Aug 10th, 12:00 AM

Alternative Approaches to Data Architecture and Data Governance: Graduate IT/IS Curriculum Considerations

Diane Murphy
Marymount University, dmurphy@marymount.edu

Michelle Liu
Marymount University, xliu@marymount.edu

Laura Vera
Marymount University, lcv54996@marymount.edu

Alternative Approaches to Data Architecture and Data Governance: Graduate IT/IS Curriculum Considerations

Emergent Research Forum (ERF)

Diane Murphy
Marymount University
dmurphy@marymount.edu

Xiang Michelle Liu
Marymount University
xliu@marymount.edu

Laura Vera
Marymount University
lcv54996@marymount.edu

Abstract

With data architectures evolving from data warehouses to data lakes to data fabric, there is a pressing need for organizations to upgrade their data management and data governance strategies. A workforce gap appears inevitable as knowledge and skills in these emerging technologies is currently not emphasized in the IT/IS discipline. From an educator's perspective, the question becomes whether we are preparing our students for this increasingly significant workforce need. A strategic curriculum model to determine when and how to incorporate new technology coverage into the curriculum was applied to the potential implementation of these emerging concepts into our curriculum. The authors examined the alternative approach to data architecture, data management, and data governance, and proposed a new certificate in Data Architecture and Governance to be inserted in the graduate IT/IS curriculum at their school.

Keywords

Data lakes, data fabric, DataOps, IT/IS curriculum, graduate-level, certificate.

Introduction

Data has become a significant asset in most organizations, from commercial businesses, to government agencies, to educational institutions. However, constant changes in the volume, variability, veracity, and velocity of this data are putting pressure on organizations to upgrade their data management and governance techniques to maximize value to the organization and to keep costs under control. Three generations of data management are postulated: data warehousing as the first generation, data lakes as the second, and data fabric as the third (Wells 2019). In addition, cloud computing is significant as organizations try to meet both data growth and increased data access. Following trends in other parts of the technology world, new agile techniques (DataOps) are also being implemented (Harrington 2021).

As an academic institution, how do we ensure that our technology degree offerings stay current with workplace needs, ensuring students have the knowledge and skills necessary to meet the emerging data management challenges of tomorrow? Most educational institutions today offer database courses, but the focus is largely on the first generation of data management - the traditional relational database model and, perhaps, the structured data warehouse. There are programs that offer courses in data science but their primary focus is on the analysis of data, not its management or governance.

We explore whether we should upgrade our current graduate IT/IS curriculum to address the evolving data architecture and governance domain, and if yes, through what channel. We apply a strategic curriculum model to determine when and how to incorporate new technology coverage into the

curriculum (Liu and Murphy 2012). We first examine these new data concepts then use the curriculum decision model to determine the best implementation strategy.

Alternative Approach to Data Architecture and Data Governance

The Data Lakes Context

The data lake is a data repository created by ingesting many types of datasets from various sources and storing the data in its native format, generally in the cloud (Llave 2018). The data lake then facilitates on-demand data analysis without the need for the extensive pre-processing (schema on write) associated with data warehouses. User requirements are not completely defined when the data lake is implemented, recognizing the constantly changing technology, data, security and usage (Murphy and Forbes 2022).

Datasets come in many forms including raw data, streaming data, application data (including in existing databases or data warehouses), textual data, or archived data. In the data lake each dataset is associated with a set of extended metadata, usually added on ingestion to describe its lineage, form, format, and any relationship to other datasets. The metadata is later used to interpret the data, to access the data, and to maintain query effectiveness across multiple datasets (schema on-read).

Data lakes are designed to manage the increasing data volumes, the variety of data encountered (structured, semi-structured, and unstructured), the quality of the data (veracity), as well as the high velocity of data (expanded availability, complexity, and change). At the same time, they must effectively meet the advanced ad-hoc needs of business intelligence, machine learning, and artificial intelligence. Cloud technology has become a significant component of data lakes (Denny-Gouldson 2017). One major consideration is the elasticity of both the storage and computing resources enabling the data lake to grow with more and larger datasets but maintaining performance as queries increase in complexity. The cloud also allows access from anywhere and many cloud providers offer a variety of technology tools.

