MATILLION

Guide to the Lakehouse:

Unite your data teams in the cloud to bridge the information gap

Contents

From punch cards to semiconductors to the cloud

The evolution of data storage and analytics has spanned decades. There have been tremendous advances in the technologies used to store and process data. But it's the increasing need for data-driven insights that propels people from both the business and technical side of the organization to try to take greater advantage of all the data that our systems, products, and services create. Although we're changing the way we store and compute data, a few challenges remain the same. The first is preparing data in a timely manner. The second is providing engineers, data scientists, and business users with quality data that they can use to derive insights and make transformative decisions for the business. The last is bringing different types of data workers together by providing a truly collaborative environment in which data culture can flourish.

This last challenge, the collaborative environment, may get less attention, but is perhaps the most important foundation of modern analytics. Gaps in information—both data and communication gaps—can hinder progress. Building deeper customer relationships is one of the most transformative things a business can do. And for deeper relationships to happen, organizations need to bridge the gaps with a comprehensive understanding of buying and consumer behavior, as well as the business systems necessary to support this process. Data is critical, yes, and obtaining and leveraging the data generated across the entire business requires cross-functional effort and cooperation.

However, according to NewVantage Partners' "Big Data and AI Executive Survey 2021," 92.2 percent of leading companies continue to identify culture, including people and processes, as the biggest impediment to becoming truly data-driven organizations. The good news is that the right technology can help drive the kind of cultural change that transforms business: Bringing together data engineers, data scientists, business analysts, and others whose jobs and performance depend on having quality data, to work together to reduce costs, drive innovation, and shorten time to market.

The convergence of two worlds of data

With the ability to get data out of silos and centralized into one high-performance, highly scalable environment, the cloud has long held the promise that different branches of the modern data team can work more closely together. But until recently, even in the cloud, data teams worked in their own domains, with their own data. Data engineers housed their structured data in a data warehouse, where they could use it for reporting, analytics, and business intelligence. With its ability to combine both structured and unstructured data in its raw form, the data lake has become the domain of data scientists who use it to find new opportunities through deep insights, predictive analytics, and, more lately, machine learning and AI pattern recognition.

To some extent, this makes sense. These are different use cases for different types of data that until now required different platforms to manage the different workloads. But data teams working in silos of separated data, with their own sets of tools, created great inefficiencies in many areas that need to be overcome in order to realize true business value from data. For modern analytics, it's critical for data engineers and data scientists to be closer together, not apart. That also holds true for their data.

A shift toward data science, machine learning, and artificial intelligence

Machine learning and artificial intelligence are no longer on the far horizon – they're getting closer to being a reality for all organizations. There are several reasons for the shift. First of all, with both the volume and diversity of data rapidly increasing, it's simply no longer possible for humans to analyze all of it. Organizations are turning to machine learning and artificial intelligence in order to keep up with that enormous volume of data and make sense of it. Second, as customers demand more customized, personalized experiences, companies need to model more data and identify more attributes to be able to provide customers with what they want, when they want it…even before they know they want it. ML and AI are potentially becoming the keys to unlock the insights in their data to enhance existing revenue streams or develop new ones. As a result, more companies are relying on data scientists to quickly build models and data products.

Because it's a relatively new discipline for the majority of enterprises, data science doesn't necessarily offer a one-size-fits-all solution. An explorative mindset needs to be cultivated and supported. So it's critical that data scientists adopt a flexible approach that facilitates quick iteration. Fast access to data is only possible through fast ingestion, transformation, and orchestration of data pipelines in an automated, managed way. Speed and collaboration are the vital ingredients for organizations on the data journey who wish to mature their business reporting and analytics practices from descriptive, through predictive, and eventually prescriptive insight.

Getting data products to production faster

Which brings us back to the importance of collaboration between data engineers and data scientists. One of the most critical barriers to business productivity and innovation is where the two teams overlap–in production and new product creation. Data scientists often create experimental data products that then have to be rebuilt by data engineers before they can be used in production. This division of labor duplicates effort and creates unnecessary steps that significantly slow the release of new products to production.

