LOGISTICS FOR DATA

Getting battlefield data to the right place at the right time: The mess versus the mesh.

by Thom Hawkins and Andrew Orechovesky

Part Two of a three-part series.

The Great Library of Alexandria, which flourished in Egypt during the Ptolemaic dynasty of the third and second centuries B.C.E., took any manuscripts found on ships that docked in its port, copied them, and then returned the copies to the ship while retaining the originals. Viewing an original and a copy as interchangeable relies on the notion that the parchment's information provides the value, not its medium—a giant step forward toward our modern, digital world.

While the first entry in this series ("Logistics for Data," Army AL&T Fall 2023) discusses the demand signal for data, this entry will focus on inventory and warehousing of data. Making use of a physical resource, like an ancient parchment or a shiny new case of ammunition, requires geographic co-location and a limit on the number of simultaneous users. In contrast, an unlimited number of people can use a digital resource remotely at the same time. Even still, as we discuss in this article, the library is an apt metaphor, and ties data to the same physical logistics as something like ammunition.

DATA STORAGE

Army logisticians seeking to replenish a unit's ammunition must know where the ammunition is stored and have access to view the current inventory, specifically the desired items. The same applies to data (e.g., tactical or strategic data products) stored in a data repository, or "data platform."

The location of where data is used has evolved in recent years. A structured database, often referred to as a "data warehouse," is co-located with the application that uses that data. This is inefficient because the same data may be needed by different systems, but each system may have its own source or frequency of update, resulting in the potential for discrepancies. Databases can use a process called federation to create links that synchronize those data points. However, this can lead to chaotic and disorganized connections that databases cannot maintain, especially as applications change.

From data warehouses, industry's response to the limitations of data warehouses was data lakes, which are a common pool of raw data, only structured or federated as needed to serve set purposes. Data lakes have their own set of limitations, including the potential for disparate or conflicting sources of data. More recently, data fabric, which acts as a common data warehouse, has gained popularity. Instead of assigning each application a database with its own data, all applications rely on data in the common

database. This solution offers more efficiency, but also additional challenges. In a degraded network environment with intermittent or unstable connectivity, systems may not have access to the shared data resource. Novel data also may require an update to the structure of the fabric, just as with a database or data warehouse.

Despite all of these various data "buildings"—the term "data architecture" can generalize to include the flow of data through systems and the inventory schema, expanding beyond the storage schema. Data architects can achieve data individuality, cataloging, discovery, accessibility, governance, analytics and retention periods with a well-planned and optimized data storage capability. Data individuality or uniqueness is one of the more essential characteristics; duplicate data brings on unwanted technical debt in the form of poor system performance, storage costs, data confidence concerns, data lineage problems and, eventually, archival issues. While digital forms make no distinction between original and copy, duplication is the process of replicating data and storing it separately, just as Alexandrian librarians did thousands of years ago.

MESS VS. MESH

The problem with any physical or digital warehouse is decreasing efficiency with more content. Like Jorge Luis Borges' fictional "Library of Babel," which contained not only every book but every possible book, the vastness of the collection degrades access to either hardware or data. Amazon mitigated this problem in its warehouses through an inventory method called "random stow." When goods arrive in the warehouse, Amazon employees shelve them wherever there is available space, with both the item tag and bin tag scanned and linked. This reduces time wasted adjusting allocated space to keep like items together. When items are picked for delivery, employees follow automated guidance to the closest instance of an item, thus reducing travel time and effort to retrieve the next listed object within the warehouse.

Libraries have adopted the same policy. Faced with the increasing volume of books and the cost of expanding publicly accessible storage space, larger libraries have made using the catalog, rather than browsing the shelves, the primary method of locating a book. The book's location no longer matters—including behind locked doors in the building or in cheaper offsite storage. The University of Nevada, Las Vegas' Lied Library competes for attention with the fountains at the Bellagio with a glass-walled warehouse, where similar-sized books are stored together in bins to minimize wasted space. A user clicks a button to send a request for a book to the Lied Automated Storage and Retrieval system, which sends a robotic arm to fetch the associated bin and deliver it to a retrieval desk.

