague legal requirements
-- Cycle time and throughput acceleration
-- Consolidation and decom

DISPOSAL

Analyze for abandoned data, applications and accounts; reclaim and consolidate hardware, license fees

---

E-Discovery external vendor spend

4 **External Legal vendor spend that can be reduced, avoided and eliminated**
-- Pre-assess data before and after collection
-- More precise hold scoping with higher transparency
-- Cull and review internally

DISCOVERY

Solving over-retention results in smaller data sets, and lower discovery cost. Legal Self-serve for discovery and hold speeds response, lowers IT involvement
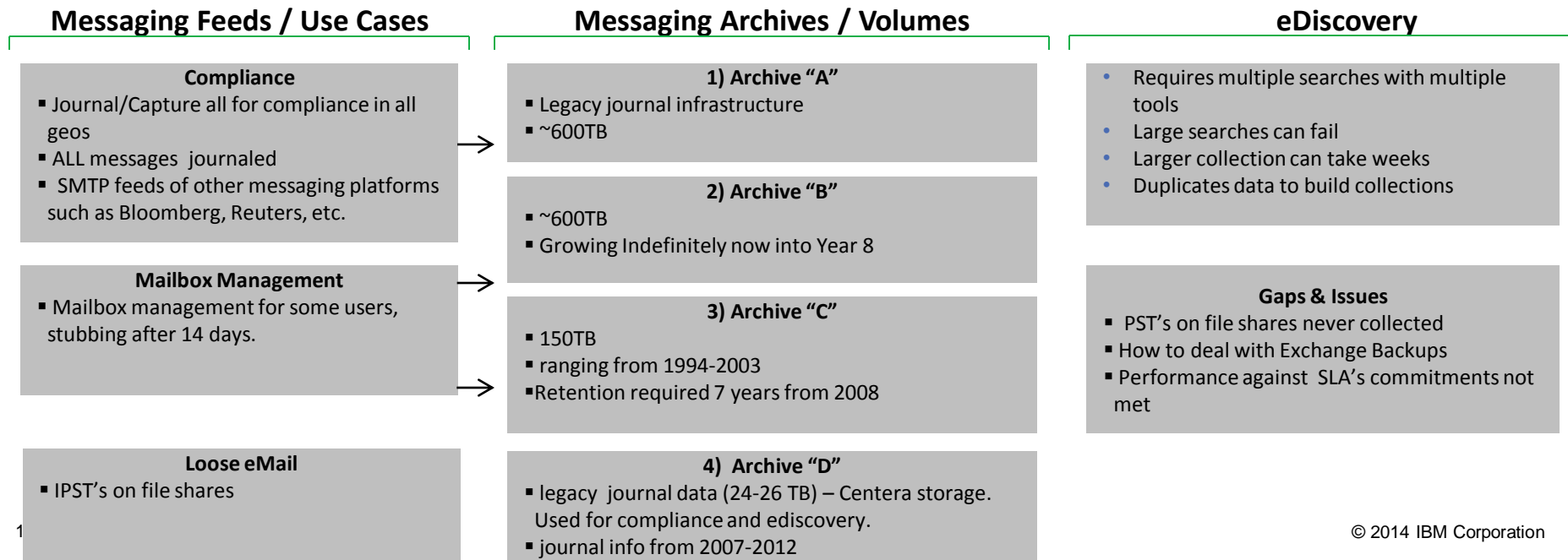
Unstructured Data Archiving

CASE STUDY

The Company's Messaging Current State

Stakeholders

	Legal	Records / Compliance	End User
Mode of Operation	▪ Hold assumed from **100% retention**, no holds applied in archive ▪ Conduct collections and export only when need arises ▪ No Early Case Assessment available for attorneys – maximizes cost and time.	▪ No message based record identification or classification ▪ No ability for users to declare records from Messaging System(s)	▪ Attachments moved to 1 Archive after 14 days for some users, however legacy PST's proliferate on workstations
Impact	▪ Collect more messages than required ▪ Unable to distinguish which messages are subject to any hold and therefore volume accumulates in perpetuity ▪ Backup restores needed	▪ Single retention period, if any, must be applied to all messages ▪ Can't apply any disposition to messages because no classification of records and no mechanism for defining business retention (co-mingled)	▪ Local PSTs are unmanaged, exist in multiple places and are expensive to collect ▪ User quotas impact business

Information Flow

Messaging Feeds / Use Cases

Compliance
▪ Journal/Capture all for compliance in all geos
▪ ALL messages journaled
▪ SMTP feeds of other messaging platforms such as Bloomberg, Reuters, etc.

Mailbox Management
▪ Mailbox management for some users, stubbing after 14 days.

Loose eMail
▪ IPST's on file shares

Messaging Archives / Volumes

1) Archive "A"
▪ Legacy journal infrastructure
▪ ~600TB

2) Archive "B"
▪ ~600TB
▪ Growing Indefinitely now into Year 8

3) Archive "C"
▪ 150TB
▪ ranging from 1994-2003
▪ Retention required 7 years from 2008

4) Archive "D"
▪ legacy journal data (24-26 TB) – Centera storage. Used for compliance and ediscovery.
▪ journal info from 2007-2012

eDiscovery

• Requires multiple searches with multiple tools
• Large searches can fail
• Larger collection can take weeks
• Duplicates data to build collections

Gaps & Issues
▪ PST's on file shares never collected
▪ How to deal with Exchange Backups
▪ Performance against SLA's commitments not met

3 requirements or gates to govern & dispose of messages in archive

Stakeholders & Governance Requirements to Dispose

1 Legal Holds / eDiscovery	**2** Records / Compliance	**3** End User Experience
Legal must be satisfied that messages subject to any existing holds are not disposed and no unique messages can exist in PST's. This entails:	**Compliance must be satisfied that messages with a regulatory retention obligation are not disposed. This entails:**	**Business units/users must have acceptable access to messages w/ business value. In addition, no disposed messages can have copies that exist in PST files. This entails:**
1. Accurate inventory of all existing holds	1. Routine policy refresh	1. Optimal user experience with transparent access, ability to search and retrieve archived messages and records
2. Apply holds to messages	2. Align information classes to employees and enable automated or semi-automated record classification	2. Universal rollout across Geographies
3. Refine, revise, and apply holds over the course of matter and as new holds come in (continuous and real time)	3. Extract records from existing archives still requiring retention and execute disposal schedule on both records and archive	3. Provide an easy to use mechanism for business users to declare records and messages required for longer than the policy retention period (i.e. Credit Default Swaps for life of swap plus 5 years)