To increase agility, businesses need to be able to move from data science experiments to production more quickly. Identifying and creating new revenue streams is a key success metric, but it's often impeded by a lack of capacity to ingest and enrich data sources at scale (this requires automation and skills, both of which can be in short supply).

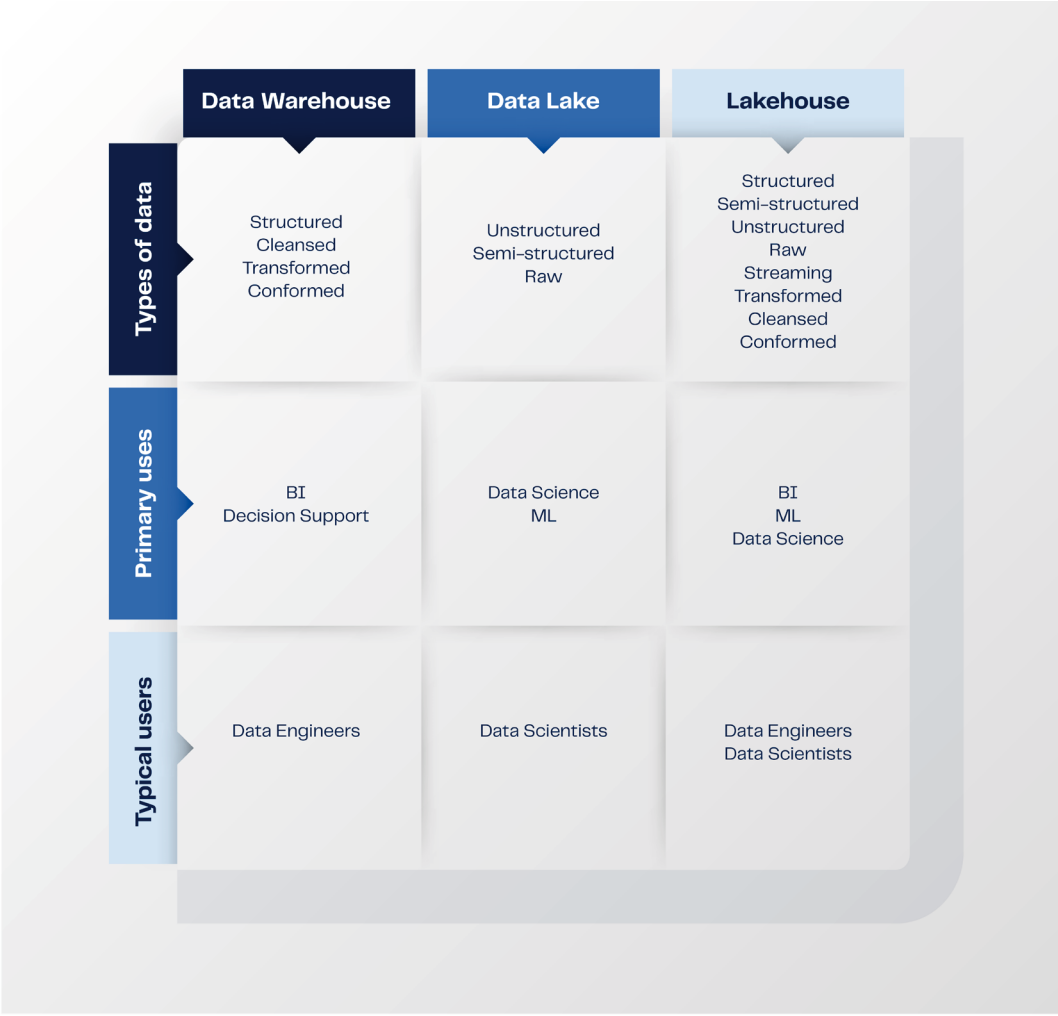
The need for speed in development and productization is especially urgent for businesses wanting to get value from their data scientists. It's estimated that data scientists spend the majority of their time prepping data, rather than doing what they're actually being paid to do: model that data and derive insight from it. In the 2020 State of Data Science report from Anaconda[2], data scientists reported that they spent nearly half of their time on data loading and data cleansing. Further, 44 percent of the data professionals surveyed reported that they plan to seek a new job within the next year. If a data scientist's job is taken up by busy work, and a data scientist changes jobs every few years, the math adds up to a harsh reality: Data scientists can't get up to speed and add real value before they are on to the next gig. To increase efficiency and retain talent, there needs to be a shift in how data scientists and data engineers work together with data.