Similarly, the Army is implementing a data mesh. A mesh links producers and consumers of data via a central catalog that lists the available data products. By the library analogy, the catalog connects a reader to a published book. According to Data.world, a data product is defined as "a reusable data asset, built to deliver a trusted dataset, for a specific purpose," and thus a book is a physical version of a data product. Other examples of data products could include an operational order, a firing target and its coordinates, or a research dataset. One data product could derive another, just as a nonfiction author may

OLD-SCHOOL DATA
Data management and storage often encounter similar problems to libraries—how do you best store vast amounts of data while making it easy for users to access? (Photo by Skitterphoto, Pexels)

consult references while writing a new book. Those references and their citations prove crucial for data trustworthiness and security, making the data product's lineage traceable to its source.

The data mesh does not concern itself with data storage, so anything can store the data including a warehouse, lake or fabric. This construct makes it easier to share data across organizations that may have their own ways of storing data, without the need to change those methods. Once a potential consumer identifies a data product in the data catalog, they can request it from the producer. For the Army, different domains, separated by subject, organization or area of operations, do not have to store their data in the same way if it can be shared upon request.

The speed and security that catalogs offer support timely enhanced data-driven decisions.

DATA INVENTORY

Metadata describe an individual data product and is stored in a data catalog. While metadata standards vary, most include author, subject, data domain, classification, releasability, temporal (time) coverage, spatial (location) coverage, confidence, lineage and governance policies. A well-crafted data product will contain each piece of metadata a user can employ to discover the data product within the catalog of all products. Metadata extraction and cataloging for each data product can begin once data has been ingested into the data platform.

Separating the data catalog from the data warehouse, although inefficient, provides an added layer of security by isolating one system from the other. The standalone digital data catalog system offers efficiencies such as fast searches, categorization, location links, data relation links and restriction tags for sensitive data products. The speed and security that catalogs offer support timely enhanced data-driven decisions.

DATA RETENTION

Just as book retention is a hot topic among librarians, who debate the criteria used to cull their collections, data retention can also be contentious, with some declaring that all data is perishable and others wanting to hoard the data forever. The problem is that data may be used differently by various stakeholders. For some, only the most current data is relevant; others want to evaluate data trends over time. For example, data ingested into a data platform will be normalized to align with a particular standard (format, scale), and each new update will overwrite previous data, because the data platform's utility is to provide the latest data for decision or action. However, at the same time, updating machine-learning algorithms requires data in its raw, unadulterated form, including how that data has changed over time.

CONCLUSION

When a library makes a decision to move some books to offsite storage, it does so based on a prediction about how frequently a particular book will be consulted. Data platforms may make that same calculation based on considerations of data usage and accessibility. Data platforms may lose the ability to synchronize due to denial or degradation of signal, just as area denial may interrupt a supply line.

Parchments arrived in ancient Alexandria by ship, but modes of transportation for data have changed quite a bit in the intervening millennia. Modern data transportation and synchronization will be discussed in the third and final article of this series.

For more information, contact Thom Hawkins at ***jeffrey.t.hawkins10.civ@army.mil****.*

THOM HAWKINS is the team lead for data architecture and engineering with Project Manager Mission Command, assigned to the Program Executive Office for Command, Control and Communications – Tactical at Aberdeen Proving Ground, Maryland. He holds an M.S. in library and information science from Drexel University and a B.A. in English from Washington College.

ANDREW ORECHOVESKY is a senior systems engineer for the data architecture and engineering team within Project Manager Mission Command, assigned to the Program Executive Office for Command, Control, and Communications – Tactical at Aberdeen Proving Ground. He holds a Doctor of Science from Capitol Technology University, an M.S. in cybersecurity from the University of Maryland, Baltimore County and a B.S. in information technologies from the University of Phoenix.