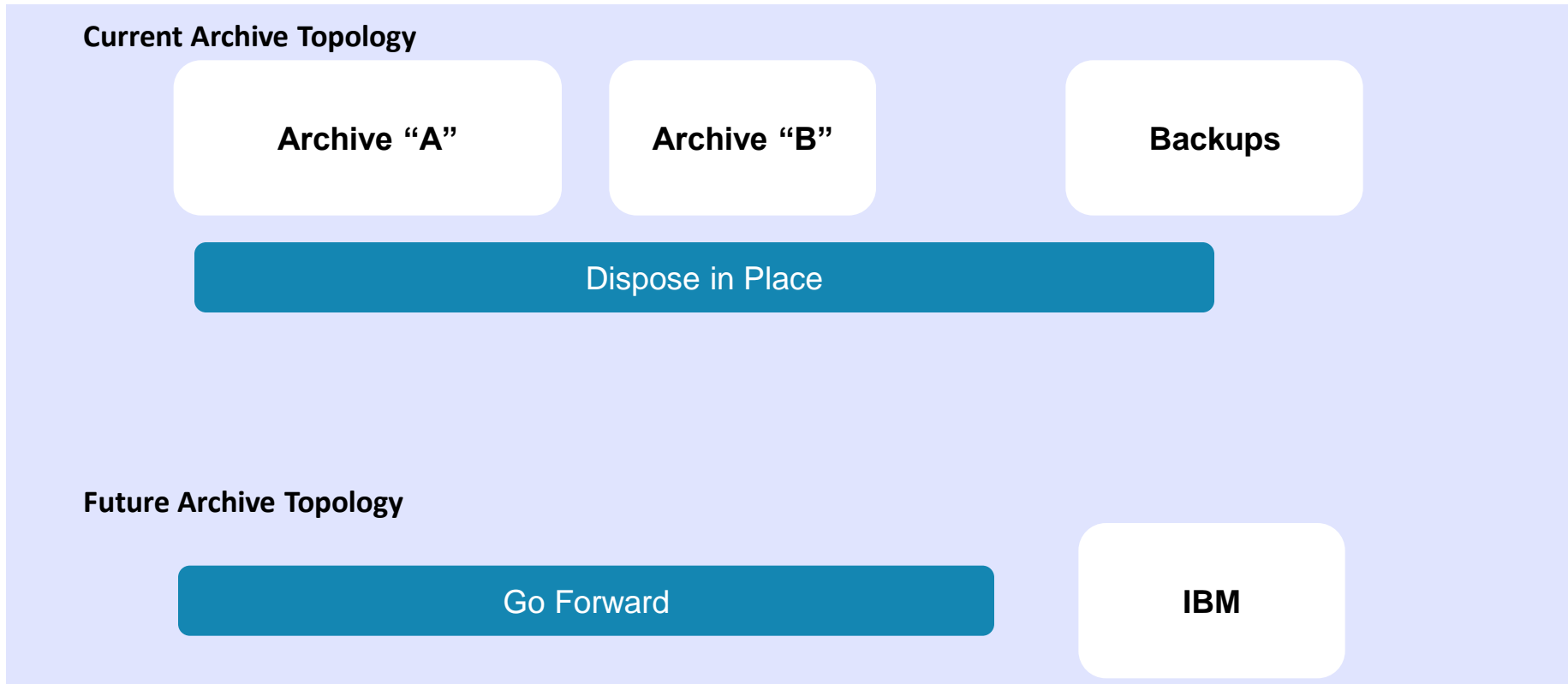# 3 Options to Dispose of Messages & Optimize ILG Messaging

Option	Description	**1** Legal: Ensure messages subject to hold are not disposed	**2** Compliance: Ensure messages with a regulatory retention obligation are not disposed	**3** End User: Business must have acceptable access to messages w/ value and no PST's with unique messages can remain
1) Ad hoc disposal of messages by mapping longest significant umbrella retention & preservation requirement (e.g. 5, 6 or 7 years)	Determine longest retention requirement: (1) Policy Approach: Create a superset of all holds and retention categories to be executed in EV archive (2) Execution Approach: engage EV to classify and protect relevant data, then delete all "else" (3) Manually set policy in Zantaz. (4) Roll all non-archived users into one of the archives	Manually map/refine legal holds to custodian/date range (~6 yrs.): ▪construct map of all existing holds by custodian and date range ▪refine map to ensure completeness (higher bar with disposal) through attorney review of holds (query of messages to refine custodians) ▪create updated official legal holds map of messages in archive	Create a standard retention period for eMail, identify exceptions and extract Refresh policy for The Company's obligations for messages and determine standard retention period (applies to ~90%) - (~6 years) ▪Likely Exceptions: (1) Time period is longer than standard and (2) Event-based retention exceptions identified and moved ▪Create updated retention map of messages in archive	Notify business of disposal period, flag exceptions, remove PSTs ▪Business notified, exceptions flagged – notify disposal range, when exceptions occur over period must create administrative hold in archive ▪Ensure removal of applicable PSTs from file shares.
2) Automate disposal policy and execution in IBM Content Management Platform on a "go-forward" basis; retain 1, 2 or 3 legacy archives and dispose in annual batches.	Maintain 1,2 or 3 archives but cease ingestion. For new messages, archive in platform capable of automatic hold and retention execution	Automatically set flags for holds in new archive from the syndication enabled legal holds tool. Run searches in archives for messages subject to hold that are longer than retention period about to be applied and store in new archive.	Automatically apply retention policy to IBM Platform including support for event based retention	Incorporate and apply business value period automatically alongside retention policy and legal holds in IBM Platform, ease of access to stubs enables discontinuation/remediation of PST creation or additions going forward
3) Automate disposal policy and execution in governance enabled archive; migrate messages from legacy archives	Automated policy application (holds, retention, business value) directly to data in IBM ECM platform, migrate archives to IBM ECM Platform and apply policy there	Automatically set flags for holds in new archive from the legal holds tool. Migrate existing holds and archives, refine using search in archive	Automatically apply retention policy on all messages including support for event based retention	Incorporate and apply business value period automatically alongside retention policy and legal holds, ease of access to stubs enables discontinuation of PST creation or additions. Engage in PST remediation work stream described below