The Data Fabric Context

According to Gartner, the data fabric is one of the top technology trends for 2022 (Beyer et al. 2021). The data fabric goes further than the data lake by not ingesting data at all but leaving it where it is, weaving together data from multiple sources. In addition, it continuously analyzes user activity and feeds that back into its metadata to optimize for reuse. Like the data lake, it requires no preprocessing of the datasets. However, metadata must be available on each of the data assets, in a metadata repository or data catalog.

The data fabric uses machine learning techniques to develop a knowledge graph to map usage including who is using the data, how they are using it, and to monitor changes to this usage on an on-going basis. It analyses users' behavior and optimizes ease of use by creating standardized data access mechanisms. Unlike the data lake with its defined border, the data fabric can be used to access data stored in multiple locations, on-premise, on multiple clouds, or on third-party application providers, with access being managed from the active metadata repository. There is no need to replicate any dataset because all of the integration happens at the computation level and not at the data storage level, so connecting the data wherever it resides. The analogy is to a blanket that conforms to the structure that it covers.

The data fabric focuses primarily on those parts of the metadata that describes the data itself and gradually builds on this initial data by becoming increasingly more informed about the data and its usage. Unlike the data warehouse and the data lake, the data fabric takes the data from where it is and relies on intelligent metadata to simplify the use of these resources to simplify access and drive business value.

DataOps

Data management has become more challenging as data is now more diverse, and in more locations, while data users want results from trusted data sources delivered faster, as often results will be used for important decision-making where time is of the essence. This emphasis on speed is no different from other parts of the technology field where DevOps has emerged as a framework to support the need for faster and more reliable applications development (Wiedemann et al. 2019).

Similarly, DataOps techniques are evolving to simplify data management processes and focus data management on the fast, flexible, and trustworthy delivery of results. It encompasses similar approaches used in agile application development and DevOps (Friedman and Thanaraj 2021). Collaboration is key between all stakeholders throughout the data pipeline, including providing an infrastructure for the sharing and version control of the metadata. Automation is also key to improve timeframes for data management and to minimize the costs of managing the increasingly large data repositories.

Data Governance

IT/IS data governance has drawn attentions from both higher education and industry practitioners (Abraham et al. 2019; Tiwana and Konsynski 2009; Topi et al. 2017). Data repositories must be subject to strict data governance policies and procedures if the organization is to take maximum advantage of data and trust any results derived. In addition, to obtain the most flexibility and performance from the data, users must be able to find the right data, at the right time, and integrate reliably across multiple data sets.

The data governance policy and implementation rules must balance the management and control of the data itself with the needs of the potential users of the data repository to ensure trust, lineage (traceability), privacy and data security. However, data governance processes must not have a significant adverse effect on the performance of the data repository, including the ingestion of the data from multiple sources and the discovery processes. As data lakes and data fabric become more valuable to organizations, so the need for effective data governance increases in complexity.

The New Data Management Workforce

The new second and third generation data architectures (data lakes and data fabric) and the new data management frameworks such as DataOps pose new challenges to organizations - there must be skilled and knowledgeable personnel. The Computing Technology Industry Association (CompTIA) recently published a whitepaper on the need for additional data skills and announced an entry level certification (Data+) in spring 2022 (CompTIA 2021). They introduce the concept of data teams and various roles, including data analyst and data architect, and acknowledge that most companies are seeking mid-level professionals in these areas. They note that the increased need for data skills presents a pipeline problem.

Data Management and Data Governance in the IT/IS Curriculum

How do we educate mid-level personnel in data management and governance with the data architecture of today and tomorrow? Most graduate IT/IS programs include at least one course on relational databases with the same, or an additional, course covering data warehouses. Is this the right time to also incorporate data architecture, management, and governance? A separate domain or part of the existing courses? To answer these questions, we applied an existing curriculum model (Liu and Murphy 2012) to help us make an informed decision as for "when" to incorporate these topics. In the model, several "forces" (i.e., factors) were integrated as a foundation for making the "when" decision.

