tcgdigital

The Data Lakehouse:

A Foundation for Scaling AI-Driven Innovation and Enterprise-wide Value Realizatio**n**

This paper explores the components, features and benefits of the data lakehouse technology in the context of enterprise class AI deployments across organizations, emphasizing its role in enhancing data accessibility, integrity, and collaboration. It also highlights how the tcgmcube platform can be leveraged to accelerate the value impact of AI solutions across organizations.

Table of Contents

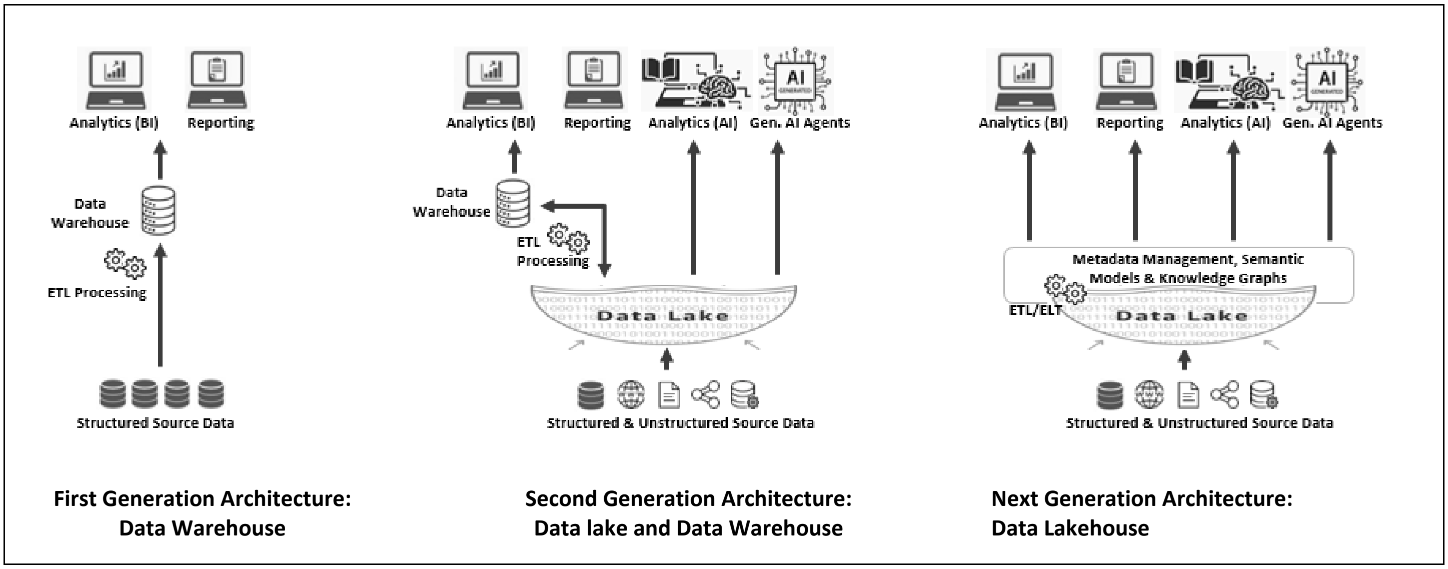
The need for the Data Lakehouse

In the era of big data, advanced analytics and AI, the need for efficient data management systems becomes critical. Traditional data warehousing and data lake architectures have their limitations, particularly in navigating through diverse and voluminous datasets, making it extremely difficult for users to get to relevant, contextualized data. Traditional data architectures suffer from these problems:

- Data Accessibility: Running analytical queries on large and diverse datasets is challenging and it becomes extremely difficult for users to find and get contextualized data out. This also means that the existing architecture can only provide limited support for advanced analytics and AI, as these algorithms need to process large datasets using complex querying.
- Collaboration Bottlenecks: Lack of a shared, unified and contextual data view causes challenges for team collaboration across the organization, often leading to redundant data acquisition and data management activities. In most cases, the data does not adhere to the FAIR (acronym for Findable, Accessible, Interoperable, and Reusable) principles and hence does not allow users to exploit the full potential of the data.
- Data Integrity Issues: Keeping the data lake and data warehouse consistent is difficult and costly because of redundancies. Lack of a semantic layer impacts analysis integrity.

