DATA LAKEHOUSE FOR HIGHER EDUCATION

Voice of the CIO

Michael Berman - Former CIO
Bruce Maas - Former CIO

KOVEXA Waterstons technology + growth partners

Contents

Executive Summary: Why Universities Need a Data Lakehouse

The data lakehouse architecture provides the single most effective architecture for meeting the complex and evolving needs of the modern higher education institution. The lakehouse enables the institution to consolidate all important campus data in a single cloud-based data lake which can meet a wide range of use cases, from the traditional reporting and analytics that were provided by the legacy data warehouse, to supporting the most forward looking Artificial Intelligence (AI) applications.

The data lakehouse is vendor-agnostic and largely built on open source tools, but the Databricks Data Lakehouse offers a coherent and efficient set of integrated tools that enable institutions to start small and build to a comprehensive institutional data repository. Instead of complex, trouble-prone, and risky processes that move data around and implicitly encourage users to keep local copies of the data – which often become shadow data systems – the data lakehouse helps IT provide the data that campus users need, when and where they need it and using the tools that those users demand.

While many institutions will find that the agile advantages of a data lakehouse over a data warehouse provide tremendous benefits, the ability of the lakehouse to enable machine learning and other advanced AI techniques can provide the biggest long-term payoff. The Databricks Data Lakehouse was built by major university researchers to enable the use of AI tools. This singular ability to use a single source of data - the lakehouse - to solve problems from the simplest reporting to the most complex experimental AI techniques is unlike any other approach available today.

Information technology leaders in higher education have been faced for decades with the challenge of securely providing the right data at the right time to our campuses. We know that our systems contain vast amounts of important and sensitive information, but most of us have struggled with getting it out of our ERP and other key systems, providing good tools for analytics, assuring that the right people have access and others don't, and ensuring that a consistent “single source of truth” data can provide the insight that our campuses were promised when they invested in the development of these enterprise systems over the last thirty years.

While this was never an easy proposition, it’s gotten significantly harder in recent years. In order to improve student outcomes, boost engagement, promote enrollment, advance the research agenda, and support the essential initiatives of our institutions, we are expected to help provide visibility into multiple disparate systems that were never designed to work together. It’s not just the ERP – it’s the LMS, door access control and video camera systems, library systems, Wi-Fi data, student residence data, student life participation records, web analytics, security logs, and so on. And if we can get the cooperation to pull that data together in a single place, understanding, managing, and sharing that data across teams, and assuring that it remains secure gets harder all the time. And then, our leaders ask “What about AI? Can we apply

machine learning to our data to get new insights and predict student outcomes and institutional demands? How can we leverage LLMs to improve efficiency?"

The traditional architecture – a data warehouse – is too complex, too cumbersome, too difficult to manage to meet these challenges, because they work most effectively with structured data. With a lot of work, you may be able to pull the key data together and generate reports, but having the data in a flexible enough format to meet the needs of everyone on campus is rarely practical. Ultimately, we end up providing (or have people compile on their own) data extracts which then find their way into dozens of shadow databases – convenient for end users, but impossible to manage, and a security nightmare. For those campuses that have data scientists, this generates large new masses of data organized in new ways never anticipated by traditional data warehouse designs.

Another opportunity – and challenge – has been the development of the cloud-based data lake. Data lakes offer unprecedented flexibility to manage unstructured data and – if planned right – can be economical. They offer easy access to the machine learning tools offered by major cloud providers, and they can be used to put together data from many different sources in a single place. But navigating this data, protecting it, and correlating it with the highly structured and inflexible data warehouse environments we have built is extremely complex. And now, we've got two headaches – the data warehouse and the data lake – that both require care, maintenance, separate tools requiring separate expertise, and attention.

The data lakehouse architecture is a promising strategy that can help to simplify the organization of our sprawling data enterprise. By providing an overarching and consistent framework for managing and using all our structured, semi-structured, and unstructured data, the data lakehouse can make it possible for us to get all the data in one place – and even make it look like its in one place virtually if we want to have the flexibility of using multiple cloud providers. The data lakehouse can provide different and tailored views of the data to different campus customers, in effect simulating the experience of a shadow database while drawing live data from current and authoritative sources on demand. With this, we can establish a single framework for governing and protecting all our data. Finally, the data lake house can assist the campus with becoming "machine learning ready" by supporting the patterns and tools data scientists need and prefer.

In this whitepaper we will explore:

- **Lakehouse Architecture:** The basic components of a data lakehouse, and how they relate to traditional university data architectures.
- **Data Lakehouse in Higher Education:** Several higher education use cases supported by Databricks Lakehouse.
- **Data Lakehouse as a Foundation for Artificial Intelligence Innovation:** How Databricks Lakehouse can prepare a university to use Machine Learning and other AI technologies with its data.
- **How to Get Started:** Contact us for a conversation.