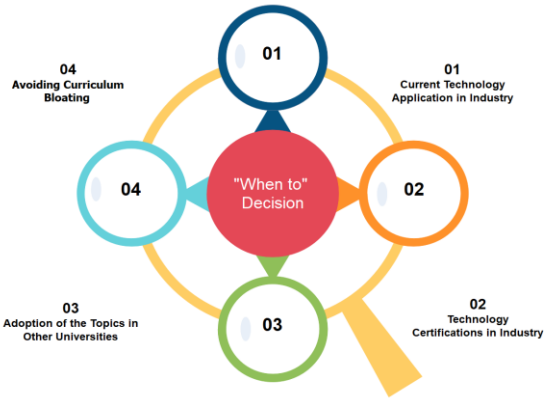

Figure 1. Strategic Model to Make "When" Decisions

We focused on four factors: current technology status in industry, technology certifications, adoption of the topic in other universities, and avoiding curriculum "bloating".

Current Technology Application in Industry

A new topic is considered higher priority if it is in use in by multiple industries. Data lakes are increasingly being adopted. For example, biotech and health research enterprises have used the data lake ecosystem to advance their operations such as increasing drug production, improving traceability of the materials throughout the manufacturing process, and improving yield (Maroto 2020). Data lakes have also been implemented by higher education institutions to improve students' academic achievement and experiences (Nanjiani 2020). The oil and gas industry were one of the earliest adopters of data lakes in the cloud, increasing production 20+ times (Thuma 2018). Finally, government agencies, such as the Department of Defense (DoD), are using data lakes in their digital transformation initiatives (Bullman 2020). Implementations of data fabric and DataOps are also becoming commonplace.

Technology Certifications

The availability of technology certifications by reputable organizations is considered important and there are a small number of existing certifications in data management and data governance. Moreover, the announcement of the CompTIA Data+ is significant in technology certification landscape. Sample certificates are summarized in Table 1.

Certification	Organization
Certified Data Management Professional (CDMP)	Data Management Association International (DAMA)
Certified Data Professional (CDP)	Institute for Certification of Computing Professionals (ICCP)
Data Governance & Stewardship Professional (DGSP)	Institute for Certification of Computing Professionals (ICCP)
Certified Public Sector Data Governance Professional (PSDGP)	Institute for Certification of Computing Professionals (ICCP)

Table 1. Sample Data Management and Data Governance Certificates

Adoption of the Topic in Other Universities

Master's level specializations/certifications in data management and governance are beginning to appear but are not commonplace. For example, Northeastern University offers a M.S. program in Data Architecture and Management. A small number of universities offer data management graduate certificates including Villanova University and SUNY Buffalo State College. However, most of the programs available are from independent training vendors, teaching bootcamps to meet the certifications.

Avoiding Curriculum Bloating

This factor is important in any decision to introduce new topics. For competitive reasons, we do not want to increase the credit requirements for the program. One of the strategies is, however the introduction of certificates that can be taken either as stand-alone or as part of the electives for a master's degree. This gives our master's students options and also allows for the reskilling of IT/IS professionals currently in the workplace in this domain.

Our Curriculum Decision

We decided, based on the model, that it was time to introduce a graduate certificate on data architecture and data governance and that it would serve students well in the competitive job market. Our initial curriculum for the planned Data Architecture and Governance certificate is envisaged as four courses in

the following areas: data architectures; database technology; data governance; and data analytics. Two of these courses exist but will be updated and two would be newly developed. Given our agile curriculum process, we believe we can implement this certificate in the next 12 month.

