Information Lifecycle and the Actionability Test

By scoring information generated internally and externally, and assessing its importance and maturity, organizations can better understand how and where to store key data for quick access and applicability to strategic planning.

Executive Summary

For the past several years, enterprises have been drowning in a deluge of data, and the sheer volume is growing exponentially. That translates to downstream reporting environments, where, for instance, data warehouses have become behemoth data dumps containing multiple terabytes and even petabytes of information. While it is fortunate that new technologies have made it possible to store this volume of information at a significantly reduced cost, professionals who are dealing with this data deluge still suffer from information overload.

Information overload is caused by too much data created by too many sources. In fact, the sheer number of sources is increasing geometrically, as is the volume and complexity of information (i.e., formats, metadata, frequency, etc.) available for analysis. (See Figure 1, next page, for illustrative examples.)

For years, businesses have focused on moving nonoperational data into archives through an orderly automated process or archiving the contents of retired systems. But rarely have they used information lifecycle techniques to ensure that what is easily accessible in their operational systems is actionable. This white paper defines a process to ensure exactly that.

According to IDC, the volume of information available will grow 50-fold over the next several years (see Figure 2, next page). But data volume is only part of the problem. The sheer number of tables, overlapping sources, references to differing formats of information and other issues are causing acute problems for organizations that utilize traditional techniques for analysis and reporting. The unbridled growth of warehouses as a means of servicing all constituencies has reached its limit of usefulness, and a better approach to managing the growth of data sources is required, such as allowing warehouse environments to grow organically to service the needs of all. This unmanaged organic growth is causing difficulty to those striving to glean insight from or discover anything useful in these behemoth warehouses.

While big data makes it possible to store and retrieve insight from large data pools at an accelerated rate, it does not help people focus on actionable information nuggets. The economist...
Herbert A. Simon had it right when he wrote, “In an information-rich world, the wealth of information means a dearth of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.” It is exactly such focused attention that organizations need to properly derive insights from the explosion of data. And, most importantly, this is where an appropriate information lifecycle program comes in handy.

Most information lifecycle programs tackle basic needs, such as segregating data into several broad buckets. They do not generally cull out organizational trivia from warehouses, under the premise that this information may be needed at some future date.

According to Omar Tawakol, CEO of BlueKai, the data warehouse’s historical premise was to provide a pool of storage where trustworthy data could be parked for analysis. However, the explosive growth of data has made this pooling of data superfluous; the enormity of the end data state made the warehouse virtually unusable.

An example of an actionable data warehouse is the one used by the U.S. Internal Revenue Service, which manages a historical base of 230 million annual tax returns and 50 billion tax-

The Reality of Information Overload

<table>
<thead>
<tr>
<th>42%</th>
<th>11</th>
<th>15.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organizations that say the paper records they are generating continue to increase. Source: AIIM survey of 512 AIIM members conducted in January and February, 2013.</td>
<td>Seconds to generate one petabyte of information, or the equivalent of 13 years of high definition video. Source: Intel Corporation</td>
<td>Millions of dollars in health insurance premiums overpaid for fiscal year 2012/13, half of which were to cover workers who were no longer employed. Source: Wisconsin Legislative Audit Bureau</td>
</tr>
</tbody>
</table>

Figure 1

The Growing Data Universe


Figure 2
related telephone requests in a single warehouse. Queries made against this warehouse help to:

- Estimate the U.S. tax gap.
- Predict identity theft, fraud and other tax issues.
- Model financial risks.
- Measure taxpayer burdens.
- Simulate new tax laws and their effects on tax behavior.
- Optimize the IRS staffing model.

Information Lifecycle Maturity

Information lifecycle management (see Figure 3) can be viewed in terms of levels of maturity:

- **Immature**, where the information lifecycle management (ILM) process begins.
- **Aware**, where in most cases basic ILM maturity emerges by partitioning data onto static (not logged for updates) partitions.
- **Basic managed**, where a basic ILM program is established.
- **Advanced managed**, where actionable information is managed.
- **Governed**, where the ILM program is governed.