When data teams need to access data in both a data warehouse and a data lake, it often requires creating ad hoc systems to access data from both places and combine it. Unfortunately, that process takes time and gets messy, often creating even more complexity to manage data in both the data warehouse and data lake. Combining data from multiple systems takes time. And the process of doing so can potentially result in duplicated data, increased infrastructure costs, or security issues. As the roles of data professionals have shifted and more people are requiring both structured and unstructured data in large volumes, modern data teams find that they need an easier way to access and share data across a more unified data ecosystem.

The lakehouse: closer collaboration in a unified environment

A lakehouse is a new data management paradigm that combines the capabilities of data warehouses and data lakes. The lakehouse has the data structures and data management features of a data warehouse, but stores data directly on the kind of low-cost storage used for data lakes. Because a data lake can house different types of data in a single location in its raw form, users can more easily perform BI and advanced analytics such as ML/AI on all of their organization's data.

The lakehouse model has the potential to change the way data teams work together. With a lakehouse, data engineers and data scientists can work in the same system, using the same tools. When data teams no longer work in cloud silos, they can work together much faster while reducing risks to data fidelity. Plus, with a single, consolidated location for data, teams always have the most complete and up-to-date data available for data science, machine learning, and business analytics projects. The lakehouse approach enables data engineers, data scientists, and data analysts to share assets such as data sets, dashboards, and models, so they're all looking at the same assets in the same tools.



One platform, many benefits

The route to rapid lakehouse adoption and success lies in a unified approach to data and the people that work with and benefit from it. This is a new mindset to cultivate for many organizations. One way to increase efficiency is to foster collaboration and a common language among data scientists, data engineers, and the internal "customers." The latter group is often drawn from the functional part of the business that stands to benefit, typically sales, marketing, and product research and development. Creating a cross-functional team responsible for guiding the work in the lakehouse can help unify different teams by creating shared goals and a common vision. Business stakeholders are another critical part of your cross-functional teams, not just data workers, to ensure a business-outcome-based focus.

Consolidating technical skill sets for data ingestion, transformation, and orchestration on a reduced number of technology platforms will open up the available resource pool that organizations can draw from and potentially eliminate some of the process bottlenecks typically found across siloed teams. A readily accessible Extract, Transform, Load (ETL) or Extract, Load, Transform (ELT) solution, coupled with a capable baseline set of skills, will have a stronger accelerator effect on the overall data transformation journey of businesses, compared to isolated pockets of highly skilled individuals using complex or niche tools. Maintaining pockets of specialization is sometimes necessary, but can be costly while capacity is not easy to expand. And it can easily lead to the creation of a less flexible and resilient data organization.

A strong bridge between data science and data engineering matters because better collaboration:

Helps shorten time to market

According to Databricks, the data scientists and machine learning engineers at ABN AMRO used the lakehouse architecture to democratize access to data and support greater collaboration. Their efforts resulted in a 10x increase in speed to market, with use cases being deployed in just two months.[4]

Drives innovation

Shell, another Databricks customer, used a shared workspace to democratize access to data and also foster cross-team collaboration across data engineering, data science, and the analyst team found that this approach helped refine the company's product roadmap and unlock new opportunities to engage with customers.[5]

Accelerates time to value

Regeneron accelerated the discovery of new drugs and therapies by reducing the time needed to run queries on its datasets, and by accelerating its data pipelines.[6]

Decreases costs

Global retailer H&M cut operational costs by 70 percent by using lakehouse features in Databricks such as auto-scaling clusters to improve operations.[7]