The concept of a data lakehouse, which integrates the best features of both data lakes and data warehouses, and adds a semantic layer for contextualization emerges as a compelling solution.

A data lakehouse is an open data management architecture that combines the flexibility, cost-efficiency, and scale of data lakes with the data management capabilities of data warehouses. It enables dashboarding, traditional AI, generative Ai and AI based applications on accessible and transparent data.



First Generation Architecture: Data Warehouse

Second Generation Architecture: Data lake and Data Warehouse

Next Generation Architecture: Data Lakehouse

Figure 1: Evolution of the data architectures for decision support

Key Components of the Data Lakehouse

The following are the core components of a holistic data lakehouse strategy. The technology helps elevate the data strategy of organizations and accelerates velocity to value across the value chain:

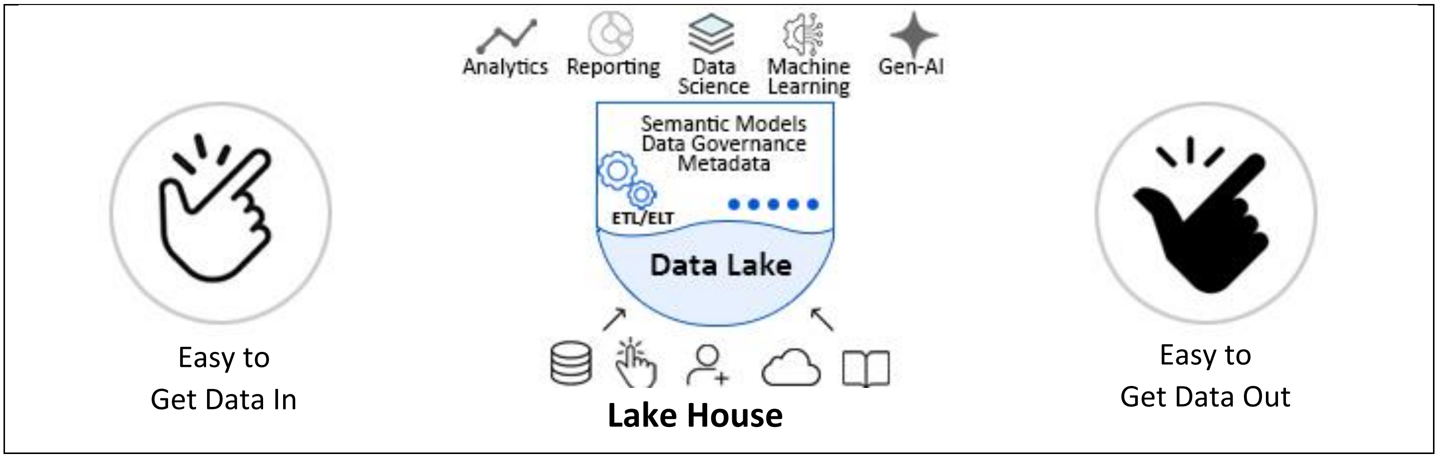
- **Data ingestion (Easy to get data in):**

  The data lakehouse makes it "easy to get data in", coming with pre-built standard connectors to various systems & instruments, supporting both real time and batch ingestion, and providing features for data transformations at various stages. The overlay of a semantic layer enables data ingestion processes utilize the semantic definitions. Knowledge graphs can integrate data from various sources, including structured, semi-structured, and unstructured data, and help create a cohesive representation of information stored in the lakehouse.

- **Data leverage (Easy to get data out):**

  The data lakehouse comes with robust data management features. The business metadata management is powered by knowledge graphs, providing ontology management and knowledge modelling capabilities. It adheres to the FAIR principles (i.e. makes data Findable, Accessible, Interoperable, and Reusable), thus making it "easy to get data out".

  - By defining semantic relationships and hierarchies between data entities, knowledge graphs provide rich domain context that enhances data understanding and usability. This allows users to navigate through data based on relationships rather than just relying on raw data of technical data dictionaries.
  - Connecting the Semantic Layer to the Analysis layer allows the use of contextualized semantic business terms for analytics. It enables efficient querying of data in natural language and providing contextual responses that are easy to use, understand & interpret.
  - Knowledge graphs can enrich data by linking it with external datasets or ontologies, providing additional context that can improve analysis and insights.



