TreeHive Strategy

WHITEPAPER

The Analytics Lake

Donald Farmer, March 2024

Over the last few years, the use of data in business decision-making has changed significantly. When I started working in Business Intelligence in the 1990s, the industry was moving beyond standardized reports to enable exploratory dashboards. Later, rich visualizations were requested by every business. But still, deep data analysis was reserved for specialists like data scientists, quants and actuaries. Today, however, we are advancing rapidly in new directions. What we thought of as mature BI in 2020 now looks like baby steps compared to the technical promise of Artificial Intelligence. At the same time, users across the spectrum are more data-literate and demanding than ever before. This advancement brings its challenges, including stringent data management and compliance requirements. Enterprises must navigate all of these organizational and technological complexities. This paper introduces **Analytics Lake**, an architectural approach developed by **GoodData**.

The Analytics Lake is a data platform that consolidates raw and transformed data, data science models, metadata, and front-end tools. This consolidation of analytics assets into one location makes the data, insights, and tools accessible and usable for both human and automated data consumers.

This architecture adeptly meets the many needs of modern businesses that need to deliver reporting, analytics, data science and AI for a new and demanding generation of users and engineers.

GoodData has been at the leading edge of innovation in analytics for over 15 years. In particular, they were among the pioneers in cloud-native analysis and the first with a purpose-built embeddable platform. As a result, GoodData has had advantages for years in scalability and security, alongside innovations in advanced analytics for the business user. You may think of GoodData as a business intelligence tool, but they are a **data analytics platform**.

The need for change

Before looking at this critical new architecture in detail, it is essential to understand where existing solutions are lacking.

At the most superficial level, the proliferation of data sources and a surge in data volume have led to deteriorating performance and escalating, unpredictable costs. There are many reasons for this intensification of data management problems. Digital transformation has seen many offline processes shift to digital platforms, so even traditional businesses now create more data than ever thanks to eCommerce, social media and IoT. One of the most common complaints I hear from clients is that they move their data to the cloud for efficiency. However, they then need help to get the best value from this new resource, often because they are still taking a traditional approach to analysis in a radically different infrastructure.

More data is a powerful asset for business, and storage today is cheap. However, it's not enough to manage the data: without using it for insight or analysis, it's a wasted asset. But this means more analytics - and more analysts - and thus the greater difficulty. Operational costs can become unpredictable when users can serve themselves to computationally intensive processes. With more users, businesses must navigate the intricacies of data privacy, compliance, and security over a broader and more diverse data landscape where users do everything from basic reporting to application integration and data science over the same sources.

Traditional Data Warehouses and lakes, designed initially for more static and uniform data environments or specialized usage, must be improved in this new role. We have seen other architectural concepts introduced, notably the Data Lakehouse - a hybrid model that combines characteristics of both Data Lakes and Data Warehouses. However, the Data Lakehouse still tends to remain a one-size-fits-all solution and needs help with performance and data management.

Given the different messages from different vendors, it's tempting to muddle through with your data architecture, scaling it here, tweaking it, and deploying new tools everywhere. But the consequences of getting things wrong are severe.

Increased operational costs are a simple and significant consequence, driven by the need for additional storage, complex processing, and specialized personnel.

The consequent inefficiencies can be even more damaging. If you need to catch up with data-savvy rivals, you can't stay caught up on product development, marketing strategies, and customer experience. In short, you can lose the ability to innovate and capitalize on market trends.

For many CTOs, CISOs and CFOs, compliance risks and data security vulnerabilities also weigh heavily on their minds. In some verticals, the potential for hefty fines, legal issues, and reputational damage is as troubling as data breaches and cyber-attacks.

The Analytics Lake is an essential contribution to these demanding concerns. Let's not pretend it is a complete or perfect answer in itself. You still need sound policies, best practices, training and skills to run an adequate data infrastructure. But if you see yourself, in Tom Davenport's memorable phrase "Competing on Analytics," you must also get the analytic architecture right.

I have always believed that adopting a new architecture must be a decision made with a strong sense of business purpose. I am wary of taking a new approach only to save money or to scale more quickly. The question should be not "What can we do better?" but "What can we do with our new architecture that we couldn't do before?"