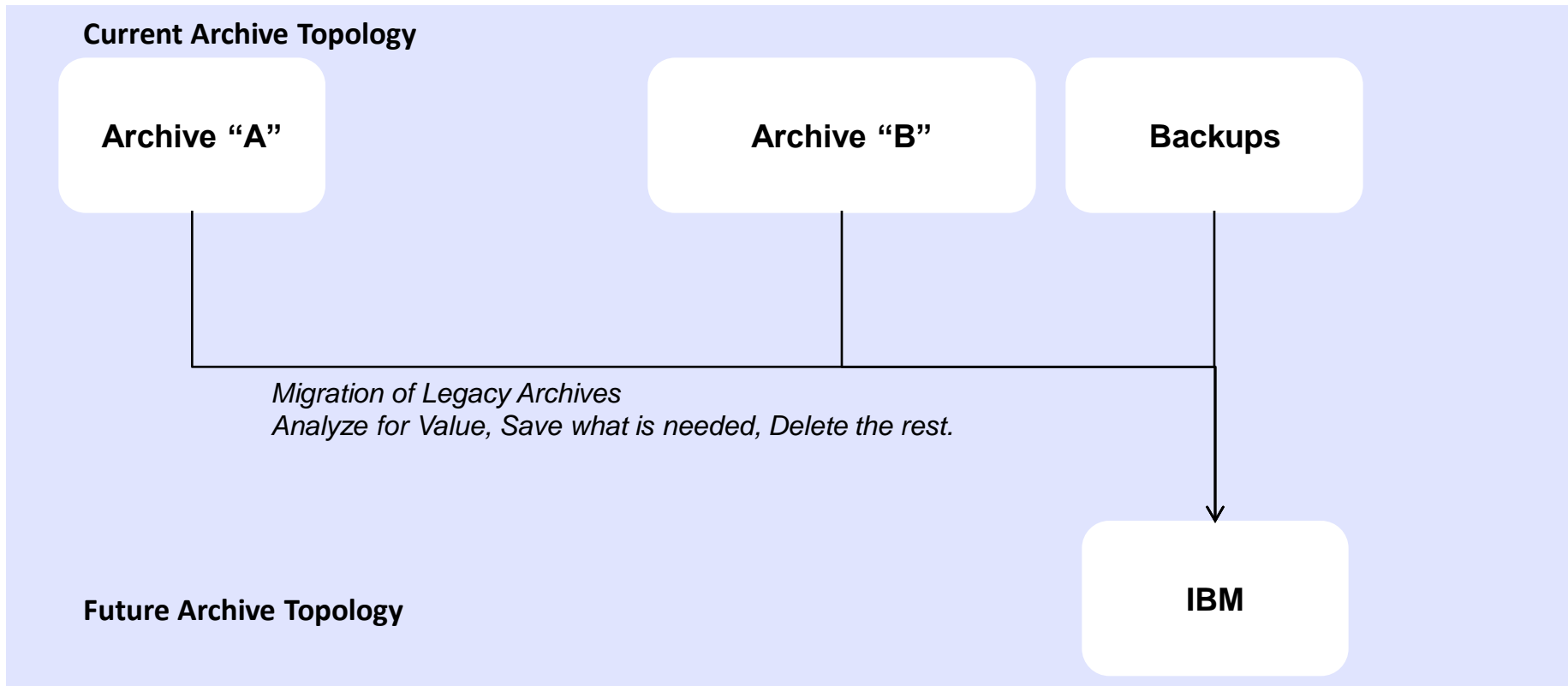# Option 1: Apply Retention/Disposition to Archives in-place

Current Archive Topology

| Archive "A" | Archive "B" | Backups |

| Archive "A" | Archive "B" | Backups |

Future Archive Topology
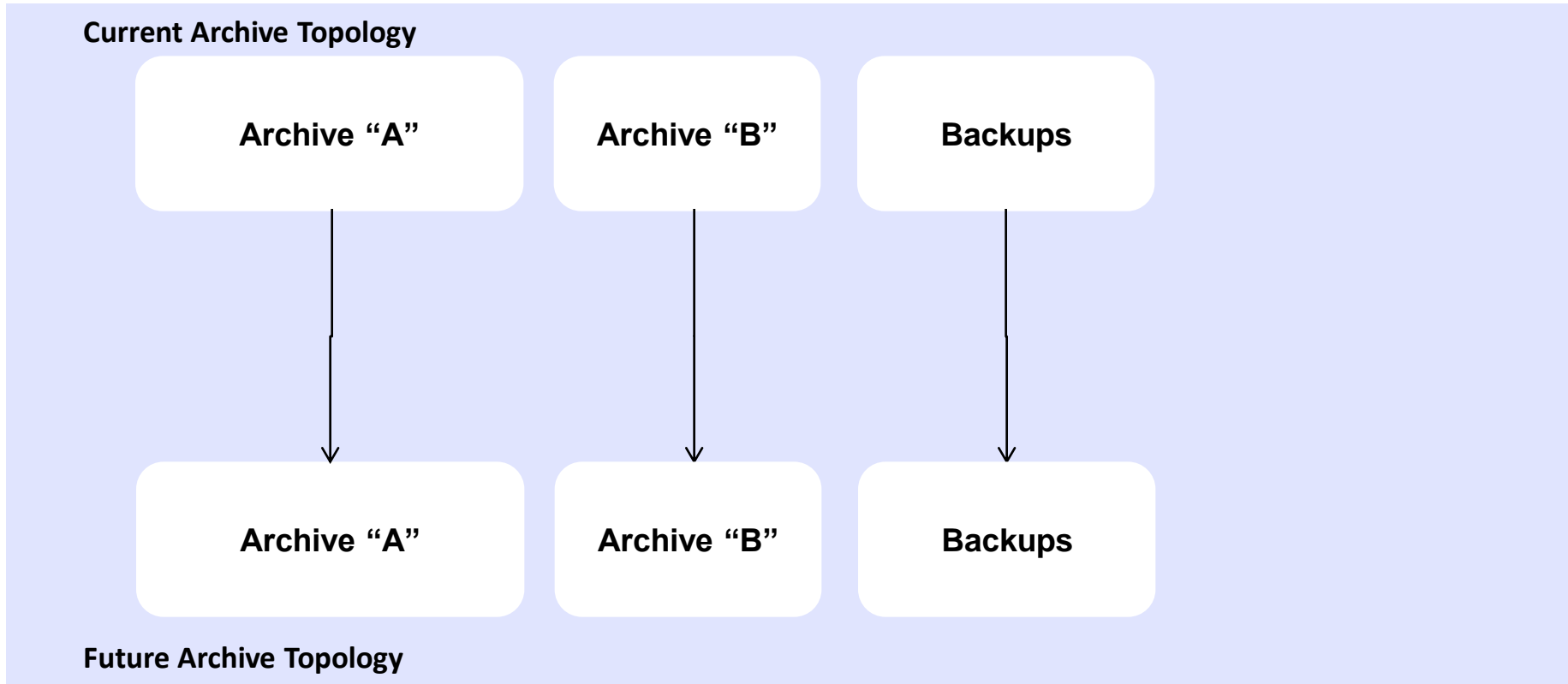
Investment	Savings	Change Required	Risk Reduction