Understanding Data Consumption

Organizations should assign a status to the candidate information to be managed by an information lifecycle program. These information priority statuses are as follows:

- Information that is actionable and routinely used for decision-making, such as sales information, customer activities, working capital availability, disputed invoices stuck in receivables, consumer sentiment, market capitalization, fixed and variable costs of operations, competitive insight and innovation programs that help differentiate an organization from its competitors.
- Information that is actionable for a specific business need that may be routine required, such as changes in long-term debt portfolios, changes in credit ratings, changes in market valuation, caustic social media coverage and changes in the regulatory climate.
- Information that is actionable for a short timeframe and may be historical trivia shortly afterwards, such as tracking the construction of a new plant, monitoring the launch of a new product, addressing raw materials reject rates from a supplier, monitoring spikes in healthcare costs and monitoring online groups for entries made by a disgruntled employee.
- Information that is operationally relevant but not necessarily required as a primary source for decision-making, such as the dispensation of contact center calls, the safety reports of elevators and the detailed attendance logs of employees.
- Information seen as historical trivia, such as long-lost organizational contacts, building maintenance reports for divested facilities, old detailed call detail records, ancient detailed journal entry logs that no longer align with the

<table>
<thead>
<tr>
<th>Information Maturity Lifecycle Stage</th>
<th>Description of Stage</th>
<th>Typical Stage Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Immature</td>
<td>Database tables grow linearly with only a fraction of their contents usable.</td>
<td>The information lifecycle management journey normally begins here.</td>
</tr>
<tr>
<td>Aware</td>
<td>A need for lifecycle is present.</td>
<td>The utilization of database partitioning is normally a first activity in satisfying ILM needs.</td>
</tr>
<tr>
<td>Basic Managed</td>
<td>A repeatable Information lifecycle management program is present.</td>
<td>Tools are introduced to orchestrate the movement of information through the ILM processes.</td>
</tr>
<tr>
<td>Advanced Managed</td>
<td>The information lifecycle management program is directed to deal with the deluge of information, much of which is never utilized.</td>
<td></td>
</tr>
<tr>
<td>Governed</td>
<td>A proactive governance body orchestrates the information lifecycle processes.</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3
chart of accounts and competitive insight on defunct companies.

**Understanding Information Gleaned**

Candidate information for an organization’s information lifecycle program can be obtained from a variety of sources:

- **Indirect sources**, such as Web sites maintained by partnering organizations, distributors and retailer reports, advertisers, regulators and watchdogs.

- **Direct sources**, such as applications, logs and other information from Web sites, logs and other information from mobile facilities and logs and other information from call centers that your organization originates.

- **Externally sourced information**, such as data collected from RFID tags, mobile devices, social media, modifications made on tracked Web sites, traditional and nontraditional news media, public relations bulletins, acquired intelligence, affiliate referrals, etc.

Each of these sources may contain superfluous information, or may signal a change to the status quo – which would require a swift determination of whether action is warranted and, if it is, execution of the appropriate coordinated action to protect, extract, originate or capture value for the organization (see Figure 4).

Data can originate from:

- A **log**, or formatted activity logs made available from a call center, Web site, RFID, mobile device or other facility that records its activity into a log.

- A **message**, received from an internal or external source, such as a SWIFT message, an EDI transaction or other message based sources.

- A **stream**, or an asynchronous source of information that streams information to a specific listening target, such as Twitter, a market ticker, newswires and other sources.

- A **document**, or any form of object that is available as an information source, such as an XML file, a public relations briefing, a weather map, a schematic or other non-tabular information source.

### Information Source Formats

<table>
<thead>
<tr>
<th>Information</th>
<th>Delivery Models</th>
<th>Possible Source Formats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indirect</td>
<td>Partner Web</td>
<td>Log, Message, Stream, Document</td>
</tr>
<tr>
<td></td>
<td>Distributor Reports</td>
<td>Log, Message, Stream, Document</td>
</tr>
<tr>
<td></td>
<td>Retailer Reports</td>
<td>Log, Message, Stream, Document</td>
</tr>
<tr>
<td></td>
<td>Advertisers</td>
<td>Log, Message, Stream, Document</td>
</tr>
<tr>
<td></td>
<td>Watchdogs and Regulators</td>
<td>Log, Message, Stream, Document</td>
</tr>
<tr>
<td>Direct</td>
<td>Internal Sources</td>
<td>Log, Message, Stream, Document</td>
</tr>
<tr>
<td></td>
<td>Web Sites</td>
<td>Log, Message, Stream, Document</td>
</tr>
<tr>
<td></td>
<td>Mobile Facilities</td>
<td>Log, Message, Stream, Document</td>
</tr>
<tr>
<td></td>
<td>Call Centers</td>
<td>Log, Message, Stream, Document</td>
</tr>
<tr>
<td>Externally Obtained</td>
<td>Mobile Facilitated</td>
<td>Log, Message, Stream, Document</td>
</tr>
<tr>
<td></td>
<td>RFID</td>
<td>Log, Message, Stream, Document</td>
</tr>
<tr>
<td></td>
<td>Social Media</td>
<td>Log, Message, Stream, Document</td>
</tr>
<tr>
<td></td>
<td>News and PR Media</td>
<td>Log, Message, Stream, Document</td>
</tr>
<tr>
<td></td>
<td>Acquired Intelligence</td>
<td>Log, Message, Stream, Document</td>
</tr>
<tr>
<td></td>
<td>Affiliate Referrals</td>
<td>Log, Message, Stream, Document</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>Log, Message, Stream, Document</td>
</tr>
</tbody>
</table>