Conclusion

Data, its management, governance, and analysis, is increasingly important as organizations move to data-driven decision making, business intelligence, and improving the efficiency of their organizations. Newer data architectures and pipelines are being implemented to ensure that relevant, current and accurate data is available when it is needed to whomever needs it. Knowledgeable and skilled professionals must be available to enable this advanced data infrastructure. Employers are looking for mid-level professionals who understand the hardware, software, and security necessary to implement the data infrastructure, but who also understand the current emerging data architectures. As such, we believe it is important to introduce additional data topics in the IT/IS certificate, such as through a certificate or optional courses.

REFERENCES

Abraham, R., Schneider, J., and Brocke, J.v. 2019. "Data Governance: A Conceptual Framework, Structured Review, and Research Agenda," *Int. J. Inf. Manag.* (49), pp. 424-438.

Beyer, M., Zaidi, E., Thanaraj, R., and De Simoni, G. 2021. "Top Strategic Technology Trends for 2022: Data Fabric." from https://www.gartner.com/doc/reprints?id=1-27VRG4MN&ct=211103&st=sb

Bullman, H. 2020. "DoD and U.S. Census Bureau Are Using Data Lakes for Modernization and Greater Insights." Government Technology Insider, from https://governmenttechnologyinsider.com/the-dod-and-u-s-census-bureau-turn-to-data-lakes-for-modernization-and-greater-insights/

CompTIA. 2021. "Closing the Data Skills Gap." from https://www.comptia.org/content/whitepapers/closing-the-data-skills-gap

Denny-Gouldson, P. 2017. "Data Lakes and Cloud Computing: Data Needs to Be Continually Enriched and Augmented with Learnings," *Scientific Computing World* (157), pp. 516-524.

Friedman, T., and Thanaraj, R. 2021. "Introducing DataOps in Your Data Management Discipline." from https://www.gartner.com/doc/reprints?id=1-26JBRQMH&ct=210616&st=sg

Harrington, J. 2021. "DataOps Fundamental for Industrial Transformation," *InTech* (68:1), pp. 28-32.

Liu, X., and Murphy, D. 2012. "Tackling an Is Educator's Dilemma: A Holistic Model for "When" and "How" to Incorporate New Technology Courses into the Is/It Curriculum " *Proceedings of the Southern Association for Information Systems Conference, March 23rd-24th, 2012*, Atlanta, GA.

Llave, M. 2018. "Data Lakes in Business Intelligence: Reporting from the Trenches," *Procedia Computer Science* (138:2018), pp. 516-520.

Maroto, C. 2020. "A Data Lake Architecture for Modern Analytics and Bi." from https://www.accenture.com/us-en/blogs/search-and-content-analytics-blog/data-lake-architecture-analytics-bi

Murphy, D., and Forbes, O. 2022. "Securing Data Lakes in a Data Centric World," in: *Cloud Security Alliance*. https://cloudsecurityalliance.org/blog/2022/01/14/securing-data-lakes-in-a-data-centric-world/.

Nanjiani, N. 2020. "Building a Data Lake at Your University for Academic and Research Success." from https://aws.amazon.com/blogs/publicsector/building-data-lake-your-university-academic-research-success/

Thuma, J. 2018. "Five Classes of Use Cases for the Data Lake." from https://www.arcadiadata.com/blog/use-cases-for-the-data-lake/

Topi, H., Karsten, H., Brown, S.A., Carvalho, J.A., Donnellan, B., Shen, J., Tan, B.C.Y., and Thouin, M.F. 2017. "MSIS 2016: Global Competency Model for Graduate Degree Programs in Information Systems," 9781459354325, Association for Computing Machinery.

Tiwana, A., and Konsynski, B. 2010. "Complementarities between Organizational It Architecture and Governance Structure," *Information Systems Research* (21:2), pp. 288-304.

Wells, D. 2019. "The Continuing Evolution of Data Management." from https://www.eckerson.com/articles/the-continuing-evolution-of-data-management

Wiedemann, A., Forsgren, N., Wiesche, M., Gewald, H., and Krcmar, H. 2019. "Research for Practice: The Devops Phenomenon," *Communications of the ACM* (62:8), pp. 44-49.