With that in mind, let's consider what can be done with the Analytics Lake ...

The Analytics Lake

It would be a mistake to think of the Analytics Lake only as a marketing variation on the Data Lake or Data Lakehouse. Its purpose is very different.

As their names suggest, the basic principles of the Data Warehouse and the Data Lake focus on data, particularly storage for long-term use.

Data Warehouses offered a centralized repository, specifically structured, designed and modeled for efficient business reporting and decision-making. Storage and compute resources are optimized to support data consistency and reliability with efficient business reporting over predefined models.

Data Lakes enabled the storage of vast amounts of raw data in whatever native format the data arrived in, regardless of structure and modeling. Storage is optimized for unparalleled scalability at low cost, and compute resources are optimized for data scientists who require distributed computing over raw, unstructured data.

Conversely, the Analytics Lake focuses on analytics and machine learning, not long-term data storage. In the Analytics Lake, data is a means to an end, not an end in itself. There are some critical implications following this approach:

1. **Optimized Storage and Compute for Analytics and Machine Learning:** The Analytics Lake is precisely engineered to optimize [1] [DF2] analytics and machine learning operations. This optimization ensures that users can efficiently process large datasets, run complex algorithms, and extract insights in real-time without the overhead of modeling or distributing data for storage. Architects need help choosing highly distributed or centralized data, so the Analytics Lake cuts through that decision for analytic use cases.
2. **Integrated Metadata Modeling and a Headless Semantic Layer:** A notable feature of the Analytics Lake is the integration of metadata modeling and a headless semantic layer within the lake's architecture. This design enables seamless mapping and interpretation of diverse data sources. The headless semantic layer provides a flexible, unified view of data across the organization, enhancing data discoverability and usability while maintaining data integrity and consistency. There's no redundancy in this approach and no inflexible metamodel.
3. **APIs for Analytics and Data Engineers:** For analytics and data engineers, the Analytics Lake provides robust APIs that support standard software engineering practices including CI/CD and code-based automation, along with hundred of Python libraries and tool,. This analytics-as-code architecture is new and energizing in Business Intelligence, but it is the natural workflow for engineers: streamlined, efficient, and familiar. With this approach, engineers can directly integrate data operations in the Analytics Lake into their existing development pipelines. An API-driven approach promotes collaboration, version control, and a higher degree of automation in data operations, aligning with modern software development methodologies.
4. **No-Code/Low-Code Environments for Analysts and Consumers:** Not all users are engineers. So, the Analytics Lake integrates intuitive no-code and low-code environments. These user-friendly interfaces allow analysts and business users to interact with data, build models, and generate insights without programming knowledge.

5. **Comprehensive BI and Visualization Environment:** Besides this integrated capability, the Analytics Lake features a robust BI and visualization environment. This component is crucial for translating complex data into detailed reports, dashboards, and visualizations. It makes it easier for decision-makers and stakeholders to interpret the data and enables the Analytics Lake to play a full role in standardized reporting and innovative exploration. What's more the BI artifacts themselves are stored as objects in the lake, making them available for use by other systems alongside the data objects.

The emphasis on APIs and no-code/low-code is notable because it reflects an essential change in the analytics market. Previously, analytics solutions were developed mainly by specialists, for example, database programmers rather than application programmers. Today, coders, engineers, and even designers from every field are working more closely and purposefully with data than before. APIs and design tools are critical, but they also need to create a new class of artifact: the data product.

A Composable Data Service Layer

Business users increasingly ask developers to deliver data products for consumption, and the Analytics Lake enables the most compelling scenarios: full-code applications, low-code machine learning and AI, and no-code dashboards with augmented insights.

You can regard the core platform offering as a *Composable Data Service Layer* designed around Apache Arrow (see the sidebar **Beneath the Surface**), forming a unified and adaptable layer of data services. The various services within the layer are interconnected, allowing them to automatically communicate and retrieve necessary data from each other based on client requests without requiring extra integration efforts.

As you can see, there's a lot to this, but besides these technical capabilities, there are significant business advantages.

Business benefits of an Analytics Lake

Engineers will love the Analytics Lake for optimizing the many workflows and processes they must manage. But elegant engineering on its own is rarely a compelling business proposition. However, as business users might say, Analytics Lake offers some significant bottom-line advantages.