Figure 4
• A tabular source of information, normally housed in databases and originated from sources internal to an organization (internal ERP systems, spreadsheets, CRM systems, etc.) or external sources (such as vendors, customers and industry associations).

The information lifecycle management program must be able to decipher which of these data sources are usable for long-term analysis and which ones are relevant only for situational use and temporal event detection. For example, the session statistics from a Web log may be usable for long-term use; however, the Web log detail may be usable only until it is transformed. At this point, this data may be very well a candidate for archiving.

A Model for Deciphering Information Actionability

Information captured for analysis can be used for managing operational risks, diffusing disruptive events, enhancing sustainable value and defining enterprise strategies. The information lifecycle must determine which of the data captured about customers, media, competitors, markets, financing and geographies will be useful in identifying patterns or deriving insight for the organization.

The business information model should serve as a primary tool for determining placement of information classes in the information management lifecycle (see Figure 5).

Scoring information consumed by an information lifecycle program prioritizes information deemed mandatory for operational use, its importance in managing operational risks, its importance in enhancing sustainable value (capturing, extending or creating sustainable value), defining organizational strategies or diffusing disruptive events.

Data Attributes and the Information Lifecycle

The information lifecycle program should consider the following characteristics of data:

• The ability to recreate information should it be incorrectly or inadvertently destroyed.
• The veracity of sources transformed into what is being archived.
• Actionability of the candidate information.
• The information's priority status.

The scoring will facilitate whether the information is:

• Actively maintained (and thus should reside in active online partitions).

Business Information Model

![Figure 5](image-url)
• Required regularly but historical in nature (static online partitions).
• Required, but has a low information lifecycle pool scoring (near-line storage).
• Required infrequently but must be retained for regulatory or other needs (archived).
• Is no longer required for regulatory needs and is not accessed (to be destroyed).

A viable approach for utilizing data attributes in an information lifecycle program is automating the lifecycle scoring algorithm (see Figure 5) within the information lifecycle orchestration application.

### Assigning Scores for Lifecycle Placement Data

The primary reason for focusing on the actionability of information—based on its format and ease of recreation, its age and consumption pattern—is to derive a mechanism to assign a score that can be used to place information into the buckets of storage managed by an information lifecycle program.

It is important to note that there will be differing retention requirements for information utilized by an organization, which may serve as information lifecycle overrides. For example, vested

### Scoring Data Attributes

<table>
<thead>
<tr>
<th>Factors in Scoring the Placement of Information</th>
<th>Managed Information Lifecycle Pools Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Active Online (100-86)</td>
</tr>
<tr>
<td>Operationally Critical (1=not critical, 15=highly critical)</td>
<td></td>
</tr>
<tr>
<td>Information Priority Status (1=organizational trivia, 15=highly actionable)</td>
<td></td>
</tr>
<tr>
<td>Analytic Criticality (1=not required, 10=critical)</td>
<td></td>
</tr>
<tr>
<td>Able to Recreate (1=easily, 5=not able)</td>
<td></td>
</tr>
<tr>
<td>Veracity of Sources (1=Low, 10=High)</td>
<td></td>
</tr>
<tr>
<td>Storage Platform (1=other, 2=Mainframe, 3=portal, 4=database, 5=big data)</td>
<td></td>
</tr>
<tr>
<td>Regulatory Availability (1=not required, 15=critical)</td>
<td></td>
</tr>
<tr>
<td>Age of Data (1=&gt;3 years, 3=1-3 years, 5=&lt;1 year)</td>
<td></td>
</tr>
<tr>
<td>Information Format (5=Tabular, 4=Document, 3=Message, 2=Log, 1=Stream)</td>
<td></td>
</tr>
<tr>
<td>Used in Active Program (1=Not used, 5=actively used)</td>
<td></td>
</tr>
<tr>
<td>Frequency of Access (1=rarely if ever, 10=regularly)</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5
pension recipient information must be saved for employees who have left an organization because they will be receiving pensions at a future date. Thus, the information about them has operational (funding needs for pension plans) and regulatory (pension regulations) retention requirements that will override some information lifecycle placement scores — i.e., you cannot destroy data on potential pensioners who are less than 60 years old, vested and have not yet received pensions.