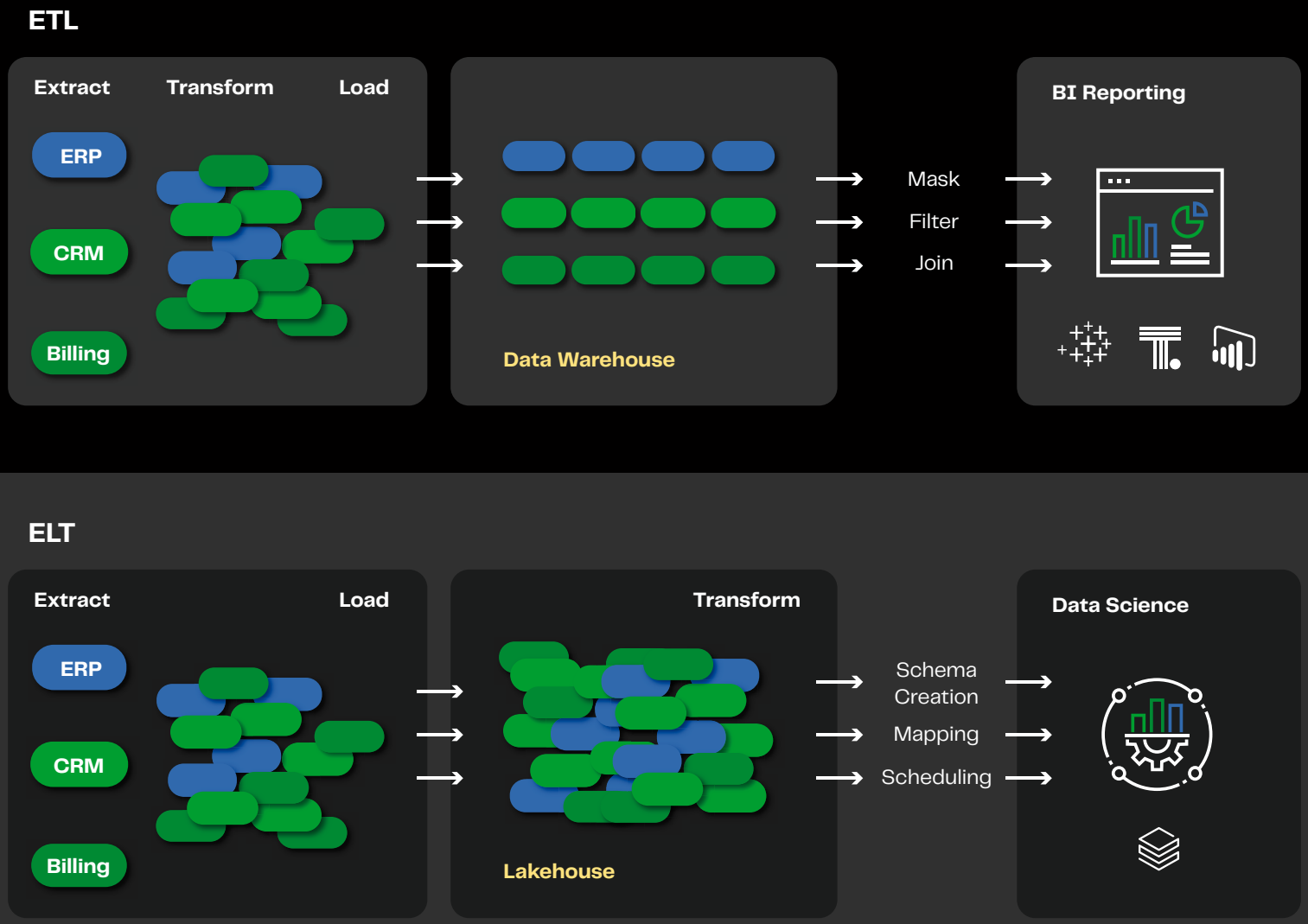
Cloud-native ELT and the lakehouse: Bridging the gap between data engineers and data scientists

Using ETL methods to consolidate data for analytics is nothing new – the process has been around for decades. But as our data warehouses have modernized and moved to the cloud, our ETL solutions have lagged behind, with capabilities more suited to on-premises data architectures than the modern cloud data platform. If you're analyzing data in the cloud, you need a cloud-native solution that supports a slightly different take on the data transformation stage. The Extract, Load and Transform ELT approach takes advantage of the power of the cloud platform's compute capabilities by "pushing down" the transformation step, the "T" in "ELT," within the cloud platform. Using the power of the cloud to help transform data is faster, more economical, and better suited to supporting modern analytics.