The Lakehouse Architecture

In this section we provide a high-level overview of the Databricks Lakehouse Architecture and describe how it can be used to replace and greatly expand the functionality of an institutional data warehouse. For an institution with a functioning data warehouse, the data lakehouse offers a more agile and modern architecture; and for institutions that have struggled or failed in their attempts to build a data warehouse, going directly to a data lakehouse offers a faster and more reliable path to a successful data analytics environment. We emphasize, to the extent possible, an abstract view that comprises components assembled from one or more suppliers including supportable open source options.

The tools provided by Databricks of course represent a flexible and easy framework for these components. We know that CIO's and IT architects generally prefer solutions that avoid vendor lock-in and give an institution control over its own data. One of the very attractive features of the lakehouse is that it's vendor-agnostic and can be built on a combination of commercial and open source tools, but using the well-established, integrated architecture available through Databricks is a fast and proven path to a predictable outcome.

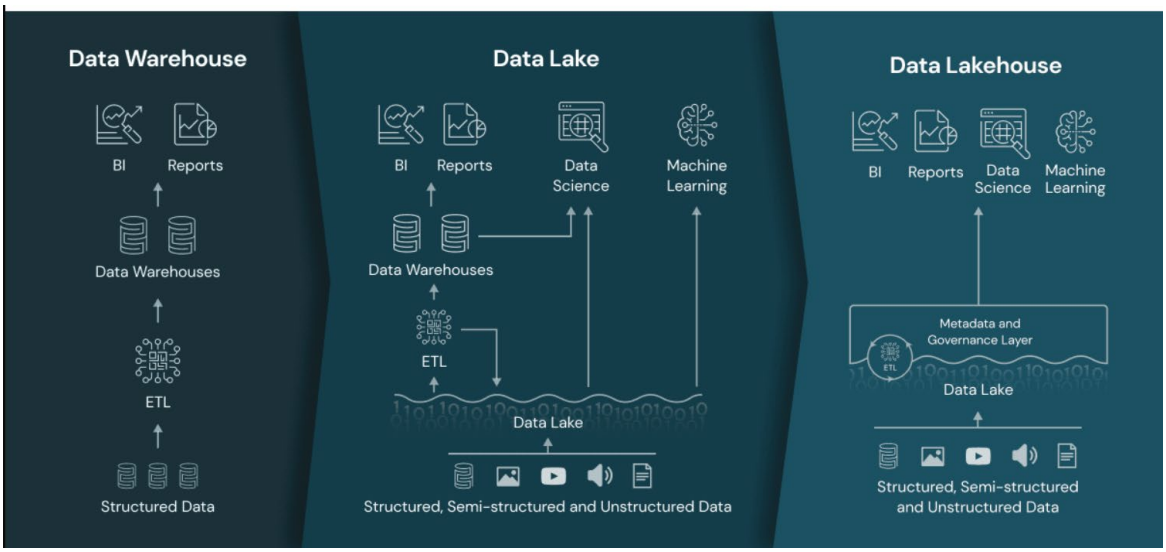

Figure 1.0 Differences of Data Architectures[1]

From Data Warehouse to Data Lakehouse

A traditional data warehouse environment uses an ETL (Extract, Transform, Load) to copy and process data from a relational database to a series of data cubes that are efficient for processing. While the data warehouse has been the mainstay of data analytics for the last thirty years, it has a number of negatives:

[1] https://www.databricks.com/wp-content/uploads/2020/01/data-lakehouse-new.png

1. ETL processes can be long and trouble prone, and typically occur only once every 24 hours.
2. Cubes are difficult to design, complex to implement, and expensive to modify.
3. Adding new data sources to a data warehouse is a slow, expensive process.
4. Implementation depends upon assumptions about the kind of questions users will want to ask of the data; innovative use of the data is discouraged.
5. Data warehouses are typically built with proprietary tools requiring expensive licenses.
6. Unstructured data can't really be part of a traditional warehouse.

Data warehouses were developed for processing and storage hardware that was slower and more expensive by many orders of magnitude. Modern cloud-based processing enables institutions to use their data with excellent performance at a reasonable cost, without brittle and expensive legacy solutions.

In the data lakehouse tools such as Amazon Redshift, Databricks Photon, or Snowflake can enable structured queries to be executed in a manner that's analogous to a data warehouse, while the governance layer provides a level of protection and logging that is much more consistent and flexible than the tools provided by a legacy warehouse. IT reporting teams can use standard tools to build reports and dashboards, but with a far higher degree of agility than that provided by the warehouse. Where appropriate for the institution, end users can be enabled to construct queries with popular tools such as Tableau or Microsoft Power BI. And they can be given access to the entire range of data in the data lake, enabling reports that draw from multiple domains rather than being forced to extract the data into a local repository for processing. IT teams benefit by replacing slow, complex and high maintenance ETL processes with relatively straightforward ingestion routines that can be operated in a batch manner or updated in real time depending on operational needs. During the ingestion process in a lakehouse, data flows through a "medallion architecture" that focuses on continuous improvement of data quality. For more information on this process see the Appendix I: The Lakehouse and Medallion Architecture.

