The Art of Intelligent Data Management

Lessons for discovering, securing and exploiting the value of your enterprise content while minimizing cost and risk

An Intellyx eBook by Jason English, for Aparavi

TABLE OF CONTENTS

INTRODUCTION

Data is what a modern business lives by. But maintaining the data management status quo for a rapidly expanding data estate is driving enterprises completely out of balance. Too much content clutter makes it hard to extract real value from data – and exposes the company to untold cost and risk.

This eBook will guide you on a simple six-step path to a harmonious relationship with your company's data, based on the learnings of other everyday practitioners who became masters of the art of intelligent data management.

WHAT YOU’LL LEARN

Intelligent data management isn’t just about moving data to the cloud. Nor is it solely the domain of people titled as data architects and database managers anymore. Virtually any business professional generates, shares, uses and stores expansive amounts of data every workday, so how can we help them improve the reward-to-risk ratio on this critical data?

This guide will help you learn how to:

- Identify the opportunities and risks underneath your enterprise data.
- Understand the lifecycle of data across all of its uses and applications.
- Distinguish between data you need, and what should be thrown away.
- Successfully apply intelligent data management principles with real-world use cases.

STEP 1 – FINDING THE DATA

Knowing yourself is always the first step toward mastery.

Most organizations are awash in data, with little awareness of where it lives. Employees may know of the data within a core application, or of a shared file system in the cloud, but such knowledge is generally tribal in nature and limited to a particular role or departmental view.

Where to start: An optional initial data assessment activity can scope out the potential quantity of data in motion and at rest within the organization. A starting survey of data may not be 100 percent accurate, it can highlight the total opportunity for ROI and risk reduction, as well as help estimate project phases, cost and timeline.

The work: Discover primary storage resources for databases as well as unstructured data silos throughout the enterprise. Primary storage could exist in on-prem servers, DAS/NAS/SAN resources, cloud-based data warehouses and data lakes.

Unstructured data can also exist in endpoints including shared drives, email servers, files, emails, chats, and application data – on edge and even end user devices.

Lesson: Intelligent data management starts with shining a light on the entire distributed data estate as it exists today, since as much as 80% of enterprise data is unstructured, sitting outside of a database, and never analyzed. During discovery, one retail company found that employee usage of a high-cost cloud storage bucket for sharing internal files had increased more than 3X with no discernible business justification, and possible data exposure risk.

STEP 2 – IDENTIFYING THE DATA

Data without context doesn’t mean much.

How much do we really know about our data? Structured databases, by nature, may already offer a limited idea of their context, if they have defined schema and table names. But the much larger pools of unstructured data offer much less insight into content.

Where to start: The heart of intelligent data management practices starts with rapid and effective classification and identification of data across the enterprise, by labeling data sources and elements with metadata that provides context into how each datum should be organized and handled. This indexing should happen without needing to extract or move the data away from its sources for external processing, so data gravity is respected.

The work: Full content indexing applies metadata labels to identify the network addresses, geo-location and basic characteristics of each datum such as file names, timestamps, types and sizes.

More importantly, indexing can map metadata across dozens of other important dimensions, such as identifying private or personal data, classifying it according to regulatory status and security policies, or marking the content’s relevance to particular groups, departments and work processes.

Lesson: A technology conglomerate grew through a series of ten corporate acquisitions, and is now having trouble unifying disparate data lakes and CRM SaaS tools to support customer service and sales. They sought to normalize all of the unique data structures they inherited into an understandable, searchable whole so they could gain a 360-degree view of each account.

Using the cloud-based Aparavi platform, they scaled to process a million or more records per hour, achieving a highly compressed full content indexing of their enterprise-wide multi-petabyte data estate in under two days, without the wasted time and cost of migrating all of that data.

STEP 3 – CONDUCTING DATA HYGIENE

According to IDC, for every 1000 people in an organization, an average of $5.7M in labor costs is wasted every year searching for and not finding data.

Once the organization has a full content index with metadata in place, it's time to start tidying up. Data hygiene involves gaining an understanding of the properties of all data, and curtailing data sprawl that causes unnecessary costs, process friction and risk.

Where to start: Data hygiene usually begins with a series of searches against the enterprise's data assets, looking for things like duplicate files and orphaned data. This can be a daunting task involving manually formulating regular expressions and SQL queries, only to get back millions of potential matches.

Fortunately, an enterprise-wide content index allows even non-data-scientists to run automated searches by selecting policies that filter against dozens of metadata properties.

The work: Create data hygiene policies with predefined automations that capture the intended goals of complex searches. For instance, one policy may be to purge trash files, or delete duplicate files. The policy can basically act as a boolean of several automated searches to produce a limited list of data without human error that falls within the criteria of being a 98% match with desired criteria, or 100% identical, so further action can be taken.

Lesson: One large insurance company ran automated searches with a policy of finding and discarding data that had not been opened or used within the last three years, and ended up removing more than 30% of their data which proved redundant or obsolete. This prevented additional CapEx outlays in their on-prem datacenter, and reduced the rate of OpEx increases in the cloud.

STEP 4 – SECURING THE DATA ECOSYSTEM

Only the paranoid survive.
- Andy Grove

Many organizations embark on a data transformation initiative simply because they are required to comply with industry and regulatory standards to avoid penalties. Or perhaps they are more concerned about cybersecurity threats and data leaks than cost efficiency. That's ok, because intelligent data management can provide coverage for these risk-averse use cases too.

Where to start: Strong security event monitoring and authorization, identity and access controls are extremely useful starting points for securing enterprise data – but these tools should also inform the data management platform. Data stakeholders need advance notification of incoming threats, latent or introduced data vulnerabilities, and potential privacy or compliance issues as soon as possible, so incidents can more likely be.