Option 2: Dispose From Legacy Archives in Place
Deploy Governance Aware Archive Going Forward

Current Archive Topology

| Archive "A" | Archive "B" | Backups |

Dispose in Place

Future Archive Topology

Go Forward   **IBM**

Investment	Savings	Change Required	Risk Reduction

Option 3: Deploy Governance Enabled Platform to Apply Policy to eMail & Migrate/Decommission Legacy Archives

Current Archive Topology

Archive "A"	Archive "B"	Backups

Migration of Legacy Archives
Analyze for Value, Save what is needed, Delete the rest.

Future Archive Topology

IBM

Investment	Savings	Change Required	Risk Reduction

Optimal Recommendation #3: Governance Aware Platform for New Content, Migrated Messages, Existing PSTs and Backup Tapes

Stakeholders

	Legal	**Records / Compliance**	**End User**
Mode of Operation	• Legal users define holds which are automatically applied directly to messages • Legal users can search, analyze, pre-cull and refine scope of holds in archive – with no IT involvement	• Policy automation system employed so record classes can be tied to employees, classification enabled and business value incorporated for disposal execution	• Archive is transparent to user, no additional clicks or interfaces. • Enjoys an "infinitely" sized mailbox making PST's obsolete. Easily declares records with a single click or auto declares.
Impact	• Messages in archive are automatically flagged as "held" and protected from disposal until all holds are lifted • Less data is collected, processes and reviewed, no backup tapes	• Retention schedule is a source of truth for disposal and executed in archive • Records are routed to record repository for holistic records management	• No negative impact on the business users • Improves user experience by removing debris, providing search across files and messages

Information Flow

Messaging Feeds

Exchange & Notes Feeds
• All required messages

Other Feeds
• LinkedIn, Twitter and Facebook messages
• Bloomberg and Reuters direct feed

IBM Content Collector captures based on policy, applies deduplication en-route to Archive & trips relevant flag such as "Declared as Record"

Single Messaging Archive

Archive provides unified approach for messages and files and SharePoint

• One archive for both retention and mailbox management
• Retention schedule applied in place
• Holds applied in place
• Event based records disposition
• Seamless stubbing for no change in employee experience
• Early Case Assessment & Culling in place for Legal
• Self-service search for users
• Common FINRA Supervision enterprise wide

Risk Remediation & Defensible Disposal

Summary Email Volume Reductions
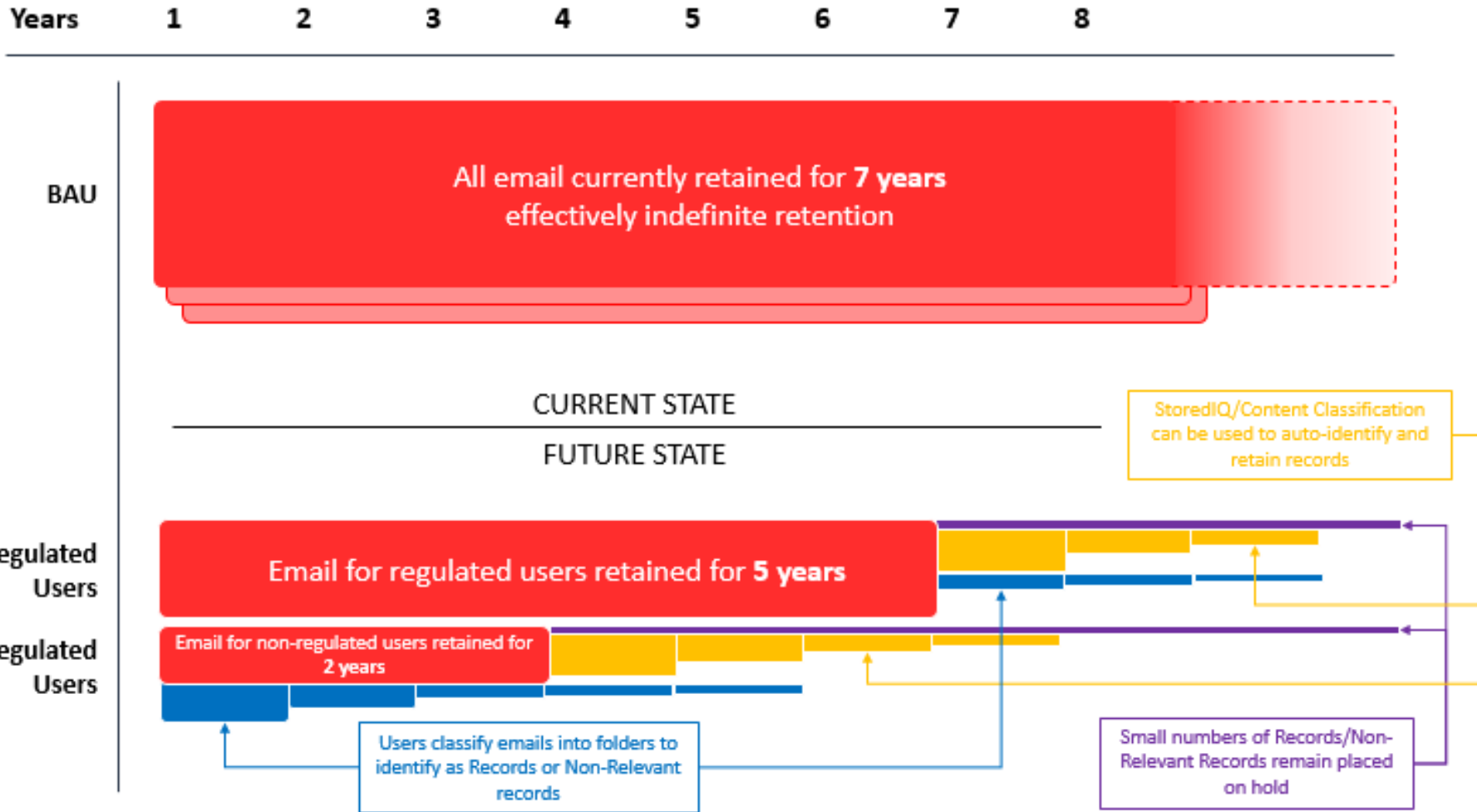
Capabilities for IT