- **Improved responsiveness for the Metrics Layer through Integrated Analytics-Ready Data**
    - A metrics layer standardizes business metrics' definition, calculation, and usage across an organization. It ensures consistency and accuracy by providing a single source of truth for all, enhancing collaboration and consistency across different platforms. By consolidating analytics-ready data on the same platform as the metrics layer and data products, performance is significantly improved. This integration facilitates faster data processing and more efficient analytics operations. It eliminates the latency often encountered when data resides across multiple platforms, ensuring quicker access to analytics-ready data. The proximity of data to the analytics tools also enables a more rapid generation of insights. If that sounds too technical, what it means for every user is a more responsive, high-performance analysis of all kinds of data, meaning that decision-making can be informed with the most up-to-the-minute data.
- **Increased Cost Efficiency with Optimized Data Processing**
    - This system's streamlined approach to data processing leads to a lower Total Cost of Ownership (TCO) for data solutions. You may substantially lower operational costs by reducing the data processing required in your cloud Data Warehouse and minimizing data movement (both are more expensive than most businesses realize). As a result, companies can achieve a more cost-effective and sustainable model for handling their data needs.
- **Commitment to Openness with Open-Source Technologies**
    - GoodData's embrace of open-source technologies and an analytics-as-code approach marks a significant step towards integration and flexibility in your data ecosystem. This openness ensures compatibility and ease of integration with numerous tools and platforms. Interoperability with existing tools and future-proofing are critical considerations for IT procurement. By leveraging open-source technologies, the platform supports innovation and customization. It aligns with modern data practices, offering a robust, adaptable solution for various data analytics needs and familiar tools and platforms for engineers.

- **Access to data in the formats developers and data scientists need**
    - The ability of GoodData to deliver data as data frames is vital for data scientists, as data frames are a fundamental tool in data analysis, offering a familiar, tabular format to efficiently handle large datasets, perform complex data transformations, and efficiently conduct statistical analyses. For developers of business applications, integration with MotherDuck (an enhancement of the in-process OLAP database DuckDB) significantly amplifies the potential for advanced analytics.

- **Robust Governance for High-Quality Data Across Applications:**
    - Developing a comprehensive, metrics-driven data platform is central to ensuring high data quality for all downstream uses, including Business Intelligence (BI), Artificial Intelligence (AI), and custom applications. The Analytics Lake ensures that the data utilized in BI analyses, AI models, and bespoke applications is reliable and trustworthy. Analytics is too often the weak component of the enterprise governance framework. This architecture not only enhances the quality of insights derived but also reinforces users' confidence in the data and the decisions based on it.

Beneath the surface ...

In the movies, something mysterious always lies under the lake's surface. Let's look at what enables the Analytics Lake to deliver its impressive power - the FlexQuery analytics cache. It's remarkable, but no mystery.

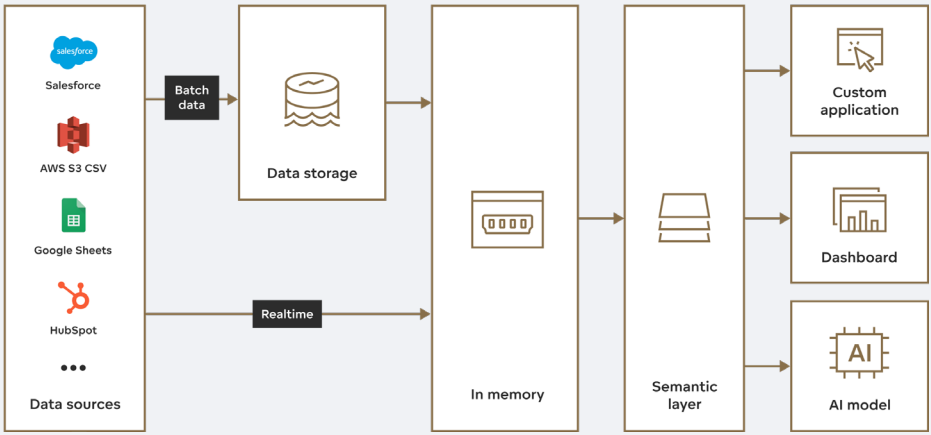
GoodData built FlexQuery on the cutting-edge in-memory Apache Arrow open-source project. Arrow's columnar storage format is at the core of FlexQuery's operation, enabling the efficient handling and processing of large volumes of data. Apache Arrow is a platform for in-memory analytics. This format is optimized for high-speed analytics and machine learning rather than just for storage, thus aligning perfectly with the needs of a modern analytics platform like GoodData's Analytics Lake.