While many ETL vendors have been around for years, their products are not designed or well suited to the cloud. For example, some vendors tried to modernize by forcing their older tools to generate heavy Spark and Scala code. But tools that are scripted and text-based don't remove any complexity. If anything, they add to it. Sometimes it seems that the only way to adopt these tools is to get a computer science degree and/or learn a new language, which is basically akin to learning a new programming framework and acts as a barrier to entry for fresh talent.

Don't try to future-proof your organization with tools that weren't designed for the future. Older tools weren't created with the flexibility, features, and scale of the cloud in mind. They are harder to use and require a lot of work that can be made obsolete by automated features and visual interfaces in modern ELT platforms. They simply can't keep up in the cloud, which is the last thing a time-strapped data team needs. In short, be wary of tools that take older paradigms and try to shoehorn them into a very different modern context.

ETL vs. ELT Comparison



The ELT connection

ELT plays an important role in creating a unified approach to data ingestion and enrichment in the lakehouse, helping bridge the gap between data engineers and data scientists. SQL is the one protocol that is consistent in almost every cloud data platform. SQL is the layer in ELT that abstracts the complexity of these very technical yet powerful data platforms. With the right set of unified tools or platforms that can abstract the complexity of Spark and Scala using SQL and low-code processes, data engineers can eliminate hours of hand-coding and produce repeatable data pipelines that enable less technical data professionals to easily collaborate and work with data in the cloud.

Besides cloud-native architecture and features that uncomplicate the ETL/ELT process, data engineers and data scientists need an ELT platform that will:

- Orchestrate data workflows that ingest data from a variety of sources.
- Support overlapping dataset needs between data analysis and data science applications without creating duplicate data and transformation logic across multiple systems.
- Rationalize data transformation workflows with rapidly changing business logic to support the democratization of data in the lakehouse for everyone in the organization.
- Provide support for modern datasets that are by default shared, secured, and unified.
- Promote collaboration between the two teams, for example when moving data science projects into a production environment.
- Nurture a common language and common skillset to achieve mutually beneficial goals.

Technical capabilities and flexibility are huge considerations when choosing ETL tooling. But the right choice can also enhance team cohesion and collaboration and streamline the creation of common ground between data engineering and data science. The right toolset will develop and nurture a common language for describing and communicating data requirements, helping to reduce misunderstanding and wasted effort and accelerate productivity.

Data teams of all kinds and technical abilities benefit from common, easily transferable skills. A highly visual interface, drag-and-drop components, and a common underlying language, such as SQL, all facilitate faster cross-functional collaboration and deliver more value for your business, faster. Increasingly, organizations cannot afford the time it takes to develop and maintain highly specialized skills like Scala or Java, using niche tools and code intensive solutions. And such solutions are not easily scalable beyond a few key team members. Getting more data workers on board, faster and easier than before, and getting the value of their expertise right away, is a key step in unlocking your data-driven business potential.

Future-proofing your data-driven business with a unified cloud environment

We don't often see fundamental changes in the way we work with data. The relational database was first conceived in 1970 and it has remained the predominant database model since then. While there have been many database-related innovations over the decades, the lakehouse architecture represents one of the biggest fundamental changes in the way we work with data in a long time. And it's fair to say that the lakehouse paradigm is very likely to be a shift that will be with us for a very long time to come.

By adopting a lakehouse architecture, organizations are future-proofing their data needs. Because a lakehouse lets teams manage both structured and unstructured data, it creates greater resiliency when responding to new trends in data. You won't have to worry about increasing data diversity, because structured or unstructured, you can manage it all. By combining structured and unstructured data, you're also less susceptible to data loss. Data recovery and high availability are simpler when all of your data is managed in a unified solution.