IBM VALUE-BASED ARCHIVING

Analyzing Data In-Place Before Preserving or Collecting

Analyze content across organizational data sources and identify owners to determine people and systems with potentially relevant information.

- Abandoned File Systems
- Explore for local PSTs
- Examine data for risk (SSN Numbers, Credit Card Numbers)
- Monitor unauthorized file type
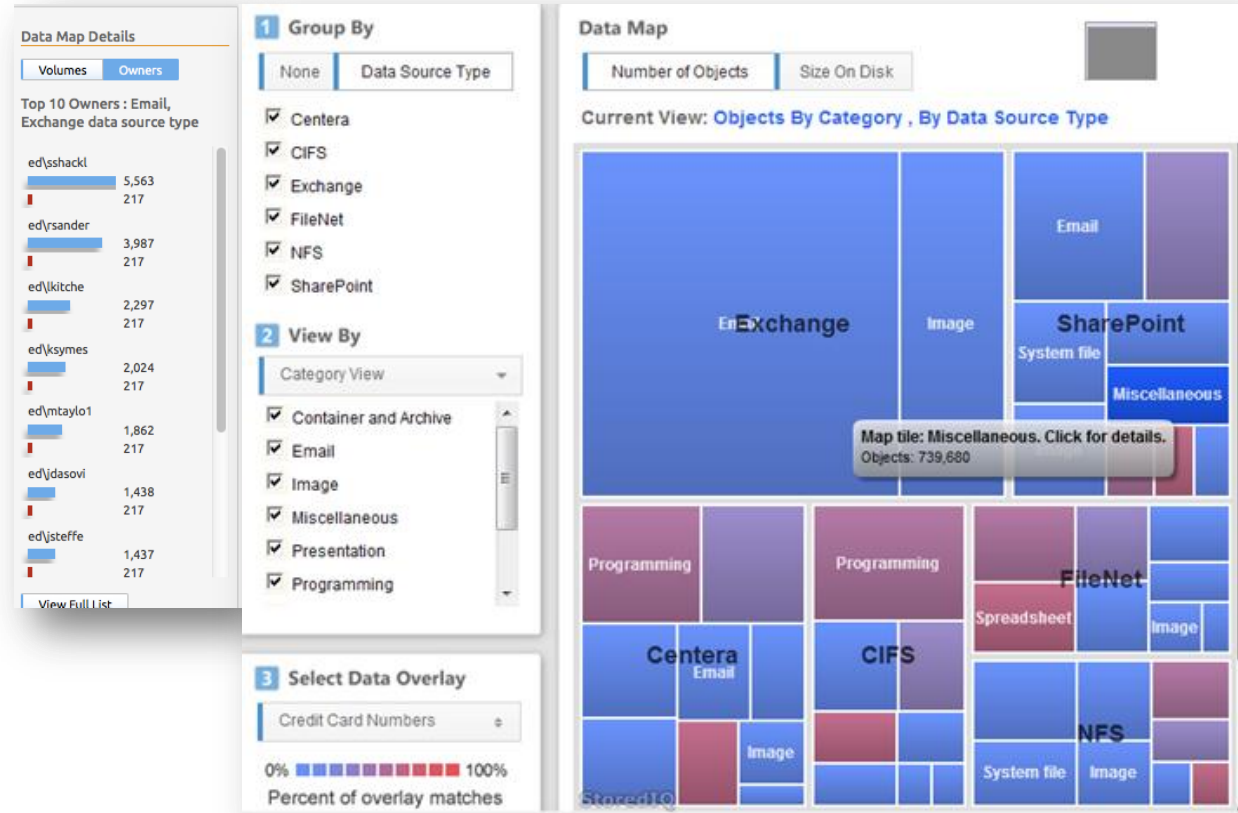
- IBM eDiscovery Identification and Collection
- IBM Policy Assessment and Compliance
- IBM Unstructured Data Identification and Management
- IBM Desktop Data Collector

Archive Content Based On Its Business Value or Obligation with IBM Content Collector

Selectively archive from email, file systems, social media and collaboration systems according to VALUE

Filter on source to determine what to archive:
a. Age
b. Size
c. Content Type
d. Metadata, properties and attributes
e. Classification based on message/file contents

De-duplicate to reduce volume

Automatically Apply Retention & Expiration Policies according to value or legal obligation

Data-source specific retention schedules are defined centrally by experts, and syndicated to other applications