There are numerous benefits to this architecture.

- It allows for advanced analytics on diverse data formats, such as CSV or Parquet, enhancing the system's flexibility.
- Arrow is a widely supported format integral to numerous technologies and provides a far-reaching infrastructure to meet diverse technical needs, acting as a gateway to a rich ecosystem of data and analytics libraries and tools.
- Apache's Arrow Database Connectivity (ADBC) offers performance improvements when working with various data source types, including Snowflake and PostgreSQL. Analytics Lake can, therefore, interact seamlessly with different databases and Data Warehouses you may already have, saving the cost and hassles of migration.
- Another critical aspect is enabling analytics on top of lake houses like Iceberg and Delta and providing SQL interfaces to the GoodData platform. These capabilities ensure that FlexQuery is not just a query engine but a comprehensive analytical tool that can operate in various data environments.
- And, of course, high-performance caching is also critical. It enhances the system's ability to rapidly respond to complex queries, which is necessary for near-real-time analytics, exploratory analytics and data science experiments.

With FlexQuery, the Analytics Lake can seamlessly process data from different sources, whether traditional databases, real-time streams, or IoT sensors: a versatility that not only simplifies data integration but also enhances the overall utility of the Analytics Lake. And, while caching is powerful, it cannot be the only answer for high-performance analytics. So, FlexQuery employs sophisticated algorithms to streamline query processing, reducing the computational load and improving response times. This optimization is particularly beneficial when dealing with complex analytical operations that require aggregating, filtering, or transforming large volumes of data.

Finally, scalability matters too. GoodData designed FlexQuery so that as data volumes grow and query demands become more intricate, FlexQuery scales accordingly, ensuring that the Analytics Lake's performance remains consistently high.



FlexQuery schema

AI and the Analytics Lake

Naturally, even more than BI, AI is top of mind for many executives today. However, BI remains critical to the governance and management of their business. For these scenarios, the Analytics Lake can uniquely transform artificial intelligence (AI) deployments while enabling better BI.

The analytics lake acts as a repository containing data, data transformations, analytics models, visualizations, and descriptive metadata. The consolidation of these assets enables AI services to discover and retrieve any component as needed to generate insights. As AI capabilities advance, the components can be leveraged in innovative combinations; AI might even generate new assets within the lake.

This robust combination of solutions, engineered to address the nuanced demands of AI, addresses the key data challenges inherent in AI implementations, accelerating time-to-insight and enhancing model accuracy.

The Analytics Lake will catalyze AI deployments, acting as a robust and integrated data solution that establishes a single source from which to draw both BI reporting and AI algorithms. It acts as a high-performance cache layer that significantly reduces data retrieval times, ensuring that AI models are fed a continuous data stream with minimal latency. This cache is vital for real-time AI applications that rely on swift data analysis to make immediate decisions, such as fraud detection systems or dynamic pricing models.

Large Language Models (LLMs) promise to revolutionize our work with data and visualizations. But for LLMs to achieve this, they need much more business context for each scenario. The semantic layer, with its rich metadata about content and relationships between objects, can provide this context for each business uniquely, reducing errors and hallucinations.

Another area of innovation is the automated generation of insights and entire dashboards with minimal user input.Soon, classical static dashboards will be outdated as new AI capabilities enable more intuitive, dynamic experiences. By simply describing the critical business question or metrics of interest in natural language, AI will soon be capable of producing a fully functional dashboard customized to that goal. Relevant charts can be automatically selected to display the data, alerts configured to notify users of changes, and interactive data stories generated to guide analysis.

Beyond basic dashboard assembly, at GoodData, we see LLMs providing enhanced context around the raw data, suggesting optimal chart choices, summarizing key trends in writing, explaining anomalies, identifying relationships between metrics, and improving the dataset by merging supplemental information from external sources. Users can have natural language conversations to uncover more profound meaning with their data.