Data Storage - The Cloud Data Lake

All data in a typical data lakehouse is stored within commodity public-cloud object storage, typically Amazon S3, Azure Blob Storage, or Google Cloud Storage. These highly available, managed-cost storage containers can be scaled and offer price-latency tradeoffs to reduce the costs of storing less-used archival data sets.

Moving data into commodity cloud storage offers tremendous operational advantages and flexibility and can be highly cost effective when managed securely and intelligently. S3 and its competitors use common API's and offer very similar services, making it relatively easy to move data from one provider to another. A single data lakehouse can support data from more than one provider at the same time, thus providing an unusual amount of flexibility. These object stores are the institutional data lake; once there, the goal of the Databricks Lakehouse

Architecture is to securely provide all the services you need to process, compare, track, correlate, and manage your data without ever having to move it.

Data stored in traditional sources such as Oracle or SQL Server databases needs to be ingested into object storage. There are proprietary and open solutions to support ingestion, including tools for single-time ingestion and for ongoing updating.

The Data Reliability and Processing Layer: Delta Lake

Delta Lake is the glue that binds the data lake together and allows it to be processed with reliability comparable to a traditional relational database. Offered by Databricks, but also available as an Apache open source product, Delta Lake can provide ACID (atomicity, consistency, isolation, durability) transactions, enforce schemas, and use metadata to integrate with a data catalog. Delta Lake organizes data in a format based on Apache Parquet, which assures a high degree of efficiency in operations and storage while avoiding a proprietary format. Delta Lake keeps complete transaction logs to support auditing as well as rollback. Operations such as delete, update, and merge, difficult in a typical data lake, are managed by Delta Lake, greatly simplifying operations.

Delta Lake provides direct support for analytics tools of multiple flavors, from Apache Spark, typically used in highly parallelized workloads such as machine learning, to Amazon Redshift and Snowflake, which allow SQL queries on the data as if it were in a traditional database or data warehouse. Delta Lake is serverless and provides a range of options that drive the value of the data lake, making it possible to get the analytical value from all the data in the lake with a level of performance comparable or better than legacy data warehouse solutions. Even better, because Delta Lake is open source and non-proprietary, there are no licensing fees or vendor lock-in as in legacy solutions. This shift from legacy systems with expensive licenses, software patching, fussy tuning, and long development times frees up an IT data support team to provide better services with fewer person-hours of support and often at much lower cost.

Data Governance and Sharing Layer

The data lakehouse enables a level of unified and consistent access control that was nearly impossible to achieve in previous architectures. When institutional data is scattered across multiple proprietary systems and locations, it’s not possible to assure consistent rules are applied across these systems and locations. In addition, the challenges that campus users face with accessing data stored in multiple locations practically forces them to make local copies to meet their needs. The inevitable result is shadow databases, which increase the challenge of protecting data from misuse and abuse while increasing risk.

Shadow databases erode trust in a “single source of truth” since the data are often processed in undocumented and inconsistent ways. Furthermore, once the data has been moved from the central data repository, you can assume it will be shared or simply be accessible within the

institution and sometimes beyond without regard to data governance or information security oversight. Nearly every institution is familiar with this scenario, and legacy tools have not been helpful in resolving them. People want to use data to understand their institution and if it is not provided in a manner that's both consistent and controlled, but still convenient and flexible for the appropriate users, out-of-control data sprawl and outdated data will always be the result.

Data governance can be a scary word for an institution, and many IT leaders have been frustrated by the difficulty of getting good data policies and oversight in place. However, whether your institution has a best-practices governance process in place, or operates on a seat-of-the-pants, ad hoc manner – or something in between – IT ends up with responsibility for the systems that control who can access the data and protecting it from exploitation by internal misuse and outside cyber-attack. The data governance and sharing layer within Databricks doesn't solve all your data governance challenges but it does give you a set of unified tools that manage access and maintain all the logging you need to respond to audit requests and address incidents of misuse.

Databricks Unity Catalog offers a well-integrated and ambitious set of tools for finding, managing, and controlling access to data. The Unity Catalog can manage both the structured data that you would have in a traditional data warehouse, as well as unstructured data such as activity logs and other big data sets that are popular for machine learning and other advanced analytics. The centralized metadata in Unity Catalog enables users to discover the data in the lake, and access only the data that they are authorized to access.

Unity Catalog operates in conjunction with Delta Sharing, an open-source protocol designed to allow controlled access (and ingestion) of data from outside the institution. It supports sharing across platforms and gives outside entities the ability to perform analytic operations without making a local copy, while logging every access for audit. This is a promising approach for supporting open science and other multi-institutional efforts while reducing the potential for misuse of the data compared to shipping around copies of data.