The work: At a high level, compliance and security concerns share a decision workflow for data, wherever it exists on-premises or within cloud infrastructure or services.

First, organizations need to identify what data needs to be retained, whether it is essential to conducting business, or to meet the company's compliance regimes – for instance, financial data for a SOX audit needs to be held onto for 7 years, whereas GDPR statutes in Europe dictate that user data should be eliminated as soon as it is no longer needed.

Then, determine when and how data should either be securely locked down in place, or discarded. For some businesses, a master record or archive of certain sensitive data should be retained, so long as it cannot be copied or viewed by unauthorized actions. In most cases, historical and idle data is risky (and costly) to keep, and should be discarded eventually by default if it is not needed.

Lesson: A US-based healthcare firm must support HIPAA requirements for its patients, as well as PII requirements for its billing and payment systems.

By setting automated data search policies to match the current protocols for both compliance regimes, the company can rapidly identify and resolve issues such as sharing medical data without signed forms, or failing to dispose of account information when it is no longer needed. A great time and cost saving site benefit? They can almost instantly generate a report on the current status of all the data when asked.

STEP 5 – OPTIMIZING THE DATA

Chefs call it 'mise en place' – meaning everything is in its proper place to begin work.

Several unique application silos exist that move and store data – from major cloud vendor repositories, to SaaS-based productivity apps, to streaming data services, core enterprise applications, network and system-level security tools, integration and deployment pipelines, and backup and recovery tools aligned with incident management.

Where to start: Instead of requiring the business to rip-and-replace any of these essential tools, intelligent data management should fully index the data as it exists within these sources and destinations, so higher level optimization can take place.

The work: In today's distributed cloud-based and API-driven application environments, data can elastically scale even faster than its underlying infrastructure, as different tools will constantly generate more data at scale.

Smart policies can seek out areas where data sprawl is ripe for pruning. More so, it can help companies regain the concept of a 'single source of truth' in a fragmented data ecosystem. This extends the cost-saving aspects of data hygiene, of course, but more importantly, it can drastically reduce the time needed to fulfill data requests, with fewer version conflicts between all of these systems.

Lesson: A mid-sized MSP responsible for managing the cloud data estates and productivity apps of more than 100 companies, leveraged Aparavi as a multi-tenant data management platform for indexing all of their clients' data stores and productivity apps like MS365, Salesforce and Slack.

By leveraging their own best practices as automated policies within the Aparavi platform, they were able to demonstrably reduce data costs and improve data utility across clients, while keeping the data views and domains separate.

STEP 6 – EXPLOITING THE DATA

Companies that take advantage of data insights grow 30% each year. - *Forrester*

Cost and risk will increase as the organization's data estate expands, but it's important to remember that this data is also essential for the enterprise to survive and thrive. Fundamentally, we want to extract maximum value from data, whether it is used to make employees more productive, improve our strategic insight for better decisions, or deliver newer and leaner services for customers.

Where to start: Prioritize the alignment of data around the organization's most critical use cases first, then proceed to optimize other important processes. For instance, a pharma research company may prioritize machine learning, where a property insurance firm may lean on improving incident management and claims resolutions.

The goal for every use case is to create an on-demand data experience, so that high performance responses to searches and application queries will meet employees and customers where they are in real time.

The work: Get more ROI from data by optimizing its performance, reuse and resiliency characteristics. Cleaning up the data estate with tagging and optimization, and adding a fast data index already gets us much of the way there, so at this phase we use data analytics to observe the workloads stakeholders are trying to accomplish. We then take action by setting policies for copying, moving, archiving, retrieving and deleting data so it is more adaptive and responsive to those workloads.

Employees can stream on-demand data directly into their productivity apps like PowerBI or Excel. Since different locations and countries have different work and compliance requirements, migrating data to on-premises private clouds or nearby public cloud data centers can deliver on real-time performance needs.

Lesson: A global multi-brand manufacturer improved the performance of their business planning and analytics workloads while improving its responsiveness to compliance requests. Using Aparavi's data management platform, the firm mapped workloads directly to data where it resides, and prioritized cloud data lake investments only when necessary for a particular division or region. The firm now has a more resilient data estate, while avoiding most of their expected cloud migration costs.

THE INTELLYX TAKE

Intelligent data management is no longer a lost or secret art, available only to the most technically advanced data architects and scientists.

If planned correctly, this data transformation should be an easy process to socialize and share with all stakeholders in the organization, without setting a high technical bar or requiring intensive retraining. With real visibility and knowledge, everyone can better understand the nature of the data they interact with every day.

ABOUT THE AUTHOR

Jason "JE" English (@bluefug) is Principal Analyst and CMO at Intellyx, a boutique analyst firm covering digital transformation. His writing is focused on how agile collaboration between customers, partners and employees can accelerate innovation.

In addition to several leadership roles in supply chain, interactive and cloud computing companies, Jason led marketing efforts for the development, testing and virtualization software company ITKO, from its bootstrap startup days, through a successful acquisition by CA in 2011. JE co-authored the book Service Virtualization: Reality is Overrated to capture the then-novel practice of test environment simulation for Agile development, and more than 60 thousand copies are in circulation today.

ABOUT APARAVI

Aparavi is the data intelligence and automation platform helping organizations find and unlock the value of data, no matter where it lives. Aparavi cloud-based platform with deep intelligence finds, automates, governs, and consolidates distributed data easily. We ensure secure access for modern data demands of analytics, machine learning, and collaboration connecting business and IT to transform data into a competitive asset. Aparavi is a privately funded company headquartered in Santa Monica, Calif. For more information, visit aparavi.com, and stay informed by following Aparavi on LinkedIn and Twitter.

©2021 Intellyx LLC. Intellyx retains complete editorial control over the contents of this document. At the time of writing, Aparavi is an Intellyx customer.