The Reference Architecture for the Data Lakehouse

The reference architecture of the data lakehouse platform, in the context of the end to end analytic/AI platform, shows the various components described above in more detail.

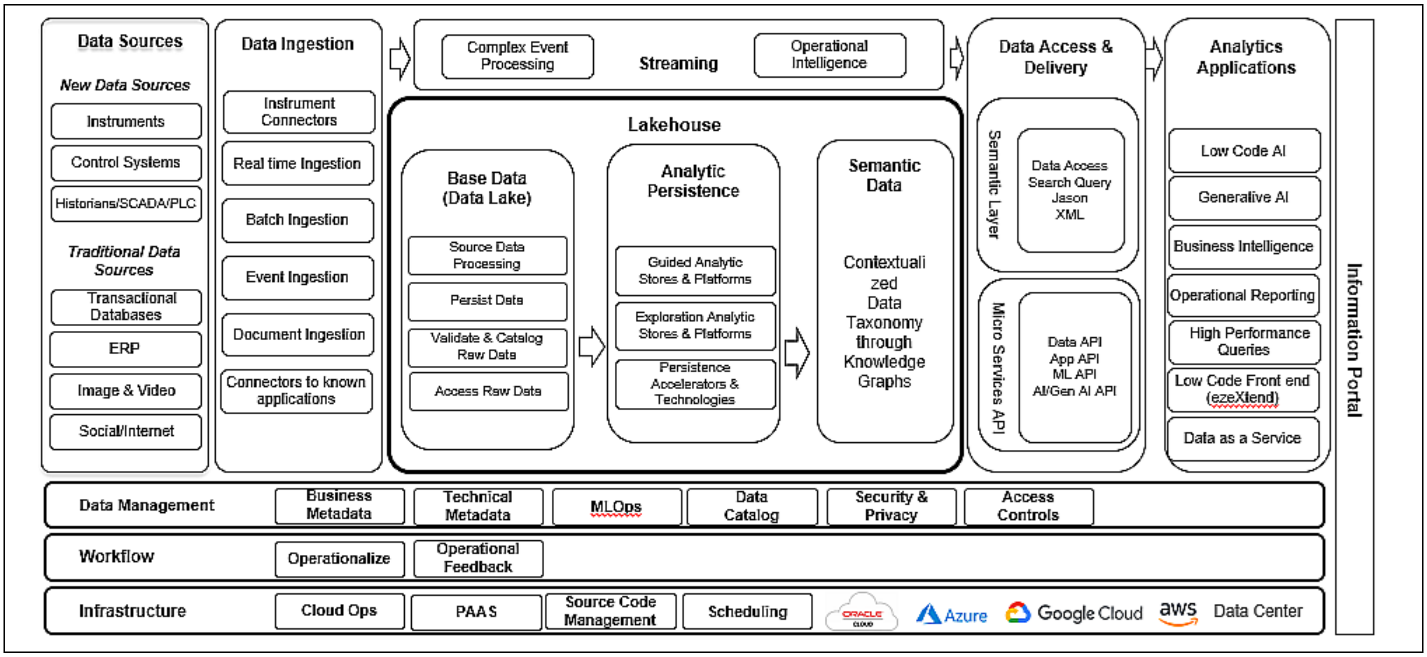


Figure 2: Reference Architecture for the data lakehouse

This reference architecture attempts a comprehensive and complete view of all possible components that can contribute to a Data Lakehouse implementation. Depending on the scope, type of data and the analytical processes that need to be supported, your mileage might vary in terms of functionality and required elements.

Benefits of the Data Lakehouse

- **Accessibility:** Facilitates actionable insights and analytics by ensuring that users have easy access to the right data at the right time through the right user interface.
- **Collaboration**: Enables teams to work together more effectively by providing a shared view of data across the organization.
- **Integrity:** With better data management practices, version control and semantic consistencies, data lakehouses enhance the analysis integrity. To sustain the impact and analytical integrity, the lakehouse provides the ability to manage model drift and data drift seamlessly within the boundaries of the contextual model.