Open-source tools can provide key governance features like keeping your data secure and accurate; building and supporting a fully featured open source stack, however, might take more time and effort than fully featured proprietary systems. For example, a campus already using Microsoft Purview can enable it as the data governance layer in the Lakehouse; alternatively, Unity Catalog and Microsoft Purview can interface via API to share metadata. There are a number of other similar tools that can serve this role. The main advantage here is that by keeping all types of data in a single (virtual) repository with a single access point, end users have the advantage of being able to perform analytics or research on the entire range of institutional data, using their favorite query tools or analytics platforms, without making a copy of the data. With these tools, a single source of truth, managed, protected, and governed becomes much closer to reality

Data Lakehouse in Higher Education

Architecture and infrastructure only offer a return on investment to the extent in which they support applications in the enterprise. The data lakehouse offers the promise of getting applications to value more quickly and completely, by supporting a more modern data architecture that is “artificial intelligence and machine learning capable." In this section we outline how a Data Lakehouse Architecture can provide value for the higher education institution. But first, let’s explore infrastructure a bit more by discussing IT Modernization.

IT Modernization

Many institutions have built data lakes in the public cloud. Others are just exploring this option now. The data lake offers the opportunity to consolidate data from a wide range of applications in a single place, enabling data exploration and analytics that span the enterprise in a way that requires inefficient and insecure copying, extracting, filtering, and moving data in ways. These legacy approaches reduce consistency and increase the vulnerability of the data to internal or external misuse or attack. The cloud-based data lake offers much greater resiliency in the event of a disaster and enables a more agile approach to managing and building applications.

One of the key considerations - and challenges - in creating a data lake is managing secure, auditable user access. Traditionally, application security has been baked into the application; even when using an Identity and Access Management (IAM) system, it’s usually the application that tracks who has access to read and write specific fields. This application security is not only cumbersome and difficult to manage, it’s also different for different applications, and a consistent approach is not enforced across the various applications an enterprise uses. It’s a common scenario that an individual leaves or moves into a new role, and their access gets adjusted in some applications but not in others, and that downloaded data is not completely expunged.

The Databricks Lakehouse Architecture offers a different approach. The built-in governance layer enables a comprehensive and consistent approach to data access that can be managed from one place. Databricks’ Unity Catalog implements a rich set of tools for managing an institution's data, by interfacing with an IAM via Active Directory or other options, maintaining access control for each authorized user, and maintaining as much audit information as needed for specific classes of data. Thus, the management is simpler and more reliable, and the user experience is improved by offering a unified gateway for data access. A good user experience reduces the

> “The California State University Chancellor’s Office is strategically utilizing Databricks to streamline data processing across all 23 campuses within the nation's largest public four-year university system. This initiative will enhance our capabilities in managing HR, student, and financial systems efficiently.”
>
> Subash D’Souza
> Director, Cloud Data Engineering
> CSU Chancellor’s Office
>
> CSU The California State University

incentive for data customers to build their own shadow data systems to get the flexibility they desire.

As institutions modernize their IT estate, they are increasingly adopting software-as-a-service applications. This results in even less access to structured data than before, as we often cannot access underlying relational databases. To enable core business intelligence within a modern IT environment, we need a solution that can support semi-structured and unstructured data. The Databricks Lakehouse Architecture is an open architecture that combines the best elements of data lakes and data warehouses such that analytics teams can work with semi-structured and unstructured data seamlessly using familiar languages like Databricks SQL.

Recruitment/Enrollment

According to the National Student Clearinghouse, total enrollment in postsecondary institutions remains significantly below pre-pandemic levels, falling by approximately 1.09 million students overall and 1.16 million undergraduates alone compared to the spring of 2020. While the trend appears to be reversing as the number of freshmen increased by 9.2 percent from the spring of 2022 to the spring of 2023. However, spring percentage increases are predicated on a much smaller scale than fall increases. In the UK, where student numbers have increased or remained stable, there is a need to recruit international students or more students from specific demographics.

Most institutions of higher education live or die on their ability to recruit, enroll, and retain the right mix of students. A data lakehouse offers the opportunity to pull together a broader and richer range of data sources than previously available in traditional systems. Unlike their experiences with closed, proprietary data management provided by third party vendors, an admissions office can pull together data from recruiting CRMs, but also from web logs, student communications, and social media, into a single data repository to which advanced analytics and Machine Learning can be applied.

In a unique use case, a post-secondary institution created their data lakehouse to connect student applications, courses offered by their colleges with admission and transfer requirements and leveraged AI to read the transcripts and automatically assess eligibility for admission or transfers. Leveraging data and AI has improved accuracy, and time to respond to students quickly. The costs of manual labor significantly decreased while student communications increased.