Schedules

Name	Description	Organization	Records Cordinator
ADM120, Purchasing and Requisition	Records related to the request, authorization, and procurement of products and services. Excludes records for payment of products or services. See FNA100 for Accounts Payable	GWM-NYC (US)	Ross, Alex
CDR100, Departmental Administration	Records documenting business unit administration, planning, and management activities, and other miscellaneous administrative. Records not covered elsewhere.	GWM-NYC (US)	Ross, Alex
ENV100, Accident and Injury Records (OSHA & Non-OSHA)	Records related to general liability and on-the job accidents.	GWM- Private Bank US - Chicago (US)	Mason, Judy
HUM420, Personnel Files	Records providing a history of employment from initial hiring through termination/resignation.	GWM-NYC (US)	Ross, Alex
SAL140, Market Research and Analysis	Records related to the development, administration, and analysis of market research.	GWM- Private Bank US - Chicago (US)	Mason, Judy
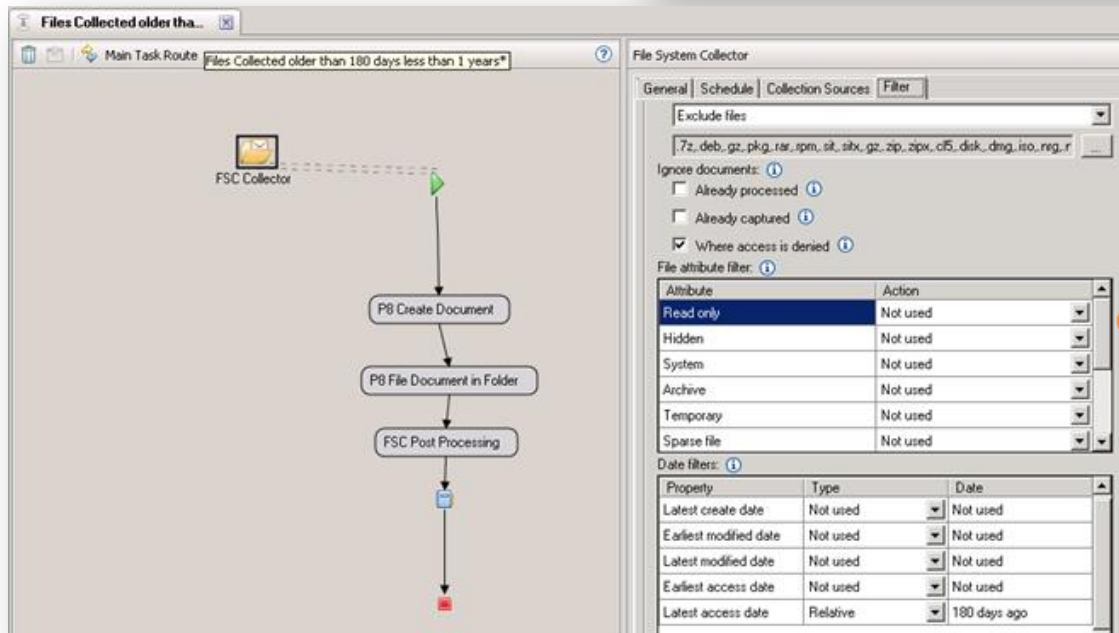
Retention instructions attached to data during collection and executed automatically at the appropriate time.

Journal & Archived/Stubbed Email Stored Only Once

Archived email appears in the user's inbox for a seamless experience.

Unlike First Generation archives, no web page opens and no additional searching is required.