These days, a strong data posture has become imperative for improving overall organizational readiness and resilience. Organizations that adopt data lakehouse architectures now, while also addressing the cultural changes needed to succeed, should be well-prepared for the future.

Preparing for the future

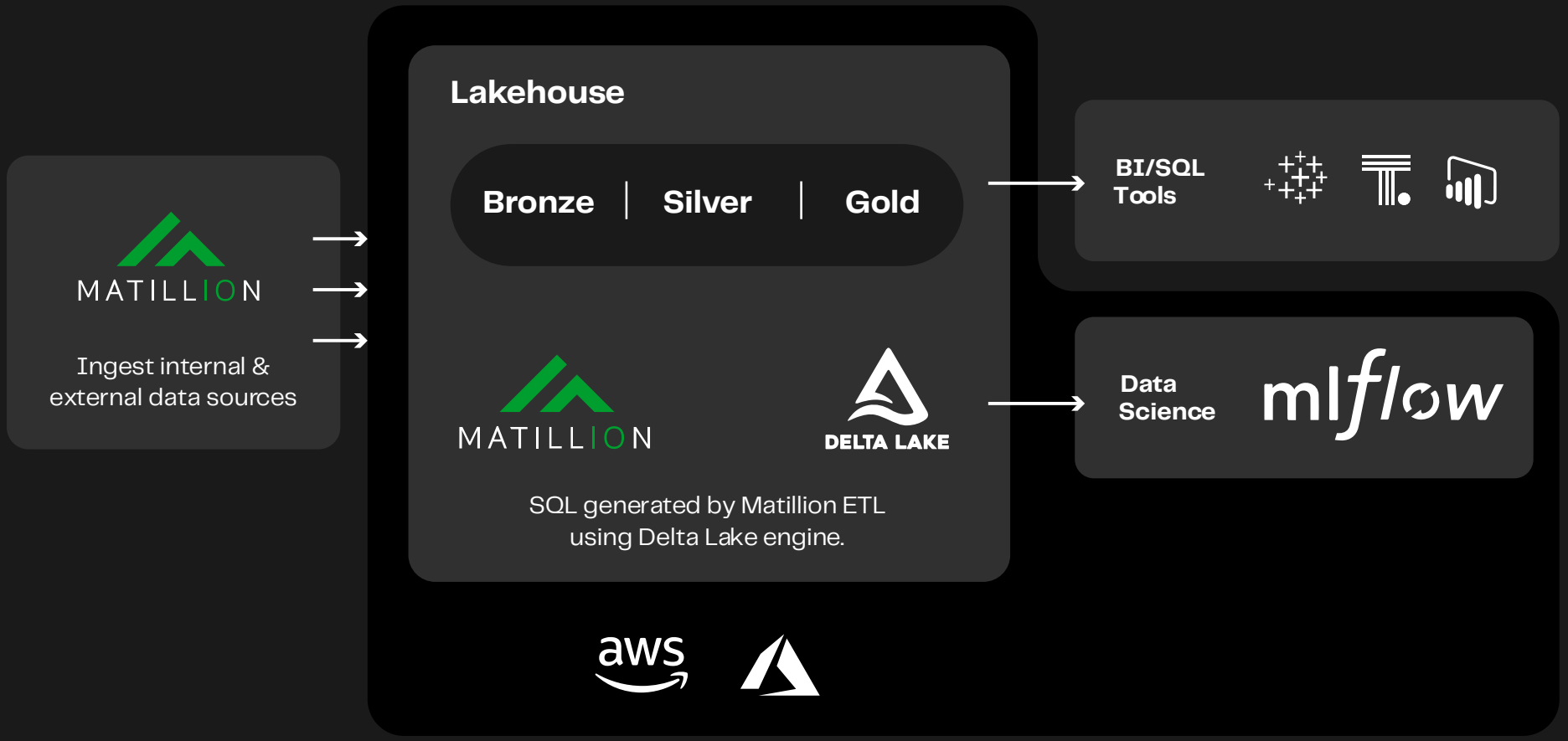
A lakehouse can help propel many of our most strategic initiatives for the future:

Reduced operational costs. A lakehouse can help an organization consolidate its data infrastructure, accelerate data processing, and automate data pipelines, cutting operational costs.