Using a matrix similar to the one in Figure 5, a numeric scoring will be derived, which should be used to determine the appropriate placement of data within the information lifecycle. To control the scoring of information for use in analytics and regulatory retention requirements, a simple scoring algorithm can be deployed to enter data into the organization’s information management lifecycle management software.

Just as important as moving data through the lifecycle into its archival target is the destruction of data no longer required nor usable by the organization. There are countless examples of organizations that have had old records subpoenaed and used for tax and HR disputes and patent infringement disagreements, among other reasons. When information is no longer required, it should be disposed of.

ILM Implementation and Testing Approaches

The processes that govern an information lifecycle program should be automated in such a way that are repeatable and auditable (see Figure 6). These processes should ensure that information, particularly that which is moved to near-line storage or archived, can be retrieved and used and that information flagged for destruction cannot be utilized after deletion.

ILM Governance

The role of governance is simple in the information lifecycle program. Governance facilitates the scoring of the placement of information into the appropriate information lifecycle pools and informs the proper execution of the information lifecycle program.

Tasks specific to the administration of the information lifecycle program are as follows:

- Scoring the managed information lifecycle pools matrix.
- Measuring and communicating the program’s cost and the associated cost avoidance.
- Measuring the risk associated with the scoring of specific regulated and high-risk information and rescoring should an elevated risk be derived.

ILM Construction

![Diagram of ILM Construction](image)
• Measuring and communicating information lifecycle maturity model placement of the organization.

Organizations that have effective information lifecycle governance programs and high degrees of information lifecycle maturity routinely revisit their information lifecycle program and adjust the information lifecycle pools matrix as part of the annual planning cycle. This enables organizations to:

• Ensure that a partitioning program administered through an information lifecycle product (Soliq, Optim, Informatica ILM, Autonomy, OpenText, StoreTrends, etc.) exists and is regularly monitored for effectiveness.

• Review key technologies in the information lifecycle program and monitor storage, floor space, manpower and power consumption statistics; and regularly update to newer technologies where they make economic sense and are not operational disruptions.

• Regularly monitor the effectiveness of third-party information lifecycle products and align to make the capabilities of the information lifecycle program effective and make optimal use of the information lifecycle capabilities of vendors participating in the program.

Footnotes
1 IDC Digital Universe Study, December 2012.
4 Internal Revenue Service Compliance Data Warehouse case study, General Accounting Office, 2011.

About the Author
Mark Albala is the Service Line Lead of Cognizant’s Data On-Demand Service Line, an organization that provides services and products that support the effectiveness and efficiency of databases and other information platforms used for managing information as part of Cognizant’s overall capabilities in enterprise information management. A graduate of Syracuse University, Mark has held senior consulting, thought leadership, advanced technical and business development roles for organizations focused on the disciplines of business intelligence, governance and data warehousing. He can be reached at Mark.Albala@cognizant.com.

About Cognizant
Cognizant (NASDAQ: CTSH) is a leading provider of information technology, consulting, and business process outsourcing services, dedicated to helping the world’s leading companies build stronger businesses. Headquartered in Teaneck, New Jersey (U.S.), Cognizant combines a passion for client satisfaction, technology innovation, deep industry and business process expertise, and a global, collaborative workforce that embodies the future of work. With over 50 delivery centers worldwide and approximately 164,300 employees as of June 30, 2013, Cognizant is a member of the NASDAQ-100, the S&P 500, the Forbes Global 2000, and the Fortune 500 and is ranked among the top performing and fastest growing companies in the world. Visit us online at www.cognizant.com or follow us on Twitter: Cognizant.