As these AI capabilities take hold, we will fundamentally change how we build and utilize dashboards. Static pixel-perfect dashboards will transition to flexible canvases with contextual insights that adapt to users' needs. Data analysis will become a conversation, with LLMs actively collaborating with users to uncover insights. Rather than merely visualize data, tomorrow's experiences will reveal the stories, meanings and predictions buried within.

These scenarios are, of course, sensitive to compliance and trust: the Analytics Lake ensures that data governance, security, and quality are consistent across the board, an imperative for the success of any AI initiative.

The road ahead leads to the lake

I started this paper looking back to the roots of business intelligence as a better form of reporting: more interactive, visual and engaging. Now, those dashboard applications are as much a commodity as reports: they are still super helpful for day-to-day business but don't offer much differentiation or opportunities for innovation.

Today, businesses are looking for an advantage in data science and, of course, in AI. However, these approaches are very demanding regarding the volume of data, the speed of analysis and the need to interoperate with the numerous platforms and applications that integrate with your data throughout the business process. It's not enough for AI and analytics to be an isolated process: the greatest value will be found when embedding AI into the daily workflow. And it's not enough for analytics to be a specialist practice: analytics-as-code integrates analytics development into their existing development methodologies.

This is why the road ahead leads to the analytics lake. The new analytics and the unique demands of governance all demand storage, metadata, performance, APIs and integration that the Analytics Lake can uniquely deliver.

Read more about GoodData's solution here and their vision here.

Comparison Table

	Data Warehouse	Data Lake	Analytics Lake
Primary Focus	Business reporting and decision-making	Scalable raw data storage	Analytics and machine learning
Storage Format & Optimization	Structured data optimized for query performance	Optimized for low-cost storage of large raw, unstructured data	Optimized compute for ML workflows
Metadata & Semantic Layers	Typically has integrated business metadata	Limited metadata and semantics	Integrated metadata and headless semantic layer

	Data Warehouse	Data Lake	Analytics Lake
BI & Visualization Capabilities	Full support for BI workflows and visualizations	Minimal BI support	Integrated BI and visualizations environment
Advance Analytics Support	Basic predictive modeling capabilities	Strong support for advanced analytics	Optimized for advanced analytics and ML
Data Science & ML Support	Limited support for data science workflows	Supports data science through big data tools	Tailored for data science and ML with compute optimization and APIs
Compliance & Governance	Strong auditing, security, and governance capabilities	Limited governance capabilities	Robust governance through metadata integration
APIs & Programmatic Access	Traditionally accessed through SQL	Access via big data APIs and notebooks	APIs designed specifically for analytics engineers
Cost Efficiency	Predictable but relatively high storage costs	Very low-cost storage but unpredictable analytics costs	Optimized processing reduces overall costs
Flexibility & Future Proofing	Schema-on-write limits flexibility	Flexible but typically requires migration for BI use	Designed to interoperate with multiple data platforms

	Data Warehouse	Data Lake	Analytics Lake
Key Strengths	Performance, consistency, reliability	Scalability, cost efficiency for storage	Purpose-built for advanced analytics and ML
Key Limitations	Limited flexibility and ability to handle messy, large or streaming data	Challenging to apply governance, reuse data for BI and reporting	Emerging architecture

Donald Farmer is a seasoned data and analytics strategist, with over 30 years of experience designing data and analytics products. He has led teams at Microsoft and Qlik Technologies and is now the Principal of TreeHive Strategy, where he advises software vendors, enterprises, and investors on data and advanced analytics strategy. In addition to his work at TreeHive, Donald serves as VP of Innovation and Research at Nobody Studios, a crowd-infused venture studio.

Donald has a passion for innovation and has worked on some of the leading data technologies in the market, as well as in award-winning startups. He has a diverse background, having worked in fish farming, archaeology, and forestry, but data has always been at the heart of his work.

Donald is an independent advisor to clients globally, specializing in innovation, productization, and analytics strategy. He advises clients ranging from the world's largest software vendors to small startups and investors in both public and private markets. Donald is a prolific writer and teacher, and he speaks to audiences worldwide on the subjects of data analysis and innovation.