More accurate demand forecasting. Based on patterns and trends discovered through data science, you can build better forecasting models that allow you to better anticipate and respond to peaks in demand.

More personalized customer experiences through accurate recommendation engines. Tailored offerings that are well informed by data lead to delighted customers. That personalization is only possible with the ability to more quickly and easily analyze trends and history across customers and customer segments.

Analytics beyond the dashboard. Modern data teams want to move beyond descriptive reporting (describing the present state) and even predictive reporting (forecasting the future state). Prescriptive reporting, which advises us on possible outcomes and tells us what to do next, is becoming the ultimate goal. In a lakehouse, where data and data practices can be shared across teams, it's possible to build both quality data and data science agility that are essential to prescriptive analytics.



Example Lakehouse architecture using Databricks and Matillion ETL for Delta Lake

Matillion and the lakehouse: Better together

Matillion provides several key features that can help you build out your lakehouse

If the lakehouse architecture can help you bring data engineering and data science together, then Matillion ETL can help ensure that they work productively together. Matillion ETL replaces data prep tools and provides the optimal way to build out your lakehouse and unify your data teams.

Cloud-native ELT helps create a unified approach to data ingestion and enrichment and bridges the gap between data engineering and data science. With the right shared ELT platform, data engineers and data scientists can produce repeatable data pipelines, creating resource efficiency by removing manual processes. With more timely access to data, your data teams can experiment more freely, generate insights, and unlock a new world of business opportunities.

Cloud-native design

Unlike traditional ETL tools, Matillion ETL was born in the cloud and takes advantage of the architecture and compute capabilities of the lakehouse. Its ELT capabilities let you use the full power of the cloud platform.

Low-code UI

Matillion simplifies the creation of a unified analytics platform. With a low-code UI, data professionals of all technical abilities can quickly prepare and cleanse data. This UI also makes it easier to transfer knowledge to new data workers, while experienced data workers can still use and access the underlying complexity should they wish to do so.

Automated data ingestion

Matillion ETL helps automate data ingestion workflows from numerous sources and leverages transformations to create instantly shareable datasets that, for example, align with Databricks design patterns. Data is in the lakehouse and ready to use in minutes, not days. Your highly paid data engineers and data scientists can spend less time getting the data into the lakehouse and more time focused on advanced analysis that benefits the business.

Pre-built source connectors

Save valuable time for your data engineers with Matillion's pre-built data source connectors for common APIs, files, applications, NoSQL, and databases. In addition, the ability to create your own custom connector makes it easy to expand your data reach.

Learn more about Matillion ETL for Delta Lake on Databricks

Matillion ETL for Delta Lake on Databricks allows data professionals to easily extract business-critical data from operational databases, files, NoSQL, and API sources, and then load it into Delta Lake. Matillion ETL has a graphical interface that empowers users across the business to take ownership of data and build out the lakehouse, creating a centralized storage repository. With this graphical interface, users can quickly transform data and prepare it for advanced analytics.

Learn more about Matillion ETL for Delta Lake on Databricks here:
https://www.matillion.com/technology/lakehouse/databricks

1. NewVantage Partners, Big Data and AI Executive Survey 2021, https://c6abb8db-514c-4f5b-b5a1-fc710f1e464e.filesusr.com/ugd/e5361a_d59b4629443945a0b0661d494abb5233.pdf.
2. Anaconda 2020 State of Data Science, https://www.anaconda.com/state-of-data-science-2020
3. Anaconda 2020 State of Data Science, https://www.anaconda.com/state-of-data-science-2020
4. Databricks ABN AMRO case study https://databricks.com/customers/abn-amro
5. Databricks Shell case study https://databricks.com/customers/shell
6. Databricks Regeneron case study https://databricks.com/customers/regeneron
7. Databricks H&M case study https://databricks.com/customers/hm

MATILLION

Visit us at **www.matillion.com** to learn more.