This gives users an "infinite" mailbox, so they do not need PST's. PST existence is the gate to defensible disposal of eMail.
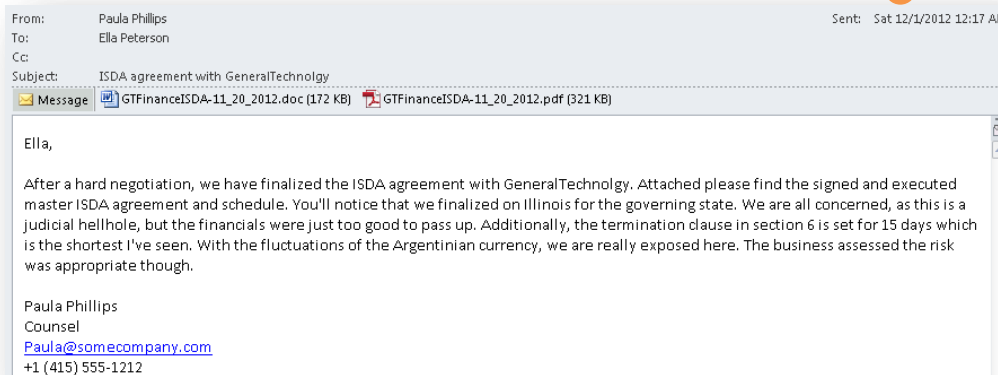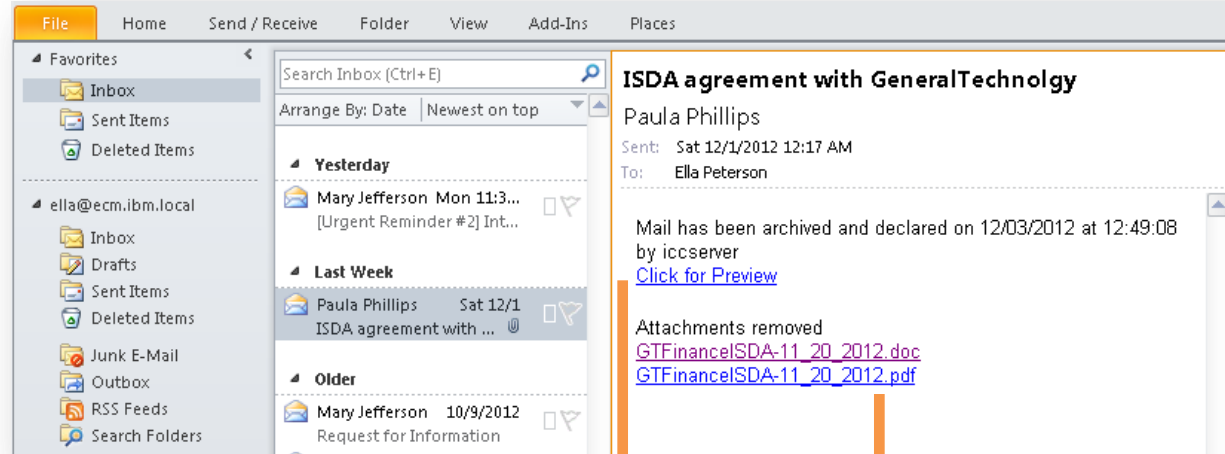
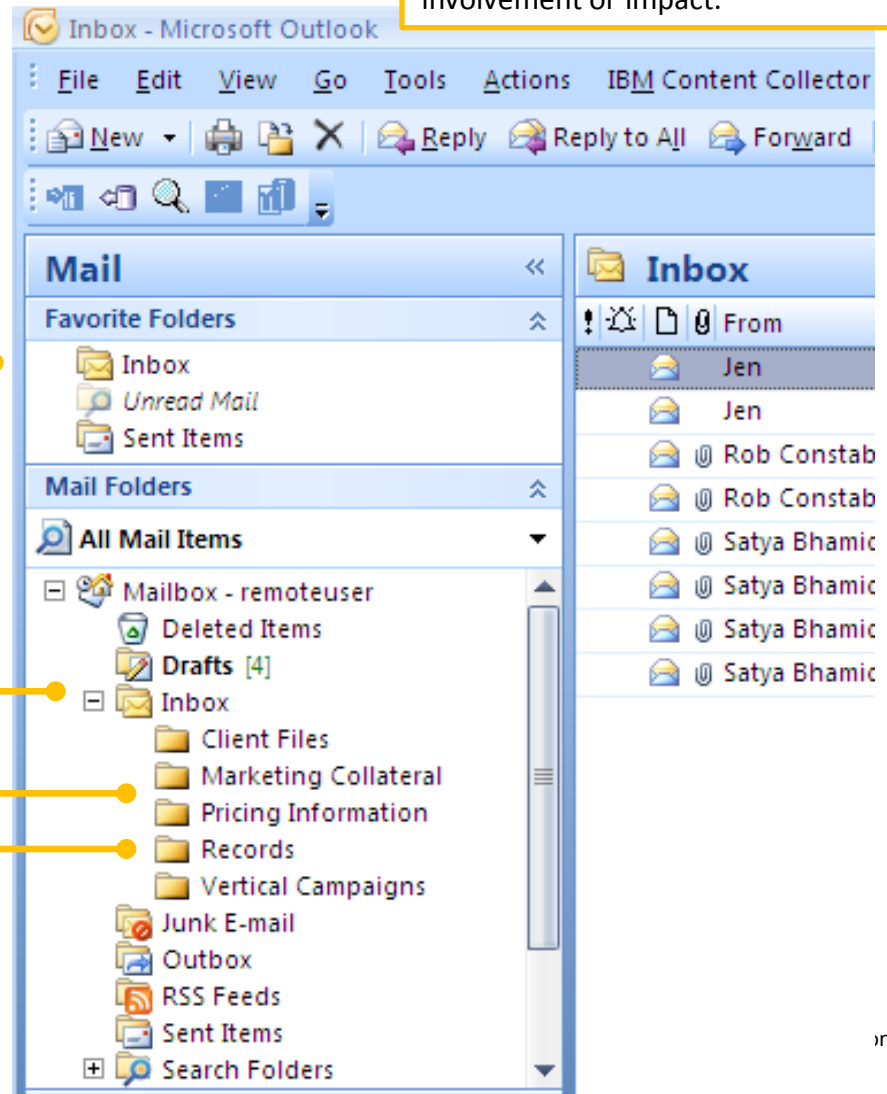# Deep integration with email client for ease of use: Instantaneous Retrieval; Classification & Application of Retention Policy Performed in the Background

Note: Messages for regulated users are journaled without user involvement or impact.

Messages in the root Inbox that are not foldered and are purged after 60 days in-place.

Messages Deleted from the production system completely are held in the Dumpster for 15 days in case users need deleted items.

Messages Foldered in any way (all non-root folders other than records folders) are stubbed to the archive after 60 days. Retention is set at the user's business unit's business value period.

Messages moved to a Records Folder are stubbed to the archive and automatically declared as records.

24

Reduce risk in archived content by blocking non-compliant messages before they are sent

Define policies that are then executed against outgoing messages to determine adherence to corporate policies &.procedures.

From: Administrator
To: Debbie Patrick
Cc:
Subject: Warning Advisory [ID=CGCZLBGVGJVBQNBCCG]

Sent: Thu 5/16/2013 1:46 PM

Message   [a] Order Details - ForwardInc to BackwardInc International   (10 KB)

** You received this message because your email breaches compliance policy and has not been sent: **

Date: Thu, 16 May 2013 13:45:26 -0400
Subject: [a] Order Details - ForwardInc to BackwardInc International

You must take action on this message, or your email will not be sent.
If you still want to send your email, please reply to this message without making any changes.
If you want to comply with policy, you must resend a modified version of the original email.

** WARNING INFORMATION **

Warning Advisory
This message or the associated attachment(s) you are attempting to send is in possible violation of a policy called "Credit Card Information" which is considered personally identifiable information (PII) and can be used for identity theft, planning robberies, stalking etc.
Possible recommended actions that can be taken to eliminate or mitigate risk:

1. Remove words or phrases that you consider being in violation of this policy before sending the message.
2. Do not send message.
3. Password Protect or use appropriate encryption policy.
4. Confirm all recipients are approved for communication.

Contact Jeff Andrews, security officer on x7335 if you have questions regarding this message. [DC1]

- ▷ Corporate and Regulatory Compliance
- ▷ Customer/ Supplier Treatment
- ▷ Employee Behavior
- ▷ Intellectual Property (IP)
- ▷ ☑ Legal
- ◢ ☑ Non-Public Information (NPI)
  - ▷ ☐ Board Minutes and Discussions
  - ▷ ☐ Corporate Contracts
  - ▷ ☑ Customer Lists
  - ▷ ☐ Draft Documentation
  - ▷ ☐ Financial Information - Balance Sheet
  - ▷ ☐ Financial Information - Income Statement
  - ▷ ☑ Financial Information - Projections
  - ▷ ☑ Information Security Label Control
  - ▷ ☐ Inside Information: Front-Running/ Trading Ahead
  - ▷ ☐ Inside Information: Non-Public Company Information Loss
  - ▷ ☐ Inside Information: Non-Public Financial Information Loss
  - ▷ ☐ Inside Information: Rumors and Secrets
  - ▷ ☐ Inside Information: Trading Ahead of Research
  - ▷ ☐ Internal Investigations
  - ▷ ☐ Internal IT Support Documents

Non-compliant Emails are not sent but returned to sender for remediation or quarantined for further review

Apply real time enforcement to ensure compliance with electronic communication policies. Report on performance.