The data lakehouse architecture presents a transformative approach to data management and helps foster a data-driven culture across the organization. By bridging the gap between data lakes and data warehouses, it provides users with the tools necessary for efficient data accessibility, collaboration, and integrity. As the various user communities continue to generate vast amounts of data, the adoption of data lakehouses will likely play a pivotal role in advancing innovation.

tcgmcube: taking the Data Lakehouse to the next level

Leveraging our end-to-end AI platform, tcgmcube, organizations can create robust data lakehouses, with the aim to streamline data management by integrating various data processing and analytics needs into one architecture. This approach helps avoid redundancies and inconsistencies in data, accelerates analysis throughput, and minimizes costs.

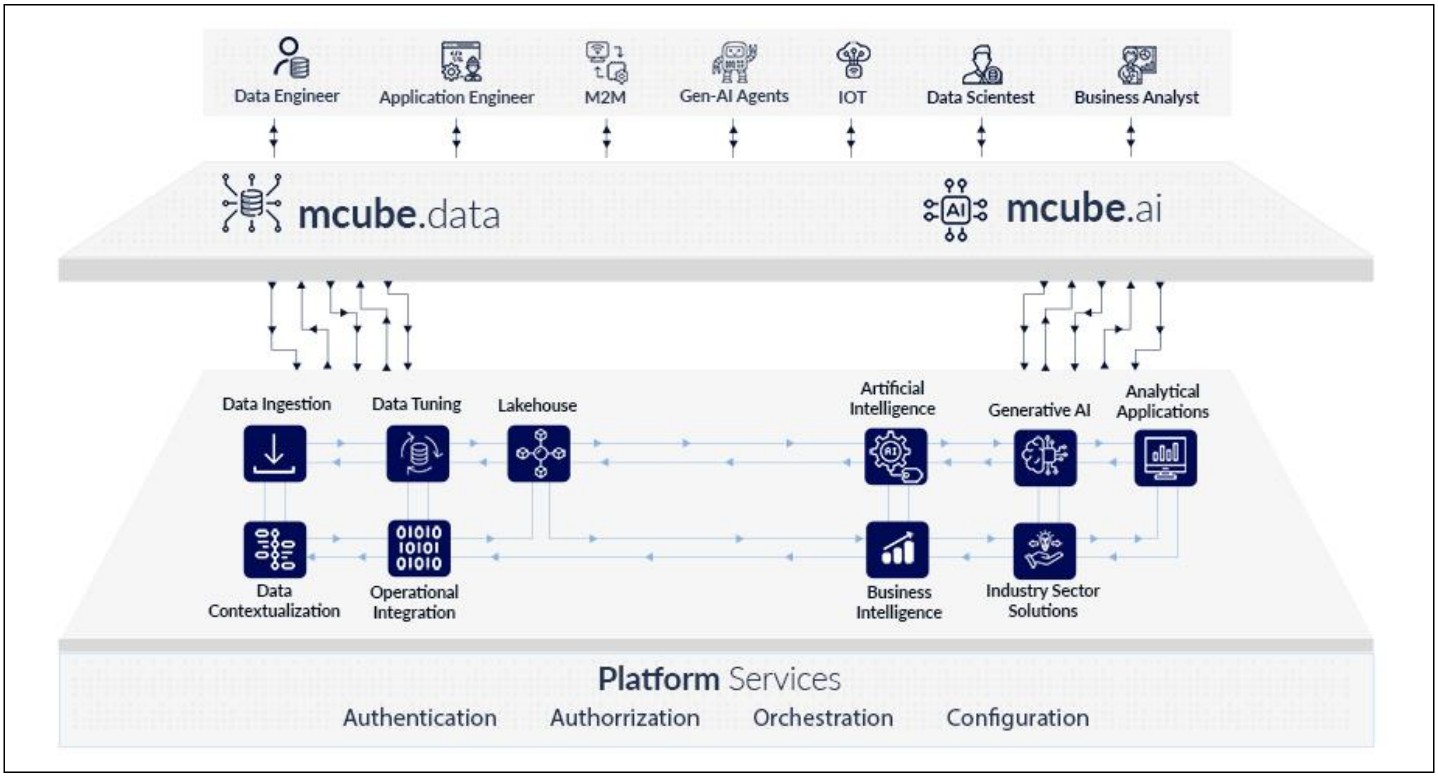


Figure 3: Functional Overview of tcgmcube

The platform tcgmcube comes with mcube.data and mcube.ai, thus providing advanced analytics and AI capabilities and data management on the same platform managed by common platform services. This makes it an extremely powerful platform for implementing the lakehouse and deploying analytical and AI applications on top of the lakehouse.