Student Success and Engagement

Tracking and improving student success and engagement is far easier with a data lakehouse than legacy approaches. Typically, student success support is offered as a new enterprise application that becomes a new data silo, pulling information from the Student Information System (SIS), Learning Management System (LMS), and perhaps other auxiliary systems that

are supported by the application. Many applications use their own proprietary strategies for managing and analyzing the data and create the need for another security scheme to protect it.

By starting with a data lakehouse, the institution can build a flexible environment for analysis that easily accommodates new data sources as they are identified. Suppose, for example, the institution wishes to explore the effectiveness of the student recreation center for promoting student well-being and retention. With a student success application, you have to first hope that the software provider has, or can build, an interface with your recreation center data, and then someone has to figure out how to pull that data into the student success system. In a data lakehouse, the log of student usage becomes a new object in the data lake and can then be used for analysis in conjunction with other data objects in the same place. Sensitive data can be protected through the data governance layer to protect student privacy. This is just one example of the many possible scenarios that can be implemented relatively quickly and inexpensively once the data lakehouse is in place.

If your institution is ready to use machine learning with your student success data to build predictive models, the data lakehouse enables direct access to the wide array of ML tools available from cloud providers. We discuss these options further in the next section. Even if you're not using ML now, by selecting the Databricks Lakehouse as your architecture you are ready for this phase of analysis when your institution desires to go there. ***Traditional architectures require that you create an entirely new environment to support ML and other AI applications, whereas with Databricks the ML functionality is built in from the outset.***

Academic Research

The Databricks Lakehouse Architecture comes from the work of academic researchers at the University of California, Berkeley, and other institutions, and offers many attractive features for organizing, protecting, and sharing of research data. In this section we'll take a look at a couple: protecting and sharing and using machine learning.

Big data research has increasingly become a multi-institutional activity, with large data sets shared within and among international research teams. In this environment, controlled access to many data sets is critical to assure privacy, protect intellectual property, and to meet federal regulatory requirements. As important as these considerations are, the protection of the data is often applied as a patch on top of data organization, which creates risk of data compromise and also makes it difficult or impossible to demonstrate in an audit how the data is being protected and shared. Different data sets may have different methods of tracking user identity and data access, and require complex administration from faculty members or graduate students.

By contrast, the Databricks Lakehouse Architecture enables a reliable and auditable scheme of access, integrated with institutional or federated Identity and Access Management (IAM) schemes. Using a tool such as Databricks' Unity Catalog allows fine-grained access control and

complete logging, improving security and assuring that logs are available for audit and incident response. While the term “Data Governance” may not always be popular in the research world, it’s a reality that institutions incur significant liability when data is misused.

In addition, by implementing Delta Sharing, the Unity Catalog can provide broad access to data that is meant to be shared, while enforcing access controls such as federal restrictions on international export of data. Having a comprehensive architecture that’s reusable across multiple research groups and schools provides a power strategy for data management and governance while supporting researchers in getting the access they need.

The roots of the Databricks Lakehouse come from Machine Learning, and this architecture, built on the commonly used open source tool Apache Spark, provides direct access to popular tools such as PyTorch and TensorFlow. A Databricks Lakehouse enables Machine Learning researchers to use common cloud-based tools with which they are familiar, reducing the time it takes to achieve their research goals. Simultaneously, the same data can be used with analytic tools to perform more traditional analytics. This hybrid nature enables the institution to provide appropriate risk management while enabling innovation and discovery for the researcher.

Advancement, Alumni, Athletics

The traditional campus approach to fundraising systems has been to purchase closed, proprietary software that provides limited and siloed approaches to managing this critical data. The data lakehouse approach offers the opportunity to store a wider range of data sets, structured and unstructured, and to support traditional data analytics which expands the opportunity to use Machine Learning and other advanced tools to perform a deeper dive into this potential well of rich data.

Athletic programs have increasingly been using Machine Learning tools to help athletes achieve peak performance. They do so with a range of point solutions specifically designed to address needs in a particular area. Having a secure data lake house with both structured and unstructured data available to analyze athletic performance is a goal of all athletic programs. Having the ability to work with researchers who are pushing the envelope creates a synergy that benefits both the athletes and researchers and requires a common but expanded data pool that is highly secure and auditable.

In addition, we are seeing the most advanced professional and university athletic programs using data to decide which athletes should be recruited. The data lake house provides a far more flexible and less costly way for local innovators to pursue new ways of identifying athletic talent that is in alignment with an institution's goals.

Cyber Security Data Lakehouse

For years campuses have struggled with the challenge of handling the vast quantities of data generated by security logging tools such as Splunk and many others. The data lakehouse offers an environment ideally suited for managing these volumes of unstructured data. In the data lakehouse, the analytics tools available can be used for traditional reporting, while Machine Learning can be applied for more traditional analysis, including the ability to find patterns in the data that simply wouldn't be accessible through older methods.