Identify potential regulatory breaches in email

Gain real time insight into compliance performance

Legal Users can create cases directly on the archive & apply holds or export case data directly to outside counsel without involving IT

Reduce process steps and avoid duplication of data.

Export data in its native format or in the industry standard EDRM XML

Information Lifecycle Governance

Proprietary and Confidential – Do not distribute without consent from IBM

Retention Schedules Syndicated and Executed Automatically on Structured & Unstructured Data for Efficient, Defensible and Routine Disposal

Access and execute approved schedules on records repositories, structured and unstructured archives, SharePoint and other data sources.

Changes in business or regulatory requirements initiated by policy owners are automatically syndicated for reliable, defensible and efficient execution.

Application archive automatically receives retention and disposition schedule to execute on records.

Learn More!

IBM Education courses available

IBM Redbooks

Chapter 1. Value-based archiving and defensible disposal overview
Chapter 2. Example use cases
Chapter 3. Dimensions of content archiving themes
Chapter 4. Designing, adapting, and deploying task routes
Chapter 5. Retention management
Chapter 6. Document classification
Chapter 7. Records management integration
Chapter 8. IBM Connections integration

Keep the conversation going!
Talk further with the ILG team at your organization and at these upcoming events!

Setup a call with IBM's ILG experts to assess your archiving maturity

o Brent Pohl – bpohl@us.ibm.com - +1 (512) 286-3225

3rd Party Industry Events:

o Wisconsin IT Symposium | Milwaukee | May 7

o AIIM Seminar – Resetting SharePoint Expectations | Toronto | May 15

o MER Conference | Chicago | May 19 – 21

o AIIM Seminar – Resetting SharePoint Expectations | NYC | May 20

o AIIM Seminar – Resetting SharePoint Expectations | DC | May 22

o ARMA Canada | Ottawa | June 8-11

IBM Lunch & Legacy Data Cleanup Event Series – coming to a city near you!

o Lunch & Learn| Dallas | May 6

o Lunch & Learn | Kansas City | May 14

o Lunch & Learn | St. Louis | May 28

o Lunch & Learn | DC | TBD

o Lunch & Learn | Buffalo | TBD

o Lunch & Learn | Rochester | TBD

Engage! Join the Conversation!
Compliance, Governance and Oversight Council

Join the CGOC!

- Forum of over 2,000 corporate legal, IT, records and information management professionals. CGOC conducts primary research, has dedicated working groups on challenging topics, and hosts meetings throughout the U.S. and Europe where practice leaders convene to discuss discovery, retention, privacy and governance.

- Mission: To provide executives the opportunity to benchmark and exchange case studies; its practice groups focus on discreet areas in preservation, retention, and information governance to deliver work products that help our members best approach the challenges in maintaining best-in-class programs.

Online and in person events
Regional and International summits
Published materials

Benchmark Report on Information Governance in Global 1000 Companies (CGOC Publication, 2010)

**Information Lifecycle Governance Leader Reference Guide** (CGOC Publication, 2012)

German State Orders Businesses to Close Facebook Pages Over Privacy and Data Protection Issues (CGOC Advisory, 2011)

Elements of the Modern, Executable Retention Schedule (Work Product of the CGOC RIM Practice Group, 2011)

Only IBM ILG Solution Improves Information Economics

IBM Information Lifecycle Governance
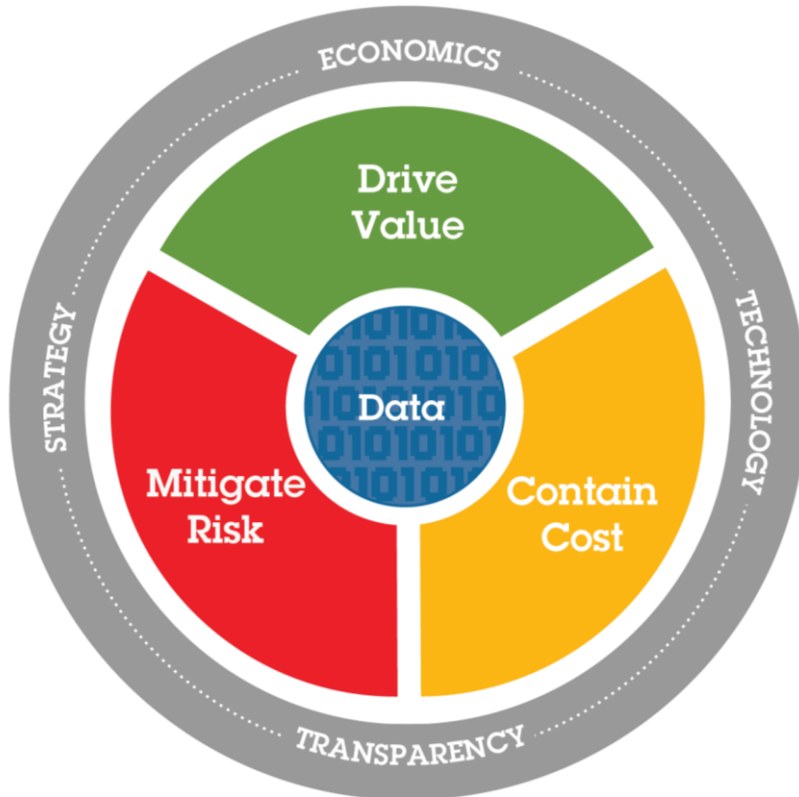
ILG lowers the total cost of information while increasing the value derived from it:

Eliminate unnecessary cost and risk by defensibly disposing of data debris

Align cost to value through value-based archiving and tiering

Reduce information risk by instrumenting privacy, e-discovery and regulatory policy across the data environment

Enable business to **realize information value** as context erodes with analytics-in-place, content management and collaboration

Please note

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion.

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.