mcube.data

mcube.data is designed to handle the complexity of today's data landscape covering structured, semi-structured and unstructured data in real-time, near real-time and batch environments. It comprises of data ingestion, data storage and data management features and comes with a semantic layer powered by knowledge graphs.

- The Data Ingestion Layer of tcgmcube comes with pre-built standard connectors to various systems & instruments. It is highly interoperable and has pre-built connectors to instruments such as balance, DNA sequencer, gas chromatographs, pH meter, thermo-cycler, titrator, etc. It supports real time data ingestion as well as batch ingestion, providing features for data transformations at various stages.

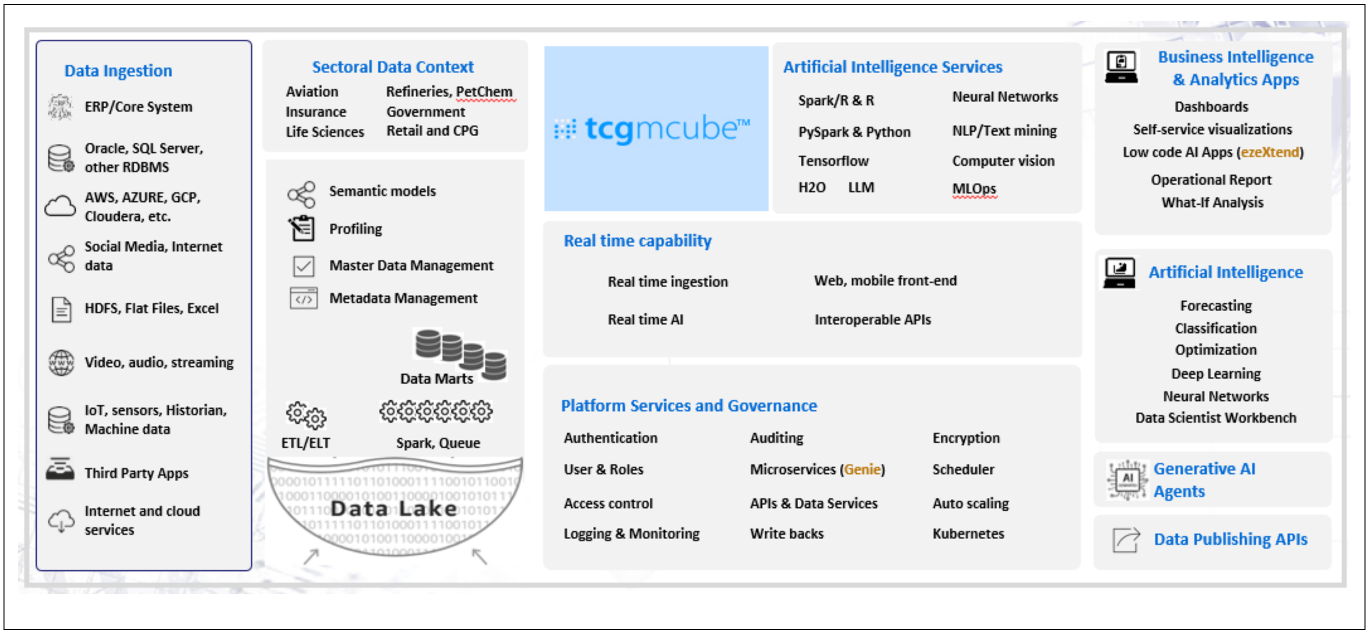
- For real time AI (e.g. manufacturing 4.0 scenarios), the lakehouse supports data collection and data management at the edge before data gets transferred to the central lakehouse. This process handles network interruptions and other unforeseen events through data caches and synchronization capabilities. Built-in connectors of tcgmcube have been designed to support both OPC UA and OPC DA protocols, enabling the ingestion of near real-time tag data.
- The versatile data storage layer of tcgmcube comes with robust data management features. It leverages ontology management and knowledge modelling capabilities, making it “easy to get data out” and has the following layers:
    - **Base data layer** for source data processing, providing features to validate and catalogue the raw data
    - **Analytic Persistence layer** with processed datasets for optimizing analytical queries and AI driven processes
    - **Semantic Persistence Layer** with contextualized data taxonomy through knowledge graphs