But while it is important to note that a data lakehouse provides a way for cybersecurity staff to better analyze threat intelligence, it is also a far more secure way of protecting an institution's data from outside threats than legacy systems. To best protect data, the Unity Catalog allows for not only the classification of data (some needs to be protected more than others), but has built-in controls that reduce the risk of downloaded data hanging around in residual files that may or may not be known to cyber security staff, internal auditors, or risk management officers.

Data warehouse projects are notoriously difficult to execute well, and many institutions have experienced failed projects in recent memory. The challenges of these projects are numerous, but one of the most significant is finding a tool that can support the wide requirements of a university while also being cost-effective to run. When these projects fail, the data needs of our institutions don't go away. Instead, they are often addressed with shadow IT databases or spreadsheets with limited controls. Unfortunately, human error in handling data is a significant factor in the rise of data breaches that institutions face today. As more core business processes fall out into shadow IT this risk increases.

Control Functions

One of the most challenging aspects of legacy data systems is that they do not make it easy for control functions to do appropriate auditing. For example, most institutions have a few individuals who are known to have the ability to create value from downloading data, and creating shadow systems that are entirely different from the main application systems that support students, HR, Finance, and research to name a few. These shadow systems do not have appropriate levels of security around them, often being built with tools like Microsoft Access or Microsoft Excel, and being managed by someone with great familiarity with those tools, but less familiarity with enterprise tools and needs.

The Office of Risk Management, the Internal Audit Office, and the Office of Cybersecurity all have an interest in quantifiably reducing the risks that come from the proliferation of data outside of normal controls. HR and Finance staff traditionally rely on data feeds to discover patterns, both good and bad, in the actions (hiring, promoting, retaining employees-procuring goods and services efficiently and effectively) taken by these offices to advance the institutional mission.

Risk management has largely been considered to be an insurance function. More and more, institutions are relying on Risk Managers to help identify risks and mitigate them. For example, insurance companies started selling "cyber security insurance" starting a few decades ago. Recently, many have stopped offering insurance because of their inability to properly assess risk. This was not due to lack of information about major enterprise systems. Rather, it was about non-enterprise areas of risk associated with shadow systems, and shadow data. Again, a data lakehouse architecture allows for enterprise level management and control of all data, while allowing for easy access to approved users. Access is logged, enterprise systems and Databricks handle the security, and internal audit can periodically audit how things are working. We believe that insurance companies would lower costs for cyber liability insurance if they fully understood all the risks, instead of building in a cushion for the risks they do not know about.

> "We need to empower staff through decentralisation, whilst maintaining centralised control of core functions and governance"
>
> Mark Wantling
> CIO at the University of Salford

Beyond the institutional context, the use of a data lakehouse can provide additional support to maintain legal compliance with data regulations such as GDPR. The data lakehouse both allows for the implementation of appropriate governance and controls but also provides potential for retention schedules to be managed and automated. With data siloed in different local systems institutions are either retaining data for longer than they should be, without oversight that this is happening, or cleansing data manually at considerable cost in terms of time and resources.

Data Lakehouse as a Foundation for Artificial Intelligence Innovation

More than ever before, IT leaders are being challenged to apply Artificial Intelligence (AI) to promote the mission of their institution. The combination of legitimate breakthroughs and exciting applications along with a drumbeat of hype and opportunism has created an urgent need to find a way to respond that will add lasting value to the services provided to the institution. At the same time, legitimate concerns about the potential risks of AI provide challenges to experimentation and innovation.

While the role of AI in the academic enterprise in the future is hard to predict and controversial, one thing is clear: most effective applications of AI are grounded in data. It's the ability of AI models, whether predictive analytics for learning, generative AI, or any other application based on machine learning, to draw upon a large volume of data, much of it by nature unstructured, that makes AI's future both interesting and challenging. Having a data architecture that's managed, scalable, cost-effective, and flexible may be the single best investment that higher education IT leaders can make to prepare for current and future applications of AI.

S&P Global' s 2023 Global Trends in AI [2] report surveyed AI decision makers at large companies worldwide. According to an article by Axios[3], many companies are finding their data is not well organized and ready for AI and need to reconsider how data is managed and processed. Similar to education institutions a different way of thinking about how data is stored, managed and processed is also needed.

Consider a couple of strategies that are common but may not be optimal in the long run. First of all, building an environment for AI innovation using on-prem equipment may seem cheap at first, but the limitations of a fixed sized environment and the challenges of providing access to the ever growing range of open and proprietary tools and languages mean that most institutions are going to run out of equipment and support long before they run out of ideas. Furthermore, the boom in interest in AI has made it hard to even acquire the special purpose hardware needed to stand up and expand a dedicated environment.

Operating an AI environment in the cloud is the likely future path for most institutions, but building a cloud environment that's tailored for AI without a lakehouse architectural strategy ends up creating yet another data silo that needs to be fed and managed. Adding new data sources to this environment is going to revert to the age-old but deeply flawed method of extracting data from somewhere and copying it to this new environment. Once the data is in this environment, there's a whole new set of data to protect, most likely including Personally Identifiable Information (PII). You can build a pilot environment and try to manage and protect it using ad-hoc methods but in the short run it is risky, and in the long run it is not scalable.

[2] https://www.weka.io/resources/analyst-report/2023-global-trends-in-ai/

[3] https://www.axios.com/2023/08/19/ai-corporate-barriers-cost-data

The Mosaic Platform and the Databricks Lakehouse

This is why Databricks has chosen the Mosaic platform architecture for AI infrastructure. The Mosaic platform is purely a computational substrate, providing a highly scalable, multi-cloud system for orchestrating the large quantity of GPUs and other specialized processing resources. No new persistent storage is added to complicate data governance, as the Mosaic platform enables streaming data directly from the Lakehouse so efficiently that there is no compromise to the size of the task, or the speed at which AI and machine learning workloads can be computed.

Through the integration of the Mosaic platform, the Databricks Lakehouse represents the most promising architecture for supporting the needs of AI research. The very roots of the Databricks Lakehouse as an architecture and Databricks as a company were in the needs of managing large volumes of data for machine learning. The tools that machine learning and data engineering researchers use for ingesting, cleaning, managing, and sharing large volumes of data are native to the lakehouse environment. The wide range of flexibility in data lakehouse - from providing traditional business analytics functions to advanced data engineering to machine learning – is the hallmark of the architecture and what it was built to do.

Not only does the Databricks Lakehouse environment provide the tools and techniques that support AI applications, it does so in a way that supports data governance. It sounds like an oxymoron – a single architecture that spans the needs of data governance and cutting-edge innovation and flexibility. But that’s a big part of what makes the Databricks Lakehouse environment so interesting and exciting.

Databricks recognizes the concern that many institutions have when creating, deploying, and maintaining not just large language models, but all AI models. Handling personally identifiable information (PII), maintaining models in production, having access to state-of-the-art models, and ensuring data does not leak out, are all points of contention. When handling potentially very sensitive student data, it is important to be confident of the security of your model. Databricks, and the lakehouse architecture, provide tools to mitigate these concerns. The toolset available allows you to spin up your own AI models in your own environment, giving you confidence about where your data is going. Through MLflow, a platform developed by Databricks, you are able to track and maintain models in production easily. With MosaicML you are able to get easy access to state-of-the-art AI models to use and fine-tune your own data in a secure manner.

As mentioned above, many higher education institutions are considering pilot applications of AI. Let's suppose for example that you’ve been tasked with supporting the development of a specialized Large Language Model (LLM) to serve the needs of your university. One of the first challenges is collecting the data to draw upon for the LLM and putting it in a place and format that can be used to build the model. The nature of an LLM suggests that a large portion of that data is going to be sample text, but the specific needs of the model may also require more traditional, structured data. IT now has a strategy for putting all that data together, providing

access to best-in-class tools, and providing a framework that governs and protects access to the data. Using a data lakehouse as the organizing principle behind this project results in a long term strategy that can grow with demand rather than a one-off pilot that's a potential dead end.

There's more good news – standing up a data lakehouse in the cloud is not a long-term project; rather it's something that can be accomplished in weeks or months, perhaps with some help to get you started. The costs of a data lakehouse implemented through Databricks are "pay as you go", so you can start small at a reasonable price and grow capacity as you need it. This is the attraction of the cloud model – speed, agility, flexibility, scalability, and managed costs – as applied to modern data management through a Databricks Lakehouse Architecture.

When faced with an uncertain future, the innovative IT leader is always looking for strategies that can provide flexibility and capacity for whatever may come. Right now a data lakehouse is clearly one of the most valuable strategies available for creating an "AI-ready environment" that can provide value now and has the potential to continue to serve well for a number of years.

We can already see universities using AI to automate tasks and free up staff time for value adding activities or using AI to supplement staff and increase their capacity. For example, institutions are implementing AI support as the first line of student-facing IT helpdesk with escalation to staff at the second line. This same approach could be applied to a broader range of student support processes, or the review of entrance applications. As institutions explore the art of the possible, the insights offered by data will help them to identify areas in which they can implement AI and maximize their return on investment in new technologies.

How can you get started?

Discussions of data are not new to present and past CIO's. There are clear challenges of managing both evolving and current needs as budgets continue to shrink. What is new is that for the first time we have a unified tool set which flexibly and securely supports both structured and unstructured data. While it is difficult to think of taking on new initiatives, the need for data and its use will become more important.