mcube.ai

mcube.ai comprises advanced multi-modal AI integrating deep learning, traditional AI, computer vision, private LLM’s enhanced with knowledge graphs. It has a repository of 1000+ AI algorithms, supporting real-time AI and comprehensive model management. The analysis layer is powered by the semantic layer that makes it “easy to get data out” for analysis needs as it provides deep ontologies for domain contextualization. This block provides:

- **Traditional AI at scale** with a wide assortment of statistical, machine learning, deep learning, and optimization algorithms
- **Comprehensive Generative AI algorithms** covering traditional LLM (private and public LLMs for text data), multimodal LLM (to include image data) and RAG models for fast information retrieval and complete source traceability
- **Insights dissemination layer** providing multiple user interfaces such as dashboards with easy business user self-service, operational reports and low-code “upgrade safe custom screen painting”. The analytical applications and dashboards leverage the semantic layer for data interpretation and reporting
- **Action dissemination layer** providing inputs to automated operational processes such as alerts, recommendations, action triggers, etc.

The Platform Services and Governance layer of tcgmcube helps implement enterprise class governance practices to ensure data quality, security, and compliance.



Conclusion

The need for a holistic approach

Establishing a data lakehouse is not a value proposition on its own. It is the analytical processes and applications that it supports, which determine the actual value impact to the organization. It is therefore crucial to keep use cases and business processes that need to be optimized in mind when starting the build out of a data lakehouse. Data need to be organized in fit for purpose data structures to balance cost and performance. Refresh cycles, real- or right-time requirements determine the approach to ingestion processes and the analytical/AI based result delivery processes to humans and other applications drive the approach to integration. Only a holistic approach and a technology platform which allows for the required flexibility and integrated approach between the data lakehouse the analytical/AI based processes and applications can provide the speed and agility to minimize time to value.

The impact of tcgmcube

As end to end data and AI/GenAI platform tcgmcube is designed from bottom to the top to conquer the ever changing needs of organizations, which are embarking onto the journey of their digital transformation. The functional components within mcube.data and mcube.ai cover the breadth of capabilities needed for accelerated deployment cycles of traditional AI and generative AI driven applications and business processes. The underlying platform services allow for enterprise class management, monitoring, and compliance. As business and AI innovation cycles accelerate, tcgmcube, the open, interoperable, but strongly governed platform is able to deploy the analytical applications businesses need today at the speed required: Be it to reduce time to value for customers or improve the EBIT of their business processes running in the back or front office.

TCG Digital is the digital & AI arm of The Chatterjee Group (TCG), a multi-billion dollar conglomerate with a diverse portfolio in Pharmaceuticals, Biotech, Petrochemicals, and Real Estate across the US, EU, and South Asia. Our umbrella includes companies such as LabVantage, Lummus Tech, and TCG LifeSciences. At TCG Digital, we are driven by our mantra of delivering "Velocity to Value", helping enterprises transform faster and smarter. Our AI Analytics platform tcgmcube is at the heart of these transformations. We enable organizations unlock the full potential of their data, and by seamlessly integrating AI/ML capabilities into their business processes, we empower businesses to accelerate their digital transformation journey, enhancing agility and driving impactful results.
https://www.tcgdigital.com

tcgmcube is TCG Digital’s flagship Data, AI and Analytics platform. Built with a domain driven design at the cross-roads of industry knowledge and digital prowess, our architecture is designed to handle the most disparate data landscapes with AI 2.0 being at the heart of it combining powerful and advanced models to solve the most complex business problems. The platform integrates mcube.ai and mcube.data, delivering AI capabilities and data management seamlessly through unified platform services.
https://www.tcgdigital.com/tcg-mcube/