> "It's common for universities to invest lots of resource in a big data warehouse project, which then over-runs or fails entirely. Databricks has allowed us to start small and articulate the value to the wider institution"
>
> Mark Wantling
> CIO at the University of Salford

If you have any questions, comments, or thoughts on data lakehouses, we welcome an introductory conversation. Please reach out to us at:

North America: Info@kovexa.com
LATM: Info@kovexa.com
APAC: Info@waterstons.com
EMEA: Info@waterstons.com
For specific Databricks questions: University@databricks.com

Appendix 1: The Lakehouse and Medallion Architecture

The architecture of the lakehouse and the toolset of Databricks enables a wide range of projects to happen in the same ecosystem. Here, data engineers, data analysts and data scientists are empowered to use a similar set of tools to each other, lowering the barrier between each field. Since structured and unstructured data are all located in the same place, it means whether you are running a traditional reporting project (fed usually from structured information) or a state-of-the-art AI deployment (typically trained on unstructured data), the root of the data you are using is the same.

During the ingestion process in a lakehouse, data flows through a "medallion architecture". This is a core concept of how the lakehouse is structured and enables enterprise data products to be sourced with high-quality data. Data goes through a bronze, silver and then gold layer, each indicating the increasing cleanliness and usefulness of the data. Different users can then siphon data from different layers depending on their needs, confident it is all coming from a single source of truth.

Bronze: The initial layer is where raw data is landed from external sources. It aims to retain its original structure and can serve as a historical backup (in cold storage) allowing us to rerun any process without re-querying the source.

Silver: Here, data begins to be transformed for reporting and analysis use. Data will be cleaned, merged and changed to conform with your business's structure. Corrections can be made here too – actions like changing column names and object types. The silver layer will provide a validated view of our data that will be relied upon further downstream.

Gold: In this final layer, data is made ready for consumption into business intelligence platforms. Specific values you will want to report on are calculated and made read-optimized. It is from here we will do data modeling techniques (such as star schemas) and data quality checks.

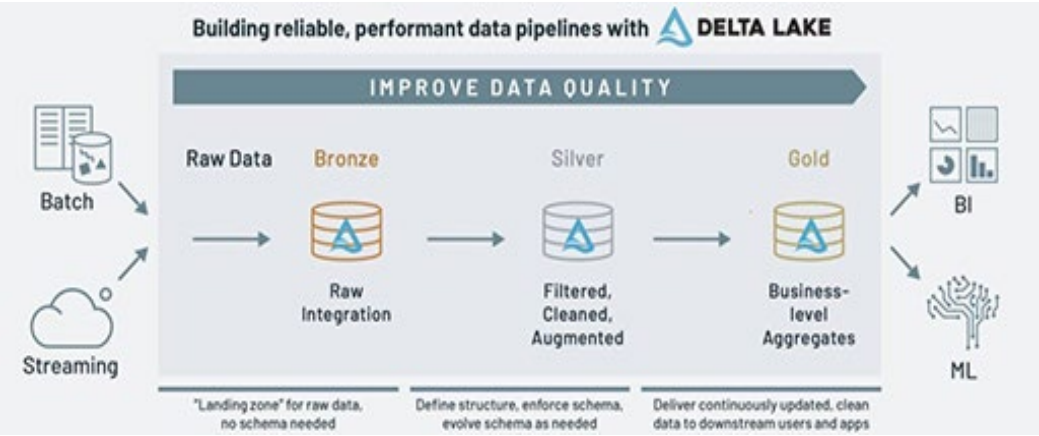

Figure 2: 'Delta Lake Medallion Architecture', Databricks[4]

The medallion architecture structure allows for a simple and understandable way to apply our ETL processes within the lakehouse. It creates a single source of truth for users – as data professionals will all pull their data from the same silver and gold layers. Developers creating dashboards and reports can access very clean data from the gold layer, whereas developers looking to create AI and machine learning models (who may need more raw data) can plug into the silver layer. As discussed in the section Using a Data Lakehouse as a Foundation for Artificial Intelligence Innovation, this is one reason that the lakehouse is a versatile tool that can be used effectively for the creation of large language models.

[4] https://cms.databricks.com/sites/default/files/inline-images/delta-lake-medallion-architecture-2.png

databricks

Thank you for downloading this Databricks resource! Carahsoft serves as the Master Government Aggregator® and Distributor for Databricks offering expertise in government procurement processes and practices with purchasing available via GSA, SEWP V, ITES-SW and other contract vehicles.

To learn how to take the next step toward acquiring Databricks' solutions, please check out the following resources and information:

For additional resources:
carah.io/DatabricksResources

For additional Databricks solutions:
carah.io/DatabricksProducts

To set up a meeting:
Databricks@carahsoft.com
(877)-742-8468

For upcoming events:
carah.io/DatabricksEvents

For additional Open Source solutions:
carah.io/OpenSourceSolutions

To purchase, check out the contract vehicles available for procurement:
carah.io/DatabricksContracts